The genome and DNA puff sequences of the fungus fly, Sciara coprophila, and genome-wide methods for studying DNA replication.

by
John M. Urban
B. S., William Paterson University, Wayne, NJ 2010

A dissertation submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy in the Department of
Molecular Biology, Cell Biology, and Biochemistry at Brown University

Providence, Rhode Island
May 2017

This dissertation by John M. Urban is accepted in its present form by
the Department of Molecular Biology, Cell Biology, and Biochemistry as satisfying the dissertation
requirement for the degree of Doctor of Philosophy.

Date _____ _____
Dr. Susan A. Gerbi, Thesis Advisor

Recommended to the Graduate Council

Date _____ _____
Dr. Erica Larschan, Reader

Date _____ _____
Dr. David Mark Welch, Reader

Date _____ _____
Dr. Benjamin Raphael, Reader

Date _____ _____
Dr. Mark Johnson, Reader

Date _____ _____
Dr. David MacAlpine, Outside Reader

Approved by the Graduate Council

Date _____ _____
Dr. Andrew G. Campbell
Dean of the Graduate School

Abstract of "The genome and DNA puff sequences of the fungus fly, Sciara coprophila, and genome-wide methods for studying DNA replication." by John M. Urban, Ph.D., Brown University, May 2017.

DNA replication initiates from any given "origin" at most once per cell cycle to ensure perfect duplication of the hereditary material. DNA re-replication occurs when an origin initiates more than once. This can lead to DNA damage, gene amplification, and genomic rearrangements, all hallmarks of cancer. Whereas it is difficult to study re-replication in cancer cells, the fungus fly Sciara coprophila is a tractable alternative where re-replication occurs in the DNA puffs of giant polytene chromosomes in larval salivary glands during normal development. The DNA sequence for only one of many DNA puffs was known, preventing identification of shared motifs. Progress has also been thwarted by the absence of a genome sequence, necessary for genomics approaches. In this thesis, I present the first draft of the Sciara coprophila genome sequence, assembled using Illumina short reads, long single-molecule reads from Pacific Biosciences and Oxford Nanopore, and optical maps from BioNano Genomics. Using the Oxford Nanopore MinION, I established protocols to obtain high quality, ultra-long reads that exceed 100 kb in some cases, which were important in assembling multi-megabase contigs. RNA sequencing of the transcriptomes of both sexes across the lifecycle facilitated annotation of genes hidden in the genome sequence. High throughput sequencing of the salivary gland genome throughout the DNA re-replication stages identified multiple DNA puff sequences, which were confirmed with qPCR, and in some cases mapped by Fluorescence in situ Hybridization (FISH) to corresponding DNA puffs. Mapping the origins where DNA re-replication begins is still needed to locate cis-regulatory elements. Toward this goal, I have refined the genome-wide origin mapping technique Nascent Strand sequencing (NS-seq) by inclusion of important controls and found that nucleosomes are phased around a subset of origins near G-quadruplex motifs in the human genome. Moreover, I describe my progress in developing new single-molecule, genome-wide origin mapping techniques. Overall, this thesis presents the genome and DNA puff sequences of Sciara coprophila as well as methods to map the re-replication origins, thereby establishing a foundation to unravel the mechanism of site-specific DNA re-replication.

Vita

John Michael Urban

185 Meeting Street
Providence, Rhode Island 02912
Phone: (973) 919-2844
E-mail: john_urban@brown.edu
E-mail: mr.john.urban@gmail.com

Education

Brown University Providence, Rhode Island	*Doctor of Philosophy, Defended 2016* Genomics Track in Molecular Biology, Cell Biology, and Biochemistry Research on mapping DNA replication origins in the human genome, developing new genomic and single-molecule approaches to mapping origins, genome assembly, and DNA amplification in *Sciara coprophila*
William Paterson University Wayne, New Jersey	*Bachelor of Science, 2010* Honors Biology and Psychology Summa Cum Laude, William Paterson University Research on DNA replication in *Trypanosoma brucei*
University of Cambridge Cambridge (U.K.)	*2009* Summer research in bioinformatics

Dissertation

The genome and DNA puff sequences of the fungus fly, *Sciara coprophila*, and genome-wide methods for studying DNA replication.

Thesis Advisor: Dr. Susan A. Gerbi, PhD.

Publications

Foulk MS*, **Urban JM***, Casella C, Gerbi SA, (2015) Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. Genome Research 25:725-35. PMC4417120 (* signifies co-first author).

Urban JM, Foulk MS, Casella C, Gerbi SA. (2015) The hunt for origins of DNA replication in multicellular eukaryotes. F1000Prime Rep 7:30. PMC4371235

Urban JM, Bliss J, Lawrence CE, Gerbi SA. (2015) Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. bioRxiv. http://dx.doi.org/10.1101/019281

Ip, C. L., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., Leggett, R. M., Eccles, D. A., Zalunin, V., **Urban, J. M.**, Piazza, P., Bowden, R. J., Paten, B., Mwaigwisya, S., Batty, E. M., Simpson, J. T., Snutch, T. P., Birney, E., Buck, D., Goodwin, S., Jansen, H. J., OGrady, J., Olsen, H. E., Consortium, M. A., Reference, Ip, C. L., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., Leggett, R. M., Eccles, D. A., Zalunin, V., Urban, J. M., Piazza, P., Bowden, R. J., Paten, B., Mwaigwisya, S., Batty, E. M., Simpson, J. T., Snutch, T. P., Birney, E., Buck, D., Goodwin, S., Jansen, H. J., OGrady, J., Olsen, H. E. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research, 4.

Grants, Awards, and Honors

2016-2017	NSF RI EPSCoR Funding and Priority Access to Super-Computing Resources
08/2016	MCB Retreat Poster Competition Winner: invited talk for MCB retreat 2017.
2016	NSF EAGER Grant (wrote/edited proposal)
2015-2016	Office of the Vice President for Research Seed Grant (wrote/edited proposal)
2015-2016	Sponsored research contract (wrote/edited proposal)
2014-2015	NSF EPSCoR Predoctoral Fellowship
08/2014	MCB Retreat Poster Competition Winner: invited talk for MCB retreat 2015.
02/2014	Won a spot in the first round of Oxford Nanopore's MinION Access Program to beta test this new nanopore sequencing technology after writing and submitting 2013 proposal describing novel applications for the MinION
2011-2014	NSF Graduate Research Fellowships Program (GRFP) Individual Fellowship
2010-2011	National Institutes of Health (NIH) Graduate Traineeship (MCB Training Grant)

Summer 2010	Summer Undergraduate Research Project (SURP) funding for summer research with Dr. Patnaik at William Paterson University (WPU)
Spring 2010	Summer Undergraduate Research Project (SURP) travel award from WPU for Experimental Biology meeting (Anaheim, CA)
Spring 2010	ASBMB Student Travel Award for Experimental Biology meeting
11/2009	98^{th} percentile in Biochemistry, Cell and Molecular Biology GRE
Summer 2009	Summer Undergraduate Research Project (SURP) funding from WPU to intern at the University of Cambridge (UK)
Spring 2009	Inducted into Beta Beta Beta Honors Society for Biology
Spring 2009	Student Undergraduate Research Grant from WPU for research supplies
04/2009	1^{st} prize for Undergraduate Research Symposium Poster Competition hosted by WPU and represented by many universities in 4 hour radius
2009-2010	NSF Increasing Student Success in Biology and Biotechnology (ISSBB) Scholarship
2009-2010	CK Warner Scholarship for essay in biology
2009-2010	AL Rubin Scholarship for competitive honors student
2008-2009	Margaret Landi Scholarship for competitive science student
2008-2009	NSF ISSBB Scholarship
2008-2009	CK Warner Scholarship for essay in biology
2008-2009	AL Rubin Scholarship for competitive honors student
2007-2009	Federal SMART grant

Invited Talks

Urban JM, Yutaka Yamamoto, Leo Kadota, Audrey Lee, Michael Foulk, Jacob Bliss, and Susan Gerbi. (2017/TBA). High contiguity de novo genome assembly from long single-molecule data allows effective mapping of intrachromosomal DNA amplification sites in the fungus gnat, *Sciara coprophila*. Brown University MCB Graduate Program Retreat. (Invited to speak as an award for winning 2016 poster competition).

Urban JM, Foulk MS, Casella C, Gerbi SA. (2016). G4 motifs and nucleosomes are phased around a subset of human replication origins. William Paterson University Biology Department.

Urban JM, Foulk MS, Casella C, Gerbi SA (2015). G4 motifs and nucleosomes are phased around a subset of human replication origins. Cold Spring Harbor Laboratory Meeting on Eukaryotic DNA Replication and Genome Maintenance, p. 8. Invited platform talk.

Urban JM, Foulk MS, Casella C, Gerbi SA (2015). Lambda exonuclease, an enzyme used to enrich

RNA-protected nascent DNA for genome-wide mapping of replication origins by depleting parental DNA, inefficiently digests GC-rich DNA and G-quadruplexes. Brown University MCB Graduate Program Retreat. (Invited to speak as an award for winning previous year's poster competition. Had to decline in order to speak at CSHL).

Foulk MS*, **Urban JM***, Casella C, Gerbi SA (2015). Characterizing and controlling intrinsic biases of Lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. 19th Annual Buffalo DNA Replication and Repair Symposium (Buffalo, NY).

Urban JM (2013). Discovering our origins: the search for DNA replication initiation sites along human chromosomes. Brown University MCB Graduate Program Data Club.

Foulk MS*, **Urban JM***, Casella C, Gerbi SA (2013). Mapping DNA replication origins in the human genome. Cold Spring Harbor Laboratory Meeting on Eukaryotic DNA Replication and Genome Maintenance, p. 3. Invited platform talk – transferred privilege to a colleague on this project, Michael Foulk.

Urban JM, Foulk MS, Casella C, Gerbi SA (2013). Mapping DNA replication origins in the human genome. ASBMB annual meeting/Experimental Biology 2013 meeting (Boston, MA), abstract # 759.1. Invited platform talk.

Urban JM (2012). Discovering our origins: the search for DNA replication initiation sites along human chromosomes. Brown University MCB Graduate Program Data Club.

Urban JM (2011). Detecting initiation sites of DNA replication on human chromosomes. CCRI/Brown University Joint Symposium.

Urban JM (2010). WPU Honors thesis defense on "The only known promoter on an autonomous replicon that is essential for reporter gene/selectable-marker transcription as well as DNA replication in procyclic trypanosomes is dispensable in bloodstream forms".

Urban JM (2009). WPU Independent Study proposal defense on "Confirming that the only known promoter on an autonomous replicon that is essential for reporter gene/selectable-marker transcription as well as DNA replication in procyclic trypanosomes is dispensable in bloodstream forms".

Poster Presentations

Urban JM, Yutaka Yamamoto, Leo Kadota, Audrey Lee, Michael Foulk, Jacob Bliss, and Susan Gerbi. (2016). High contiguity de novo genome assembly from long single-molecule data allows effective mapping of intrachromosomal DNA amplification sites in the fungus gnat, *Sciara coprophila*. Brown University MCB Graduate Program Retreat.

Urban JM, Bliss J, Foulk MS, Gerbi SA (2015). Genome-wide analysis of developmentally regulated DNA re-replication in larval salivary glands of the fungus fly *Sciara coprophila*. Cold Spring Harbor Laboratory Meeting on Eukaryotic DNA Replication and Genome Maintenance, p. 199.

Urban JM, Gerbi SA (2015). Minimal changes to the standard MinION library preparation protocol to obtain 100 kb 2D reads. Oxford Nanopore Technologies MinION Community Meeting (New York City).

Urban JM, Gerbi SA (2015). The obstacles to great runs with ultra long DNA. Oxford Nanopore Technologies London Calling spring meeting (London, U.K.)

Urban JM, Bashir A, Sebra R, Howison M, Foulk MS, Gerbi SA (2015). Re-replication origins in *Sciara* DNA puffs revealed by new and old genomic technologies including nanopore sequencing. American Society for Biochemistry and Molecular Biology annual meeting; Experimental Biology 2015 meeting (Boston, MA), abstract # 6831.

Urban JM, Foulk MS, Casella C, Gerbi SA (2014). Lambda exonuclease, an enzyme used to enrich RNA-protected nascent DNA for genome-wide mapping of replication origins by depleting parental DNA, inefficiently digests GC-rich DNA and G-quadruplexes. Brown University MCB Graduate Program Retreat.

Foulk MS*, **Urban JM***, Casella C, Gerbi SA (2014). Lambda exonuclease, the cornerstone of nascent strand sequencing to map replication origins genome-wide, digests GC-rich DNA inefficiently and stalls at G quadruplex structures. 55th Annual Meeting of the American Society for Cell Biology/International Federation of Cell Biology (Philadelphia, PA), abstract # P321, p. 19.

Urban JM, Foulk MS, Casella C, Gerbi SA (2013). Mapping DNA replication origins in the human genome. ASBMB annual meeting/Experimental Biology 2013 meeting (Boston, MA). Poster and Invited Talk.

Urban JM, Foulk MS, Casella C, Gerbi SA (2011). Mapping DNA replication origins to the human genome. Cold Spring Harbor Laboratory meeting on Eukaryotic DNA Replication and Genome

Maintenance, p. 222a.

Urban JM (2010). A 10 bp element that is essential for autonomous plasmid replication in procyclic Trypanosoma brucei is dispensable in bloodstream forms. American Society for Biochemistry and Molecular Biology annual meeting/ Experimental Biology 2010 meeting (Anaheim, CA).

Urban JM (2010). A 10 bp element that is essential for autonomous plasmid replication in procyclic Trypanosoma brucei is dispensable in bloodstream forms. WPU Research Day.

Urban JM (2009). A 10 bp element that is essential for autonomous plasmid replication in procyclic Trypanosoma brucei is dispensable in bloodstream forms. 3rd annual WPU Undergraduate Research Symposium in Biological Sciences (with undergraduate representation from multiple schools). 1st Prize in Genetics category.

Teaching and Outreach

2016 - Present	Training Brown undergraduate Ben Doughty in RNA and DNA work as well as bioinformatics; guiding him in his undergraduate Honor's thesis work.
2015 - Present	Training Brown undergraduates Audrey Lee, Leo Kadota, and Julia Leung in molecular biology (e.g. working with DNA, gels, qPCR) and basic computation in R for qPCR analysis. Gave guidance for their undergraduate theses, and currently providing guidance to Leo and Audrey for their Master's theses.
April 2016	Poster Judge for William Paterson University Undergraduate (WPU) Research Symposium.
April 2016	Alumni Career Panel session at WPU Undergraduate Research Symposium.
April 2016	Guest lecturer for WPU Advanced Undergraduate and Master's level Molecular Biology Course.
Summer 2015	Trained visiting undergraduate, Sevde Nur Karatas, in DNA work.
2012-2015	Started and maintained the Brown University Genomics Club (GC). The goal of GC was to form a network amongst people engaged in genomics research at Brown University and other nearby universities, and included a Google group where members were able to seek advice, a web-site to help members learn about genomics, monthly talks, and occasional workshops/tutorials. GC included 60 participants from Brown University, Marine Biological Laboratory, Roger Williams University, and University of Rhode Island.
2013–2014	Trained undergraduate researcher, Julia Borden, in RNA-seq protocols

June 2013	HHMI Brown University 5 day summer course on "Basic Computation and Statistics for Genome-wide Studies". I taught 32 undergraduates the basics of Unix/Linux and installing bioinformatics software from source code through mapping reads and calling peaks with real ChIP-seq data.
Fall 2011	Brown University Teaching Assistant/Lab Instructor for Genetics (Bio 47).
March 2010	Spoke at WPU Undergraduate Career Day about applying to Graduate Schools.
March 2010	Promoted undergraduate research at the WPU Scholarship Luncheon where I spoke to over 200 WPU scholarship donors and recipients on how receiving scholarships helped me succeed by offsetting the need for an additional job.
2008–2010	Tutor in Genetics and Molecular Biology for WPU Science Enrichment Center.
2008–2009	Teaching Assistant for WPU courses in Genetics and Molecular Biology.
May 2010	Blessed Sacred Heart Elementary School, Paterson, NJ. Science Fair Judge for 5th–8th grade students
Fall 2009	Gave a presentation at a WPU Biology Department seminar for undergraduates majoring in and thinking about majoring in Biology. It was titled, "Do stuff", and in it I encouraged the undergraduates to get involved with research early on.
2008-2010	Biology Representative for Student Advisory Board to the WPU Dean of the College.

Interactions with Industry

Oxford Nanopore Technologies	Won spot in the first round of MinION Access Program (MAP) and received free equipment, reagents, and flow cells to develop proposed applications. Designed and promoted modified protocols for pushing MinION read lengths beyond 100 kb. Developing replication origin mapping technique.
Genomic Vision	Collaboration on proposed applications for DNA combing.
NabSys	Collaboration on obtaining "electronic maps" for the *Sciara* genome and exploring applications of electronic mapping technology.
Intact Genomics	Collaboration on attempts to obtain Pulsed-Field Gel size-selected 100 kb libraries for the Oxford Nanopore MinION. The aim of this collaboration was to create libraries with mean lengths of >100 kb whereas my previous protocols pushed the tail of the read length distribution beyond 100 kb.

Experiences with Large Consortia

MARC 2014-2016 — The **M**inION **A**nalysis and **R**eference **C**onsortium formed early on in the MinION Access Program, and was initially lead by Ewan Birney. I participated in the formation of MARC, in discussions on the weekly phone calls and in-person meetings leading up to the first MARC paper [Ip et al, 2015], and helped to draft and edit the paper (I am a co-author).

Additional Education

Coursera MOOC Platform — "Computing For Data Analysis" with Dr. Roger Peng of Johns Hopkins University. Certificate of Completion: 02/2013.

"Data Analysis" with Dr. Jeff Leek of Johns Hopkins University. Certificate of Completion: 03/2013.

"Computational Molecular Evolution" with Dr. Anders Gorm Pedersen of the Technical University of Denmark. Certificate of Completion: 09/2013.

"Bioinformatics Algorithms" with Dr. Pavel Pevzner of UC San Diego. Certificate of Completion: 02/2014.

EdX MOOC Platform — "MITx 6.00x: Introduction to Computer Science and Programming" with Drs. Eric Grimson, John Guttag of MIT. Certificate of Completion: 06/2013.

Broad Institute Workshop — Computational Genomics Workshop at the Broad Institute funded by the NHGRI Center for Excellence in Genomics (CEGS) Center for Cell Circuits. Topics: RNA-seq, ChIP-seq, Single-Cell genomics. September 22-23, 2014.

Other popular press

Urban JM "How does bowtie2 assign MAPQ scores?" Published at the Biofinysics Blog on May 28, 2014. >8200 readers as of October 2016. http://biofinysics.blogspot.com/2014/05/how-does-bowtie2-assign-mapq-scores.html. Summary: a tutorial/guide and analysis of one of the most popular short read mappers for genomics research.

Urban JM "The Craft of Homebrew-ing for Mac OS". Published at the Biofinysics Blog on January 7, 2014. >4300 readers as of October 2016. biofinysics.blogspot.com/2014/01/the-craft-of-homebrew-ing-for-mac-os.html. Summary: a tutorial/guide on using 'homebrew' to install bioinfomatics software on a Mac Book Pro.

Urban JM "The slow death of the term 'uniquely mappable' in deep sequencing studies and resisting the 'conservative' urge to toss out data". Published at the Biofinysics Blog on May 28, 2014. >1400 readers as of October 2016. http://biofinysics.blogspot.com/2014/05/the-slow-death-of-term-uniquely.html. Summary: explores what 'uniquely mappable reads' means to different people in different contexts.

Urban JM "Where does bowtie2 assign true multireads (AS=XS)?". Published at the Biofinysics Blog on May 29, 2014. >1100 readers as of October 2016. http://biofinysics.blogspot.com/2014/05/where-does-bowtie2-assign-true.html. Summary: a tutorial/guide and analysis of one of the most popular short read mappers for genomics research that extends upon the previous blog titled, "How does bowtie2 assign MAPQ scores?".

References

PhD Thesis Advisor
Dr. Susan A. Gerbi, PhD.
George Eggleston Professor of Biochemistry
Brown University BioMed Division
Department of Molecular Biology, Cell Biology, and Biochemistry
Brown University
185 Meeting Street, Sidney Frank Hall Room 260
Providence, Rhode Island, 02912
Susan_Gerbi@Brown.edu, (401) 863-2359

Undergraduate Research Advisor
Dr. Pradeep Patnaik, PhD.
Professor of Biology
Department of Biology
William Paterson University
Science East 4055
Wayne, New Jersey, 07470
patnaikp@wpunj.edu , (973) 720-3454

PhD Thesis Committee Chair
Dr. Erica Larschan, PhD.
Assistant Professor of Biology
Department of Molecular Biology, Cell Biology, and Biochemistry
Brown University
185 Meeting Street
Providence, Rhode Island, 02912
Erica_Larschan@Brown.edu, (401) 863-1070

Acknowledgements

I dedicate this thesis to my grandparents, the Smiths and the Urbans.

I never knew my grandmother and grandfather from the Urban lineage, and I never had the opportunity to have conversations with my grandparents from the Smith lineage as an adult. However, I know that my successes and achievements rest, in part, on all of their shoulders. They were able to overcome their own sets of challenges, achieve their own successes, and to instill hard-working principles into their children, my parents.

I dedicate this thesis to my parents, Kathleen and Daniel Urban.

As a child my mother would read to my twin brother and I every night at bedtime until we could read on our own. Then she brought us to the library and always kept us excited about books. My mother has been selfless in her support. As an undergraduate researcher, my advisor helped orchestrate an opportunity to do research for the summer of 2009 at Cambridge University in England with his colleague. Days before leaving, my mother went hiking with her friends and she fell off a 40 foot cliff into a shallow river. As far as I knew upon receiving the news, she was dead. I drove to the hospital, slamming my fists against the steering wheel, wondering if I had been a decent son. Miraculously, she survived the fall without any long-lasting injuries or conditions. Nonetheless, I thought I should cancel my trip to England. She was insistent that it was too important to pass up though - she would be fine. That summer I was introduced to bioinformatics – an experience that changed my perspective on the future of biological research completely, leading me to use my time in graduate school to learn programming and probability.

My father is an extremely hard worker. Growing up, I saw his commitment to running "Urban's Auto Sales", a business that he inherited from his father. He would wake up very early and leave to work. Then he would get home in the evening, hungry for supper. A hard day's work behind him, he still found time to do a lot of work around the house and the yard, building the decks and

staircases that decorate our backyard's descent to the lake behind our house. I like to think I work as hard as he does, just on different things.

My parents have been extremely supportive in each era of my life. When I wrestled my freshman year in high school, my parents made it out to see my matches. When my brother and I wanted to learn music, my Mom brought us to a music store, and before long I was a decent bass player. My brother and I formed a band and my parents came to all the shows. When I was the young Executive Chef at a local restaurant, my parents came for lunch and dinner. When I wanted to leave the culinary industry around the age of 24 to pursue science like my brother, they were encouraging. I felt old at 24. I thought I was too old to go to go to college - that ship had surely sailed - but they reminded me that I could be 28 with a degree or 28 without one. During my undergraduate endeavors into biology research, my mother and father came to the poster competitions to cheer me on, and they have been supportive during my PhD.

I dedicate this thesis to my brother Kevin and my older sister Lorretta. One of my earliest memories is a conversation with my identical twin brother, school boy age, pondering if the universe had walls and, if so, asking, "What was behind them?" I also remember wrestling with the idea that we "shared a body" before we were born, from the single-cell stage until our small clump of cells split, forever separating us and thereby leading to the formation of two 'clones', albeit one clone is cooler than the other. He has been my partner in "philosophizing" ever since. In one of my earliest memories with Lorretta, she was sitting in front of me while watching TV. When I complained that I couldn't see, she proceeded to "teach" me how vision curves around objects and that I should be able to see just fine! She kindly suggested that perhaps something was wrong with me. She was my big sister and knew everything, so she must be right..right? That experience made me contemplate whether or not I should believe her or anyone else just on the basis that they are older than me.

I dedicate this to my nephews and nieces – Hunter, Avia, Jack, and the newcomer Harrison. Perhaps you will one day read this thesis 15–20 years from now and think, "Wow. For his thesis, he only sequenced, assembled, and annotated one insect's genome, mapped amplicons in it, and mapped replication origins in the human genome? PhDs were easy back then, though that nanopore sequencing sounded laborious and time-consuming."

I dedicate this thesis to my wife, Jennifer Urban, who has been doing a PhD in parallel with me. We have been through many challenges in the last six years, from living far apart to the death of her Mom and the consequences on her family life thereafter. I hope that I have been supportive at every step. She has been for me.

I also dedicate this thesis to the Universe, which is having at laugh at my expense. My twin brother was born one minute before me. Ever since, any time we argue I have to endure, "Don't

worry, you'll understand when you're older and wiser like me." Now my wife Jen is defending her PhD one day before I defend mine. Good one Universe. You win.

Finally, I dedicate this to my undergraduate and graduate thesis advisors, Pradeep Patnaik and Susan Gerbi. Pradeep is an incredibly smart and rigorous scientist who took the time to painstakingly train me as an undergraduate to work with cell culture and DNA. The long conversations we had about our work and life prepared me for graduate school. Susan Gerbi is also brilliant and has been a champion of my budding career in science. When we go to conferences together, she brings me around to meet all of the best and brightest scientists in our field. She was the perfect graduate advisor for me, as she creates an environment where her students can develop independence. She calmly waited while I took time to become proficient in bioinformatics. I like to think her investment paid off, as I obtained the skills to assemble the genome of the fungus fly, *Sciara coprophila*, that she has a deep conviction about. One of her goals is to popularize *Sciara* as a model system. I am only too happy to help her achieve that goal and like to think my contributions in her laboratory will set the stage for this to happen. Susan always encouraged me to do things my own way, even at times that I have to assume other advisors would discourage my way. To illustrate, the passage below opened up my NSF fellowship application. Others who read it thought it was too risky and poetic, and encouraged its removal. In contrast, Susan said, "How wonderful!".

"Here I am. A swarm of highly ordered, differentiated clones convinced I am an individual, perhaps a god among these basic units of life, perhaps a servant, but without a doubt, a product. One evolution has wrought. Here sits and writes the result of carbon, water and time; a conscious molecular mass that emerged in the late 1900s, a time in which a clump of cells began to differentiate and collaborate so I could one day contemplate my organic fate. It is almost cruel despite its staggering beauty. My only purchase on eternity is making sure that portions of 'my' DNA go on to see another day in the same way it has almost always been." –John Urban (NSF GRFP application)

I wrote that passage in my first semester of doing a PhD, and here in my last semester I sit and write again, still a result of carbon and water, but 6 years more time. As far as I can tell, the only reason the department allowed me to join Susan's laboratory, which was underfunded at the time, was because I won that fellowship.

Thank you all for being part of my life and helping to mold me in your own ways.

Contents

List of Figures

List of Tables

Part I:

Introduction

This thesis is broken up into four parts. In this first part, eukaryotic DNA replication and DNA re-replication in insects is thoroughly reviewed. All other background information necessary for each chapter is present in the chapters themselves. Prior to this work, the sequence for only one re-replication locus in *Sciara coprophila* was known, though many exist at sites called DNA puffs. Little other DNA sequence information for *Sciara* was available. The central problem tackled in this thesis is focused on expanding the studies of locus-specific re-replication in *Sciara coprophila* to the many DNA puffs. To enable modern genomics approaches for studying *Sciara coprophila*, I sequenced and assembled its genome using an array of single-molecule and high throughput technologies. This allowed me to use high throughput sequencing to identify the majority of re-replication loci in salivary glands across the reference genome. These projects are detailed in Part II of this thesis, specifically in Chapters 2 and 3 that describe the genome assembly and DNA puff mapping, respectively. In addition to identifying the DNA puff loci within the larger genome sequence, a goal is to understand exactly where DNA replication initiates from within each. As using low-throughput methods on all 14 or more DNA puff loci identified would be laborious and expensive, I also pursued the development of genome-wide origin mapping methods that can be used in future studies on *Sciara coprophila*. This is detailed in Chapters 4, 5, and 6 found in Part III. Chapter 4 reviews genome-wide origin mapping techniques. Chapter 5 describes my work in mapping replication origins across the human genome. Chapter 6 discusses the progress I have made in developing novel methods to study DNA replication. Finally, Part IV is the epilogue.

CHAPTER 1

Introduction to DNA re-replication in insects

DNA replication and DNA re-replication

DNA is the hereditary molecule. For the survival of any species in all three domains of life, it is crucial that DNA is faithfully passed on. This is the essential process known as DNA replication. In the linear chromosomes of eukaryotes, DNA replication initiates at multiple sites called "origins". Origins are tightly regulated such that each is used at most once per cell cycle [Dutta and Bell, 1997, Bielinsky and Gerbi, 2001, Bell and Dutta, 2002, Aladjem and Fanning, 2004, Aladjem, 2007]. This is accomplished by selecting origins for use as initiation sites in G1 and restricting their activation to S-phase. When an origin is used, it is said to have "fired". Limiting an origin to firing once per S-phase ensures that only two copies of the genetic material are made, one for each daughter cell. When this strict once-and-only-once regulation is lost at an origin, it can "re-fire". DNA re-replication occurs when an origin re-fires during the same S-phase leading to nested replication bubbles containing extra copies of that locus [Schimke et al., 1986, Dutta and Bell, 1997, Bielinsky and Gerbi, 2001, Bell and Dutta, 2002, Aladjem and Fanning, 2004, Aladjem, 2007]. It is important to understand (i) how re-replication is normally prevented, (ii) what the consequences are when the normal regulation against re-replication is perturbed, (iii) how re-replication can occur when the normal regulation is intact, and (iv) how re-replication can be directed to specific sites.

How re-replication is normally prevented has been well-studied. Eukaryotic cells employ several regulatory mechanisms to ensure that DNA polymerase is recruited to each replication origin

The introduction for my thesis has been written entirely by me and will be the basis for a review of DNA re-replication in insects to be submitted in Spring 2017.

to initiate DNA synthesis at most once per cell cycle (Fig. 1.1). The regulation of initiation takes place in two usually mutually exclusive phases: (i) an origin selection and licensing phase when the pre-Replication Complex (preRC) can bind to form a competent origin that cannot fire, and (ii) an origin activation phase when competent origins can fire but cannot form [Dutta and Bell, 1997, Bielinsky and Gerbi, 2001, Bell and Dutta, 2002, Aladjem and Fanning, 2004, Machida et al., 2005, DePamphilis et al., 2006, Aladjem, 2007]. Specifically, origin selection and licensing is typically restricted to G1 whereas activation can only occur in S-phase. The mechanisms to prevent competent origins from re-forming in S-phase, and often through G2 and M-phase, vary somewhat between different species, but include phosphorylation, ubiquitination, degradation, sequestration, downregulation, and nuclear export of preRC proteins that are not needed for elongation, all of which depend on the same conditions that lead to the initiation of DNA synthesis in S-phase [Aladjem, 2007, DePamphilis et al., 2006, Aves, 2009]. At the start of this process, the preRC forms and can only do so when cyclin dependent kinase (CDK) levels are low. PreRC formation starts by early G1 with the Origin Recognition Complex (ORC) binding to potential origins, which then recruits Cdc6 (cell division cycle 6), Cdt1 (cdc10 dependent transcript1; Chromatin Licensing And DNA Replication Factor 1), and the MCM2-7 double-hexamer (MiniChromosome Maintenance 2-7). At the end of G1 phase when S-phase CDK activity rises, preRC components become phosphorylated by CDK as well as Dbf4-dependent kinase (DDK; Cdc7 kinase and its regulatory subunit Dumbbell-forming 4 [Chapman and Johnston, 1989]). Cdc6 and Cdt1 leave the preRC, which is converted into the pre-initiation complex (preIC) [Aladjem, 2007]. The formation of the preIC requires loading of Cdc45 (cell division cycle 45) and GINS (Go, Ichi, Nii, and San; Japanese for five, one, two, and three [Takayama et al., 2003]) to the Mcm2-7 complex [Heller et al., 2011] as well as the recruitment of DNA polymerases. The combination of Cdc45, MCM2-7, and GINS is known as the CMG complex, sometimes called the unwindosome, and is thought to make up the replication fork helicase [Pacek and Walter, 2004, Moyer et al., 2006, Pacek et al., 2006]. Loading of the Cdc45 and GINS activates the helicase activity of Mcm2-7 [Ilves et al., 2010, Costa et al., 2011] and allows the origin to unwind. Recruitment of DNA polymerases and other factors, including the procsseivity factor PCNA (proliferating cell nuclear antigen) that clamps polymerases to DNA during elongation, allows the origin to fire, thereby initiating DNA synthesis. From S-phase typically through M-phase, new preRCs cannot form to produce a new preICs while CDK activity remains high. The affinity of ORC1 for chromatin is strong in G1, but decreases in S-phase through mitosis [DePamphilis et al., 2006]. The loss of affinity for chromatin is at least in part due to ORC1 phosphorylation, which can also lead to export from the nucleus in hamster cell lines [DePamphilis et al., 2006]. In some human cell lines, ORC1 is also targeted for ubiquitin-dependent degradation during S-phase [Machida et al., 2005, DePamphilis et al., 2006]. CDK prevents Mcm2-7 helicase loading by inhibiting the interaction of Cdt1 with Orc6 [Chen and Bell, 2011]. During S-phase, the protein geminin sequesters Cdt1 and other pathways lead to Cdt1 proteolysis through ubiquitination [McGarry and Kirschner, 1998, Quinn et al., 2001, Arias and Walter, 2005, DePamphilis et al., 2006]. In yeast, phosphorylated MCM complexes that are not bound to chromatin during S-phase–M-phase are exported from the

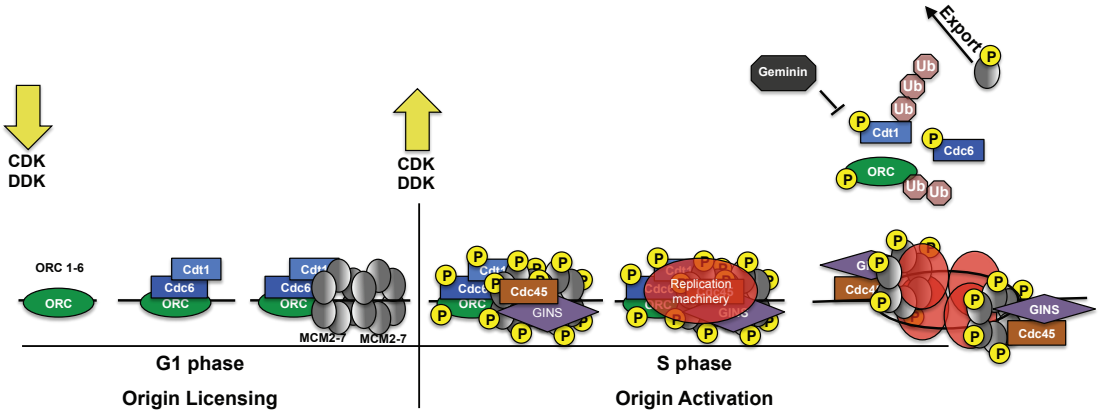


Figure 1.1: Regulation against re-replication by separating origin licensing from origin activation.

nucleus [DePamphilis et al., 2006]. Metazoan MCMs are phosphorylated as part of activation, but do not seem to be exported. Instead, they are unable to bind again during S-phase, although this may reflect the negative regulation of ORC and Cdt1 [DePamphilis et al., 2006]. Though these mechanisms are extensive, they may not be completely redundant leaving the possibility that, though below most detection limits, re-replication and its consequences might be more common than is currently appreciated [Green et al., 2006, Dorn et al., 2009].

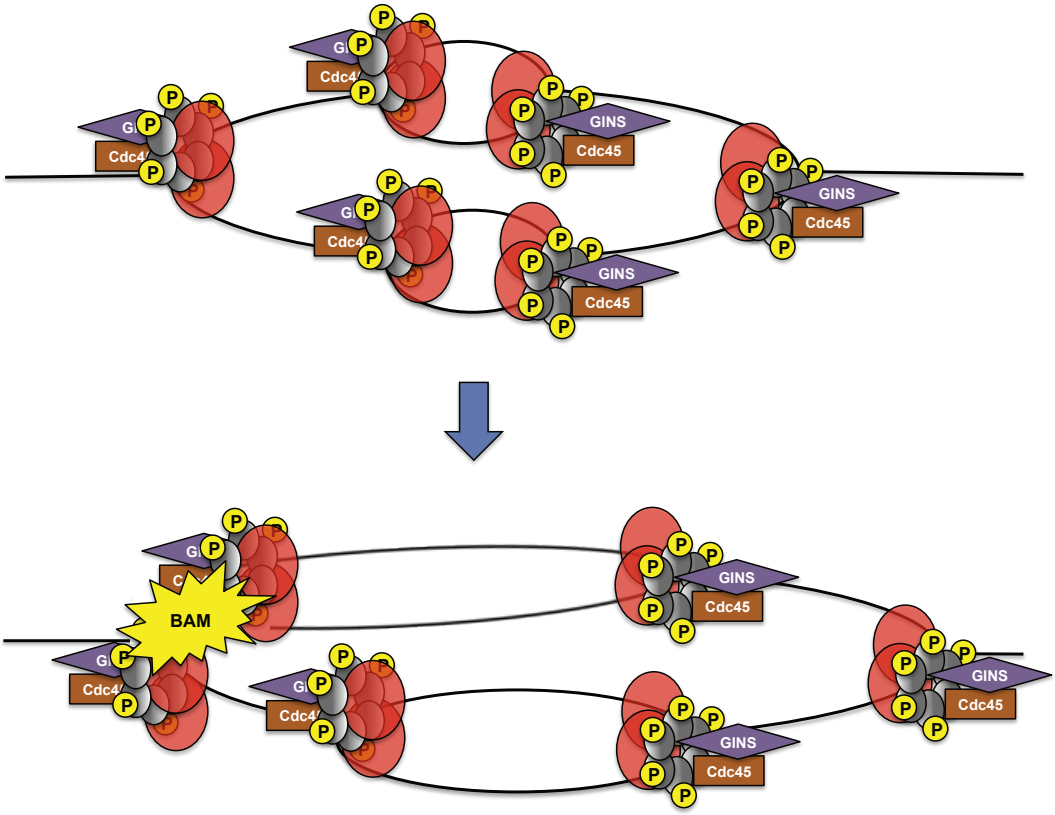
The consequences of re-replication have also been documented. Deregulation of the multiple controls that prevent re-replication of the genome has been implicated in genomic instability and cancer progression [Blow and Gillespie, 2008, Diffley, 2010]. Over-expression of Cdt1 or depletion of geminin in frogs, flies, and mammals results in re-licensing, re-replication, and general over-replication of the genome [Saxena and Dutta, 2005, DePamphilis et al., 2006]. DNA replication stress and amplification of oncogenes are hallmarks of cancers [Blow and Gillespie, 2008, Negrini et al., 2010, Hanahan and Weinberg, 2011, Macheret and Halazonetis, 2015]. However, oncogene amplification is detected as multiple copies inserted throughout the genome, often in tandem arrays. Though this superficially appears different than nested re-replication forks where extra copies exist in parallel sibling DNA molecules, it has been suggested that these precarious nested fork structures may be part of the initiating events that lead to oncogene amplification [de Cicco and Spradling, 1984]. Directly supporting this notion, it has been demonstrated in yeast that removing global controls against re-replication results in re-firing of origins genome-wide [Nguyen et al.,

2001, Gopalakrishnan et al., 2001, Green et al., 2006, Kiang et al., 2010], which leads to replication fork break down [Finn and Li, 2013], blocked cell proliferation [Green and Li, 2005], extensive DNA damage [Green and Li, 2005], non-allelic homologous recombination [Green et al., 2010], stable gene amplification [Green et al., 2010], other copy number changes and genomic re-arrangements [Green et al., 2010,Brewer et al., 2011,Finn and Li, 2013,Brewer et al., 2015], as well as chromosome instability and aneuploidy [Hanlon and Li, 2015]. Moreover, DNA damage, double-strand breaks, crossover structures from homologous recombination, and large rearrangements have been detected during DNA re-replication in *Drosophila* [Heck and Spradling, 1990, Yarosh and Spradling, 2014, Alexander et al., 2015] and *Sciara* [Liang et al., 1993]. In addition to the fragility of the nested replication bubbles, head-to-tail replication fork collisions resulting from a pursuant fork catching up to the one in front of it can actively lead to some of these effects, including DNA fragmentation (Fig. 1.2) [Davidson et al., 2006].

How re-replication can occur when the normal regulation is intact and how it can be directed to specific sites are harder questions to study in yeast and human cancer cell culture models. Re-replication is an extremely rare event in most eukaryotic DNA replication programs due to the effectiveness of the normal regulation against it. Knocking out these controls in yeast allowed the identification of origins that preferentially re-replicated [Kiang et al., 2010, Richardson and Li, 2014], but it is unclear if these origins can also preferentially direct re-replication in the wild type background or if they are just opportunistic and efficient when the controls have been taken away. In contrast, insects present examples of locus-specific re-replication during normal development [Gerbi and Urnov, 1996, Claycomb and Orr-Weaver, 2005]. Importantly, the specific rogue origins that re-fire in insects do so when all other origins in the genome remain under the controls against re-firing. Therefore, insects present unique opportunities to elucidate how re-replication can occur when the normal regulation is intact and how it can be directed to specific sites.

Differential DNA replication in insects

Insects have polyploid tissues where each nucleus harbors many copies of the genome. This is mediated through a process called endoreplication where the genome is repeatedly replicated without cell division. When the many resulting sister chromatids are held in tight register, they form polytene chromosomes that are visible under a microscope when stained, first observed by Balbiani in 1881 in the larval salivary glands of *Chironomus* midges [Balbiani, 1881]. Polytene chromosomes became a bigger focus in the 1930s when they were observed in other dipteran insects, such as the fruit fly *Drosophila melanogaster* and the fungus fly *Sciara coprophila*, in addition to *Chironomus* [Heitz and Bauer, 1933, Painter, 1933, Metz and Gay, 1934a, Metz and Gay, 1934b, Doyle and Metz, 1935a, Doyle and Metz, 1935b]. In 1938, Poulson and Metz observed giant puff structures in the Sciarid salivary gland polytene chromosomes [Poulson and Metz, 1938]. It wasn't until almost two decades later,



The fragile nested bubble structure, parallel copies, and fork collisions can result in DNA damage, recombination, genomic instability, and gene amplification: all fodder for genome evolution and cancer progression

Figure 1.2: Consequences of re-replication.

in 1955, when it was first noticed that the puffs of a related species, *Rhynchosciara angelae*, were sites of intense DNA synthesis and contained "extra DNA" as suggested by Feulgen staining [Breuer and Pavan, 1955], ^{3}H-thymidine incorporation [Ficq and Pavan, 1957], and microspectrophotometry [Rudkin and Corlette, 1957]. The DNA in these puffs seemed to be actively amplified. This was quite surprising at the time as it was in violation of the proposed "rule of DNA constancy" where all cells of an organism, even polyploid cells, were expected have the same DNA content for each copy of the genome [Gerbi and Urnov, 1996]. Soon thereafter, in 1962-1970, the giant puffs in *Sciara coprophila* were also confirmed to be sites of disproportionate DNA synthesis [Swift, 1962, Gabrusewycz-Garica, 1964, Crouse and Keyl, 1968, Rasch, 1970a, Rasch, 1970b]. Puff structures in polytene chromosomes that contain amplified DNA were termed "DNA puffs" to differentiate them from other puffed out structures that do not have extra DNA or intense DNA synthesis, but that are sites of intense RNA synthesis called "RNA puffs" [Gerbi and Urnov, 1996]. The genes amplified in Sciarid DNA puffs seem to encode proteins for the pupal coat that are needed in abundance in a short period of time. In the 1980s, amplified DNA was also found in the follicle cells of *Drosophila melanogaster* at the chorion loci, which contain genes that are needed for proper eggshell development [Spradling and Mahowald, 1980, Claycomb and Orr-Weaver, 2005]. All of these insects demonstrate intrachromosomal DNA amplification as opposed to the extra chromosomal amplification of rDNA in *Xenopus* or minichromosomes containing rDNA in *Tetrahymena* [Gerbi and Urnov, 1996, Claycomb and Orr-Weaver, 2005]. It was established in both *Sciara* and *Drosophila* that DNA amplification was due to differential DNA replication, raising the question: how can specific DNA replication origins be regulated to generate multiple copies of specific loci? It was posited that the answer could give insight into how re-replication might occur aberrantly in other systems and potentially give mechanistic insight for the amplification of genes in cancer and evolution [Spradling and Mahowald, 1980]. Therefore, researchers went to work on unraveling re-replication in insects.

DNA re-replication in the follicle cells of the fruit fly, *Drosophila melanogaster*

Early characterization of chorion gene amplification

The most work towards understanding re-replication in insects has been in studying chorion gene amplification in the follicle cells of *Drosophila melanogaster*. In this system, the regions of the genome that undergo amplification have been descriptively named, "Drosophila Amplicons in Follicle Cells" (DAFCs) [Claycomb et al., 2004, Claycomb and Orr-Weaver, 2005]. In Drosophila, DNA amplification of specific loci occurs directly after a separate process called endoreplication where the entire genome undergoes four rounds of DNA replication without intervening mitoses resulting in 16 genomic copies [Spradling and Mahowald, 1980]. In 1980, the chorion gene locus on the X chromosome (DAFC-7F) was shown to amplify approximately 12-fold by Alan Spradling and Anthony Mahowald using quantitative blotting [Spradling and Mahowald, 1980], and they proposed

that amplification was due to differential DNA replication. In 1981, Spradling showed that a second cluster of chorion genes on chromosome 3 (DAFC-66D) also amplified. He demonstrated for both gene clusters that there was a decreasing amplification gradient in both directions for a total of 90-100 kb of flanking sequence centered around each [Spradling, 1981]. The original amplicon, DAFC-7F, was estimated to amplify 16-fold in this study, and the newer amplicon on the third chromosome, DAFC-66D, was estimated to amplify 60-fold [Spradling, 1981]. That same year Spradling and Mahowald observed that a small inversion on the X chromosome that moved DAFC-7F to a different locus also resulted in shifting the amplification gradient such that DAFC-7F remained at the peak of amplification [Spradling and Mahowald, 1981]. Chromosomal sequence that is near DAFC-7F in wild type, but not in the inversion mutant, no longer amplified. This strongly suggested that DAFC-7F contained cis-regulatory elements that directed DNA amplification. In 1984, the Spradling laboratory used P-element transformation to insert fragments from DAFC-66 into new chromosomal locations to study ectopic DNA amplification [de Cicco and Spradling, 1984]. They found that only transposons that contained a specific 3.8 kb fragment were able to confer amplification potential to ectopic loci, which amplified in follicle cells during the normal amplification stages. Overall, this supported the notion from the DAFC-7F inversion that specific cis-regulatory elements drove amplification, and it was in this paper that the term "Amplification Control Element" (ACE) was coined to describe the 3.8 kb fragment from DAFC-66D. Specifically, this ACE is called ACE3, which specifies it as the "ACE on chromosome 3". ACE3 was narrowed down to just 510 bp in 1986 by Terry Orr-Weaver and Alan Spradling [Orr-Weaver and Spradling, 1986]. This also demonstrated that the nearby genes were not essential to the amplification process and that the important element was likely only a special amplification origin or a control sequence that caused nearby origins to re-replicate. The following year the amplification control element for DAFC-7F on the X chromosome was also identified as a 467 bp fragment able to confer ectopic amplification. The DAFC-7F ACE is known as ACE1. ACE1 shares a 32 bp stretch of homology with ACE3 that is comprised of imperfect AATAC repeats, though it could simply be the AT-rich nature that is important. These repeats were on the border of the ACE1 fragment and appear to be important cis-elements as deleting all repeats perturbed ectopic amplification. In contrast, most repeats could be removed before affecting amplification levels.

Electron microscopy complemented the genetics studies of the 1980s [Osheim and Miller, 1983,Osheim et al., 1988, Osheim and Beyer, 1991]. The "onion skin" model of intrachromosomal DNA amplification (Fig. 1.3) was confirmed by Osheim and Miller in 1983 when they presented electron micrographs of chromatin spreads from amplification stage follicle cells that contained multi-forked structures where a strand of chromatin forks into two strands, which subsequently fork into two more strands each [Osheim and Miller, 1983]. In contrast, branching structures were not found in pre-amplification stage chromatin spreads. These images of nested re-replication forks resulting from an origin firing multiple times were consistent with the gradient of amplification observed by

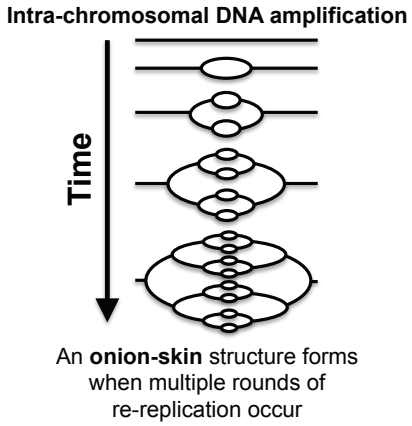


Figure 1.3: Onion-skin model of amplification.

Spradling. Importantly, this was strong evidence against the possibility that amplification was extra-chromosomal. In 1988, Osheim and colleagues used electron microscopy of chromatin spreads again and could classify onion-skin structures as either from DAFC-7F or DAFC-66D based on the size, direction, spacing, and transcriptional timing of locus-specific genes and the presence of replication forks [Osheim et al., 1988]. Their results were consistent with a centrally located replication origin that is used during amplification for both DAFCs, observing in one case a nascent replication bubble at the expected ACE1 location (DAFC-7F). Ultimately, where they approximated the DAFC-66D origin to be with respect to ACE3 was confirmed a year later by 2D gels [Delidakis and Kafatos, 1989]. They also observed that the re-replication forks to either side of the origin were not equidistant from the origin, suggesting different fork speeds. Unequal fork rates was also suggested from 2D gel analyses a year later [Delidakis and Kafatos, 1989]. Interestingly, electron micrographs showed that the amplified genes were actively transcribed throughout the re-replication process, indicating that any disruption of transcription from passing replication forks is transient [Osheim et al., 1988].

In 1989, Delidakis and Kafatos confirmed the importance of ACE3 for driving amplification [Delidakis and Kafatos, 1989], but failed to detect replication activity from within the ACE3 element itself using the recently developed 2D gel method [Brewer and Fangman, 1987]. They identified four other cis-regulatory elements called amplification enhancer regions (AERs) and showed evidence that replication initiated from at least two sites nearby. They had direct evidence from 2D gels that replication bubbles emanated from a particular site that they named ori-C3, consistently located with previous electron microscopy estimates [Osheim et al., 1988]. However, those gels also showed evidence of replication forks coming into that site from one or more nearby origins. This was the first indication that there was more than one initiation site in the centrally amplified region of DAFC-66D. Whereas it was clear that ACE3 contained information that uncoupled it from regulation against re-replication as shown by its ability to induce ectopic amplification, ACE3 was never

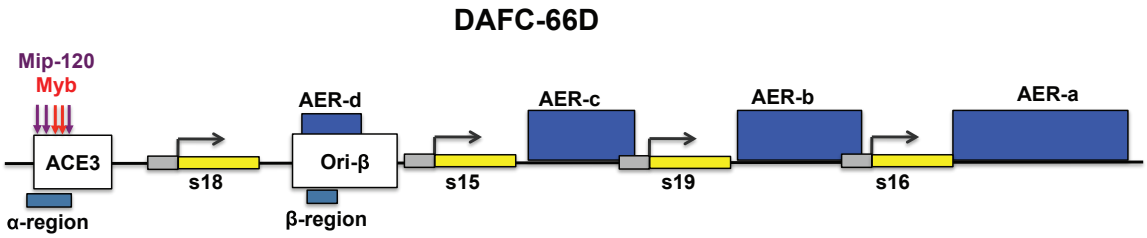


Figure 1.4: Anatomy of DAFC-66D, the chorion amplicon on chromosome 3.
DAFC-66D has the highest amplification levels and is the most studied amplicon.

able to fully reproduce endogenous amplification levels. This was typically explained by position effect variegation, where different areas in the genome exert differential negative regulation against re-replication. However, Delidakis and Kafatos presented a model where the lower amplification levels at ectopic sites were also at least partially due to the absence of the four AERs (AER-a, AER-b, AER-c, AER-d). Each AER slightly lowered amplification levels when deleted, but the effect was more dramatic when deleting multiple AERs. The structure of the amplification origin emerging at this point with respect to the four AERs was ACE3–d–c–b–a, with chorion genes intercalated between the AERs and with ori-C3 somewhere between ACE3 and AER-d (Fig. 1.4). Later that year, data from the same laboratory suggested that ACE3 is important but not essential for DAFC-66D amplification since some constructs without it could amplify at marginal but detectable levels [Swimmer et al., 1989]. This was in opposition to the earlier results from Orr-Weaver and Spradling [Orr-Weaver and Spradling, 1986], but the authors offered that their experiments were a little more sensitive for various reasons.

Though nearby genes were shown to be dispensable for amplification activity when ACE3 was narrowed down to 510 bp [Orr-Weaver and Spradling, 1986], it was still possible that ACE3 contained transcriptional regulatory elements, such as an enhancer. In 1989, Orr-Weaver and colleagues demonstrated that transcriptional control could be completely separated from the elements that control replication, narrowing down ACE3 even further to 320 bp [Orr-Weaver et al., 1989]. The majority of amplification-inducing potential was localized to two non-essential sub-elements within the 320 bp ACE3 fragment. They also showed that ACE3 functions orientation-independently relative to other elements in the DAFC-66D locus and could be moved at least 1.5 kb away without affecting amplification. In conjunction with the earlier 2D gel data showing that the majority of replication activity was at least 1 kb downstream from ACE3, the picture that started to come into focus portrayed ACE3 not as a replication origin, but as an important regulatory element that can engage in long-range interactions with origins. In 1990, Heck and Spradling performed extensive 2D gel analyses across the DAFC-66D locus, locating 3 non-random initiation sites, none of which were inside ACE3 [Heck and Spradling, 1990]. One of the origins, named Ori-β, correlated with

the AER-d region and was responsible for 70-80% of all replication activity. Nonetheless, the Orr-Weaver laboratory demonstrated that a 440 bp ACE3 sequence, independent of other cis-elements, is sufficient for regulating ectopic DNA amplification with normal developmental specificity. They reported that the majority of initiation events in the ectopic location seem to emanate from ACE3 or directly adjacent to it [Carminati et al., 1992]. They could not strictly rule out that it was activating a very close origin, but concluded with the possibility that ACE3 could also act independently as an origin. Overall, ACE3 is sufficient to direct ectopic amplification, but is more effective when coupled with Ori-β and protected from position effects with the transcriptional insulator element "suppressor of hairy wing", su(Hw) [Lu and Tower, 1997]. In fact, with su(Hw) protection against position effects, it was shown that the 320 bp ACE3 element and an 884 bp sequence spanning Ori-β were both necessary and sufficient for DNA amplification [Lu et al., 2001]. This apparent contradiction to previous results that showed ACE3 is sufficient for ectopic amplification does not actually refute those conclusions, but potentially explains them. Lu and colleagues first showed that having only ACE3 and Ori-β between the insulating Su(Hw) sites was sufficient for ectopic amplification. Then to test whether ACE3 acts as a replication-promoting element for Ori-β in cis, they inserted a Su(Hw) site in between them, which diminished ectopic amplification levels. They then showed that deleting either ACE3 or Ori-β inside the Su(Hw)-protected constructs greatly reduced amplification. Thus, when using ectopic constructs that are flanked on both sides by Su(Hw) elements, amplification becomes dependent on only the sequence between them, preventing ACE3 from interacting with other nearby origins. They conclude by demonstrating with 2D gels that in these insulated constructs, there is no detectable replication initiation inside ACE3 and that the majority of initiation events emanate from Ori-β. Overall, these data strongly supported the hypothesis that ACE3 is not a replication origin, but is a cis-regulatory element that can interact with origins, sometimes over relatively long distances.

Trans-acting factors and their interactions with cis-elements

There were few studies on possible trans-acting factors in the early years of studying DNA amplification in Drosophila. In 1984 and 1988, Kafatos and colleagues described female-sterile mutants that suppressed DNA amplification of both DAFC-7F and DAFC-66D in trans [Orr et al., 1984, Komitopoulou et al., 1988]. However, trans-acting factors received more attention in later years. In 1997, it was shown that the *Drosophila* homolog for the yeast origin recognition complex subunit 2 (ORC2) was required for chorion gene amplification [Landis et al., 1997], also shown for ORC1 a decade later [Park and Asano, 2008]. Interestingly, the endoreplication cycles preceding chorion gene amplification proceeded normally in the ORC2-mutant background raising the possibility that ORC2 is not needed for endoreplication [Landis et al., 1997]. In support, ORC1 and ORC2 were apparently dispensable for normal endoreplication in *Drosophila* salivary glands [Park and Asano, 2008]. In 1999, the *Drosophila* homolog for Dbf4, the regulatory subunit for the Dbf4-dependent

kinase (Cdc7/DDK), was shown to be essential for chorion gene amplification [Landis and Tower, 1999] as was the *Drosophila* homolog for Cdc7 kinase very recently [Stephenson et al., 2015]. In 2000, the *Drosophila* homolog for Cdt1 ("double-parked") was shown to co-localize with ORC2 at amplification sites [Whittaker et al., 2000]. Soon after, in 2002, the homolog for MCM6 was demonstrated to be required for amplification as well [Schwed et al., 2002]. The *Drosophila* geminin homolog that normally sequesters Cdt1 as a control against re-replication is present in amplification stage follicle cell nuclei too [Quinn et al., 2001]. Geminin mutants had four sites of intense BrdU incorporation during late amplification stages and amplification continued into later stages than it does with wildtype geminin. Therefore, geminin may also have a role in regulating the end of normal DNA amplification consistent with its role in regulating against re-replication. It was suggested that other co-factors may be necessary for geminin's role in amplification, though. A protein named humpty dumpty (Hd) was shown to be required for amplification, but did not appear to have a direct role [Bandura et al., 2005]. Calvi and colleagues established that amplification stages are characteristic of high cyclin E levels normally associated with preventing re-replication [Calvi et al., 1998]. This suggested that the locus specific re-replication occurs when the normal controls are intact. Their results indicated that not only is cyclin E needed to negatively regulate re-replication globally, but that it is also required in positively regulating amplification. Nonetheless, they saw that high cyclin E levels were not sufficient to induce amplification. Interestingly, cyclin E also appears to regulate elongation during amplification. A cyclin E mutant caused the re-replication forks to move faster and travel twice as far from the replication origin during gene amplification, indicating there are not necessarily strict barriers where the re-replication forks terminate [Park et al., 2007]. In support of this, a loss-of-function mutation of the Suppressor of Under-Replication (SUUR) also results in re-replication forks traveling farther within the same developmental window [Sher et al., 2012, Nordman et al., 2014] and SUUR over-expression had the opposite effect of forks moving a shorter distance [Nordman et al., 2014]. However, neither the mutant nor over-expression affected amplification copy number [Sher et al., 2012,Nordman et al., 2014] and tethering SUUR to an amplicon origin did not affect amplification [Sher et al., 2012]. SUUR is not enriched at the amplification origins. However, SUUR tracks with replication forks emanating from DAFC-66D [Nordman et al., 2014], as does the replication protein CDC45. The data indicate that SUUR is recruited to the replication forks distal from the amplification origin, though CDC45 is enriched at the origin before the forks move out. Overall, there is abundant evidence that DNA amplification uses standard components of the pre-Replication Complex and is regulated by some of the same proteins that regulate normal DNA replication.

The parallel threads of cis-acting elements and trans-acting factors involved with gene amplification began to converge at the turn of the century. In 1999, Royzman and colleagues used antibodies against ORC2 to visualize its localization in the nucleus before and during amplification. During the endocycles that precede DNA amplification, ORC2 is seen everywhere in the nucleus, despite the possibility raised previously that ORC2 may not be necessary for endoreplication. Immediately

before the onset of DNA amplification, ORC2 localizes specifically to the chorion clusters. There is also a shift in the nuclear distribution of ORC1 at the onset of amplification [Asano and Wharton, 1999]. Royzman and colleagues introduced two other trans-acting factors into the amplification story as well. Specifically, they found that mutations in E2F and DP, two proteins that form a heterodimer and typically act as a transcription factor, result in the loss of specific ORC recruitment to the chorion clusters as well as diminished amplification levels [Royzman et al., 1999]. They suggested that what may be important mechanistically is that ORC is retained at amplification origins when it is cleared from the majority of origins across the genome. That same year ORC was shown to directly bind ACE3 as well as AER-d, which co-localizes with Ori-β, and ACE1 from DAFC-7F [Austin et al., 1999]. Thus, ACE1 and ACE3 function at least in part as ORC binding sites. Deletion mapping indicated that ACE3 contains multiple partially redundant elements important to ampification [Zhang and Tower, 2004], and deletion mapping of Ori-β identified two essential elements: one at the 5' end and another A/T-rich stretch at the 3' end. The latter element, referred to as the β-region, shares sequence homology with ACE3 and was shown to bind ORC in vitro [Austin et al., 1999]. A mutant of the trans-acting factor named satin has reduced amplification levels and eliminates the detection of ORC foci by immunofluorescence [Zhang and Tower, 2004]. The dbf4 homolog (chiffon) was shown to be important in ORC localization [Zhang and Tower, 2004] as well, in contrast to its typical placement downstream in the initiation pathway. The diffuse ORC staining in the nucleus during endocyles does not become focalized at the onset of amplification in dbf4 mutants [Zhang and Tower, 2004]. The authors favored a two-step model where a small amount of ORC binds ACE3 and the β-region of Ori-β followed by dbf4-dependent catalysis of ORC recruitment and spreading, which forms an ORC-containing chromatin structure responsible for the large foci at both endogenous and ectopic amplification loci detected by ORC staining. The minimal amount of ORC binding needed to nucleate this ORC spreading could be a result of selective retention of ORC at ACE3 and Ori-β when it is cleared elsewhere in the genome [Royzman et al., 1999].

In 2004, Remus and colleagues demonstrated that *Drosophila* ORC only had a six-fold higher affinity, at most, to ACE3 and Ori-β than non-specific DNA in vitro [Remus et al., 2004]. Interestingly, the affinity to ori-β was higher than to ACE3. Nonetheless, specific binding sites were not observed with entire DNA fragments containing these elements being protected from DNase. ORC appeared to be a general DNA binding protein that lacked sequence-specific discrimination, potentially consistent with the ORC spreading model [Royzman et al., 1999, Zhang and Tower, 2004]. In contrast, ORC had a 30-fold higher affinity for negatively super-coiled DNA than linear DNA, demonstrating that DNA topology may be important [Remus et al., 2004]. Overall, these data suggested that ORC was not capable on its own to target specific sequences for replication, such as the chorion origins, and that other factors must play a role. In 2001, Bosco and collaborators looked at mutants for retinoblastoma (RB), a protein that interacts with the E2F-DP heterodimer. RB mutants failed to limit DNA replication resulting in more endoreplication and higher amplification levels [Bosco et al., 2001]. They showed that E2F and RB co-immunopreciptate with ORC

in follicle-cell containing ovary extracts, suggesting they form a complex with ORC. Moreover, E2F was shown to bind the chorion amplification locus in vivo with chromatin immuno-precipitation (ChIP). ORC and E2F were shown to directly bind ACE3. Another E2F isoform, E2f2, was shown to regulate the transition from endoreplication to locus-specific gene amplification [Cayirlioglu et al., 2001]. E2f2 mutants complete endoreplication, but DNA synthesis does not become restricted to amplification loci thereafter, and amplification levels are reduced. Instead DNA synthesis appears more widespread, though an additional S-phase is not completed. Consistently, the *Drosophila* homologs for the replication proteins ORC2, ORC5, and CDC45 become distributed more broadly throughout the nucleus in E2f2 mutants instead of localizing to amplification loci. They propose that the loss of E2f2 might lead to an increase in expression of replication genes that are normally limiting, consistent with a model of limited preRC assembly confined to amplification loci. A more comprehensive microarray study supported the idea that E2F mutants result in up-regulated expression of replication genes and more dispersive replication initiation [Cayirlioglu et al., 2003]. Interestingly, a related idea was considered in the 1980s by de Cicco and Spradling [de Cicco and Spradling, 1984]. Essentially they wondered if there is a limiting amplification-specific trans-acting factor that is titrated out as amplification levels rise, thereby eventually stopping the amplification process. However, de Cicco and Spradling noted that the additional amplification at ectopic loci did not decrease amplification levels at the endogenous locus, suggesting that the trans-acting amplification factors are in excess [de Cicco and Spradling, 1984]. Other trans-acting proteins were found to have roles in locus-specific amplification and direct interactions with cis-elements as well. Beall and colleagues identified a Myb-containing complex of 5 proteins that associated with ACE3 and Ori-β and found that the absence of Myb seems to prevent amplification even though members of the preRC (ORC2 and Cdt1) appear to become localized presumably to the amplification loci [Beall et al., 2002]. These data demonstrated a role for a transcription factor in regulating DNA replication. They speculated that Myb seems to act after preRC recruitment, potentially recruiting an acetylase or deacetylase consistent with known interactions of Myb proteins. Mutants for Mip130, a member of the 5 subunit Myb-containing complex, show dispersive rather than locus-specific DNA synthesis in the gene amplification stages [Beall et al., 2004]. Mip130 and Myb were suggested to have direct roles in amplification origin activation and repression.

The early research consisted of transposon-based genetics, electron microscopy, and eventually 2D gels. In the late 1990s, fluorescence microscopy became a popular way to visualize gene amplification. For example, in 1998, Calvi and colleagues introduced a new microscopy technique to detect chorion genes during amplification by combining in vivo incorporation of the nucleotide analog, BrdU, by active DNA polymerases with Fluorescence In Situ Hybridization (FISH), which allows one to label specific loci using complementary nucleotide sequences [Calvi et al., 1998]. This allowed them to correlate DNA replication activity with specific loci, namely the chorion clusters. In addition to confirming the incorporation of BrdU at DAFC-66D and DAFC-7F, they also noticed two loci that seemed to undergo lower levels of amplification that were distinct from where their chorion

FISH probes were bound, raising the possibility of more amplicons to be discovered. They also were able to show that DAFC-66D, which is the larger amplicon, begins amplifying before the end of the endoreplication cycles contrary to previous data from less sensitive Southern blotting techniques. In 2001, Calvi and Spradling used 3D confocal microscopy to investigate whether the position in the nucleus has a role in regulating DNA amplifcation in follicle cells. This was in response to a rising concern that metazoan replication origins did not have specific sequence requirements and might instead respond to chromatin cues or privileged compartments inside the nucleus. However, they did not find support for the privileged compartment hypothesis, finding that amplification could take place in a diverse set of nuclear positions. In 2002, Claycomb and colleagues demonstrated that initiation of gene amplification and elongation occur in two distinct developmental windows [Claycomb et al., 2002]. Using quantitative Polymerase Chain Reaction (qPCR), they demonstrated that re-replication initiation near ACE3 ends in a relatively early developmental window, with all subsequent amplification of adjacent sequences in later stages arising strictly from elongation. Moreover, as opposed to the ~64-fold amplification levels near ACE3 originally reported using Southern blots, qPCR demonstrated maximum amplification levels closer to ~32-fold [Claycomb et al., 2002], an estimate also supported by microarrays [Claycomb et al., 2004, Kim et al., 2011], though it was estimated closer to 48-fold using Illumina sequencing [Yarosh and Spradling, 2014]. Employing high-resolution confocal and deconvolution microscopy, the Claycomb study showed that ORC is only present for initiation, whereas the elongation factors PCNA and the MCM complex track with the "double bar" structures that arise from visualizing the incorporation of BrdU by elongating forks moving bi-directionally away from the origin [Claycomb et al., 2002]. To their surprise, the Cdt1 homolog, which typically functions as an initiation factor, also moved along with the elongating forks. This may be novel, but later reports have also suggested that, in normal replication, PCNA can recruit Cdt1 to chromatin in a pathway targeting Cdt1 for proteolysis [DePamphilis et al., 2006]. The N-terminus of *Drosophila* Cdt1 has ten potential CDK phosphorylation sites, which have an inhibitory effect on amplification when mutated [Thomer et al., 2004]. The Cdt1 mutant even inhibited DNA synthesis in the later elongation-only stages, supporting the notion that it has a unique role in elongation during amplification. In contrast, the C-terminus of Cdt1 when expressed without the N-terminus enhances amplification. The authors propose that, taken together, this suggests the N-terminus is typically phosphorylated during amplification in a way that enables the C-terminus to act, possibly by inducing a conformational change. Interestingly, mutating the potential phosphorylation sites causes re-replication in cycling cells, the opposite of what would be expected given it inhibits locus-specific chorion gene amplification. Nonetheless, this suggests distinct phosphorylation patterns of Cdt1 could differentially regulate normal replication and gene amplification.

The rise of system-wide approaches to studying gene amplification in Sciara

In the past 10-15 years, the genomics revolution has been important in delineating system-wide characteristics of gene amplification. In 1998, Calvi and colleagues noticed two additional loci that appeared to be amplifying that did not co-localize with their FISH probes to the two known amplicons, DAFC-66D and DAFC-7F. In a subsequent study, four amplification loci were also observed in geminin mutants [Quinn et al., 2001]. Claycomb and her collaborators used microarrays in 2004 to identify two new amplicons called DAFC-30B and DAFC-62D, which were shown by FISH to likely be those that were observed in previous studies [Claycomb et al., 2004]. Two additional amplicons, DAFC-22B and DAFC-34B, for a total of 6, were identified by Jane Kim and colleagues in 2011 using a microarray that represented a larger area of the genome [Kim et al., 2011]. All four new amplicons span a total of 75-100 kb as seen for the original amplicons. DAFC-62D, 30B, 22B, and 34B amplify 6-fold, 4-fold, 3-fold and 5-fold respectively [Claycomb et al., 2004, Kim et al., 2011], which is lower than the 14-fold and 30-fold amplification detected for the original amplicons. DAFC-62D and DAFC-34B undergo amplification initially coincident with the others in the early amplification stages (10b), but have additional amplification initiation activity in a later stage (stages 12-13) formally thought to only be characterized by elongation and that has no detectable ORC by immunofluorescence [Claycomb et al., 2004, Kim and Orr-Weaver, 2011]. An origin in DAFC-62 was identified and named ori62, which is bound by ORC in early amplification stages (stage 10a-10b) like the other amplicons [Xie and Orr-Weaver, 2008]. However, in contrast to the other DAFCs, ORC is also bound to ori62 in late stages consistent with its late round of re-replication. ORC was also found bound 3 kb upstream in all stages and to another site 3.5 kb downstream only in the early stages. All three ORC binding sites were shown to be essential for both rounds of amplification. The MCM complex was broadly distributed around ori62 in the early stage, dissociating in the later stage 12 and reappearing at ori62 and the upstream site in stage 13. In contrast to the other amplicons that are intergenic, ori62 resides inside of a gene that is transcribed strictly in stage 12, which appeared to be critical for the late round of amplification (stage 13) in α-amanitin experiments [Xie and Orr-Weaver, 2008], but not in promoter deletion experiments [Hua et al., 2014]. Inhibition of transcription prevents late stage MCM loading in DAFC-62D as well, but α-amanitin does not appear to negatively affect any of the other amplicons [Xie and Orr-Weaver, 2008, Hua et al., 2014]. In contrast the late replication round for DAFC-34B is not dependent on transcription, and it seems to happen in the absence of ORC binding [Kim and Orr-Weaver, 2011]. Though late amplification of DAFC-34B did not show ORC enrichment, an ORC mutant abolished both early and late amplification. Moreover, both early and late rounds show MCM enrichment. However, MCM enrichment was absent during the intervening stages between the first and second round of re-replication, ruling against MCMs being pre-loaded in an earlier stage. Overall, the analyses on DAFC-62D and DAFC-34B demonstrated that amplicons are differentially regulated, each offering unique opportunities to study DNA replication [Claycomb et al., 2004, Kim and Orr-Weaver, 2011].

Analysis of RNA levels in amplification stage using RNA-seq demonstrated that amplified genes

are not all highly expressed and that most highly expressed genes are not amplified [Kim et al., 2011]. Using ORC ChIP-chip on early amplification stage follicle cells, nearly 100 regions were identified as ORC binding sites [Kim et al., 2011]. This indicated that ORC localization is not enough to specify amplification. It also indicates that ORC is not only selectively retained at amplification origins [Royzman et al., 1999]. However, ORC may serve other functions at non-amplification loci, such as cohesin loading or in maintenance of heterochromatic silencing [Pak et al., 1997,Huang et al., 1998,Takahashi et al., 2004]. The amplification loci had some of the strongest ORC signal consistent with the strong ORC foci detected by immunoflourescence [Zhang and Tower, 2004]. However, it is pointed out that rather than being a cause, stronger ORC signal could be a consequence of amplification producing more binding sites for ORC for subsequent rounds of re-replication [Kim et al., 2011]. ORC ChIP-chip signal spans 10-30 kb centered on the peak of amplification at all 6 amplicons [Kim et al., 2011], consistent with the ORC-spreading hypothesis presented previously [Zhang and Tower, 2004]. Nonetheless, this result was in contrast to observing ORC binding at specific elements in the amplicons [Austin et al., 1999, Xie and Orr-Weaver, 2008]. ORC localization to the two original amplicons, DAFC-66D and DAFC-7F, dramatically increased from pre-amplification stage into the amplification stage suggesting an active recruitment mechanism at these loci [Kim et al., 2011]. In contrast, ORC localization remained similar between the stages for the smaller amplicons.

In addition to DNA topology [Remus et al., 2004] and interactions with transcription factors [Bosco et al., 2001,Cayirlioglu et al., 2001,Beall et al., 2002,Cayirlioglu et al., 2003,Beall et al., 2004], the focus of research has also turned to chromatin regulation to help explain locus-specific re-replication. Follicle cells have been immuno-labeled at the beginning of amplification using antibodies against acetylated histone H4 (AcH4), acetylated histone H3, and the specific acetylation of lysine 8 on H4 (H4K8Ac) [Aggarwal and Calvi, 2004] as well as H4K16Ac and H4K21Ac [Liu et al., 2012]. In contrast to a diffuse staining before amplification, at the onset of amplification four salient acetylated histone foci were seen, amongst lower levels of label across the nucleus, that co-localized with ORC2 and that likely corresponded to four amplicons. The intensity of acetylation at these foci was present at the onset and during re-replication, persisting at the origin after the forks have left. Similar results were seen specifically for acetylation of lysines H4K5, H4K8, H4K12, H3K14 [Hartl et al., 2007]. ACE3 and Ori-β were specifically shown to harbor AcH4 with ChIP [Aggarwal and Calvi, 2004], and in later studies it was shown that histone acetylation is broadly distributed at the amplicon loci similar to ORC [Kim et al., 2011, Liu et al., 2012]. Directly supporting these observations, histone H4 acetylation and H4K8 acetylation in particular were most enriched at the two original chorion amplicons in a genome-wide study [Kim et al., 2011], as was also seen for H4K16Ac and H4K21Ac [Liu et al., 2012]. Histone acetylation staining of DAFC-66D was rapidly lost at the late amplification stage (stage 12), a time when ORC is gone and there is no more replication activity.

Aggarwal and Calvi showed that a mutant of the histone deacetylase (HDAC) Rpd3 led to hyperacetylation of histones, dispersive DNA replication, and redistribution of ORC throughout the

genome [Aggarwal and Calvi, 2004]. They also tethered Rpd3 to DAFC-66D, which inhibited amplification [Aggarwal and Calvi, 2004]. Similarly, tethering Polycomb (Pc), a repressive chromatin complex that associates with HDACs, to DAFC-66D also inhibited amplification. In contrast, tethering the histone acetyltransferase HAT1/Chameau (the *Drosophila* ortholog to HBO1) enhanced amplification. Interestingly, Myb, which was shown to bind ACE3 [Beall et al., 2002], is known to act with HATs and HDACs, and could potentially be important for regulation of histone acetylation at DAFC-66D.

Normal H4 acetylation in the amplicon loci is seen in ORC mutants that are defective for amplification, but H4 acetylation is delayed in Cdt1 mutants [Hartl et al., 2007]. This suggested that high levels of amplification and ORC were not necessary for H4 acetylation. In contrast, H4K12ac and H4K56ac may partially depend on ORC binding, suggesting that there may be more than one HAT involved [Liu et al., 2012]. Supporting this idea, both Hat1/Chameua and another HAT called CBP bind to amplification origins [McConnell et al., 2012]. Depletion of each separately has mild effects on amplification, but the double deletion more drastically reduces histone acetylation at amplification origins as well as amplification levels. Nonetheless, general H4 acetylation was determined to not be specific to amplicons in a genome-wide study indicating it is not sufficient for inducing amplification [Kim et al., 2011]. In addition, H4 acetylation was shown to be unnecessary for amplification at the four newly identified amplicons although it appears to potentially be necessary for inducing amplification at the chorion amplicons and correlates with amplification levels [Kim et al., 2011, Liu et al., 2012]. Of interest, one study proposed that amplicon origins are acetylated while they are active at a time when nearby genes are inactive, and become deacetylated when amplification ends and the nearby genes turn on [Liu et al., 2012]. This is in contrast to early electron microscopy results that suggested transcription is highly active during amplification [Osheim et al., 1988].

Other studies have looked at other areas of the histone code. High levels of another histone modification, phosphorylation of H1, in follicle cells was shown to be correlated with gene amplification, though it is not necessarily at the amplicons [Hartl et al., 2007]. The histone variant H3.3 is abundant at amplicon origins prior to origin activation, and at DAFC-66D abundance is highest over ACE3 and Ori-β [Paranjape and Calvi, 2016]. H3.3 is deposited at DAFC-66D in pre-amplification stage follicle cells, but ORC is bound there as well, and the authors were unable to conclude H3.3 deposition precedes ORC binding. Even so, this raised a possibility of H3.3 designating DAFC-66 for amplification. However, null mutations for H3.3 did not affect DNA amplification [Paranjape and Calvi, 2016]. Therefore, H3.3 cannot be necessary for origin selection nor for origin activation during re-replication in *Drosophila* follicle cells.

A related focus on nucleosome positioning in gene amplification loci raised some interesting observations [Liu et al., 2015]. They saw that ORC was typically associated with nucelosome depleted regions (NDRs), but NDRs were not sufficient to specify ORC binding sites. ACE3 and

Ori-β, previously shown to bind ORC in vitro and in vivo, were NDRs consistent with their possible role in nucleating ORC spreading along chromatin [Zhang and Tower, 2004]. However, not all cis-regulator elements previously studied at the DAFCs were depleted of nucleosomes [Liu et al., 2015]. It is possible that amongst the many copies of DNA at the amplification loci, some of the copies are depleted for nucleosomes at these sites and some of them are not. Variation across the population of cells could mask part-time NDRs as well. However, it is also possible that those cis-regulatory elements attract nucleosomes and their importance to amplification is in helping to ensure proper positioning of neighboring nucleosomes. Strong nucleosome positioning may in turn ensure that other cis-elements are exposed. Consistent with this notion, Liu and colleagues used an algorithm that predicts nucleosome positioning given a DNA sequence and found that sequence alone was sufficient to predict the nucleosome positioning profiles across the amplicons [Liu et al., 2015]. This indicates that the sequences of the nucleosome-covered cis-elements favor that state. In addition, using in vitro ORC binding data from Remus and colleagues, they saw that the profile of slight preferences that ORC had to naked DNA fragments spanning DAFC-66D was inversely related to the nucleosome profile there [Liu et al., 2015]. This suggested that ORC slightly favors sequences that are strongly disfavored by nucleosomes. Interestingly, they suggest that the topology of the AT-rich stretches in ACE3 and Ori-β might be what is favorable to ORC and unfavorable for nucleosomes. Consistently, bent DNA, which typically disfavors nucleosomes and correlates with origins, has been mapped inside ACE3 and Ori-β [Gimenes et al., 2009]. Overall, the exploration for amplification-specific chromatin modifications and configurations is ongoing.

An integrated view of amplification in follicle cells

Given the vast amount of work, the picture that is starting to emerge for locus-specific re-replication in *Drosophila* follicle cells is one of nucleosome-depleted elements, such as ACE3 and Ori-β, nucleating ORC binding, which may be facilitated by interactions with transcription factors such as E2F-DP. Nucleated ORC binding then spreads 10-30 kb across the amplicon locus possibly facilitated by dbf4 and histone acetylation, the latter of which may be regulated by Myb interactions with HATs and HDACs. However, a non-mutually exclusive possibility for Myb and regulation of histone acetylation may be in the downstream step of re-replication origin activation. Since histone acetylation is not necessary at the more minor amplicons for re-replication, it is possible it just makes re-replication more efficient, consistent with the higher amplification levels of acetylated amplification loci. In part, limiting preRC components may help ensure that there is no re-replication at non-specific origins across the genome, consistent with over-expression studies of preRC components that result in over-replication. However, the machinery needed for amplification is not so limiting in wild type that ectopic amplification can titrate enough factors away from the endogenous locus for a noticeable effect on amplification. Instead, the end of amplification seems to be actively regulated by factors such as geminin and Rb. High cyclin E levels appear to be necessary for blocking

non-specific re-replication. Rb, which binds ACE3 and interacts with both MCM7 [Machida et al., 2005], ORC and E2F [Bosco et al., 2001], can inhibit cyclin E [Hartl et al., 2007], so it will be interesting to explore all of those relationship further. Perhaps Rb both helps prevent the normal regulation against re-replication at DAFC-66D while also taking that role for itself to limit the amount of amplification there, maybe by eventually blocking E2F-DP and ORC from further preRC recruitment and assembly. Why both ACE3 and Ori-β are needed for amplification when Ori-β is the major initiation site and is also able to bind ORC on its own might reflect cooperative local looping that facilitates stronger ORC binding. It is also possible that ACE3 may be the main site that nucleates ORC recruitment and preRC assembly prior to spreading, which becomes inefficient without it, and that the ORC-chromatin near Ori-β is the most suitable for full preRC assembly and activation. The relatively long distance between ACE3 and the major initiation site Ori-β could also potentially reflect MCMs loading at ACE3 and MCM sliding [Gros et al., 2015, Powell et al., 2015] along the chromatin with Ori-β reflecting an area where MCMs tend to concentrate. Though this is attractive, the presence of ORC across this locus seems to suggest that MCM sliding need not be invoked to imagine how MCM concentration can build up at Ori-β. Why Cdt1 travels with the re-replication forks is unknown, but perhaps Cdt1 is needed at the fork to load MCMs if the fork breaks down assuming there is not an abundance of dormant origins to deal with such replication stress as there is in the S-phase of cycling cells [Woodward et al., 2006, Ge et al., 2007, Blow et al., 2011, McIntosh and Blow, 2012]. N-terminal phosphorylation mutants also affected elongation, and this may reflect the inability to re-initiate broken down forks. Despite all the progress, particularly for DAFC-66D, the search continues for amplification-specific factors and regulatory marks. Studies on new amplicons has made it clear that there may not be only one way locus-specific amplification is achieved.

DNA re-replication in the salivary glands of the fungus fly, *Sciara coprophila*

Early characterization of the DNA puffs

Charles Metz began studying the giant polytene chromosomes in the larval salivary glands of *Sciara coprophila* and a related species *Sciara ocellaris* in 1934, noting that the unusually large size of Sciarid polytene chromosomes made them quite suitable for understanding the nature of polytene chromosome structure and organization [Metz and Gay, 1934b,Metz and Gay, 1934a,Doyle and Metz, 1935a,Doyle and Metz, 1935b]. Explaining the large size, it was later observed that Sciarid polytene chromosomes undergo up to 12 rounds of endoreplication, reaching 8,192 total copies of each chromosome held closely together [Rasch, 1970a]. Sciarid salivary gland polytene chromosomes differ than polytene chromosomes seen in *Drosophila* in that they are "free" and not attached together by a "chromocenter". Like *Drosophila* polytene chromosomes, the chromosomes can be differentiated

by banding patterns. In 1938, Poulson and Metz observed giant DNA puff structures in Sciarid salivary gland polytene chromosomes in the fourth larval instar, shortly before pupation [Poulson and Metz, 1938]. The loci of the puff structures were dynamic, sometimes blending in with the rest of the polytene chromosome and other times appearing swollen, having expanded out into massive puffs. They observed at least ten regions that undergo this expansion, indicating that if one region was puffed, then they all tended to be puffed. These loci contained 10 to 30 bands each, but the banding pattern vanished in the puffed state. Some of the puff structures were better described as 'bulbs', and sometimes the bulbs were present when other puffs were not. Over two decades later, the DNA puff structures were determined to be sites of "extra DNA" that underwent disproportionate DNA synthesis from "extra replications" [Swift, 1962, Gabrusewycz-Garica, 1964, Crouse and Keyl, 1968, Rasch, 1970b, Rasch, 1970a]. These observations followed suit after similar conclusions were made by Breuer, Pavan, Ficq, Rudkin, and Corlette for a related species, *Rhynchosciara angelae* [Breuer and Pavan, 1955, Ficq and Pavan, 1957, Rudkin and Corlette, 1957]. These insects were responsible for discounting the 'rule of DNA constancy', demonstrating that specific parts of a genome can become amplified over others. The important work done in parallel in other dipteran insects, such as *Rhynchosciara* and *Bradysia hygida*, by many researchers including Pavan has been summarized elsewhere [Simon et al., 2016], and the focus here will be on the work done in *Sciara coprophila*. In large part, elucidating the nature of DNA amplification in the DNA puffs of *Sciara coprophila* was championed by four women: Helen Crouse, Natalia Gabrusewycz-Garcia, Ellen Rasch, and later, Susan Gerbi. The first three conducted a series of separate studies using autoradiographic [Gabrusewycz-Garica, 1964], microspectrophotometric [Crouse and Keyl, 1968], and two-wavelength cytophotometric [Rasch, 1970a, Rasch, 1970b] approaches on *Sciara coprophila* that concretely demonstrated DNA amplification in the DNA puffs. Whereas these studies were cytological and often descriptive in nature, later studies lead by Gerbi would begin to unravel the molecular nature of DNA amplification at one of the DNA puffs, II/9A, demonstrating that the repeated re-firing of a replication origin was responsible for intrachromosomal DNA amplification at that locus.

In 1964, as part of her PhD thesis work, Gabrusewycz-Garcia prepared cytological maps of the polytene chromosomes spanning 5 stages of the fourth larval instar when puff expansion takes place [Gabrusewycz-Garica, 1964]. The haploid complement of the *Sciara* genome in somatic nuclei has four chromosomes, one of which is a sex chromosome and three that are autosomes, that were named X, II, III, and IV, respectively, by Helen Crouse in 1943 [Crouse, 1943]. Gabrusewycz-Garcia divided the four chromosomes up into zones assigned based on chromosome length with 14 zones for X and II, 15 for III, and 20 for IV. She observed eleven puffs: II/2B, II/9A, II/6A, II/11A, III/2B, III/10A, III/11A, III/15B, IV/12A, IV/15B, and IV/19B. These puffs spanned 9–20 bands on the polytene chromosomes at maximum expansion. The first six, from chromosomes II and III, were noted to be the largest. Moreover, she noted that the anterior part of the salivary gland had some different puffs than the posterior part. For example, the DNA puff at locus 11A on chromosome

II (II/11A) as well as puff IV/15B are small or absent in the anterior portion, but large in the posterior. In contrast, puffs III/2B, III/15B, and IV/12A are large in the anterior and small or absent in the posterior. She observed 10 'bulbs' on chromosome X, 8 on II, 7 on III, and 10 on IV. Whereas puffs were Feulgen-stain positive with increasing staining in puff expansion, the bulbs were largely non-staining structures with tiny threads of stain scattered throughout. However, at the time of maximum puffing, the bulbs become Feulgen-positive and simultaneously increase uptake of tritiated thymidine at a comparable rate to puffs. Moreover, bulb-like structures sometimes form in pre-puffing stages at puffing loci. These observations lead her to question the distinction between puffs and bulbs. Perhaps some of the bulbs could be explained as late stage DNA puffs. Later studies seemed to recognize that bulb structures were typically RNA puffs [Gabrusewycz-Garcia and Kleinfeld, 1966, Gabrusewycz-Garcia, 1968, Gabrusewycz-Garcia and Mariano Garcia, 1974]. In 1971, Gabrusewycz-Garcia revisited her chromosome maps and puff charting. Each chromosome was divided into zones again, though apparently differently. Fortunately, the re-organization does not appear to have affected puff locations from her previous work except IV/19B, which became IV/19A. In this study, there were 9 zones for chromosome II, 10 for III and X, and 13 for IV. The zones were subdivided into as many as three sub-regions per zone denoted A, B, and C. This time she identified 9 major and 9 minor puffs, somewhat arbitrarily defined. Major puffs expanded to an average diameter of 10 microns. Minor puffs were defined as having smaller than 10 micron expansions and a higher variability in both size and occurrence. She identified seven new puffs: II/13A, II/14B, IV/5C, IV/8C, IV/10B, X/7A and X/11B. All were in the minor puff category, and only one varied across the anterior and posterior sections of the salivary gland. That was the posterior-located II/14B. It was tempting to assume the smaller puffs amplified DNA to a lesser extent, but there was no evidence to support this notion.

At the time of the 1964 study, the recent results from *Rhynchosciara* and *Sciara* were being debated. There was disagreement on whether or not they demonstrated disproportionate DNA synthesis in the puffs relative to the rest of the chromosome [Breuer and Pavan, 1955, Ficq and Pavan, 1957, Rudkin and Corlette, 1957, Swift, 1962]. The previous results from Rasch and Swift (1962) in *Sciara* were from microspectrophotometry. To weigh in on the debate, in this study Gabrusewycz-Garcia employed an autoradiographic approach where salivary glands are incubated in medium supplemented with the DNA precursor tritiated thymidine for 10-30 minutes to be incorporated into the polytene chromosomes by active replication forks. The polytene chromosomes were then exposed to autoradiographic film for 1-2 weeks before being developed. The distribution of autoradiographic signal, or "grains", along the polytene chromosomes was then used to determine if there was significantly more signal in the DNA puffs or if it was uniformly distributed throughout. She used female larvae specifically because their polytene chromosomes were larger and stained better. Female larvae have been the primary source of data on DNA puffs ever since. To stage the larvae, she used the rows and columns of "eye spots", which are the anlage of the adult eye. The eyespots are located on the top of the larvae, right behind the head capsule, and can be used to estimate

what stage in the puffing process the salivary gland polytene chromosomes are in, essentially by counting the number of columns and rows that make up a triangular matrix of black dots. The five stages of larvae she looked at were pre-eyespot, 7x2, 7x2–10x5, 10x3–13x7 where puffing becomes apparent, and the final stage of maximum puff expansion, 10x5–14x7. Using the autoradiographs of tritiated thymidine uptake, Gabrusewycz-Garcia was able to plot relative grain densities as histograms along chromosomes using the zones as bins [Gabrusewycz-Garica, 1964], not unlike read depth in the modern genomics era. In doing so, polytene chromosomes could be characterized by one of three patterns. In the most frequent pattern called 'E', grains were distributed across the entire chromosome lengths. In the second pattern called 'C', grain density peaked at centromeres, some chromosome ends, and some other bands that are characterized by heavy Feulgan staining. The loci encompassed by pattern C were largely heterochromatic. In the third pattern called 'P', tritiated thymidine uptake peaked selectively in puffs, bulbs, and some other lighter staining bands that are seen to be swollen on occasion. Whether or not a puff locus had grains depended on it being in the expanded state or not. In some ways the profiles for pattern C were the inverse of pattern P. There were chromosomes that seemed to show intermediate steps, connecting P with E and E with C. Gabrusewycz-Garcia favored the model that these patterns followed a developmental ordering of P, E, C. She posited that P marked the beginning of the DNA synthesis period inside DNA puffs, E marked an elongation period, and C marked the termination of DNA synthesis, particularly because heterochromatic regions were known to replicate late in normal S-phases. However, other orderings could not be ruled out. In fact, it appeared that pattern C occurred in earlier eyespot stages and P occurred in later ones. Moreover, in most stages there was minimally a small amount of DNA synthesis distributed across the polytene chromosomes, and in the final stages, only pattern P was observed with DNA precursor uptake exclusively in puffs and bulbs. It is possible that she was observing the ending stages of endoreplication in patterns E and C and pattern E might consist of both endoreplication elongation and amplification elongation stages. Overall, this study demonstrated asynchronous, differential, localized DNA synthesis in the DNA puffs, consistent with the findings of Rasch and Swift (1962).

In 1966, two years later, Gabrusewycz-Garcia used similar techniques to study active transcription in the polytene chromosomes with a radio-labeled RNA precursor [Gabrusewycz-Garcia and Kleinfeld, 1966]. She found that immediately before puff expansion, a period of intense RNA synthesis occurs in nucleolar and micronucleolar compartments as well as bulbs. Interestingly, the heterochromatic sites of pattern C were associated with micronucleolar compartments. RNA synthesis begins in the DNA puffs at the onset of puffing, and is associated with micronucleoli, RNA-containing bodies adjacent to the overall chromosome structure that often appear to be attached by chromatin strands that traverse the body. After maximal puff expansion, intense RNA synthesis remains high in puffs and bulbs, but is low in the pattern C sites. At this stage, the RNA synthesis in the puffs is entirely intrachromosomal, occurring within the bounds of the puffed out structure, and micronuclei are absent. In later studies, puffing was broken up into 4 stages (I-IV), from pre-puff (I),

to initial puff-expansion (I), maximal puff expansion (III), and puff regression (IV) [Gabrusewycz-Garcia, 1968, Gabrusewycz-Garcia and Mariano Garcia, 1974]. Radio-labeled DNA precursor was intensely taken up in stages I-II and declined through stages III-IV whereas radio-labeled RNA precursor was taken up in stages II-III and became undetectable in stage IV [Gabrusewycz-Garcia, 1968, Gabrusewycz-Garcia and Mariano Garcia, 1974]. Overall, puffing morphology and DNA amplification was associated with RNA accumulation and puff expansion was correlated with active RNA synthesis. Other systems, such as *Xenopus*, were also known at the time to generate extra DNA to facilitate intense RNA synthesis of rRNA. Thus, one hypothesis from these studies was that amplification in DNA puffs could be of rDNA as well. However, early studies of in situ hybridization by Susan Gerbi in the laboratory of Joseph Gall demonstrated that the DNA in these puffs did not seem to contain rDNA [Pardue et al., 1970]. Thus, it was still an open question on the function of amplified genes in Sciara. A study from Been and Rasch on protein content in salivary glands during the larval-pupation transition suggested that no new proteins were made in correlation with RNA and DNA puffing, nor were proteins present in previous stages lost during puffing [Been and Rasch, 1972]. However, there were quantitative differences between stages. Later molecular studies of the genes in puff II/9A suggested they were structural proteins important for the pupal coat in the next stage of development [DiBartolomeis and Gerbi, 1989].

In 1968, Crouse and Keyl used a spectrophotemetric approach, which uses light absorbency to measure DNA content, to study three DNA puffs on chromosome II: II/2B, II/6A, and II/9A [Crouse and Keyl, 1968]. After the series of endoreplication ends, they argued that if extra DNA arises by additional, localized rounds of replication, then one should be able to detect the DNA increase in a stepwise, geometric series. Higher resolution spectrophotometric equipment was available to Crouse and Keyl than had been used by Rasch and Swift (1962), prompting them to re-evaluate the nature of the extra DNA in the DNA puffs by measuring the DNA content of single bands along the polytene chromosomes. They compared the DNA content in puff regions to that in a non-puff control region spanning zones 4B-5A. Crouse and Keyl used only the anterior portions of salivary glands, and appear to have monitored DNA amplification across larval stages 10x5 to 14x7, though details were limited on eye spot staging. The study focused mainly on puff II/2B, analyzing a band called 'p', which goes through four doublings, reaching 16-fold its starting value in the final stage. The subsequent data in the paper are not so clean. Measurements of the non-puff region seem to undergo one round of doubling, possibly from a final round of endoreplication. The data for II/6A and II/9A seem to undergo 5.2-fold and 7.7-fold amplification at most. However, in that particular dataset, II/2B only amplified 5.8-fold. Therefore, it seems possible that the first measurements in this dataset were taken after 1-2 doublings had taken place. Finally, they were unable to show the perfect stepwise pattern for the other puffs. This was similar to the increases detected in an earlier study by Rasch and Swift that were not stepwise, but continuous. Detecting a continuous increase from a series of staged larvae is actually in some ways a more expected result than the stepwise doublings. To observe strict doublings, larval selection would have to be perfect, larval development

would need to be perfectly synchronized with respect to staging, salivary gland cells would need to re-replicate synchronously, and salivary gland dissection would need to strictly take place only after each doubling. Moreover, the geometric series they observed at II/2B in the first experiment depends on the perfect synchrony of both the firing of the many copies of the re-replication origin and of all corresponding elongating forks. Nonetheless, this study demonstrated differential DNA amplification levels at puff and non-puff regions, and provided early estimates of amplification levels.

Molecular nature of the re-replication origin in DNA puff II/9A

In the late 1980s, Susan Gerbi began dissecting the molecular nature of DNA puffs and amplification. In 1989, Susan DeBartolomeis with Susan Gerbi, identified the sequence of DNA puff II/9A, potentially the largest, earliest expanding, and longest lasting DNA puff. To do so, they cloned cDNA from salivary glands during the stage of most active RNA synthesis in the DNA puffs. They identified a 1-2 kb clone that mapped to the II/9A locus cytologically by hybridization and Southern blot confirmed that complementary genomic DNA underwent amplification by 5–20 fold. This will be referred to as cDNA-II/9. A genomic lambda library, where restriction fragments from the entire genome are ligated into lambda phage vectors and propagated in E. coli clones, was screened with cDNA-II/9 for clones that carried complementary genomic DNA (gDNA). This gave them a clone with a 14.7 kb gDNA insert. Hybridization of the gDNA clone to polytene chromosomes confirmed that it mapped to puff II/9A. This clone will be referred to as λ-II/9. They mapped the directionality of cDNA-II/9 across λ-II/9, and in doing so, identified two complementary genes therein, referred to as genes II/9-1 and II/9-2 (see Fig. 1.5 for the anatomy of the II/9A locus). They manually sequenced overlapping fragments of λ-II/9 as well as the II/9-1 and II/9-2 cDNA clones using the Sanger di-deoxy sequencing method. This confirmed the directionality and ordering of the genes across λ-II/9 and showed that II/9-1 and II/9-2 shared 85% identity. Sanger sequencing of total RNA and poly-A RNA using primers for II/9-1 and II/9-2 allowed them to identify a 65-67 bp intron in each. An RNase protection assay indicated there were no other introns in II/9-1, but were unable to apply this to II/9-2 due to its much lower abundance than and high sequence similarity with II/9-1. Nonetheless, bioinformatics identified only the introns that were found experimentally. Alignment of 134 bp promoter regions of both genes showed they shared 89% identity, and that each contained a TATA-box transcriptional element, a 20 bp A-rich consensus sequence for ecdysone-response genes in *Drosophila*, and two sets of short inverted repeats. Another consensus sequence was found further upstream of each gene. With the DNA sequences, they used bioinformatics to predict the protein sequences and their potential properties. The predicted proteins are each 286 amino acids (AA) long and share 76% identity with no gaps in the alignment. They differ in that II/9-1 is more alanine-rich and II/9-2 is richer in argine and lysine. Both are enriched in cysteine residues, have hydrophobic N-terminal regions, and contain a predicted α-helical coiled coil domain in a large central region. They were predicted to be secreted proteins or membrane proteins. Overall,

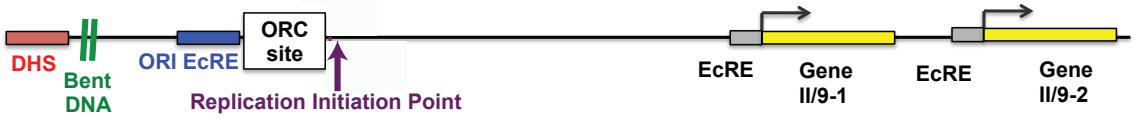


Figure 1.5: Anatomy of DNA Puff II/9A.
II/9A undergoes the most amplification of all DNA puffs and has been the only DNA puff studied at the molecular level.

the authors concluded that these proteins were most likely secreted structural proteins for the pupal coat as was seen in other insects.

In 1993, Liang, Gerbi, and colleagues used 2D gels to locate where DNA replication initiates from on II/9A [Liang et al., 1993]. A genomic library of cosmid clones was screened with gene II/9-1 to identify a clone, cII/9, with a 35 kb insert that had genes II/9-1 and II/9-2 near the center. Sub-cloned fragments of cII/9 were used to probe 2D gels of EcoRI-digested genomic DNA to identify where replication bubbles emanate from. Ultimately they found replication bubbles in a 5.5 kb EcoRI fragment from cII/9 as well as in a 4.5 kb Msc-I fragment inside it. Nonetheless, both fragments also showed fork arcs, suggesting that there was at least one more initiation site nearby. Neutral/alkaline 2D gels demonstrated that replication largely proceeds bi-directionally from within the EcoRI fragment, and they were able to narrow down the the estimated origin location to a 1 kb fragment. In a follow up study in 1994, Liang and Gerbi developed a novel 3D gel method to interrogate the origin further. The first 2 dimensions are from the standard neutral/neutral 2D gel procedure. Vertical slices are then taking from the 2D gel from where the replication intermediates are rotated 90° and ran on separate alkaline gels. This separates replication intermediates first into forks and bubbles, then into parental and nascent DNA from each. Overall, the 3D gel method led them to confirm the location of an origin within the 1 kb fragment and to conclude that only one initiation event from within the 1 kb origin occurs per DNA molecule. This ruled out competing hypotheses at the time that (i) replication initiates from many sites inside a large melted DNA region before being ligated together and (ii) DNA replication may initiate from multiple nearby sites, forming microbubbles that merge. The 3D gel method also allowed them to rule out that the fork arcs present in all their 2D gels were from broken bubbles. Thus, the 3D gel method supported the possibility of at least one more nearby initiation site. The exact start site of DNA synthesis was mapped in the II/9A origin in 2001 by Anja Bielinsky with Susan Gerbi and colleagues [Bielinsky et al., 2001]. They developed a method called Replication Initiation Point (RIP) mapping that employs the 5'–3' DNA exonuclease λ-exo. In this method, nascent strands are enriched over the depleted parental DNA background due to the RNA primers at their 5' ends that λ-exo is much less efficient at digesting. Primer extension is then used with dideoxy Sanger sequencing gels to identify the exact transition point between continuous and discontinuous DNA synthesis inside a

known origin. On the 1 kb *Sciara* origin, RIP mapping identified an exact transition point when interrogating the top strand, but a single transition point on the bottom strand has not yet been mapped. A binding site for recombinant *Drosophila* ORC as well as *Sciara* ORC in nuclear extract was identified immediately adjacent to the transition point [Bielinsky et al., 2001] (see figure 1.5). DmORC bound to this sequence in an ATP-dependent matter in vitro. ORC ChIP that queried the origin site, an upstream site, and a downstream site near gene II/9-1 suggested that ORC was only bound at the origin. This was the first time both a discrete ORC binding site and a discrete replication start site were mapped for a metazoan replication origin.

Relationship of puffing, amplification, and transcription

In 1993, Wu, Gerbi, and colleagues studied the developmental progression and temporal relationships among morphological puffing, DNA amplification, and puff gene transcription [Wu et al., 1993]. This was the first study to definitively correlate molecular events with eyespot stages. Puff expansion was seen around stage 12x6 on chromosome II at two major puffs that appeared to expand before the other DNA puffs, which were less prominent at this stage. All puffs reached maximum expansion by stage 14x7, which occurs 24 hours after 12x6. The two early-expanding DNA puffs, II/9A and II/2B, were the largest. All puffs regressed in subsequent pre-pupal stages called Edge Eye and Drop Jaw (EEDJ to describe both for shorthand). Autoradiography of radio-labeled precursor signified that the the most DNA synthesis occurred at stage 12x6 before puff expansion for most DNA puffs. They identified cDNA clones that were specific to puff stage larval salivary glands and In Situ Hybridization (ISH) mapped them to the major DNA puffs II/9A and II/2B, a minor DNA puff IV/5C, and to RNA puff III/9B. The RNA puff appeared to morphologically puff at stage 12x6. Quantitative blotting confirmed no DNA amplification for the RNA puff in pre-puff and puff (combined 12x6 to Edge Eye) stages, though it expands around stage 12x6. Northern blots showed that RNA puff III/9B transcripts appear in Edge Eye, after puff expansion. However, other genes at that locus may be actively transcribed earlier. Southern blot estimated DNA puffs IV/5C, II/2B, and II/9A to amplify 5-fold, 18-fold, and 17-fold respectively. Since IV/5C amplification was lower than the two major puffs, it was not considered further. Amplification at II/2B and II/9A was detected at stage 10x5 and plateaued by stage 14x7. The transcripts from these two DNA puffs were barely detectable until stage 12x6, peaking in abundance in stage 14x7. Thus, amplification begins before puffing and transcription correlates with puff expansion. Nonetheless, in 2002, Mok and colleagues showed that morphological puffing was independent of both active transcription and replication using inhibiting reagents and conditions for each [Mok et al., 2001].

The amplification levels of 17-18 fold detected by Southern blot [Wu et al., 1993] roughly suggest 4 rounds of re-replication, which is consistent with Helen Crouse's geometric series for II/2B [Crouse and Keyl, 1968]. In contrast, in the other part of the early Crouse paper, amplification estimates for II/2B and II/9A were 5.8-fold and 7.7-fold, respectively. Wu and colleagues suggested that the difference between their 17-18 fold estimate and Crouse's lower estimates was because they used

a high resolution technique whereas microspectrophotometry measures an average over hundreds of kilobases. Since it was known that amplicons in *Drosophila* only reached 75-100 kb, they argued that her lower resolution technique would be unable to estimate the amplification level at the peak assuming *Sciara* amplicons are of similar sizes. However, Crouse's geometric series clearly demonstrated four rounds of extra replication in one of her experiments, suggesting it was able to estimate the amplification level. It seems more likely that Crouse's lower amplification estimates in those experiments are due to beginning DNA content measurements after the first round of extra replication. In support of this view, she aimed to start her measurements at the stage right before puffing, assuming this was the baseline and that DNA amplification began with puffing. However, Wu and colleagues showed amplification precedes puffing. This indicates that Crouse started measuring during or after the first (or maybe second) round of re-replication. Perhaps coincidentally, her estimates were off by less than 2 rounds. If it is the case that her low-resolution technique is capable of accurately estimating amplification, then that would also suggest that the amplicons in *Sciara* are much wider than in *Drosophila*. This follows from the argument that the low resolution would be unable to accurately estimate the peak amplification levels of a 75-100 kb amplicon gradient, largely suffering from averaging the copy number across the gradient. If one accepts that the technique did accurately estimate the peak amplification level in the first experiment showing the geometric series, then the variation and average of the copy numbers across the measured site must have been small and close to the maximum copy number, respectively. In other words, the entire measured site roughly corresponds to the peak of the amplicon, which is 16-fold amplified, indicating the amplicon peak is wider than would be expected for a 75-100 kb Drosophila amplicon.

This work seems to raise the question of whether amplification continues into or past stage 14x7. Wu and colleagues suggested that the amplification plateaus they detected in stage 14x7 for II/9A and II/2B are likely real since (i) 2D gels detected only replication forks at II/9A, not bubbles after 14x7 in EEDJ stages [Wu et al., 1993, Liang et al., 1993], and (ii) there is less DNA precursor uptake in stage 14x7 than in 12x6. However, both points actually suggest that there was in fact DNA synthesis detected in 14x7, just less than in 12x6 and just not from within the fragment they were studying. Re-initiation could still occur in these late stages and it remains possible that the plateau in amplification that they detected is a technical error due to detection limits or is biological and from a lower rate of re-initiation. In turn, a counterpoint could be argued from them that DNA synthesis detected in 14x7 was strictly from elongation, not re-initiation. This is given apparent support from the fact that only forks were detected by 2D gels in EEDJ stages. However, the data only let us conclude that replication forks traveled from outside the sequences that were interrogated in those stages. If we accept that there were replication forks coming through puff II/9A in EEDJ, then next we can attempt to identify the most reasonable hypothesis as to where they emanate from. Since replication forks only emanate from re-replication origins after the endoreplication cycles, these forks must have come from a re-replication origin at some point in time. Since the nearest re-replication origins other than those at II/9A are very far away, on the megabase scale, and since

those long-traveling forks would have to survive a collision with replication forks heading towards them from II/9A, probability weighs in favor of the hypothesis that the forks detected by 2D gels that pass through the II/9A origin must be coming from somewhere nearby in II/9A, just not the sequence they used for 2D gels. It seems likely that the initiation zone either shifts over time, possibly as transcription begins or ends, or that there are more initiation zones clustered in the nearby area that become preferred in later stages when re-initiation becomes less efficient, explaining the plateau. Overall, it remains possible that amplification occurs at a lower rate in 14x7 and EEDJ given their data.

A role for ecdysone in re-replication in insects

In 1968, Helen Crouse was the first to investigate the role of ecdysone, also called the molting hormone [Schneiderman and Gilbert, 1964], in promoting DNA synthesis in the DNA puffs of *Sciara coprophila* [Crouse, 1968]. Up until then, Rasch and Swift had pubished their work showing extra DNA content in the puffs and Gabrucewycz-Garcia published her work detailing the disproportionate DNA synthesis in the puffs during normal development. The debate that arose was whether this was a programmed part of development or just reflected a misfiring of sick and dying cells as the salivary glands began breaking down with the onset of pupation. Crouse argued against this interpretation, observing that the DNA puffs and disproportionate DNA synthesis inside them occurred at the same specific sites in all larvae and followed a prescribed order of puffing. She reasoned that if they were developmentally programmed, then she should be able to induce them. Ecdysone is a steroid hormone released from the prothoracic gland that regulates many stages of insect development, usually stimulating growth or molting [Schneiderman and Gilbert, 1964]. There is a spike of ecdysone levels in late-stage larvae, which is the time of puffing in salivary gland polytene chromosomes, that leads them into pupation. Moreover, at the time, recent experiments from Clever and Karlson on RNA puffs in *Chironomus* showed that morphological puffing and RNA synthesis were stimulated by ecdysone injection in a dose-dependent manner [Clever and Karlson, 1960, Schneiderman and Gilbert, 1964]. She hypothesized that ecdysone might regulate puffing and DNA amplification. To test this, she injected young larvae with ecdysone. Using Feulgen staining and autoradiography, she measured puffing and DNA synthesis in polytene chromosomes from normal, ecdysone-injected, and mock-injected larvae. She was able to confirm earlier results in DNA puff staged normal larvae that the highest grain density was in the DNA puffs, though it was strewn across the polytene chromosomes as well, possibly reflecting elongating forks from the last round of endoreplication. For testing the effects of ecdysone on DNA amplification, she first used autoradiography with tritiated thymidine to determine a period when there was no DNA synthesis. Otherwise, she reasoned it would be impossible to conclude that ecdysone induced DNA synthesis. She found that at 3–4 days after the third molt, 98% of the nuclei in larval salivary glands had no detectable DNA synthesis. In contrast, when larvae that were 3 days past the third molt were injected with ecdysone, 87% of the salivary

gland nuclei showed DNA synthesis a day later. Importantly, the DNA puffs had extremely high grain density and there was evidence of synthesis across the entire lengths of polytene chromosomes. In contrast, larvae injected with a control solution were indistinguishable from uninjected larvae. In addition to disproportionate DNA synthesis, ecdysone induced the full and normal complement of morphological puffing. Overall, she saw the same results from ecdysone injection across a range of differently aged larvae from young larvae that were three days past the third molt all the way up to those with early eye-spots just prior to puff expansion. In all cases, ecdysone injection induced clear morphological puffing and DNA synthesis as measured 24 hours later and within 48 and 72 hours, the puffs were characteristic of the final stages before pupation and in early pupation, respectively. Altogether it was clear that ecdysone could prematurely induce both puffing and DNA synthesis up to seven days early with respect to the normal course of development. This demonstrated that these features were coordinated and could be triggered by a developmental cue.

In 1970, Amabis and Cabral tested the effects of limiting ecdysone exposure on puffing [Amabis and Cabral, 1970] in a related Sciarid fly, *Rhynchosciara*. Ecdysone trickles towards the salivary glands into the hemolymph from the larval brain. When the anterior portion of *Rhynchosciara* larvae, behind the brain, was ligatured to prevent the flow of ecdysone, RNA and DNA puffing was prevented compared to various control groups that were not ligatured, were ligatured and cauterized in front of the brain, or were only cauterized. They concluded that this prevented DNA amplification and transcription at the puff loci. However, strictly speaking, the evidence only supports that the ligature prevented morphological puffing. In 1971, Amabis added to this work by implanting salivary glands into larvae of different ages with one gland in the pair kept out as a control [Amabis and Simrs, 1971]. Implanting a salivary gland from young larvae into older larvae induced puffing at some sites and not at others. They reported that extra DNA synthesis was observed, but do not quantify this other than by showing a picture of chromosomes. Glands from older larvae implanted into younger larvae had mixed results, but in some cases puffs did not develop on time nor any further if they had already started. This was reported to be consistent with their previous results from culturing salivary glands in vitro.

Stocker and Pavan conducted experiments on *Rhynchosciara hollaenderi* in 1974, showing that ecdysone injection induced DNA synthesis, but it was not clear that DNA puffs were sites of markedly higher induced DNA synthesis [Stocker and Pavan, 1974]. Ecdysone injection experiments in *Rhynchosciara americana* from Berendes and Lara in 1975 showed two DNA precursor incorporation patterns. The first was a weak continuous pattern across chromosomes possibly from endoreplication. In the second pattern, grain density was more, but not fully, localized typically at DNA puffs indicative of DNA amplification there, perhaps occuring while endoplication forks from the last endoreplication are still present. When actinomycin D injection was used to inhibit RNA synthesis following the ecdysone injection, it blocked the higher incorporation rate of radio-labeled DNA precursor at amplification loci, but the continuous pattern was still just as frequently present albeit with

lower grain density. Overall, RNA synthesis seemed necessary for amplification, but not necessarily replication across the polytene chromosomes. This experiment was in part motivated by previous results from their group that injecting DNA puff stage total RNA into pre-puff salivary glands induced DNA puff formation in them [Graessmann et al., 1973]. Berendes and Lara also showed that ecdysone injection causes new discrete "fast green staining" bands at the DNA amplification sites before puffing, indicative of new nonhistone proteins locating there.

In 1984, Amabis demonstrated that ecdysone injection into the related species, *Trichosia pubescens*, induces three puff responses: early-response puffing (<2 hours post-injection), late-response puffing (>2 hours post-injection), and repressed puffing [Amabis and Amabis, 1984a]. The DNA puffs in this species primarily resided in the late-response class. Ecdysone injection also induced disproportionate DNA synthesis as seen by uptake of radio-labeled DNA precursor into the DNA puffs. Blocking protein synthesis with cyclohexamide largely did not affect puffing of early-response puffs, but completely prevented puffing of late-response puffs. They inferred based on chromosome thickness and staining intensity that DNA amplification was likely blocked, but did not do the more definitive autoradiography experiment. Nonetheless, the late-responding puffs and potentially DNA amplification are dependent on protein synthesis whereas early-responding ones can use factors in the already available protein pool. Another set of ligature experiments were done in 1984 by Amabis on *Trichosia pubescens* [Amabis and Amabis, 1984b]. This time they also did the control of injecting ecdysone-blocked larvae with exogeneous ecdysone and showed that puffing could still occur identical to ecdysone-injected larvae that did not have the ligature, both of which showed normal puffing. As with the previous study, ligations behind the brain prevented puffing when exogenous ecdysone was not injected. In subsequent experiments, the larvae were ligated in the middle of the body to compare puffing in anterior and posterior portions of the same salivary glands under different conditions. Whereas the anterior portion that was in contact with endogenous ecdysone went through puffing, the posterior portion cutoff from ecdysone did not. Injection of ecdysone into the posterior portion of the larvae rescued puffing. They went beyond morphological puffing to determine if DNA synthesis was affected by checking the grain density in the DNA puffs after radio-labeled DNA precursor uptake. Whereas larvae that were not ligatured as well as the anterior portion of ligatured larvae demonstrated high grain densities in DNA puffs representing disproportionate DNA synthesis there, the polyenes from the posterior salivary glands did not. They identified a late stage at which ligature no longer made a difference on DNA precursor uptake. Similar results were found for uptake of radio-labeled RNA precursor, but the late stage had mixed results. The anterior and posterior portions of some ligatured larvae showed similar RNA precursor uptake while other larvae still demonstrated differences where the anterior showed intense RNA synthesis and the posterior did not. This suggests that RNA synthesis depends on ecdysone a little bit longer than the disproportionate DNA synthesis.

In 1991, Brigitta Bienz-Tadmor with Susan Gerbi, showed that the promoter of II/9-1 was

inducible by ecdysone. As discussed above, studies in the 1960s showed that ecdysone injection affected puffing. For example, it induced RNA synthesis in *Chronomus* RNA puffs [Clever and Karlson, 1960, Schneiderman and Gilbert, 1964] and DNA amplification in *Sciara* DNA puffs [Crouse, 1968]. Later work from Ashburner painstakingly showed ecdysone's involvement in the RNA puffing cascade of *Drosophila* [Ashburner, 1971, Ashburner, 1972, Ashburner, 1973, Ashburner, 1974, Ashburner and Richards, 1976, Ashburner et al., 1990]. In brief, ecdysone was shown to regulate transcription in sets of early-responding puffs that were proposed to encode gene products that regulated the transcription of later-responding puffs. To test whether the *Sciara coprophila* II/9-1 promoter was regulated by ecdysone, a 718 bp II/9-1 promoter-containing fragment was fused to a reporter gene that was inserted into the *Drosophila* genome by P-element transformation. Testing whole animals for the reporter gene, they found it only in the larval stages associated with puffing in the salivary glands. A spike in reporter gene activity was detected at the time when ecdysone titers rise to stimulate early puffs. They dissected out the puff stage salivary glands from *Drosophila* larvae and found reporter gene expression only in the salivary gland fraction, not in the larval remains. Conversely, reporter gene activity was not detected when the II/9-1 promoter was left off the construct or when other promoters were tested. These experiments showed that the transgenic *Sciara* promoter was capable of being regulated by *Drosophila* in the same tissue and stage as in *Sciara*, specifically at the onset of ecdysone-induced early puff activity. Culturing dissected transgenic *Drosophila* salivary glands in vitro with and without ecdysone confirmed that reporter gene activity was rapidly responding to this hormonal cue. Inhibiting protein synthesis with cyclohexamide shortly after ecdysone-induction suggested that a short burst of transcription is responsible for most of the reporter gene activity. Amplification of the II/9 sequences and morphological puffing of the ectopic locus containing them did not occur in *Drosophila*. Overall, this study showed that II/9-1 gene activity is likely regulated by ecdysone.

In 2006, Foulk and colleagues demonstrated that injection of ecdysone induces premature DNA amplification at DNA puff II/9A as detected by qPCR [Foulk et al., 2006]. In some cases, amplification reached normal levels estimated to be approximately 17-fold in 14x7 staged larvae. Several ecdysone response elements (EcREs) were found at the II/9A locus, some which are in the gene promoters, but others that are in the origin area. There are *Sciara* strains where two of these are mutated and amplification proceeds fine (Susan Gerbi, Eric Gustafson, and Yutaka Yamamoto unpublished). There are no strains with a polymorphism over an EcRE that sits directly adjacent to the ORC binding site in the II/9A origin called the ORI EcRE (Fig 1.5). Foulk showed that the Ecdysone Receptor (EcR) was able to bind the ORI EcRE in vitro whereas it did not bind some of the other EcREs at this locus [Foulk et al., 2006]. *Sciara* EcR-A and EcR-B isoforms were subsequently cloned and antibodies were prepared against the N-terminal domain specific for each [Foulk et al., 2013]. EcR-A was shown to be the predominant form in salivary glands and increases at amplification stage. Immunofluorescence demonstrated EcR-A binding sites in the polytene chromosomes, including at DNA puffs, both at amplification and post-amplification stages [Liew et al.,

2013]. Preliminary EcR ChIP results have confirmed that EcR binds the ORI EcRE only during amplification (Susan Gerbi, Michael Ezrokhi, Victoria Lunyak, unpublished).

Similar results as described for *Sciara* were seen for a related species, *Bradysia hygida*, where transgenic lines of *Drosohila* were able to control gene expression from a DNA puff promoter in salivary glands in response to ecdysone [Monesi et al., 1998, Monesi et al., 2001, Monesi et al., 2003]. Conversely, comparable experiments with *Rhynchosciara* sequences showed no amplification and low constitutive expression in *Drosophila*, suggesting other factors were needed for regulation of each [Soares et al., 2003]. The C4 DNA puff in *Bradysia hygida* was shown to amplify 21-23 fold in salivary glands and 4.8-fold simultaneously in the prothoracic gland [Candido-Silva et al., 2015]. The amplified gene in puff C4 (BhC4-1) has been shown to be secreted in the saliva where it is present in fibrous structures [Monesi et al., 2004]. In 2002, Basso and colleagues showed that injection of ecdysone into pre-puff larvae of *Bradysia hygida* induced both gene expression and DNA amplification at that locus as detected by PCR and Southern blot respectively [Basso et al., 2002]. DNA amplification was detected approximately 24 hours after injection and peaked at 36 hours. Interestingly, more recent developments in studying amplification in *Drosophila* follicle cells suggested a role for ecdysone as well. Both isoforms of the ecdysone receptor (EcR-A and EcR-B) are expressed in the follicle cells [Hackney et al., 2007, Sun et al., 2008] and increased EcR activity is detected during the amplification stages [Hackney et al., 2007]. Mis-expression of a dominant negative mutant for EcR in amplification stage follicle cells is associated with reduced chorion gene expression as well as lower amplification levels, suggesting it is either directly or indirectly essential to amplification. There are EcRE half sites in each of the chorion gene promoters and ACE3 contains a likely EcRE motif. Moreover, activation of the EcR pathway seems to be involved with the switch from endoreplicaton to DNA amplification [Sun et al., 2008], potentially through regulation of the microRNA miR-318 [Ge et al., 2015], though the microRNA miR-7 and zinc-finger protein Tramtrack69 are also involved in the switch [Sun et al., 2008, Huang et al., 2013]. Overall, these studies in *Sciara*, *Drosophila*, and other insects suggest that not only is ecdysone involved upstream of amplification, but that the Ecdysone Receptor might be directly involved in amplification. However, there has not yet been direct evidence of the latter.

Trans-acting factors, chromatin, and DNA topology at DNA puff II/9A

An early study in 1973 on non-histone chromosomal protein methylation in the polytene chromosomes of *Sciara* during the eyespot stages using autoradiography showed stage-specific non-histone chromosomal protein methylation, seemingly during amplification in the DNA puffs [Goodman and Benjamin, 1973]. Early eyespot stages had virtually no methylated non-histone chromosomal proteins. The methylation was confined to specific bands that included DNA puffs among them, but not RNA puffs. When histones were not selectively removed, methylation was seen more uniformly

across the polytene chromosomes. The radio-labeled non-histone methylation had a turnover rate such that most was gone in an hour and all was gone in 2 hours in a pulse-chase experiment. Studying DNA precursor uptake in one gland of a pair while simultaneously observing non-histone chromosomal protein methylation in the other gland allowed them to make correlations. When one occurred so did the other and when one was absent so was the other. Moreover, the patterns of DNA synthesis and chromosomal protein methylation were correlated whereas no such correlation occurred with RNA synthesis. They provided an image demonstrating this on puffs II/9A and II/IIB, the largest puffs. They also looked at ethylation of chromosomal proteins, but found it was uniformly distributed on polytene chromosomes and that the turnover was much slower. Regarding the methylation of non-histone proteins in the DNA puffs during amplification, PCNA has been shown to be regulated by lysine methylation [Takawa et al., 2012, Hamamoto et al., 2015] where methylation of a specific lysine enhanced its interaction with the flap endonuclease FEN1. Loss of PCNA methylation affected Okazaki fragment synthesis, slowed down replication forks, and potentially lead to DNA damage. Other proteins commonly thought of as transcription factors have also been shown to be targets of methylation [Biggar and Li, 2014, Hamamoto et al., 2015], an example being RB1, the methylation of which prevents it from being phosphorylated by CDK complexes and affects its interaction with E2F. E2F1 is also a methylation target. Interestingly, both E2F and RB are involved in DNA amplification in *Drosophila* follicle cells (discussed in the *Drosophila* section above). Could they also be at puff II/9A, and if so, could they be the targets for methylation?

In 2002, it was shown that the 1 kb amplification origin in DNA puff II/9A is also used as a replication origin in other stages of development. Using PCR analysis of nascent strand enrichment, an 8 kb zone at locus II/9A was enriched in the endoreplication cycles preceding amplification as well as in mitotic embryonic cells (Fig 1.5). The left boundary stays the same, but the right boundary shrinks far back during amplification. RNA polymerase II (RNAP-II) was detected in the downstream promoter region of gene II/9-1 only during amplification and is possibly responsible for the loss of initiation activity in that area. Overall, this established a relationship between a metazoan initiation zone and transcription machinery. On the left boundary, a ~400 bp DNase hypersensitivity site (DHS) was identified approximately 600 bp upstream of the main origin in a region where initiation activity drops off [Urnov et al., 2002]. The DHS contains a 72-100 bp stretch of identity with a *Rhynchosciara* DNA puff, suggesting it is important for amplification in both systems. The sensitivity of the DHS to DNase was correlated with both origin activity and binding of unidentified nuclear protein(s). It lost sensitivity to DNase when amplification ended. The DHS did not seem to flank or coincide with a gene as per Northern blot analysis. The 1 kb origin itself did not contain any DNase hypersensitivity sites. Bent DNA was mapped at two sites between the origin and the DHS (Susan Gerbi and Fyodor Urnov, unpublished data). In contrast to previous results suggesting a discrete binding site for ORC, subsequent experiments have indicated that ORC might be enriched more dispersively as in *Drosophila* gene amplification, though it is most enriched at the origin during amplification (Susan Gerbi and Eric Gustafson, unpublished). As seen at DAFC-66D

in *Drosophila*, histone acetylation is observed at II/9A, with stronger enrichment in the origin region during amplification and a broader distribution in later stages (Fig. 1.5). Finally, Northern blot analyses have identified two RNA species that are complementary to the 1 kb origin, one 200 nt long and the other around 20 nt (Susan Gerbi and Janell Johnson, unpublished).

Synthesis of information from Sciara

The future of DNA Puff research in *Sciara*

Given the work that has been done so far, and taking hints from the work in *Drosophila*, what may direct locus-specific re-replication at the II/9A origin? The upstream DHS is reminiscent of ACE3, which was shown to be a nucleosome depleted region [Liu et al., 2015]. Amplification levels are low or not detected in both ACE and the DHS and they are both relatively far away from the main initiation site. The two sites of bent DNA, which are AT-rich, are interesting as well. Urnov has hypothesized the bent DNA sites may serve as a histone magnet that depletes histones from the ORC binding site. However, bent DNA and AT-rich sequences have been shown to exclude nucleosomes. Perhaps these bent DNA sites induce a strong nucleosome positioning profile over the locus that leaves the ORC binding site exposed in a linker region. The ORI EcRE is adjacent to the ORC binding site and might be bound by EcR at the right time. It is tempting to speculate that EcR may interact with one or more members of the preRC to target them specifically to the amplification origin or that EcR prevents other factors from destabilizing preRC proteins there. However, EcR can easily be hypothesized to have other roles as well. In *Drosophila*, ligand-bound EcR forms a complex with a histone acetyltransferase (HAT) and demonstrates HAT activity [Ghbeish et al., 2001, Zhu et al., 2006, Kirilly et al., 2011]. Therefore, one of the EcR isoforms in *Sciara* might bind the ORI EcRE and other EcREs nearby to recruit HATs, which are needed for the histone acetylation found across II/9A. The *Sciara* lines where some of the nearby EcREs were mutated did not affect amplification, but this could be due to the redundancy of EcREs at this locus. Nonetheless, one wonders if acetylation was affected in those mutants. Since the 1 kb origin sequence hybridized to two RNA species in a Northern blot experiment, another possible role of the ORI EcRE could be in recruiting EcR to act as a transcription factor. The role of this origin RNA could be to target the PreRC to II/9A similar to an RNA that guides ORC to amplification origins in *Tetrahymena* [Mohammad et al., 2007]. One study was consistent with ORC having a discrete binding site. However, other data suggests ORC spreads out during amplification. This is reminiscent of the model proposed for *Drosophila* where ACE3 and/or Ori-β act as nucleation sites for ORC binding, which then spreads out forming an ORC chromatin structure [Royzman et al., 1999, Kim et al., 2011]. The larger ORC binding profile is also more consistent with the H4 acetylation profile across the locus. It was shown in *Drosophila* that ORC and histone acetylation have similar distributions across DAFC-66D [Kim et al., 2011, Liu et al., 2012]. Though it has been postulated that amplification ends by stage 14x7, DNA synthesis was still detected at this stage and replication forks were observed passing through

the 1 kb origin sequence in later stages [Wu et al., 1993]. Therefore it remains possible that a slower rate of re-initiation occurs from elsewhere in the II/9A locus. The stronger signal of ORC spreading across II/9A in later stages supports this possibility. Since the initial claim that amplification ends in 14x7, all studies since have estimated the final copy number in this stage. RNA polymerase II (RNAP-II) seems to help define the narrower initiation zone observed during amplification relative to other developmental stages where initiation occurs more broadly across II/9A. RNAP-II could be poised in the II/9-1 promoter ready to begin transcription. It could also potentially be transcribing a gene in that area that has not been identified yet. Active transcription has been shown to mold the locations of MCMs [Powell et al., 2015, Gros et al., 2015]. RNAP-II could also help anchor preRC components as it has been shown to anchor ORC to rDNA replication origins [Mayan, 2013].

Although there is an abundance of evidence for a role for ecdysone in DNA amplification in insects, all of it is correlative and can be interpreted as indirect effects. Ecdysone may just trigger a cascade of events leading to DNA amplification. For example, ecdysone injection may stimulate RNA synthesis, protein synthesis, or both, of a gene product that is more directly involved with amplification. It may even be farther removed, stimulating the production of a transcription factor that is responsible for regulating the amplification factor. Similarly, ligature experiments might prevent ecdysone from stimulating the synthesis of the more direct factors. Correlation of high ecdysone titers with puffing is circumstantial as the high ecdysone titers are also correlated with many other developmental events occurring at the same time. For *Sciara coprophila*, one crucial experiment will be to mutate the ORI EcRE next to the ORC binding site that is hypothesized to be important for EcR binding and amplification. If that abolishes amplification site-specifically, then a more direct role can be inferred. This has not been possible since *Sciara* has not had transgenic techniques available. However, gene insertion has recently been demonstrated for *Sciara* [Yamamoto et al., 2015], and the experiment of mutating the ORI EcRE is underway (Yutaka Yamamoto, personal communication). It will also be important to determine if EcR interacts with ORC or another preRC member. Alternatively, since there may be an RNA complementary to the II/9A origin, perhaps EcR stimulates the transcription of this RNA at the origin. If so, then mutating the EcRE might abolish both transcription of this RNA as well as amplification. The presence of this origin-RNA still needs to be confirmed and those experiments are currently being carried out by Leo Kadota. Studies on re-replication in *Sciara* have also been limited by the lack of a genome sequence to identify genes for tagging, mutating, or knocking down. Moreover, there are numerous other re-replication origins that could be studied. Having the sequence for more than one could help identify shared motifs to target for mutation to assess the effects on amplification. Does ecdysone stimulate amplification in all DNA puffs or only some? Does it inhibit amplification in any of the puffs? These questions have been asked, but there has not been a way to answer them. In this thesis, I present the genome sequence for *Sciara coprophila* and the locations of at least 14 amplicons within it, thereby unlocking experiments and analyses that have not been possible before now. This thesis represents a turning point for *Sciara* as a model system.

Part II:

The Genome and DNA Puff Sequences of Sciara coprophila

Part II of this thesis covers the journey to determining the sequences of the genome, transcriptome(s), and the salivary gland DNA puffs for the fungus fly, Sciara coprophila. It is broken up into three chapters. A large part of the work for determining the genome sequence initiated when we were accepted as early testers of new nanopore sequencing technology in the MinION Access Program (MAP) hosted by Oxford Nanopore Technologies (ONT). Intrigued by the possibility of obtaining extremely long reads, I was motivated to modify and optimize the standard protocol for preparing MinION sequencing libraries. My early results from sequencing genomic DNA from Sciara with the MinION included high quality reads that exceeded 100 kb. This generated excitement in the MAP community and we were encouraged to submit our protocols and results as a preprint on bioRxiv. That preprint is represented here as the first chapter. The long read protocols established in the first chapter became the basis for continued optimization and pursuit for ultra-long reads to facilitate the Sciara genome assembly. In the second chapter in this section (Part II), I present the Sciara genome assembly, which features Illumina short reads as well as single-molecule long reads from both Pacific Biosciences (PacBio) and Oxford Nanopore's MinION. The discussion on obtaining high-quality reads that surpass 100 kb is continued in the second chapter of Part II where I demonstrate that they align to PacBio-only assemblies with up to 91.1% identity. The Sciara genome assembly resulting from a combination of PacBio and ONT datasets was highly contiguous and was further scaffolded with single-molecule recognition sequence mapping data from the BioNano Genomics (BNG) Irys platform. The kinetics information and ionic current signal from PacBio and Minion data, respectively, were then used to identify signatures of DNA base modifications throughout the Sciara genome. In the third chapter of Part II, the Sciara genome sequence is used

as a reference to map high throughput Illumina datasets to in order to identify regions of the salivary gland genome that increase in copy number throughout the course of the DNA amplification stages. In doing so, I was able to identify at least 14 of the 18 DNA puffs. These studies set the stage for unraveling the mechanism of site-specific DNA re-replication in Sciara coprophila. Nevertheless, we need higher resolution estimates of where DNA replication initiates inside the DNA puffs to specifically determine the anatomy of the re-replication origins. Therefore, in parallel, I worked on methods to identify origins of replication genome-wide, which is discussed in Part III of this thesis. The long read protocols I established in the first two chapters of Part II will be invaluable for others to facilitate genome assemblies, identify structural variants, and aid other applications that could benefit from ultra-long reads, including an application discussed in Part III of this thesis for studying DNA replication.

CHAPTER 2

Sequencing Ultra Long DNA Molecules with the Oxford Nanopore MinION

John M. Urban[1,3*], Jacob E. Bliss[1], Charles E. Lawrence[2,3], and Susan A. Gerbi[1,3*]

1 Division of Biology and Medicine, Brown University, Providence, RI, USA
2 Division of Applied Mathematics, Brown University, Providence, RI, USA
3 Center for Computational Molecular Biology, Brown University, Providence, RI, USA
*Correspondence to John_Urban@Brown.edu and Susan_Gerbi@Brown.edu

This chapter is adapted from:

Urban JM, Bliss J, Lawrence CE, Gerbi SA. (2015) Sequencing ultra long DNA molecules with the Oxford Nanopore MinION. bioRxiv. http://dx.doi.org/10.1101/019281

This manuscript was prepared to release as a preprint on bioRxiv to provide the MinION community with details on how I was obtaining 100 kb 2D reads after I wrote a popular blog on the Oxford Nanopore community website titled, "The obstacles to great runs with ultra long DNA". Within days of posting to bioRxiv, this became one of Altmetric's highest ranked articles of all time. Currently, it is in Altmetric's top 5% and has been tweeted by 105 people to a potential audience size of >168,000 people. It has also been featured in 3 blogs, and was written about by GenomeWeb, who interviewed us about it.

I conceived, designed, and carried out the experiments, analyzed the data, wrote and utilized the associated "poreminion" software (`https://github.com/JohnUrban/poreminion`), and prepared the manuscript. J.B. performed mass-matings of flies for embryo collection and contributed to discussions. C.E.L. provided guidance in statistics. S.A.G. provided guidance in molecular techniques, and helped edit the manuscript. All authors read and approved the manuscript.

A couple illustrative figures that portray how MinION sequencing works are borrowed from:
Ip, C. L., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., Leggett, R. M., Eccles, D. A., Zalunin, V., **Urban, J. M.**, et al (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research, 4.

I participated in the formation of the MinION Analysis and Reference Consortium (MARC) and in discussions on the weekly phone calls and in-person meetings leading up to this first "Phase 1" paper. Moreover, I helped draft and edit the paper. The MARC paper has an even higher Altmetric score than my preprint and was highlighted in many blogs and news pieces. Nonetheless, I did not perform the experiments[1] nor analyses, and am only using visual aspects (no text) from MARC figures that help illustrate the descriptions of MinION sequencing from my bioRxiv preprint. The MARC figure is denoted as such in the chapter.

[1] The ultra long read protocol(s) described in my preprint were going to be re-evaluated in the subsequent "MARC Phase 2" paper where I would have re-done them on E. coli genomic DNA so that the results could be compared to the standard protocol results from Phase 1. Thus, my major contribution to MARC would have been performing those experiments. However, the technology was evolving too quickly, outpacing the speed at which we could coordinate experiments for whatever the current pores and chemistries were at the time. This lead MARC to decide to abandon a protocol development paper, and instead focus on biological applications that could span multiple iterations of the MinION technology.

2.1 Abstract

Oxford Nanopore Technologies' nanopore sequencing device, the MinION, holds the promise of sequencing ultra long DNA fragments >100 kb. An obstacle to realizing this promise is delivering ultra long DNA molecules to the nanopores. We present our progress in developing cost-effective modifications to the library preparation protocol to overcome this obstacle. Our resulting MinION data contain multiple reads >100 kb, including a 103 kb 2D read. Importantly, our modified library preparation protocols result in shifting the entire distribution of reads up to longer lengths, especially the tail of the distribution. We also demonstrate a new way to deplete DNA that is smaller than 10 kb. Subsequent analyses comparing the ionic current signal events to the resulting base-called read sequences demonstrated that molecules occasionally have events that accumulate in the neighborhood of certain 5mers in the base-called sequence, suggesting that DNA molecules may occasionally stall while traversing the nanopores. Interestingly, these localized accumulations of events are enriched near G-quadruplex (G4) motifs in the base-called sequences, indicating that G4 structures may form and temporarily prevent further passage through the nanopores until they unfold. This indicates that the MinION may be useful for future studies on G-quadruplex folding. Finally, our open source software used to analyze our MinION data is freely available at: `https://github.com/JohnUrban/poreminion`.[2]

Key Words: Oxford Nanopore MinION; nanopore DNA sequencing; long DNA preparations; long reads

[2]Poreminion will only properly work on Oxford Nanopore data from up to the time this preprint was released (May 13, 2015). The "fast5" files output by the MinION and corresponding MinKNOW and Metrichor software have had multiple changes that break some poreminion functions since then. I have developed a new set of tools called Fast5Tools (`https://github.com/JohnUrban/fast5tools`) that are much more flexible and can currently handle all fast5 file formats up to the latest R7.3 versions. It is yet to be tested on and updated for R9 though.

2.2 Introduction

High-throughput DNA sequencing is at the cusp of two paradigm shifts: where and how sequencing is performed. The MinION from Oxford Nanopore Technologies (ONT) is a pocket-sized, long-read (>1 kb) DNA sequencing device used in individual laboratories and fieldwork [Check Hayden, 2015, Kilianski et al., 2015, Quick et al., 2015, Loman et al., 2015] that detects 5mer-dependent changes[3] in ionic current, or "events", as single DNA molecules traverse nanopores (Figure 2.1). The events are then base-called using ONT's Metrichor cloud service. Base-calling accuracy and total output per flow cell have been increasing [Mikheyev and Tin, 2014, Quick et al., 2014, Ashton et al., 2014, Jain et al., 2015, Goodwin et al., 2015a], now up to 85% [Jain et al., 2015] and 490 Mb [Goodwin et al., 2015a]and are expected to exceed 90% accuracy and 2Gb per flow cell soon[4]. Importantly, since MinION reads can be re-base-called, older reads can inherit the benefits of better base-calling.

The protocol for preparing a MinION sequencing library is still evolving, but currently includes shearing genomic DNA using a Covaris g-TUBE, an optional "PreCR" step to repair damaged DNA, end repair, dA-tailing, adapter ligation, and His-bead purification. When properly ligated, a double-stranded (ds) DNA molecule has a Y-shaped (Y) adapter (also called the lead or leader adapter) at one end and a hairpin (HP) adapter at the other (Figure 2.1D). dsDNA is pulled through a pore one strand at a time starting at the 5' end of the Y, followed by the "template" strand, and, ideally, the HP and "complement" strand. The information from either strand can be used for 1-directional (1D) base-calling, and integrating the information from both strands can be used for 2-directional (2D) base-calling, which results in higher mean quality scores (Q; Supplementary Table A.1) and higher accuracy [Quick et al., 2014, Ashton et al., 2014, Jain et al., 2015, Ammar et al., 2015, Madoui et al., 2015, Goodwin et al., 2015a]. Molecules with a Y at each end or a nick in the template only give 1D template reads. It is also possible to obtain information from both strands but have no 2D base-calling (Supp. Fig. A.1, Supp. Table A.2). There is an approximate ratio of 1 event per base in 1D reads (especially when $Q \geq 3.5$) and 2 events per base in 2D reads (Supp. Fig. A.2). 2D reads with mean quality scores $Q \geq 9$ are considered high quality by ONT. Nonetheless, most other 2D reads are also valuable: 83-91% of all 2D reads align to their corresponding reference genome [Ashton et al., 2014, Madoui et al., 2015]. Moreover, 1D reads are also likely to be quite valuable in various applications [Kilianski et al., 2015, Madoui et al., 2015, Warren et al., 2015a, Karlsson et al., 2015, Bolisetty et al., 2015, Greninger et al., 2015, Cao et al., 2015, Szalay and Golovchenko, 2015, Sovic et al., 2015] especially if they are first error-corrected or if one works directly with the

[3] The Kmer size, in part, depends on the pores used, speed of translocation, and modeling choices. For example, 5mers were used before and during the time this manuscript was written, but more recently 6mers were used to explain the ionic current changes.

[4] As of Fall 2016, R9 and R9.4 pores, their associated chemistries, and the newer Recurrent Neural Network (RNN) base-calling approach (formerly a Hidden Markov Model approach was used) seem to have reached these benchmarks.

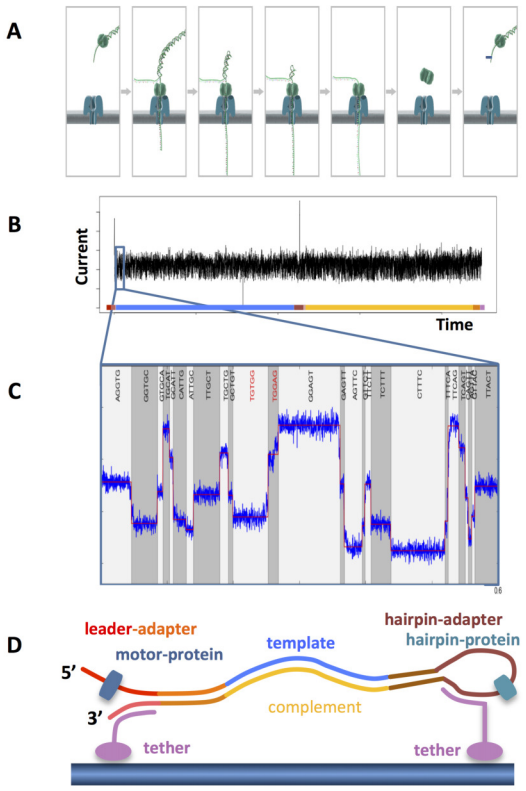


Figure 2.1: Illustrations of Nanopore Sequencing.
(A) Double-stranded DNA is pulled through a nanopore one strand at a time. **(B)** A sensor measures ionic current going through the nanopore, which becomes impeded when DNA enters the pore. **(C)** As the DNA is pulled through the pore the changes in ionic current correspond to the changes in local base composition. The ionic current signal (blue squiggly line) is then segmented into 'steps' (red line inside blue squiggly line) with the goal that each step represents a single increment of the DNA traversing the pore. However, some neighboring increments will be combined into the same 'step' (especially for long homo-polymers) and some increments will be segmented into more than one 'step'. Each step is called an event, and consists of 4 parameters: mean, standard deviation, start time, and duration. After learning the parameters that describe the ionic current distribution for each kmer in a set of Kmers, the sequence of events can be matched to the most likely sequence of kmers that describe them and translated into a nucleotide sequence. This has been done with both Hidden Markov Models and Recurrent Neural Networks. **(D)** A schematic representing a 2D-capable read. One end has a lead adapter and the other end has a hairpin adapter. The DNA is pulled through the pore from the 5' end of the lead adapter (see **(B)** where the colors on bottom correspond to colors here). The hairpin is abasic resulting in an identifiable spike in the ionic current signal. After base-calling the template and complement strand independently (1D reads), they can be aligned and the event information from both can be used to create a higher accuracy sequence called a 2D read. The motor proteins help control the speed of translocation. The tethers help tether DNA molecules to the membrane where nanopores are embedded. This reduces the search space from 3 dimensions to 2 dimensions thereby reducing the input material needed by orders of magnitude.
Illustrations taken from the MARC paper [Ip et al., 2015], but the text is my own.

ionic current events underlying them [Loman et al., 2015, Szalay and Golovchenko, 2015][5].

The MinION has the promise to sequence ultra long DNA fragments >100 kb [Check Hayden, 2012]. However, early reports suggested it falls short of this promise [Mikheyev and Tin, 2014, Check Hayden, 2014] with maximum read lengths near 10 kb. In more recent reports, the majority of reads were concentrated below 30 kb, but maximum 2D read lengths were approximately 31.6 kb [Ashton et al., 2014], 48.5 kb [Jain et al., 2015], 57.4 kb [Goodwin et al., 2015a], 58.7 kb [Madoui et al., 2015], and maximum 1D read lengths were 66.7 kb [Ashton et al., 2014], 123 kb [Madoui et al., 2015], and 191 kb [Goodwin et al., 2015a]. Nonetheless, 1D reads >100 kb and 2D reads >30 kb have been rare and no instances of 2D reads >60 kb have yet been reported. Here we describe modified protocols to harness the MinION's potential for sequencing ultra long molecules and present three resulting MinION runs (A, B, and C) with many reads exceeding 100 kb, including a ~103 kb 2D read, and with multiple 2D reads exceeding 60 kb (Fig. 2.3, Table 2.1, Supplementary Tables A.3–A.6). Importantly, these protocols result in sequencing longer molecules in general with the majority of data coming from molecules > 10 kb (in contrast to early reports where no reads exceeded 10 kb) and molecule N50s up to 25.2 kb, where 50% of the summed length of base-called molecules comes from molecules $\geq$ 25.2 kb. Finally, bioinformatics analyses of the MinION reads and ionic current signal events suggest that DNA molecules stall at G-quadruplex (G4) motifs while traversing the nanopores, indicating that the MinION could potentially be used for future studies on G-quadruplex folding.

2.3 Results

First, for Run A, we sought to maximize read lengths by (i) skipping the Covaris shearing step specified by ONT, (ii) minimizing DNA breakage, (iii) using AMPure ratios that deplete smaller DNA, and (iv) performing elutions that facilitate the release of long DNA from AMPure beads (Fig. 2.2). Run A had 9,935 base-called event files featuring 139 molecules >50 kb, 21 molecules >100 kb, a max 2D length of 102.9 kb (Q=8.74) and a max 1D length of 304.3 kb (Q=2.12), although the longest 1D length with Q>3.5 was 202.3 kb (Fig. 2.3, Table 2.1, Supplementary Tables A.3–A.6)[6]. The summed molecule length was 49.8 Mb with a molecule N50 of 25.2 kb (Table 2.1, A.3). 79.2% of the summed molecule length came from molecules greater than 10 kb, which made up 14.4% of all base-called molecules (Table 2.1).

There were three side effects when keeping the DNA long (Table 2.1): (1) a low proportion of

[5]Many papers have come out since this manuscript was released on biorxiv that demonstrate the utility of high noise 1D reads, but the new Canu [Koren et al., 2016] long read genome assembler might be the best example. Moreover, the new pores, chemistry, and base-calling have become much better, giving rise to the majority of 1D reads having >80-90% accuracy. Thus, the problem of low quality nanopore reads is becoming a thing of the past.

[6]See Fig. A.2 to see why Q > 3.5 was used here.

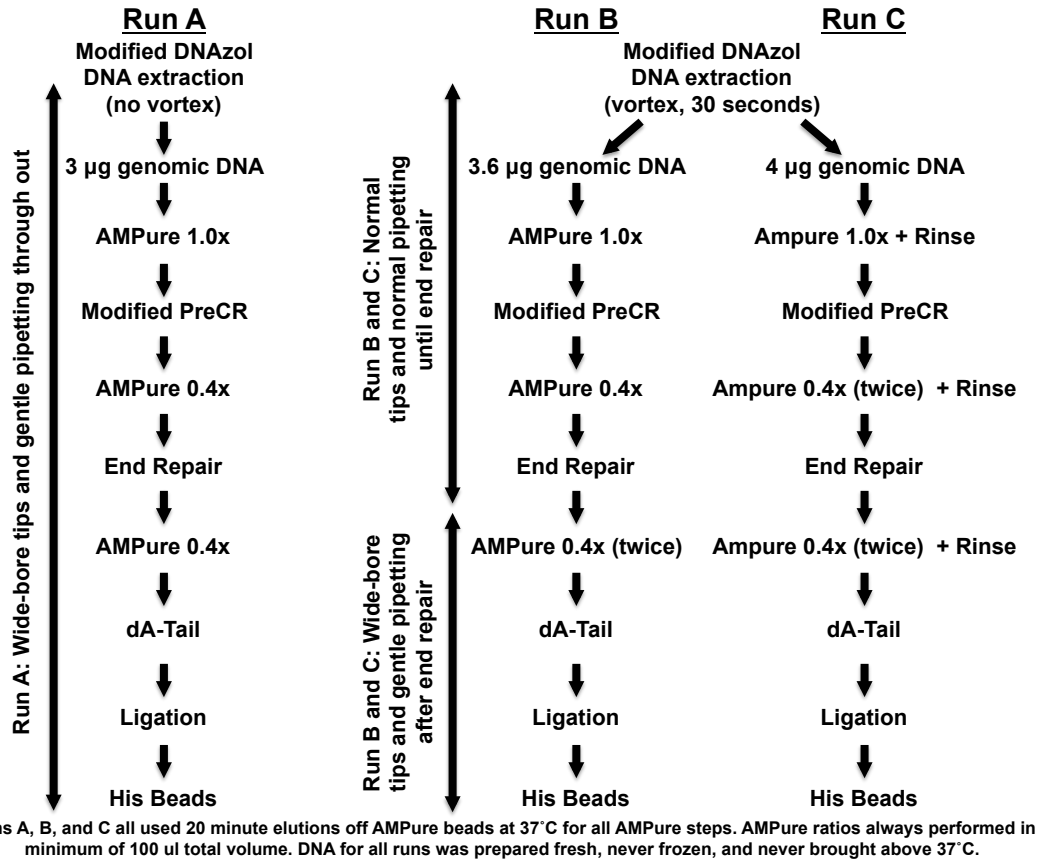


Figure 2.2: Overview of protocols for each run.

For **Run A**: we gently let freshly obtained precipitated DNA re-suspend in TE (pH 8.0), skipped the Covaris shearing step in the standard protocol, used wide-bore tips with gentle pipetting to minimize DNA breakage, and started with 3X the recommended starting material (3 μg instead of 1μg) to compensate for differences in molarity. PreCR was done in double the volume with double the reagents for double the time since we started with more DNA than recommended. Both AMPure steps after PreCR were done at 0.4x to help deplete small DNA molecules (e.g. <1 kb).

For **Run B**: we vortexed the DNA (full speed, 30 seconds) during and after DNA extraction and used normal pipette tips until end-repair, after which wide-bore tips and gentler pipetting were employed. Moreover, to account for the possible increase of molecules <1 kb due to vortexing, we started with more material (3.6x) and performed a 0.4x AMPure bead clean-up after PreCR and two sequential 0.4x AMPure bead clean-ups after end-repair.

For **Run C**: we used 4x the recommended input amount of DNA (from same source as for Run B), did two sequential 0.4x AMPure clean-ups before and two after end-repair, and did a new rinse step (see Supplementary Methods and Supplementary Fig. A.3b) at the end of all AMPure steps before the final elution of DNA off the beads to deplete DNA <10 kb. In the rinse, smaller DNA preferentially falls off the beads.

For **all runs**: in all AMPure elutions, the beads were incubated at 37°C for >20 minutes to facilitate and wait for long DNA to come off the beads. DNA was never subject to temperature extremes below 4°C or above 37°C and was re-suspended in 1X TE (pH 8.0) when isolated. In AMPure bead steps, the ethanol washes were performed with buffered 80% ethanol (10 mM Tris-Cl, pH 8.0).

See Methods and Supplementary Methods for more information.

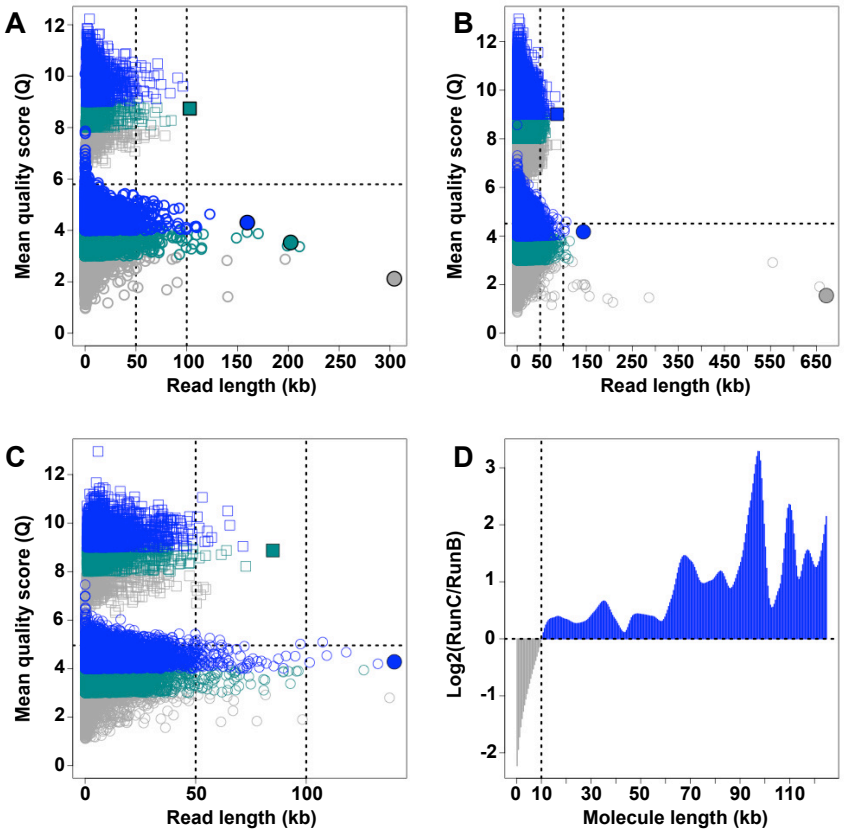


Figure 2.3: Read lengths and mean quality scores (Q) across runs.
Read length vs. Mean Quality Score (Q) for **(A)** Run A, **(B)** Run B, and **(C)** Run C. In each, the distributions for both 2D reads (squares; concentrated above Q=6) and 1D reads (circles; concentrated below Q=6) are shown. For 2D reads, blue is used for reads with Q $\geq$ 9, cyan for Q between 8 and 9, and grey for Q $<$ 8. For 1D reads, blue is used for Q $\geq$ 4, cyan for Q between 3 and 4, and grey for Q $<$ 3. Filled-in squares and circles indicate reads highlighted in the main paper for their length and Q. The horizontal dashed line marks the minimum 2D Q. The vertical dashed lines denote 50 kb and 100 kb. **(D)** The log2 fold change of Run C over Run B of the proportion of total summed molecule length plotted as a function of molecule length. This shows that, despite coming from the same DNA source as Run B, Run C is enriched for DNA molecules $>$ 10 kb and depleted for molecules $<$ 10 kb compared to Run B.

2D reads (21.9%); (2) a high proportion (87.2%) of pre-base-called event files with <2,000 events (though this made up only 3.41% of all events obtained); and (3) lower output than might have been achieved if the DNA was sheared (100-400 Mb routinely achieved by others in the MinION Access Program (MAP) at the time of our experiments). One possible explanation is that ultra long DNA is fragile and can break at various steps: after end repair leading to problems in ligation, after ligation leading to His-bead enrichments of HP-ligated DNA that cannot be sequenced, and while being injected into the MinION leading to additional issues such as Y-ligated DNA that can only give 1D reads. Therefore, we proceeded to find a balance between read length, total output, and proportion of 2D reads.

Since long DNA will break at each step even with gentle handling, we sought to minimize breakage after end repair at which point breaks affect sequencing. Therefore, for Run B we vortexed the DNA (keeping most >10 kb) to help ensure most breaks would occur prior to end repair (Fig. 2.2, Supplementary Fig. A.3A). Run B had a bigger proportion of files with 2D reads (49.8%), a smaller proportion (40.2%) of event files with <2,000 events (2.48% of all events), and a much higher summed molecule length (386.9 Mb) albeit with a lower yet still impressive molecule N50 of 13.6 kb (Table 2.1). Run B had proportionally fewer reads >50 kb than Run A, but had a higher count owing to the higher output (396 molecules >50 kb, 24 >100 kb). The longest 2D read was 86.8 kb (Q=9.01) and the longest 1D read was 671.2 kb (Q=1.55), but the longest 1D read with Q>3.5 was 143.8 kb (Fig. 2.3, Table 2.1, Supplementary Tables A.3–A.6). The majority (58.5%) of the summed molecule length came from molecules >10 kb (15.8% of all base-called molecules).

For Run C, we sought to improve upon Run B by increasing the amount of data from molecules >10 kb. We first explored ways to deplete DNA molecules smaller than 10 kb with modifications to the standard AMPure bead protocol (Supplementary Fig. A.3B) and found that it was sufficient to add a gentle rinse step just prior to elution. We integrated rinses into the AMPure steps for Run C (Fig. 2.2). Although Runs B and C started with the same source of DNA, the amount of molecules <10 kb in Run C was greatly depleted compared to Run B (Fig. 2.3D, Supplementary Fig. A.4). Run C had proportionally fewer events in files with <2,000 events (1.24%) than both Runs A and B (Table 2.1). There were >2-fold more base-called molecules >50 kb and >100 kb in Run C than in Run B. Moreover, Run C had the highest mean and median molecule sizes of all runs (Table 2.1, Supplementary Table A.3). The longest 2D read was 84.9 kb (Q=8.87) and the longest 1D read was 139.9 kb (Q=4.28) (Fig. 2.3, Table 2.1, Supplementary Tables A.3–A.6). The base-called molecule N50 was 20.8 kb and there was a higher proportion (25.9%) of base-called molecules >10 kb than both previous runs, which made up 77.2% of the summed molecule length. The percent of molecules with 2D reads (28.2%) and the summed molecule length (70.1 Mb) were both intermediate between Runs A and B, suggesting that there is a trade-off between read length and output/2D reads. It is also possible that these differences were due to variation in library preparations and flow cell quality, though Runs B and C had similar estimates of active channels (Supplementary Table A.7).

Table 2.1: Statistics and Values for Runs A, B, and C

	Run A	Run B	Run C
Number event files (prior to base-calling)	27,667	82,040	12,536
Number successfully base-called (% of starting number of event files)	9,935 (35.9%)	64,282 (78.35%)	8,941 (71.3%)
Number of molecules >50 kb	139	396	119
Number of molecules >100 kb	21	24	8
Longest 2D read (Q)	102.935 kb (8.74)	86.797 kb (9.01)	84.898 kb (8.87)
Longest 2D read (Q>9)	96.237 kb (9.6)	86.797 kb (9.01)	71.830 kb (9.04)
Longest 1D read (Q)	304.309 kb (2.12)	671.219 kb (1.55)	139.864 kb (4.28)
Longest 1D read (Q>3.5)	202.293 kb (3.53)	143.763 kb (4.17)	139.864 kb (4.28)
Total summed molecule length	49.8 Mb	386.9 Mb	70.1 Mb
Molecule N50	25.238 kb	13.553 kb	20.824 kb
Percent of base-called molecules >10 kb	14.4%	15.8%	25.9%
Percent of summed molecule length from base-called molecules >10 kb	79.2%	58.5%	77.2%
Percent of molecules with 2D reads	21.9%	49.8%	28.2%
Percent of event files with <2,000 events	87.2%	40.2%	55.5%
Percent of total summed events in event files with <2,000 events	3.41%	2.48%	1.24%

Values pertaining to "molecules", "base-called molecules", and "reads" are from successfully base-called fast5 files only. Values pertaining to "event files" are from the set of all pre-base-called fast5 files.

Each run produced event files (69 total, Supplementary Table A.8) with >1 million events, which is too many events for the base-caller, a limit set by ONT due to memory constraints. Given that there is typically 1-2 events per base, we investigated whether these files with millions of events represented megabase molecules. First we eliminated all multi-million event files that contained blocks of repeated events (Supplemental Fig. 2.4). This error is rare across all files in Runs A, B, and C (0-0.017%; Supplementary Table A.8), but was prominent in files with >1 million events (28.6-65.2%), leaving 30 of the 69 without this error (19 of which are shown in Fig. 2.5A). Another concern is that large event files might arise from a faulty pore independent of a DNA molecule traversing the pore. To rule this out, we discarded 11 of the 30 remaining files that did not show evidence of a lead adapter profile (Fig. 2.5B). A third concern is that DNA molecules can become temporarily stuck in the pores leading to an accumulation of events from the same region of a DNA molecule. To determine how pervasive this issue might be in the remaining multi-million event files that could not be base-called, we looked in base-called files with <1 million events to see how many times the base-caller decided two or more adjacent events corresponded to the same kmer (move=0, "stay") instead of advancing to a new kmer (move>1) (Fig. 2.5C, Supplementary Fig. A.5). In general, all base-called files with >531,779 events had an average of >80% "0 moves" in the template and complement (Fig. 2.5C), indicating there were many more events than bases. Indeed, files with >500 thousand events corresponded to 12.7-196.4 kb molecule sequences, demonstrating high event:base ratios of 2 to 76 events per base (Supplementary Fig. A.5) rather than the average of ~1 event per base. Thus, it is more probable that the remaining 19 files that contain 1.1-5.2 million events correspond to sub-megabase DNA molecules that were temporarily stuck in the pores leading to an accumulation of events than the possibility that they correspond to megabase DNA molecules.

Finally, we sought to understand how DNA of any size might get stuck in pores. We hypothesized one possibility is that a highly stable DNA secondary structure known as a G-quadruplex (G4) [Huppert, 2010] may form and block further translocation through the nanopore until it unfolds (Fig. 2.5D, Supplementary Fig. A.6), resulting in an accumulation of measurements of a 5mer or set of 5mers slightly upstream (i.e. 5') of the G4. Indeed, when analyzing all base-called template and control reads, there is a significantly higher number of "0 moves" near G4 motifs than near randomly selected locations even when controlling for read-specific effects (Supplementary Table A.9). Moreover, there is a significantly higher number of "0 moves" near G4 motifs on complement strands than near G4 motifs on template strands consistent with the higher propensity of G4-folding in single-stranded DNA [Huppert, 2010] (Supplementary Table A.10). Indeed, there is also a significantly higher number of "0 moves" near G4 motifs that have >4 poly-G tracts than G4 motifs with only 4 poly-G tracts (Supplementary Table A.11) consistent with their higher probability of forming a G4 structure. Finally, in an aggregate analysis looking at all base-called template and complement reads, there is a clear enrichment of "0 moves" near G4 motifs with the highest enrichment (Run A) and shoulders (Runs B, C) slightly upstream (-9 to -27 nt) of the G4 motif (position

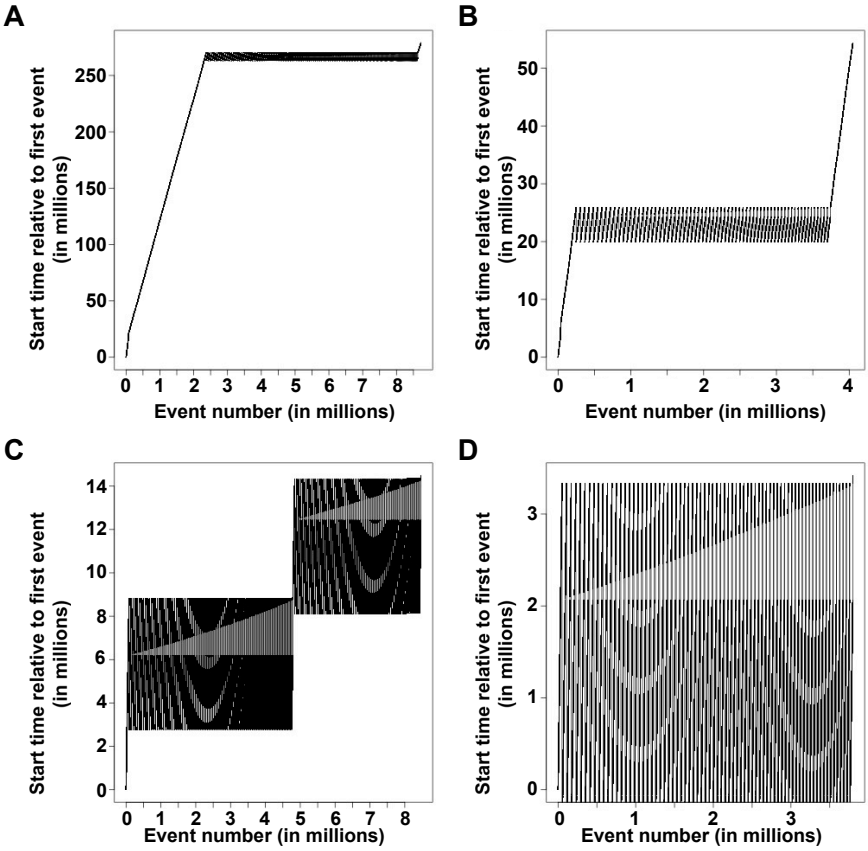


Figure 2.4: Examples of multi-million event files that contained "Time Errors" (repeated blocks of events).

(A) Example from Run A. **(B)** Another example from Run A. **(C)** Example from Run B. **(D)** Example from Run C. Events are ionic current measurements that have 4 parameters: mean, standard deviation, start time, and duration. The start time of each subsequent event in a sequence of events is necessarily later than the start time of the immediately preceding event (and all before it). Repeated blocks of events are easily identified (e.g., using the timetest subcommand in poreminion) since there are events with earlier start times than previous events. Repeated blocks of events are also easy to identify visually. When there are no repeated blocks of events, the "event start time" as a function of "event number" is a monotonically increasing function (see Fig. 2.5A). However, in files with repeated blocks of events, the function instead exhibits periodic patterns (A-D). This error is rare (0-0.017% of all files; Supplementary Table A.8), but was more prominent in files with 100 thousand to 1 million events (0-5.51%) than it was in files with <100 thousand events (0%) and was most prominent in the files with >1 million events (28.6-65.2%). It is also important to note that the MinKNOW software issue that caused this error has since been fixed (MinKNOW 0.49.2 and later versions) and is not expected to occur in future MinION runs.

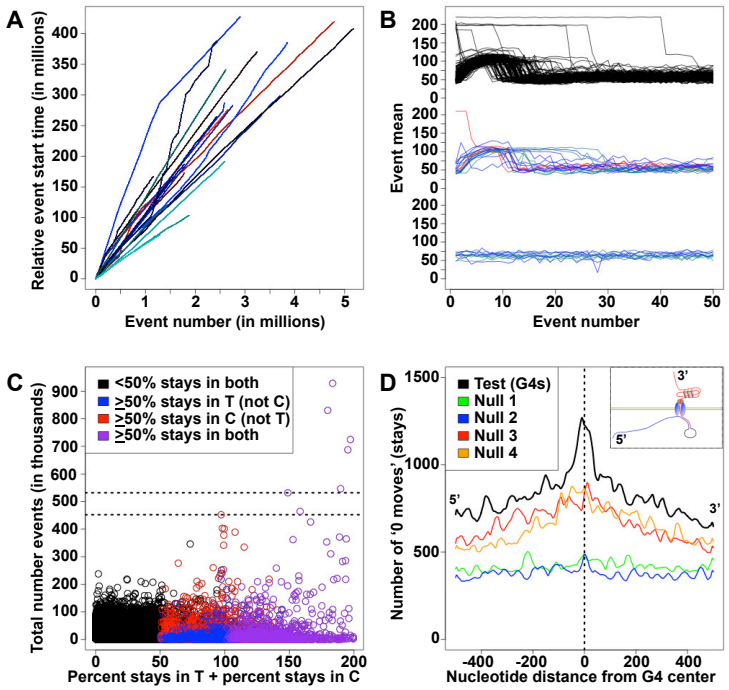


Figure 2.5: Dissecting files with > 1 million events and a role of G-quadruplexes in DNA stalling.

(A) Event start times (relative to the first event) as a function of event number for the 19 multi-million event files that did not have repeated blocks of events and that had evidence of the lead adapter (i.e. middle row of **B**). The monotonically increasing nature of each curve visually demonstrates the absence of repeated blocks of events, which create periodic patterns (as seen in **??**). Different colors represent different molecules. Shades of red are from Run A, shades of blue are from Run B, and shades of cyan are from Run C. **(B)** The event means of the first 50 events for: (Top) 150 randomly sampled (50 from each run) high quality (Q $\geq$ 9) 2D reads; multi-million event files that did not have repeated events and (Middle) do show or (Bottom) do NOT show evidence of the lead adapter. The lead adapter event mean profile at the beginning of the sequence of events for a read signifies that adapter-ligated DNA has entered the pore whereas lack of a lead adapter event mean profile is consistent with (cannot discount) the faulty pore hypothesis. Line color scheme as in A. **(C)** The total number of events plotted as a function of the summed percentages of base-called events that had "0 moves" (stays) in all base-called template, "T", and complement, "C", reads (note that these are files with <1 million events). Notably 452,301 events (bottom dashed line) was the highest number of events where the average of base-called events with "0 moves" in T and C reads was < 50% and 531,779 events (top dashed line) is the highest number of events where the average was < 80%. All points above the latter had an average of > 80% "0 moves" in the base-called events. This suggests that most or all of the fast5 files that could not be base-called due to having > 1 million events are likely comprised of mostly "0 move" events. **(D)** The number of "0 moves" as a function of proximity to G4 motif centers (test condition) or the centers of randomly selected positions (null 1–4 conditions described in Supplementary Methods) in all base-called T and C reads (note that these are files with < 1 million events) in Run A. The inset at the top right depicts the orientation (5' – 3') of a DNA molecule going through a pore from the top to bottom chamber and how a G4 might cause a DNA molecule to stall resulting in an accumulation of measurements of a 5mer or set of 5mers slightly upstream (5') of the G4. The template strand is blue, the hairpin adapter is black, and the complement strand is red.

0). Since DNA is pulled through the nanopore 5' – 3', an accumulation of "0 moves" around the region immediately upstream (i.e. 5') to the G4 motif is what would be expected if a G4 structure temporarily blocked translocation (Fig. 2.5D, Supplementary Fig. A.6). Nonetheless, G4 motifs are likely not the only sites where DNA molecules get stuck since there are many "0 moves" from molecules without detectable G4 motifs. Fortunately, since the base-caller can identify "0 moves", it is able to deal with DNA stalling, though there is a slight decrease in Q with each increase in the proportion of called events that have "0 moves" (Supplementary Fig. A.5). Consistently, reads with G4 motifs appear to have a slightly lower average Q than all reads (Supplementary Table A.12). Nonetheless, in future studies, "0 moves" might serve as an in vitro indicator of whether, how often, and for how long all possible G4 motifs in a genome form in the context of both double-stranded DNA (when the motif is in the template read) and single-stranded DNA (when the G4 motif is in the complement read) in a single MinION experiment.

2.4 Discussion

In conclusion, our data demonstrate that the MinION can sequence ultra long DNA molecules >100 kb that make it intact to the nanopores despite early reports that suggested the MinION fell short of this promise. Specifically, this can be achieved with our modified protocols. Importantly, we demonstrate that it is possible to obtain high quality 2D reads >100 kb (e.g. 102.9 kb, Q=8.74) with the MinION. Indeed, using our modified protocols, ONT internally obtained a 192 kb high quality 2D E. coli read that mapped to the reference genome. The modified protocols presented here will help others obtain similar read size distributions. Importantly, these protocols require no additional equipment or reagents with respect to what is already needed to prepare a MinION library. Ultra long reads will be instrumental in highly contiguous genome assemblies of complex genomes as well as in spanning long arrays of tandem repeats and detecting other complex structural variants.

In addition to demonstrating that the MinION is capable of producing 100 kb reads, we also explored event files that had millions of events, which cannot be base-called, to determine if it was likely that they came from megabase molecules. However, we found that molecules that produced hundreds of thousands of events and could be base-called usually corresponded to molecules that gave rise to very high proportions of "0 moves" in base-calling. These "0 moves" indicate that an event most likely came from the same kmer as the event before it, which in turn indicates an error in segmenting the raw ionic current data stream. One interpretation for specific regions in reads that have numerous "0 moves" is that the DNA molecule became temporarily stuck causing long dwell times and perturbations that lead to over-segmenting that region of the data stream. Given that molecules with hundreds of thousands of events in our dataset were those with the highest densities of "0 moves", it is rational to conclude that the molecules that produced millions of events likely corresponded to sub-megabase molecules that became temporarily stuck. Nonetheless, there is no

reason to assume that the MinION is not capable of sequencing megabase molecules. This may be achievable with specialized protocols that perform size-selection just before sequencing to remove the accumulation of sub-megbase DNA or that keep the DNA embedded in agarose plugs or agarose microbeads throughout the library preparation[7]. However, the MinION device itself may also pose a problem since sequencing requires injecting the library through a small hole in the flow cell, which can fragment an otherwise intact megabase library.

Finally, given that we found many examples of reads with localized regions that had a high density of "0 moves", we explored whether there were any specific sequence features in the vicinity of those regions. We uncovered evidence that G4 motifs were one such feature, indicating that G4 structures may temporarily obstruct passage of the DNA through the nanopore during MinION sequencing[8]. This observation could be used in future studies, including studies that test small molecules thought to stabilize or destabilize G4 structures, as a novel application of the MinION.

2.5 Materials and Methods

MinION Library Preparations

Genomic DNA was extracted from 2-day old, male *Sciara coprophila* (fungus gnat) embryos using DNAzol. DNA was cleaned with AMPure XP beads (Beckman Coulter Agencourt), then repaired with PreCR Repair mix (NEB). The repaired DNA was then cleaned with AMPure XP beads, end-repaired with NEBNext End Repair Module (NEB), cleaned with AMPure XP beads, and dA-tailed with NEBNext dA-Tailing Module (NEB). ONT SQK-MAP004 adapters were ligated on to the end-repaired/dA-tailed DNA using Blunt/TA Ligase Master Mix (NEB). Hairpin-adapter-ligated DNA was enriched using His-beads (Dynabeads His-tag Isolation and Pulldown; Life Technologies). This yields the Pre-Sequencing Mix (PSM). Before the PreCR step, Run A started with 3000 ng gDNA, Run B started with 3600 ng, and Run C started with 4080 ng. All had A[260/280] and A[260/230] $>$1.8 prior to PreCR and $>$2.0 after subsequent clean-up steps during the library preparation as determined by Nanodrop (Thermo Scientific). Each sample preparation ended, following His-bead enrichment, with between 300-400 ng DNA in the PSM as measured by Qubit dsDNA HS (Life Technologies).

[7] I worked on an agarose microbead library preparation protocol as well as collaborated with Intact Genomics on Pulsed-Field Gel size-selection strategies. I discuss these attempts at creating libraries with mean read lengths $>$100 kb in the appendix.

[8] Since this preprint was released, I talked to employees from ONT at conferences who said potassium is part of the buffer in their flow cells. That is important because potassium (K^+) ions strongly stabilize G4 structures giving more credibility to our observations [Kankia and Marky, 2001, Shim et al., 2009]. Moreover, the employees from ONT said they were also aware that G4s caused issues early on in their development of nanopore sequencing. This is in agreement with our observations and interpretations as well. Finally, there is a paper that demonstrates detection of G-quadruplex folding by observing their blockage of ionic current through single nanopores [Shim and Gu, 2012]. The effect of G4 structures on ionic current can lead to segmentation and base-calling issues as discussed in our preprint.

Important differences for preparing the libraries for Runs A, B, and C include whether or not vortexing was employed to partially shear DNA, when wide-bored tips were used, AMPure bead ratios, sequential AMPure steps, and the introduction of rinses in AMPure steps. The differences in the workflows for each run is portrayed in Figure 2.2, and are extensively detailed in the Supplementary Methods.

Loading the Sequencing Mix (SM) and Sequencing

Sequencing Mix (SM) was made using 3-6 μl PSM, 3-4 μl Fuel, and EP buffer up to 150 μl (Oxford Nanopore Technologies, SQK-MAP004). SM was made fresh for loading/re-loading at various intervals 6-7 times throughout each 48 hour run. See Supplementary Table A.13 for exact details. All libraries were loaded onto R7.3 flow cells, and run on the MinION using MinKNOW versions 0.47.2.6, 0.48.2.6, 0.48.2.12 for Runs A, B, and C respectively.

Base-calling and Bioinformatics overview

MinION events data for each DNA molecule is stored in an individual "fast5" file. All fast5 files were base-called using the Metrichor 1.12 r7.X 2D-basecalling XL protocol. Metrichor returns fast5 files that are updated with base-calling information and sorts these base-called fast5 files into two folders: 'pass' and 'fail'. Despite the name, even the fail folder contains successfully base-called fast5 files. The 'pass' folder simply includes the fast5 files that contain 2D reads with mean quality scores ≥ 9. The 'fail' folder contains all other fast5 files including those that have 2D reads with $Q < 9$, those that only have 1D reads, and those that completely failed base-calling and contain no reads. In our analyses, we characterize the full distribution of reads (including 1D and 2D reads) in all successfully base-called fast5 files from both folders as well as the ionic current events therein. After base-calling, fast5 files were filtered to remove those with errors in event timing (repeated blocks of events) and those that were unsuccessfully base-called due to having too few events ($<$ 200; arbitrarily set by Metrichor), too many events ($>$1 million, set by Metrichor due to memory constraints), or "no template found" (a characterization made by Metrichor) (Supplementary Table A.8). The rare error that leads to blocks of events being repeated numerous times is identified by events with earlier start times than preceding events. It is important to point out that the MinKNOW software issue that led to this error in a small number of files has since been resolved (versions 0.49.2 and later) and is not expected to occur in future experiments. Each base-called fast5 file from a MinION run describes data from a single molecule, yet there can be up to 3 reads per file: template, complement, and 2D. We define the molecule size as the length of the 2D read if present, the length of the template read if there is only a template read present, and the longer length between the template and complement reads when both are present in the absence of a 2D read. Filtering, characterization of the data in each fast5 file (including molecule and read lengths, events information, and G4 motif locations, etc), and other data analyses were carried out using our open source MinION toolset called "poreminion" (`https://github.com/JohnUrban/poreminion`)

as well as in R and using poretools [Loman and Quinlan, 2014] and BEDtools [Quinlan and Hall, 2010]. For more information on bioinformatics analyses, see the Supplementary Methods.

Supplementary Materials

Supplementary Figures, Tables, and Methods are available in the Appendix.

Data deposition

MinION data will be available on NCBI SRA.

Funding

This work was supported by National Science Foundation predoctoral fellowships to JMU: NSF GRFP DGE-1058262 and NSF EPSCoR Grant# 1004057.

Conflict of Interest

JMU and SAG are members of MAP and have received free reagents from ONT.

Acknowledgements

We thank Oxford Nanopore Technologies for including us in the MinION Access Program (MAP), the MAP community for online discussions, Ben Raphael for providing computing resources, and Mark Howison for helpful discussions.

CHAPTER 3

Single-molecule sequencing of long DNA molecules allows high contiguity de novo genome assembly and detection of DNA modification signatures for the fungus fly, Sciara coprophila

John M. Urban [1] , Michael S. Foulk [1,2], Jacob E. Bliss [1], C. Michelle Coleman [3], Nanyan Lu [3], Reza Mazloom, Susan J. Brown [3], Susan A. Gerbi [1]

[1] Brown University Division of Biology and Medicine, Department of Molecular Biology, Cell Biology and Biochemistry, Providence, Rhode Island 02912, USA

[2] Present Address: Mercyhurst University, Department of Biology, Erie, PA 16546, USA

[3] Kansas State University Division of Biology, KSU Bioinformatics Center, Ackert Hall, Manhattan, Kansas 66502, USA

Corresponding authors: Susan_Gerbi@Brown.edu and John_Urban@Brown.edu

This chapter represents a manuscript in preparation for submission. Michael Foulk prepared the Illumina DNA library. Jacob E. Bliss did all Sciara mass matings. I collected all embryos, larvae, pupae, and adult Sciara needed for my experiments. I obtained high molecular weight genomic DNA and sent it to the Technology Development Group at the Institute of Genomics & Multiscale Biology at the Icahn School of Medicine at Mount Sinai, where PacBio sequencing libraries were prepared and sequenced. Though I prepared many DNA plugs for BioNano Genomics data with embryonic DNA, Michelle Coleman ultimately obtained successful DNA plugs from Sciara pupae that I sent, and performed the BioNano preparations and sequencing. Reza Mazloom and Nanyan Lu performed BioNano hybrid scaffolding with selected assemblies sent to them. Sue Brown provided guidance in our acquisition of BioNano data and provided oversight to Michelle Coleman, Reza Mazloom, and Nanyan Lu. I prepared all MinION libraries and performed all MinION sequencing and analyses. I wrote the suite of tools for working with MinION data, called fast5tools (`https://github.com/JohnUrban/fast5tools`). I performed all short and long read assemblies, genome annotation, genome polishing, and assembly evaluations with DNA puff II/9A, Illumina, PacBio, MinION, and BioNano data. I did all RNA work and library preparations for the 12 RNA-seq samples representing replicates from both sexes at different stages. I performed all transcriptome assemblies and RNA-seq data analysis. I performed DNA modification analyses with PacBio single molecule kinetics data and MinION single molecule ionic current data. I conceived the experiments and analyses. I wrote the manuscript.

3.1 Abstract

Biological studies of the fungus fly, Sciara coprophila, began in the early 1900s and gave rise to an early example of an organism that broke the rule of DNA constancy. In Sciara, specific-loci are amplified, entire chromosomes are eliminated, and a single nucleus can contain over 8,000 copies of the genome. The phenomenon of imprinting was also discovered first in Sciara when it was observed that it was the paternal chromosomes that are specifically targeted for chromosome elimination in normal development. Moreover, Sciara females have either only male or only female offspring and since matings can be controlled to produce either, this makes Sciara a useful model system for sex-specific early development studies. However, studies into these and other interesting biological features of the fungus fly have been hampered by a lack of transgenic techniques and the lack of a genome sequence. Recently we demonstrated the first targeted gene insertion for Sciara and now we present the first genome sequence. We approached assembling the genome with multiple technologies. Using short read, paired-end Illumina data, we generated 40 different assemblies from seven popular assemblers. After evaluating the assemblies with several metrics, we were able to identify the best short read assemblers for the Sciara genome with our data. Nonetheless, we were unable to produce a highly contiguous genome sequence from short read data even after filtering for contamination and re-assembling. In contrast, we were able to produce assemblies with multi-megabase contigs using long reads from single molecule sequencing technologies including PacBio and the Oxford Nanopore MinION. As part of the MinION access program (MAP), we developed protocols to obtain longer reads, obtaining reads that exceed 100 kb. Not only do these ultra-long reads map to PacBio-only assemblies, they were very useful in assembling the Sciara genome. We generated 50 assemblies from 6 long read assemblers. Since we had 44.1X coverage from PacBio and only 6-11X from the MinION, these assemblies were made using either only PacBio reads or combinations of PacBio and MinION reads. Through extensive evaluation using reference-free metrics we identified the assemblers that produced the best assemblies given our datasets and this particular genome sequence, although all assemblies were arguably of high quality. Importantly, the assemblies generated from the combination of PacBio and Oxford Nanopore datasets typically outperformed PacBio-only assemblies in the majority of metrics. This effect may be a result of additional coverage. However, whereas the additional coverage was modest, the status among the rankings was not. Therefore, it is tempting to posit that the presence of many extremely long MinION reads, the combination of different technologies, or both played roles in these results. Optical maps from BioNano genomics were used to scaffold a subset of the best assemblies. RNA-seq datasets from a combination of embryos, larvae, pupae, and adult flies from both sexes were used to facilitate annotation of the final genome sequence. Finally, both PacBio and Oxford Nanopore data gave us the opportunity to explore DNA modifications in the Sciara genome sequence to potentially begin to unravel the mechanism of imprinting.

3.2 Introduction

The fungus gnat, Sciara coprophila, is both an old and emerging model system, rich with opportunities for studying fundamental biology, especially chromosomal biology. Similar to other insects, different cells have different genomic copy numbers, from canonical diploid tissues to the endocycling larval salivary glands that result in cells with over 8000 copies of each chromosome closely held together to form giant polytene chromosomes [Rasch, 1970a, Rasch, 1970b]. In Sciara, there are also differences in the ratio of chromosomal copy numbers in different cells [Rasch, 1970a, Rasch, 1970b, Rasch, 2006]. An interesting example comes from the tissue-specific germ-line limited L chromosomes that are eliminated from somatic cells before cellularization in the early embryo [Gerbi, 1986]. Another example is seen in sperm, which have a single copy of all chromosomes except the X chromosome, for which there are two copies [Metz, 1925, Metz, 1930, Metz, 1934, Schmuck, 1934, Abbot and Gerbi, 1981, Gerbi, 1986]. The fusion of a sperm with an egg gives rise to yet a third example where the zygote and early embryo have 3 copies of the X per nucleus before the elimination of 1 copy for females or 2 copies for males as part of the sex determination pathway [Metz, 1938, Gerbi, 1986, Sánchez, 2008, Sánchez, 2014]. Chromosome elimination is itself an interesting feature of Sciara, and yet it is made more fascinating since the eliminated chromosomes are paternal in origin [Metz, 1938, Gerbi, 1986]. This distinction of maternal versus paternal chromosomes was the first example of and gave rise to the term "imprinting" [Crouse, 1960a, Rieffel and Crouse, 1966, Crouse et al., 1971]. The modifications to DNA or chromatin that may be responsible for distinguishing paternal and maternal chromosomes in Sciara have been elusive [Sánchez, 2008, Sánchez, 2014], but evidence from an antibody against 5-methylcytosine suggests it may be present throughout the genome, particularly in heteorchromatic regions, as visualized by immunofluorescence on polytene chromosomes [Eastman et al., 1980, Greciano et al., 2009]. DNA modifications in Sciara is otherwise unexplored. In addition to genome and chromosome copy number regulation throughout Sciara development, Sciara also presents examples of locus-specific copy number regulation [Gabrusewycz-Garica, 1964, Crouse and Keyl, 1968, Rasch, 1970b, Gabrusewycz-Garcia, 1971, Gerbi et al., 1993, Gerbi and Bielinsky, 2002, Simon et al., 2016]. There is developmentally regulated locus-specific increases in the copy numbers of up to 18 different sites termed "DNA puffs" that occurs at the end of endocycling in the larval salivary gland genome. Our laboratory has extensively studied one DNA puff on chromosome II at locus 9A, termed "DNA puff II/9A" that amplifies to over 100,000 copies, at least 17-fold over the polytene background [DiBartolomeis and Gerbi, 1989, Bienz-Tadmor et al., 1991, Wu et al., 1993, Liang et al., 1993, Gerbi et al., 1993, Liang and Gerbi, 1994, Bielinsky et al., 2001, Mok et al., 2001, Lunyak et al., 2002, Urnov et al., 2002, Foulk et al., 2006, Liew et al., 2013, Foulk et al., 2013]. This is an example of intrachromosomal DNA amplification, which is also studied in Drosophila follicle cells where several sites in the genome amplify after endocycles as well [Claycomb and Orr-Weaver, 2005, Nordman and Orr-Weaver, 2012]. Studies from Drosophila and Sciara have demonstrated that this regulation of locus-specific copy number in insects is a direct result from repeated DNA replication initiation at replication origins, known as DNA re-replication, and is governed by components of the DNA replication system [Osheim et al., 1988, Underwood et al., 1990, Landis et al., 1997, Asano and Wharton,

1999, Schwed et al., 2002, Tower, 2004]. Both systems offer opportunities to study the regulation of locus-specific DNA replication in metazoans. For Sciara, although there has been a lot of progress made in studying the re-replication origin at DNA puff II/9A, little is known about the other DNA puffs at the molecular and sequence level. Similarly, the dearth of DNA sequence information for Sciara has slowed down the progress of searching for base and chromatin modifications in the genome that may underly imprinting, chromosomal elimination, and sex determination. Overall, these and other rich opportunities that Sciara presents for studying fundamental biological processes have been inhibited by two factors: (i) a lack of transgenic techniques to enable thorough genetic manipulation and (ii) a lack of a genome sequence to identify relevant genes. Recently, transgenic techniques have been developed for Sciara for the first time that will be useful for future investigations [Yamamoto et al., 2015]. In this manuscript, we present a high quality draft genome sequence for Sciara coprophila.

The complete Sciara genome is spread across three autosomes (chrII, chrIII, and chrIV), an X chromosome, an X' chromosome that differs from the X by a large paracentric inversion, and the germ-line limited L chromosome(s) [Metz and Schmuck, 1929, Metz, 1938, Crouse, 1960b, Crouse et al., 1977, Gerbi, 1986]. Studies of nuclear DNA content suggest the Sciara genome has ~38% GC content [Gerbi, 1971] and is ~292 Mb in somatic cells and ~369 Mb in germ cells that have L chromosomes [Rasch, 2006] (Supp Table B.1). L chromosomes are eliminated from nuclei destined to become somatic cells around the 5th or 6th nuclear division, roughly 3 hours after egg deposition [Dubois, 1932, DuBois, 1933, Metz, 1938, Gerbi, 1986, de Saint Phalle and Sullivan, 1996, Goday and Esteban, 2001]. Although we included some early embryos to capture L sequence if possible, we focused primarily on the somatic genome and do not expect L sequences to be well-represented in our assembly. In addition, we sequenced only the male genome to avoid complications from the X' in some females. The X' chromosome and L chromosomes will be the focus of future studies. Our goal for this first release of the Sciara genome was for it to be contiguous enough to unambiguously identify DNA puff sequences in future studies by looking for areas of the salivary gland genome that have increasing copy number across DNA amplification stages. To satisfy this, DNA puff sequences would need to reside in contigs long enough such that the peak of amplification, where the re-replication origin exists, can be unequivocally discriminated from flanking sequences for the given puff. Moreover, the assembly would have to be contiguous enough such that the number of contigs with amplification detected is close to the expected number of DNA puffs. Otherwise, validation by quantitative PCR (qPCR) and mapping each sequence to its corresponding DNA puff cytologically by fluorescence in situ hybridization (FISH) would be too exhaustive and too expensive. In Drosophila follicle cells, up to 75-100 kb centered on the chorion locus amplifies [Spradling, 1981, Claycomb et al., 2004, Claycomb and Orr-Weaver, 2005, Kim et al., 2011]. Therefore, assuming that Sciara DNA puffs may encompass similar widths, we reasoned we would be able to unequivocally identify up to 50% of the DNA puffs if 50% of the expected genome size was captured by contigs of at least 100 kb. This is a well-defined quantitative goal for assembly contiguity, requiring a minimum NG50 of 100kb to be

satisfied. We also used the known 9 kb sequence of the well-studied puff II/9A to determine if the assemblies could give us more information about the genomic sequence context of this locus. Finally, when evaluating assemblies, we aimed to choose those that appear to be supported the best by the data and to have the minimal number of mis-assemblies, using several reference-free metrics.

When we began the endeavor of assembling the Sciara genome, we obtained high coverage with 100 bp paired-end Illumina reads. We generated 44 short-read assemblies using 7 popular assemblers [Zerbino and Birney, 2008, Simpson et al., 2009, Simpson and Durbin, 2010, Bankevich et al., 2012, Luo et al., 2012, Kajitani et al., 2014, Li et al., 2015]. Evaluating these assemblies with several reference-free evaluation tools [Vezzi et al., 2012, Clark et al., 2013, Ghodsi et al., 2013, Hunt et al., 2013, Simão et al., 2015] allowed us to discriminate which assemblies were comparatively of the highest quality. Nonetheless, all were fragmented assemblies with up to hundreds of thousands of contigs where most were less than 1000 bp in length. In our case, filtering for bacterial contamination and re-assembling with the filtered data [Kumar et al., 2013, Laetsch et al., 2016] did not improve the contiguity of the resulting genome assembly. Overall, the Illumina-based assemblies would be too fragmented to map amplified DNA puffs in the Sciara genome that each potentially span more than 100 kb. In those assemblies, thousands of contigs would show signs of amplification in copy number experiments. To obtain more contiguous assemblies, there are other short read assemblers that perform better given the right prescription of Illumina data sets [Butler et al., 2008, Maccallum et al., 2009, Gnerre et al., 2011, Weisenfeld et al., 2014], including longer Illumina reads (e.g. 250 bp), PCR-free Illumina data, paired-end libraries of different insert sizes (typically between 200-800 bp), and mate-pair jumping libraries representing longer distance information (typically 2-20 kb). In all cases, this meant that we would need to acquire more short read data. However, the highly fragmented nature of our Illumina assemblies suggested there was a high copy number of some repetitive element(s) strewn throughout the genome that short reads cannot resolve [Koren and Phillippy, 2015]. Moreover, though these approaches could produce higher contiguity assemblies, it would likely be a result of more comprehensive scaffolding of contigs assembled from relatively short reads. Thus, although these assemblies would represent the genome stitched together in longer sequences, each scaffold would comprise many gaps of unknown sequence represented by Ns.

Another route to obtaining more contiguous assemblies is incorporating Single Molecule Real Time (SMRT) sequencing data from Pacific Biosciences (PacBio) [Eid et al., 2009], which produces average read lengths that are much longer than the average contig lengths of our Illumina assemblies. These long reads are more error-prone than Illumina, but the errors are randomly distributed allowing high quality consensus sequences with enough coverage [Eid et al., 2009]. It was shown that these long reads could be used in hybrid approaches with Illumina data several ways. PacBio reads were used to fill gaps in Illumina assemblies during or after the assembly process [Ribeiro et al., 2012, English et al., 2012], to resolve paths through Illumina-based graph structures [Ribeiro et al., 2012, Bankevich et al., 2012, Deshpande et al., 2013], and to scaffold higher accuracy Illumina

contigs produced by de Bruijn graph-based assemblers [Bashir et al., 2012]. In addition to using the PacBio reads to supplement short read assemblies, higher accuracy short reads were also used to error-correct the error-prone long reads before assembling them using the Overlap-Layout-Consensus (OLC) approach [Koren et al., 2012] that preceded short read assemblers and was originally used for the human and Drosophila genomes [Adams et al., 2000,Myers et al., 2000,Venter et al., 2001,Lander et al., 2001]. This was the first example of a hierarchical genome assembly process, which paved the way for non-hybrid PacBio-only assembly approaches where the PacBio reads are used to error-correct each other before assembling them through the OLC paradigm [Chin et al., 2013,Koren et al., 2013]. Both the hybrid and non-hybrid approaches produce assemblies that are far superior to those produced from short reads alone [Koren and Phillippy, 2015]. Due to these impressive advances in genome assembly using long noisy reads, we opted to acquire PacBio data rather than pursue more intricate Illumina-based approaches.

Around the same time of choosing to pursue long reads from PacBio, we were accepted into the MinION Access Program (MAP) from Oxford Nanopore Technologies (ONT) when it opened in 2014, granting us early access to their single-molecule nanopore sequencing technology. Nanopore sequencing is accomplished by measuring the changes in ionic current passing through a nanopore as a DNA molecule translocates through it [Branton et al., 2008]. Under the right conditions these ionic current signals can be translated into a base sequence by using hidden Markov models [Timp et al., 2012, David et al., 2016] or recurrent neural networks [Boža et al., 2016], both of which have been employed by ONT in their own base-callers (www.metrichor.com, https://github.com/nanoporetech/nanonet). The MinION is a hand-held sequencing device that contains an array of nanopores to sequence many DNA molecules in parallel. In MinION sequencing, one strand of DNA is pulled through at a time resulting in 1-direction (1D) reads. When a DNA molecule has a hairpin adapter on one end, both strands of the same molecule can be sequenced resulting in higher quality 2-direction (2D) reads [Ip et al., 2015]. By base-calling simultaneously with sequencing, researchers can obtain DNA sequences in real time within minutes of starting a run. In 2012, Oxford Nanopore reported single reads that spanned the entire 48.5 kb genome of the Lambda bacteriophage, and it was suggested that their technology held the promise of producing reads that exceeded 100 kb [Check Hayden, 2012]. However, in the initial stages of MAP, early reports suggested that the MinION fell short of this promise, describing maximum read lengths near 10 kb [Mikheyev and Tin, 2014, Check Hayden, 2014]. In contrast, even in our first attempts at sequencing Sciara genomic DNA in the first year of the MinION Access Program, the majority of our data came from reads $>$ 10 kb and we were able to obtain a 2D read that exceeded 100 kb as well as many 2D reads that exceeded 50 kb [Urban et al., 2015a]. To do this, we made slight modifications to the standard Oxford Nanopore procedure that excluded DNA shearing steps, involved gentle handling of the long DNA, and manipulated AMPure bead procedures to deplete DNA $<$ 10 kb and to elute long DNA from the beads. These protocols and early results were posted as a preprint on bioRxiv for others in the MAP community to access [Urban et al., 2015a]. Importantly, these ultra-long reads map to our PacBio-only genome assemblies,

which we did not initially report since the Sciara genome was unpublished. We have continued to use and update those protocols enabling us to reproducibly obtain extremely long MinION reads, including additional 2D reads that exceed 100 kb, across multiple iterations of chemistry and platform changes. In this manuscript, we demonstrate that these long reads are real and quite valuable. MinION reads have been shown to be useful in genome assembly both in combination with Illumina reads in hybrid approaches [Goodwin et al., 2015b, Risse et al., 2015, Warren et al., 2015b, Madoui et al., 2015] as well as in non-hybrid nanopore-only approaches [Loman et al., 2015, Koren et al., 2016]. We can confirm their value for aiding the assembly of an insect genome.

A third single molecule technology that can improve genome assembly contiguity is in the form of optical mapping from the BioNano Genomics (BNG) Irys platform [Lam et al., 2012]. BNG refers to their technology as Next Generation Mapping (NGM) perhaps to contrast it with the first generation of optical mapping technology [Schwartz et al., 1993, Lin et al., 1999] where DNA molecules were stretched out on slides before being cut with a restriction enzyme and subsequently stained to estimate the sizes and ordering of restriction fragments, demarcated by dark regions at restriction sites, along ultra long DNA molecules. With enough coverage from long enough molecules, overlaps of the patterns of restriction fragment lengths can be detected between DNA molecules and the restriction patterns can be assembled to represent full chromosome restriction site maps [Lin et al., 1999]. BioNano Genomics has offered a newer method that is higher throughput where DNA molecules are simply nicked with a nicking-endonuclease and briefly nick-translated in the presence of fluorescent nucleotides at recognition sequences [Lam et al., 2012]. In an iterative fashion, labeled DNA molecules are pulled into nano-channels where they can be visualized and ejected before moving on to a new set of DNA molecules. Nonetheless, the goal is the same: to visualize patterns of restriction sites on DNA molecules to determine genome-scale restriction maps. These restriction maps can be used to scaffold the megabase contigs produced from long read assemblies and help approach the goal of one scaffold per chromosome [Lam et al., 2012, Pendleton et al., 2015, Cunningham et al., 2015, VanBuren et al., 2015]. Optical mapping data can also be used to compare the structural integrities of a group of assemblies to identify which assemblies are most consistent with long-range optical maps [Bradnam et al., 2013, Koren and Phillippy, 2015].

Overall, in addition to acquiring 44X coverage with respect to the expected male somatic genome size from PacBio subreads, we obtained 10-11X coverage from DNA molecules sequenced by the MinION. As the MinION can produce up to three reads per molecule, this coverage estimate results from counting only one read per molecule. Approximately, 6X coverage was from 2D reads. During the time-frame of our data collection, there have been multiple advances in hybrid [Ye et al., 2016] and non-hybrid genome assemblers that take hierarchical approaches [Berlin et al., 2015, Koren et al., 2016, Chin et al., 2016] as well as those that directly assemble the long noisy reads by overlap-layout approaches [Li, 2016] (`https://github.com/ruanjue/smartdenovo`) and generalizations of de Bruijn graphs [Lin et al., 2016]. In total, using these recent hybrid and non-hybrid assemblers, we

generated 50 long read assemblies using data from both PacBio and Oxford Nanopore, producing megabase contigs and NG50s. These assemblies were iteratively polished using the Quiver algorithm [Chin et al., 2013], which incorporates the signal level information from all aligned PacBio reads to generate a de novo consensus sequence. The assemblies were then polished using the high quality Illumina data [Walker et al., 2014]. We evaluated the final polished long read assemblies as well as assemblies at multiple steps during the polishing process using Illumina-based metrics [Vezzi et al., 2012, Clark et al., 2013, Ghodsi et al., 2013, Hunt et al., 2013], the number of universal single copy orthologs in arthropods [Simão et al., 2015], and metrics from PacBio and MinION read alignments. Moreover, we incorporated single-molecule optical mapping data from the BioNano Genomics platform in both our evaluations of long read assemblies and for hybrid scaffolding, the latter of which improved the contiguity of our final assemblies. Our final assembly was annotated with the inclusion of transcriptome information from embryos, larvae, pupae, and adults from both males and females. Finally, both PacBio SMRT and Oxford Nanopore single molecule sequencing technologies come with raw signal information that can be used to identify base modifications [Flusberg et al., 2010, Clark et al., 2012, Loman et al., 2015, Simpson et al., 2016, Rand et al., 2016, Suzuki et al., 2016] and even secondary structure formation [Sawaya et al., 2015, Urban et al., 2015a]. With both technologies, anomalies in the signal can be used to infer the presence of modifications even if the identity of the modification is unknown. For PacBio data the incorporation of fluorescent bases by DNA polymerase have characteristic pulse widths (PWs) and interpulse durations (IPDs), and the kinetic variation in these metrics can be used to predict the occurrence of 4-methylcytosine, 5-methylctosine, and 6-methyladenine [Flusberg et al., 2010, Clark et al., 2012, Suzuki et al., 2016] as well as base J in trypanosomes [Genest et al., 2015]. For the MinION, modified bases often result in different ionic current disturbances than their unmodified counterparts and predicting 5-methylcytosine, 5-hydroxymethylcytosine, and 6-methyladenine has been demonstrated [Simpson et al., 2016, Rand et al., 2016]. Therefore, we leverage both technologies to weigh in on recent evidence of DNA modifications in the Sciara genome. Moving forward, the genome sequence will enhance studies in Sciara including re-replication, imprinting, and chromosome elimination.

3.3 Results

3.3.1 Short Read Assemblies

To begin exploring genome assembly approaches, we obtained ~103x coverage of paired-end Illumina data from a ~430 bp insert library. This data was assembled using several popular genome assemblers: Abyss [Simpson et al., 2009], Megahit [Li et al., 2015], Platanus [Kajitani et al., 2014], SGA [Simpson and Durbin, 2010], SOAPdenovo2 [Luo et al., 2012], SPAdes [Bankevich et al., 2012], and Velvet [Zerbino and Birney, 2008]. We also attempted to use MaSuRCA [Zimin et al., 2013]. However, after several attempts that failed due to exceeding allotted time limits, with the final attempt given >8 days, we decided to not pursue it further. For all other assemblers, we tried assembling the raw data as well as a read set that was quality-filtered using Trimmomatic [Bolger

et al., 2014]. Moreover, SPAdes and SGA both error-correct the input reads followed by assembling the error-corrected reads. Therefore, we also tried the error-corrected raw reads and error-corrected quality-filtered reads produced by BayesHammer [Nikolenko et al., 2013] in the SPAdes pipeline with the other assemblers. Finally, some assemblers, such as Megahit, Platanus, and SPAdes, employ pipelines that iterate over multiple k-mer sizes whereas others, such as Velvet, SOAPdenovo2, and Abyss, require a value of K to be selected. Therefore, we tried K=55 and K=77 for each set of reads for the latter set of assemblers.

To evaluate the resulting assemblies, we used LAP [Ghodsi et al., 2013], ALE [Clark et al., 2013], FRC^{bam} [Vezzi et al., 2012], REAPR [Hunt et al., 2013], and BUSCO [Simão et al., 2015]. LAP and ALE provide probability measures that a given set of reads came from each assembly. LAP requires finding all mappings for each read, which can be time-consuming and demanding of computational resources. However, the authors note that fairly small sample sizes of reads tend to correlate well with larger samples for eukaryotic genome assemblies, which can drastically differ from one assembler to the next [Ghodsi et al., 2013]. Therefore, we first tried two independent samples of ~15,000 paired reads, followed by a sample of ~150,000, and finally ~1.5 million. All samples largely agreed with each other and with the other metrics (Supp Fig. B.1 A), so we were satisfied with a final sample size of 1.5 million. FRC^{bam} flags potential errors (called features) throughout each assembly given a set of reads, with lower numbers being better than higher ones. We also checked to see if normalizing to the number of features per megabase changed the conclusions, but found that it was highly correlated with the total number of features (Spearman's rho = 0.88) and that the rank leaders remained the same (Supp Fig. B.1 B). Therefore, we used only the number of features in calculating mean rankings. REAPR outputs the percent of bases in the assembly that are error-free as well as a score for each base in the assembly, which can be used to calculate the mean base score. We found that the percent of error-free bases was highly correlated with the mean base score (Spearman = 0.62), so chose to use only the former when calculating mean rankings (Supp. Fig. B.1 C). Finally, BUSCO reports the percentage of complete single-copy orthologs (SCOs) found in each assembly given a set of SCOs expected to be present in the genome. The higher the percentage of complete SCOs found, the more complete and correctly put together an assembly is likely to be. We used the 2,675 SCOs from arthropods.

The assemblies ranged from ~226-348 Mb in size, with a mean assembly size of ~280 Mb, which is close to the expected genome size of ~292 Mb. NG50 was highly correlated with assembly size and assemblies from SPAdes tended to have the largest values for both whereas Velvet tended to have the smallest values (Supp. Fig. B.1 D). The five reference-free evaluation approaches were typically in agreement, though FRC^{bam} showed no correlation with LAP and REAPR ranks, and was negatively correlated with BUSCO and NG50 ranks (Fig. 3.1 A). The mean ranking of the assemblies had no correlation with NG50, although individual ranking metrics did (Fig. 3.1 A). Specifically, assemblies with larger NG50s tended to be ranked higher by BUSCO and LAP, but

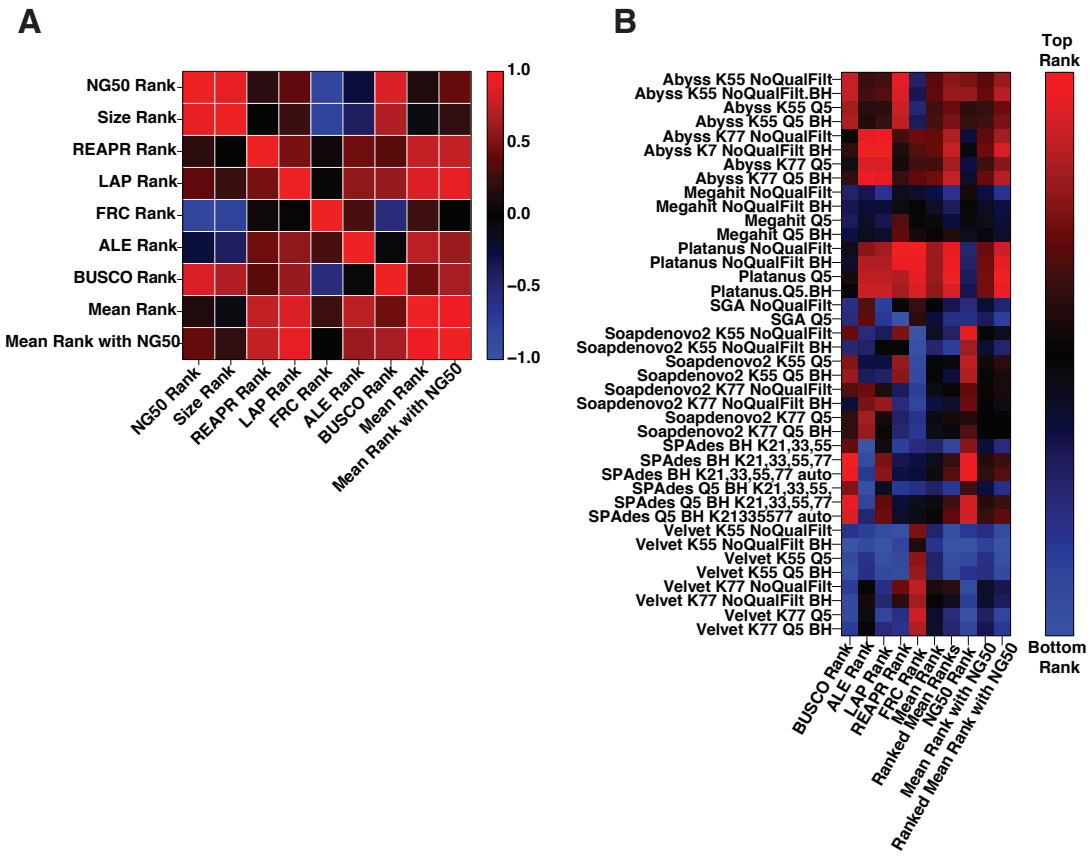


Figure 3.1: Illumina-based short-read assemblies.

(A) Correlation matrix of evaluation metrics used on short-read assemblies. **(B)** Rank matrix for the 40 Illumina assemblies (rows) with the ranks for each metric (columns). There were 5 evaluation metrics. Also displayed is the mean rank taken by averaging the ranks of the 5 metrics, the ranking of the mean rank that converts the means into ranks from 1–40, the NG50 ranking, the mean rank when NG50 rank is included in the calculation, and the ranking of those mean ranks that converts them from means to 1–40.

lower by FRC^{bam} and ALE. The assemblies produced from quality-filtered reads did not do better or worse than their raw data counterparts as determined by their mean ranks as well as ranks from each metric independently (Supp. Fig. B.1 E). Similarly, the error-corrected reads did not do consistently better or worse than their uncorrected counterparts. For the relevant assemblers, the bigger value of K (77) consistently did better than the lower value (55) according to LAP, FRC^{bam}, and ALE. However, BUSCO and REAPR favored the K=55 assemblies. Ultimately, the biggest differences in ranks were due to the assembler used, not how the input data was processed or the choice of Kmer size (Supp. Fig. B.1 E). In general, the NG50 and other size statistics for a given assembler did not drastically change with the different conditions. Finally, Platanus most consistently performed with the best rankings across metrics, followed very closely by Abyss. Though individual metrics correlated one way or another with NG50, Platanus was a consistent exception, having a higher percent of error-free bases, higher LAP and ALE scores, and a lower number of features detected by FRC^{bam} than would be predicted by a linear model of NG50 versus these metrics (Fig. B.2). Given that Platanus is a diploid-aware assembler, it is satisfying that it received many of the best scores. However, it is plausible that it performed better than other assemblers since it iterates over values of K up to 80, whereas others used K=77 as the highest value. In contrast, Megahit iterates up to values including K=81 and K=99 and did not perform as well, suggesting the value of K is not the only reason Platanus received many of the best scores.

The NG50 values of all of the assemblies were quite low at 2.5-7.3 kb, and the assemblies were highly fragmented, distributed across tens to hundreds of thousands of scaffolds. Nonetheless, the largest scaffolds in the assemblies reached up to the megabase range, which was initially exciting. However, BLAST demonstrated that the longest scaffolds were all bacterial, something not uncommon to assemblies from whole animals. In fact, the longest scaffolds of apparent insect origin were in the 50-60 kb range. It has been shown that filtering out the contaminating reads can improve an assembly [Kumar et al., 2013]. Therefore, to remove reads from contaminating species, we adopted a procedure similar to that used recently for the Tardigrade genome [Koutsovoulos et al., 2016] with the help of BlobTools [Kumar et al., 2013, Laetsch et al., 2016]. We focused on the Platanus assembly produced from quality filtered, error-corrected reads. Instead of using the final gap-closed scaffolds, we used the Platanus contigs (largest contig ~77 kb) to allow each contig to be characterized separately and avoid discarding data due to misjoins. The BlobPlot for the initial assembly demonstrates two major clusters of contigs, one of which corresponds to bacterial sequences and one that mainly harbors eukaryotic sequences (Fig 3.2 A). There were 1564 contigs explicitly annotated as being bacterial that made up ~6.8 Mbp of the 283 Mb assembly. Of the bacterial-labeled contigs, the majority (1152, 6.2 Mb) was labeled as Delftia at the species level.

Ideally one could use coverage and/or GC content information associated with annotated bacterial contigs to also eliminate the likely bacterial contigs that were not annotated. Unfortunately, the annotated bacterial contigs were of equivalent coverage compared with those annotated as arthropod

(Fig 3.2 A). Moreover, the GC content of the bacterial and eukaryotic contig clusters overlapped enough to limit its usefulness (Fig 3.2 A). Therefore, we used an alternative set of reads from pre-amplification stage salivary glands [Urban et al., 2016], reasoning that since this dataset was from a different tissue, from a different stage, and prepared by a different person, the contaminating contigs would have much lower coverage. Over 95% of these reads mapped to the assembly, suggesting most or all of the Sciara genome is represented in the Platanus contigs. As expected, the BlobPlot demonstrates that the coverage of the bacterial cluster is much lower in this dataset (Fig 3.2 B). We therefore removed non-Arthropod-labeled contigs (and associated reads) from the assembly that were either labeled as super-kingdom "Bacteria" or had coverage < 0.1. There were 83327 such contigs ranging 79 bp to 77.2 kb in size and making up ~16 Mb of the 283 Mb assembly. Interestingly, there were 1785 contigs, 79 to 4963 bp in size, labeled as Eukaryota that had 0 coverage from the pre-amplification stage salivary gland reads, the majority of which (1745) were arthropoda, which in turn were mostly labeled (1719) as "Bradysia coprophila" (an alternative name for Sciara coprophila). Such contigs are consistent with L-chromosome sequences that would be present at low copy numbers in the embryo data, but would be absent from the salivary gland data. Exploring these further will be the subject of future investigation. Of the few Eukaryota-labeled contigs that were discarded, 8 were labeled as "Lentinula edodes", more commonly known as Shitake mushroom, which is part of what our cultured Sciara is fed.

To filter out contaminating reads for re-assembly, we removed all paired-end reads where both mates mapped to contigs marked for removal, and retained pairs when at least one mate mapped inside a retained contig or when both mates did not map to the assembly. The retained paired-end reads were used for assembly with Platanus four ways: (i) no quality filtering nor error-correction, (ii) no quality filtering but error-correction with BayesHammer, (iii) quality filtering but no error-correction, and (iv) both quality filtering and error correction. Overall, the contiguity statistics of the re-assembled Platanus contigs did not improve over simply eliminating the contaminating contigs from the original assembly, with the longest contig being ~28 kb in both scenarios. The scaffold contiguity statistics were slightly smaller than the original Platanus assemblies, most likely due to the absence of the very large bacterial scaffolds. Evaluation scores were similar among the four Platanus re-assemblies from different sets of filtered reads. REAPR's percent error-free bases metric improved drastically from ~56% error-free bases in the original assemblies to ~80% error-free bases in the assemblies produced from contamination-filtered reads. The assembly from quality-filtered, error-corrected reads had the highest percentage of error-free bases. We attempted a second round of contamination-filtering on this assembly, but this time the pre-amplification coverage offered no additional benefits as offending low-coverage, non-Arthropod sequences were already removed. There were only 15 contigs labeled as Bacteria, all Rickettsia at the genus level, summing in length to only 2,277 bp.

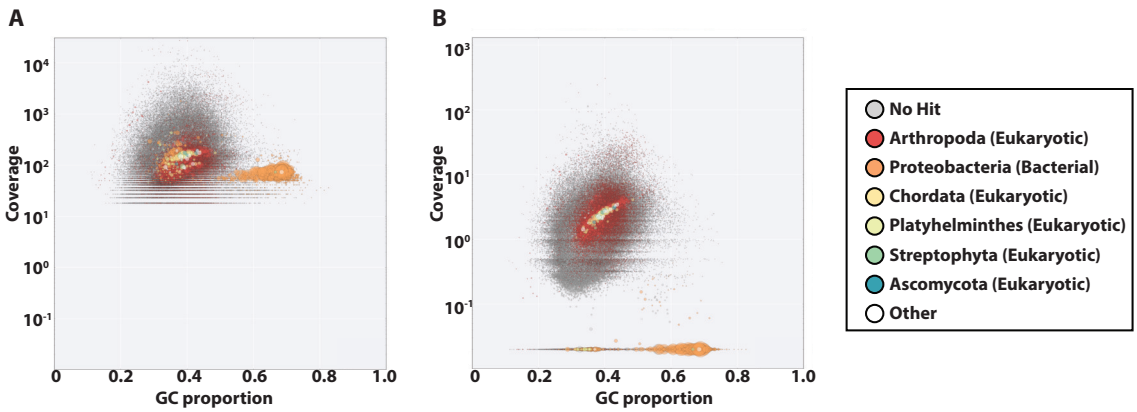


Figure 3.2: Filtering out non-Arthropod, contaminating reads using Taxonomy-annotated GC plots.

Taxonomy-Annotated GC (TAGC) plots can be used to visualize the GC proportion and coverage for each contig in an assembly. In the TAGC plots, each circle represents a contig. The size of each circle is proportional to contig size. The colors correspond to phylum-level taxonomy assignments. Contaminating genomes from the microbiome on embryos, for example, are expected to be at different copy numbers than the target organism. Similarly, contaminating genomes can also have differing GC content. Therefore, when contaminating genomes are present, more than one cluster of contigs is typically seen, allowing coverage and GC proportion cutoffs to be chosen for filtering unannotated contigs. It is not always the case that these two parameters will be enough to differentiate clusters. **(A)** TAGC plot of the chosen Platanus assembly with kmer coverage from the reads input into the assembly on the y-axis and GC proportion of contigs on the x-axis. The bacterial cluster has similar coverage and though it has a higher average GC content, the clusters overlap in that dimension as well. **(B)** TAGC plot with read coverage from pre-amplification stage salivary glands on the y-axis and GC proportion of contigs on x-axis. In this case, using a different sample from a different tissue and prepared by a different person, resulted in differential coverage between the Sciara genome and the bacterial cluster.

The size statistics of our Illumina-based assemblies, including our final one, suggested that they would be unfit for unequivocally mapping the centers of DNA amplicons in the Sciara genome. Nonetheless, since some of the arthropod-labeled scaffolds reached 60 kb, we made a hopeful attempt to possibly extend the known 9 kb DNA puff II/9A sequence by mapping it to all 44 Illumina assemblies we generated. It mapped to contigs ranging from ~1-13 kb in size. In the Platanus, Megahit, and SOAPdenovo2 assemblies, it mapped to a single contig of ~12-13 kb. One ABYSS assembly mapped it across 2 contigs summing to ~16 kb. On the other end of the spectrum, it was divided across 5-6 contigs of ~1-5 kb summing to 9-12 kb for Velvet assemblies. Thus, even in the best cases, we did not gain much sequence information for a major focus of our studies, DNA puff II/9A.

3.3.2 Sequencing Ultra-long DNA molecules with the Oxford Nanopore MinION

It was clear that, for future studies, we would need a higher contiguity assembly to unambiguously map the locations of amplified onionskin structures produced by repeated DNA re-replication in the post-amplification salivary gland genome. Long read technologies were attractive since the majority of reads from PacBio and Oxford Nanopore are longer than most of our Illumina scaffolds (Fig. 3.3 A-C). For example, whereas the scaffold N50 of our final Platanus assembly is 5.3 kb, the sub-read length N50 and molecule length N50 for our complete PacBio and Oxford Nanopore datasets, respectively, were ~10 kb (Fig. 3.3 B-C). We obtained 44.1x coverage from PacBio reads and ~10.77x coverage from molecules sequenced with the MinION (using one read per molecule and including sequences regarded as both pass and fail by ONT), though just ~6.2x of this is from 2D reads and only 2.4x from 2D reads with mean quality scores ≥ 9 (i.e. the subset of reads considered to be of passing quality by ONT).

Our MinION data were collected over the course of 18 months and span various reagent, flow cell, software, and MinION upgrades. From the beginning of that timeframe we modified the standard Oxford Nanopore protocols in order to increase the read lengths, obtaining 2D reads that exceed 100 kb, and reported our early results and protocols in a preprint on bioRxiv [Urban et al., 2015a]. As the ONT protocols changed, we needed to adapt our modifications. The basic principles have stayed the same though and the results are presented here (Fig. 3.3 D-H). The principles we apply to our protocols are the following: (1) Start out with more DNA than required for the standard protocol that assumes 8 kb molecules to target similar molarities, (2) Skip the Covaris shearing step to keep DNA long, (3) Always perform a DNA repair step to repair damaged bases and single-stranded nicks, (4) Use wide-bored tips and very gentle pipetting throughout the protocol, (5) Use 0.4x ratio of AMPure beads in all clean-up steps, (6) Add a rinse step before elution of AMPure beads to deplete DNA $<$ 10-12 kb in all clean-up steps, and (7) Elute DNA off the AMPure beads while adding heat into the system (37-50°C) for extended periods of time (10-20 minutes) in all clean-up

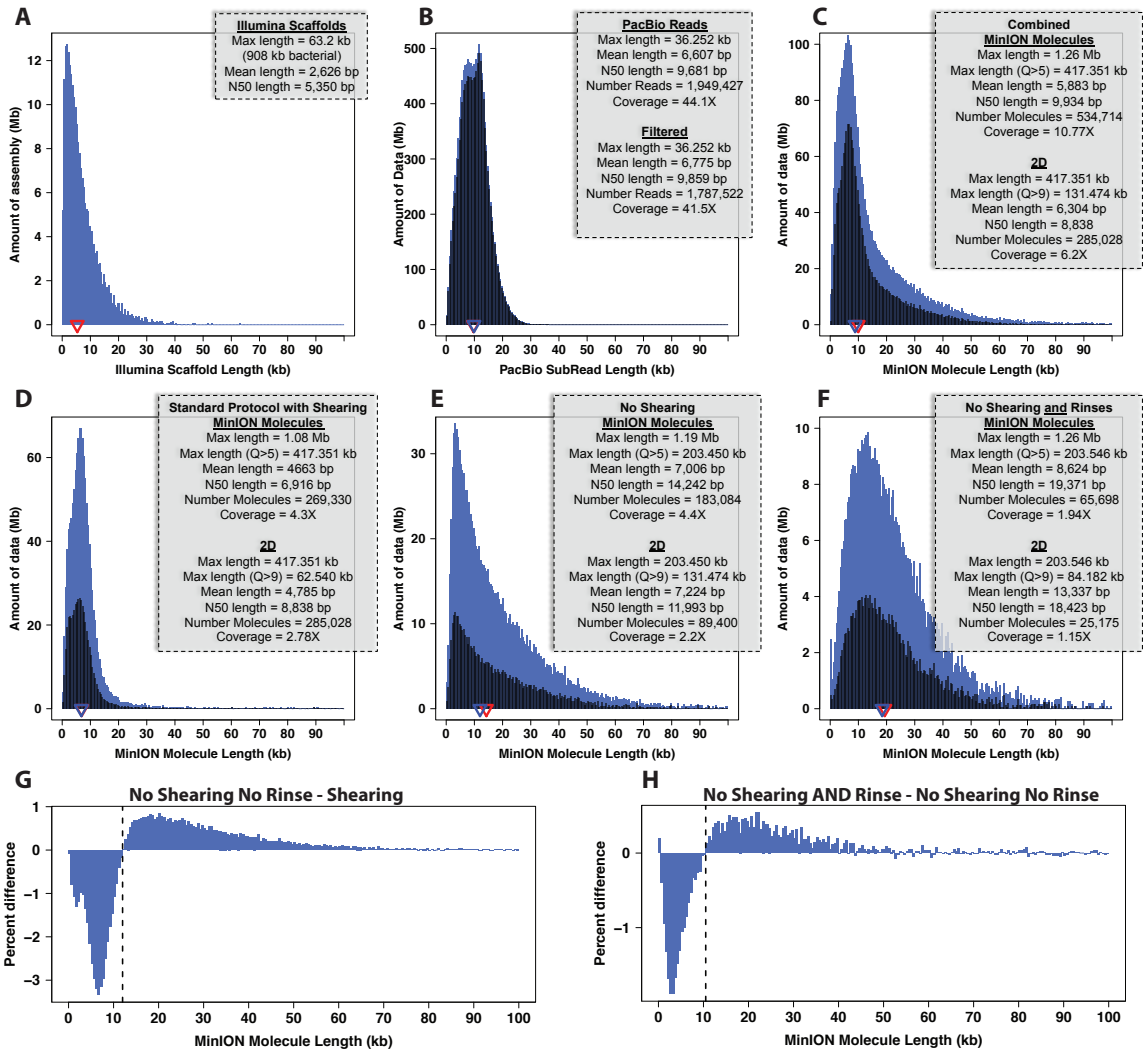


Figure 3.3: Length Distributions for Illumina Scaffolds, PacBio Reads and MinION Molecules.

(A) Platanus scaffold lengths from Illumina data. **(B)** Sub-read length distribution for all PacBio subreads. Shaded area in distribution comes from filtered subreads used. **(C)** MinION molecule length distribution from combining all libraries. Shaded area represents the proportion from 2D reads. **(D)** MinION molecule length distribution from libraries that followed the standard ONT protocol with shearing, targeting 8kb. **(E)** MinION molecule distribution from libraries that skipped shearing and used other long read principles, but did not include rinse steps. **(F)** MinION molecule distribution from libraries that skipped shearing AND included rinse steps. **(G)** Libraries that skipped shearing are depleted for molecules around 10-12 kb and enriched for longer molecules with respect to the libraries from the standard protocol. **(H)** Libraries that skipped shearing AND included rinse steps are additionally depleted for molecules $<$ 10-12 kb and additionally enriched for longer molecules, as demonstrated by comparing to the libraries that only included shearing, but no rinse. Note that libraries were partitioned for D-F, but one library was excluded since it did not fit into these categories as it skipped shearing, but did not include DNA repair, an important step for ultra long reads. Note that MinION molecules are selected by using the 2D read if present, the longer of the template and complement if both are present in the absence of a 2D read, or the template read when only a template read is present.

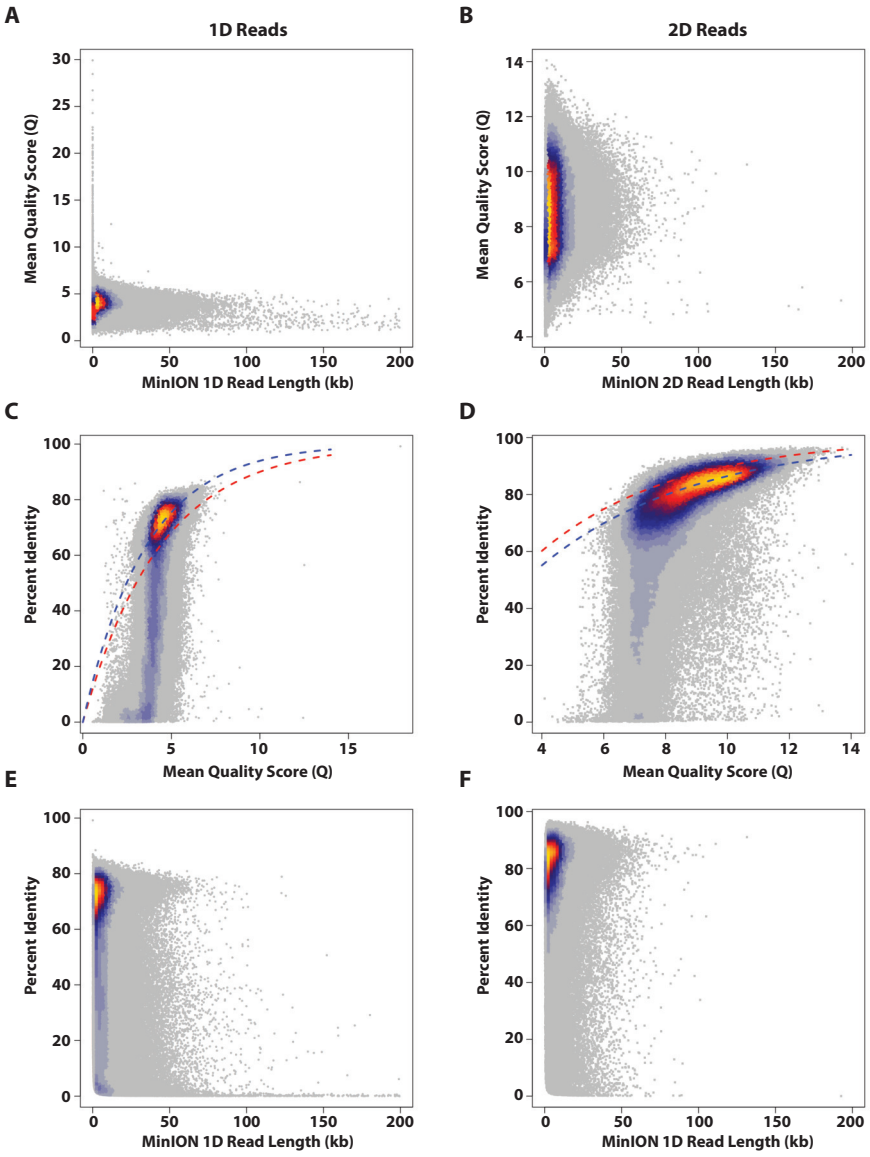


Figure 3.4: Quality scores and percent identities of MinION reads.
Read Length vs Mean quality score (Q) of **(A)**1D reads and **(B)** 2D reads. Mean quality score (Q) vs Percent Identity of **(C)**1D reads and **(D)** 2D reads. In red is the percent identity line that would be predicted from $PctId = 1 - 10^{-q/10}$. We found that slight adjustments to the denominator in the exponent produced predictions more consistent with out data. Specifically for 1D reads the blue line represents $PctId = 1 - 10^{-q/8.2}$ and for 2D reads the blue line represents $PctId = 1 - 10^{-q/11.5}$. Overall, at lower values of Q, reads are too low quality to be aligned from end to end. The percent identity is the result of summing the identities in the section of the read that aligned and dividing by the length of the read. Since we did not perform a re-alignment step [Jain et al., 2015], low quality reads with only small sections mapped appear to have a much lower percent identity than is likely the case. In **(E)** and **(F)**, the Read length vs Percent Identity is shown. for 1D and 2D reads, respectively. For A-B and C-D, the limit on the x-axis was set to 200 kb to better visualize the data shorter than that length. The several reads longer than 200 kb appear to be quite low quality.

steps. It is also important to minimize the amount of handling needed, which has been facilitated by updates to the standard protocol by combining End-Repair and dA-tailing into one step. Previously, we reported that targeting ultra long reads had a trade-off with total data output. ONT has since advised that loading the entire library when targeting long reads helped them reach outputs similar to the standard protocol. We now typically load half the library and assess pore occupancy about 1 hour into the run. If >80-90% of the pores are occupied at any given time, then we wait to add the second half of the library until a later time point (e.g. halfway through the 48 hour protocol). Of the 5 libraries we produced using MAP006 reagents, 3 were from the standard protocol and 2 were from our modified protocols. The average genome coverages per run from molecule lengths for the standard protocol runs and our modified protocol runs were 1.43X and 1.41X, respectively. Thus, we can confirm that there does not need to be a loss in output when pursuing longer reads. All things being equal, or even if they were not, it is arguable that having the longer reads is more valuable to a genome assembly.

Across the iterations of kits, flow cells, and MinIONs, we continued to obtain read distributions with larger N50s than produced with the standard protocol as well as reads that exceed 100 kb (Fig. 3.3 D-H). Specifically, we employed the standard protocol for three libraries, all from the higher throughput MAP006 kit and MkI MinION, which gave a combined molecule N50 of 6.9 kb and 2D read N50 of 8.8 kb (Fig. 3.3 D). This is in agreement with the standard protocol target of 8 kb molecules. When we introduce a subset of the modifications above excluding the rinse steps, the molecule and 2D N50s from 10 libraries, one from MAP006/MkI, rose to 14.2 kb and 11.9 kb respectively (Fig. 3.3 E). For the 3 libraries that included rinses in the clean up steps designed to deplete DNA smaller than 10 kb [Urban et al., 2015a], one of which was MAP006/MkI, the molecule N50 and 2D N50 rose to 19.3 kb and 18.4 kb respectively, over double that from the standard protocol (Fig. 3.3 F). The pooled modified-protocol libraries that skipped shearing, but did not include rinse steps, are depleted for molecules < 10-12 kb and enriched for longer molecules with respect to the standard protocols (Fig. 3.3 G). Importantly, the modified-protocol libraries that included the rinse step are additionally depleted for molecules < 10-12 kb and are enriched for longer ones relative to the modified-protocol libraries that did not include the rinse steps (Fig. 3.3 H). The reproducibility of this effect is shown in Supplementary Figure B.3. Our longest 2D reads ranged from 203.5 to 417 kb and the longest 2D reads with mean quality scores (Q) greater than 9 ranged from 62.5–131.5 kb. The longest base-called 1D reads were in the megabase range, though these seem to represent noise. Amongst all of our libraries, the longest 1D reads with mean qualities greater than 3.5 were $\leq$ 240.4 kb and the longest 1D reads with Q > 5 reached up to 107.5 kb. All in all, for 1D reads with Q > 3.5, we obtained 742 that exceeded 50 kb and 45 that surpassed 100 kb. Moreover, amongst 2D reads, we obtained 522 and 18 that exceeded 50 kb and 100 kb, respectively. Importantly, nanopore reads, including many of these ultra long ones, were validated on PacBio-only assemblies.

To estimate nanopore read accuracy, we aligned the MinION reads to the top ranked PacBio-only

assembly, as defined by metrics in the next section, after both Quiver and Pilon polishing. BWA was able to align 85.7% of 2D reads and 57.9% of 1D reads to this high quality assembly from Canu [Koren et al., 2016]. Percent identities for the alignments were obtained using MarginStats [Jain et al., 2015] using the entire read to define the percent identity, not only the local regions that aligned. Optimized re-alignment was not performed, which can increase the percent identity reported. For this reason and since the assembly may have some errors itself, the percent identities reported here are conservative estimates. Percent identities for both 2D and 1D reads were correlated with mean quality scores (Spearman's rho for 2D = 0.76 and for 1D = 0.70; Fig 3.4 A–B). We found that Q could be used to reasonably predict percent identity with slight adjustments to the normal equation used to convert between the quality scores and percent identity (Fig 3.4 C–D). Similar to mean quality scores, percent identity is relatively consistent across all read lengths for both 1D and 2D reads (Fig 3.4 E–F). The major exceptions seem to be the exceptionally long 1D reads that exceed 200 kb, which are strictly of low quality and low percent-identity, though the latter estimate is also affected by the inability to align long proportions of the extremely noisy reads to the genome assembly. For aligned 2D reads, the median percent identity was 82.1% and it reached as high as 96.9%. Ten percent of the 2D alignments had percent identities higher than 89%. The percent identities of ultra-long 2D reads were similar. For example, we obtained 2D reads of length 131.5 kb (Q=10.3), 111.2 kb (Q=9.9), 105 kb (Q=9.3), and 100.7 kb (Q=10.4) that aligned across their full lengths with identities of 91.1%, 88.7%, 63.2%, and 84.2%. The 102.9 kb 2D read that we previously reported as high quality (8.74) [Urban et al., 2015a] also aligned in full at 84.2% identity, representing a high quality, ultra-long 2D read derived from early MinION sequencing reagents (MAP004). A total of 494 2D reads longer than 50 kb aligned and had a median accuracy of 78.3%. There were 291 1D reads $>$ 50 kb and Q$>$4.5 that aligned with a median percent identity of 67% and ~10% of which aligned with higher than 75% identity. For all 1D reads, the median percent identity was 68% and reached as high as 99.2%, though this was an outlier, the 99th percentile being 80.1%. In sum, we have demonstrated that the validity of the nanopore reads that we have obtained, including many of the extremely long ones, are supported by their alignment to a high quality PacBio assembly. Next we demonstrate that they also make valuable contributions to long read assemblies.

3.3.3 Long read assemblies from single-molecule data

We generated hybrid assemblies starting with our final Illumina-based Platanus contigs using DBG2-OLC [Ye et al., 2016] as well as long-read-only assemblies with several emerging long read assemblers including Canu [Koren et al., 2016], Falcon [Chin et al., 2016], Miniasm [Li, 2016], ABruijn [Lin et al., 2016], and SMARTdenovo (`https://github.com/ruanjue/smartdenovo`). Assembling the Sciara genome with HINGE [Kamath et al., 2016] was also attempted, but it required more computational resources than we had available. For each assembler we tried different inputs: (i) quality-filtered PacBio subreads ("PBfilt", Fig. 3.3 B), (ii) all PacBio subreads ("PBall", Fig. 3.3 B,

shaded proportion), (ii) quality-filtered PacBio subreads with all ONT 2D reads ("PBfilt+ONT2d", Fig. 3.3 B–C and shaded portions), (iii) all PacBio subreads with one ONT read per molecule ("PBall+ONTmol", Fig. 3.3 B-C), and (iv) only for Miniasm, all PacBio subreads and all ONT reads for each molecule ("PBall+ONTall") where each molecule can have up to three reads (template, complement, 2D). In total, there were 50 assemblies chosen to polish and evaluate after selecting several of the best candidates from each assembler for each dataset using size statistics. Canu and Falcon follow similar paradigms where long reads are first error-corrected by mapping them against each other to construct higher quality consensus sequences. Each then uses the error-corrected reads to find overlaps and ultimately generate contigs. Miniasm, ABruijn, and SMARTdenovo share in common the absence of an initial error-correction step, and find overlaps in the raw long reads instead. ABruijn, Canu, Falcon, and SMARTdenovo have consensus steps built into the end of their assembly pipelines. RaCon, a rapid consensus caller [Vaser et al., 2016], was used with Miniasm assemblies to generate a consensus and the consensus approach using pbdagcon (https://github.com/PacificBiosciences/pbdagcon) [Chin et al., 2013] was used for Platanus+DBG2OLC assemblies as recently described [Chakraborty et al., 2016]. All evaluation metrics for Miniasm, which skips both pre-assembly error-correction and post-assembly consensus steps, improved with RaCon as expected [Vaser et al., 2016]. For example, BUSCO found 0 SCOs in Miniasm assemblies before RaCon and 29-32% afterward. Similarly, only 29-31% of Illumina reads mapped and 0% of bases were judged error-free by REAPR in Miniasm assemblies prior to RaCon whereas 94-95% of Illumina reads mapped and 70-71% of bases were judged error-free after RaCon.

The consensus sequences for all assemblies were polished with up to 7 rounds of Quiver using PacBio data [Chin et al., 2013]. We employed extensive polishing under the assumption that the number of variants that Quiver detects in an assembly is the combination of true variants and consensus errors that it can still fix. The assemblies started with a range of variants from 178.6 thousand for ABruijn to 3.2 million for Platanus+DBG2OLC. On average, the variants consisted of 8.7% substitutions, 74.3% insertions, and 17.0% deletions. The only assemblies with variant profiles that slightly deviated from this were from Miniasm that had more deletions than insertions, both differing from the global means by ~1.8 standard deviations. PB-only and PB+ONT assembly-specific mean percentages were very similar to the global means, suggesting there was little difference in error-profiles from different combinations of data inputs. We iterated Quiver polishing until the number of variants stabilized for all assemblers, the majority converging to around 3000-5000 variants (Fig. 3.5 A). Nonetheless, the majority of error-correction from Quiver polishing occurred after the first round where the range dropped from 0.2-3.2 million down to 6.4-16.1 thousand, and as is seen in all metrics (Fig. 3.5 B-G). In the first round, Quiver dropped out contigs that have no coverage from the raw PacBio reads and we chose to leave them out. We noticed that, although contigs were dropped from Canu, DBG2OLC, Falcon, and Miniasm assemblies, there were no contigs dropped from ABruijn and SMARTdenovo assemblies. Moreover, Quiver reports lower-case letters for bases that do not have enough coverage (< 5 here) to compute a consensus on whereas consensus-computed bases are

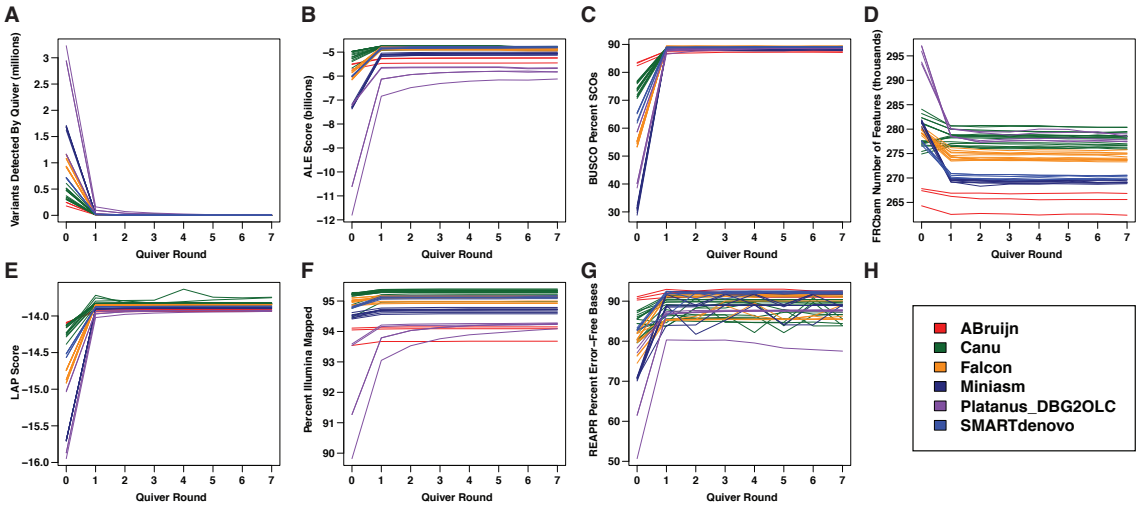


Figure 3.5: Evaluations across Quiver polishing rounds.
(A) Number of variants detected by Quiver. **(B)** ALE scores. **(C)** BUSCO percent complete SCOs. **(D)** Number of features detected by FRC^{bam}. **(E)** LAP scores. **(F)** Percent of Illumina dataset that maps to the assembly.**(G)** REAPR percent error-free bases. **(H)** The legend for all plots. Finer structure of scoring placements can be seen when viewing only rounds 1-7 and leaving out the initial consensus scores (round 0) ((SEE SUPP -TODO)).

uppercase. Most assemblers contained contigs that were 100% lower-case, though these were typically quite small (e.g. $\leq$ 20 kb). ABruin and SMARTdenovo did not contain any contigs that were reported as 100% lower-case. These initial observations suggested that ABruijn and SMARTdenovo assemblies might be of higher quality than others, but the differences in Quiver output discussed above likely reflect decisions made by each assembler about what to do with low coverage contigs. ABruijn and SMARTdenovo likely remove such contigs whereas the other assemblers are more conservative and retain them.

Before Quiver polishing and after each round, the assemblies were evaluated with BUSCO as well as the percentage of Illumina reads that mapped to the assembly, REAPR, LAP, FRC^{bam}, and ALE with the same input Illumina reads used for evaluating the short read assemblies to directly compare scores. For all metrics, the majority of long read assemblies had better scores than short read assemblies either at the initial consensus stage or directly after the first round of Quiver polishing (Fig. 3.6 A-E). Prior to Quiver polishing, ABruijn dominates the rankings, beginning with the lowest number of variants, highest BUSCO scores, lowest number of features, highest LAP scores, and highest percent error-free bases (Figures 3.5 and 3.6). In all cases, the top 10 started with the three ABruijn assemblies and was immediately followed by 7 Canu assemblies. Canu starts out with the highest ALE scores and percent mappable Illumina reads (Figures 3.5 and 3.6), capturing the entire top 10 ranks for both. This suggests that ABruijn has the best consensus module, followed by Canu. The majority of top 10 ranks for Quiver variants, LAP, and ALE comprised PacBio-only

assemblies (Fig. 3.6). Whereas ALE and LAP were Illumina-based metrics, this result for Quiver is hardly surprising since Quiver is using only the PacBio data for this determination. That PacBio-only assemblies have the lowest number of variants is essentially circular, demonstrating that the PacBio-data has the fewest differences with assemblies produced by, and only by, the same PacBio data. The top 10 ranks for BUSCO, FRC, percent mappable Illumina reads, and percent error-free bases by REAPR were dominated by assemblies that combined both PacBio and Oxford Nanopore reads (Fig. 3.6). Thus, for 4 out of 6 PacBio-independent metrics, or 4 out of 7 metrics total, the combination of PacBio and Oxford Nanopore data prevailed prior to polishing.

For the 6th and 7th round of Quiver polishing we lowered the minimum coverage required for Quiver to compute a consensus and identify variants from 5 to 3. This did not seem to impact most metrics for most assemblies in a negative or positive way. However, ALE measures for Canu and Falcon seemed to degrade whereas the measures for all metrics except REAPR continued to improve for PlatanusDBG2OLC assemblies. Therefore, for each assembly, we compared the evaluation metrics from each Quiver round and selected the version from the round with the highest average rank, opting for later rounds in cases of a tie. This resulted in selecting assemblies from Quiver rounds 3-7, with 13 and 19 of the 50 assemblies from rounds 5 and 7, respectively. Canu and Falcon assemblies tended to be selected for earlier rounds whereas ABruijn, Miniasm, PlatanusDBG2OLC, and SMARTdenovo assemblies tended to be selected in later rounds, seemingly reflecting whether or not the assemblers had pre-assembly read correction steps. In the final selections of Quiver polished assemblies, ABruijn no longer was the overwhelming leader (Fig. 3.6 A-E, 3rd columns). All assemblers except PlatanusDBG2OLC were represented in the top 10 for Quiver variants, which ranged from 2.7–3.4 thousand, with Canu in the top two positions. Canu dominated 9 spots in the top ten ALE ranks and the entire top ten of both LAP scores and percent mappable Illumina reads. All assemblers except ABruijn and Miniasm were in the top ten for BUSCO, with Falcon in the top 3. ABruijn and Miniasm were in the top ten for lowest number of features detected by FRCbam. However, this was strictly due to their smaller assembly sizes. When normalizing to assembly size, PlatanusDBG2OLC took the top 3 ranks, followed by Canu and Falcon assemblies. All assemblers except Canu and PlatanusDBG2OLC were in the top ten as judged by REAPR, with the top being the ABruijn assembly combining PacBio and 2D nanopore reads. Six of the ten assemblies with the lowest number of variants detected by Quiver were PacBio-only. This again seems consistent with the fact that Quiver is using only the PacBio data to evaluate the assemblies. However, of the top ten assemblies as ranked by both ALE and REAPR, 6/10 for each were PacBio-only as well (Fig. 3.6 A and E, 3rd columns). Assemblies that combined PacBio and Oxford Nanopore data consumed the top ten ranks for the other metrics, with 8/10 for BUSCO, 7/10 for FRC, 10/10 for normalized FRC, 6/10 for LAP, and 6/10 for percent mappable Illumina reads (for example, Fig. 3.6 B, C, and D, 3rd columns). Thus, the top of the ranks for the majority of metrics were commanded by assemblies that combined PacBio and Oxford Nanopore data once again after Quiver polishing.

After selecting the best version of each assembly from the Quiver rounds, all 50 assemblies were subject to 2 rounds of Pilon polishing with our Illumina dataset. The number of changes Pilon made in the first round ranged from 19.2–25.8 thousand for ABruijn and PlatanusDBG2OLC assemblies at the extremes, respectively. On average, 5.5% of the changes were substitutions, 49.5% were insertions, and 45.0% were deletions. Most assembler-specific mean percentages were within less than 1 standard deviation from these means. The only assemblies to slightly defy this trend were from PlatanusDBG2OLC and SMARTdenovo, which had slightly more substitutions and deletions respectively, both containing fewer insertions. In round 2 the range dropped to 0.9–2.4 thousand Pilon changes made, with PlatanusDBG2OLC again taking on the most corrections and with a Falcon assembly using PBall+ONTmol taking on the fewest changes in this round. The average proportions for types of corrections looked different than in round 1. On average, 43.9% were substitutions (up from 5.5%), 25.6% were insertions, and 30.4% were deletions, with most assembler-specific mean percentages less than 1 standard deviation away from these means. For both rounds, PB-only-specific and PB+ONT-specific mean percentages did not deviate from the global means, again suggesting few to no differences in the error models for these different inputs as seen before polishing. Overall, the the top ten assemblies with the fewest changes introduced in the second round was dominated by Falcon, which took the first 7 ranks followed by Canu assemblies. The top 3 assemblies with fewest changes were from combined PacBio+ONT assemblies, though PacBio-only assemblies also captured 5 of the top ten ranks.

The same set of Illumina reads used to evaluate Illumina assemblies and the long read assemblies during Quiver rounds were used again to evaluate our final Pilon-polished assemblies for direct comparison of the relevant metrics. Canu assemblies captured 9/10 of the highest ALE rankings, 10/10 or highest LAP rankings, 10/10 percent mappable Illumina reads, and 6/10 of normalized number of features, though PlatanusDBG2OLC was in the top 3 positions for the latter (for example, Fig. 3.6 A, C, and D, 4th columns). Falcon took the top 3 ranks for BUSCO, followed by SMARTdenovo and PlatanusDBG2OLC assemblies (Fig. 3.6 B). Abruijn scored the top position for percent error-free bases from REAPR, though all assemblers except Canu and PlatanusDBG2OLC were in the top ten (Fig. 3.6 E). In the top ten ranked assemblies for ALE, 6/10 were from PacBio-only assemblies. Moreover, though the assembly with the highest percent of error-free bases was from a combination of PBfilt+ONT2d, 6 of the top ten were PacBio-only. In contrast, the percent of mappable Illumina reads, LAP, BUSCO, FRC, and normalized FRC favored combinations of PacBio and Oxford Nanopore data. Specifically, combination datasets scored the top 6 ranks for percent mappable Illumina reads, the top 3 and 6/10 of the top BUSCO ranks, the top 5 and 6/10 leading LAP ranks, 7/10 of the best FRC ranks, as well as the top 4 and 9/10 of the highest normalized FRC ranks.

We also evaluated our final assemblies with a variety of other metrics incorporating PacBio, Oxford Nanopore, and BioNano data. We used the restriction map aligner called “maligner” [Mendelowitz et al., 2015] to align the raw BioNano optical maps to all 50 assemblies. Maligner reports an

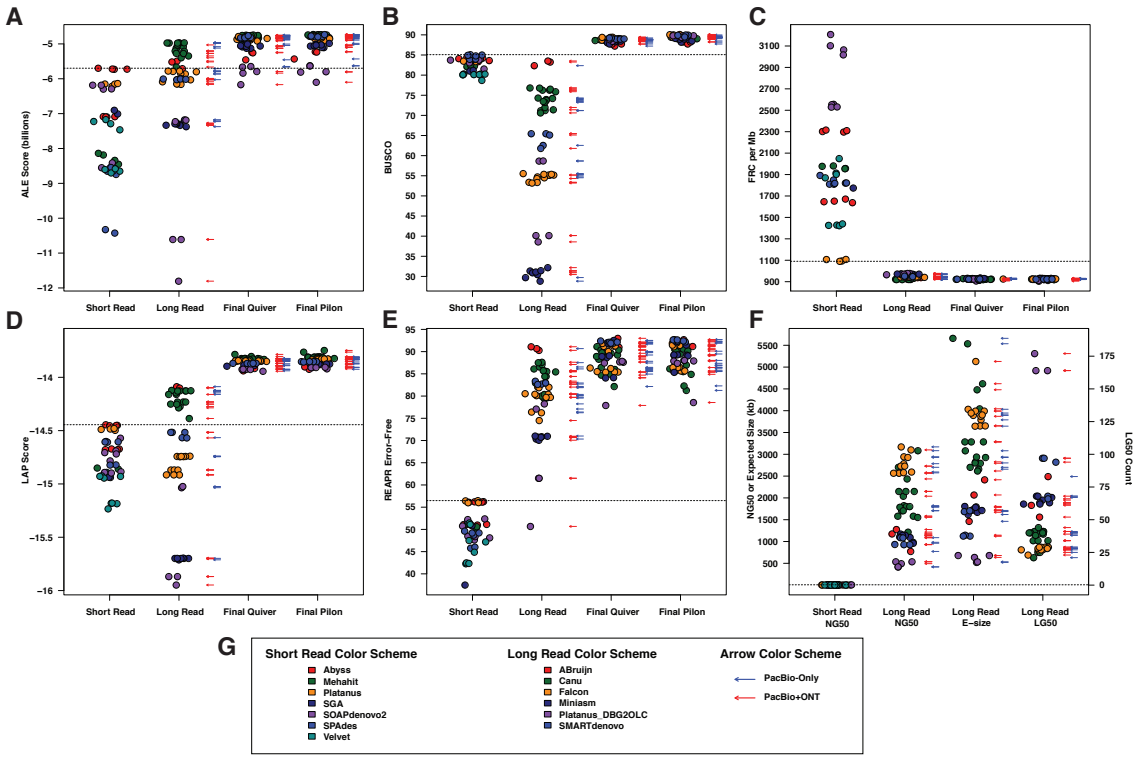


Figure 3.6: Comparing evaluations of short read assemblies to long read assemblies. **(A)** ALE scores. **(B)** BUSCO percent complete SCOs. **(C)** Number of features detected by FRC^{bam} normalized to number of features per megabase of assembly. **(D)** LAP scores. **(E)** REAPR percent error-free bases. **(F)** Size statistics. NG50 = size of contig such that at least 50% of the expected genome size is on contigs of that size or larger. Expected contig size is as defined in the Genome Assembly Gold-standard Evaluations (GAGE) paper [Salzberg et al., 2012], which is designed to give the expected size of contig containing a base in the assembly selected at random. The expected genome size was used in the denominator of the equation instead of individual assembly sizes for direct comparisons as done in the GAGE paper. LG50 is the number of contigs it takes to reach at least 50% of the expected genome size when selecting contigs from longest to shortest. Lower LG50s are better. **(H)** Legend for A-F.

"M-score" for each alignment. We ranked assemblies using combined M-scores, total span across the assembly from alignments, total coverage as determined by summing the reference interval lengths for each alignment, and the number of BioNano maps that could be aligned. We also used the long reads from PacBio and Oxford Nanopore both separately and in combination to evaluate the 50 assemblies. We used the long read aligner BWA [Li and Durbin, 2009] and the structural variation (SV) algorithm called "Sniffles" [`https://github.com/fritzsedlazeck/Sniffles`] to detect the number SVs in each assembly. Reasoning that the number of SVs is a combination of both real variants and errors in the assemblies, assemblies were ranked higher for having lower numbers of reported SVs. We also separately considered the the sum of the interval lengths reported to be involved in SVs and refer to this as SV span. For the same reasoning as above, assemblies were ranked higher for having shorter SV spans. In some cases, assemblies can have short spans, but high numbers of reported SVs and vice versa. We called SVs separately for PacBio and ONT reads as well as when combining them. We also simply looked at the percentage of PacBio and Oxford Nanopore reads that aligned to each assembly, ranking assemblies higher for having a higher percentage. Finally, in addition to considering only the percentage of long reads that aligned, we also considered the ratio of the number of alignments to the number of unique reads represented in those alignments. Since reads are sometimes broken to align in separate places or in different orientations, the ratio is greater than 1. Similar to the more formal SV results from Sniffles above, we reasoned that split reads result from true biological variation, errors in the reads, and mis-assemblies. Since the first two are constant across all conditions, the only variable is the number of mis-assemblies. The closer to 1 the ratio was for an assembly, the higher it was ranked. All of the BioNano, PacBio, and Oxford Nanopore metrics were highly correlated with each other and with most Illumina-based metrics including percent mappable reads, ALE, LAP, the number of Pilon changes made in round 2, and to a lesser extent the normalized number of features reported by FRC (Fig. 3.7 A). There was no correlation with REAPR metrics and a slightly negative correlation with FRC without normalization. The single-molecule data had the same low correlation levels with BUSCO as seen for most Illumina metrics except normalized FRC. Overall, it was satisfying that the metrics from 4 different technologies largely agreed with each other.

For all 4 BioNano map metrics, Falcon assemblies scored 10/10 of the highest ranks. For the highest numbers of alignments, a PB-only assembly took the top position, but both PB-only and PB+ONT assemblies took 5 of the top ten ranks. Similarly, two PB+ONT assemblies ranked highest for M-scores, but both types captured 5/10. However, for total span and total coverage, PB+ONT assemblies were favored, capturing the top 2 positions and 6/10 of the highest ranks. BioNano maps spanned a range of 181–252 Mb across the 50 assemblies, with the shortest span in the top 10 being approximately 249 Mb. Canu assemblies were close to Falcon, taking most of ranks between 10 and 20 with the first 6 Canu assemblies being from combinations of PB+ONT where BioNano maps spanned 237-242 Mb of the assemblies. Importantly, although the assemblies from earlier versions of Canu (1.0-1.2) often had longer size statistics, assemblies from the most recent release of Canu

used (1.3) had the most support from all 4 BioNano metrics. Following Falcon and Canu in the ranks, BioNano maps spanned 214-229 Mb of Miniasm assemblies, 225—228 Mb of SMARTdenovo assemblies, 202–221 Mb of Abruijn assemblies, and 181–200 Mb of PlatanusDBG2OLC assemblies. Two of the 5 SMARTdenovo assemblies were PacBio-only, and these were the two highest ranks for this assembler. In contrast, the top 6 of 8 Miniasm assemblies, the top 2 of 3 ABruijn assemblies, and the top 3 of 5 PlatanusDBG2OLC assemblies were from combining PacBio and ONT datasets. The same trends were seen for BioNano M-scores, total coverage, and number of alignments. Overall, this re-iterates the emerging theme that favors assemblies derived from a combination of data types. Though this might simply be from higher coverage, the amount of extra coverage from Oxford Nanopore data was modest at an extra 6.2X for assemblies that incorporated only 2D reads.

For metrics from long reads, including number and span of SVs detected by Sniffles, percentage of reads that mapped, and the alignment ratio, Canu primarily dominated the top ten of all metrics. For all PacBio-specific metrics, it encapsulated all 10/10. For ONT-specific metrics, it captured 7/10, 7/10, 10/10, and 9/10 for number of SVs, SV span, alignment ratio, and percentage of reads that mapped, respectively. Falcon took 3 of the top 5 ranks for fewest detected SVs as well as the top 2 positions and 3/10 total for having smallest SV span. The PlatanusDBG2OLC assembly using PBall+ONTmol made a surprise visit to the number one spot for percentage of nanopore reads aligned. When combining PB and ONT data to call SVs, Canu had 8/10 of the lowest number of SVs and 9/10 of the shortest SV spans. Nonetheless, the SMARTdenovo assembly that used PBfilt+ONT2d took the top position for both metrics, and a Falcon assembly that used PBall+ONT2d made it into the top 10 of fewest SVs. Regarding what datasets were favored, unsurprisingly ONT-specific measurements tended to favor assemblies that incorporated ONT data. The top 7 assemblies with fewest SVs, 6/10 of the shortest SV spans, the top 6 alignment ratios, as well as the top 3 and 7/10 total of highest percentage of ONT reads mapped were from assemblies combining PB and ONT data. These trends were seen at the assembler-specific level for most assemblers as well. For example, for all 4 ONT metrics, the top assemblies from Canu, Falcon, Miniasm, ABruijn, and PlatanusDBG2OLC are all from PB+ONT datasets, strictly ranking before all PacBio-only datasets for the latter two assemblers. In contrast, SMARTdenovo favored PacBio assemblies for all ONT metrics. There was less favoritism in PacBio-specific metrics to PacBio-only assemblies where they captured the top 2 and 5/10 of the fewest SVs, the top 4 and 5/10 of the shortest spans, the top 1 and 3/10 of the lowest alignment ratios, and 5/10 of the assemblies with highest percentage of PacBio reads mapped. PB+ONT assemblies had 7/10 of the lowest PacBio alignment ratios as well as the top 2 assemblies with most PacBio reads mapped. For ranks specific to each assembler, when evaluating with the 4 PacBio-based ranks, PacBio-only assemblies claimed the top rank for Canu 3 times, Falcon once, Miniasm 3 times, platanusDBG2OLC once, and SMARTdenovo three times. However, PB+ONT assemblies took the majority of top positions ranking first in the 4 PacBio-based metrics once for Canu, 3 times for Falcon, 4 times for ABruijn, once for Miniasm, 3 times for PlatanusDBG2OLC, and once for SMARTdenovo. When SVs were called from combined

datasets, perhaps unsurprisingly, the assemblies produced from the combined datasets won out with the top 4 and 8/10 of the fewest SVs as well as the top 1 and 7/10 of the shortest SV spans. At the assembler-specific level, assemblies from the combined PB+ONT datasets came in first place for 5/5 of the assemblers for lowest number of variants and 4/5 of the assemblers for shortest span. In general, the frequency of SVs was quite low in the top ranking assemblies. ONT datasets called only 8-50 SVs spanning 100-1108 bp, though this number could be higher with higher coverage or setting the parameters for SV calling to be less stringent to compensate for the lower coverage. The higher coverage PacBio dataset called 26-446 SVs spanning 0.3-9.9 kb. When combining PacBio and MinION data for even higher coverage, there were 49-604 SVs in the assemblies spanning 0.8-13.2 kb. Overall, the highest ranked assemblies had few SVs detected, suggesting few mis-assemblies.

The final Pilon-polished assemblies ranged from 281.5 Mb for ABruijn to 306.6 Mb for Platanus-DBGOLC. The average absolute distance of the assembly sizes from the expected male somatic genome size of 292 Mb was 6.14 Mb. Falcon and Canu had the largest NG50s, with Falcon taking the top 2 spots for a PacBio-only and a PBall+ONTmol assembly with NG50s of 3.17 and 3.1 Mb respectively. The highest Canu NG50 was 3.08 Mb for an assembly using filtered PacBio reads and an early version of Canu (1.0) after employing special parameters in the Bogart step for diploid genomes. This Canu assembly had the lowest LG50 of 21 contigs (number of contigs with at least 50% of the genome) followed by the PBall+ONTmol Falcon assembly, which had 23. Two Canu 1.0 assemblies using the diploid approach had the highest normalized expected contig sizes [Salzberg et al., 2012] of 5.5-5.7 Mb with the PBall+ONTmol Falcon assembly following close behind at 5.1 Mb. The top 3 assemblies for maximum contig size were from Canu 1.0-1.1 assemblies with diploid settings using PBfilt, PBall, and PBall+ONTmol that had longest contigs of 20.9–26.7 Mb. The Falcon PBall+ONTmol assembly was close behind with a 20.1 Mb contig, followed by a Canu 1.2 assembly using PBall+ONTmol containing a 19.5 Mb contig. For all size metrics above, the top ten contained only Falcon and Canu. Falcon had slightly higher representation than Canu in the top ten positions for NG50 and LG50, but they each had 5/10 for expected size and longest contig. PB-only assemblies took 7/10 of the top ten NG50 positions, but PB+ONT took 6/10 of the top LG50 positions, 6/10 of the top expected size ranks, and 6/10 of the assemblies with longest contigs. Importantly, all long read assemblies were orders of magnitude more contiguous than short read assemblies with 56–1200-fold larger NG50s when considering all assemblies from both groups. All of the long read assemblies exceed our target NG50 of 100 kb. Moreover, our known 9 kb sequence of DNA puff II/9A mapped to contigs ranging in size from 232 kb to 26 Mb, serendipitously mapping to the longest contigs in many assemblies. Therefore, even before further scaffolding with BioNano optical maps, our goals for the first release of the Sciara genome have been met with long reads.

To select a final subset of assemblies for BioNano hybrid scaffolding, we sorted the assemblies by taking mean ranks across combinations of all 27 evaluation metrics used on the assemblies. Since it was possible that some assemblies have higher mean ranks due to a intrinsic weighting biases present

across this set of 27 evaluations, we tried 40 different combinations of the 27 metrics with as few as 6 metrics in one combination and with as many as all 27 in another. The first twenty combinations did not include the size statistics ranks of NG50, LG50, expected size, and longest contig. This was done to uncover assemblies that ranked high without giving preference to more contiguous assemblies. The second set of 20 combinations did include size rankings. For both the first and second sets of twenty, the trend was roughly to use fewer metrics from the first to last combination. Moreover, since different categories (BUSCO, Illumina, PacBio, ONT, BioNano, size statistics) tended to favor different assemblers, the trend was to converge to more similar representations from each category for each set of twenty. At least one metric from each category was represented in all combinations except in the first twenty where size statistics were excluded. Reflecting the results of individual metrics discussed above, both Canu and Falcon assemblies dominated the top of all of these averaged ranks (Fig. 3.7 C). The averaged ranks were highly correlated (Fig. 3.7 D), differing primarily in whether Canu or Falcon rose to the top. By category, Canu seems to dominate Illumina and long read metrics from ONT and PB. Since 18 of the 27 metrics fit into these categories, more of the metrics that we used favored Canu than Falcon. It could be argued that some of these metrics were redundant. For example, percent mappable Illumina reads tracked nearly perfectly with the Illumina-based LAP and ALE scores with correlations above 0.9. Similarly, the SV results from combining ONT and PB data are very similar to the SV results from PB-only, with correlations above 0.86. This is likely because PB made up 80-88% of the reads and both were processed by the same algorithm. Consequentially, combinations that included most metrics or many more Illumina metrics than BioNano or NG50 metrics, for example, were weighted towards Canu. In contrast, Falcon was categorically favored by BUSCO, BioNano, size statistics, and some ONT metrics. Since the BioNano metrics are highly correlated with NG50, it could be argued that this could possibly explain the slight BioNano preference for Falcon assemblies. Nonetheless, when fewer metrics were used, when size statistics were introduced, and when categories became more balanced, Falcon became the rank leader at least as often as Canu. Altogether, both assemblers produced extremely high quality assemblies that are difficult to choose between. In both cases, when you look at the actual scores between the top performing Canu and Falcon assemblies for a given metric, the difference is not dramatic. For BUSCO, the difference is 0.37% and the differences between percentages of Illumina, PacBio, and ONT reads that mapped are 0.2%, 0.12% and 0.11%, respectively. The top Falcon and Canu assemblies only differ by rates of 3.6 features per megabase from FRCbam, 0.6 base corrections per megabase in the final Pilon round, 0.14 SVs per megabase as determined by combined data, and 2.7 BioNano map alignments per megabase. Moreover, the mean base scores of the top ranked Canu and Falcon assemblies from REAPR are nearly identical. Similarly, the top NG50s of each assembler differ by only 87 kb and the difference in LG50 between the averages of the top 5 for each assembler is 0.4. For the metrics mentioned, Canu and Falcon win equal shares. Moreover, the top ranked assemblies for any given metric were often virtually the same assemblies with slight parameter tweaks. If that redundancy is accounted for, it is an even closer contest. For both assemblers, the assemblies using combinations of PacBio and ONT data were the clear rank

leaders. Given the high performance of both assemblers and specifically the high ranks of assemblies from combined datasets, we chose two PB+ONT assemblies for each Canu and Falcon for BioNano hybrid scaffolding.

In our BioNano data, there were 438,139 molecules >100 kb and 217,194 >150 kb. The molecule N50s for each subset were 173.8 and 214.1 kb, respectively. For each subset, the optical maps were aligned to each other to find overlaps and develop a consensus map for the Sciara genome. As using the assemblies helps create the consensus maps, this process was repeated for each of the 4 assemblies independently. The resulting consensus maps were slightly more contiguous when using only molecules longer than 150 kb with a map N50 of 712 kb and a cumulative length of 325.5 Mb. Therefore, these were used going forward. The BNG consensus maps spanned 266-278 Mb of the assemblies, with the longest span belonging to the Falcon assembly from PBall+ONTmol. Assemblies were turned into in silico maps and aligned to the BioNano consensus maps to develop the hybrid scaffold map. In hybrid scaffolding, BioNano consensus maps are further joined when supported by contig overlaps and contig maps are joined when supported by consensus map overlaps. The result is a genome-wide map that is typically more contiguous than both of the input maps. The assembly contig maps and BioNano consensus maps had similar spans across the resulting hybrid scaffold map of approximately 275–280 Mb. The N50s of the final scaffolded assemblies more than doubled in all cases. The N50 for the Falcon PBal+ONTmol assembly more than tripled with a final N50 of 9.3 Mb.

3.3.4 DNA modification signatures in single-molecule data

The term "imprinting" was originally coined to describe the observation in the fungus fly of paternal chromosomes being specifically targeted for elimination [Crouse, 1960a, Rieffel and Crouse, 1966, Crouse et al., 1971]. However, it is still unclear how paternal chromosomes are differentiated from the maternal set [Sánchez, 2014]. Differential DNA modification is an attractive hypothesis, and there is evidence of 5-methylcytosine (5mC) as detected by an antibody along polytene chromosomes, particularly in heterochromatic regions [Eastman et al., 1980, Greciano et al., 2009]. DNA methylation is also present in many insects, where it locates primarily to gene bodies [Field et al., 2004, Lyko and Maleszka, 2011, Glastad et al., 2011]. It is unknown whether Sciara has other base modifications as well, but recent evidence from Drosophila suggests that it has N6-methyladenine (6mA) [Zhang et al., 2015] in addition to 5-methylcytosine [Capuano et al., 2014, Takayama et al., 2014] and 5-hydroxymethylcytosine (5hmC) [Rasmussen et al., 2016]. Moreover, 6mA has been identified in other eukaryotes as well, including green alga and C. elegans [Heyn and Esteller, 2015, Fu et al., 2015, Greer et al., 2015]. Kinetic signatures in PacBio Single Molecule Real Time (SMRT) sequencing data can be used to identify various base modifications, including 5mC, 6mA, and 4-methylcytosone (4mC) [Flusberg et al., 2010, Clark et al., 2012, Suzuki et al., 2016]. The less common 4mC was detected in eukaryotes using SMRT kinetic signatures [Ye et al., 2016]. Anomalies in SMRT kinetic data were also used to detect base J in trypanosomes [Genest et al., 2015]. Similarly, ionic current

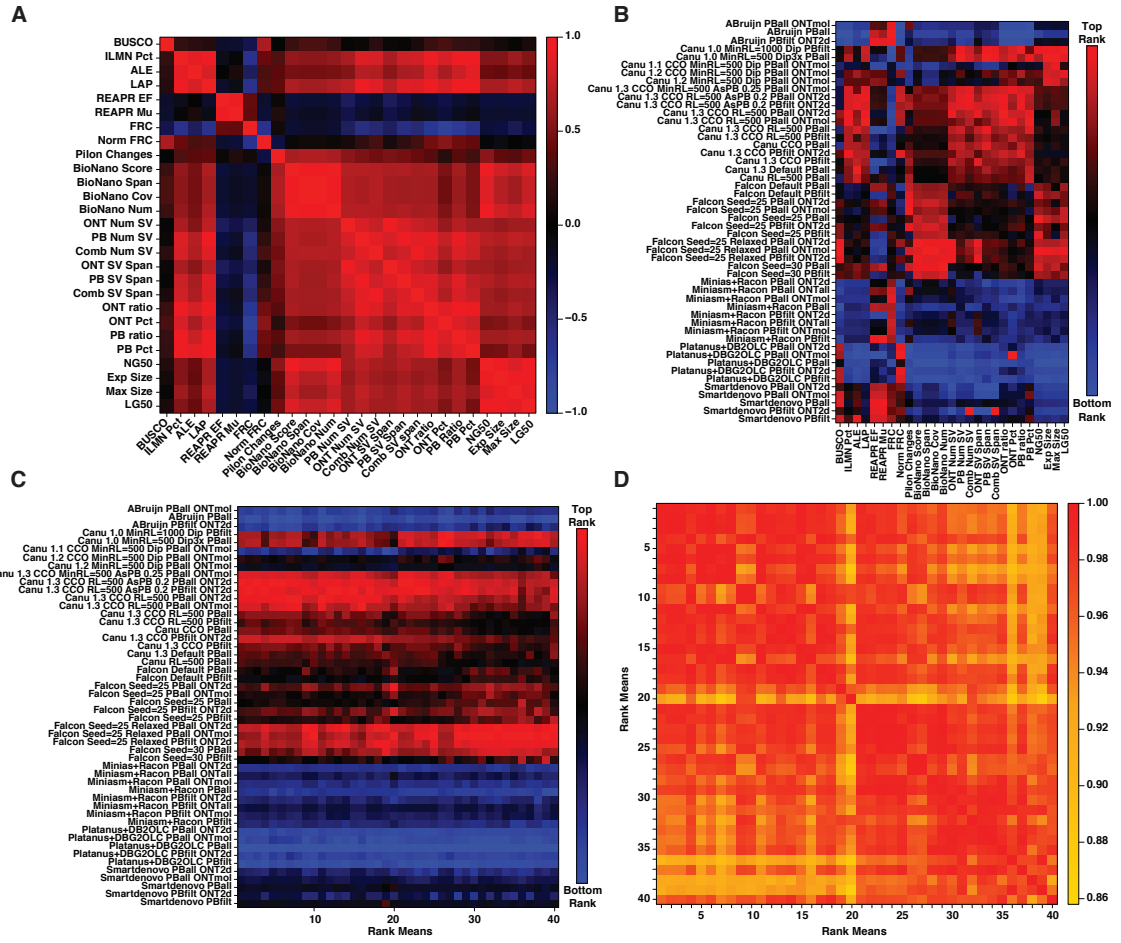


Figure 3.7: Comprehensive evaluation of the long read assemblies.
(A) Correlation matrix of all ranks used for evaluating long read assemblies. **(B)** Rank matrix for the 50 long read assemblies (rows) with the ranks for each metric (columns). There were 27 metrics total. **(C)** Rank matrix for the 50 long read assemblies (rows) and their rankings after averaging various combinations of individual ranks (columns). The average ranks of first 20 did not include size statistics whereas the second 20 did. From 1-20 and from 21-40, the trend was to use fewer metrics and to reach a better balance between the metric catgories: BUSCO, Illumina, BioNano, PB, ONT, and size statistics (only in 21-40). **(D)** Correlation matrix of rankings for the 40 different rank sortings after averaging various combinations of ranks. All are highly correlated, mainly differing in whether Canu or Falcon assemblies were the rank leaders.

signatures in MinION data have been used to predict 5-methylcytosine, 5-hydroxymethylcytosine, and 6-methyladenine [Simpson et al., 2016, Rand et al., 2016]. Since we have access to both SMRT kinetic and ionic current signal from the MinION, it was an opportunity to see if two completely orthogonal approaches could provide evidence of base modifications in the Sciara genome. Using the SMRT kinetics data, we identified 67,901 sites in the genome demonstrating the 6mA signature. Similarly, we identified 272,796 and 119,137 sites with the signatures for m5C and m4C, respectively. There was over 1 million sites, many with high scores, that were reported as "modified base" without being assigned to one of the three categories. Using 40 bp sequences surrounding the candidate modification sites, CG-rich motifs arise from m5C and m4C sites whereas m6A sites identify A-rich motifs with 2-4 As typically flanked by 1 or more Gs. An orthogonal algorithm that uses the kinetics data to query blocks of CpG dinucleotides [Suzuki et al., 2016] interrogated over 10.5 million CpG dinucleotides in the Sciara genome and marked 7.3% as hypermethylated, leaving stretches of hypomethylated regions. In the MinION data, ionic current is translated into the 6mers that likely gave rise to it, which can then be interpreted as a DNA sequence. The long DNA sequences were aligned to the genome, subsequently allowing the ionic current events corresponding to each 6mer to be aligned. With events aligned to the genome, emissions parameters for each 6mer can then be updated and the events realigned [Simpson et al., 2016]. Updating the parameters and realignment were iterated for 5 rounds. At the end, we were able to compare the emissions parameters from each 6mer from the Sciara data to the original ONT emissions parameters to find 6mers that were significantly different. When we tested this procedure on E. coli data, 32 of 34 kmers that had differences in the emission means greater than 1 contained GATC, the known motif for adenine methylation. When we ran it on Sciara using only the passing quality 2D reads, there were 196 6mers in the template strand model with a difference greater than 1, 119 of which contained AA and 64 that contained AAA. Running them through MEME to identify motifs resulted in CG and AC rich motifs. There are two complement strand models to train. Model 1 contained 108 6mers and model 2 contained 111 6mers that had emission parameters that differed by more than 1 pA from the ONT model. Similar to the template model, these were rich in CG dinucleotides and poly-A sequences. When divergent 6mers from all 3 models are combined, MEME identifies primarily CG motifs whereas DREME identifies poly-A/poly-T motifs. All in all, the motifs identified by both PacBio and Oxford Nanopore data are consistent. Both technologies identify CG as a motif for DNA modification, the known motif for 5-methylcytosine methylation in eukaryotes, consistent with studies that found cytosine methylation on polytene chromosomes. Moreover, both identify A-rich signatures. Complementing the detection of base modification signatures for cytosine and adenine methylation, the Sciara transcriptome contains transcripts with high degrees of homology with DNA methyltransferase (DMNT), Ten-eleven transloation (TET), and DNA N6-methyl adenine demethylase (DMAD) sequences from Drosophila, C. elegans and humans, demonstrating that Sciara has the machinery commonly associated with 5-methylcytosine and 6-methyladenine regulation.

3.4 Discussion

The genome of the fungus fly Sciara coprophila presents many opportunities to study unique fundamental biological questions. Sciara has different genomic, chromosomal, and locus-specific copy numbers in different cells and has tissue-specific chromosomes limited to the germ line after somatic chromosome elimination. Sex determination in Sciara is governed by females via a modified X-chromosome [Metz and Schmuck, 1929, Metz, 1931, Sánchez, 2008, Sánchez, 2014]. Specifically, females that have two copies of the X give rise to male offspring whereas females with one copy of the X and one copy of the X' give rise to only female offspring. Sciara has great resistance to genomic damage by X-irradiation [Metz and Boche, 1939, Bozeman and Metz, 1949, Crouse, 1949] and stressors such as being kept in the cold induce a dauer-like, reversible state of suspended animation, increasing Sciara lifespan >10-fold (our unpublished observations). Finally, chromosome "imprinting" was first discovered in Sciara [Crouse, 1960a] when it was observed that all paternally-derived chromosomes are eliminated in male meiosis on a unique monopolar spindle. Similarly, embryonic X chromosome elimination is restricted to paternally derived X-chromosome in normal development. Studies into these fascinating biological features have been inhibited by the lack of both a genome sequence and transgenic techniques. We recently demonstrated precise and targeted gene insertion in the Sciara genome using the preferred pathway of non-homologous end-joining [Yamamoto et al., 2015]. Now we present a high quality genome sequence to add to the growing toolbox for studying Sciara coprophila.

We approached the problem of assembling the genome of Sciara coprophila from several angles. First, we obtained high coverage Illumina data and generated 44 assemblies using several assemblers. We found that Platanus and Abyss out-performed the other assemblers in most metrics. However, SPAdes produced the most contiguous assemblies with the most detectable single copy orthologs. Nonetheless, we set a target NG50 of 100 kb in order to ensure feasibility of mapping sites of intrachromosomal DNA amplification in the salivary gland genome in future studies and none of the Illumina assemblies met our needs. We therefore obtained long, single-molecule data from Pacific Biosciences and Oxford Nanopore Technologies. We modified library preparation protocols from Oxford Nanopore in order to obtain 2D reads exceeding 100 kb. The majority of both 1D and 2D Oxford Nanopore reads mapped to an assembly produced from only PacBio reads, showing median percent identities of 68% and 82.1% respectively. These analyses both confirmed the validity and value of the ultra long nanopore reads as well as supported the structural integrity of the PacBio assembly. Of note, a 131.5 kb 2D read had an impressive 91.1% identity. From the point of view of a nanopore, a 2D read of this length is equivalent to sequencing >230,000 bases. Such read lengths may become more common, particularly with 1D library preparations. All of the long read assemblies were very close to the expected genome size of 292 Mb. The orthogonal BioNano Irys optical mapping data spanned up to over 250 kb of these approximately 292 Mb assemblies when mapping the raw data with Maligner. The consensus BioNano maps made during the Hybrid Scaffolding process spanned up to 278 Mb of the final assemblies chosen for scaffolding. Overall,

the majority of metrics employed favored assemblies produced with the combination of PacBio and Oxford Nanopore reads. Though this may only reflect the increase in long read coverage from the nanopore reads, it is tempting to speculate that the presence of extremely long nanopore reads and having a combination of long reads from different technologies played roles in this effect. The addition of a modest 6X coverage of only 2D reads resulted in assemblies that typically outperformed PacBio-only assemblies for all assemblers except SMARTdenovo. All assemblies produced from long read technologies exceeded our NG50 target, with 1000-fold larger NG50s in some cases. Although all of the long read assemblers produced assemblies that were of high enough quality for our purposes after polishing, the hierarchical non-hybrid assemblers, Canu and Falcon, outperformed the assemblers that skip pre-assembly error-correction steps and the assembler that employed a hybrid approach with short read contigs. Canu and Falcon were top contenders for assembling the Sciara genome with our datasets and were chosen for BioNano hybrid scaffolding. The final scaffolded assembly had an NG50 exceeding 9 Mb. This assembly is more than suitable for mapping the 18 DNA puffs that undergo DNA amplification in the larval salivary glands and this is already underway.

The Sciara coprophila assembly can be improved further in the future by evaluating it with and incorporating more long reads and other long range information. For example, Oxford Nanopore has since released R9 and R9.4 pores that yield accuracies exceeding 90% in conjunction with their new recurrent neural network base-caller. Moreover, ONT has demonstrated long read protocols using size-selection that produce libraries with average read lengths of 50 kb and the throughput of the latest MinION flow cells could potentially double our nanopore data in a single run. Ideally, in the near future we will be able to obtain >10X high accuracy coverage per flow cell coupled with our long read protocols. Pacific Biosciences has also released a new higher throughput machine called the Sequel that produces 7x more data per SMRT cell [`http://www.pacb.com/products-and-services/pacbio-systems/sequel/`]. Generally speaking, obtaining another 50X long read coverage would allow more stringent quality filtering as well as permit higher length cutoffs and higher minimum overlap requirements [Koren and Phillippy, 2015]. NabSys employs a recognition sequence mapping method similar to BioNano, but detects the sequence-specific tags as they are driven through semiconductor-based nanodetectors at over 1 million bases per second [`http://www.nabsys.com`]. Instead of optical mapping, NabSys refers to their technology as "electronic mapping" or "high definition whole-genome mapping" and claim that it has much higher resolution, a lower error rate, and higher throughput than its optical counterparts. We have begun to explore this technology for future updates to the Sciara genome. Linked reads from 10X genomics [Mostovoy et al., 2016] and long-range interaction data from from HiC [Kaplan and Dekker, 2013, Burton et al., 2013, Marie-Nelly et al., 2014] or the Dovetail Genomics Chicago method [Putnam et al., 2016] have also been demonstrated to increase assembly contiguity, and represent possible future directions. Leveraging the data we already have to update the assemblies with scaffolding and finishing tools [English et al., 2012, Lam et al., 2015] as well as assembly-merging and reference-guided tools [Chakraborty et al., 2016, Kolmogorov et al., 2016] might be avenues to explore as well.

Nevertheless, it is important to maintain a balance between assembly contiguity and the number of mis-assemblies introduced, and we will continue to leverage data from orthogonal techniques to ensure the structural integrity in future releases of the Sciara genome. We will soon begin to use probes designed from the longest contigs in the assembly for 1- and 2-color FISH to help anchor, order, and orient contigs along the giant polytene chromosomes in salivary glands. These experiments can also flag mis-assemblies when two probes of different color from the two ends of a contig map to different chromosomes. We will also begin to explore the X' chromosomes found in gynogenic females and the germ-line limited L chromosomes.

Despite our plans to continue improving the Sciara reference genome, this first release of the Sciara coprophila genome sequence is more than sufficient in its current state for our purposes of expanding our research from one to multiple DNA puffs, and is ready for a variety of genomic analyses, including ChIP-seq. It is the product of single-molecule datasets from PacBio, Oxford Nanopore, and BioNano Genomics. Furthermore, it was polished with high quality Illumina short reads and annotated with the incorporation of mRNA-seq data across multiple life stages from both sexes. As an example of its utility we show DNA puff II/9A in its genomic context for the first time where it resides on a multi-megabase contig. Moreover, researchers recently demonstrated evidence of pervasive cytosine methylation in the Sciara genome by immunofluorescence using an antibody against 5-methylcytosine. We were able to weigh in on these observations from two angles. The signal level of both PacBio and MinION data present opportunities to test the possibility of DNA modifications. Both technologies strongly support the presence of base modifications throughout the Sciara genome. The kinetics variations in the SMRT sequencing data identified 5mC, 4mC, and 6mA, but mostly it identified signatures in the kinetics data that it could not assign to those three modifications. A little over 7% of the CpG dinucleotides in the genome appear to be methylated. In general, the SMRT data gave rise to CpG dinucletode and poly-A motifs. Analyses on the nanopore sequencing signal also identified CpG and poly-A motifs enriched in 6mers that had different ionic current levels than expected. It will be interesting to explore the context and function of these modifications, as well as the fraction of molecules that were modified at each site. Given the prevalence of base modifications, it seems like a promising avenue for the study of imprinting and chromosome elimination in Sciara coprophila.

3.5 Epilogue

This work is ongoing. The published version will also include the annotation of the final genome assembly, which will be supplemented with RNA-seq data from the transcriptomes of embryos, larvae, pupae, and adults for both sexes. The annotation phase will also shed light on the repeat content of the genome. We will also more extensively explore the PacBio and MinION data to weigh in on DNA modification in the genome. This chapter included the early results of those investigations. For future publications, we have begun using the genome sequence to interrogate

interesting facets of Sciara biology. For example, one future direction involves identifying where the "controlling element" is within the genome [Crouse, 1960a, Gerbi, 1986]. The controlling element is involved in the regulation of X chromosome elimination. It was shown to reside among an array of rDNA repeats [Crouse et al., 1977]. It is currently unknown how long of a sequence interrupts the rDNA repeats on the X-chromosome. I have identified candidate contigs in the assembly with rDNA repeats. Moreover, in collaboration with Jack Bateman, we have also used the MinION to sequence a cosmid clone thought to contain sequence either corresponding to the controlling element or nearby it. I assembled that data to create a consensus sequence for the cosmid to align to the genome. I also tried simply aligning the MinION reads to the genome to see what region becomes enriched. Both approaches identify the same area of the genome, which was also one of the candidates identified from search for rDNA sequence. It will be interesting to explore this region of the genome further. For example, what genes, if any, may reside at or near the controlling element? The germ-line limited, somatically-eliminated L chromosomes are also of interest and we will look further into candidate L contigs as well. Candidate L contigs are expected to have little coverage from our embryonic data and no coverage from other tissues. We have Illumina data from salivary gland genomic DNA to help select candidates. Can we begin to unravel the mystery of the L chromosomes? Do they have a function? Do they have genes? Differentiating X chromosome sequence from X' sequence is also of interest to us. Male-producing females have two copies of the X whereas female-producing females have 1 X and 1 X'. The X' chromosome therefore seems to be involved in determining how many X chromosomes are eliminated from embryonic nuclei and ultimately the sex of the offspring. To our knowledge, the X' differs from the X only by a long paracentric inversion. Thus, identifying the differences between the two turns into a structural variation problem. We have paired-end Illumina data from mixed females. This data can be compared to the male genome sequence that only contains sequence from the X chromosome with many established SV algorithms for this data type [Koboldt et al., 2012]. We will collect MinION data using our ultra-long read protocols for androgenic and gynogenic females separately to perform comparative long read structural variation analyses using Sniffles as performed in this manuscript. Finally, we have been collaborating with NabSys who produce genome-wide maps similar to BioNano. They have been developing structural variation algorithms for their data and are interested in weighing in on this problem. Once the breakpoints for the paracentric inversion are confidently identified we can explore, for example, whether or not there are nearby genes and whether their expression becomes silenced or enhanced by the inversion. Perhaps a new fusion gene is made or a gene is destroyed. We will also use the NabSys data from our collaboration to update the genome assembly. In particular, some algorithms are able to use multiple genome-wide map datasets. Therefore, we can attempt to scaffold the Sciara assemblies with data from both NabSys and BioNano at the same time. This is expected to both increase the contiguity, perhaps dramatically with enough maps from different enzymes, as well as reduce erroneous scaffolding events. Finally, we plan to further explore the general transcriptome data that we have to begin to untangle Sciara's sex determination pathway.

3.6 Methods

Embryo Collection

Sciara is monogenic: the females produce either only male or only female offspring. Male-producing females have 2 copies of the X chromosome while female-producing females have an X and an X'. The two types of females can be differentiated when they are adult flies based on the wavy wing phenotypic marker that is associated with the X': female-producers have wavy wings, whereas male-producers have normal straight wings. However, the different females cannot be distinguished as embryos, our primary target life stage for sequencing. Therefore, we mated strictly with male-producers for embryo collection to avoid sequencing the X' that can complicate the assembly. Mass matings consisted of 6 female and 4 male flies per vial. One day after combining the flies for mating, females were separated from males. Embryo laying was induced by squishing the thorax with forceps and plating on 2.2% bactoagar (2.2g per 100 mL). After 2-4 hours, adults were removed from the plate and embryos were transferred to an antibiotic/antimycotic plate, where they incubated at 20-21°C for up to 2 days, unless early (2-4 hour embryos) were being collected. Prior to DNA extraction, embryos were washed in TE (10 mM Tris-HCl pH 8, 1 mM EDTA) serially 10 times to remove external contamination.

DNA extraction

Genomic DNA (gDNA) was isolated from Sciara using DNAzol (ThermoFisher) and following the manufacturers instructions with some modifications. Either a glass dounce homogenizer or a blue pestle was used each with 10 strokes for the homogenization step. Prior to precipitating the DNA, RNase A was added to the DNA sample and incubated at 37°C for 10 minutes, followed by Proteinase K, incubated at 37°C for 10 minutes. After adding 100% ethanol, the tube was slowly inverted 50 times, incubated at room temperature for 2 minutes followed by ice for 2 minutes, then centrifuged at 18000g for 10 minutes. The supernatant was removed. The pellet washed twice with 75% ethanol, very briefly air-dried, and re-suspended in TE. The gDNA was cleaned with AMPure beads (Beckman Coulter) until NanoDrop reported purity levels of A60/280 $>$ 1.8 and A260/230 of ~2.0. It was re-suspended in Tris-HCl (pH 8.0) before beginning library preparations.

Illumina library

gDNA from mixed stage embryos (2 hour – 2 day old) was sonicated to a size range of 100-600 bp, and Illumina libraries were prepared using the NEBNext kit (New England Biolabs) following the manufacturers directions. The library was run on a 2% NuSieve agarose (Lonza) gel, size-selected near the 500 bp marker, gel purified (Qiagen), and sequenced on the Illumina HiSeq 2000 platform to obtain 100 bp paired-end reads.

PacBio sequencing

Genomic DNA was brought to the Technology Development Group at the Institute of Genomics & Multiscale Biology at the Icahn School of Medicine at Mount Sinai for PacBio library construction and sequencing. Two DNA libraries were prepared and sequenced according to the manufacturers instructions and reflects the P5-C3 sequencing enzyme and chemistry, respectively. Details are in supplementary methods.

MinION sequencing

In total, 17 libraries were prepared over the course of 18 months spanning a few iterations of reagent, flow cell, software, and MinION upgrades. The libraries are named/numbered 01-09, 11-16, 20-21. Library 01 used SQK-MAP002 reagents. Libraries 02-08 used SQK-MAP004 reagents. Libraries 09, 11, 12, and 13 used SQK-MAP005 reagents. All aforementioned libraries used the original MinION and the R7.3 pore model. Libraries 14, 15, 16, 20, and 21 used SQK-MAP006 reagents, the MinION MkI, and the R7.3 70 bps 6mer model. Libraries 04, 05, 09, and 11 did not perform well due to flow cells with very limited numbers of pores available for sequencing and consequentially gave little data, but we included it anyway. Library 11 was attempted on two different flow cells that both started with very few available pores. Moreover, though fragmentation was not performed, the DNA was lower molecular weight than expected as viewed on an agarose gel. Libraries 14, 20, and 21 were constructed following Oxford Nanopore's satndard protocol. Libraries 02-09, 11-13, and 15-16 were constructed using various iterations of our ultra-long reads protocols. Importantly, libraries 07, 08, 09 (poor flow cell), and 15 included rinses in the clean-up steps [Urban et al., 2015a] whereas the other libraries did not. Moreover, libraries 06, 07, 08, 09, 12, 13 and 14 were all prepared from the same DNA source to be able to directly compare size distributions after different protocols. Similarly, libraries 15 and 16 were prepared from the same source to make direct comparisons of the effects of including rinse steps in the AMPure clean-ups. Library 01 was an early attempt to obtain longer reads by simply skipping the shearing step, but the genomic DNA was not handled gently, wide-bored pipettes were not used, there was no DNA repair step, and manipulations to AMPure bead steps were not performed. Therefore, we do not include this library in either of the standard protocol or longer read protocol categories.

Library preparations varied through development of long read protocols and with the evolution of the technology and Oxford Nanopore's protocols. For standard protocol libraries (014, 20, 21), we followed Oxford Nanopore's protocol from start to end for SQK-MAP006 reagents and sequencing on theMinION MkI. The only libraries that included DNA shearing were these.

For libraries prepared with modified protocols (01-09, 11-13, 15-16), shearing was always skipped and we started out with 2.2–17 μg, using the lower end for early libraries (2.2–5 μg in 01-13) and the higher amounts more recent libraries (15-17 μg in 15-16). All libraries except 01 were subject

to DNA repair steps. PreCR (NEB) was performed on earlier libraries (02-09,11-13) and FFPE Repair (NEB) was performed on more recent libraries (14-16, 20-21). For most early libraries prior to MAP006 (excluding 01 and 03), End-Repair and dA-tailing were done in separate steps using the NEBNext End-Repair Module (NEB) and the NEBNext dA-Tailing Module (NEB). For the 01 and 03, these steps were combined using the NEBNext Ultra End Repair/dA-tailing module (NEB). For all libraries using MAP006 MinION reagents, end repair and dA-tailing were performed as a single step using the NEBNext Ultra II End Repair/dA-tailing module (NEB). Ligation was always carried out using MinION kit-specific adapters and volumes with Blunt/TA Ligase Master Mix (NEB). We only needed to manually add the tether and motor proteins for our earliest MinION library (01). In all others, the tether and motor proteins are combined with other reagents such as the adapters and elution buffer. The first library did not have a hairpin enrichment step. All other libraries did, concurrent with MinION kits and protocols. Seven libraries (02-08) used His-beads with MAP004 reagents and four libraries (09, 11-13) used His-beads with MAP005 reagents according to Oxford Nanopore's instructions. The five MAP006 libraries (14-16, 20-21) used MyOne C1 streptavidin beads (Dynabeads/Thermofisher) for hairpin enrichment following Oxford Nanopore's protocol. In AMPure bead steps, a 0.4x ratio was typically used (details in Supplmentary methods). For some libraries (07, 08, 09, 15), a novel rinse step was added in AMPure clean-ups as we described in our preprint [Urban et al., 2015a] and demonstrate the reproducibility of here. For all AMPure beads steps for our protocols, DNA was eluted off the beads by incubating for 10–20 minutes at 37–50°C (details in Supplement). For early libraries we did 37°C for 20 minutes and in more recent libraries we were doing higher temperatures for short periods of time. Wide-bored tips and gentle pipetting were used throughout unless specified otherwise in the supplement. Sequencing was conducted following standard procedures. For early libraries, smaller amounts of DNA were loaded more frequently (e.g. 4 times) throughout the run. For more recent libraries, half of the library was loaded at the beginning and the second half was added 24 hours in.

BioNano Irys data

Many attempts were made for obtaining plugs with high molecular weight genomic DNA from male embryos. However, the plugs either had too little gDNA or it was too fragmented. We obtained successful plugs in our first attempt at obtaining plugs from male pupae. To isolate ultra high molecular weight DNA, pupae were flash frozen and ground in liquid nitrogen. The powder was resuspended in Nuclear Isolation Buffer (NIB)(10 mM Tris pH 9.4, 60 mM NaCl, 10 mM EDTA, 0.15 mM spermidine, 0.15 mM spermine 0.5% Triton-X 100, 1% beta-mercaptoethanol) on ice and filtered through 100 μM mesh cell strainer. Cellular debris was removed by sedimentation for 15 sec at 1K rpm, 4°C. Nuclei were recovered from the supernatant by sedimentation for 3min at 1800xg, 4°C. Nuclei washed 3x with NIB, were resuspended in Cell Suspension Buffer (Chef Mammalian Genomic DNA plug kit, Bio Rad) and embedded in low melt agarose, final concentration 0.8%. Nuclei were lysed in Alternate Lysis Buffer (BioNano Genomics) containing1.6 mg Proteinase K (Qiagen)

at 50°C for 24 hr, and RNAs degraded by addition of RNAse A (80 μg, Qiagen). Plugs were liquefied using Gelase and the resulting high molecular weight DNA was membrane dialyzed against 10mM Tris HCL, 1mM EDTA, pH 8.0). DNA was harvested with wide bore tips and stored at 4°C.

High Molecular Weight (HMW) DNA was nicked, labeled and repaired according to the IrysPrep protocol (BioNano Genomics). In brief, HMW DNA was digested with the single-stranded nicking endonuclease BssSI (CACGAG, NewEngland BioLabs). Fluorescently labeled nucleotides were incorporated by nick translation. The backbone of the labeled DNA was stained with YOYO-1. Individual labeled DNA molecules were imaged on the Irys platform (BioNano Genomics). Optical maps of the genome were assembled using an overlap layout consensus method. A p-value threshold of 2.6e-9 was used with BioNano Pipeline Version 2884 and RefAligner Version 2816 (BioNano Genomics).

The BNG optical maps were used with hybridScaffold.pl version 4741 (BioNano Genomics) to link contigs in the PacBio sequence assemblies. BssSI restriction maps of PacBio contigs were generated in silico. Consensus Maps (CMAP) were only created for scaffolds $>$ 20 kbp that contained $>$ 5 BssSI sites. A p-value of 1e-10 was used as a minimum confidence value to output initial alignments (BNG CMAP to in silico CMAP) and final alignments (in silico CMAP to final hybrid CMAP). A p-value of 1e-13 was used as minimum confidence value to flag chimeric/conflicting alignments and to merge alignments. The final assemblies include all super-scaffolds from hybridScaffold.pl and those that were not super-scaffolded.

Strand-specific RNA-seq

Total RNA from male and female embryos, larvae, pupae, and adult flies was extracted using TRIzol (Invitrogen/Thermofisher). RNA quantity and purity were measured with the NanoDrop (ThermoScientific) and Qubit (ThermoFisher). Total RNA was treated with DNase (Qiagen) and subject to RNeasy column clean up (Qiagen). The cleaned total RNA quantity and purity were checked using the NanoDrop and Qubit. RNA integrity was evaluated on 1.1% formaldehyde 1.2% agarose gels. Poly-A RNA was enriched using Oligo-dT DynaBeads (LifeTechnologies). Qubit was used to measure the quantity of poly-A RNA. Poly-A RNA was fragmented with NEB's Magnesium Fragmentation Module for 3 minutes at 94°C, which was selected after optimizing for conditions for 200-500 bp fragments as determined by the Bioanalyzer. Fragmentation reactions were cleaned up with RNeasy columns. First strand synthesis was performed with SSIII (Invitrogen). Briefly, 1 μl Random Primer (3 μg/μl), 1 μ 10 mM dNTP mix, and 10 μl fragmented RNA were incubated at 65°C for 5–10 minutes and transferred to ice for 5 minutes. A mix of 4 μl 5X First Strand Synthesis (FSS) Buffer, 1 μl 0.1 M DTT, 1 μl Murine RNase Inhibitor, 1 μl 0.5 μg/μl Actinomycin D, and 1 μl

SSIII (200 units) was added to the mixture of RNA, dNTPs, and Random Primers. This was incubated in the thermocycler at 25°C for 5 minutes to anneal the random primers, 50°C for 60 minutes to extend from the random primers, and 70°C for 15 minutes to inactivate SSIII. The reaction was cleaned up with AMPure beads using a ratio of 2.0. For Second Strand Synthesis (SSS), a mixture of 10 mM each of dATP, dCTP, dGTP, and 20 mM of dUTP (instead of dTTP) was made. For a single reaction, 64 μl of cleaned FSS cDNA:RNA in Ultra Pure Water, was combined with 4 μl ACGU mix, 8 μl of NEB dNTP-free SSS Reaction buffer and 4 μl NEB SSS Enzyme mix from the SSS module. The reaction was incubated for 1 hour at 16°C, then cleaned with AMPure beads using a 1.0x ratio to begin eliminating DNA smaller than 200 bp, and quantified with the Qubit. There was typically 100-200 ng at this step. The double-stranded cDNA was then subject to End Repair (NEBNext), and cleaned with AMPure beads using a 0.9x ratio to deplete DNA $<$ 300 bp. The End-Repaired cDNA was then dA-tailed (NEBNext), and cleaned with AMPure beads using a 0.9x ratio. For adapter ligation, 38 μl of DNA was combined with 10 μl 5X NEBNext Quick Ligation Reaction Buffer, 1 μl NEB Adaptor, and 2 μl Quick T4 Ligase. The reaction was incubated at room temperature for 15 minutes, then cleaned with a 0.9x ratio of AMPure beads. The library was then size-selected with AMPure beads before proceeding to the PCR step. To obtain adapter-ligated fragments in the 300-600 bp range, the DNA was first incubated with 0.6x AMPure beads by adding 60 μl AMPure beads to 100 μl DNA in UPW. The beads were pelleted on a magnet and the supernatant containing DNA smaller than approximately 600 bp was transferred to a new tube (DNA longer than 600 bp stayed on the beads). To make the final ratio 0.9x to select for DNA $>$ 300 bp on the beads, 30 μl more AMPure beads was added to the 160 μl supernatant. From there the AMPure clean-up proceeded as normal. USER enzyme digestion to cut the DNA at uracils (in the second strand and in the hairpin adapters) and PCR were then performed as follows: 20 μl of cleaned DNA was combined with 25 μl NEBNext High-Fidelity 2X PCR Master Mix and 3 μl NEBNext USER enzyme. This reaction was incubated for 15 minutes at 37C to ensure uracil cutting occurs before addition of primers. Then 1 μl indexed primer (NEBNext) and 1 μl universal primer were added. The reaction was put in the thermocycler for 37C for 15 minutes to ensure USER digestion went to completion, followed by 98C for 30 seconds and 12 cycles of: 98C for 10 seconds, 65C for 30 secons, 72C for 30 seconds. The PCR products at this stage are approximately 122 bp longer than the target insert size. We adjusted AMPure ratios accordingly for a final clean up and size-selection. The PCR reactions were cleaned with 0.85x AMPure beads to deplete DNA smaller than 350 bp. The DNA was eluted in 100 μl UPW and AMPure size selection was initiated by incubating with 55 μl AMPure beads (0.55x) to precipitate DNA longer than 650 bp onto the beads. The beads were pelleted on a magnet and the 155 ul of DNA shorter than 650 bp in the supernatant was transferred to a new tube where another 30 μl of AMPure beads was added for a final ratio of 0.85x. The AMPure procedure then continued as normal to obtain DNA $>$ 350 bp. The estimated insert sizes at this step was 230-530 bp. DNA samples were quantified with Qubit and purity was measure by NanoDrop. There was typically 600 ng at the end of this protocol. Bioanalyzer traces suggested the mean estimated fragment sizes was around 420 bp putting the mean insert sizes near 300 bp.

Supplementary Materials

Supplementary Figures, Tables, and Methods are available in the Appendix at the end of this thesis.

Data deposition

Illumina, PacBio, MinION, and BioNano data will be available on NCBI SRA.

Funding

This work was supported by National Science Foundation predoctoral fellowships to JMU: NSF GRFP DGE-1058262 and NSF RI EPSCoR Grant# 1004057. NSF EPSCoR also provided extended access to computational resources necessary to support this project in a timely manner.

Conflict of Interest

JMU and SAG have been members of the MinION Access Program since 2014 and have received free reagents from ONT over the time frame of this study. JMU is also a member of the MinION Access and Reference Consortium (MARC), an international group that conducts experiments partially funded by ONT.

Acknowledgements

We would like to thank Oxford Nanopore for granting us early access to the MinION and for strong continual support, particularly from Michael Micorescu, Sissel Juul, Daniel Turner, Stuart Reid, David Stoddart, Margherita Coccia, Richard Ronan, and Jackie Evans. We would also like to thank the members of the nanopore community, and particularly members of the MinION Access and Reference Consortium (MARC), for lively discusions. Thanks to Benjamin Raphael for helpful discussions on genome assembly and the Center for Computational Molecular Biology (CCMB) at Brown University for providing computing resources for our MinION. We would like to thank the Technology Development Group at the Institute of Genomics & Multiscale Biology at the Icahn School of Medicine at Mount Sinai, particularly Gintaras Deikus for help in obtaining PacBio data, and Ali Bashir and Robert Sebra for discussions on PacBio chemistries, genome assembly, and example assemblies from HGAP3 and an early version of Falcon (though neither is featured in this manuscript). Thanks to Adam Phillippy, Sergey Koren, and Brian Walenz for plenty of helpful correspondence and guidance about Canu and genome assembly in general, and for incorporating

our datasets into weekly Canu assembly regression testing (assemblies from regression testing are not featured in this manuscript and were only used by them, amongst many other assemblies, to ensure updates to Canu maintained similar assembly statistics). Thanks to Mark Howison, Stefano Lonardi and Stephen Richards for helpful correspondence and sharing experiences with various assemblers and tools. Thanks to Jared Simpson and Arthur Rand for providing guidance on Nanopolish and SignalAlign, respectively. Thanks to the Center for Computation and Visualization (CCV) at Brown University for computational resources and support. Many thanks to the NSF RI EPSCoR committee who granted us access to a massive amount of computing resources to carry out this project.

CHAPTER 4

The DNA puffs of Sciara coprophila before, during, and after developmentally programmed intrachromosomal DNA amplification

John M. Urban, Yutaka Yamamoto, Leo Kadota, Audrey Lee, Benjamin Doughty, Julia Leung, Jacob E. Bliss, Heidi S. Smith, Susan M. Dibartolomeis, Susan A. Gerbi

[1] Brown University Division of Biology and Medicine, Department of Molecular Biology, Cell Biology and Biochemistry, Providence, Rhode Island 02912, USA

Corresponding authors: Susan_Gerbi@Brown.edu and John_Urban@Brown.edu

This chapter represents ongoing work for a manuscript expected to be submitted in Spring 2017. I did all RNA and DNA work and library preparations for Illumina sequencing of the salivary gland genome and transcriptome over 5 developmental stages. I wrote the pufferfish program designed to find DNA 'puffs for FISH' analysis, which uses a hidden Markov model approach to identify regions with increased copy number, and applied it to map DNA puffs in the Sciara genome. I performed all bioinformatic and genomic analyses including those on the height, width, and timing of DNA puff amplification and inferred fork speed, DNA puff comparisons, and RNA-seq assembly, differential expression, expression profile over time, etc. I designed and validated the 24 qPCR primer pairs used to confirm DNA puff copy number and for the ecdysone inject experiments. Audrey Lee performed qPCR of late stage DNA puffs and control regions under my guidance to validate sequencing results. Leo Kadota independently optimized the ecdysone injection protocol for larvae. He performed qPCR of ecdysone- and mock- injected larvae under my guidance to test responsiveness of newly identified puffs to ecdysone compared to control regions. I performed all qPCR analyses and visualization. Yutaka Yamamoto and Julia Leung performed FISH experiments with probes designed from DNA puff sequences that I provided. Leo Kadota, Audrey Lee, and Julia Leung helped optimize FISH. With my guidance, Benjamin Doughty did validations on previously unknown EcR isoforms that I identified in the salivary gland transcriptome. Heidi Smith, Susan Dibartolomeis, and Yutaka Yamamoto contributed to cloning and sequencing genomic and cDNA inserts that hybridized to puff II/2B and was used as a positive control in addition to II/9A for my DNA puff analyses. I wrote the manuscript.

4.1 Abstract

Eukaryotic DNA replication is normally regulated to ensure each region of the genome is duplicated just once such that the ratio between all loci stays constant. This regulation is enforced at DNA replication initiation sites, called origins, of which there are many in a given genome. An origin can only be used once per cell cycle. Loss of this regulation can result in re-replication and extra copies of DNA where this occurs, which has been shown to lead to gene amplification, DNA damage, and genomic instability. The controls against re-replication and the consequences when the controls are knocked out have been well-studied. An open question is, how can locus-specific re-replication occur when the controls are intact? This question is difficult to address in most systems because the controls against re-replication successfully make it a rare event. In contrast, dipteran flies present examples where DNA re-replication is used for gene amplification as a normal part of development. Importantly, these rogue amplification origins repeatedly initiate DNA replication when the majority of origins in the genome are prevented from doing so. One such model is the intrachromosomal DNA amplification that occurs in the 18 "DNA puffs" of late larval salivary gland polytene chromosomes in the fungus fly, Sciara coprophila. Though all DNA puffs have been studied cytologically, only DNA puff II/9A has been sequenced. The re-replication origin in DNA puff II/9A has been well-characterized, and the steroid hormone ecdysone plays a role in inducing amplification of this locus. A comparison of the DNA puff II/9A sequence to the sequences of other DNA puffs would allow a search for shared regulatory elements as well as an exploration of ecdysone's role in DNA re-replication more generally. Recently, we released the first draft genome sequence for Sciara coprophila, which is highly contiguous and valuable for genome-wide studies. We now use the reference to map the sequences of the DNA puffs, and explore their developmental timing patterns, their final copy numbers, the distance re-replication forks travel, and whether ecdysone promotes premature DNA amplification at each.

4.2 Introduction

Replication of the genome is carefully controlled by dividing eukaryotic cells to ensure that the ratio amongst all genomic loci remains constant. Although replication initiates from thousands of origins in a given genome, each is strictly regulated to "fire" at most once per cell cycle by restricting origin 'licensing' to G1 and origin activation to S-phase. Origins can re-fire in the same S-phase if this regulation is lost in the cell or overridden at a specific locus. When an origin re-fires, a new replication bubble forms nested inside the previous one with new replication forks chasing behind the previous forks. This chase condition can result in head-to-tail collisions and DNA fragmentation if the pursuant fork overtakes the one in front of it [Davidson et al., 2006]. The extra DNA replication from an origin re-firing is called re-replication, which has been shown in yeast to lead to replication fork break down [Finn and Li, 2013], blocked cell proliferation [Green and Li, 2005], extensive DNA damage [Green and Li, 2005], non-allelic homologous recombination [Green et al., 2010], stable gene amplification [Green et al., 2010], other copy number changes and genomic re-arrangements [Green

et al., 2010, Brewer et al., 2011, Finn and Li, 2013, Brewer et al., 2015], as well as chromosome instability and aneuploidy [Hanlon and Li, 2015]. Moreover, DNA damage, double-strand breaks, crossover structures from homologous recombination, and large rearrangements have been detected during DNA re-replication in flies [Heck and Spradling, 1990, Liang et al., 1993, Yarosh and Spradling, 2014, Alexander et al., 2015]. Overall, re-replication is associated with genomic instability in human cells and cancer progression [Blow and Gillespie, 2008, Diffley, 2010]. Moreover, DNA replication stress and oncogene amplification are hallmarks of cancers [Blow and Gillespie, 2008, Negrini et al., 2010, Hanahan and Weinberg, 2011, Macheret and Halazonetis, 2015]. The linear distribution of the extra copies of oncogenes found in cancer cells differs from the parallel nature of the extra copies of DNA in nested re-replication forks. However, the aforementioned consequences of re-replication strongly suggest that re-replication structures can initiate the cascade of events that lead to oncogene amplification [de Cicco and Spradling, 1984]. Therefore, it is crucial to elucidate: (i) how re-replication is normally prevented by the eukaryotic cell, (ii) the consequences of re-replication when normal regulation is perturbed, (iii) how re-replication can happen when the normal regulation is intact, and (iv) how re-replication can be directed to specific sites. How re-replication is normally prevented and the consequences when its not have been well-studied [Green and Li, 2005, Blow and Gillespie, 2008, Diffley, 2010, Green et al., 2010, Brewer et al., 2011, Finn and Li, 2013, Brewer et al., 2015, Hanlon and Li, 2015]. The latter two questions are still open, and there is likely no single answer. They are more difficult to study in most systems where the normal controls successfully decrease re-replication rates to nearly undetectable levels. Moreover, how re-replication can occur site-specifically needs to be carefully distinguished from scenarios that demonstrate apparent specificity. For example, a low basal rate of non-specific re-replication in a cell population can result in specific, reproducible copy number increases in their descendants by conferring a selective advantage under test conditions, such as drug selection or nutrient deprivation. Though specific loci are amplified, the underlying process is not itself specific. Fortunately, dipteran flies demonstrate multiple rounds of locus-specific re-replication as a part of normal development. This results in a hierarchical structure of nested replication forks, called an onion-skin structure, and is an example of intrachromosomal DNA amplification as opposed to extrachromosomal amplification from repeated origin firing on episomes or rolling circle amplification [Claycomb and Orr-Weaver, 2005]. A famous example of intrachromosomal DNA amplification is that of the chorion genes in the follicle cells of Drosophila melanogaster [Spradling and Mahowald, 1980]. The first example of this ever observed, though, was in the DNA puffs of the giant polytene chromosomes in the late larval salivary glands of Sciarid flies [Poulson and Metz, 1938, Breuer and Pavan, 1955]. We study the 18 DNA puffs in the fungus fly, Sciara coprophila.

The DNA puff structures in Sciarid polytene chromosomes were first observed by Poulson and Metz in 1938 [Poulson and Metz, 1938]. In the two decades spannning 1950-1970, it was observed that Sciarid DNA puffs were sites of disproportionate DNA synthesis and contained “extra DNA” [Breuer and Pavan, 1955, Ficq and Pavan, 1957, Rudkin and Corlette, 1957, Swift, 1962, Gabrusewycz-Garica,

1964, Crouse and Keyl, 1968, Rasch, 1970b, Rasch, 1970a]. These were controversial observations at the time since they violated the proposed "rule of DNA constancy" where the ratio of all DNA in a cell was expected to stay constant [Gerbi and Urnov, 1996]. In Sciara, DNA amplification is preceded by up to 13 rounds of endoreplication, where the entire genome is replicated without intervening cell divisions. This results in polytene chromosomes containing up to 8,192 copies held in close register. The DNA puffs then continue with additional rounds of locus-specific re-replication. Starting in 1989, the molecular nature of one DNA puff called II/9A began to be dissected and is well-characterized. Starting with cDNA clones that hybridized with puff II/9A, a lambda clone containing a genomic DNA insert was identified and confirmed to map to II/9A [DiBartolomeis and Gerbi, 1989]. Two genes (II/9-1 and II/9-2) were identified therein, which seem to encode proteins for the pupal coat that are needed in abundance in a short period of time [DiBartolomeis and Gerbi, 1989]. The Sciara II/9A amplification origin has been mapped by 2D gels and 3D gels to a 1 kb region [Liang et al., 1993, Liang and Gerbi, 1994]. These studies also demonstrated that there is only one replication bubble per DNA fragment and that replication is bi-directional. PCR analysis of short nascent DNA abundance demonstrated that the amplification origin is within the initiation zone used for endoreplication and normal DNA replication in mitotic cells [Lunyak et al., 2002]. A specific binding site for the Origin Recognition Complex (ORC) was identified and the exact transition point from continuous to discontinuous DNA replication was found to be directly adjacent to the ORC binding site [Bielinsky et al., 2001]. A DNase I hypersensitive site (DH-1) of 400 bp is located 600 bp upstream of the II/9A amplification origin [Urnov et al., 2002]. DH-1 has an approximately 100 bp sequence conserved with DNA puff C3 of the related species, Rhynchosciara. Chromatin decondensation for DNA puffing correlates with the burst of transcription at II/9A, both of which begin after DNA amplification [Wu et al., 1993], but the maintenance of puff expansion is independent of active DNA and RNA synthesis [Mok et al., 2001]. It was demonstrated that the morphological puffing as well as DNA and RNA synthesis inside the DNA puffs were stimulated by ecdysone injection [Crouse, 1968]. Ecdysone regulates transcription from the gene II/9-1 promoter and induces premature DNA amplification of II/9A up to normal levels [Foulk et al., 2006]. Two Sciara Ecydysone Receptor (EcR) isoforms (EcR-A and EcR-B) were cloned and antibodies against the N-terminal domain specific for each were prepared [Foulk et al., 2013]. EcR-A was shown to be the predominant form in salivary glands with an increased presence during DNA amplification stages [Foulk et al., 2013]. Interestingly, an Ecdysone Response Element (EcRE) was identified adjacent to the ORC binding site, which binds the Ecdysone Receptor (EcR) in vitro [Foulk et al., 2006]. EcR-A was shown to bind DNA puff II/9A in polytene chromosomes during DNA amplification [Liew et al., 2013]. Overall, these results suggest that ecdysone may play a direct role in DNA amplification at locus II/9A.

How far the re-replication forks travel from amplification origins in Sciara remained uncharacterized, though they may travel up to or over 100 kb [Claycomb and Orr-Weaver, 2005], indicating that DNA amplification in Sciara may also provide an opportunity to study elongation. Moreover,

although there has been a lot of progress made in studying the re-replication origin at II/9A, there have been no molecular studies on the other 17 DNA puffs. Elucidation of re-replication in Sciara has been generally inhibited by (i) a lack of transgenic techniques to enable thorough genetic manipulation, (ii) a lack of a genome sequence to identify relevant genes, and (iii) no sequence information from other DNA puffs for comparison. Recently, we published the first example of gene insertion in Sciara [Yamamoto et al., 2015] as well as the first draft genome sequence for Sciara coprophila [Chapter 3]. Now we present the sequences and locations of the DNA puffs within the genome assembly. The high contiguity of our assembly enabled us to identify at least 14 DNA puff sequences that we validated through qPCR and fluorescent in situ hybridization (FISH). The long contigs in the assembly, reaching megabases in length, allowed determination of how far re-replication forks travel, of developmental timing of amplification in each DNA puff, and of the genes inside each amplicon. Obtaining the sequences for numerous DNA puff loci let us search for common motifs shared between the DNA puffs that may be important to explore in future genetic studies. Finally, we were able to test the hypothesis that ecdysone induces premature DNA amplification at each of these puffs.

4.3 Results

4.3.1 The sequence of DNA puff II/2B

DNA puff II/9A is the largest DNA puff and has been the focus of many of our studies. DNA puff II/2B is the second largest DNA puff and has been shown to amplify 16-fold [Crouse and Keyl, 1968, Wu et al., 1993]. cDNA clones were obtained for DNA puff II/2B and used in previous studies to identify lambda clones with complementary genomic DNA inserts and to study II/2B by in situ hybridization and Southern blots [Wu et al., 1993]. However, the sequence to II/2B has been unknown. We have Sanger-sequenced the cDNA and genomic DNA inserts that hybridize with DNA puff II/2B. Therefore, in addition to DNA puff II/9A, DNA puff II/2B was used as a positive control in identifying DNA puffs de novo with high throughput sequencing, discussed in the next section.

4.3.2 Mapping sites of DNA amplification using high throughput sequencing

To identify the locations of amplified DNA within the reference genome sequence, fourth instar female larvae were first staged according to the eyespots that progressively develop prior to pupation [Gabrusewycz-Garica, 1964,Foulk et al., 2006]. Essentially, the rows and columns of a triangular matrix formed by the eyespots are correlated with amplification and transcription of the DNA puffs in the salivary gland polytene chromosomes. We considered early eyespot stages to be larvae with eyespots up until and including the 8x4 stage. Amplification is thought to begin by stage 10x5, continue through 12x6, and end by 14x7 [Wu et al., 1993, Foulk et al., 2006]. The larvae then proceed into two eyespot stages called Edge Eye and Drop Jaw, where the eyespots migrate from the

top to the sides of the larval head and the jaw parts fall off as the larvae begin molting. Previous studies have found that the largest DNA puff, II/9A, amplifies 16-17 fold by stage 14x7 [Wu et al., 1993,Foulk et al., 2006]. It has been suggested that re-initiation events end by 14x7 [Wu et al., 1993]. Nevertheless, in the same study it was mentioned that tritiated thymidine uptake was seen in the DNA puffs in later stages, though this is attributed to elongation only. Moreover, it was mentioned that 2D gels did not show bubble arcs from the 1 kb origins sequence in EEDJ stages, meaning replication did not initiate from the known origin in these later stages. However, fork arcs were seen in the 2D gels meaning replication forks traveled through the origin. We reasoned that these forks must have come from nearby, else one would need to posit that they came from re-replication origins in another DNA puff likely to be megabases away. Therefore, we chose to sequence larval salivary gland DNA from 10x5, 12x6, 14x7, and a combination of Edge Eye and Drop Jaw stages (EEDJ) to allow for the possibility that a higher copy number would be detected. To control for biases in sequencing and collapsed repeats in the genome assembly, DNA from pre-amplification early eyespot ($\leq$8x4) larval salivary glands was also sequenced. We sequenced an extra replicate of EEDJ to be sure of the final copy numbers of each amplification locus identified. Moreover, we included a sample from larvae that were between 8x4 and 10x5 in case amplification proceeded earlier than expected. If not, then this sample would essentially be an additional pre-amplification sample. For simplicity, this sample will be referred to as 9x4. We Illumina sequenced (50 bp, single end) salivary gland DNA from each stage for a total of 8-12 million reads each.

Amplification sites were identified in stages 9x4, 10x5, 12x6, 14x7, and EEDJ with respect to the pre-amplification stage ($\leq$8x4) in the following way. For all samples, the reads were mapped to the reference genome and the number of reads in 500 bp bins was counted. The number of reads in each bin was internally normalized by dividing it by a read count that was representative of loci with a relative copy number of 1 to obtain the relative copy number of each bin. The bins from later stage samples were then externally normalized to the early stage ($\leq$8x4) bins. This step gives the final relative copy number with respect to the pre-amplification endoreplicated genome and controls for spurious coverage spikes and drops from sequencing and read mapping biases as well as from collapsed repeats in the genome assembly. These final relative copy numbers are called the RCN values. Re-replication proceeds bi-directionally, forming a nested onion skin structure. This structure should be reflected in an RCN gradient of amplification that increases with closer proximity to the origin and decreases moving away from it, forming a giant peak (Fig. 4.1). Each complete round of re-replication effectively doubles the copy number for as far as the replication forks travel. However, the replication forks likely travel at different rates and stop at different sites. Moreover, re-replication rounds and elongation are not perfectly synchronized between the cells in a pair of salivary glands and, further still, we sequenced salivary gland DNA from multiple larvae at each stage. Therefore, one would not expect perfect 2-fold steps of increased coverage, but a continuous distribution. Nonetheless, in addition to mapping the locations of amplicons, we sought to approximate the boundaries of where replication forks from each round of re-replication tend to travel to on

average to approximate the 2-fold steps. To do both simultaneously, we used a hidden Markov model approach that scanned the final normalized RCN values from each stage and segmented the 500 bp bins into one of seven states representing copy numbers of 1, 2, 4, 8, 16, 32, and 64, reasoning that since the highest copy numbers were previously expected to be 16-fold, 64-fold would be a higher copy number than was expected to be seen. After segmenting the genome into these 7 copy number states, we merged bins that were above a copy number of 1 and manually inspected all loci with particular attention to those that spanned more than 50 kb. Most of these sites were determined to be developmental amplicons simply by observing that they progressively increased in copy number and in width across the eyespot stages (Figures 4.2, 4.3, and 4.4). Some however, arose late in 14x7 and EEDJ stages (e.g., Fig. 4.4). Both DNA puffs for which we had known sequence for, II/9A [DiBartolomeis and Gerbi, 1989, Urnov et al., 2002] and II/2B (this study), were identified. Puffs II/9A and II/2B, twelve putative amplicons, and eight control loci (RCN=1) were selected for confirmation with qPCR analysis of EEDJ staged larval salivary glands, which demonstrated strong agreement with the sequencing data (Fig. 4.5 A-B).

Previously, II/9A was reported to begin amplification at eyespot stage 10x5 and end by stage 14x7 with a final amplification level near 16-fold. We found that II/9A begins amplifying earlier than previously appreciated, with up to 2.2-fold amplification detected in stage 9x4 (Figure 4.2). This makes DNA puff II/9A the region of earliest amplification. Moreover, we found that II/9A continues amplifying into the Edge Eye and Drop Jaw stages (EEDJ) that occur after 14x7 and reaches an amplification level around 32-fold (Figure 4.2). The data show that puff II/2B is the second genomic region to begin amplifying (Figure 4.2). II/2B amplification begins after II/9A, but before 10x5 since it has already amplified to 2-fold by this stage. As with II/9A, DNA puff II/2B and other newly identified amplicons continue amplifying into the EEDJ stages. In general, final amplification levels appear to be correlated with how early amplification begins for each locus. The regions that appear to begin amplifying in the middle to late eyespot stages become reach 2-4 fold amplification levels whereas the five regions that appear to begin amplification the earliest reach 4-32 fold (Figures 4.2, 4.3, and 4.4). The amplification gradients for all of the amplicons extend much farther than we anticipated based on *Drosophila* data that show the amplification gradients spanning 75-100 kb [Spradling and Mahowald, 1980, Claycomb et al., 2004, Kim et al., 2011]. In *Sciara*, the amplification gradients span at least 200-600 kb (Figures 4.2, 4.3, and 4.4).

To test if ecdysone induced premature DNA amplification at all amplicons as it is known to do for II/9A [Foulk et al., 2006], pre-amplification stage larvae (< 8x4) were injected with either ecdysone or a mock control. Twenty-four hours later their salivary glands were dissected out for genomic DNA extraction and quantitative PCR for all fourteen amplicons and eight control sites. Ecdysone induced premature amplification at all fourteen amplicons with resulting copy numbers proportional to their copy numbers in normal development (Fig. 4.5 C). In contrast, ecdysone did not induce amplification at any of the control sites. Moreover, the mock injection did not induce

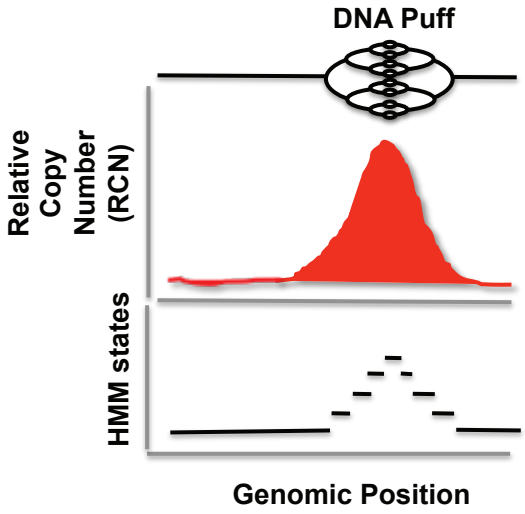


Figure 4.1: The onion-skin structure and expected relative copy number distribution. The top shows the onion-skin structure from intrachromosomal DNA amplification. Re-replication initiates from the origin and forks move away from it bi-directionally, placing the origin in the region of highest copy number. The red shows an example of the relative copy number (RCN) that might be expected given the onion-skin structure. The bottom shows an example of what the Viterbi state path tmight look like given the RCN values. The state boundaries mark areas replication forks tend to travel on average.

premature amplification at any of the sites (Fig. 4.5 D). The small amount of amplification seen at II/9A is consistent with what we found in the sequencing results, namely that it begins amplifying by or before the stage between 8x4 and 10x5 (9x4), which occurs within 24 hours of 8x4. Therefore, some of the older larvae in the batch likely entered into normal DNA amplification. In any case, the amplification level is far lower than that induced by ecdysone injection. The resulting copy number of II/9A from ecdysone injection is in perfect agreement with a previous study [Foulk et al., 2006]. Interestingly, the copy numbers twenty-four hours post-injection are more similar to those seen in stage 14x7 than in EEDJ (Fig. 4.5 E-F).

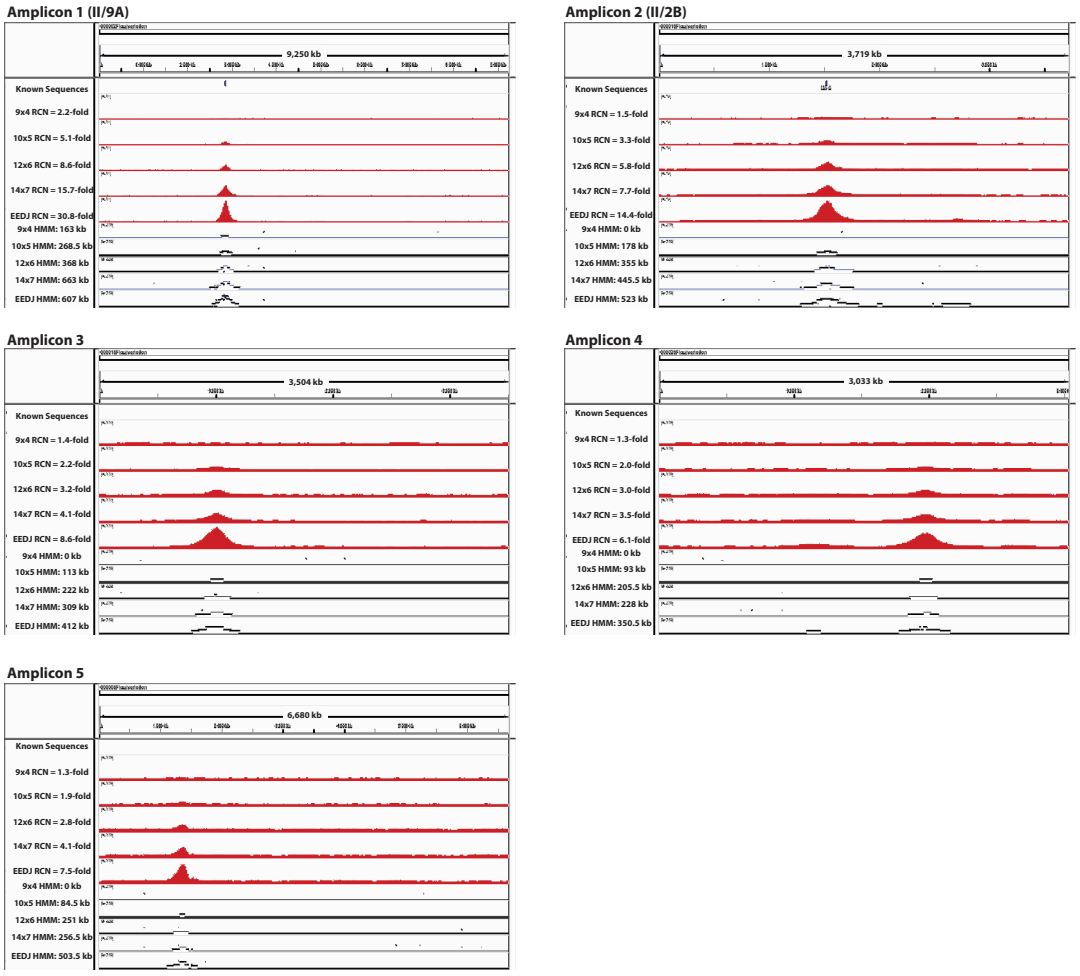


Figure 4.2: The largest and earliest amplicons.

The top part of each panel (red) shows the relative copy number (RCN) across developmental stages as specified on the left. Next to the stage label, the maximum fold-amplification detected in the amplicon is reported. The bottom part of each panel (black) shows the Viterbi state path that segments the amplicons into copy number corresponding to complete doublings. The outside boundary of each step is our estimation of where the replication forks from that round of re-replication tend to travel to on average. The left labels each stage and gives the width of the amplicon in that stage as determined by the boundaries defined by the hidden Markov model.

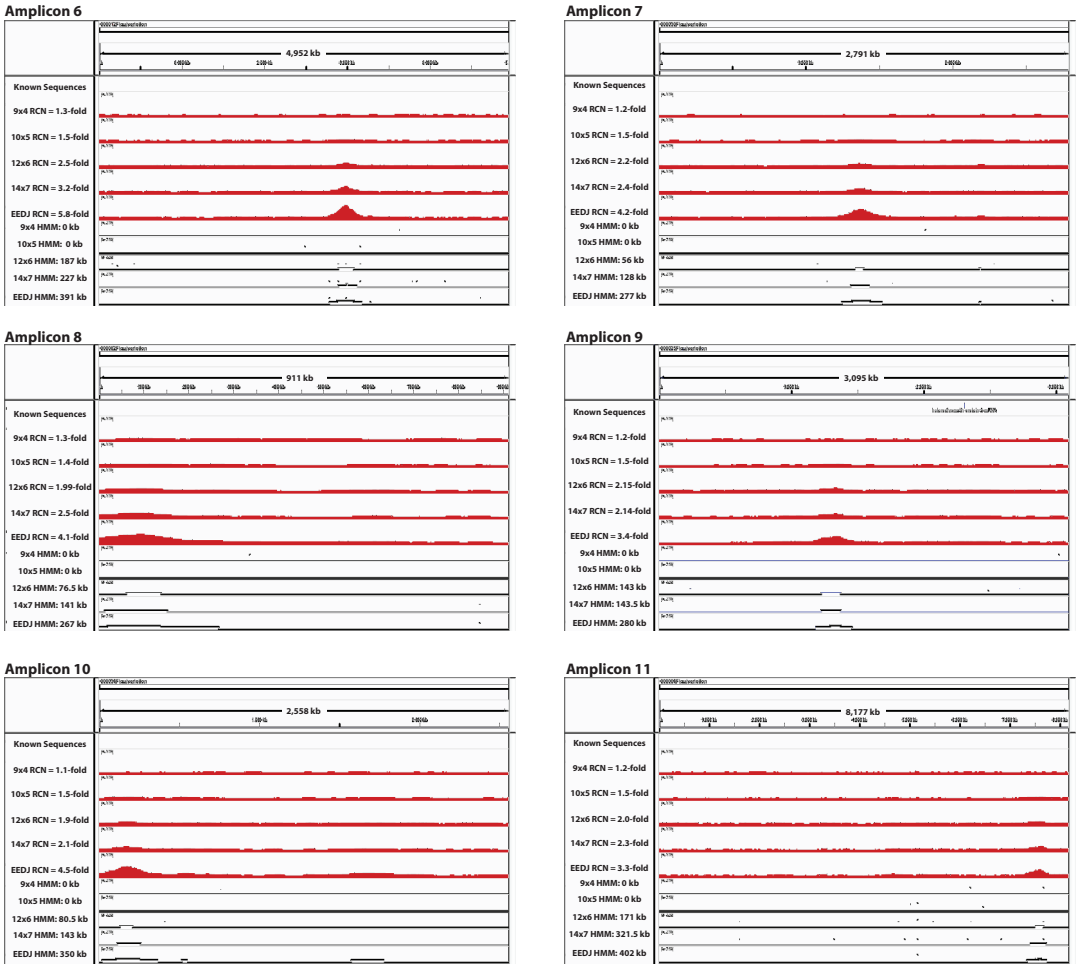


Figure 4.3: The middle rising amplicons.

The panels are as described in figure 4.2.

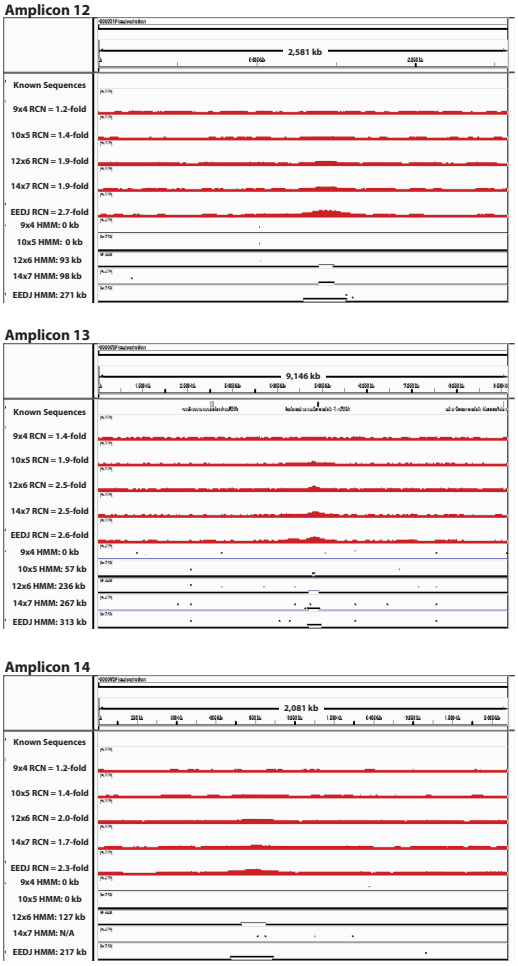


Figure 4.4: The middle-late smallest amplicons.

The panels are as described in figure 4.2.

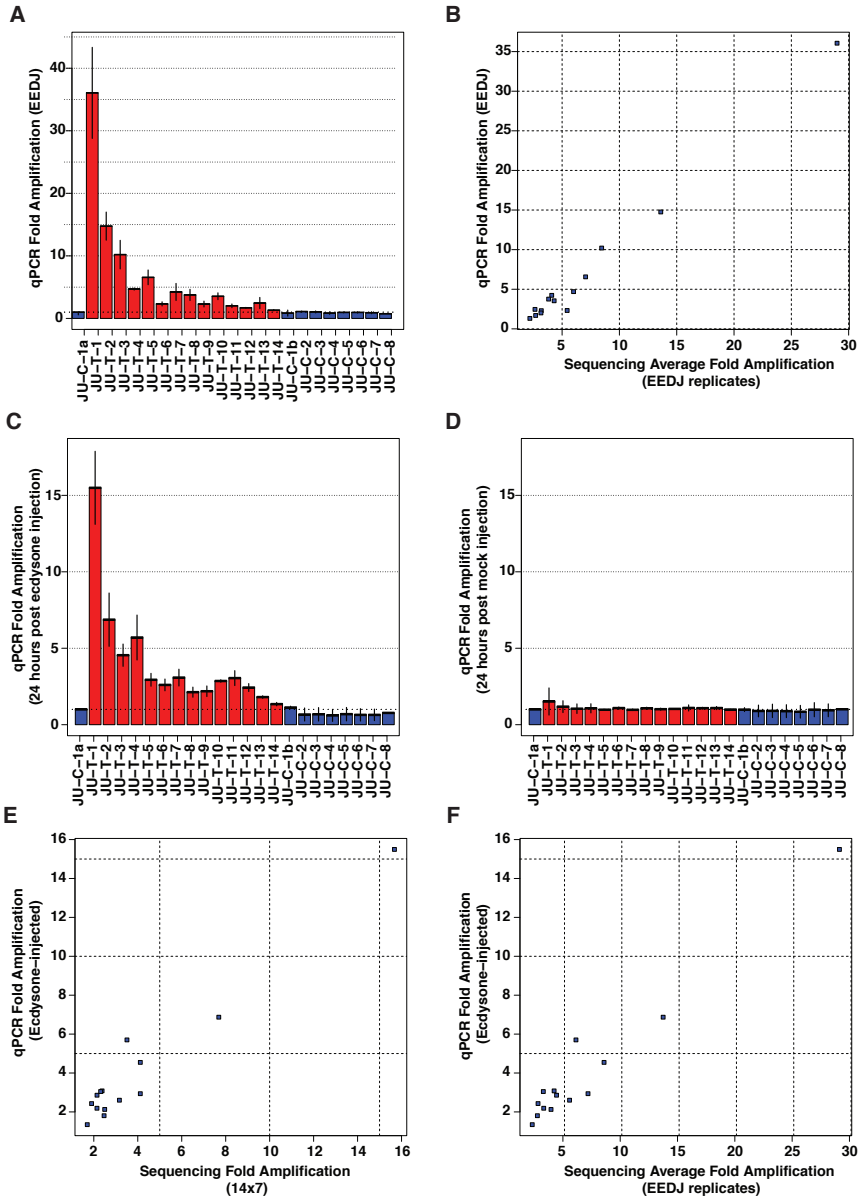


Figure 4.5: qPCR validation of RCN and the effect of ecdysone on DNA amplification. **(A)** qPCR validation of copy number estimates from sequencing in Edge Eye Drop Jaw (EEDJ) staged salivary glands. T1-T14 correspond to amplicons 1–14. C1–c8 are the eight control sites. **(B)** Scatter plot of the mean copy number from two EEDJ sequencing replicates versus the mean copy number from qPCR. The sequencing replicates had a Pearson correlation of 0.9994 with each other. The mean copy number from the sequencing replicates had a Pearson correlation of 0.993 with the copy numbers determined by qPCR. **(C)** qPCR copy number estimates 24 hours after ecdysone injection of early eyespot (< 8x4) larvae. **(D)** qPCR copy number estimates 24 hours after mock injection of early eyespot (< 8x4) larvae. **(E)** Scatter plot of the maximum copy numbers in 14x7 as determined by sequencing versus the copy numbers from ecdysone injection. Each point represents one of the 14 DNA amplification sites identified. Pearson correlation = 0.972. **(F)** Scatter plot of the average maximum copy numbers in EEDJ as determined by two sequencing replicates versus the copy numbers from ecdysone injection. Each point represents one of the 14 DNA amplification sites identified. Pearson correlation = 0.975.

4.3.3 The salivary gland transcriptome

The developmental transcriptome across 5 stages was investigated using 100 bp paired-end strand-specific RNA-seq. Specifically, poly-A RNA was sequenced in biological triplicate from pre-amplification ($\leq$ 8x4), 10x5, 12x6, 14x7, and EEDJ. The data was combined to perform a de novo transcriptome assembly with Trinity. Trinity assembled 57.8 Mb of sequence with a GC content of 38.8%, containing 40,369 Trinity genes and 50,600 Trinity transcripts. The assembled transcripts were aligned to the reference genome, keeping strand information intact for visualization. The RNA-seq reads were also aligned to the genome using the splice-aware read aligner, HISAT [Kim et al., 2015], and are visualized with the transcript alignments to view expression levels. The agreement of our RNA-seq data with previous studies is encouraging. As was seen previously [Wu et al., 1993], the highly expressed genes in DNA puffs II/9A and II/2B are barely detectable in earlier stages and begins to increase in expression by 12x6, peaking in 14x7 (Fig. 4.6 A–B). Moreover, as was shown in the same study, the intense amount of RNA synthesis in RNA puff III/9B does not occur until the Edge Eye and Drop Jaw stages (Fig. 4.6 C). The developmental relationship between amplification and transcription at most other puffs reflects that of the two major puffs, II/9A and II/2B, where amplification precedes the bulk of transcription as was shown by previous studies that looked at the uptake of radiolabeled DNA and RNA precursors in the polytene chromosomes [Gabrusewycz-Garica, 1964, Gabrusewycz-Garcia and Kleinfeld, 1966]. The majority of amplicons have a spike in RNA levels in similar patterns to II/9A and II/2B or a later spike in EEDJ as seen for RNA puff III/9B. Some have additional genes that spike in EEDJ after the first spike in 14x7. In contrast, two of the later smaller amplicons (12 and 13) show the highest RNA levels in the pre-amplification stage with a steady decrease through 14x7, and relatively nothing in EEDJ. Amplicon 14 had little to no RNA expression throughout all stages.

A previous study demonstrated that the initiation zone for II/9A is confined to a smaller region during amplification than seen during the endocycles and in mitotic cycling cells [Lunyak et al., 2002]. Specifically, the right boundary shifted to the promoter region of gene II/9-1 and it was observed that RNA polymerase II (RNAP-II) was upstream of gene II/9-1, thought to be poised in the promoter, ready for the spike in transcription around 12x6–14x7. Interestingly, we can now see that a transcript assembled from mixed-stage salivary glands aligns exactly over this region, suggesting that RNAP-II was observed in this location while it was transcribing this DNA sequence rather than being poised to transcribe II/9-1. Interestingly, when zooming in to view the RNA-seq read alignments, one can see alignments not only where this transcript aligns from stage 12x6–EDDJ, but there is also a low level of read alignments spanning the majority of the initiation zone during 12x6–14x7 (Fig. 4.6 D). It is possible that this low level of transcription through the origin during these stages shifts the initiation zone during the later rounds of amplification. Similarly, a low level of transcription seems to appear over the presumed II/2B origin (Fig. 4.6 E) as well as regions of highest copy number of other amplicons.

Since the ecdysone receptor may play both indirect and direct roles in DNA amplification, we used the known Sciara sequences of EcR-A and EcR-B to pull out matching transcripts from the de novo Trinity transcriptome assembly. There were five transcripts that aligned with the EcR sequences across their lengths with high percent identity (Fig. 4.6 F). Trinity reported that they are different isoforms of the same gene. Three of the isoforms are A-like, though only one had the most 5' exon from EcR-A. The other two are B-like. De novo transcriptome assemblies from short reads can give rise to false positive isoforms. Therefore, we used PCR and Sanger sequencing to test if the exon junctions unique to each isoform were real. This orthogonal evidence suggests that all five isoforms do exist in larval salivary glands. Moreover, the Sanger sequences aligned better to the de novo assembled transcripts than to the pre-established sequences of EcR [Foulk et al., 2013]. Looking at all Trinity transcript alignments, it appears that there are other genes on both strands that overlap the EcR locus. This makes the expression profile hard to interpret for the time being. However, the stage with the most RNA-seq reads aligning to the EcR locus is 14x7 (Fig. 4.6 F).

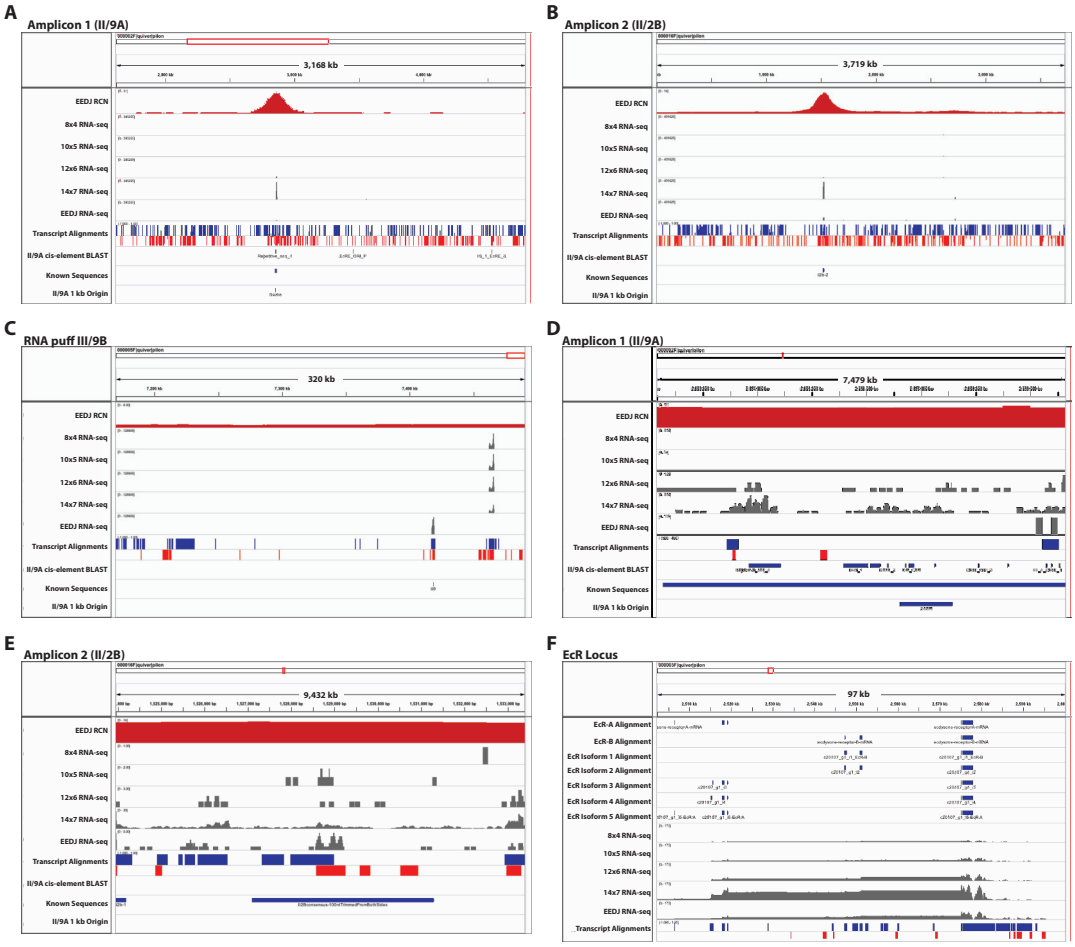


Figure 4.6: RNA levels in late larval salivary glands.

(A) Zoomed-out view of II/9A showing RNA levels across stages. Top red track is the relative copy number (RCN) of DNA in EEDJ to help frame where the amplicon is located. The next 5 grey tracks are RNA levels in salivary glands across the eyespot larval stages. The y-axis is set to the maximum expression level of the genes in view across all stages. Since the expression of gene II/9-1 is so high, other expression levels are imperceptible here. The next track has blue for Trinity-assembled transcripts that align to the positive strand and red for those that align to the negative strand. The next three tracks show pre-established features and sequences where relevant. **(B)** Zoomed-out view of II/2B showing RNA levels across stages. Panel details same as in (A). **(C)** Zoomed-out view of RNA puff III/3B showing RNA levels across stages. Panel details same as in (A). Note the known III/9B gene in the known sequences track. That is the region discussed in the text. **(D)** Zoomed-in view of II/9A showing RNA levels across stages. Panel details same as in (A) except that each stage's expression track (grey) is autoscaled to show RNA levels there. The scales between stages are not necessarily comparable. **(E)** Zoomed-in view of II/2B showing RNA levels across stages. Panel details same as in (D). **(F)** Ecdysone Receptor (EcR) locus. The top two tracks show the alignments of the two known EcR transcript sequences (EcR-A and EcR-B). The next five tracks show the alignments of EcR isoforms from the Trinity assembly. The following 5 tracks (grey) are the expression levels across stages. The final track shows all Trinity transcripts that aligned to this locus.

4.4 Discussion

There appears to be little doubt that the amplicons we identified with Illumina sequencing are real. Many show developmental increases in copy number and in amplicon length, consistent with expectations. Fourteen were selected for qPCR confirmation, and all had detectable copy numbers in agreement with the sequencing results. Moreover, all fourteen were stimulated to have copy number increases by ecdysone, but not by the mock control injection. Finally, some were mapped to their corresponding cytological DNA puff with FISH (Fig. 4.7), which also helped anchor long contigs from our recent assembly into chromosomes. The amplification levels at II/9A are at least double what was previously thought. We found this by following a hunch that amplification might continue into the later Edge Eye and Drop Jaw stages, potentially at a lower rate after 14x7. Whereas the copy number of approximately 16-fold was found in 14x7 in agreement with previous studies, the copy number in EEDJ reached at least 32-fold. The final copy number of II/2B was in agreement with previous studies. In general, amplification continued and even began in later stages than previously appreciated. It is known that some DNA puffs only arise in the anterior salivary gland, and others only in the posterior [Gabrusewycz-Garica, 1964]. It is unknown, however, if DNA amplification levels differ at these sites between the anterior and posterior. It is tempting to assume it does, and that for those particular loci, our copy number estimates are lower than the true amplification levels that would be detected if the anterior and posterior sections of salivary glands were sequenced separately. That will be a focus of future work. It is also unclear at the moment if there is loss of amplification levels through nucleolytic degradation as the larvae enter pupation during or near the end of Drop Jaw. If so, that would also dampen the copy number estimates. In future experiments, Edge Eye and Drop Jaw stages should be separated since we now know amplification continues into them.

The amplicons are much wider than anticipated, demonstrating that the re-replication forks in Sciara travel up to 250 kb or more in each direction. In contrast, the amplicons in Drosophila are 100 kb, with forks traveling up to 50 kb in each direction. However, cyclin E and Suppressor of Under-replication (SUUR) mutants lead to the forks traveling twice as far in Drosophila [Park et al., 2007, Sher et al., 2012, Nordman et al., 2014], suggesting there are no set barriers. The difference in how far the replication forks travel in Sciara is probably explained by time. Whereas amplification in Drosophila follicle cells happens over the course of hours, amplification in Sciarid salivary glands happens over the course of days. The relatively long developmental window of re-replication fork elongation and the extremely far distances they travel make Sciara a highly suitable system to study elongation, complementing recent studies in Drosophila [Sher et al., 2012, Nordman et al., 2014, Yarosh and Spradling, 2014, Alexander et al., 2015].

Ecdysone injection was able to stimulate amplification in all fourteen amplicons tested. The resulting copy numbers were in proportion to their normal copy numbers, which are correlated with developmental timing. Earlier amplifying sequences reach higher copy numbers. The question is

still open, though, on whether ecdysone is directly involved with DNA amplification or indirectly involved. Ecdysone-injection studies that simultaneously blocked RNA or protein synthesis seemed to block the effects of ecdysone on amplification [Foulk et al., 2006]. This suggests that it initiates a cascade of events that lead to DNA amplification. That the effects of ecdysone injection seem to recapitulate the developmental and size relationships between the amplicons identified in this study seems to support the cascade initiation model. In contrast, there is an Ecdysone Response Element (EcRE) directly adjacent to the binding site for the Origin Recognition Complex (ORC) at the II/9A origin, called the ORI EcRE, that binds the Ecdysone Receptor (EcR) in vitro [Foulk et al., 2006]. It is tempting therefore to speculate that ecdysone may also play a direct role. We identified 5 isoforms of EcR in our salivary gland transcriptome data, the unique exon junctions of which were validated with PCR and Sanger sequencing to ensure they were not de novo assembly artifacts. It is an open question on whether these isoforms have redundant functions or if the isoform diversity represents functional diversity. Perhaps one is specifically involved with amplification. Overall, it remains possible that ecdysone and the EcR isoforms play both indirect roles in stimulating the transcription and translation of factors needed for amplification as well as direct roles at some of the amplification origins, particularly II/9A.

4.5 Epilogue

There are still several analyses to do for this paper. I have done preliminary analyses to compare puff sequences. There are numerous shared motifs identified by MEME, most are AT-rich, when looking at large regions. However, most of the motifs are probably not related to amplification. Identifying the ORC and nascent strand distributions inside each amplicon would possibly help narrow down the regions to perform motif searches in. Preliminary BLAST results for cis-elements thought to be important for II/9A show that none are universally represented in the amplicons, nor particularly enriched. However, a closer inspection is required to conclude that they are not present. Most or all amplicons had sequences that looked like Ecdysone Response Elements (EcREs). However, a preliminary analysis suggested that were not enriched relative to other areas of the genome. EcREs are ubiquitous. Nonetheless, given that there are at least five isoforms of the Ecdysone Receptor (EcR), it is possible that there are different classes of EcREs, one of which is unique to amplicons. Since there are probably many more EcRE-like sequences than there are EcR binding sites, the only real way to pursue the question of whether EcR is bound at all amplicons is performing ChIP-seq specific to each isoform. I have also begun to explore fork speeds (discussed more below). For example, if you take the 607–663 kb II/9A ampicon and assume that the forks have been traveling for 5 days, then the average speed is somewhere between 84–92 bases per minute, over ten times slower than normal fork speeds. The slowness may in part be due to traversing the polytene chromosomes. It may also reflect a fair amount of replication fork breakdown and restarting. In the following subsections, I outline work in progress and possible avenues to explore.

Mapping more amplicons to their corresponding puffs

We are still in the process of using FISH to connect the deep sequencing data with the longer history of cytological observations of these puffs. So far we have mapped four new amplicons to their corresponding DNA puffs (Fig. 4.7). However, these will need to be repeated to be certain and the others will need to be mapped as well.

Refining estimates of initiation zones and motifs therein

The amplicons are very broad, spanning as much as 500 kb or more. The summits, or areas with the highest relative copy numbers, in the amplicons shift are the best estimate of where re-replication initiates. However, they shift considerably depending on smoothing. This makes it difficult to define small areas where initiation may begin in order to find motifs. Alternatively, initiation may occur over large areas. Nonetheless, to try to narrow down candidate initiation zones, a few approaches can be attempted. First, we can look at the stage when each amplicon first arises in to see if it helps confine the bounds of searching. The relative copy numbers from each stage can also be summed to help give the central origin area a boost over regions that flank it. Overall, defining narrower bounds supported by the data would help defining motifs shared by the DNA puffs since motif searches now seem to be extremely abundant in what are probably motifs regulating transcription of the nearby genes. It will also be worth trying different motif search parameters. Since a 72-100 bp stretch in the DNase Hypersensitivity Site (DHS) near the II/9A origin [Urnov et al., 2002] is conserved with a DNA puff sequence from a related species, it may help to set long minimum motif sizes.

Replication fork speeds and amplicon asymmetry

The partitioning of the amplicons into copy numbers that are multiples of two, using the pufferfish HMM, allows an estimation of how far replication forks from previous initiation events tend to travel. Since the stages are separated by approximately 24 hours, this allows an estimation of average fork speeds over large areas. Moreover, since it is possible to follow the average fork placement of a given round of re-replication over multiple stages, one can also estimate the fork speed of each over time. Do forks move slower as they travel farther or are there regions where forks tend to slow down before resuming a more average speed? These analyses can be bolstered by the fact that in some cases we can observe the estimated fork speeds over given regions from forks that arise out of multiple rounds of re-replication. We can check whether a given region consistently causes forks to slow down. Such regions could later be correlated with other genomic features, such as chromatin marks, in future studies. Are these regions heterochromatic? We can also determine whether forks in subsequent rounds of re-replication are slower on average than forks from earlier rounds.

A related analysis arises from the observation that the shapes of the amplification curves clearly show asymmetry, which reflects either differential fork speeds on opposing sides of the amplification origin or in extreme cases genomic regions that act as barriers to fork progression. An analysis

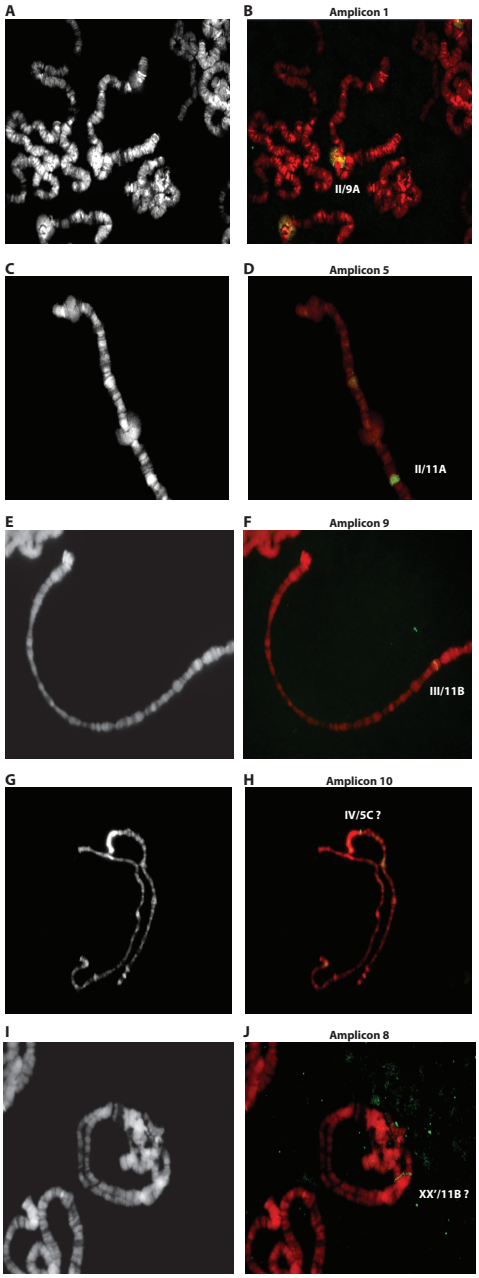


Figure 4.7: Mapping amplified DNA sequences to corresponding DNA puffs with FISH. **(A)** and **(B)** show sequence from the II/9A origin (amplicon 1) mapped to its location on chromosome II. **(C)** and **(D)** show sequence from the amplicon 5 mapping to chromosome II at locus 11A. **(E)** and **(F)** show sequence from the amplicon 9 mapping to chromosome III at locus 11B. **(G)** and **(H)** show sequence from the amplicon 10 mapping to chromosome IV perhaps at locus 5C. **(I)** and **(J)** show sequence from the amplicon 8 mapping to chromosome X/X' perhaps at locus 11B.

on Drosophila amplicons measured the asymmetry by calculating the distance to each side of the origin it takes to reach half the maximal copy number [Nordman et al., 2014]. This analysis will be applied to Sciara amplicons. However, there are many ways to expand upon it. The amplicons do not necessarily become asymmetric at 50% the maximum height. Therefore, it would be better to analyze the amplicons in a way that points to where the asymmetry begins. One way to do this would be to would be to plot the height of each arm of the amplicon as a function of distance from the summit to see at what distance and at what percent of the maximum height the arms diverge. Similar analyses to define points of asymmetry would involve plotting the percent height of the left arm vs the right arm at fixed distances, or plotting the distance away of each arm for fixed percent heights. Slope changes in the curves for each would reflect distances away from the origin where asymmetry-causing events occur. One can also look at areas transition points of slope changes in the amplicon curves. Since we have 5 or more stages, results from each stage can be compared. Overall, these analyses should give consistent results with those discussed above that use the HMM segmentations to look at average fork speeds in larger bins.

The salivary gland transcriptome over development

There are a number of analyses to be done with the RNA-seq data. First, the RNA-seq data allows us to weigh in on the question of whether amplification precedes transcription at all of the DNA puffs. Preliminary results suggested that this is the case for all but 2–3 amplicons. Second, it was seen in Drosophila that not all amplified genes are highly expressed and not all highly expressed genes are in amplicons [Kim et al., 2011]. It will be interesting to see if that conclusion generalizes to another amplification system. A preliminary view of RNA levels suggest that are definitely highly expressed genes outside of amplicons. This is an interesting possibility because it re-opens the question of why DNA amplification is used at all if it is not necessary for high gene expression. In Drosophila, amplification seems to be necessary since under-amplification leads to female sterility. Third, differential expression analyses between the different stages can highlight potential amplification factors. It is known that the induction of amplification by ecdysone-injection needs active RNA and protein synthesis to take place [Foulk et al., 2006]. Therefore, transcripts that are at higher copy numbers during or after amplification than before it become an interesting place to start looking. Fourth, since we have RNA-seq data across five stages, we can go beyond pairwise differential expression analyses and begin to look at developmental expression patterns. For examples, some transcripts may oscillate up and down, some may show a bell curve across the stages, others may progressively increase, and others may steadily decrease. One way to break transcripts up into groups by there developmental patterns will be hierarchical and k-means clustering approaches. Fifth, we have identified and validated that there are five isoforms of the ecdysone receptor. The next questions are, what are their expression profiles across the stages and how do they correlate with amplification and transcription in the DNA puffs? Presumably the different isoforms have different functions and

at least one may be involved with amplification instead of transcription. Do any of their expression profiles suggest a role in amplification? The same analyses can be done for ultra-spiracle, which heterodimerizes with EcR, as we have found multiple isoforms for it as well.

Nature of the genes in DNA puffs

The genes at II/9A are thought to encode structural proteins that comprise the pupal coat construction and that are needed in a short developmental window. What do the transcribed genes inside the other DNA puffs encode? Do they have homology to any known transcripts or proteins? What are the characteristics of the predicted protein sequences?

Shifting initiation zones with the turning on and off of local genes

Looking for amplification summits in all stages will allow us to begin to look at whether initiation zones tend to change position over time, though nascent strand analysis in the future will be more definitive. An earlier study on II/9A [Wu et al., 1993] noted that (i) amplification plateaued in 14x7, (ii) there was still DNA synthesis detected in stage 14x7 by tritiated thymidine uptake, and (iii) in stages later than 14x7, 2D gels detected replication forks coming through what is thought to be the central II/9A origin. Their interpretation was that amplification ends in 14x7 and DNA synthesis and forks detected were from elongation. However, an alternative interpretation we posit is that the replication forks detected by 2D gel came from II/9A in a different area and that the plateau did not mark the end of amplification, just a slower amplification rate. Our data in this manuscript confirm this latter interpretation, showing that amplification levels in 14x7 are those that have been seen previously, but that they double again in the later stages. Therefore, the initiation zone at II/9A must shift such that amplification continues, but not from where the 2D gels would detect bubble arcs. We can make low resolution inferences about where the initiation zones tend to be in each stage. Then we can correlate that with the turning on and off of genes in the DNA puff that may shape the landscape of where MCMs go [Powell et al., 2015, Gros et al., 2015] and where corresponding initiation may begin. It would be intriguing to find that the initiation zones shift position when genes turn on and off. Does initiation zone shifting occur? If so, does it correlate with transcription, and does it tend to maintain a bidirectional pattern where expressed genes in a given stage face away from the amplification zone?

Homology of DNA puffs with amplicons from other insects

There are six amplicons in Drosophila follicle cells, and the sequences for a few salivary gland amplicons from Rhynchosciara and Bradysia hygida are available. All can be combined in an attempt to find shared motifs. Conversely, motifs found in Sciara puffs can be searched for in the other amplicons, and motifs from the other amplicons can be used to scan the DNA puff sequences. Any pairwise local alignments can be considered further as well.

DNA topology

In addition to predicting motifs in the DNA puffs, it is also possible to predict DNA bends, DNA stiffness, and nucleosome positioning. The results of those analyses can be compared with the experimental data and conclusions for II/9A. These analyses, in addition to the location and timing of genes, could help refine what motifs are most likely involved in amplification versus transcription. This will also help define whether the bends at II/9A are predicted to exclude nucleosomes or not.

4.6 Methods

4.6.1 Detecting, validating, and interrogating DNA puff sequences

DNA sequencing library preparation

Sciara coprophila were maintained in the laboratory at 21C. Fourth instar female larvae from the Holo2 stock were staged according to eyespot progression [Gabrusewycz-Garica, 1964, Foulk et al., 2006]. Salivary glands were dissected from larvae in 250-300 ul of Robert's CR Buffer (87 mM NaCl, 3.2 mM KCl, 1.3 mM CaCl2, 1 mM MgCl2, 10 mM Tris-HCl; pH 7.3) (Robert, 1971) with care to remove all or the majority of fat tissue that attaches to the glands. Larvae were typically dissected in batches of five. After cleaning the fat away from all glands in each batch, the batch of glands were transferred to 100 μl DNAzol (Invitrogen/ThermoFisher). DNAzol containing salivary glands was kept on ice while working and was stored at 4°C over the course of collecting all salivary glands for each library preparation, which was always less than 3 days as per manufacturer's instructions. Larvae were partitioned into five developmental eyespot stages: early eye-spot $\leq$ 8x4 (pre-amplification), 10x5, 12x6, 14x7, and EEDJ (post-amplification: the combination of edge eye and drop jaw) to compare with RNA-seq data from the same stages. An extra replicate of the post-amplification sample (EEDJ) was prepared to confirm final copy numbers. A sample of larvae that were between 8x4 and 10x5 (that we termed 9x4) was included to check if sequencing could detect amplification prior to 10x5 when it was thought to initiate. The number of larvae dissected for each sample is as follows: 34 for $\leq$ 8x4, 37 for >8x4/<10x5 (intervening stage), 36 for 10x5, 36 for 14x7, and 28 for each EEDJ replicate.

Experiments that applied DNAzol to salivary glands while monitoring with a microscope suggested that salivary gland cells lyse rapidly from this reagent. Nonetheless, starting with the 100 ul DNAzol that contained DNA from 28–37 staged salivary glands, 5-10 strokes using a blue pestle was performed to disrupt any cells that were not yet lysed. The blue pestle was washed off into the microfuge tube with an additional 100 ul DNAzol (200 ul total). Then 100 ul 100% ethanol was added, followed by slowly inverting the tube 50 times, then centrifugation at 21,000xg at 4°C for 10 minutes. The supernatant was discarded and 1 mL of 75% ethanol was added to the tube. The DNA pellets in 75% ethanol were stored in -20°C for a minimum of 1 hour. When resuming, a total of two 75% ethanol washes were performed. The tubes were briefly spun to collect remaining ethanol at

the bottom to remove with a pipette. The DNA pellets were air-dired for a very brief perior of < 30 seconds according to the manufacturer's directions. The pellet was re-suspended in 100 μl TE (10 mM Tris-Cl, pH 8, 1 mM EDTA), incubated at room temperature for 3 hours, then stored at 4°C until needed for preparing the sequencing libraries. Overall, the salivary gland preparations yielded 5.7–7.4 μg of genomic DNA, indicating 160–264 ng per pair of salivary glands. As expected later stages gave more DNA per pair. The stages $\leq$8x4, 10x5, 12x6, 14x7, and EEDJ gave 160, 188, 212, 230, and 239–264 (mean = 252) ng per pair of glands, respectively.

DNA samples were incubated with 3 μl RNase A at 37°C for 30 minutes, followed by 70°C for 5 minutes, and transferred to ice where 1 μl of Murine RNase inhibitor was added to each. DNA samples were then diluted into a total of 200 μl TE for sonication with the Bioruptor. Sonication conditions were optimized to fragment DNA into the 200–400 bp range as analyzed by gel electrophoresis. The DNA samples for sequencing were sonicated at 0–4°C for 30 cycles of 30 seconds on and 90 seconds off at medium power.

Sheared DNA was cleaned with a 2.0x ratio of AMPure XP beads (BeckmanCoulter) and eluted into 88 μl of UltraPure Water (UPW). Qubit was used to estimate the DNA concentration and NanoDrop was used to estimate DNA concentration and purity. The Qubit and Nanodrop gave very similar concetration estimates of 18–38 ng/μl across all samples. Based on the Qubit results, there was 1.6–3.0 μg per sample going into the library preparations.

Sequencing libraries from sheared DNA were prepared using NEBNext following the manufacturer's instructions. Briefly, end repair consisted of 85 μl DNA combined with 10 μl End Repair buffer and 5 μl End Repair enzyme, and incubated at room temperature for 30 minutes. The DNA was cleaned with 1.8x AMPure beads, and eluted into 42 μl UPW. For dA-tailing, 42 μl of DNA was combined with 5 μl 10X dA-tailing buffer and 3 μl Klenow (3'-5' exo–), and incubated at 37°C for 30 minutes. The DNA was cleaned with 1.8x AMPure beads, and eluted into 25 μl UPW. DNA samples in UPW were stored at -20°C overnight after this step, and thawed on ice the next day. For adapter ligation, 25 μl DNA was combined with 10 μl Ligation Buffer, 10 μl NEBNext adapter (the amount instructed for 1-5 μg starting material), and 5 μl Ligase, then incubated for 15-20 minutes at room temperature. To convert the NEB uracil-containing hairpin adapters into Y-shaped adapters, 3 μl of USER enzyme was added to the ligation and incubated for 15 minutes at 37°C for 15 minutes. The adapters add 65 bp to the insert lengths. The DNA was cleaned with 1.0x AMPure to remove DNA < 200 bp, and eluted into 100 μl UPW for overnight storage at -20°C. The next day, libraries were thawed on ice, then size selected with AMPure beads. Since adapters add 65 bp and we were targeting 200-400 bp DNA inserts, AMPure size selection used ratios that we have found to select for DNA between 270-470 bp. To the 100 μl DNA samples, 73 μl AMPure beads (0.73x) was added and incubated for 5 minutes at room temperature before pelleting on a magnet. The supernatant containing DNA <470 bp was transferred to a new tube where 20 μl more of AMPure beads was

added to a final ratio of 0.93x. The AMPure procedure was continued as normal at this point, keeping the DNA on the beads that was > 270 bp, and discarding the smaller DNA species in the supernatant. DNA was eluted in 20 μl UPW. For PCR, 20 μl of DNA was combined with 26 μl 2X PCR master mix, 2 μl USER enzyme (to ensure complete conversion of adapters), 2.5 μl universal primer, and 2.5 μl indexed primer. The tubes were put into the thermocycler for 15 minutes at 37°C, 30 seconds at 98°C, and 8 cycles of 10 seconds at 98°C, 30 seconds at 65°C, and 30 seconds at 72°C, before keeping at 0°C and transferring to ice. The PCR primers and adapters add 122–128 bp, so 200-400 bp inserts correspond to 325–525 bp library DNA. We cleaned the DNA from the PCR reaction with 0.87x AMPure beads to deplete DNA < 330 bp. The DNA was eluted in 100 μl UPW, and AMPure size selection was performed. Specifically, 67 μl of AMPure beads was added to the 100 μl of DNA (0.67x), incubated for 5 minutes at room temperature, then pelleted on the magnet. The DNA in the supernatant, which is approximated by our optimizations to be < 530 bp, was transferred to a new tube. Another 20 μl of AMPure beads was added for a final ratio of 0.87x, and the AMPure procedure was completed as normal, keeping the DNA > 330 bp that was on the beads. The DNA was eluted into 32 μl UPW. The amount of DNA at the end of library construction was quantified with Qubit. There was 192–963 ng in each library. The Fragment Analyzer (Advanced Analytical) showed peak DNA sizes at 353–383 bp with DNA detected typically within the 300–500 bp size range. The samples were sequenced on a single lane of the Illumina HiSeq2000 for 50 bp single-end sequencing.

Identifying amplified regions

We designed a pipeline to find DNA puff sequences for FISH testing on polytenes, called pufferfish (`https://github.com/JohnUrban/sciara-project-tools/tree/master/pufferfish`). The pipeline starts with mapping reads from pre-amplification stage salivary glands to the genome as well as the reads from a later stage (e.g., post-amplification):

```
$ bowtie2 -q –very-sensitive -N 1 -x bowtie2index -U pre-amp-reads.fastq.gz 2>pre-amp.err | samtools view -bSh -F 4 - 2>>pre-amp.err | samtools sort -o pre-amp.bam 2>>pre-amp.err
$ samtools index pre-amp.bam 2>>pre-amp.err
```

```
$ bowtie2 -q –very-sensitive -N 1 -x bowtie2index -U post-amp-reads.fastq.gz 2>post-amp.err | samtools view -bSh -F 4 - 2>>post-amp.err | samtools sort -o post-amp.bam 2>>post-amp.err
$ samtools index post-amp.bam 2>>post-amp.err
```

The genome is then broken up into 500 bp bins using BEDtools.

```
$ bedtools makewindows -g sciara.genome -w 500 -s 500 > w500.s500.bed
```

The number of reads are then counted in each bin. Only high quality alignments (MAPQ $\geq$ 30) are

considered, though the results were consistent even with no filtering:

$ samtools view -b -h -F 4 -q 30 pre-amp.bam | coverageBed -abam - -b w500.s500.bed | sortBed -i - | cut -f 1,2,3,4 > pre-amp.q30.w500.s500.bedGraph
$ samtools view -b -h -F 4 -q 30 post-amp.bam | coverageBed -abam - -b w500.s500.bed | sortBed -i - | cut -f 1,2,3,4 > post-amp.q30.w500.s500.bedGraph

In our pufferfish pipeline, the above procedure is wrapped over using:
$ pufferfish mapreads –bt2 bowtie2index pre-amp-reads.fastq.gz
$ pufferfish mapreads –bt2 bowtie2index post-amp-reads.fastq.gz
$ pufferfish getcov -g sciara.genome -w 500 -s 500 -Q 30 pre-amp.bam
$ pufferfish getcov -g sciara.genome -w 500 -s 500 -Q 30 post-amp.bam

Pufferfish was then used to segment the genome into different copy numbers using the bedGraphs described above as the starting substrate. First, for each file, the read counts in 500 bp bins were internally normalized to give the relative copy number (RCN) compared to non-amplified regions. Nearly identical results were obtained by finding normalization factors using various approaches, such as (i) using the median bin count, (ii) using areas believed to be un-amplified, (iii) using the mean or median of a cluster in k-means clustering that corresponded to RCN=1, and (iv) using the mean or median of regions identified as having RCN=1 after iterating over a hidden Markov model (as below). Therefore, we chose median normalization due to its simplicity and easy interpretation. Since the majority of bins in the genome cover non-amplified regions, the median bin count of a given file captures a value near the central tendency of non-amplified regions. Overall, all bin counts in a given sample were internally normalized by the median bin count from that sample. The internally normalized bin counts from the later stages were then externally normalized to the same bin count from pre-amplification stage to correct for biases in sequencing and limit the effect of high coverage from collapsed repeats in the reference. This gave the final RCN values. We then used a hidden Markov model (HMM) approach to segment the genome into different copy numbers (the hidden states) using the final RCN values as emissions. Since a complete round of re-replication would double the RCN, we segmented the genome into 7 different RCN states that are multiples of 2: 1, 2, 4, 8, 16, 32, 64. Though prior knowledge that II9A, which is the DNA puff that amplifies the most (via cytological data), reaches only 16-fold, we allowed for the possibility that it amplified further since we were testing later stages than previous studies. We used the Viterbi algorithm to find the most likely state path, though posterior decoding gave similar results. For initial probabilities, we assumed amplicons may reach up to 50 kb on average, and given ~18 amplicons and a 292 Mb genome, the initial probability of starting in each one of the amplified states (RCN 2, 4, 8, 16, 32, 64) was approximately 0.0005 ($50e3 * 18/292e6/6$) and the initial probability of starting in an un-amplified region was 0.997. For emission probabilities, we used normal emissions models, setting the mean RCNs for each state to 1, 2, 4, 8, 16, 32, 64; and setting the standard deviations to the

square root of the means. For transition probabilities, we found that setting self-to-self to 0.994 and self-to-other to 0.001 gave good sensitivity to amplified regions while limiting the amount of spurious state changes in unamplified regions. The bins were then filtered to retain only those with RCN > 1. Filtered bins often tile contiguous stretches, but are some times interrupted by gaps from bins that were labeled as RCN=1 and filtered out. These gaps often reflected regions of low mappability where reads with MAPQ < 30 were filtered out. On occasion this was not the case, and it may reflect a local mis-assembly in the reference. To identify amplified reigions, after bin filtering adjacent bins were merged allowing a distance up to 10 kb to separate adjacent bins (to span gaps). Only regions that spanned more than 50 kb were considered further. The following commands execute these last steps:

$ pufferfish puffcn -l postamp.q30.w500.s500.bedGraph -e pre-amp.q30.w500.s500.bedGraph -1 -c > cn.post-amp.p1.withearly.bedGraph

$ awk '$4>1' cn.post-amp.p1.withearly.bedGraph | mergeBed -d 10000 -i - -c 4 -o median | awk '$3-$2 > 50e3' > cn.post-amp.d10k.width50k.bed

qPCR primer design and validation

We chose 14 targets to test(JU-targets 1–14): 12 putative amplicons and II/9A and II2B sites. We also chose 8 control regions (JU-controls 1–8) where the relative copy number was expected to be 1 (based on the pufferfish analysis of the sequencing data as well as visual inspection in IGV). The control sites were paired with the first 8 target sites such that all control sites were on the same contig as their paired target sites at least 0.5–2 Mb away from the boundary of the target amplicon. The bedGraphs described above with final RCNs were smoothed various ways and summits within amplicons were predicted. For each amplicon, we manually viewed the smoothed curves, their corresponding summit positions, un-smoothed RCN values, and predicted HMM states in IGV, and chose the summit that seemed most consistent with the data. An example for finding summits using our pufferfish "summits" pipeline (https://github.com/JohnUrban/sciara-project-tools) was is as follows:
pufferfish summits -l EEDJ.rep1. q30.w500.s500.bedGraph -e earlyeyespot. q30.w500.s500.bedGraph –regions summits.bed -4 -bw 100000

Starting with 2 kb to each side, progressively longer sequences surrounding the chosen summits were used until highly specific qPCR primer pairs were found, going up to 10 kb to each side of the summit in some cases. These summit-centered sequences were searched for candidate primer pairs using NCBI Primer BLAST with the following settings: PCR product size 60–120, minimum primer size = 18, optimum primer size = 20–21, maximum primer size = 25, minimum Tm = 57, optimal Tm = 61–62, maximum Tm = 66, maximum Tm difference = 1-, maximum 3' self complementarity = 2, maximum self complementarity = 8 (preference given to lower), maximum poly-X = 4. In cases

where other primers needed to be tested, we increased the maximum primer size to 135, the primer size range to 16–26, the Tm range to 55–66, the maximum Tm difference to 2–3, and the maximum poly-X to 5. A minimum of twenty candidate primer pairs were returned and screened for specificity in the genome.

Since our genome sequence was not on NCBI, we could not use primer BLAST to check primer pairs for specificity. Instead, the NCBI primer BLAST results were copied into a text file to be parsed and analyzed by a custom script called "analyzePrimerPairs.py" (`https://github.com/JohnUrban/sciara-project-tools`). This script uses bowtie2 to perform several tests on the primer pairs. First, for each primer in the pair, it aligns the primers as mock single-end reads, requesting all alignments (e.g. bowtie2 -x $BT2 -U $P1 -f -a). Next it does the same test, but with less stringent alignment parameters (e.g. bowtie2 –rdg 2,2 –rfg 2,2 –mp 2 –very-sensitive -N 1 -x $BT2 -U $P1 -f -a). It then performs two pairing tests (one less stringent than the other), to align the primers to the genome as mock paired-end reads, allowing a maximum insert size of 1000 bp and requesting all alignments be returned (bowtie2 –no-discordant –no-mixed -x $BT2 -1 $P1 -2 $P2 -f -a –maxins $MAXINS ; bowtie2 –no-discordant –no-mixed –rdg 2,2 –rfg 2,2 –mp 2 –very-sensitive -N 1 -x $BT2 -1 $P1 -2 $P2 -f -a –maxins $MAXINS). It then performs the two pairing tests (stringent and less stringent) using each primer with itself as a mock paired-end read to test for spurious self-self products. Overall, there are 10 tests. For the four independent mapping tests, we only allow one alignment. For the two primer1–primer2 pairing tests, we only allow one alignment less than 1000 bp (i.e. 1 predicted PCR product). For the four self-self pairing tests (2 for primer1, 2 for primer2), we require 0 alignments less than 1000 bp (i.e. no predicted PCR products). Using the NCBI primer design parameters above and these specificity tests resulted in high quality primer pairs that passed validation tests below.

Primers were ordered from ThermoFisher/Invitrogen and re-suspended in UltraPure Water (ThermoFisher) to make 100 μM stocks and diluted to make 10 μM and 1 μM working stock aliquots. Primer validation was run using genomic DNA from adult female Holo2 Sciara where the relative copy numbers of all loci are expected to be 1 [Foulk et al., 2006]. The gDNA was diluted to 1 ng/μl, then serially diluted to obtain the final test ratios of 1:1, 1:4, 1:16, and 1:64. For each primer pair, each condition was done in duplicate. A single qPCR reaction (15 μl) contained: 2 μl DNA, 7.5 μl 2X SYBR Green PCR Master Mix (Thermo Fisher Scientific), 2.5 μl UPW, 1.5 μl of 1 μM Forward primer, and 1.5 μl of μM Reverse primer. Real-time qPCR was performed in 96-well optical plates on an ABI 7300 Real-time PCR System for 40 cycles, followed by a dissociation stage (Applied Biosystems, Thermo Fisher Scientific). The amplification curves were visually inspected for consistency of the replicates and dissociation plots were checked to ensure there was only a single PCR product present in all curves. The cycle threshold (Ct) data for each qPCR reaction was brought into R to ensure each primer pair passed 4 tests: slope, efficiency, R^2, and relative efficiency. For a primer pair to pass validation, the slope of the standard curve needed to be in

the range -3.58 – -3.10, the reaction efficiency had to be in the range 0.9–1.1, and the R^2 needed to be ≥ 0.97. For the relative efficiency test, we required each primer pair to pass with respect to JU-control-1 and JU-control-2 (JU-control-1, the paired control for II/9A was the normalizer locus for qPCR tests in this study). To perform the relative efficiency tests, a normalization locus was chosen. The average Ct values were taken for each dilution in the series for the primer pairs from both the test locus and the normalization locus. Then at each dilution in the series, the difference between average Cts was taken: $Ct_{norm} - Ct_{test}$. The requirement for this test is that the difference in Ct values between the test and normalization loci stay relatively constant across the series. Therefore, to pass, the Ct differences needed a slope between -0.1 – 0.1. In the case that a primer pair failed any of the four tests, it was inspected for outliers. If removal of the outlier resulted in it passing all four tests, it was given a conditional pass and was tested again. If it passed unconditionally the next time, it was kept. Passing all these tests allows the $\Delta\Delta$Ct method to be performed.

The primer sequences used in all qPCR in this study are:
o-JU-target1-fwd-1 CCACTGTCACATCATCATCGCC
o-JU-target1-rev-1 GCGACGTTGCCTGTCAATCC
o-JU-target2-fwd-1 TATATGAGGCGAGGCCGAGG
o-JU-target2-rev-1 CGTTTCCGGTTCTCCCTCAC
o-JU-target3-fwd-1 CTCCGCTACGCTCAACAACA
o-JU-target3-rev-1 CATGGCCATCACCGAAGACA
o-JU-target4-fwd-1 GACCACTAACGTAAGCCGTGAA
o-JU-target4-rev-1 AAATTCCCAAAGGTGGATGCGA
o-JU-target5-fwd-1 ATCGCTTGATCGCTGGCAAA
o-JU-target5-rev-1 TTACCGTCCACTACACCCACA
o-JU-target6-fwd-1 TCGGCGCATAATGAACCTGAAA
o-JU-target6-rev-1 GCCCTTGAACAGAACCTTCCC
o-JU-target7-fwd-1 GGCAACGAGCCGATAAGGTC
o-JU-target7-rev-1 TATGTCAGCGCTGGTTTGGG
o-JU-target8-fwd-1 ATTCCGTGCCTCCCGATCTT
o-JU-target8-rev-1 AGTGTATGAAGACACTGCGCC
o-JU-target9-fwd-1 CGCCCACGGACAATCCTTAC
o-JU-target9-rev-1 CTCGAAAGAGCGTTGCCAGA
o-JU-target10-fwd-1 GCTTATTCAGCGAATGGAGTGGA
o-JU-target10-rev-1 GTAGTGTACGGTGGAAACGGG
o-JU-target11-fwd-1 CCAAACGAGGAGACGCCATT
o-JU-target11-rev-1 TATGCCTGGCGAGGTCTTGA
o-JU-target12-fwd-1 CGGCTCAGCGGTTCTACCTA
o-JU-target12-rev-1 CGCGACGTTGCGTCTGAAA
o-JU-target13-fwd-1 GGCATAGACCAACATGGATCA

o-JU-target13-rev-1 TCGACACTTTTTGCAATGCTC
o-JU-target14-fwd-1 GCTACAGCTGGTGGTCGAGA
o-JU-target14-rev-1 CGAACTGGGCCTCATGCTTT
o-JU-control1-fwd-1 TGCGAGGTATGATTTACCGTCTT
o-JU-control1-rev-1 TAGTGCGATGGCGAATTGATGT
o-JU-control2-fwd-1 CGGAAGGACGGCTACGAGAA
o-JU-control2-rev-1 GGAATGTCGTCCTGCCGTAAA
o-JU-control3-fwd-1 ATACCCAGACCCAGGTGGTAGA
o-JU-control3-rev-1 AAGTGCTGTGGGATAGCGAAGT
o-JU-control4-fwd-1 GATGCCATTCGCCGATAGTGTT
o-JU-control4-rev-1 TTCACTCACGCGATTCTCACAC
o-JU-control5-fwd-1 GGGCAACCAACGAAAAGTGG
o-JU-control5-rev-1 TCGGATTCCGGCTCCATACA
o-JU-control6-fwd-1 TCCGAGCCATAATCCTCACCTT
o-JU-control6-rev-1 TCGCACGCTAACGAAGGTATCA
o-JU-control7-fwd-1 ATTGTGGCGCAAGTCGAATCA
o-JU-control7-rev-1 ATGACATCGCTCATGTCGGG
o-JU-control8-fwd-1 GTTTCATCGGGAGGAACGGG
o-JU-control8-rev-1 TGGGACACGACAACATCAACTAC

qPCR tests of amplification levels

Female larvae were staged and dissected as above. To test the final copy numbers in the late larval stages, 16 pairs of salivary glands (in three rounds of 5–6 pairs) from EEDJ staged larvae (8 of each) were dissected and transferred immediately into 100 μl of DNAzol on ice. This procedure was done in triplicate and each replicate had 16 pairs. Salivary gland DNA was extracted using DNAzol as above for sequencing except before adding the 100 μl 100% ethanol to precipitate the DNA, the DNA/DNAzol solution was incubated with 2 μl RNase A for 5 minutes at 37°C, followed by 2 μl of Proteinase-K for 5 minutes at 37°C. The DNAzol-extracted, precipitated DNA was re-suspended in 100-400 μl 10 mM Tris-HCl and stored at -20C. Similarly, female adult genomic DNA, used as a calibrator in qPCR, was extracted using the DNAzol reagent protocol, diluted to 1 ng/μl, and stored at -20C. We tested diluting the test and calibrator DNA samples to 50 pg/μl or 100 pg/μl for use in qPCR. Both gave identical results, so we continued with diluting samples to 50 pg/μl for validating DNA amplicon copy numbers. qPCR of each biological replicate was performed in technical triplicate. Each qPCR reaction consisted of 13 μl of PCR mix (7.5 μl 2X SYBR Green, 2.5 μl ultrapure water, 1.5 μl forward primer, 1.5 μl reverse primer) and 2 μl of the appropriate DNA sample at 50 pg/μl for a total of 100 pg per reaction.

Ecdysone injections and qPCR

All larval injections were performed using the Nanoject Auto-Nanoliter Injector (Drummond Scientific Company, Broomall, PA). Injection needles were pulled from 3.5" glass capillaries (#3-000-203-G/X, Drummond Scientific) using the Model 700C DKI Vertical Pipette Puller (David Kopf Instruments, Tujunga, CA). The settings for DKI Vertical Pipette Puller were as follows: heater, 45; solenoid, 50. The tip of the pulled needle was tapered under a dissecting scope using forceps. Needles were backfilled with mineral oil using a syringe, affixed to the injector, then filled with the injection solution. All ecdysone injections were done on early eyespot-stage (< 8x4), fourth instar female larvae. To prepare a single sample of salivary gland DNA, 10-15 larvae were anaesthetized with CO2 and injected with 32 nanoliters of a solution consisting of 1 mg/ml 20-hydroxyecdysone, 50% ethanol, and blue dextran as a tracer. Control larvae were injected with 32 nanoliters of 50% ethanol containing blue dextran. Injected larvae were incubated at room temperature for 24 hours on a 2.2% Bacto-agar plates (Becton, Dickinson and Company, Sparks, MD). For ecdysone-injected larvae, five larvae with a prematurely-induced edge-eye stage phenotype were selected from the plate, and the remaining larvae were discarded. Larvae that did not prematurely progress to an edge-eye phenotype were most likely not successfully injected with ecdysone solution (for example, the needle passed straight through the larva). Salivary glands from each of the five larvae were dissected in Roberts CR buffer. Salivary gland DNA was extracted using DNAzol reagent as above and eluted in UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher Scientific). Salivary gland DNA from control larvae was extracted using the same procedure. Altogether, three biological replicates, each with DNA extracted from five pairs of salivary glands, were prepared for both the ecdysone and control-injected larvae (six samples total).

Calibrator genomic DNA was extracted from female Holo2 adult flies using DNAzol as above, and eluted in UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher Scientific). Real-time qPCR was performed with primer pairs testing the 14 confirmed amplicons as well as 8 non-amplified regions. A 1X reaction mix contained 7.5 μl 2X Power SYBR Green PCR Master Mix (Thermo Fisher Scientific), 2.5 μl UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher Scientific), 1.5 μl of 1 μM forward primer, 1.5 μl of 1 μM reverse primer, and 2 μl of 50 pg/μl sample DNA. Real-time PCR was performed on an ABI 7300 Real-time PCR System for 40 cycles, followed by a dissociation stage (Applied Biosystems, Thermo Fisher Scientific). Three biological replicates for both ecdysone and control-injection salivary gland genomic DNA were analyzed by real-time PCR. For each biological replicate, three technical replicates of a sample were included on a plate for a particular primer pair.

qPCR analysis

Ct data from ABI 7300 Real-time PCR System was parsed and analyzed with custom R scripts (`https://github.com/JohnUrban/sciara-project-tools/tree/master/qpcr`). The $\Delta\Delta$Ct method was used for fold amplification analysis [Livak and Schmittgen, 2001, Schmittgen and Livak, 2008,

Weaver et al., 2010].

FISH probe design

FISH probes were designed in two ways. In the first way, 1–2 kb PCR products were used for probes. Primer pairs for PCR products were found using the same procedure as primers for qPCR above. However, NCBI Primer BLAST settings were: PCR product size = 1000–2000, Minimum Tm = 65, Optimum Tm = 67, Maximum Tm = 70, Maximum Tm difference = 1. Moreover, when checking the genome for specificity with "analyzePrimerPairs.py", the maximum alignment length between the mock paired-end reads (i.e. max product size to look for in the genome) was set to 10 kb. PCR was then carried out using Q5 HotStart Polymerase (NEB) according to the manufacturer's instructions. In the second FISH probe design approach, up to 10 kb to each side of the same summits used for qPCR primer design was used to select approximately 500 bp sub fragments. BLAST and Bowtie2 local mode were used to find the sub-fragments that were most unique in the genome:

```
bowtie2 -f -x bt2index -U subfragments.fasta -a --local
blastn -db blastdb -query subfragments.fasta -outfmt 6
```

We typically chose fragments that had only 1 local alignment of >200 bp and > 80%. Two sub-fragments for each amplicon were chosen, their sequences stitched together, and ordered for DNA synthesis from Integrated DNA Technologies (IDT).

Polytene and FISH procedure

Larvae were staged according to eye-spots. Salivary glands for puff stage larvae were dissected and squashed on microscope slides to release polyene chromosomes. The FISH protocol we used was adapted from "A simplified and efficient protocol for nonradioactive in situ hybridization to polytene chromosomes with a DIG-labeled DNA probe," by E.R. Schmidt (Roche online manual). It is described in detail in the Supplementary Methods. DNA sequences of 1 kb were synthesized by Integrated DNA Technologies (IDT) and labelled with Fluorescein-High Prime (Sigma Aldrich). Signal was visualized using rabbit Alexa 488 conjugated anti-Fluorescein Ab (1st) and goat anti-rabbit Alexa 488 Ab (2nd). DNA was stained by DAPI but colored in red.

Bioinformatics

4.6.2 Transcriptome

Strand-specific RNA-seq

Fourth instar female larvae were staged and dissected as above for DNA sequencing. Larvae were partitioned into five developmental eyespot stages: $\leq$ 8x4 (pre-amplification), 10x5, 12x6, 14x7, and EEDJ (post-amplification: the combination of edge eye and drop jaw) to compare with DNA-seq

data from the same stages. All stages were done in triplicate. The following number of larvae were dissected for their salivary glands for the three replicates of each stage: early eye-spot $\leq$8x4 (16, 22, 24), 10x5 (15, 17, 17), 12x6 (10, 12, 12), 14x7 (10, 12, 9), and EEDJ (12, 11, 11), for a total of 210 dissections for these experiments. Larvae were dissected in batches of 5–6 as described for dissections above, but the batches of salivary glands were transferred directly to TRIzol (Invitrogen/Thermofisher), and immediately further homogenized with 5–10 strokes using a blue pestle before storing in the TRIzol reagent at -80°C until needed for continuing total RNA extraction.

The 15 tubes containing TRIzol and salivary glands were thawed on ice and the TRIzol procedure was followed using the manufacturer's instructions. RNA quantity and purity were measured with the NanoDrop (ThermoScientific) and Qubit (ThermoFisher). Total RNA was treated with DNase (Qiagen) and subject to RNeasy column clean up (Qiagen). The cleaned total RNA quantity and purity were checked using the NanoDrop and Qubit. RNA integrity was evaluated on 1.1% formaldehyde 1.2% agarose gels. Poly-A RNA was enriched using Oligo-dT DynaBeads (LifeTechnologies). The Qubit was used to measure the quantity of poly-A RNA. The amount of poly-A RNA was typically 100-300 ng at this step. Poly-A RNA was fragmented with NEB's Magnesium Fragmentation Module for 3 minutes at 94°C, which was selected after optimizing for conditions for 200-500 bp fragments as determined by the Fragment Analyzer (Advanced Analytical). Fragmentation reactions were cleaned up with RNeasy columns. First strand synthesis was performed with SSIII (Invitrogen). Briefly, 1 μl Random Primer (3 μg/μl), 1 μl 10 mM dNTP mix, and 10 μl fragmented RNA were incubated at 65°C for 5–10 minutes and transferred to ice for 5 minutes. A mix of 4 μl 5X First Strand Synthesis (FSS) Buffer, 1 μl 0.1 M DTT, 1 μl Murine RNase Inhibitor, 1 μl 0.5 μg/μl Actinomycin D, and 1 μl SSIII (200 units) was added to the mixture of RNA, dNTPs, and Random Primers. This was incubated in the thermocycler at 25°C for 5 minutes to anneal the random primers, 50°C for 60 minutes to extend from the random primers, and 70°C for 15 minutes to inactivate SSIII. The reaction was cleaned up with AMPure beads using a ratio of 2.0x. For Second Strand Synthesis (SSS), a mixture of 10 mM each of dATP, dCTP, dGTP, and 20 mM of dUTP (instead of dTTP) was made. For a single reaction, 64 μl of cleaned FSS cDNA:RNA in Ultra Pure Water, was combined with 4 μl ACGU mix, 8 μl of NEB dNTP-free SSS Reaction buffer and 4 μl NEB SSS Enzyme mix from the SSS module. The reaction was incubated for 1 hour at 16°C, then cleaned with AMPure beads using a 1.0x ratio to begin eliminating DNA smaller than 200 bp, and quantified with the Qubit. There was typically 100-200 ng at this step. The double-stranded cDNA was then subject to End Repair (NEBNext), and cleaned with AMPure beads using a 0.9x ratio to deplete DNA $<$ 300 bp. The End-Repaired cDNA was then dA-tailed (NEBNext), and cleaned with AMPure beads using a 0.9x ratio. For adapter ligation, 38 μl of DNA was combined with 10 μl 5X NEBNext Quick Ligation Reaction Buffer, 1 μl NEB Adaptor, and 2 μl Quick T4 Ligase. The reaction was incubated at room temperature for 15 minutes, then cleaned with a 0.9x ratio of AMPure beads. The library was then size-selected with AMPure beads before proceeding to the PCR step. To obtain adapter-ligated fragments in the 300-600 bp range, the DNA was first incubated with 0.6x

AMPure beads by adding 60 μl AMPure beads to 100 μl DNA in UPW. The beads were pelleted on a magnet and the supernatant containing DNA smaller than approximately 600 bp was transferred to a new tube (DNA longer than 600 bp stayed on the beads). To make the final ratio 0.9x to select for DNA $>$ 300 bp on the beads, 30 μl more AMPure beads was added to the 160 μl supernatant. From there the AMPure clean-up proceeded as normal. USER enzyme digestion to cut the DNA at uracils (in the second strand and in the hairpin adapters) and PCR were then performed as follows: 20 μl of cleaned DNA was combined with 25 μl NEBNext High-Fidelity 2X PCR Master Mix and 3 μl NEBNext USER enzyme. This reaction was incubated for 15 minutes at 37°C to ensure uracil cutting occurs before addition of primers. Then 1 μl indexed primer (NEBNext) and 1 μl universal primer were added. The reaction was put in the thermocycler for 37°C for 15 minutes to ensure USER digestion went to completion, followed by 98°C for 30 seconds and 12 cycles of: 98°C for 10 seconds, 65°C for 30 secons, 72°C for 30 seconds. The PCR products at this stage are approximately 122 bp longer than the target insert size. We adjusted AMPure ratios accordingly for a final clean up and size-selection. The PCR reactions were cleaned with 0.85x AMPure beads to deplete DNA smaller than 350 bp. The DNA was eluted in 100 μl UPW and AMPure size selection was initiated by incubating with 55 μl AMPure beads (0.55x) to precipitate DNA longer than 650 bp onto the beads. The beads were pelleted on a magnet and the 155 μl of DNA shorter than 650 bp in the supernatant was transferred to a new tube where another 30 μl of AMPure beads was added for a final ratio of 0.85x. The AMPure procedure then continued as normal to obtain DNA $>$ 350 bp. The estimated insert sizes at this step was 230-530 bp. DNA samples were quantified with Qubit and purity was measure by NanoDrop. There was typically 600 ng at the end of this protocol. Fragment Analyzer (Advanced Analytical) traces suggested the mean estimated fragment sizes was around 400–420 bp putting the mean insert sizes near 300 bp. The samples were sequenced on a single lane of the Illumina HiSeq2000, combined with samples from another study, for 100 bp paired-end sequencing.

Transcriptome assembly, transcript alignment, and RNA-seq alignment

The salivary gland transcriptome was assembled from combining all replicates from all stages (15 samples total) using Trinity [Grabherr et al., 2011] following the recommended protocol for strand-specific data [Haas et al., 2013]. The command used was:

Trinity --seqType fq --JM 280G --left $L - - right$R --SS_lib_type RF --CPU 24 --trimmomatic --quality_trimming_params "ILLUMINACLIP:neb_primers.fa:2:30:10 LEADING:5 TRAILING:5 MINLEN:36" --bflyHeapSpaceMax 28G --bflyCPU 8 --bflyHeapSpaceInit 10G

Where L contained all mate1 files and R contained all mate2 files (in same order as L). Assembly statistics were obtained with the provided TrinityStats.pl script.
RNA-seq reads were aligned to the reference genome with HISAT2, sorted and indexed with SAMtools, and analyzed for coverage over each position in the genome with BEDtools:

hisat2 -p 8 --dta -x hisatidx -1 r1.fastq.gz -2 r2.fastq.gz--rna-strandness RF | samtools view -bSh - | samtools sort -T prefix > prefix.bam
samtools index prefix.bam
genomeCoverageBed -ibam prefix.bam -g sciara.genome -bg > prefix.bedGraph

Trinity-assembled transcripts were aligned to the reference with BLAST, the output of which was converted into BED format.

Other analyses

Visualization of genome tracks was done using the Integrative Genomics Viewer (IGV) [Robinson et al., 2011, Thorvaldsdóttir et al., 2013]. All other plotting was done in R.

4.7 Acknowledgements

Aisha Keown-Lang for help with slide preparations for FISH. Charles “Chip” Lawrence for discussions on hidden Markov Models.

Part III:

Development of genome-wide methods for studying DNA replication

Part III of this thesis covers my work on developing genome-wide approaches to studying DNA replication with a particular focus on metazoan systems, such as human cells, where origin mapping methods that work well in yeast have been less effective. The first chapter in this section is a review of genome-wide origin mapping techniques and datasets that we published in F1000 Prime Reports. That chapter will introduce you more broadly to DNA replication and describe some of the problems with the interpretation of these datasets. The second chapter in Part III then delves specifically into one of popular origin mapping techniques called Nascent Strand sequencing (NS-seq) that features the 5'-3' exonuclease λ-exo. The logic behind the method is that since λ-exo has little activity on RNA, parental DNA can be depleted while nascent strands remain protected from digestion due to the RNA primers at their 5' ends. Size-selection of 500-2500 bp DNA ensures the remaining RNA-primed nascent strands are centered on initiation sites. We found that λ-exo does not deplete DNA uniformly and that non-origin sequences throughout the genome may also be enriched by this technique. Therefore, we used a λ-exo control dataset to identify origin sequences throughout the genome with higher specificity. When controlling the NS-seq data in this way, interesting biological signals seem to become much clearer. In the third and final chapter of this section I discuss the progress that I have made in collaboration with others on developing novel genomic approaches to studying DNA replication using single-molecule technologies. Overall, the work I present in these chapters will lead the field of DNA replication, particularly for metazoans, in the direction of higher accuracy origin datasets from which to draw conclusions. These methods can be used in the future

to obtain higher resolution estimates of where DNA re-replication initiates inside the DNA puffs of Sciara coprophila.

CHAPTER 5

The hunt for origins of DNA replication in multicellular eukaryotes

John M. Urban[1], Michael S. Foulk[1,2], Cinzia Casella[1,3], and Susan A. Gerbi[1]

1 Division of Biology and Medicine, Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Sidney Frank Hall, 185 Meeting Street, Providence, RI 02912, USA
2 Department of Biology, Mercyhurst University, 501 East 38th Street, Erie, PA 16546, USA
3 Institute for Molecular Medicine, University of Southern Denmark, JB Winsloews Vej 25, 5000 Odense C, Denmark

This chapter is adapted from:

Urban JM, Foulk MS, Casella C, Gerbi SA. (2015) The hunt for origins of DNA replication in multicellular eukaryotes. F1000Prime Rep 7:30. PMC4371235

This manuscript was an invited review and was largely written by me with help from Dr. Susan Gerbi. Drs Foulk and Casella provided edits and helpful comments through multiple iterations.

5.1 Abstract

Origins of DNA replication (ORIs) occur at defined regions in the genome. Although DNA sequence defines the position of ORIs in budding yeast, the factors for ORI specification remain elusive in metazoa. Several methods have been used recently to map ORIs in metazoan genomes with the hope that features for ORI specification might emerge. These methods are reviewed here with analysis of their advantages and shortcomings. The various factors that may influence ORI selection for initiation of DNA replication are discussed.

5.2 Background

DNA synthesis initiates at multiple replication origins (ORIs) in eukaryotic genomes. When an ORI is used to initiate replication, it is said to have fired. Accurate duplication of the genetic material depends on a reliable mechanism that ensures that any given ORI fires at most once per cell cycle by restricting licensing to G1 phase and activation to S phase. In G1, each ORI that binds an origin recognition complex (ORC) and subsequently Cdc6 and Cdt1/MCM2-7 to form the pre-replication complex (pre-RC) is said to be licensed [Bell and Dutta, 2002, DePamphilis ML, 2006, DePamphilis and Bell, 2010]. Activation occurs when the pre-RC is converted to the initiation complex (IC) through the exit of Cdc6 and Cdt1 and the entry of Cdc45 [Zou and Stillman, 1998] and GINS to form the CMG complex (containing Cdc45, MCM2-7, and GINS) [Ilves et al., 2010, Costa et al., 2011, Gambus et al., 2011] that ultimately recruits DNA polymerase.

The classic studies of Huberman and Riggs [Huberman and Riggs, 1966, Huberman and Riggs, 1968] using DNA fiber autoradiography demonstrated many key points. DNA replication is bidirectional and starts at ORIs that sometimes fire coordinately in clusters. DNA fiber autoradiography allowed measurement of the rate of DNA replication fork progression. In its original form, this approach was unable to correlate the mapped ORIs with DNA sequence, and it remained a possibility that the ORIs were neither sequence-specific nor site-specific in the population of DNA molecules. In metazoans, the possibility that ORIs were not sequence-specific was supported by the observation that plasmid replication was independent of its sequence content [Krysan et al., 1989, Heinzel et al., 1991] and that any DNA sequence could replicate when injected into Xenopus embryos [Harland and Laskey, 1980]. Nonetheless, the site specificity of ORIs was demonstrated by investigations from the Hamlin laboratory that showed that a 6 kb restriction fragment of the amplified dihydrofolate reductase (DHFR) locus in Chinese hamster ovary cells was always the earliest one labeled by radioactive thymidine after release into S phase [Heintz and Hamlin, 1982, Burhans et al., 1986]. This ruled out the possibility that ORIs were totally random in the genome, although this might be the case in early embryos before the mid-blastula transition [Hyrien and Méchali, 1993, Blow et al., 2001, Hyrien et al., 1995].

Thus, the search for site-specific metazoan ORIs was invigorated by using a variety of methods at a few individual loci [DePamphilis, 1997]. A sequel to the earliest labeled restriction fragment approach was polymerase chain reaction (PCR) mapping of small nascent strands, which was applied to the DHFR locus [Vassilev et al., 1990, Pelizon et al., 1996, Kobayashi et al., 1998] and to other loci to map preferred start sites of DNA replication. In this same era, two-dimensional (2D) gels were developed and revealed a specific restriction fragment on a yeast plasmid where DNA replication starts [Brewer and Fangman, 1987, Huberman et al., 1987]. 2D gels were subsequently used to map ORIs in eukaryotic chromosomes. Neutral-neutral 2D gels took advantage of the differences in gel migration of restriction fragments containing a replication bubble or a replication fork [23]. Neutral-alkaline 2D gels were able to measure the direction of replication fork movement [Nawotka and

Huberman, 1988].With all of these pioneering studies showing that preferred regions of the genome were used as ORIs in both yeast and metazoa, the stage was set to elucidate what specifies eukaryotic ORIs. Most ORIs in budding yeast have been confidently mapped by a variety of methods [Siow et al., 2012] and are relatively well understood. However, metazoan ORIs have remained mysterious with regard to a universal principle, if any, that is shared by all. Initial experiments in metazoans studied a few individual ORIs [Aladjem, 2007, Hamlin et al., 2008, Masai et al., 2010]. Recent efforts have been made to map all ORIs in certain metazoan genomes (for example, Drosophila, mouse, and human), hoping to uncover global principles. This article provides an overview of the various current methods used to map ORIs genome-wide. Understanding the methods serves as the foundation to evaluate the conclusions from these studies regarding what features might define metazoan ORIs.

5.3 Methods to Map Origins

5.3.1 DNA combing and single-molecule analysis of replicating DNA

Current approaches to map ORIs tend to be improvements on earlier methods, such as the application of newer technologies to an older technique. For example, labeling with CldU and IdU and subsequent detection by fluorescence microscopy constitute a modern adaptation of DNA fiber autoradiography, which used H3-thymidine labeling. Both the older and newer approaches allow determination of replication fork rate, and pulse-chase labeling allows the ORI to be mapped at the center of the bidirectional fork pattern. The CldU/IdU-labeled DNA can be spread freehand in a technique called SMARD, short for Single-Molecule Analysis of Replicating DNA [Norio and Schildkraut, 2001], or with a DNA combing machine that ensures uniform spreading [Bensimon et al., 1994, Michalet et al., 1997, Herrick and Bensimon, 1999, Herrick and Bensimon, 2009]. DNA combing of evenly stretched DNA allows accurate genome-wide information to be obtained on replication fork speed, fork asymmetry, and the distance between ORIs that are used in the same S phase on a single DNA molecule [Bianco et al., 2012, Técher et al., 2013]. When DNA combing or SMARD is coupled with fluorescence in situ hybridization (FISH), it is possible to map the ORI at the locus defined by the hybridization probe. FISH has been combined with DNA combing to analyze replicating DNA in yeast [Pasero et al., 2002, Patel et al., 2006] and in mammalian samples [Anglana et al., 2003, Lebofsky et al., 2006]. However, the large size of mammalian genomes and the low frequency of finding replicating DNA molecules with the FISH signal make this a very time-consuming, though feasible, process. This problem has been minimized when coupling FISH with SMARD by a prior enrichment step to select the locus of interest by pulsed-field gel electrophoresis [Norio and Schildkraut, 2001]. Future developments will be required to allow analysis of spread metazoan DNA with regard to DNA sequence genome-wide rather than just at individual loci and to increase the resolution [1].

[1] One step toward this has recently been published [De Carli et al., 2016]. Moreover, I proposed another such technique to Oxford Nanopore Technologies in 2013 as an application for their MinION instrument, and was accepted into the first round of their early access program to pursue it.

5.3.2 Origin recognition complex chromatin immunoprecipitation

Several approaches have been taken to map ORIs genome-wide [MacAlpine and Bell, 2005, Gilbert, 2010, Hamlin et al., 2010]. Following the same trend as above, these often couple earlier approaches to map ORIs at individual loci with current genomics technologies. Chromatin immunoprecipitation (ChIP) to identify DNA sequences bound to ORC has been used in conjunction with DNA microarrays (ChIP-chip) or genomic sequencing (ChIP-seq) [Lubelsky et al., 2012]for yeast [Wyrick et al., 2001, Xu et al., 2006] and recently for Drosophila [MacAlpine et al., 2010] and human samples [Dellino et al., 2013]. Several difficulties face ORC-ChIP, especially for mammalian genomes that are 20 times larger than the Drosophila genome [Schepers and Papior, 2010]. As for any ChIP experiment, the quality of the data reflects the quality of the antibody; epitope tagged proteins can be used with an antibody against the tag to increase sensitivity and specificity but requires transformation [Lubelsky et al., 2012]. A challenge is that ORC has low sequence specificity and the ChIP signal is not very high compared with the background control, although CsCl gradient ultracentrifugation was recently used as a pre-enrichment step for ORC-bound DNA in a genome-wide study on human cells [Dellino et al., 2013]. An alternative enrichment approach is to biotinylate the protein of interest and then perform avidin purification of the biotinylated chromatin before doing ChIP [44]. Moreover, since ORC has other roles beyond DNA replication [Chesnokov, 2007, Duncker et al., 2009, Hemerly et al., 2009, Prasanth et al., 2010, Chakraborty et al., 2011], such as its localization to heterochromatin, a mixture of ORIs and other sequences might be enriched by ORC-ChIP alone. For higher specificity in identifying sites of pre-RC formation, MCM ChIP-seq can also be done to filter for ORC peaks that coincide with MCM peaks [Lubelsky et al., 2012, Xu et al., 2006]. However, it is known that ORC/Cdc6/Cdt1 loads many MCM double-hexamers onto DNA that can move away from the ORC site, raising the possibility of initiation sites that are not coincident with ORC [Hyrien et al., 2003, Blow et al., 2011] [2]. These potential ORC-distal MCM initiation sites would not meet the criterion of overlap with ORC sites; in some cases, the ORC sites loading the MCMs that may be involved in distal initiation might be excluded from the analysis if they do not overlap with MCMs. In addition, pre-RC ChIP studies identify licensed ORIs in the genome but do not indicate the subset that is chosen for activation [Santocanale and Diffley, 1996].

5.3.3 5-bromo-2'-deoxyuridine immunoprecipitation

Other methods have been employed to identify those ORIs that are active in metazoan genomes. Immunoprecipitation of nascent DNA strands labeled with 5-bromo-2'-deoxyuridine (BrdU) (BrIP-NS) has been coupled with microarrays (BrIP-NS-chip) to map active ORIs to the ENCODE 1% of the human genome [Karnani et al., 2010] or with deep sequencing (BrIP-NS-seq) to map active ORIs genome-wide [Mukhopadhyay et al., 2014]. In this technique, short, origin-centered BrdU-labeled nascent strands are first enriched by sucrose gradient ultracentrifugation and then further enriched by immunoprecipitation. Sucrose gradient size fractionation is used by some laboratories without

[2] In the time since this review was published, direct evidence was obtained supporting this hypothesis [Gros et al., 2015]

further enrichment steps and was the basis for an early microarray study to map ORIs [Lucas et al., 2007]. Another technique, Repli-seq [Hansen et al., 2010], enriches for BrdU-labeled DNA at consecutive time points throughout S-phase to identify the replication timing profile over the genome by tracking where replication forks tend to be at each time point. This information has been used to identify the likely regions where replication forks initiate [Hansen et al., 2010, Chen et al., 2010, Baker et al., 2012, Dellino et al., 2013]. The Repli-seq data are consistent with a computational method of calculating the nucleotide compositional skew profile over the genome, which also predicts the direction of replication forks genome-wide and the likely regions where they initiate ([Baker et al., 2012, Hyrien et al., 2013] and references therein; Olivier Hyrien, personal communication).

5.3.4 Bubble-trap

The bubble-trap method [Mesner et al., 2006, Mesner et al., 2009] is another approach to map active ORIs. It has some similarities to 2D gels since in both methods restriction fragments containing replication bubbles are retarded in their mobility on a gel. However, bubble trapping takes advantage of the circular nature of replication bubbles by allowing the matrix of a gel to form through them, thereby trapping bubble-containing fragments in the gel during electrophoresis. Visualizing blots of 2D gels can interrogate only a single restriction fragment for a bubble arc, whereas trapped bubble-containing restriction fragments are recovered from the gel for library preparation and can be coupled with DNA microarrays (Bubble-chip) [Mesner et al., 2011] or genomic sequencing (Bubble-seq) [Mesner et al., 2013] to identify bubble-containing restriction fragments genome-wide. It is possible that extrachromosomal circular DNAs (eccDNAs) [Dlaska et al., 2008, Cohen and Segal, 2009, Cohen et al., 2010, Shibata et al., 2012] that do not contain the specific restriction site will be trapped in the gel matrix too, but owing to their circular nature, they would not be cloned in the subsequent step. If future adaptations of bubble-seq bypass the cloning step by directly fragmenting and sequencing bubble-trapped material, eccDNA may become more of a concern, although it would still be possible to identify and eliminate eccDNA enrichments in downstream analytical steps by using paired-end sequencing and flagging enriched fragments with discordantly mapped reads consistent with circular DNA. In its current form, bubble trapping was shown to have very few false positives as demonstrated by 2D gel analysis [Mesner et al., 2006]. However, the current bubble-trap datasets cannot recover all ORIs in the genome since they interrogate the fragments of only a single restriction enzyme. Any given restriction enzyme will inherently disfavor the identification of certain ORIs that are too near a restriction site or that are in a small restriction fragment. Moreover, the resolution is limited to the sizes of the bubble-containing fragments of the single restriction enzyme. Constructing parallel bubble-seq libraries where each is derived from a different restriction enzyme could increase the sensitivity of this approach for ORI discovery and potentially allow higher-resolution inferences to map ORIs. It is possible, though, that the gain in information from additional restriction enzyme libraries is not enough to justify the increased workload and cost. Because each step in the bubble-trap process favors certain restriction fragment sizes, it is currently unclear whether the enrichment value of a particular fragment can accurately

be used to estimate ORI efficiency. Despite the conceptual elegance and high purity of bubble trapping, it has not been widely adopted perhaps because its many steps make it a time-consuming and technically challenging method.

5.3.5 Mapping the transition between leading and lagging nascent strands

There is a transition from leading (continuous) to lagging (discontinuous Okazaki fragment) strand synthesis at each ORI of bidirectional replication (Figure 1). The region of the transition from leading to lagging strand was mapped [Handeli et al., 1989] at the DHFR locus by using emetine, a protein synthesis inhibitor thought at the time to cause nucleosomes to segregate to only the leading strand. Micrococcal nuclease digestion of the assumed naked lagging strand was then used to map the strand-specific transition of the leading strand. However, shortly thereafter, the DePamphilis laboratory showed that emetine actually inhibits Okazaki fragment synthesis and not nucleosome segregation contrary to what was previously assumed [Burhans et al., 1991]. Nonetheless, the conclusion that the emetine approach could map the transition from leading to lagging strand synthesis remained intact. Another study mapped the transition from lagging to leading strand synthesis at the DHFR locus by using strand-specific dot blotting [Burhans et al., 1990], an adaptation of an earlier technique that mapped this transition with single nucleotide resolution in replication origins of animal viruses using sequencing gels [Hay and DePamphilis, 1982, Hendrickson et al., 1987]. The conceptual approach of mapping the transition point is also the basis for the more recent strand-specific sequencing of Okazaki fragments to identify ORIs and estimate ORI efficiencies genome-wide; this mapping technique holds great promise as seen by its application to the budding yeast genome, where isolation of Okazaki fragments was accomplished by DNA ligase I repression in a degron-tagged construct [Smith and Whitehouse, 2012,McGuffee et al., 2013]. Published data have yet to be released for deep sequencing Okazaki fragments from metazoan genomes, although preliminary results suggest that deep-sequenced Okazaki fragments in the human genome track replication forks and predict initiation regions in agreement with Repli-Seq timing and nucleotide composition skew profiles (Olivier Hyrien, personal communication [3]).

5.3.6 Lambda exonuclease enrichment of nascent strands

Mapping the transition point from continuous to discontinuous synthesis is also the basis of replication initiation point (RIP) mapping that has been used to map the start site of DNA synthesis to the nucleotide level [Gerbi and Bielinsky, 1997, Bielinsky and Gerbi, 1998, Gerbi et al., 1999, Gerbi, 2005, Das-Bradoo and Bielinsky, 2009] in specific ORIs from budding yeast [Bielinsky and Gerbi, 1998, Bielinsky and Gerbi, 1999], fission yeast [Gómez and Antequera, 1999], fungus gnats [Bielinsky et al., 2001], and humans [Abdurashidova et al., 2000]. The key to RIP mapping is to obtain enriched preparations of nascent DNA that are not contaminated by broken parental DNA. The inspiration for this approach came from studies in the DePamphilis laboratory that mapped the start

[3] The experiments showing this have been published since the release of this review [Petryk et al., 2016].

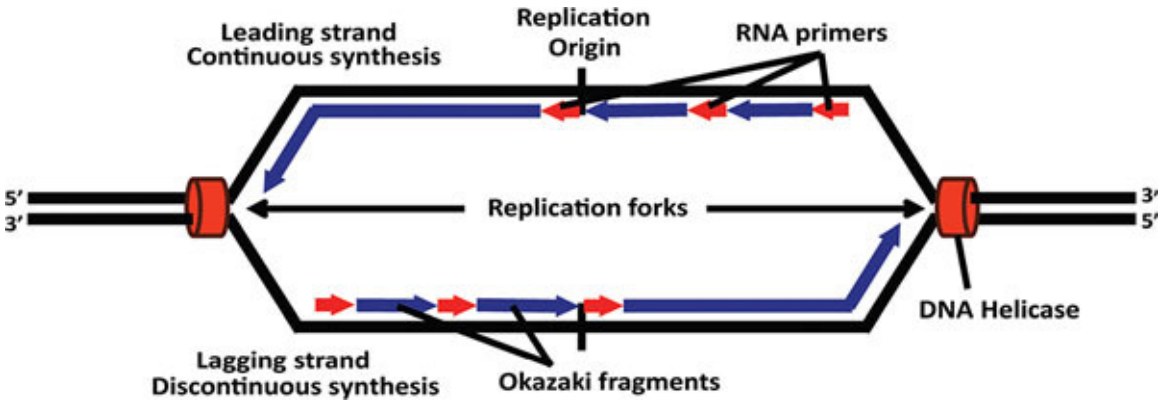


Figure 5.1: Origin of bidirectional replication
The transition between leading, continuous strand synthesis and lagging, discontinuous strand synthesis marks the origin of bidirectional replication. DNA polymerase can only extend nascent DNA in the 3' direction, and Okazaki fragments are used to allow net growth of the lagging strand in the 5' direction.

site of DNA synthesis with nucleotide resolution in SV40 and polyomavirus [Hay and DePamphilis, 1982, Hendrickson et al., 1987]. Those studies phosphorylated the 5' end of all DNA fragments (including parental DNA that was broken) and then identified nascent strands by removal of their 5' primer to expose a 5'-hydroxyl group that was then end-labeled with P^{32}. This approach worked well to map ORIs in viruses but did not have sufficient sensitivity to map ORIs to the nucleotide level in eukaryotic genomes. RIP mapping was able to retain single nucleotide resolution in eukaryotic genomes because of the increased sensitivity gained by employing the enzyme lambda exonuclease (λ-exo) that digests parental DNA from its 5' end [Radding, 1966, Little, 1967] but leaves intact the nascent DNA containing a 5' RNA primer. A more recent adaptation of RIP mapping bypasses λ-exo digestion and simply performs cell lysis in the well of an alkaline gel to minimize breakage of parental DNA before selection for small nascent strands [Romero and Lee, 2008]. Sequencing the primer-extended products from nascent DNA templates derived from an asynchronous population of cells allows the origin of bidirectional replication to be mapped as the transition from continuous to discontinuous synthesis.

The use of λ-exo to enrich nascent strands has become a preferred method for nascent strand purification, which has been used with PCR analysis of nascent strand abundance to map individual ORIs at a lower level of resolution than RIP mapping. λ-exo-enriched nascent strands (λ-exo-NS) also provide the foundation for genome-wide ORI mapping when coupled with DNA microarrays (λ-exo-NS-chip) or DNA sequencing (λ-exo-NS-seq) [Cayrou et al., 2012b]. λ-exo-NS-chip has been applied to Drosophila [Cayrou et al., 2011, Cayrou et al., 2012a], mouse [Cayrou et al., 2011, Cayrou et al., 2012a, Sequeira-Mendes et al., 2009], and human [Cadoret et al., 2008, Karnani et al., 2010, Valenzuela et al., 2011] genomes. λ-exo-NS-seq has been used to map ORIs in the human genome [Mukhopadhyay et al., 2014, Martin and Wang, 2011, Besnard et al., 2012, Foulk et al., 2015]

[4]. In these approaches, the transition point between continuous and discontinuous synthesis is not determined, but instead ORIs are mapped where short nascent strands are enriched. Nascent strands that are about 500 to 2500 nucleotides long (the size range depends on the laboratory) are used to avoid inclusion of 200 nt Okazaki fragment sequences that occur throughout the entire genome. Exclusion of Okazaki fragments is usually accomplished by the isolation of short single-stranded DNA through sucrose gradient fractionation. Alternatively, BND-cellulose column chromatography can be used to enrich replicative intermediates [Gerbi and Bielinsky, 1997, Bielinsky and Gerbi, 1998, Gerbi et al., 1999, Gerbi, 2005, Das-Bradoo and Bielinsky, 2009], with a size selection step later in the procedure [Foulk et al., 2015]. Depending on the computational analysis, λ-exo-NS-seq has the potential to give nucleotide-level resolution estimates for ORIs that map to fixed locations. It has also been used to gauge ORI efficiency, where an increased number of reads has suggested greater efficiency of early firing ORIs [Cayrou et al., 2011, Besnard et al., 2012, Picard et al., 2014]. However, λ-exo has base compositional and other biases that result in enriching various non-origin DNA sequences and influencing these ORI efficiency estimates in favor of GC-rich ORIs as discussed below.

5.4 Further Insights Into Genome-wide Origin Mapping

The ORI mapping results from the various methods described above are not in full agreement with one another, likely because most or all methods enrich both origin as well as non-origin DNA with various degrees of sensitivity and specificity. This makes it imperative that we understand the proclivities of each method to fully appreciate their outputs. Overall, it is to be expected that all origin mapping techniques will have some degree of overlap at least where origins are concerned. What are enigmatic are those putative ORIs unique to one method even after reaching saturation. Statistically significant but still incomplete overlap between two or more orthogonal datasets should not be taken as uniform validation of each individual putative ORI in those datasets. Although a significant number of overlaps between two approaches may increase our belief (in the Bayesian sense of the word) in the validity of putative ORIs that are represented in both approaches, it should not necessarily increase our belief in the validity of those ORIs absent from one approach. Conversely, some putative ORIs unique to one approach may be true positives, such as ORIs detected by nascent strand techniques that reside too close to a restriction site for bubble-trap detection. Thus, while studying only ORIs that overlap from several different origin-mapping techniques helps to highlight the most likely putative ORIs, it can also falsely discard true ORIs unique to one method. Therefore, perhaps meta-analyses that consider all datasets could instead use the degree of belief (e.g. posterior probability) of each potential ORI site in the genome to weight downstream analyses in favor of the most probable ORIs rather than allowing all putative ORIs to uniformly influence results and rather than discarding all putative ORIs with lower support. There is a possibility that some putative ORIs that overlap in orthogonal methods are not guaranteed to be true positives

[4] I am co-first author on this publication and it is a chapter in this thesis [Foulk et al., 2015]

if both orthogonal methods contain the same systematic false positives. However, this is likely to be rare and the weighting scheme above would account for this if additional orthogonal methods were included in the analysis. These possibilities highlight the need to better understand the outputs of each method to fine-tune our approaches and analyses to hunt for ORIs in metazoan genomes.

Despite the incomplete agreement between methods, comparing the outputs of orthogonal approaches is the best way to have an estimate on the reliability of a given method in ORI discovery. Bubble-trap was orthogonally validated by 2D gels and can be used to make orthogonal comparisons with nascent strand enrichments. SMARD with FISH, though of lower resolution, is orthogonal to both bubble-trapping and nascent strand techniques. Strand switch detection methods and ORC-ChIP are both orthogonal to SMARD, nascent strand enrichment techniques, bubble trap, and 2D gels, though ORC-ChIP requires making the assumption that ORC binding sites always flank or overlap initiation sites. Other comparisons are supportive but do not orthogonally validate the ORI status of a query selected from a genome-wide dataset. For example, quantitative PCR (qPCR) analysis of the abundance of λ-exo-enriched small nascent strands can be used to test the reproducibility of a λ-exo-based enrichment and could provide more confidence if the qPCR enrichment is present only in S phase, but since it uses λ-exo, is identical in preparation, and differs only at the readout stage, it is not orthogonal to λ-exo-NS-seq. Using λ-exo-NS and BrIP-NS techniques to validate each other is not strictly orthogonal either because both typically begin with the same sucrose gradient fractionation step to obtain short single-stranded DNA. Nonetheless, since one technique further enriches nascent strands based on enzymatic logic directed at RNA primers and the other based on immunoprecipitation targeted at incorporated BrdU in nascent DNA, their agreement can still be satisfying. Overall, orthogonal comparisons are important for both building confidence in ORI mapping techniques and in quantifying our degree of belief in putative ORI sites across the genome in order to paint a refined picture of the ORI landscape.

Previous λ-exo-NS-chip datasets did not have high overlap with other ORI mapping techniques such as Bubble-chip (10-14% of bubbles overlapped 26-35% of λ-exo-NS [Mesner et al., 2011]) and BrIP-NS-chip (2.2-12.8% BrIP overlapped 6.4-33.4% λ-exo-NS [Karnani et al., 2010, Cadoret et al., 2008]). λ-exo-NS-chip datasets did not even overlap well with each other (approximately 5-6% [Karnani et al., 2010, Cadoret et al., 2008]). Poor overlap was often attributed to a lack of saturation (that is, incomplete sets of ORIs), which was supported by a recent deep-sequencing λ-exo-NS-seq peak set that identified more than 350,000 putative ORIs across four different cell lines that encompassed 89-92% of most previous λ-exo-NS datasets (25.3% for one outlier) [Besnard et al., 2012]. Nonetheless, even after saturation, overlap of these λ-exo-NS-seq peaks with Bubble-chip was not as high (50-65% bubbles) and with BrIP-NS-chip was still low (9-20% BrIP-NS-chip) [Besnard et al., 2012]. More recent analyses of λ-exo-NS-seq datasets have shown that 45-46% of λ-exo-NS-seq peaks overlap with 36-37% of the bubble-seq ORI map [[Picard et al., 2014] and that 56.5% of the BrIP-NS-seq peaks overlapped 50.2% of the λ-exo-NS-seq peaks from the same study [Mukhopadhyay

et al., 2014]. Preliminary results from Okazaki fragment sequencing in the human genome had better agreement with results from bubble-trap than from λ-exo-NS methods (Olivier Hyrien, personal communication [5]). The Okazaki fragment results appear to be supported by their agreement with the bubble-trap method, which was shown to be highly specific [Mesner et al., 2009]. Conversely, more confidence can be given to bubble-trap since the Okazaki fragment sequencing profile was in agreement with Repli-Seq and nucleotide skew profiles. However, the lower concordance between Okazaki fragment sequencing and λ-exo-NS-seq supports the idea that non-origin peaks in addition to origins are systematically enriched in λ-exo-NS-seq datasets.

Data from single-molecule studies [Perkins et al., 2003, van Oijen et al., 2003, Conroy et al., 2010] and from deep sequencing of λ-exo-digested genomic DNA isolated from non-replicating G0 cells [Foulk et al., 2015] demonstrated that λ-exo is more efficient in digesting AT-rich DNA than GC-rich DNA and that this enriches GC-rich regions of the genome. Moreover, *in vitro* experiments and genomic analyses revealed that λ-exo digestion is obstructed by G-quadruplexes (G4s) [Foulk et al., 2015], reminiscent of exonuclease 1 digestion [Yao et al., 2007] and DNA polymerase elongation [Han et al., 1999, Lemarteleur et al., 2004], both of which are also impeded by G4s and are used as diagnostics to detect G4 formation *in vitro*. Therefore, λ-exo-NS-seq may also be enriched for G4-protected and GC-rich DNA independent of nascent strands, and efficiency estimates may be highly dependent on base composition. There may also be other nascent strand independent (NSI) λ-exo biases. For instance, since parental DNA contains ribonucleotides, with potentially as many as one million ribonucleotide insertions per genome duplication [Potenski and Klein, 2014], λ-exo digestion could be obstructed when it reaches a ribonucleotide in the parental DNA in accordance with its inability to digest RNA-protected DNA. Whether or not *in vivo* ribonucleotide incorporation events are randomly distributed or occur at hotspots is unknown, but the latter seems to be the case *in vitro* [Potenski and Klein, 2014]. λ-exo will also not digest eccDNA, such as "microDNA" circles that have been described in mammalian cells and range in size with a peak at 200 to 400 bp [Shibata et al., 2012]. eccDNAs of all sizes appear to be generated from non-random genomic sites such as tandem repeats [Dlaska et al., 2008, Cohen and Segal, 2009, Cohen et al., 2010]. MicroDNA sequences were enriched in CpG islands (CGIs) and GC-richness, as well as at the 5' end of genes [Shibata et al., 2012]. In general, non-random genomic sites associated with ribonucleotide insertions, eccDNA, G4s, and GC-rich DNA could pose as ORIs in λ-exo-NS-seq datasets and could potentially be more highly reproducible than true ORIs, which are plastic and stochastic.

In contrast to the evidence of pervasive, reproducible NSI λ-exo biases [Foulk et al., 2015], other λ-exo-NS studies that looked at non-replicating DNA (mitotic NS) as well as RNase-treated nascent strands concluded that there were no enrichments independent of nascent strands [Cayrou et al., 2011, Cayrou et al., 2012b]. Why one NS-λ-exo study [Cayrou et al., 2011], which looked at mitotic

[5] Now published [Petryk et al., 2016]

NS by microarray came to different conclusions on nascent strand independent enrichments of λ-exo is somewhat enigmatic. However, finding no enrichment after performing qPCR [Cayrou et al., 2012b] on sites of known origins in λ-exo-treated non-replicating DNA does not qualify as ruling out nascent strand independent enrichments in genome-wide datasets; it only supports the existence of nascent strand dependent enrichments. NSI λ-exo biases by definition enrich non-origin sites, which would still be enriched in non-replicating DNA and RNase controls. Despite the inconsistency between these studies, it still seems plausible that NSI λ-exo biases affect all previous λ-exo-NS studies at least to some extent. Indeed, 47% [Foulk et al., 2015], 72% [Cadoret et al., 2008],74-77% [Martin and Wang, 2011] and 70-94% [Besnard et al., 2012] of the λ-exo-NS peaks from published datasets overlap peaks from λ-exo-enriched non-replicating DNA from G0-synchronized cells ([Foulk et al., 2015] and Gerbi lab, unpublished data]) indicating that NSI λ-exo biases may permeate λ-exo-NS-seq datasets or that these datasets preferentially enrich origins in regions favorable to λ-exo enrichment. The former possibility (non-origin enrichments from NSI λ-exo biases in λ-exo-NS datasets) might explain why preliminary Okazaki fragment sequencing results in the human genome agreed better with Bubble-seq than λ-exo-NS-seq (Olivier Hyrien, personal communication [6]). NSI λ-exo biases might also explain the apparent discrepancy between the original Okazaki fragment sequencing studies in yeast [Smith and Whitehouse, 2012, McGuffee et al., 2013] and a more recent study that used λ-exo to enrich Okazaki fragments before sequencing [Yang et al., 2013].

It is desirable to overcome the aforementioned concerns that can result in λ-exo-NS DNA preparations being a mixture of true positives (nascent strands) and systematic false positives (NSI λ-exo enrichments). One possibility to overcome G4 protection and GC-rich λ-exo biases could be to use very high enzyme-to-DNA ratios to completely eliminate the parental DNA, leaving intact only RNA-protected nascent strands. However, 50 units/μg at pH 9.4 was shown to be insufficient to completely digest G4 structures in plasmid DNA [Foulk et al., 2015], yet higher λ-exo-to-DNA ratios have been shown to sacrifice the specificity of λ-exo and lead to the digestion of RNA-primed DNA [Yang et al., 2013] [7]. When 1 μg of DNA was mixed with 50 ng of RNA-primed DNA oligos, the RNA primers conferred protection against λ-exo digestion (10 units/μg DNA), but when 100 units/μg DNA was used to digest 50 ng of DNA with 50 ng of RNA-primed DNA, the RNA primers no longer conferred protection [Yang et al., 2013]. Therefore, starting the digestion with very high enzyme-to-DNA ratios should be avoided in order to preserve the RNA protection of nascent strands. This indicates that other ways are needed to overcome the NSI λ-exo biases.

Fortunately, there are several ways to improve λ-exo-NS-seq. Most previous NS-λ-exo studies have used genomic DNA to account for copy number variation in the genome and to control against biases introduced during library preparation and sequencing, but this does not control for

[6] Now published [Petryk et al., 2016]

[7] Also, more recent personal communication from John Yates

biases introduced by λ-exo. Thus, the detected enrichments can be summarized into three categories: (1) enrichments that arise from nascent strands alone (true positives), (2) enrichments that arise solely from NSI λ-exo biases (systematic false positives) and (3) enrichments that arise from some combination of both (true positives within NSI λ-exo-biased regions) [Foulk et al., 2015]. One needs a way to eliminate category 2 while retaining category 3 enrichments. Many of the NSI λ-exo biases can be overcome by using λ-exo-digested non-replicating DNA from G0 cells (λ-exo G0) as a control [Foulk et al., 2015], which simultaneously accounts for nucleotide composition biases, the G4 protection bias, and other possible issues related to λ-exo digestion while also controlling for copy number and biases introduced during library preparation and sequencing. Importantly, this approach corrects for the nascent strand independent enrichment components introduced by λ-exo across the genome by detecting origins as enrichments of nascent strands over the nascent strand independent λ-exo-digested background. Thus, with this approach, regions enriched from NSI λ-exo biases alone (category 2) no longer pose a problem, but it is still possible to discover origins in λ-exo-biased regions (for example, regions with G4-containing and GC-rich sequences) (category 3). Moreover, if both undigested and λ-exo-digested non-replicating genomic DNA controls are performed, it is also possible to compare the two peak sets that arise in λ-exo-NS-seq after separately applying each control to analyze what λ-exo-based enrichments are lost (or gained) when correcting for NSI λ-exo biases. λ-exo-digested non-replicating DNA will control against ribonucleotide incorporation and eccDNA hotspots too if the abundance and sites/sequences of these features are similar in S phase and G1/G0. Regardless, if paired-end sequencing is performed, then eccDNA hotspots can be identified in λ-exo-NS-seq data by searching for read pairs inside peaks that map in the wrong orientation (the signature of a circle would be outward- instead of inward-facing read pairs) or by removing such discordantly mapped reads before peak calling. Similarly, strand-specific sequencing could aid in identifying ribonucleotide incorporation hotspots and G4-protected DNA enrichments if these features lead to detectable imbalances in the number of reads mapped to Watson and Crick strands since origin-centered nascent strands should contain relatively balanced enrichments. Using emetine in conjunction with strand-specific sequencing could allow identification of leading strand transition points (similar to Okazaki-fragment sequencing that identifies lagging strand transitions), which provides a biological signature to use in addition to enrichment values. Thus, in the future, λ-exo-NS-seq experiments could have significant gains in specificity from novel strand-specific, paired-end strategies and analyses. Furthermore, future λ-exo-NS-seq studies might be able to circumnavigate the G4 bias altogether by changing the λ-exo digestion buffer from glycine-KOH (pH 9.4) to glycine-NaOH (pH 8.8) [Foulk et al., 2015]. This buffer simultaneously replaces K^+ with Na^+ and lowers the pH from 9.4 to 8.8, both of which were shown to lower the effect of G4s on λ-exo digestion [Foulk et al., 2015]. Controlling against NSI biases with the λ-exo G0 control, changing the buffer conditions to reduce G4 formation, and using strand-specific, paired-end sequencing strategies will reduce the number of systematic false positives to begin with, aid in identifying and eliminating persistent systematic false positives, and permit ORI efficiency estimates that are less influenced by base composition biases. Given that the potential issues in λ-exo-NS-seq can be solved, it is likely

to remain a powerful and popular technique in the coming years, although these issues may also motivate the innovation of new ORI mapping methods altogether.

5.5 What Features Define an Origin of Replication?

In contrast to the punctuate ORIs of budding yeast, metazoan initiation zones contain many potential ORIs, some of which are preferred (reviewed by [Aladjem and Fanning, 2004, Gilbert, 2005, Aladjem, 2007, Hamlin et al., 2008, Hamlin et al., 2010, Borowiec and Schildkraut, 2011]). These observations led to the Jesuit model (Many are called but few are chosen [DePamphilis, 1993]) that posited that although there may be multiple potential origins in a local genomic region (an initiation zone), only one or a few of those origins are activated in any given DNA molecule (Figure 2). If one ORI fires on a given DNA molecule in an initiation zone, its replication forks could inactivate other potential ORIs that are nearby, reflecting the phenomenon of origin interference [Brewer and Fangman, 1993, Brewer and Fangman, 1994, Marahrens and Stillman, 1994, Santocanale et al., 1999, Vujcic et al., 1999]. A well-studied example is the 55 kb intergenic initiation zone of the DHFR locus [Vaughn et al., 1990], where 20% of the initiation events occur at two preferred sites called ori β and ori γ [Dijkwel and Hamlin, 1995, Dijkwel et al., 2002], although other studies concluded that most initiation events occur at ori β [Burhans et al., 1990, Vassilev et al., 1990, Pelizon et al., 1996]. Even when ori β is deleted, initiation still occurs within the 55 kb initiation zone [Dijkwel et al., 2002]. Moreover, deletion of 45 kb of the 55 kb intergenic initiation zone still allowed initiation from the remaining 10 kb [Kalejta et al., 1998]. These observations reinforce the notion that there are multiple potential initiation sites within the broad initiation zone. Nonetheless, some metazoan ORIs are more tightly circumscribed, such as at the human lamin B2 [Giacca et al., 1994, Abdurashidova et al., 2000, Paixão et al., 2004] and β-globin [Wang et al., 2004] loci. Moreover, the width of the initiation zone can change during development, as observed for the 8 to 9 kb initiation zone of the II/9A locus from the fly Sciara that shrinks to 1 to 2 kb during locus-specific re-replication during DNA puff amplification [Lunyak et al., 2002].

Regardless of whether the initiation zone is broad or narrow, the question remains of what features specify the potential ORIs in these regions. Since there are limitations and biases in various methods to map ORIs, the interpretations of results are intimately linked to the methods used. What conclusions have these ORI mapping studies come to about ORI specification? Recurring themes in genome-wide replication origin studies have been proximity to genes and gene promoters, transcription of nearby genes, open versus compact chromatin, GC and AT content, CGIs, and G4 structures.

5.5.1 Metazoan Origins, Genes, and Transcription

ORC binding might be random in early Xenopus embryos, where ORC has a periodic spacing of 9 to 12 kb that appears to be DNA sequence independent and governed by an unknown mechanism

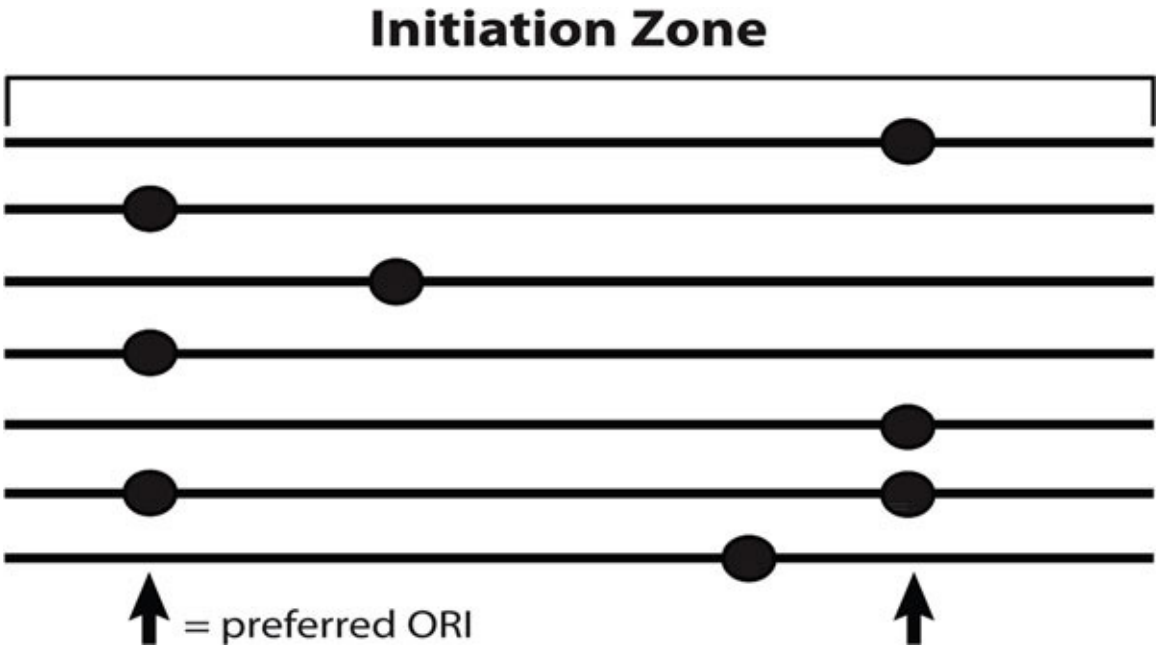


Figure 5.2: Multiple potential origins in an initiation zone

Each line represents an individual DNA molecule spanning the region of an initiation zone and each black oval within a line represents an origin that fired within that initiation zone on that molecule. There are many potential origins of DNA replication (ORIs) in an initiation zone for metazoan DNA replication, but oftentimes only one will be used on a given DNA molecule. Some of the potential ORIs are used more often than others, leading to preferred initiation sites of DNA synthesis in the initiation zone.

[Hyrien and Méchali, 1993, Blow et al., 2001]. However, after the mid-blastula transition and the onset of zygotic transcription, ORIs are found in defined locations [Hyrien et al., 1995, Sasaki et al., 1999]. In addition, deletion of the DHFR promoter to down-regulate DHFR transcription reduced ORI efficiency in the downstream intergenic initiation zone [Saha et al., 2004]. These observations suggest a link between transcription and ORI specification and efficiency. ORIs were thought to be absent in transcribed regions [Maric and Prioleau, 2010, Martin and Wang, 2011], and previous studies showed that ORC was found in intergenic regions, often in promoters [MacAlpine et al., 2004, Dellino et al., 2013], suggesting that ORC cannot bind or is displaced by RNA polymerase. However, an initial genomic study identified 28 new putative ORIs (using short nascent strands from sucrose gradient size fractionation), with the majority within genes and not intergenic [Lucas et al., 2007]. Bubble-chip results also suggested that ORIs were equally distributed across genic and intergenic restriction fragments and that many of the genes that overlapped ORIs were actively transcribed [Mesner et al., 2011]. Other results showed that 85% of ORIs were associated with transcription units with nearly half locating to promoters, and they suggested that transcription in early development might be linked to the regulation of ORI efficiency in later development [Sequeira-Mendes et al., 2009]. Subsequent studies also found that ORIs were significantly enriched near transcription starts sites as well as RNA Pol II and transcription factor binding sites [Cadoret et al., 2008, Karnani et al., 2010, Cayrou et al., 2011, Valenzuela et al., 2011, Dellino et al., 2013]. Martin and colleagues [Martin and Wang, 2011] expanded on this finding showing that ORIs were specifically enriched at moderately expressed genes rather than at lowly or highly expressed ones. Conversely, other studies did not find strong links of ORIs with genes and transcription. Bubble-seq results [Mesner et al., 2013], in contrast to bubble-chip results [Mesner et al., 2011], suggested that ORIs only marginally associated with transcriptionally active genes. Other nascent strand studies reported that relatively few ORIs were located near transcription start sites [Mukhopadhyay et al., 2014, Besnard et al., 2012] and concluded that the association between ORIs, promoters, and CGIs is independent of transcription [Mukhopadhyay et al., 2014]. Cadoret and colleagues [Cadoret et al., 2008] suggested that there might be no link between ORIs and gene regulation since half the ORIs found were not associated with open chromatin. Moreover, lack of correlation between transcription and ORI selection is evident on the transcriptionally inactive X chromosome [Mukhopadhyay et al., 2014], where the same ORIs are used as on the active X [Rowntree and Lee, 2006] and only their timing [Gómez and Brockdorff, 2004] and replication program for the order of ORI firing [Koren and McCarroll, 2014] differ. Earlier plasmid experiments revealed that the presence of the transcription complex is sufficient to specify ORI location even in the absence of active transcription [Danis et al., 2004]. Overall, in terms of ORI specification, it seems that the replication program interacts with the transcription program but is not dictated by it.

5.5.2 Metazoan origins and chromatin

Histone modifications may play a role in ORI control. ORC binds to regions of open chromatin in Drosophila [MacAlpine et al., 2010, Lubelsky et al., 2014], and in mammals the DHFR initiation zone

exhibits low nucleosome density [Lubelsky et al., 2011]. However, genome-wide studies have shown that human ORIs are distributed across regions of both open and closed chromatin [Cadoret et al., 2008,Martin and Wang, 2011,Mesner et al., 2013,Mukhopadhyay et al., 2014]. One sign of open chromatin is acetylation. Histones at ORC binding sites are hyperacetylated in Drosophila, and tethering a histone acetyl transferase (HAT) increases ORI activity [Aggarwal and Calvi, 2004]. Moreover, an increase in histone acetylation correlates with ORI activation [Liu et al., 2012]. HBO1 is a HAT that has been found to be associated with pre-RC components such as MCM2 and ORC1 [Iizuka and Stillman, 1999, Burke et al., 2001]. Moreover, H4K12ac is dependent on ORC. Overall, histone acetylation is not unique to the active amplicons in Drosophila follicle cells, but there is a quantitative relationship between the level of hyperacetylation and amplicon origin activity [Aggarwal and Calvi, 2004, Hartl et al., 2007, Kim et al., 2011].

Methylation of histone H4 at lysine 20 (H4K20me) may play a role in the control of ORIs [Dorn and Cook, 2011], and ORC recruitment appears to be enhanced by methylation of H4K20 by the histone methyltransferase PR-Set7 [Sherstyuk et al., 2014]. A recent λ-exo-NS-seq analysis also found a potential role for H4K20me1 as well as H3K27me3 at ORIs [Picard et al., 2014]. This is supported by a recent integrative genomics analysis of chromatin marks across the rDNA repeat [Zentner et al., 2011], where these two marks occur over the sites where rDNA replication initiation activity occurs ([Coffman et al., 2005, Foulk et al., 2015] and references therein). There have been reports for and against a significant association of ORIs with H3K4me3 [Cadoret et al., 2008, Karnani et al., 2010, Valenzuela et al., 2011, Martin and Wang, 2011], although most support this association with the caveat that it is not present at all ORIs and perhaps mostly early ones. In addition, different ORIs at the chicken β-globin locus have different patterns of histone modifications [Prioleau et al., 2003], suggesting further complexities. In general, conclusions have been that chromatin accessibility appears to be suitable for ORI formation, but is not necessary for it [Mukhopadhyay et al., 2014], and that no single chromatin mark studied to date can guarantee the presence of an ORI (nor vice versa).

5.5.3 Metazoan origins and DNA sequence elements (G4s, CpG islands, and GC-rich DNA)

Although an AT-rich consensus sequence motif was found in budding yeast ORIs [Newlon and Theis, 1993, Theis and Newlon, 1997, Breier et al., 2004, Dhar et al., 2012], the hunt for such a motif in metazoan ORIs has been less successful. Any DNA sequence injected into Xenopus embryos could replicate [Harland and Laskey, 1980], and plasmid replication was dependent on size rather than sequence [Krysan et al., 1989, Heinzel et al., 1991]. In addition, it was shown that ORC binds most DNA equally well and has insufficient DNA sequence binding specificity to be responsible for ORI definition [Vashee et al., 2003, Remus et al., 2004, Schaarschmidt et al., 2004], although ORC had a slight preference for AT-rich double-stranded DNA consistent with the prediction that metazoan ORIs would likely be AT-rich similar to most studied ORIs across the tree of life [Aladjem and

Fanning, 2004, Méchali, 2010, Leonard and Méchali, 2013]. Conclusions drawn from other experiments suggested that DNA topology rather than a conserved DNA sequence motif may be a more important determinant for ORC binding [Remus et al., 2004]. The consensus motif possibility was recently revived in the form of G4 motifs [Cayrou et al., 2012a, Besnard et al., 2012], which can form stable secondary structures. It was shown that ORC preferentially binds G4 structures in RNA and single-stranded DNA [Hoshina et al., 2013]. However, it needs to be emphasized that this preference was equivalent to AT-rich double-stranded DNA [Hoshina et al., 2013]. Most genome-wide λ-exo-NS studies have reported GC-rich ORIs [Cadoret et al., 2008, Cayrou et al., 2011, Cayrou et al., 2012a, Besnard et al., 2012], but Karnani and colleagues [59] reported AT-rich ORIs and Foulk and colleagues [Foulk et al., 2015] reported both AT- and GC-rich ORIs with mostly AT-rich ORIs after controlling for NSI λ-exo biases. Perhaps the slight preference of ORC binding for both AT-rich double-stranded DNA and G4 motifs as well as the mixed results of AT-rich and GC-rich ORIs reflect the possibility of both AT-rich and GC-rich ORIs in metazoans, as was recently seen for the yeast species Pichia pastoris [Liachko et al., 2014].

In general, a high correlation has been noted between λ-exo-NS peak (putative ORI) density and G4 motifs [Cayrou et al., 2012a, Besnard et al., 2012, Picard et al., 2014, Valton et al., 2014], CGIs [Cadoret et al., 2008, Sequeira-Mendes et al., 2009, Cayrou et al., 2011, Cayrou et al., 2012a, Martin and Wang, 2011, Besnard et al., 2012, Picard et al., 2014], and GC content [Cadoret et al., 2008, Cayrou et al., 2011, Cayrou et al., 2012a, Besnard et al., 2012]. In one study, nearly all (91.4%) λ-exo-NS-seq peaks overlapped G4 motifs [Besnard et al., 2012]. Moreover, in λ-exo-NS studies, G4-associated ORIs appear to be among the most efficient ORIs [Besnard et al., 2012, Picard et al., 2014, Valton et al., 2014], as do CGI-associated ORIs [Sequeira-Mendes et al., 2009, Martin and Wang, 2011, Cayrou et al., 2011, Picard et al., 2014], and GC-richness tends to peak near ORI centers with G4s and G-richness occurring most frequently within 500 bp 5' to λ-exo-NS enrichments [Cayrou et al., 2012a, Valton et al., 2014, Foulk et al., 2015]. In one λ-exo-NS study [160], point mutations in origin-associated G4s that lowered the stability of the G4 also lowered the λ-exo enrichment level of the associated ORI and switching the strand the G4 was on changed the position of the λ-exo-enriched ORI signal to remain 3' of the G4. It was also seen that areas of higher G4 motif density have higher λ-exo-NS enrichments and peak densities [Besnard et al., 2012, Picard et al., 2014, Valton et al., 2014] and that 86% of λ-exo-NS peaks associated with CGI are constitutive ORIs (present in all or most cell lines studied). At least one λ-exo-NS study has claimed that the association with both CGI and genes is actually due to the association with G4s [99], which are correlated with those features. λ-exo-NS studies have also highlighted that ORIs with these three features (G4, CGI, and GC richness) are typically early ORIs and that ORI density is correlated with timing (early ORIs being most densely populated and strongest) [Cayrou et al., 2011, Cayrou et al., 2012a, Besnard et al., 2012]. However, a recent study demonstrated that significantly enriched regions of the genome from sequencing λ-exo-digested genomic DNA from non-replicating G0 cells were also associated with GC-richness, CGI, and G4s, suggesting that these observations in

λ-exo-NS studies could be explained, at least to some degree, by NSI λ-exo biases [Foulk et al., 2015].

Other techniques have been either supportive of or in conflict with the λ-exo-NS-seq results. Bubble-seq did not support the conclusion that G4s are necessary for all ORIs [Mesner et al., 2013]. The majority (59%) of bubble-containing fragments did not overlap G4s, and the number that did was only 1.05-fold enriched over the number expected at random [68]. Nonetheless, it is possible that G4s are important at the subset (41%) of ORIs that do overlap G4s. A small proportion (9.6%) of Bubble-seq enriched fragments [Mesner et al., 2013] overlapped 51.3% of CGIs. BrIP-NS-seq peaks significantly overlapped G4s, although the G4 overlap made up just 37.5% of the peaks [Mukhopadhyay et al., 2014], and 13.1% BrIP peaks overlapped 35.1% CGI. Approximately a third (34.1%) of ORC binding sites [Dellino et al., 2013][8] overlapped 2.8% of G4 motifs, and 30.6% of ORC sites overlapped 14.7% CGI. Moreover, it was concluded from nucleotide composition skew analysis that CpG-rich genes are over-represented at skew-predicted ORIs [Hyrien et al., 2013], and other λ-exo-independent methods have identified ORIs near CGI ([Delgado et al., 1998] and references therein). Given all the results from λ-exo-NS, bubble-trap, BrIP-NS, ORC-ChIP, and nucleotide skew studies, it is clear that G4s are likely near 34-41% of ORIs and that CGIs are near approximately 10% to 30% of ORIs (many of which are the same ones associated with G4s). It is also clear that estimates from recent λ-exo-NS studies stating that more than 70-90% of ORIs [Cayrou et al., 2012a, Besnard et al., 2012] are near G4s are inflated relative to other techniques. This inflation of G4 association may arise because λ-exo digestion is inefficient for GC-rich DNA [Perkins et al., 2003, van Oijen et al., 2003, Conroy et al., 2010, Foulk et al., 2015] and is impeded by G4 structures [Foulk et al., 2015] and thus may reflect a high abundance of NSI enrichments and/or preferential enrichment of G4-proximal origins in those datasets. Indeed, in a recent study, after controlling for NSI λ-exo biases, 35% of λ-exo-NS-seq peaks overlapped 6.8% of G4s [Foulk et al., 2015], which is more consistent with the BrIP, bubble, and ORC estimates. Moreover, Bubble-seq results suggested that ORI density was similar in early and late replicating regions [Mesner et al., 2013]. Thus, it is possible that the higher ORI densities in early S phase associated with λ-exo-NS studies are due to the combination of higher G4 and CGI occurrences in early replicating regions of the genome and NSI λ-exo biases.

Another interpretation for the prevalence and high efficiency of G4 associated ORIs in genome-wide studies is that ORI-proximal G4s might impede replication fork progression *in vivo*. This would leave the replication bubbles and nascent strands from G4-proximal ORIs over-represented in the population of all ORIs because of their longer half-life, which would inflate the detection sensitivity and apparent efficiency of this subset of G4-proximal ORIs. In fact, higher detection and higher apparent efficiency would be true for ORIs with ORI-proximal pausing of replication forks *in vivo* for any other reason as well, such as impediments incurred by other structures. Valton and

[8] A new ORC dataset also did not find a high degree of overlap with G4 motifs [Miotto et al., 2016]

colleagues [Valton et al., 2014] have presented some compelling arguments against this interpretation of the G4 association with many apparently efficient ORIs, but it remains a formal possibility. Replication fork pauses have been detected 9 to 35 kb away from human ORIs [Frum et al., 2008], and high abundances of 200 nucleotide pieces of apparently aborted nascent DNA from ORIs have been found [Gómez and Antequera, 2008] but neither of these will be present in nascent strand preparations after size selection (typically 500-2500 bp). For support of this concern, more evidence will be required of origin-proximal pausing of replication forks that affect nascent strands in the target size range.

It is interesting that the observation that 5' to 3' λ-exo digestion is both impeded by G4 structures and inefficient in digesting GC-rich DNA [100] explains quirks seen in λ-exo-NS-based results at least as well as other interpretations. This observation predicts that λ-exo will enrich G4-protected and GC-rich DNA. It predicts that λ-exo-NS data will correlate with G4 density and that G4s will be 5' to λ-exo-NS enrichments. It predicts that experimentally inverting a G4 to place it on the opposite strand will change the location of λ-exo-based enrichment to stay 3' to the G4. It predicts that the strongest enrichments will be GC-rich. It predicts that GC-rich features such as CGIs will be constitutively enriched in all human cell lines (explaining why constitutive ORIs are near CGIs). It predicts that deleting G4s will result in less or no enrichment and that lowering the stability of G4s will lower the level of enrichment. All in all, given these caveats and others discussed above pertaining to all genome-wide ORI mapping approaches, the conclusions drawn so far about characteristics of metazoan ORIs should be cautiously considered.

5.6 Future Outlooks

The genetic basis for ORI definition is suggested by many examples where ORI activity is retained by an ORI sequence put into an ectopic location in the genome [Aladjem et al., 1998, Liu et al., 2003, Paixão et al., 2004, Altman and Fanning, 2004, Guan et al., 2009]. Deletion of certain sequences reduces efficiency for the DHFR ori β [Altman and Fanning, 2001, Altman and Fanning, 2004, Gray et al., 2007], lamin B2 [Paixão et al., 2004], c-myc [Liu et al., 2003], and β-globin [Wang et al., 2004] ORIs. Similarly, two elements from the Drosophila chorion gene locus, ACE3 and ori β, together can direct re-replication (amplification) when inserted into other genomic locations [Lu et al., 2001, Zhang and Tower, 2004]. These observations raise the question of what features explain the basis for ORIs. As discussed above, the binding of metazoan ORC is more dependent on topology than on DNA sequences [Remus et al., 2004]. Could chromatin and higher order structure be determinants for ORI specification rather than ORC binding a specific sequence motif in metazoan genomes?

Moving forward, to continue harnessing the power of genomics for studying metazoan ORIs, a goal should be to refine and innovate origin-mapping techniques to hunt for ORIs more specifically. The challenges in genomics studies will be in overcoming systematic errors. Although the level

of false positives from random sampling error in a dataset can be controlled by setting a desirable false discovery rate or irreproducible discovery rate ([Landt et al., 2012] and references therein), these statistical procedures do not control the level of systematic false positives that arise from biases in the method used. There is also a need to use other approaches, in addition to overlap of peak coordinates, to determine if and where datasets agree since the overlap of peak coordinates can be affected by many factors in upstream analytical steps (for example, P value thresholds). With the ever-increasing number of genomic datasets and ORI maps, quantifying our degree of belief in the validity of each individual putative ORI site in the genome on the basis of the accumulation of evidence, rather than maintaining uniform confidence over all putative ORIs in a single dataset, will be of utmost importance. For example, a statistical score such as the Bayesian posterior probability that there is an ORI or ORI potential (given all trusted datasets) could be assigned to each base pair in the genome. With highly authenticated genome-wide maps of replication initiation sites and zones, clearer pictures on what defines different classes of metazoan origins may emerge.

Overall, the field is now poised to address several questions for regulation of DNA replication: (a) what defines where the potential ORIs are located, (b) what determines which potential ORIs will be activated, and (c) what determines when in S phase an ORI will be activated [29]. It is clear that ORIs occur at defined sites in the genome, but their specification could be determined by something other than primary DNA sequence. It also remains possible that we are observing apparent specificity, rather than actual specificity, where ORIs opportunistically occur in reproducibly available sites determined by other features and processes.

5.7 Abbreviations

2D, two-dimensional; bp, base pairs; BrdU, 5-bromo-2'-deoxyuridine; BrIP-NS, BrdU-immunoprecipitation enriched nascent strands; CGI, CpG island; ChIP, chromatin immunoprecipitation; DHFR, dihydrofolate reductase; eccDNA, extrachromosomal circular DNA; G4, G-quadruplex; λ-exo, lambda exonuclease; λ-exo-NS, λ-exo-enriched nascent strands; NSI, nascent strand-independent; ORC, origin recognition complex; ORI, origin of DNA replication; PCR, polymerase chain reaction; qPCR, quantitative polymerase chain reaction; RIP, replication initiation point.

5.8 Disclosures

The authors declare that they have no disclosures.

5.9 Acknowledgments

We gratefully acknowledge support from postdoctoral fellowship DOD W81XWH-11-1-0599 to Cinzia Casella and a fellowship to John M. Urban for work supported by the National Science Foundation Graduate Research Fellowship Program under grant DGE-1058262.

This review article is dedicated to Joyce Hamlin and Mel DePamphilis, two giants in the field of DNA replication, whose years of healthy debate of whether metazoan replication origins are broad initiation zones or punctate sites have enlivened numerous Cold Spring Harbor Laboratory DNA replication meetings.

CHAPTER 6

Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins

Michael S. Foulk[1,2], John M. Urban[1], Cinzia Casella[3], and Susan A. Gerbi

Brown University Division of Biology and Medicine, Department of Molecular Biology, Cell Biology and Biochemistry, Providence, Rhode Island 02912, USA

[1] These co-first authors have contributed equally to this work.

[2] Present Address: Mercyhurst University, Department of Biology, Erie, PA 16546, USA

[3] Present Address: Institute for Molecular Medicine, University of Southern Denmark, 5000 Odense C, Denmark.

Corresponding author: Susan_Gerbi@Brown.edu

This chapter is adapted from:

Foulk MS*, **Urban JM***, Casella C, Gerbi SA, (2015) Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. Genome Research 25:725-35. PMC4417120 (* signifies co-first author).

This manuscript was submitted to Genome Research on September 1, 2014 and was accepted on February 18, 2015 when it was published initially as an advanced online manuscript. Genome Research published it in print in May 2015. Cinzia Casella performed cell culture and FACS analysis. Michael Foulk performed the rest of the benchwork, including the plasmid experiments and analyses, nascent strand preparations, and preparing Illumina sequencing libraries. I conceived, designed code for (some of which is here: `https://github.com/JohnUrban/LexoNSseq2015`), and performed all genome-wide bioinformatics and data analyses, made the majority of figures and tables throughout the main text and supplement, and led this project in the direction of controlling for and characterizing enzymatic biases introduced when enriching origin-proximal nascent DNA with λ-exonuclease. Susan, Michael, and I wrote the manuscript.

6.1 Abstract

Nascent strand sequencing (NS-seq) is used to discover DNA replication origins genome-wide, allowing identification of features for their specification. NS-seq depends on the ability of lambda exonuclease (λ-exo) to efficiently digest parental DNA while leaving RNA-primer protected nascent strands intact. We used genomics and biochemical approaches to determine if λ-exo digests all parental DNA sequences equally. We report that λ-exo does not efficiently digest G-quadruplex (G4) structures in a plasmid. Moreover, λ-exo digestion of nonreplicating genomic DNA (LexoG0) enriches GC-rich DNA and G4 motifs genome-wide. We used LexoG0 data to control for nascent strand independent λ-exo biases in NS-seq and validated this approach at the rDNA locus. The λ-exo-controlled NS-seq peaks are not GC-rich, and only 35.5% overlap with 6.8% of all G4s, suggesting that G4s are not general determinants for origin specification but may play a role for a subset. Interestingly, we observed a periodic spacing of G4 motifs and nucleosomes around the peak summits, suggesting that G4s may position nucleosomes at this subset of origins. Finally, we demonstrate that use of Na^+ instead of K^+ in the λ-exo digestion buffer reduced the effect of G4s on λ-exo digestion and discuss ways to increase both the sensitivity and specificity of NS-seq.

6.2 Introduction

DNA replication is a highly regulated event whereby the genome is duplicated precisely once per cell cycle. In eukaryotes, nuclear DNA replication initiates at numerous origins of replication along linear chromosomes. What defines origins in metazoans remains unclear. Origins are AT-rich in bacteria and yeast [Méchali, 2010, Méchali et al., 2013] with a few exceptions [Xu et al., 2012, Liachko et al., 2014]. Similarly, many origins in metazoans have AT-rich elements [Aladjem and Fanning, 2004]. Thus, an attribute shared by most origins studied across the tree of life is the occurrence of AT-rich features. However, recent genome-wide studies have suggested that origins in multicellular eukaryotes may be GC-rich and correlated with motifs for G-quadruplex (G4) structures [Cayrou et al., 2011, Besnard et al., 2012, Picard et al., 2014]. Intrastrand G4 structures are highly stable DNA secondary structures that can form at physiological conditions *in vitro* when a DNA strand has four or more adjacent poly-G tracts typically defined as being separated by loops of 1–7 nucleotides (nt) [Huppert, 2010, Bochman et al., 2012].

Current insight into what defines metazoan origins suffers most from a small sample size of well-characterized origins. The search for sequence motifs, epigenetic marks, and other unifying features that specify metazoan replication origins has been ongoing for many years. Recently, several genome-wide approaches have been taken (for review, see [Gilbert, 2010, Urban et al., 2015b]) to increase the sample size of origins in an effort to finally resolve this issue. A popular method used to study metazoan origins, both genome-wide and in other applications, involves the enrichment of nascent strands by lambda exonuclease (λ-exo). λ-exo is a 5' to 3' DNA exonuclease that is used to deplete parental DNA, while nascent strands with 5' RNA primers are protected [Radding, 1966, Little, 1967] and become effectively enriched over the depleted parental DNA background. The λ-exo enrichment technique was originally developed to map the transition point from leading to lagging strand synthesis with single-nucleotide resolution in known origins [Gerbi and Bielinsky, 1997, Bielinsky and Gerbi, 1998] and has since been adopted to identify origins in metazoan genomes by pairing with microarrays (NS-chip) [Cadoret et al., 2008, Sequeira-Mendes et al., 2009, Karnani et al., 2010, Cayrou et al., 2011, Cayrou et al., 2012a, Valenzuela et al., 2011] and deep sequencing (NS-seq) [Martin and Wang, 2011, Besnard et al., 2012, Mukhopadhyay et al., 2014, Picard et al., 2014]. There are a few variations of nascent strand enrichment protocols that employ λ-exo, but the heart of each multistep procedure is the λ-exo enrichment step.

Due to the strong association of putative origins identified in NS-seq studies with predicted G4 motifs, it was proposed that G4s might have a function at origins [Cayrou et al., 2011, Besnard et al., 2012, Picard et al., 2014, Valton et al., 2014]. However, this interpretation of NS-seq data depends on the purity of the nascent strand preparation and the ability of λ-exo to efficiently digest G4 sequences in the contaminating parental DNA. Notably, single-molecule studies on λ-exo have shown that its digestion rate is dependent on base composition [Perkins et al., 2003, van Oijen et al., 2003, Conroy et al., 2010]. In particular, λ-exo was shown to digest GC-rich DNA less efficiently

and to pause at GC-rich motifs. Moreover, another exonuclease, Exo1, inefficiently digests G4s and is used as a diagnostic to detect G4s *in vitro* [Yao et al., 2007]. We hypothesized that λ-exo also inefficiently digests G4 structures that form under the conditions used to prepare nascent strands, leading to significant enrichment of G4-protected DNA, which may largely explain the association of putative origins from NS-seq studies with predicted G4 motifs. Moreover, we hypothesized that the inefficiency of λ-exo digestion of GC-rich DNA is generally responsible for the recent λ-exo-based observations that metazoan origins are GC-rich. We present (1) biochemical and genome-wide evidence supporting these hypotheses, (2) a new way to control λ-exo biases in NS-seq, and (3) a potential role of G4s near a subset of origins.

6.3 Results

6.3.1 G-quadruplexes are resistant to λ-exo digestion in a pH-dependent manner

We tested the ability of λ-exo to efficiently digest G4 structures by digesting a plasmid derived from the human MYC locus [Malott and Leffak, 1999] that contains a well-characterized G4 motif (Fig. 6.1A, see Pu27; [Brooks and Hurley, 2010]). Another putative G4 sequence (Pu30) was identified nearby on the opposite strand (Fig. 6.1A), using the QGRS mapper [Kikin et al., 2006]. The BglII linearized plasmid was 3' end-labeled, kept double-stranded (dsDNA) or made single-stranded (ssDNA), and digested overnight with λ-exo at pH 9.4, the optimal pH for the enzyme [Radding, 1966]. After λ-exo digestion of ssDNA, two prominent bands were observed (Fig. 6.1B) that correspond to the predicted size of fragments resulting from Pu27 (3260 bp) and Pu30 (4591 bp), impeding λ-exo digestion. These bands were also weakly detected after λ-exo digestion of dsDNA. Digesting plasmids from which Pu27 or Pu30 had been deleted confirmed that the two major bands were the result of the inability of λ-exo to digest these G4-containing sequences (Fig. 6.1C, cf. lanes 3, 8, and 13). When plasmids were digested at pH 8.8 (75%–80% λ-exo activity), the Pu27 band was faint in the wild-type plasmid and disappeared when Pu27 was deleted, while the Pu30 band was absent for all constructs (Fig. 6.1C, cf. lanes 2, 7, and 12). These data suggest that Pu27 forms a more stable G4 than Pu30 and that both G4s are less stable at the lower pH. Consequently, λ-exo digests these G4 sequences more efficiently at pH 8.8 despite having somewhat lower enzymatic activity (Fig. 6.1C, cf. lanes 2, 3).

6.3.2 Characterizing biases in λ-exo digestion genome-wide

Genome-wide biases of λ-exo were profiled by sequencing the DNA remaining after λ-exo digestion (pH 9.4) of sonicated genomic DNA (gDNA) from nonreplicating G0 MCF7 cells (LexoG0). We mapped 115.1, 174.6, and 153.7 million reads to the human genome from three biological replicates, resulting in a pooled total of 443.4 million mapped reads (Supplemental Table D.1). In addition, we sequenced undigested gDNA from nonreplicating G0 MCF7 cells (G0gDNA), which had 181.9 million

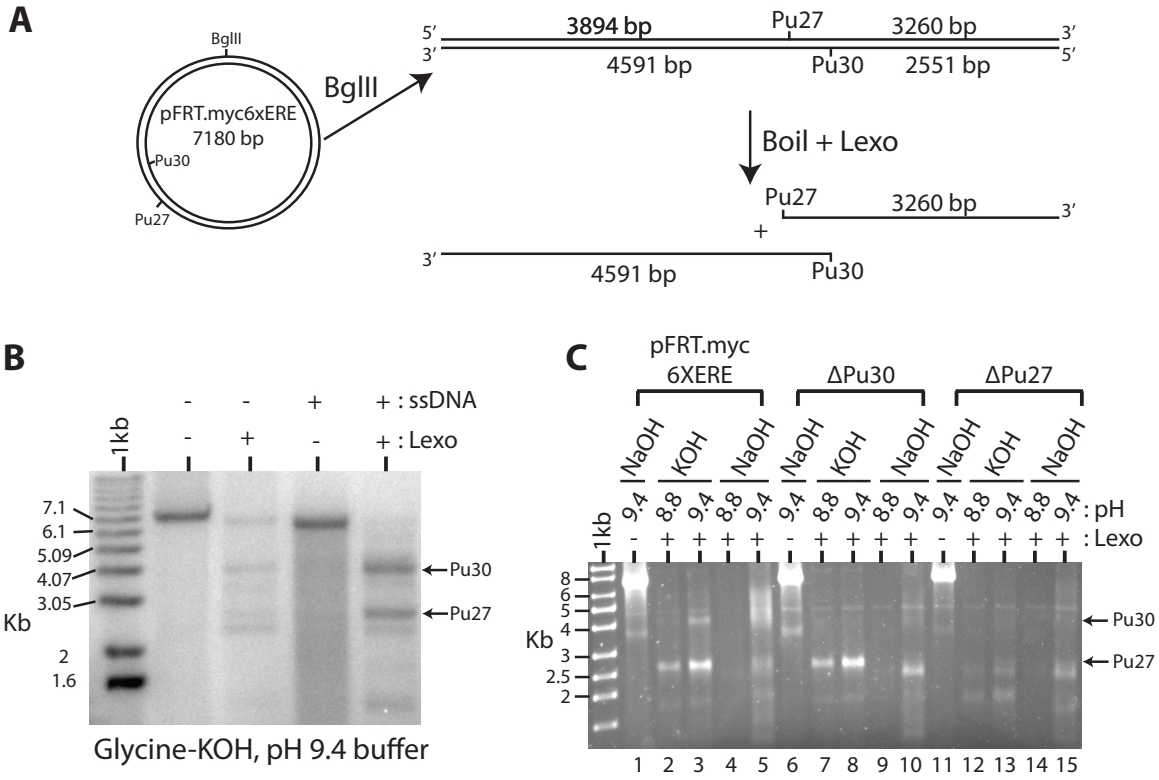


Figure 6.1: The MYC G-quadruplex (G4) impedes λ-exo digestion.
(A) Diagram of the plasmid digestion experiment. The predicted sizes of single-stranded fragments resulting from the inability of λ-exo to digest through the G4 motifs are shown. **(B)** Digestion of 3'-labeled BglII linearized pFRT.myc6xERE in glycine-KOH (pH 9.4). Double-stranded and single-stranded DNA were used as indicated. **(C)** Digestion of single-stranded DNA from pFRT.myc6xERE and deletion mutants of Pu30 and Pu27 in four different buffers. The pH of the buffer and the base used to titrate the pH are indicated.

mappable reads (Supplemental Table D.1). We identified regions of the genome enriched by λ-exo digestion of nonreplicating DNA by calling LexoG0 peaks relative to the G0gDNA control. The LexoG0 replicates were highly reproducible, and the 196,851 peaks derived from the pooled reads significantly encompassed the replicate peak sets (Supplemental Figures D.1A, D.2; Supplemental Tables D.2, D.3, D.4). The LexoG0 peak set from the pooled reads is referred to as LexoG0$_{G0gDNA}$ (following a "treatment$_{control}$" format) to distinguish it from the mappable reads prior to peak calling (LexoG0).

6.3.3 Nonreplicating genomic DNA digested with λ-exo (LexoG0) is enriched in GC-rich sequences and depleted for AT-rich sequences

If λ-exo uniformly digests gDNA, then the distribution of GC content over the mappable reads should be similar with and without λ-exo digestion. However, relative to undigested G0gDNA, AT-rich reads were depleted and GC-rich reads were enriched in all three LexoG0 replicates (Fig. 6.2A). Among the replicates, the median enrichment of GC-rich reads reached 14.2-fold and the median depletion of AT-rich reads reached 310.6-fold. These results indicate that λ-exo digests AT-rich sequences more efficiently than GC-rich sequences.

The LexoG0$_{G0gDNA}$ peak sequences were GC-rich with a single mode centered at 55% GC, while a mode centered at 39% GC is expected when the peaks are shuffled around the genome at random (Fig. 6.2C, cyan). There is a high correlation of LexoG0$_{G0gDNA}$ peak and CpG island densities in 1 Mb bins (Pearson's product-moment correlation coefficient, Pearson's r = 0.646; Spearman's rank-order correlation, Spearman's ρ = 0.746) (Supplemental Table D.6A), and 12% of LexoG0$_{G0gDNA}$ peaks overlap 92% of all CpG islands (Supplemental Table D.7). Strikingly, the –log10(P-value) signal from the LexoG0$_{G0gDNA}$ data set mimics the periodicity of 16 CpG island repeats over a 48 kb region of Chromosome 19 (Fig. 6.3A). Overall, the effect of base composition on λ-exo digestion rate results in enriching GC-rich regions of the genome.

6.3.4 Nonreplicating genomic DNA digested with λ-exo is enriched with telomere repeats and G4 sequences

The G4 counts per million reads (G4-CPMR) was calculated for the LexoG0 replicates and the G0gDNA control (Fig. 6.2E). The G4-CPMR for the G0gDNA reads was 2343, whereas the LexoG0 replicates ranged from 8,928–10,421 (3.8- to 4.5-fold enriched). Moreover, 36.9% of LexoG0$_{G0gDNA}$ peaks (15.5% expected at random) directly overlap with 48.5% of predicted G4 motifs ("G4s"; 8.6% expected at random) (Fig. 6.3D; Supplemental Table D.7). The average LexoG0$_{G0gDNA}$ fold enrichment signal was highly correlated with G4 density in 100-kb bins (Pearson's r = 0.862, Spearman's ρ = 0.776) (Supplemental Table D.6B), as were the densities of peaks and G4 motifs (Pearson's r = 0.704, Spearman's ρ = 0.704) (Fig. 6.3; Supplemental Table D.6A), as seen in whole chromosomes (Fig. 6.3 B, C). Additionally, G4s are increasingly enriched with proximity to LexoG0 peak summits,

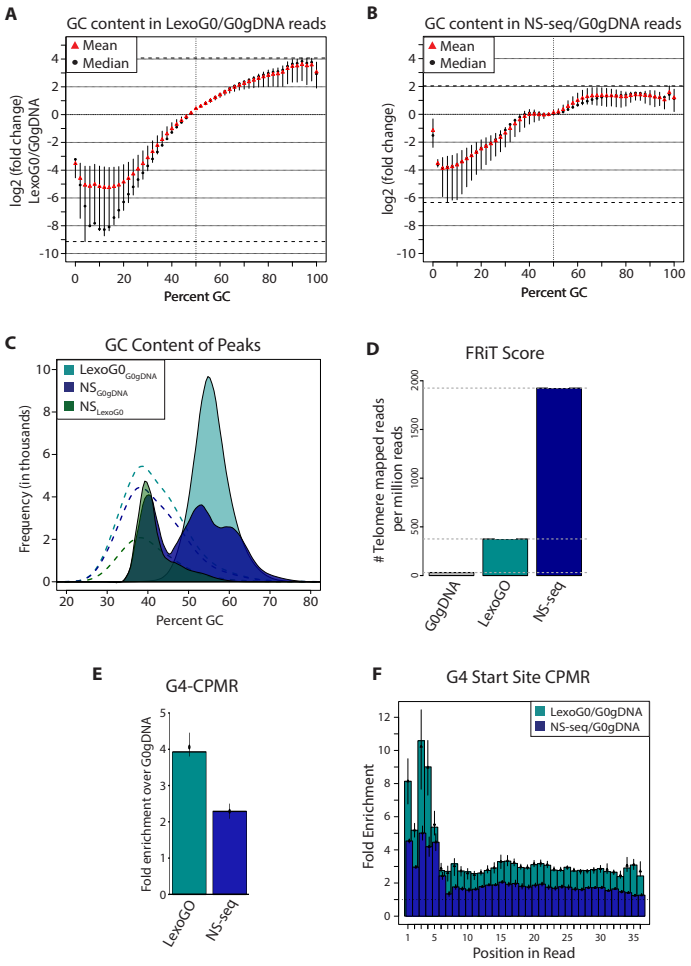


Figure 6.2: λ-exo digestion enriches GC-rich and G4-containing sequences.
(A) Log2(fold change) of the distribution of GC content in LexoG0 reads relative to that of G0gDNA reads. Over each GC%, the minimum to maximum (line segment), median (black dot), and mean (red triangle) for the replicates are shown. The dotted lines at the top and bottom represent the absolute maximum enrichment and depletion values found among the replicates. **(B)** Log2(fold change) of the distribution of GC content in NS-seq reads relative to that of G0gDNA reads. Other details as in panel A. **(C)** GC content of peaks for LexoG0$_{G0gDNA}$, NS$_{G0gDNA}$, and NS$_{LexoG0}$. Dashed lines show the distribution of GC content in randomly shuffled peaks. **(D)** Fraction of Reads in Telomeres (FRiT) scores. **(E)** G4-CPMR fold enrichment of the NS-seq and LexoG0 replicates over G0gDNA. The median (bar height), the minimum to maximum values (vertical line), and mean (black dot) for the replicates are shown. **(F)** Fold enrichment of G4-start-site-CPMR over positions 136 in 50-bp reads. Over each position, the minimum to maximum (vertical line), median (bar height), and mean (black triangle) of the replicates are shown.

suggesting that G4 structures are enriched by λ-exo (Fig. 6.4A).

To discount the possibility that G4 motifs were enriched simply due to the GC-rich nature of DNA after λ-exo digestion, we performed the following analyses. Human telomeres are composed of a repeat sequence (TTAGGG) that strongly favors formation of G4 structures *in vitro* [Huppert, 2010]. The number of mappable reads per million reads that remapped to a 6 kb telomere repeat sequence, the fraction of reads in telomeres (FRiT), was calculated. Undigested G0gDNA contained relatively few telomere reads, with a FRiT score of 30.62 (Fig. 6.2D). Conversely, the LexoG0 FRiT score of 375.67 was over 12-fold higher (Fig. 6.2D). This large enrichment cannot be explained by the telomeric GC content of 50%, which is associated with a near neutral λ-exo enrichment (Fig. 6.2A), leaving the propensity of telomeres to fold into G4 structures *in vitro* as a likely contributor to this effect.

In the LexoG0 preparation, the distribution of 5' DNA ends initially corresponds to where 5' to 3' λ-exo digestion stopped. Subsequently, after fragmentation for library construction, there is a mixture of λ-exodigested 5' ends, as well as 5' ends from breaks. Since fragment ends are sequenced from 5' to 3', this mixture is directly reflected in the Illumina read sequences. In the undigested G0gDNA control, the 5' end sequences correspond only to breaks from fragmentation. If λ-exo is impeded by G4 structures during digestion, then it would be detectable as an enrichment of G4 motif start sites concentrated at the 5' end of LexoG0 reads compared with undigested G0gDNA reads. Thus, we calculated the G4-start-site counts per million reads (G4-start-site CPMR) over each position from 1 to 36, where position 36 is the last position in a 50-bp read that a G4 motif ($G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$) [Huppert and Balasubramanian, 2005] can start. Relative to the G0gDNA control, the first 5 bp of the read profiles in all three LexoG0 replicates are much more enriched in G4 start sites than the remaining 3' end, which exhibits a uniform enrichment pattern (Fig. 6.2F).

Since λ-exo is a 5' to 3' directed exonuclease and digested fragments of 500–1500 bp were size-selected, the following would hold true if G4 structures impede λ-exo: (1) G4-protected DNA would extend 500–1500 bp 3' to the G4-protected 5' end; (2) the highest sequencing coverage would be within the first 500 bp 3' of the genomic location of each protective G4; and (3) G4 motifs would be enriched within 500 bp 5 of peak summits in an aggregate analysis of all λ-exo-enriched peaks. To test this, LexoG0$_{G0gDNA}$ peaks were aligned by their summits, and the distribution of predicted G4 motifs around the summits was plotted while preserving the strand information of whether a G4 motif was 5' or 3' to the peak summit. Figure 6.4B shows that G4s mapped preferentially within 500 bp 5' to LexoG0$_{G0gDNA}$ peak summits compared with randomly shuffled G4 motifs. Taken together, the results of the plasmid experiments, the enrichment of telomere repeats in LexoG0 reads, the enrichment of G4s at the 5 ends of LexoG0 reads, and the enrichment of G4s 5' to LexoG0$_{G0gDNA}$ peak summits strongly support the conclusion that λ-exo is impeded by G4 structures *in vitro*, in

addition to inefficiently digesting GC-rich DNA.

6.3.5 Regions of the genome enriched by λ-exo in replicating DNA (NS-seq) are many of the same regions enriched in nonreplicating DNA, but also include a distinct set of AT-rich regions

Three biological replicates of NS-seq were prepared from replicating MCF7 cells. BND cellulose-enriched replicative intermediate DNA was λ-exodigested at pH 8.8 to ensure the preservation of RNA primers [Li and Breaker, 1999] and to decrease the stability of G4s (Fig. 6.1). Peaks were called relative to the undigested G0gDNA control. The replicates were highly reproducible and the 162,098 peaks obtained by using the pooled set of reads were highly representative of the three replicates as measured by correlation and overlap (Supplemental Figures D.1B, D.2; Supplemental Tables D.1, D.2, D.3, D.5). Thus, the peak set from pooled reads (named NS_{G0gDNA} to distinguish it from the NS-seq reads) was used for subsequent analyses. Although more than half of the peaks were unique to NS_{G0gDNA}, a substantial subset (47%) (Supplemental Table D.7) overlapped with $\mathrm{LexoG0}_{G0gDNA}$ peaks, indicating that a large proportion may arise from nascent strand-independent λ-exo enrichment.

Characteristics of the NS-seq data that distinguish it from the LexoG0 data could be attributed to the presence of λ-exo-resistant DNA unique to the replicating cell population such as RNA-protected nascent strands. AT-rich reads were depleted and GC-rich reads were enriched in NS-seq compared to undigested G0gDNA, but each to a lesser extent than in LexoG0 reads (Fig. 6.2B). The GC content of the NS_{G0gDNA} peaks (Fig. 6.2C, blue) displayed a bimodal distribution with one GC-rich mode (53%–60% GC), similar to $\mathrm{LexoG0}_{G0gDNA}$ peaks, and one AT-rich mode (40% GC) not present in $\mathrm{LexoG0}_{G0gDNA}$ peaks. Given the strong depletion of AT-rich reads after λ-exo digestion in both LexoG0 and NS-seq (Fig. 6.2A,B), it was surprising to find λ-exo-enriched AT-rich peaks in NS_{G0gDNA}, indicating that λ-exo-resistant DNA emanates from a subset of AT-rich regions in the genome only in the replicating cell population, consistent with the behavior of RNA-protected nascent strands.

There is a moderate to high correlation between NS_{G0gDNA} peak and CpG island densities (1 Mb bins; Pearson's r= 0.802, Spearman's $\rho = 0.490$) (Supplemental Table D.6A) and the NS_{G0gDNA} profile mimics the CpG island repeats on Chromosome 19 (Fig. 6.3A). Direct overlap analysis showed enrichment of CpG islands with 8.3% of NS_{G0gDNA} peaks (2.0% expected at random) overlapping 44.2% of all CpG islands (11.2% expected at random) (Supplemental Table D.7). In addition, there was a moderate correlation between average NS_{G0gDNA} fold enrichment and G4 density (100 kb bins, Pearson's r = 0.692, Spearman's $\rho = 0.564$) (Supplemental Table D.6B) as well as between NS_{G0gDNA} peak and G4 densities (100 kb bins, Pearson's r = 0.692, Spearman's $\rho = 0.363$) (Fig.

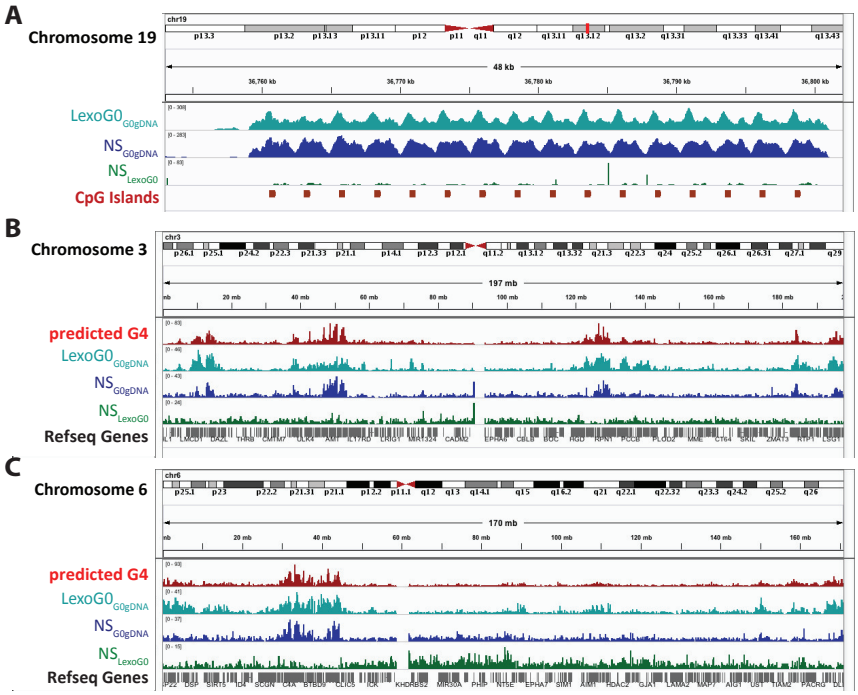
6.3B-D; Supplemental Table D.6A). 34.9% of NS_{G0gDNA} peaks (15.6% expected at random) overlapped with 26.0% of G4 motifs (7.0% expected at random) (Fig. 6.3D; Supplemental Table D.7).

NS-seq reads were highly enriched for telomere repeat sequences with a FRiT score of 1924.85 (Fig. 6.2D), over 63-fold higher than the undigested G0gDNA FRiT and about fivefold higher than LexoG0. This additional enrichment is G4-independent as the G4-CPMR for NS-seq reads is depleted relative to LexoG0 (Fig. 6.2E) and may be due to the enrichment of nascent DNA from origins in telomeres [Drosopoulos et al., 2012] (and references therein). G4 motifs were enriched in NS-seq reads compared with G0gDNA both in G4-CPMR (Fig. 6.2E) and in G4-start-site-CPMR (Fig. 6.2F), but were less enriched than in LexoG0 reads. Similarly, G4s were enriched within 500 bp 5' to NS_{G0gDNA} peak summits when oriented by strand, but less so than near $LexoG0_{G0gDNA}$ summits (Fig. 6.4, cf. D and B). Interestingly, G4 motifs were enriched 3' of the NS_{G0gDNA} peak summits (Fig. 6.4D). Mapping the G4s without correcting for strandedness revealed a periodicity of 171–224 bp both upstream of and downstream from NS_{G0gDNA} peak summits (Fig. 6.4C). These results support the conclusion that NS-seq enriches genomic regions that result from both nascent strands and nascent strand independent biases of λ-exo digestion.

6.3.6 Controlling λ-exo biases increases the specificity of NS-seq

Due to the inability of λ-exo to uniformly digest DNA, NS_{G0gDNA} peaks are of three types: (1) peaks resulting solely from nascent strands (true positives), (2) peaks resulting solely from nascent strand-independent λ-exo biases (systematic false positives), and (3) peaks resulting from some combination of both (true positives within λ-exo-biased regions). To deal with peak type 2, one could simply discard all the NS_{G0gDNA} peaks that overlap $LexoG0_{G0gDNA}$ peaks to obtain a higher fidelity but incomplete set of origins due to the elimination of true positives in λ-exo-biased regions (peak type 3). A better approach would be to call nascent strand enrichments relative to a λ-exo-digested gDNA background, such as LexoG0, to account for nascent strand independent λ-exo biases. Ideally, this approach controls against peak type 2 regions, because these regions are similarly enriched in both NS-seq and LexoG0, while not eliminating peak type 3 regions due to the additional nascent strand signal enriched over the LexoG0 background.

We tested this approach first on the human rDNA sequence, where the origin locations are known to be within the intergenic spacer (Supplemental Fig. D.3; Supplemental Table D.9). Figure 6.5A shows the signal per million reads (SPMR) over the rDNA for G0gDNA (black), LexoG0 (cyan), and NS-seq (blue), demonstrating that the G0gDNA background does not adequately represent all biases present in NS-seq. The G0gDNA SPMR is lower in magnitude compared with both the LexoG0 and NS-seq SPMRs, which both similarly respond to the presence of G4s (red and blue dots) and GC-richness (red line). The NS-seq SPMR closely tracks the LexoG0 SPMR due to the λ-exo-enriched parental DNA background but rises above the LexoG0 SPMR only in the intergenic spacer. Figure 6.5B plots the NS-seq fold enrichment across the rDNA locus over the two different



Sample	Pearson	Spearman	Peaks vs. G4	G4 vs. peaks
LexoG0$_{G0gDNA}$	0.704	0.704	36.9%	48.5%
NS$_{G0gDNA}$	0.692	0.363	34.9%	26.0%
NS$_{LexoG0}$	-0.248	-0.260	35.5%	6.8%

Figure 6.3: Correlation with predicted G4 motifs and CpG islands.
(A) The –log10 (P-value) signal for LexoG0$_{G0gDNA}$, NS$_{G0gDNA}$, and NS$_{LexoG0}$ for a 48 kb region on the q arm of Chromosome 19 containing a repeated array of CpG islands. **(B,C)** Density of G4 motifs and LexoG0$_{G0gDNA}$, NS$_{G0gDNA}$, and NS$_{LexoG0}$ peaks on Chromosome 3 **(B)** and Chromosome 6 **(C)**. **(D)** Scatterplot of genome-wide densities (counts in 100 kb bins) of LexoG0$_{G0gDNA}$, NS$_{G0gDNA}$, and NS$_{LexoG0}$ peaks and predicted G4 motifs. Pearson's r and Spearman's ρ for the densities of the indicated sample and G4 motifs are displayed in the box, along with the percentage of overlap of peaks with predicted G4 motifs (peaks vs. G4) or vice versa (G4 vs. peaks) for each data set. For all panels, cyan, LexoG0$_{G0gDNA}$; blue, NS$_{G0gDNA}$; and green, NS$_{LexoG0}$.

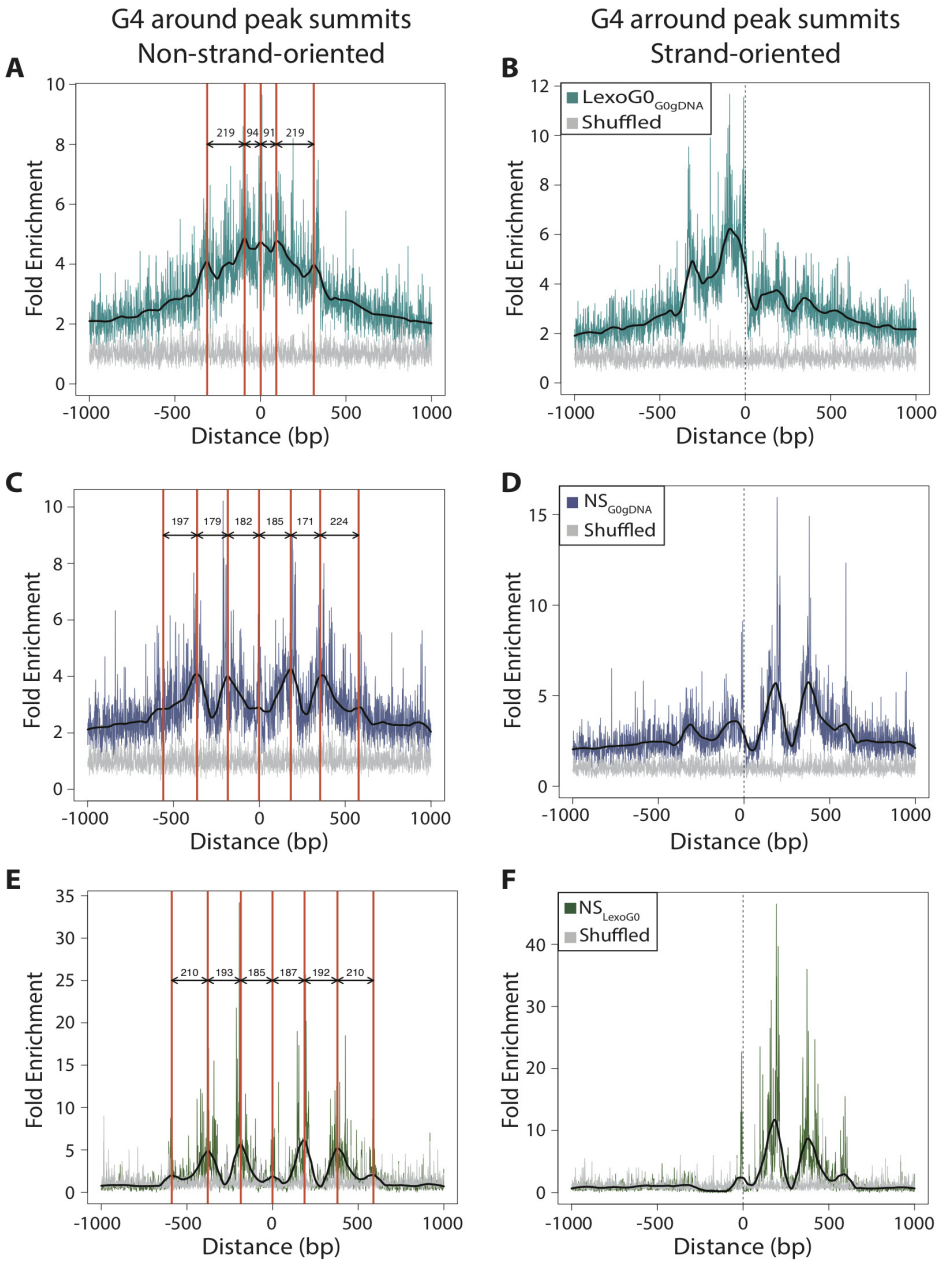


Figure 6.4: Distribution of G4 motifs around peak summits.
(Left) The distributions of G4 motifs (nonstrand-oriented) around LexoG0$_{G0gDNA}$ **(A)**, NS$_{G0gDNA}$ **(C)**, and NS$_{LexoG0}$ **(E)** peak summits. The red lines indicate the positions of wave crests; labeled arrows, the distance between adjacent crests. **(Right)** The distributions of G4 motifs (strand oriented 5'-3' left to right) around Lexo$_{G0G0gDNA}$ **(B)**, NS$_{G0gDNA}$ **(D)**, and NS$_{LexoG0}$ **(F)** peak summits.

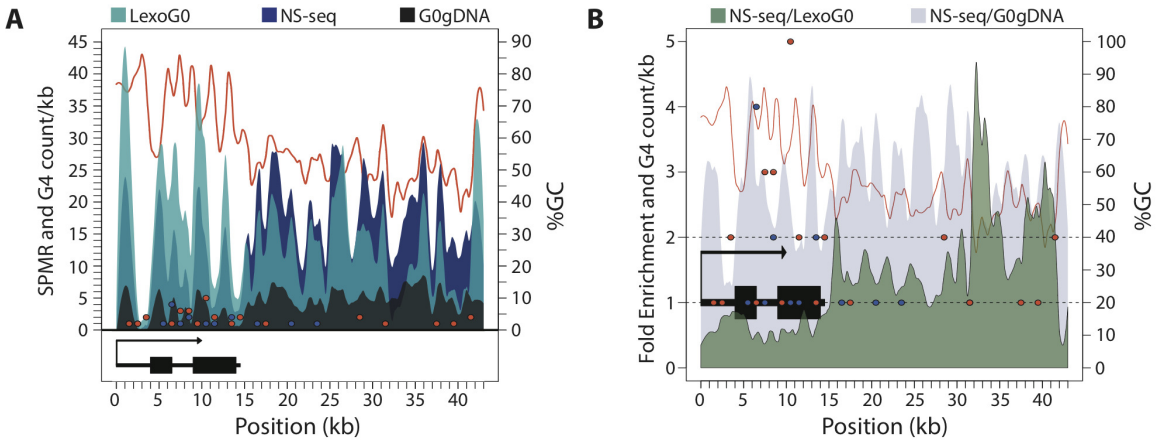


Figure 6.5: Controlling for λ-exo biases in NS-seq increases specificity for detecting replication initiation signal at the rDNA origin.
(A) The signal per million reads (SPMR) is shown for G0gDNA (black), LexoG0 (cyan), and NS-seq (blue) reads mapped to the ribosomal DNA (rDNA) sequence. The cyan and black are slightly transparent to allow visualization of the signals behind them. The lighter cyan is the LexoG0 signal alone, while the darker cyan indicates where the LexoG0 signal overlaps the blue NS-seq signal behind it. Non-zero G4 counts in 1 kb bins across the rDNA locus are shown for the plus strand (blue dots) and the minus strand (red dots). The percentage of GC across the locus is indicated by the red line. The rRNA transcription unit is shown below. **(B)** The NS-seq SPMR fold enrichment over G0gDNA (light blue-gray) or LexoG0 (green) controls. The rRNA transcription unit is shown (black). G4 counts in 1 kb bins and percentage of GC are displayed as in A. Dashed lines indicate one-fold and two-fold enrichment levels.

controls (NS/G0gDNA, blue-gray; NS/LexoG0, green). The entire locus is enriched when G0gDNA is used as the control, but when LexoG0 is used, only regions in the intergenic spacer are enriched greater than one-fold, and the only enrichments greater than two-fold correspond to known origin sites mapped by λ-exo-independent techniques (Supplemental Fig. D.3). Thus, analyzing NS-seq data relative to LexoG0 rather than G0gDNA increases the specificity of NS-seq. We then applied this approach genome-wide.

Genomic regions that are significantly enriched by λ-exo in replicating DNA compared with λ-exo-digested non-replicating DNA, hereafter referred to as NS_{LexoG0} peaks, were identified by setting the pooled NS-seq reads as the treatment and the pooled LexoG0 reads as the control. There were 66,831 NS_{LexoG0} peaks (Supplemental Table D.2), 93.3% of which overlapped NS_{G0gDNA} peaks (Supplemental Table D.7). That NS_{LexoG0} is almost entirely a subset of NS_{G0gDNA} reflects the increased specificity seen in the rDNA analysis. The GC content in the NS_{LexoG0} peaks displayed a single AT-rich mode (40% GC) in contrast to the two modes seen in NS_{G0gDNA} (Fig. 6.2C). In general, AT-rich reads were enriched and GC-rich reads were depleted in NS-seq relative to LexoG0 (Supplemental Fig. D.4). The NS_{LexoG0} and $\mathrm{LexoG0}_{G0gDNA}$ fold enrichment signals were weakly correlated with each other (Pearson's r = 0.171) (Supplemental Table D.8), demonstrating that most of the correlation with nascent strand independent λ-exo biases was broken as intended. Nonetheless,

20.2% of the NS_{LexoG0} peaks overlapped $LexoG0_{G0gDNA}$ peaks (Supplemental Table D.7), showing that using LexoG0 as the control gives higher sensitivity to detect origins in regions that overlap with λ-exo biases (peak type 3) than simply removing all NS_{G0gDNA} peaks that overlap $LexoG0_{G0gDNA}$ peaks. This is also demonstrated at the well-characterized MYC locus (Supplemental Fig. D.5).

Genome-wide, NS_{LexoG0} peak density in 1 Mb bins had a weak, negative correlation with CpG islands (Pearson's r= -0.364; Spearman's ρ = -0.472) (Supplemental Table D.6A), and the NS_{LexoG0} –log10(P-value) profile did not mimic the periodicity of 16 CpG islands on Chromosome 19 (Fig. 6.3A, green). Similarly, as visualized at the chromosomal level (Fig. 6.3B,C), the positive correlation with G4 motifs found when not controlling for λ-exo biases (NS_{G0gDNA}) was broken when accounting for them in NS_{LexoG0}, where correlations of G4 motif density with NS_{LexoG0} peak density (Pearson's r = -0.248, Spearman's ρ = -0.260) (Fig. 6.3D; Supplemental Table D.6A) and with NS_{LexoG0} average fold enrichment (Pearson's r = -0.124, Spearman's ρ = -0.004) (Supplemental Table S6B) in 100-kb bins were weakly negative. Nonetheless, although the majority of NS_{LexoG0} peaks did not overlap with G4 motifs, 35.5% did overlap (20.7% expected at random) (Fig. 6.3; Supplemental Table D.7) with 6.8% of total G4 motifs (3.8% expected at random) (Supplemental Table D.7). However, G4 motifs were no longer enriched 5' to the peak summits when oriented by strand (Fig. 6.4F), and the fold-enrichment of G4s mapping 3' to NS_{LexoG0} peak summits increased (Fig. 6.4D). When not strand-oriented, the wave-like G4 fold enrichment signal around NS_{LexoG0} summits displayed similar periodicity (185-210 bp) as that seen for NS_{G0gDNA}.

6.3.7 Phasing of nucleosomes and G4 motifs around the G4-proximal subset of NS-seq peak summits is enhanced after controlling for λ-exo biases

Both NS_{G0gDNA} and NS_{LexoG0} have wave-like G4 enrichment signals around their aligned peak summits (Fig. 6.4C,E), but the wave crests appear more prominent and phased in NS_{LexoG0} (Fig. 6.6 A and B). To quantify this, we defined the "prominence" as the difference between the mean fold enrichments of the crests and troughs (prominence = $crest_{mean}$-$trough_{mean}$) and used the crest-to-trough ratio (CTR= $crest_{mean}/trough_{mean}$) as a measure of how phased (or concentrated at the crests) the signal was. While the prominence of the crests around NS_{LexoG0} summits was 3.12 (Fig. 6.6A, Supplemental Fig. D.6A), it was only 0.95 for NS_{G0gDNA} (Fig. 6.6B, Supplemental Fig. D.6B). Similarly, the CTR for NS_{LexoG0} (4.09) was higher than that for NS_{G0gDNA} (1.35). The lower prominence and phasing of G4 motif enrichment around NS_{G0gDNA} summits may be due to a higher incidence of systematic false positives from nascent strand-independent λ-exo biases, which add a noisy, non-wave-like enrichment pattern that dampens the crest-to-trough ratio. Partitioning NS_{G0gDNA} summits into those that are and are not represented in NS_{LexoG0} decomposed the signal

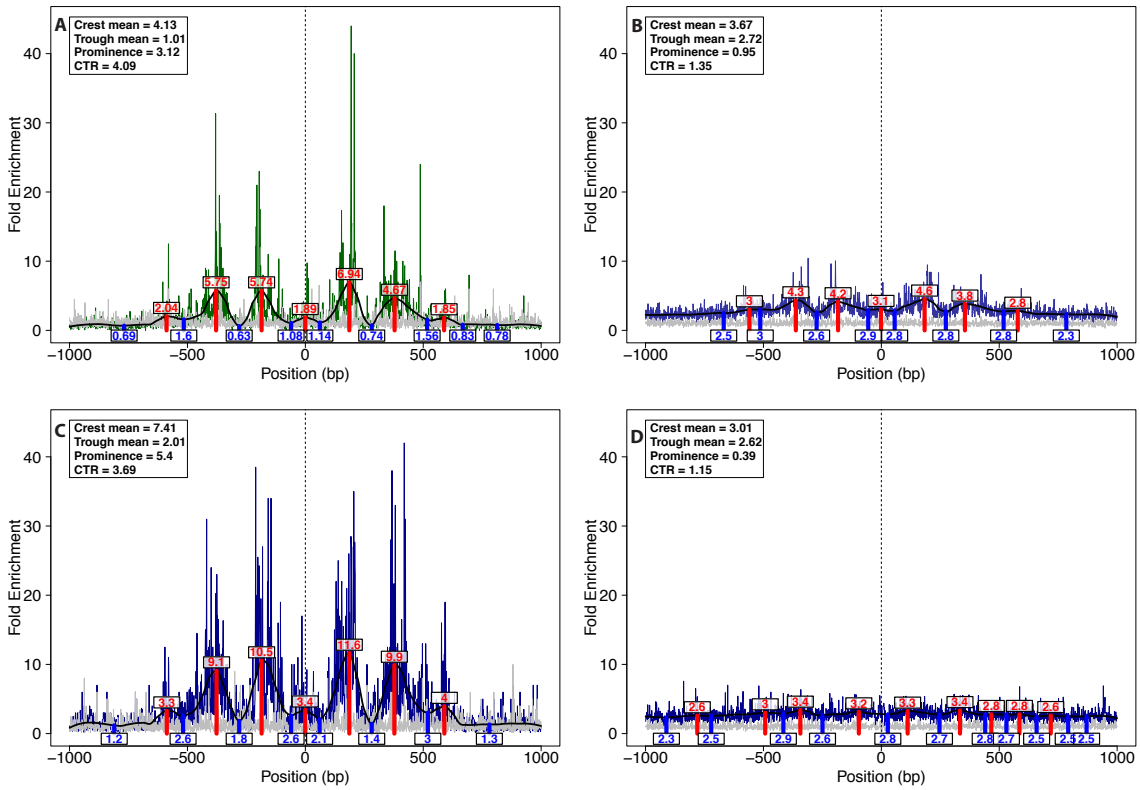


Figure 6.6: Controlling for λ-exo biases in NS-seq results in increased phasing of G4s around NS-seq peak summits.

Each panel shows the G4 enrichment signal around the specified set of summits (not strand-oriented) and, for each, measures crest heights (red vertical bars and numbers), trough heights (blue vertical bars and numbers), calculates the crest and trough means from each set of heights, and from the means computes "prominence" (crest_{mean}-trough_{mean}) and the crest-to-trough ratio (CTR = $\text{crest}_{mean}/\text{trough}_{mean}$), which is a measure of how phased the signal is around crests relative to troughs. Height, prominence, and CTR measurements were performed for **(A)** all NS_{LexoG0} summits, **(B)** all NS_{G0gDNA} summits **(C)** NS_{G0gDNA} summits represented in NS_{LexoG0}, and **(D)** NS_{G0gDNA} summits not represented in NS_{LexoG0}. NS_{G0gDNA} summits were considered to be represented in NS_{LexoG0} if they mapped inside of an NS_{LexoG0} summit window, where a summit window is a summit +/- 1 kb. Partitioning the NS_{G0gDNA} summits this way decomposes the relatively dampened wave-like NS_{G0gDNA} G4 enrichment signal (compared to NS_{LexoG0}) into a more prominent and phased signal **(C)** and a roughly uniform signal **(D)**.

into a stronger wave-like component (Fig. 6.6C, Supplemental Fig. D.6C) and a roughly uniform component (Fig. 6.6D, Supplemental Fig. D.6D), respectively. The prominence (5.4) and CTR (3.69) for the subset of NS_{G0gDNA} summits found in NS_{LexoG0} both rose (Fig. 6.6C, Supplemental Fig. D.6C) whereas both dropped (0.39 and 1.15, respectively) for the NS_{G0gDNA} summits not represented in NS_{LexoG0} (Fig. 6.6D, Supplemental Fig. D.6D). This analysis is consistent with the conclusions that the NS_{G0gDNA} summits not represented in NS_{LexoG0} are largely a product of nascent strand-independent λ-exo enrichments, that G4s in the vicinity of true nascent strand enrichments are non-randomly located with respect to peak summits forming prominently phased waves when viewed in aggregate, and that controlling for λ-exo biases increases the specificity of NS-seq.

The periodicity of G4 motif enrichment around NS_{G0gDNA} and NS_{LexoG0} peak summits was highly reminiscent of nucleosome spacing. In contrast, the enrichment of G4s around $LexoG0_{G0gDNA}$ peaks (Fig. 6.4A) appeared to be a function of proximity, with most G4s occurring near the peak summits, though with traces of nucleosomal periodicity. To test whether G4s had a relationship with nucleosomes, the nucleosome signal from K562 and GM12878 cells [Kundaje et al., 2012] was plotted around the subsets of the $LexoG0_{G0gDNA}$, NS_{G0gDNA}, and NS_{LexoG0} peak summits that contained one or more G4s within 1 kb (46.1%, 43.7%, and 34.8%, respectively) (Fig. 6.7; Supplemental Fig. D.7AC; Supplemental Table D.10). Importantly, 91.6% of the G4-proximal peak summits in NS_{LexoG0} had only a single G4 within 1 kb (Supplemental Table D.10). The G4-proximal summits in each data set were in regions of the genome with average and lower-than-average nucleosome enrichment. Nonetheless, the summits were flanked by nucleosomes, from which the nucleosome signal spread out in a wave-like fashion with crest-to-crest distances typical of nucleosome spacing. The wave-like characteristic was most pronounced around NS_{LexoG0} summits (Fig. 6.7C). As determined by the lowest "divergence" (sum of squared deviations from the mean signal) and highest correlation of K562 and GM12878 cell line signals, nucleosome positioning was most consistent around NS_{LexoG0} summits (divergence = 1.44; Pearson's r = 0.95; Spearman's ρ = 0.97) (Supplemental Tables D.11,D.12,D.13) compared with $LexoG0_{G0gDNA}$ (divergence = 28.87; Pearson's r = 0.19; Spearman's ρ = 0.34) (Supplemental Tables D.11,D.12,D.13) and NS_{G0gDNA} (divergence = 17.95; Pearson's r = 0.31; Spearman's ρ = 0.36) (Supplemental Tables D.11,D.12,D.13). Moreover, partitioning the NS_{G0gDNA} summits into those that are and are not represented in NS_{LexoG0} decomposed the nucleosome signal around the NS_{G0gDNA} summits into a stronger wave-like component with more consistent nucleosome positioning (divergence = 1.43; Pearson's r = 0.91; Spearman's ρ = 0.93) (Fig. 6.7E; Supplemental Tables D.11,D.12,D.13) and a component that looked similar to and shared a nearly identical divergence (28.85) (Fig. 6.7D; Supplemental Tables D.11,D.12,D.13) with the nucleosome signal around the $LexoG0_{G0gDNA}$ summits. Overall, controlling for nascent strand independent λ-exo biases eliminates a significant noise component in the nucleosome signal. Interestingly, plotting the distribution of both G4s and nucleosomes together revealed that G4 enrichment crests (Fig. 6.7A-C; Supplemental Fig. D.7) were offset relative to nucleosome crests in

all three data sets, raising the possibility that a role of G4s near the G4-proximal subset of origins is in nucleosome positioning.

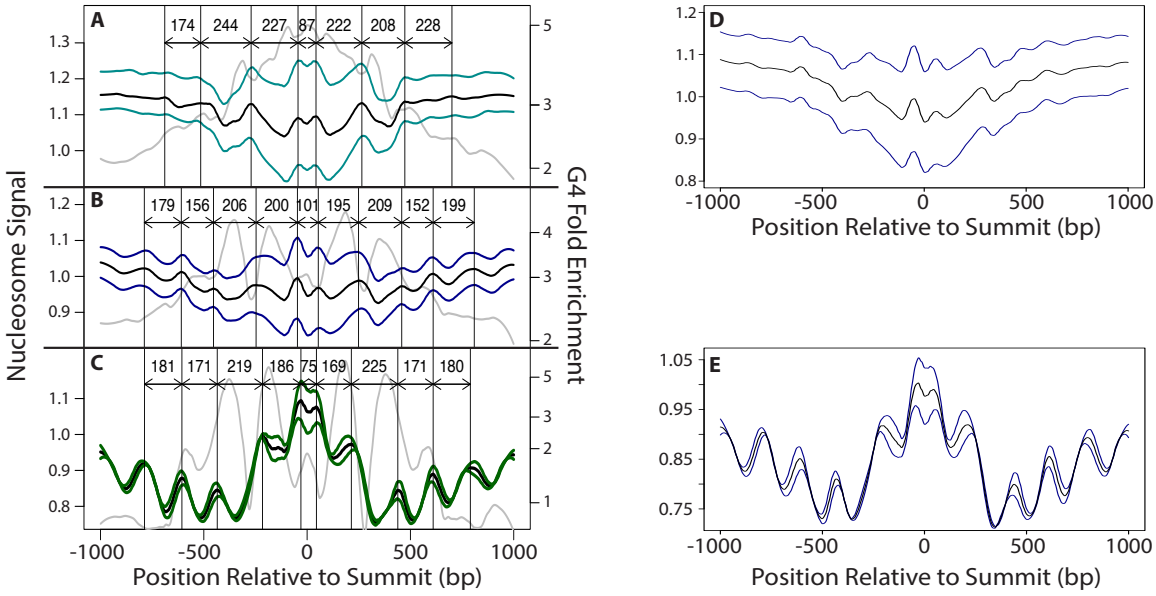


Figure 6.7: Controlling for λ-exo biases in NS-seq results in increased phasing of nucleosomes around NS-seq peak summits.
The nucleosome signal was plotted around the peak summits for **(A)** LexoG0$_{G0gDNA}$, **(B)** NS$_{G0gDNA}$, and **(C)** NS$_{LexoG0}$. The colored lines show the nucleosome signal for K562 and GM12878 cells [Kundaje et al., 2012]. The black line is the mean of the two cell lines. The vertical lines indicate the crest positions of the wave-like nucleosome signal, and the labeled arrows indicate the intercrest distances. The gray lines show the distribution of non-strand-oriented G4 motifs (log-transformed versions of Fig. 6.4 A, C, and E, respectively). Similar to the aggregate G4 analysis, nucleosomes are most phased around NS-seq peak summits when Lexo-biases are controlled (compare **(B)** and **(C)**). This is recapitulated by showing that nucleosome phasing increases around NS$_{G0gDNA}$ peak summits after removing peaks not represented in NS$_{LexoG0}$. NS$_{G0gDNA}$ summits were partitioned to show the nucleosomal signal around **(D)** NS$_{G0gDNA}$ summits that are **not** represented in NS$_{LexoG0}$ and around **(E)** NS$_{G0gDNA}$ summits that **are** represented in NS$_{LexoG0}$. NS$_{G0gDNA}$ summits were considered to be represented in NS$_{LexoG0}$ if they overlapped a NS$_{LexoG0}$ summit window (summit +/- 1 kb). Partitioning the NS$_{G0gDNA}$ summits this way decomposes the NS$_{G0gDNA}$ nucleosomal signal in (B) into a stronger, more consistent wave-like signal (E) similar to NS$_{LexoG0}$ in (C) and a less wave-like, less consistent signal (D) similar to LexoG0$_{G0gDNA}$ in (A). Also see Tables D.11,D.12,D.13.

6.3.8 Limiting the effects of G-quadruplexes in λ-exo digestions by destabilization in glycine-NaOH buffer; cation-dependent resistance to digestion

The traditional λ-exo buffer, used initially by Radding (1966) and in the nascent strand experiments reported to date (including our own) contains 67 mM glycine that is titrated to the desired pH with KOH (glycine-KOH). G4 structures are most stable in the presence of K^+ but are much less so in the presence of Na^+ [Kankia and Marky, 2001, Shim et al., 2009], suggesting that λ-exo digestion may be impeded less in buffers containing Na^+ rather than K^+. In MYC plasmid digestion experiments, titration of KCl (pH 8.8) resulted in stronger bands, suggesting higher stability of G4s, while NaCl titration had no effect (Supplemental Fig. D.8). Therefore, NaOH was substituted for KOH to titrate the desired pH of the reaction buffer (glycine-NaOH). The MYC plasmid was more efficiently digested in glycine-NaOH than in glycine-KOH (Fig. 6.1C) at both pH 8.8 (complete digestion) and pH 9.4 (partial digestion). In both glycine-NaOH and glycine-KOH, the G4s posed a greater obstacle to λ-exo digestion at the higher pH, which can be explained by their higher stability with increased concentration of monovalent cations (0.65 mM at pH 8.8 and 1.95 mM at pH 9.4).

6.4 Discussion

NS-seq is a method to map origins genome-wide that employs λ-exo to enrich nascent strands, which are protected from digestion by the RNA primer at their 5' end. However, we show here that λ-exo does not digest the parental DNA background uniformly. We report that λ-exo more efficiently digests AT-rich DNA than GC-rich DNA, which results in enrichments of GC-rich regions of the genome, extending the observations of single-molecule studies genome-wide [Perkins et al., 2003, van Oijen et al., 2003, Conroy et al., 2010]. Moreover, we show that λ-exo digestion is obstructed when it encounters G4s. Therefore, λ-exo-enriched DNA in NS-seq will contain not only RNA-protected nascent strands but also GC-rich and G4-protected DNA, which is problematic when attempting genome-wide origin discovery. This problem may also explain the apparent discrepancy between a landmark Okazaki fragment sequencing study [Smith and Whitehouse, 2012] and a more recent study that used a λ-exo-based approach [Yang et al., 2013] for enriching Okazaki fragments for sequencing.

One way to account for nascent strand-independent λ-exo biases in NS-seq is to use λ-exo-digested DNA from nonreplicating cells (LexoG0) as a control. Our analysis on the rDNA locus shows that this approach increases the specificity of NS-seq. Another indication of increased specificity is that the wave-like pattern of G4 enrichment around the NS_{LexoG0} summits is more pronounced and phased than around the summits of NS_{G0gDNA} and $LexoG0_{G0gDNA}$. Similarly, the phased nucleosomal signal around NS_{LexoG0} summits is more pronounced and consistently positioned across cell lines, showing that controlling λ-exo biases improves this recognizable biological signature that is also seen at yeast origins [Lipford and Bell, 2001, Eaton et al., 2010]. Nucleosome phasing around

tens of thousands of sites in the genome is extremely unlikely to occur at random. Moreover, using the LexoG0 control in NS-seq has the advantage of higher sensitivity to true positives in strongly λ-exo-biased regions than the alternative procedure of eliminating all NS_{G0gDNA} peaks that overlap $LexoG0_{G0gDNA}$ peaks. Genome-wide, 20.2% of NS_{LexoG0} peaks overlap $LexoG0_{G0gDNA}$ peaks. Nonetheless, the NS_{LexoG0} approach used here may not be fully sensitive to weak origins in strongly λ-exo-biased regions due to the difference in pH between LexoG0 and NS-seq. This may explain the lack of CpG island overlap in NS_{LexoG0} despite previous evidence from λ-exo-independent techniques that some origins occur near CpG islands (Delgado et al. 1998 and references therein). Moreover, we cannot exclude the possibility that digestion of replicating DNA at pH 8.8 introduces other biases that are not accounted for by the LexoG0 control digested at the higher pH. However, the results at the rDNA locus suggest that our approach is both sensitive and specific. Moving forward, important advances for NS-seq will be to (1) optimize λ-exo digestion conditions in the presence of Na^+ instead of K^+ and (2) utilize pH 8.8 for both the NS-seq sample and the LexoG0 control, both of which destabilize G4 structures, thereby minimizing the problem of G4s impeding λ-exo digestion.

In light of the biases inherent in λ-exo digestion, recent reports suggesting that G4s are hallmarks of mammalian replication origins may have had inflated estimates of the association of G4 motifs and origins: 73.9% of putative mouse origins localized with G4s [Cayrou et al., 2012a], and 67% and 91.4% of putative human origins overlapped with G4 sequences with loops of 1–7 and 1–15 nt, respectively [Besnard et al., 2012]. These estimates may be inflated by the presence of nascent strand-independent λ-exo enrichments and by preferentially enriching origins in λ-exo-biased regions. It is possible that the higher enzyme-to-DNA ratio used in previous studies lessened the impact of nascent strand-independent λ-exo biases, but it is striking that similar regions of the genome were enriched in those data sets as in our LexoG0 peaks (e.g., CpG islands, G4s, and GC-rich DNA). Furthermore, our plasmid experiments demonstrate that λ-exo digestion of G4 structures is more efficient at pH 8.8 than at pH 9.4 (the pH used in some previous studies; [Cayrou et al., 2012a]), which suggests that the higher pH may require higher enzyme-to-DNA ratios to achieve the same efficiency of G4 digestion. Indeed, even before controlling for λ-exo biases, our NS_{G0gDNA} peak set (pH 8.8) is not as strongly correlated with G4s as peak sets of previous studies [Cayrou et al., 2012a] that used pH 9.4. Moreover, in our 3' labeled plasmid experiments, we used a high ratio of 50 units of λ-exo per microgram of DNA. Still, we saw that G4s are stabilized and not efficiently digested at pH 9.4. There is also a concern that too high an enzyme-to-DNA ratio may sacrifice some of the enzyme's specificity against RNA digestion [Yang et al., 2013]. Finally, the prediction that G4s should be enriched 5' of peak summits after λ-exo digestion was borne out in our studies and those of others [Cayrou et al., 2012a]. This prediction also gives rise to an alternate interpretation of the observation that when a G4 is experimentally inverted to shift it from one strand to the other, the region attributed to have origin activity (after λ-exo enrichment) also shifted so that it remained 3' to the G4 that had been moved [Valton et al., 2014]. Moreover, the 5' G4-start-site-CPMR enrichment in LexoG0 and NS-seq reads is diagnostic of the inability of λ-exo to digest G4-protected

DNA, and one would not expect nascent strands alone to produce this effect.

Though G4s are correlated with NS-seq peaks when controlling with undigested G0gDNA (NS_{G0gDNA}), the positive correlation is broken when controlling for nascent strand-independent λ-exo biases (NS_{LexoG0}). Only 6.8% of G4s with loops of 1–7 nt in the genome overlapped with NS_{LexoG0} peaks, suggesting that the vast majority of G4 motifs are not general determinants of the location of origins of replication. Similarly, it has recently been reported that only one out of seven G4s in the human genome are associated with BrdU NS peaks [Mukhopadhyay et al., 2014] and that only 5.2% of G4 motifs are associated with nascent strand peaks from a new λ-exo-independent method called "nascent strand capture and release" (NSCR) [Kunnev et al., 2015a]. Therefore, G4s do not appear to be sufficient for origin specification as most G4 motifs are not associated with origin activity. Moreover, most of our NS-seq peaks are not near G4 motifs, suggesting that G4s are not necessary for specification of all origins. Likewise, $< 6\%$ of NSCR peaks had an orientation-specific relationship with G4s [Kunnev et al., 2015a], and using the orthogonal origin mapping technique of bubble-seq, Mesner et al. (2013) found that the majority of bubble-containing fragments lacked G4 motifs. In regions of discordance between bubble-containing DNA and NS-seq peaks mapped by Besnard et al. (2012), G4 motifs are enriched in the NS-seq peaks but are relatively depleted in bubble-containing fragments. Mesner et al. (2013) concluded that the discordance in mapping replication origins may reflect methodological problems, such as G4s impeding λ-exo activity. Our data support this hypothesis.

Despite the lack of a general correlation with G4s, a subset of the NS_{LexoG0} peaks overlapped with G4 motifs. What might be the role, if any, of G4s at this subset of origins? G4s are enriched in promoters [Huppert, 2010] and may play a role in transcriptional regulation. Since replication origins are often found in gene promoters and ORC is preferentially found in nucleosome-free regions [MacAlpine et al., 2010], it remains to be seen if G4s play an active role in the initiation of DNA replication, or if it is simply a correlation with the potential role of G4s in transcription. Mutagenesis studies [Valton et al., 2014] will have to discern if early activation of origins in S phase is just a secondary effect of G4s influencing transcription and opening chromatin structure. ORC preferentially binds to G4s in single-stranded DNA and RNA *in vitro* [Hoshina et al., 2013], suggesting that ORC might bind G4s *in vivo*. However, G4 binding of ORC was based on gel shift competition and was comparable to AT-rich dsDNA as a competitor. ORC has also been shown to preferentially bind negatively supercoiled DNA (Remus et al. 2004). Therefore, ORC may bind DNA with any of these characteristics *in vivo*.

Intriguingly, we found that in the subset of origins associated with G4s, the G4s were positioned in a phased manner reminiscent of nucleosome spacing. Moreover, nucleosomes are phased around the NS_{LexoG0} peak summits that have G4s within 1 kb, exhibiting crests of nucleosome enrichment that are offset from the crests of G4 enrichment. G4s and G-rich sequences have been

suggested to be nucleosome exclusion signals in budding yeast, Caenorhabditis elegans, and human cells [Iyer and Struhl, 1995, Halder et al., 2009, Wong and Huppert, 2009] and shown to be enriched in long nucleosome-free regions [Schwarzbauer et al., 2012]. Furthermore, G4s are predicted to form more easily in nucleosome-free regions [Hershman et al., 2008], and G4s associated with origins are in open chromatin as detected by DNase I hypersensitivity [Mukhopadhyay et al., 2014]. Similarly, origins of replication preferentially localize to nucleosome-free regions in yeast [Simpson, 1990, Lipford and Bell, 2001, Berbenetz et al., 2010, Eaton et al., 2010], and ORC localizes to nucleosome-free regions in Drosophila [MacAlpine et al., 2010]. In budding yeast, the ARS sequence establishes a nucleosome-free region, where ORC binds and then positions the flanking nucleosomes [Lipford and Bell, 2001, Eaton et al., 2010]. Given these findings, G4s may influence the positioning of nucleosomes flanking one-third of human replication origins, thus taking on some of the role played solely by ORC in budding yeast. Alternatively, as G4s are nucleosome exclusion signals, they may establish nucleosome-free regions that are then bound by ORC, which in turn positions the nucleosomes as seen in yeast. Overall, the role of G4s near a subset of metazoan origins may be involved in nucleosome positioning that results in consistently available sites for opportunistic ORC binding, giving rise to apparent specificity in origin localization.

6.5 Methods

6.5.1 Plasmid experiments

The plasmid pFRT.myc6xERE contains a 2.4 kb fragment from the human MYC promoter [Malott and Leffak, 1999] that carries two sequences shown to form G4s: Pu27 [Brooks and Hurley, 2010] and Pu30. Plasmids were linearized with BglII (New England Biolabs [NEB]), 3' end-labeled with terminal transferase (NEB) and α32P-CTP, and made single-stranded (if needed) by boiling and transferring to ice. Labeled plasmids (200 ng) were digested overnight with 10 units of λ-exo (Fermentas) in λ-exo buffer (glycine-KOH at pH 9.4, 2.5 mM $MgCl_2$, 50 μg/mL bovine serum albumin). Unlabeled plasmid (700 ng) was digested with 20 units of λ-exo in the glycine-KOH or glycine-NaOH buffer indicated. Deletion mutants were constructed with the Q5 site-directed mutagenesis kit (NEB) following the manufacturer's directions. For more details on the plasmid experiments, see Supplemental Methods.

6.5.2 Cell Culture

MCF7 breast cancer cells were obtained from ATCC and grown in Dulbecco's modified eagle medium with 10% fetal calf serum supplemented with 100 U/mL penicillin and 100 μg/mL streptomycin. For NS-seq, asynchronous cultures were grown to 70%–80% confluency. For G0gDNA and LexoG0, cells were synchronized in G0 by plating at 50% confluency and serum-starving for 24 h. The proportion of cells in S phase was determined by FACS analysis (BD FACSCalibur).

6.5.3 LexoG0 and NS-seq library construction and sequencing

gDNA was harvested from serum-starved (LexoG0) or asynchronous (NS-seq) MCF7 cells using DNAzol (Invitrogen). One hundred fifty micrograms of LexoG0 gDNA (9.6% S phase) was lightly sonicated to a size range of 200 bp to 10 kb, made single-stranded, phosphorylated at the 5' ends with T4 polynucleotide kinase (NEB), and then digested with 100 units of λ-exo in glycine-KOH (pH 9.4) buffer. For NS-seq, nascent strands were prepared from 150 μg of asynchronous gDNA (35%–40% S phase) following the protocol of Bielinsky and Gerbi (1998). Replicative intermediate DNA was enriched with BND-cellulose (Sigma), made single-stranded, phosphorylated, and then digested with 100 units of λ-exo in glycine-KOH pH 8.8 buffer. After λ-exo digestion for both LexoG0 and NS-seq, 500- to 1500-nt fragments were purified from ultrapure LMP agarose (Invitrogen), made double-stranded with random hexamers and Klenow (NEB), and sonicated to a size range of 100–600 bp. The G0gDNA control was prepared by sonicating gDNA from serum-starved MCF7 cells (6.8% S phase) to a size range of 100–600 bp. For all the samples described above, Illumina libraries were prepared using the NEBNext kit (NEB) following the manufacturer's directions, and library fragments of 200–500 bp were purified from 2% NuSieve agarose (Lonza) gels. All libraries were sequenced on the Illumina HiSeq platform. For more details on nascent strand preparation and library construction, see Supplemental Methods.

6.5.4 Analyses of reads and peaks

For each data set, reads were mapped with Bowtie2 [Langmead and Salzberg, 2012] to hg19. Peaks were called with MACS2 [Zhang et al., 2008] with "--nomodel" specified and using one data set as the treatment and the other as a control following the Treatment$_{Control}$ format. Peak and peak summit coordinates were obtained from MACS2 output files. MACS2 was used to generate bedGraphs of fold enrichment and –log10(P-value) signals for visualization in the integrative genomics viewer (IGV) [Thorvaldsdóttir et al., 2013]. Overlap analyses were performed with BEDTools [Quinlan and Hall, 2010], and significance was calculated using a binomial model in R. BEDTools was used to calculate the "percent GC" inside peak coordinates. GC content in mappable reads was obtained using Python and then analyzed and visualized in R. FRiT scores were calculated by counting the number of mappable reads per million reads that remapped to the human telomere sequence. G4-CPMR was obtained by counting the G4 motifs in reads per million reads, and G4-start-site-CPMR was obtained by counting where the motifs started in the reads. For rDNA analyses, reads were mapped to a version of hg19 that contained an rDNA repeat as an extra "chromosome." SPMR was calculated by counting the number of reads per million reads over each base with BEDTools. BEDTools was used to partition the genome into specified bin sizes (e.g., 100 kb) and to count the number of features inside each bin. G4 motifs were identified by implementing the quadparser approach in Python. CpG islands were downloaded from the UCSC Table Browser [Karolchik et al., 2004]. Counts and/or mean values of various features inside identical bins were used in correlation tests and for visualization in IGV. Pearson product-moment and Spearman's rank correlation coefficients

were calculated in R. Scatterplots of bin counts were made in R. G4 counts around peak summits were obtained with the help of BEDTools; correcting for the strand-specificity of the G4 motif, as well as visualization, was done in R. Nucleosome signals around peak summits were obtained with the help of BEDTools and visualized in R. Genomic features, such as peak coordinates, were shuffled around the genome using BEDTools. For more detailed descriptions of the bioinformatics analyses, see Supplemental Methods.

6.6 Data Access

The raw sequencing reads for NS-seq, LexoG0, and G0gDNA have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRP045284.

6.7 Acknowledgments

Illumina DNA sequencing was performed at the Brown University Genomics Core Facility supported by NIH COBRE grant P30GM103410, and we thank Christoph Schorl for support. We thank the Center for Computation and Visualization for access to computational resources and Lingsheng Dong and Mark Howison for resource support. We thank Alex Brodsky for use of tissue culture facilities and Ben Raphael for helpful discussions. We received support from DOD W81XWH-10-1-0463 research grant to S.A.G., postdoctoral fellowship DOD W81XWH-11-1-0599 to C. C., and predoctoral fellowships from NSF GRFP (DGE-1058262), NSF EPSCoR (1004057), and NIH predoctoral traineeship (5-T32- GM 07601) to J.U. This paper is dedicated to Ellen Fanning (19462013), who contributed so much to the field of DNA replication.

6.8 Epilogue

6.8.1 A survey of literature that bears on our results since publication

Since this chapter was published [Foulk et al., 2015], several papers have been published that reflect on our conclusions that GC-rich and G4-protected DNA are enriched by λ-exo in addition to nascent strands, and that NS-seq is a mixture of peak types (origins and non-origins). It needs to be re-iterated: our paper both posits that there are technical artifacts associated with the λ-exo biases as well as a subset of putative origins that are indeed associated with nearby G4 motifs that may be an important part of establishing or maintaining the chromatin environment for that subset. It often appears that our colleagues think we are saying this technique yields 100% technical artifact when we are saying it yields a mixture of enriched origin and non-origin sequences (i.e. not 100% artifact), the latter of which needs to be dealt with.

A study that performed λ-exonuclease-based NS-seq in Drosophila [Comoglio et al., 2015] reiterates some of the conclusions from earlier papers about origins associated with G4s [Cayrou et al., 2011, Cayrou et al., 2012a, Besnard et al., 2012], though only 9-22% of their putative origins are near G4 motifs in contrast to the nearly 70% reported to be near G-rich motifs in Drosophila from a previous λ-exo study [Cayrou et al., 2012a]. This paper contributes some new analyses on DNA shape that are quite interesting. Nonetheless, it remains questionable on whether they are analyzing the shape of DNA sequences associated with origins, origin-independent λ-exo enrichments such as G-quadruplexes (G4s), or a combination of both. The procedure they used is even harsher than seen previously, with 4-5 rounds (i.e. 4-5 days) of λ-exo treatment at pH 9.4. Each round has dramatically decreased levels of parental DNA compared to the previous step leaving λ-exo to spend more of its time digesting the RNA primer of nascent DNA. This makes it is hard to imagine that the RNA primers can protect nascent DNA against λ-exo for the duration of this procedure, especially since there is evidence that λ-exo degrades RNA-protected DNA, particularly when there is not a lot of competing parental DNA (Michael Foulk, unpublished observations; John Yates, personal communication; [Yang et al., 2013]). The aggregate enrichment profiles in this paper with respect to G4 motifs are exactly what one would expect if G-quadruplexes protected DNA from digestion, as seen in previous papers and discussed in our paper and in our review [Urban et al., 2015a], and as can easily be demonstrated by simulations (see Fig. 6.13F). They notice this as well, but attribute it to replication forks stalling at the G4 in vivo thereby creating an asymmetric enrichment pattern. In contrast, another paper promoting the importance of G-quadruplexes at replication origins [Valton et al., 2014] dismisses this interpretation citing studies that demonstrate evidence that replication forks do not stall near origins nor near G4 motifs in wild type cells. Indeed, it takes a Pif1 mutant to cause detectable stalling at G4 motifs [Paeschke et al., 2011]. Comoglio et al looked at λ-exo digestion of short, intermediate, and long DNA fractions from the sucrose gradient and found that the asymmetric profile was found in all fractions with the enrichment emerging 3' to the G4 and

extending longer in the longer DNA fractions. They interpreted this as evidence for stalling. However, the simpler interpretation is once again that the G-quadruplexes are protecting DNA 3' to them from λ-exo regardless of length and their results are not surprising, and are even predictable, given the results in our paper. It would take considerable stalling of the replication forks to give rise to this effect in the data, not a barely perceptible pause invisible to methods that detect fork stalling. Finally, it clearly would have been interesting to see the relationship between ORC binding sites mapped in Drosophila [MacAlpine et al., 2010] with the putative initiation sites in this study. One imagines that the authors must have looked. Given that initiation can occur at ORC-distal locations from sliding MCMs [Gros et al., 2015] and that MCMs are not necessarily adjacent to ORC binding cites in the Drosophila genome [Powell et al., 2015], there may be no need to assume the ORC binding sites and NS-seq peaks should directly overlap. However, they could be correlated in larger bins given that ORC binding sites are the nucleation points for MCM loading [Powell et al., 2015]. Either way, it would have been interesting to see that comparison.

In another paper featuring λ-exo enriched nascent strands for Trypanosomes [Lombraña et al., 2016], it was reported that 74% of origins were associated with G4 motifs. They said, "To rule out that this association could be due to the presence of G4 motifs at the 5' end of the sequencing reads, which could prevent λ-exonuclease digestion, we eliminated those reads from the analysis and repeated the peak-calling procedure". Ultimately, removing reads that overlapped G4s did not change their results. Unfortunately, although they present this as a rigorous attempt to discount the technical biases we report, it is actually an insufficient attempt and their results are completely expected. It appears they did paired-end sequencing on the 400-1500 bp DNA that came out of the nascent strand enrichment procedure without an intervening fragmentation step. Thus, to the unengaged reader, it might appear that this should have removed any nascent-strand independent bias due to G4s, since this seems to be the bioinformatic equivalent of removing molecules that had G4s at their 5' ends. It should be pointed out immediately that this does not even attempt to remove or correct for any bias from GC-content alone (which seems like the major bias in our paper). However, it also does a poor job at fully correcting for or removing the bias from G4 protection. What our paper posits is that the G4 folds and unfolds in vitro during λ-exo digestion. When the G4 is folded it protects the DNA 3' of it from digestion. Thus, the G4-protected DNA is a potential substrate for digestion, but everything 3' to the G4 has a much longer half-life during digestion than everything 5' to the G4 (and a longer half-life than other non-protected DNA in the tube). In other words, the copy number of the DNA 3' of the G4 becomes enriched compared to the copy number of background DNA. Thus, even after the G4 unfolds and λ-exo can continue digestion, the relative copy number of the 3' DNA stays much higher even if one assumes a uniform rate of digestion thereafter. However, we know digestion is not uniform and is much less efficient in GC-rich regions where G4 motifs occur. So these sequences would likely have longer half lives even without the G4 or after it unfolds. Nonetheless, to illustrate, imagine a tube starts out with 1000 copies each of two different types of molecules, one that has a sequence that can form a G4 at its 5' end and one

that does not (called G4+ and G4- respectively), and after a certain amount of digestion time there are 100 G4+ molecules and 10 G4- molecules left. In this scenario, there are more G4+ molecules remaining due to the extended half-life from G4 protection. At this time if we assume the G4 unfolds in all of the G4+ molecules and can be subsequently digested at the same rate as G4-, then the G4+ molecules will stay 10-fold enriched compared to the G4- background. For example, say the next time we checked, there were 10 G4+ molecules and 1 G4- molecule left and that only one of the G4+ molecules still had the undigested G4 motif (everything else is partially digested). What this group did is the bioinformatic equivalent of just removing the single G4+ molecule that still had an undigested G4 sequence leaving the G4+ molecules 9-fold enriched instead of 10-fold enriched, despite the fact that the entire 10-fold enrichment was from the longer half-life the G4 protection offered. To fully account for the G4-protection bias, they would need to be able to also remove all reads from the proportion of partially digested G4+ molecules that had extended half-lives due to originally being protected by the G4. Altogether, their results are consistent with what we show in our paper. First, we showed that G4s are enriched in the 5' ends of reads, which is diagnostic of the G4s folding and temporarily preventing digestion. Second, we also show that the DNA 3' to G4s in the genome is enriched, which is diagnostic of the longer half-life of these sequences. Third, we show that the predominant bias is nucleotide content with strong depletions and enrichments of low and high GC content sequences respectively. Not only did they not fully account for the G4 protection bias, they did nothing to correct for the nucleotide content bias.

Another paper was published from the Mechali laboratory [Cayrou et al., 2015], one of the original proponents of the prevalence G-quadruplexes at origins identified with λ-exonuclease [Cayrou et al., 2011, Cayrou et al., 2012a]. In this paper they offer criticisms of our methods that we also confronted at the Cold Spring Harbor meetings on DNA Replication (2013, 2015). One criticism we have encountered from this laboratory is that since we did not use as much λ-exo enzyme, our datasets are not comparable to theirs. Another criticism was that since we used BND cellulose to enrich replicative intermediates instead of sucrose gradients to size-select for nascent DNA, our results were not comparable to theirs. These arguments cannot be true if both techniques result in strictly enriching origins. Any two methods that enrich origins should be comparable. Thus, the implication of suggesting that the results are not comparable is that one method identifies origins and the other does not. Interestingly, we have found that 70-94% of the earlier human datasets of others that used sucrose gradients [Cadoret et al., 2008, Martin and Wang, 2011, Besnard et al., 2012] are captured in the LexoG0$_{G0gDNA}$ peak set resulting from using λ-exo on fragmented genomic DNA from non-replicating cells [Urban et al., 2015a]. Overlap of 70-94% indicates comparable results by most standards. What is alarming about how comparable these results are is that they are between replicating and non-replicating cells. As low as 6% of those datasets were unique to replicating cells, indicating that a large proportion likely arises from λ-exonuclease enriching parental DNA with resistant features, such as G4-protected and GC-rich DNA, our original premise. Interestingly, our NS-seq dataset has 53% peaks unique to replicating cells before controlling for λ-exo biases. If the

implication is that one method enriches origins and the other does not, those statistics seem to be in favor of our method. In any case, skipping the sucrose gradient and using much less enzyme gave comparable results to the data published earlier by others who used sucrose gradients and higher enzyme:DNA ratios. This indicates that the sucrose gradient may have irrelevant consequences for enriching nascent strands and negligible effects compared to the influence of λ-exo. Nonetheless, to directly test whether the results from our method, which used BND cellulose and less enzyme, is comparable with previous methods, we recently performed the Cayrou et al protocol [Cayrou et al., 2011] that uses a higher enzyme:DNA ratio than we used and sucrose gradients instead of BND cellulose. We did this on MCF10A cells in contrast to MCF7. Despite the different cell line and different technique, that dataset agrees with our MCF7 dataset by ~80%, dispelling the myth that sucrose gradient enrichment of nascent DNA prior to the digestion step yields different or better results after digestion than when BND is used as the pre-enrichment step. These results are further detailed in the next subsection. Despite these arguments, it would appear this group performed good controls, and even repeated some of our experiments, and came to different answers than us. It is difficult to explain why this is the case for now.

It is enlightening to look at the results of other origin mapping techniques to see if they agree with the estimates from the Cayrou et al papers that 78% of initiation sites are associated with G4 motifs [Cayrou et al., 2011, Cayrou et al., 2012a, Cayrou et al., 2015]. As discussed in our review, all other methods and most other groups performing the λ-exo protocol report lower associations of putative origins with G4 motifs, mostly below 40%. Nonetheless, new papers have been published that bear on the importance of G4s at origins. A newer method presented by Langley and colleagues [Langley et al., 2016], called Ini-Seq, is based on a cell-free DNA replication initiation assay. Late G1 nuclei that have been synchronized with mimosine are released into S-phase for a very brief incubation with labeled dUTP that can be immuno-precipitated later after DNA extraction and fragmentation. Thus, this method can only identify early origins with the assumption that the cell-free nature of the experiment and mimosine synchronization do not affect the origin distribution and efficiency. These assumptions may not hold, however. Mesner and colleagues demonstrated 56.3% of detected bubble-containing fragments after mimosine synchronization were not detected or had very low signal in the asynchronous bubble-trapped libraries. They posit that this "relatively standard synchronizing regimen" activates "a subset of origins that is otherwise inefficient or dormant in undisturbed cultures". Moreover, the same group that published on Ini-seq demonstrated previously that mimosine arrest causes pervasive double-strand breaks (DSB) concomitant with entry into S-phase [Szüts and Krude, 2004]. These caveats must be considered when interpreting Ini-seq data in addition to the caveat that late origins are largely excluded. Ini-seq identified ~25,000 putative origins, ~48% of which overlapped G4 motifs, with the majority of these G4-proximal peaks in promoters. This is a slightly higher association with G4 motifs than estimates from other techniques, but still much lower than 78%. Given that there is a prevalence of G4s in promoters, that G4 density is higher in early replicating domains than late replicating domains, and

that this technique is biased to early origins in gene promoters, the slightly higher overlap with G4s than other techniques is quite expected. Importantly, over 50% of Ini-seq peaks did not overlap G4 motifs and this suggests once again that they are not necessary for DNA replication initiation to take place through an independent technique. Similar conclusions can be found in other recent papers. Bartholdy and colleagues used BrdU-immunoprecipitated nascent strands to map DNA replication origins genome-wide [Bartholdy et al., 2015]. Importantly, they identified origins in their dataset that had single nucleotide polymorphisms (SNPs) and/or small insertions/deletions (indels). This natural variation of particular origins allowed them to perform allele-specific analyses analogous to performing 250 knock-in mutation experiments. Using this approach, they investigated the question of whether G-quadruplex (G4) motifs have an effect on the efficiency of the origins that they overlap with and found that in the majority of cases, the origin allele with the disrupted G4 motif had higher efficiency than the allele with the intact G4 motif. Overall, they assert that for the subset of origins associated with G4 motifs, the presence of a G4-forming sequence is not necessary for origin formation and is actually associated more generally with decreased origin efficiency. Petryk and colleagues published their results and analyses from sequencing Okazaki fragments strand-specifically across the human genome. They conclude that most initiation sites are not associated with G4 motifs and say, "Our data are consistent with the proposal that enrichment of SNS [[short nascent strands]] in CGI [[CpG-Islands]] and other CG-rich sequences is due to an intrinsic bias of the λ-exonuclease technique", citing our paper. Another recent paper by Miotto and colleagues [Miotto et al., 2016] mapped ORC binding sites throughout the human genome and found that only 31% of ORC binding sites contained G4 motifs and only 26% were located in CpG islands, demonstrating that these features are not required for ORC binding. Nevertheless, it has been demonstrated that initiation need not occur close to ORC binding sites [Gros et al., 2015] and that MCM are distributed widely across non-transcribed chromatin [Powell et al., 2015]. So this result does not reject the possibility that ORC loads MCMs that then slide up to or near G4s before DNA replication initiates from them, and it may be reasonable to suggest that if G4 structures are formed in vivo, they could impede MCM sliding. This would likely happen at only a subset of origins and is consistent with our data where only approximately a third of putative initiation sites have a G4 within 1 kb that is typically spaced apart from the initiation site at distances consistent with intervening nucleosomes. Given that G4s have been associated with nucleosome depleted regions and may be involved in nucleosome positioning [Iyer and Struhl, 1995, Hershman et al., 2008, Halder et al., 2009, Wong and Huppert, 2009, Schwarzbauer et al., 2012, Foulk et al., 2015], they may have roles in establishing or maintaining chromatin environments that in turn influences the locations of preferred initiation sites of nearby initiation zones, origin efficiencies therein, and when nearby origins fire during S-phase. Nonetheless, though origin-proximal G4s may modulate nearby origin activity, the majority of data overwhelmingly suggest that G4s do not have a necessary role for the initiation of DNA replication.

Another opportunity for insight into this question is determining whether or not the G4 motifs actually form structures near origins in vivo. Although it has been shown that they can form in

vitro at physiological conditions and there has been tantalizing indirect evidence that they form in vivo (such as being hotspots for double-strand breaks), the direct evidence for their in vivo formation has been scant, largely resting on the visualization of G4s in the telomeres of a ciliate [Huppert, 2010, Bochman et al., 2012]. In 2013, Biffi and colleagues [Biffi et al., 2013] isolated a G4 structure-specific single-chain antibody that they named BG4. BG4 visualization provided direct evidence of in vivo G4 formation in human cells at both telomeric and non-telomeric sites. Their results also suggested the differential occurrence of G4 formation over the cell cycle. Specifically, the fewest G4s were detected in M and G0/G1 and increasing nuclear G4 foci counts were detected as the cells entered and progressed through S-phase. The same group has now used this G4 antibody for ChIP-seq to identify specific sites where G4s actually form structures in vivo at high resolution [Hänsel-Hertsch et al., 2016]. They detected ~10,000 sites of G4 formation, far fewer than computational predictions and far fewer than were detected in vitro by G4-seq [Chambers et al., 2015], another study from the same group. What does this combination of results suggest about G4 structures at replication origins? If it is true that there are fewer G4s formed in M and G1 than in S-phase and that more G4s form as S-phase progresses, then the G4s may be more likely to form in the ssDNA regions of replication forks or in the relatively decondensed chromatin after replication forks pass. In either case, this would indicate that they are most likely to form after replication initiation occurs during a time when the pre-RC cannot re-bind, and that they largely disappear by the time the pre-RC can bind again. This makes it hard to imagine that they are used for origin selection. Moreover, if it is true that there are only approximately 10,000 sites where G4s form in vivo at all, then how does one explain the many tens of thousands or hundreds of thousands of λ-exo peaks near equivalent numbers of G4 motifs throughout the genome? One explanation might come from the in vitro G4-seq method that identified hundreds of thousands of G4 motifs across the genome that formed G4 structures in vitro, particularly in the presence of potassium. This would support our original premise that a large part of the correlation of λ-exo peaks with G4 motifs is the direct result of G4 structures forming in vitro and impeding λ-exo digestion while folded thereby increasing the half-life of 3' protected DNA, especially since the λ-exo digestion buffer contains potassium. However, the results of BG4 visualization and BG4 ChIP-seq are not definitive. It is possible that BG4 only binds well-exposed G4 structures. If this is true, then G4 accessibility may be what varied over the cell cycle rather than G4 formation and BG4 ChIP-seq may miss many G4s, particularly in heterochromatic regions, and may not lead to accurate quantitative measurements of the frequency of in vivo G4 formation at each site. Nonetheless, the BG4 results so far seem to suggest once again that in vivo G4 structures cannot possibly have a necessary function in initiation of DNA replication at the majority of initiation sites. Still, this does not reflect on possible modulatory effects G4s may have on origin location, efficiency, and timing when they are in close proximity.

6.8.2 NS-seq with sucrose gradient on MCF10A cells, a comparison with our previous protocol.

To test whether results from our protocol were comparable to the Cayrou protocol [Cayrou et al., 2011], we followed that protocol to prepare nascent strands from MCF10A cells. For the MCF10A genomic DNA control, we obtained 146,292,441 50 bp mappable reads and for the MCF10A NS-seq data, we obtained 103,521,212 mappable reads. Bowtie2 was used to map reads and SAMtools was used to work with the alignment files:

bowtie2 -p 8 -t --very-sensitive -N 1 -x hg19 -U reads.fastq | samtools view -F 4 -bSh - | samtools sort - aligned.reads.sorted

We used MACS2 to call peaks:

macs2 callpeak -t NSseq.bam -c gdna.bam --down-sample -g hs -n mcf10a.ns.gdna -q 1e-3 --keep-dup 1 --nomodel --shiftsize=175 --slocal 5000 --llocal 50000 -B

This resulted in 98,445 MCF10A NS_{G0gDNA} peaks (not controlled for λ-exo biases). We found that 76,650 (77.86%) overlapped with MCF7 NS_{G0gDNA} peaks. That is to say that there was nearly 80% overlap between the approach that used BND-cellulose and a lower λ-exo enzyme:DNA ratio with the approach that used the sucrose gradient and a higher λ-exo enzyme:DNA ratio when the approaches were used on these two different human cell lines. We also looked at the overlap between MCF10A NS_{G0gDNA} with MCF7 $LexoG0_{G0gDNA}$ to see how much of this dataset was unique to replicating cells. We found that 43,189 (43.87%) of the MCF10A peaks overlapped with MCF7 $LexoG0_{G0gDNA}$, leaving ~56% of the peak set unique to replicating MCF10A cells. Since we did not do a LexoG0 control for MCF10A, we cannot control for λ-exo as we did with MCF7. However, we found that 55,918 MCF10A peaks (56.80%) overlapped MCF7 NS-seq peaks that were controlled for λ-exo biases (NS_{LexoG0}). Thus, in our hands, the sucrose gradient protocol results in a similar amount of peaks unique to replicating cells (~56%) as the BND cellulose protocol (53%) did, both of which are higher than the 6-30% from other datasets. Nonetheless, as we saw previously for MCF7, a fairly large proportion of NS-seq peaks can be found in the non-replicating control when NS-seq is not controlled for λ-exo biases.

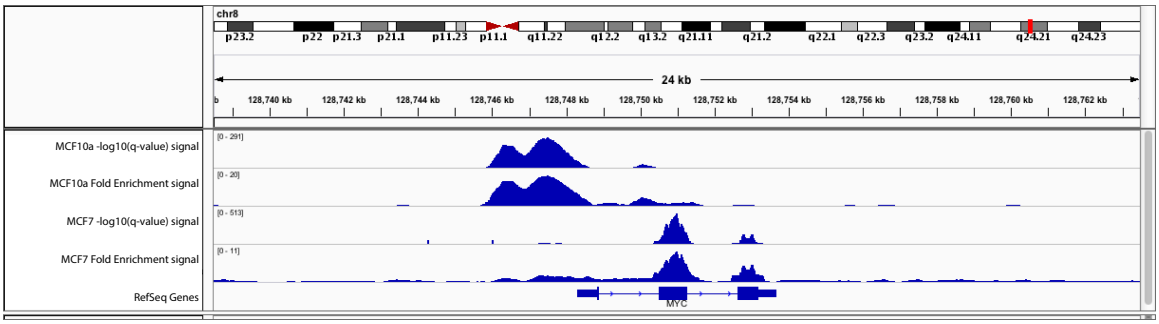


Figure 6.8: MCF10A.

The MYC locus. A look at NS-seq signal from NS_{G0gDNA} from MCF10A (top 2 tracks, -log10(qvalue) signal followed by fold-enricment signal) and MCF7 (next 2 tracks, same descriptions) at the MYC locus. Replication initiation activity is well-characterized in the promoter region and in the second exon. In MCF7 cells, we see a stronger NS-seq signal in the second exon whereas we see a stronger signal in the promoter area in MCF10A. Since nearly 80% of the peaks found in MCF10A are found in MCF7, this seems to potentially demonstrate plasticity of initiation activity rather than a difference in the methods, though it is not definitive.

6.8.3 Future Directions

Local method.

When analyzing the rDNA locus, fold enrichment values were fully local. Thus, the local effects of λ-exo were fully controlled. However, MACS2 was used to call peaks genome-wide. While the way MACS2 works can control for local λ-exo biases in enriched regions, it may result in false negatives in λ-exo-depleted regions. MACS2 is conservative when peak calling in that when it compares the treatment (NS-seq) to the control (either genomic DNA or LexoG0), it considers three different values to use for the control at each position in the genome: the global average, an average in a small local window, and an average in a larger local window. It then uses the largest of those averages to test if the current window is enriched in the treatment or not. Thus, in areas where the local average goes below the global average, the global average is chosen. Given what we know about λ-exo strongly depleting AT-rich DNA and enriching GC-rich DNA, this indicates that the global average is likely most often chosen in AT-rich regions whereas a local average is most often chosen in GC-rich regions. The latter is the desired behavior and controls for local Lexo biases. However, the former effect of using the higher global average in depleted AT-rich regions means that there is likely a higher incidence of false negatives in AT-rich regions (in this case, these are systematic false negatives due to the procedure, not random chance). Thus, moving forward, a fully local approach as done at the rDNA locus example is warranted. Such an approach would be predicted to identify more peaks, the majority of which would likely be AT-rich. A fully local approach can be done with MACS2 algorithms in the following way:

1. For both NS-seq and LexoG0 Control replicates, filter mapped reads for mapping quality and/or remove redundant/duplicate reads.
2. Optionally, if NS-seq or LexoG0 has more reads than the other, down-sample the one with more reads to equal the number of reads in the one with fewer reads.
3. Use 'macs2 pileup' for the NS-seq sample to create NS-seq bedGraph. A recommended value for --extsize is the fragment size of the sequencing library.
4. Do same thing for the LexoG0 control sample.
5. If down-sampling was not done in step 2, then normalize the sequencing depth of NS-seq and LexoG0 Control to each other by down-scaling. If NS-seq has X times more reads than LexoG0 Control, then divide the 4th column in NS-seq bedGraph by X. Conversely, if LexoG0 Control has X times more more reads than NS-seq, then divide the 4th column in LexoG0 bedGraph by X. Alternatively, for each file, divide 4th column by number of reads in that sample, then multiply by some constant (e.g. the number of reads in smallest replicate).
6. Compare the normalized bedGraphs from the previous step using 'macs2 bdgcmp -m qpois' to get a -log10(q-value) track by comparing ChIP with only local bias.
7. Call regions using 'macs2 bdgpeakcall'. Set cutoff '-c' as your qvalue cutoff (for example, use '-c 5' for a qvalue cutoff of 1e-5). It is also advisable to set minimum peak length using '-l' and a maximum distance between peaks used for merging them into the same enrichment (-g). A recommended size

for minimum peak length is somewhere between the fragment length used for sequencing and the minimum nascent strand size selected for in the experiment (e.g. 200-750 bp).

In one example of this, on the same read sets used for the original peak calling (which were down-sampled to have equal numbers of reads and already filtered for redundant reads), I did:

```
macs2 pileup -i reads.bam --extsize 200
macs2 bdgcmp -t NS.bedGraph -c LexoG0.bedGraph -m qpois
macs2 bdgpeakcall -i NSlexoG0.qpois.bedGraph -c 3 -l 200 -g 50
```

This is not strictly analogous to the original peak-calling where --extsize was set to 350 (also used as the length cutoff) and the qpois cutoff was set to 5. However, it is possible to approximately obtain the comparable set by filtering for peaks 350 bp or longer with q scores in the narrowPeak output of at least 50 (comparable to -c 5).

There were 338,593 $NS_{LocalLexoG0}$ peaks that were largely AT-rich as we saw previously (6.9 A-B). This new peak set encompassed 98.17% of the 66,831 peaks in the original NS_{LexoG0} peak set. When requiring the peaks be at least 350 bp, the number dropped to 125,591 peaks encompassing 89.24% of the original set. The percent overlap rises a bit if one first merges peaks within 200 bp of each other before removing peaks shorter than 350 bp. This results in 118,502 peaks that overlap 90.7% of the original set. In the latter example, the GC content of the ~ 9% of peaks from the original set that were not represented in the local-bias-only set was shifted toward GC-rich peaks (6.9 C-D). Nonetheless, the GC content distribution of the newer peak set was similar to that of the older one and did include some GC-rich peaks, even when looking at only the novel peaks in the set (6.9 E-F). The majority of peaks, though, had GC content that was not extreme with respect to the genomic average, which is slightly AT-rich. This result was expected given what we learned about λ-exo depleting AT-rich regions and that MACS2 would likely miss these due to defaulting to the global lambda value in these regions.

Signal correction in absence of LexoG0 control.

One practical concern is on whether it is possible to remove λ-exonuclease biases in NS-seq datasets that do not have accompanying datasets that control for them. Since we have genome-wide data that characterizes the effects of λ-exo, it is possible to use that information to correct NS-seq signal for future and past experiments. First to obtain conservative scaling factors, given local nucleotide composition we can do the following:

1. Break the genome up into 100 bp windows.
2. Get the fold-enrichment information (LexoG0/gDNA) and GC content information over each window.

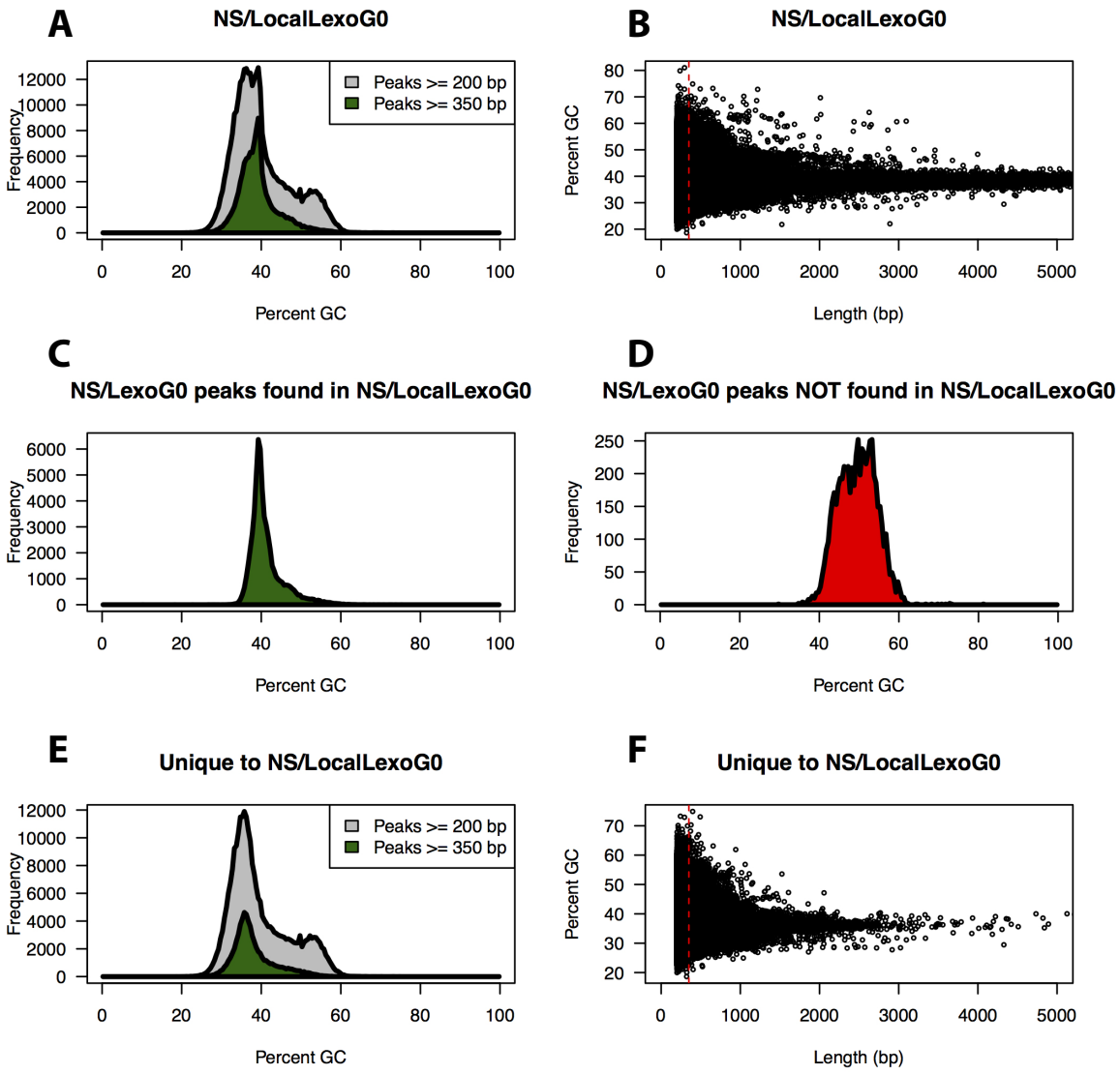


Figure 6.9: Peak calling for NS-seq when strictly controlling for local LexoG0 biases.

3. For each GC content value (0-100), obtain the median fold enrichment value (GC_m). This is the median Lexo bias for that GC content.
4. Keep these values stored for future reference.
5. These values can be updated as new LexoG0 datasets arise.

Second, to correct either the signal from NS-seq alone or from NS/gDNA, do the following:
1. Break the genome up into 100 bp windows.
2. Get the coverage or fold-enrichment information over each position.
3. Using stored GC_m values from above, correct the value (x) in all windows by: x/GC_m
4. Perform peak calling on bedGraph of λ-exo-normalized values (for example, follow the fully local method outlined above).

This is work in progress.

Potential method for identifying transition points at known origins with NS-seq data.
Transcription factors bind specific sites in the genome. For ChIP-seq, chromatin is fragmented, then an antibody against the target transcription factor is used to pull out DNA fragments bound to that factor. For a given binding site, fragmentation can occur such that the binding site is to the left side of the fragment, to the right side of it, or any where in between. Sequencing the 5' ends of these fragments therefore results in two approximately normal distributions of reads on opposite strands, one to each side of the binding site [Park, 2009]. One can then perform various methods to find the center between these two distributions, which is estimated as the approximate binding site. Similarly, if an origin of replication fires from a specific point-source-like location in the genome (as with yeast origins), size-selected origin-centered nascent strands would create two distributions of reads to either side of the origin (Fig. 6.10). The center of these two distributions would be a reasonable estimate for the approximate location of the transition point from leading to lagging strand synthesis. This would not necessarily work in large initiation zones where initiation can occur anywhere, although it may identify the transition point for preferred sites in a large initiation zone if any exist.

The distributions of 5' ends on the positive and negative strands can be processed several ways to give useful information about the transition point. In the simulation, for figure 6.10 A, for each bp (bp_x), I counted the number of 5' ends mapped to the positive and negative strands ($c_{x,+}$ and $c_{x,-}$). In 6.10B, to give each bp a single count, the number of 5' ends starting at a given bp on the negative strand is subtracted from the number starting there on the positive strand and lightly Loess-smoothed (let this be called the strand score): $SS_x = c_{x,+} - c_{x,-}$ The strand score signal crosses the 0-line at the transition point as expected. This switch from positive to negative strand score values can be converted into a peak detectable by peak-calling methods where the summit of each peak represents the most likely transition point for a given region. To convert the strand score

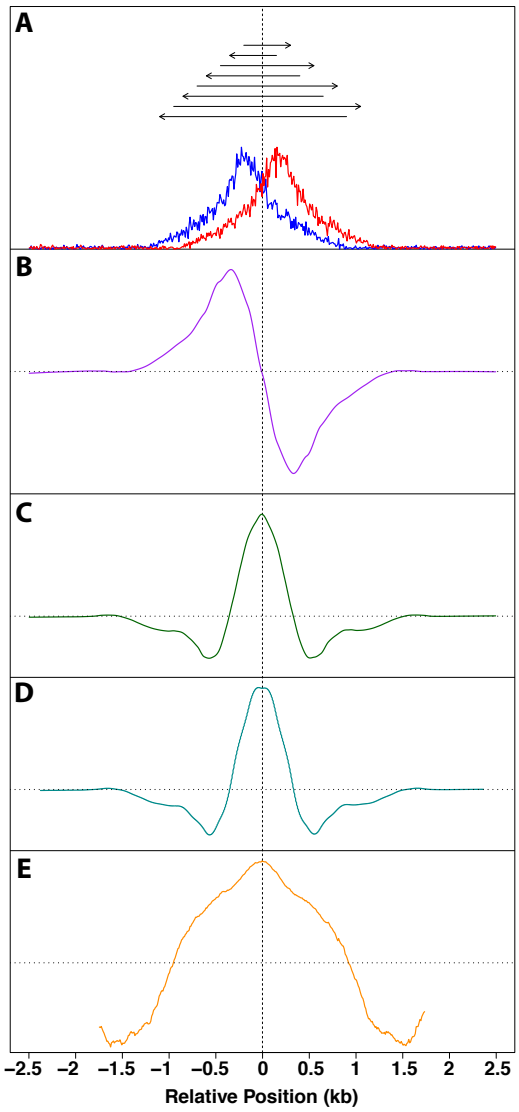


Figure 6.10: Simulated NS-seq data from a point source origin.
(A) Example of size-selected bidirectional origin-centered nascent strands from a point source origin and simulated distribution of 5' ends on the positive (blue) and negative (red) strands. **(B)** Calculating the Strand Score over each base from 5' end information on each strand over that base. **(C)** Converting the Strand Scores to Transition Point Scores using a window size of 1. **(D)** Converting the Strand Scores to Transition Point Scores using a window size of 25. **(E)** Using the Origin Efficiency Metric approach on the stranded 5' end information [Smith and Whitehouse, 2012, McGuffee et al., 2013].

signal into a transition point peak signal in Figure 6.10C, the transition point score for each bp x, TPS_x, was calculated by subtracting the strand score of bp x+1 from the strand score of bp x: $TPS_x = SS_x – SS_{x+1}$. Similarly, for Figure 6.10D, the sum of strand scores in 25 bp windows to the left and right of bp x were used for this calculation: $TPS_x = \text{sum}(SS_{(x-24):x}) – \text{sum}(SS_{(x+1):(x+25)})$. Finally, the same approach used to calculate the Origin Efficiency Metric (OEM) for Okazaki Fragment sequencing in yeast can be used as in Figure 6.10E [Smith and Whitehouse, 2012, McGuffee et al., 2013]. Negative peaks for the OEM approach on Okazaki fragment data from yeast represent termination zones. However, it is unclear for both approaches mentioned above what the negative peaks indicate for NS-seq data and are ignored for now.

I searched for transition points at the MYC and DBF4 loci (Figures 6.11 and 6.12) using our NS-seq data [Foulk et al., 2015]. This appears to have worked as expected. For example, there is a transition point inside the HindIII-XhoI fragment that contains the well-studied MYC origin, as well as a transition point in the second exon where initiation activity also seems to peak by other methods. Moreover, there seems to be a transition point in the bidirectional promoter at the DBF4 locus. Nonetheless, despite the apparent good fortune seen at these targeted sites in the NS-seq data, such transition point peaks spuriously appear quite often in the genomic DNA control and simulations show that enrichments from G-quadruplexes (or from any other feature that is enriched by Lexo) would unfortunately give false transition point signals (Fig. 6.13). Indeed, transition point peaks also arise in the LexoG0 data and this analysis faces similar problems to those faced in our paper when origins and G-quadruplexes are in the same region of the genome. Interestingly, the coverage from summing 5' ends from both strands in this G4 simulation gives a shape and offset from the G4 that is very familiar in the literature for λ-exo enrichment of nascent strands (Fig. 6.13F). Altogether, this analysis method is likely only useful to identify probable transition points in known origins, which could be piloted on yeast where all origins are known. It will take more development then these pilot analyses to make use of such a technique for de novo identification of origins. One avenue that may have potential is adjusting the transition point peak heights by coverage or fold enrichment values. As mentioned in Urban et al (2015), strand-specific information could strengthen NS-seq analyses, including this one. For example, an experimental way to have better resolution of the transition point between continuous and discontinuous replication at the bi-directional origin is to block Okazaki fragment synthesis with emetine [Burhans et al., 1991] followed by one of the aforementioned nascent strand enrichment protocols (with controls) and strand-specific paired-end sequencing to visualize the jump in continuos synthesis from one strand to the other at the origin. This is the same principle as used for Okazaki fragment synthesis but would be employed here for leading strand synthesis.

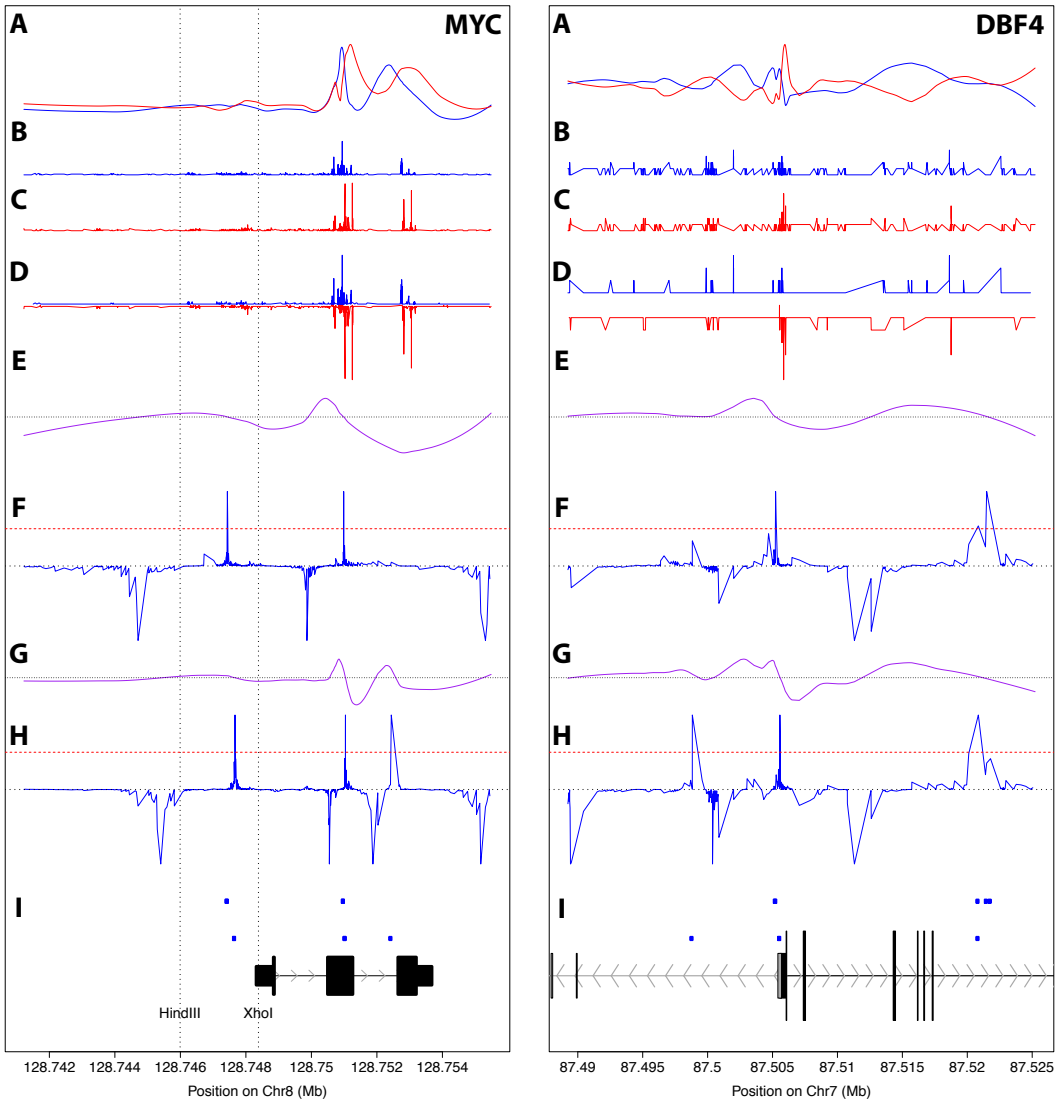


Figure 6.11: Workflow for identifying potential transition points at the MYC and DBF4 loci.

(A) Loess smoothed distribution of 5' ends on the positive (blue) and negative (red) strands. **(B)** Raw counts of 5' ends on positive strand. **(C)** Raw counts of 5' ends on negative strand. textbf(D) Raw strand scores. textbf(E) Strand score signal after smoothing. textbf(F) Transition point signal corresponding to strand score signal in E. textbf(G) Strand score signal after less aggressive smoothing. textbf(H) Transition point signal corresponding to strand score signal in G. **(I)** Blue dots are putative transition points. Gene bodies shown for MYC (left) and and SLC25A40/DBF4 (right). Dotted lines in MYC panel represent boundaries of the HindIII-XhoI fragment with known origin.

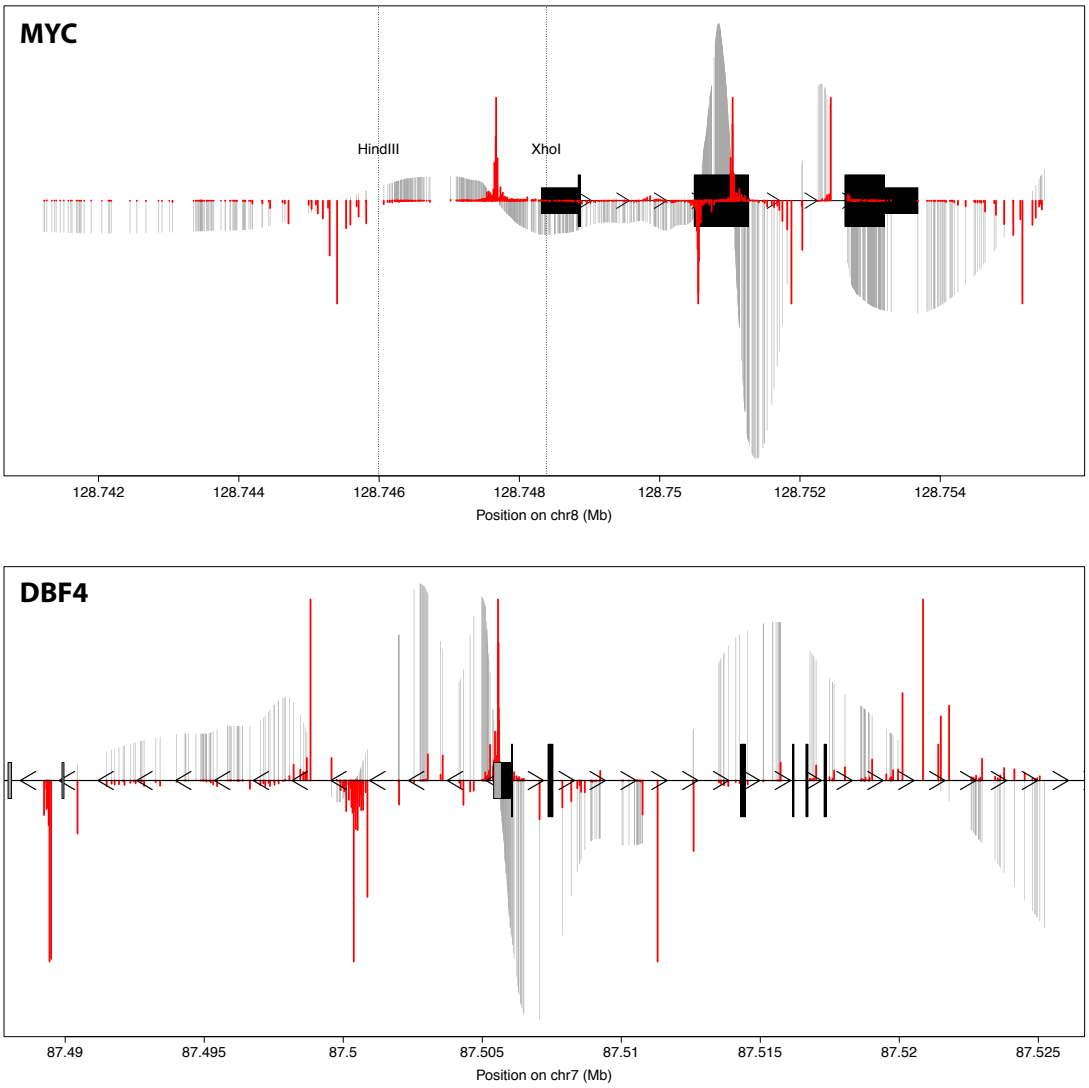


Figure 6.12: Potential transition points at the MYC and DBF4 loci.
Different view of strand score (grey) and transition point (red) signals from figure 6.11 G, H, and I.

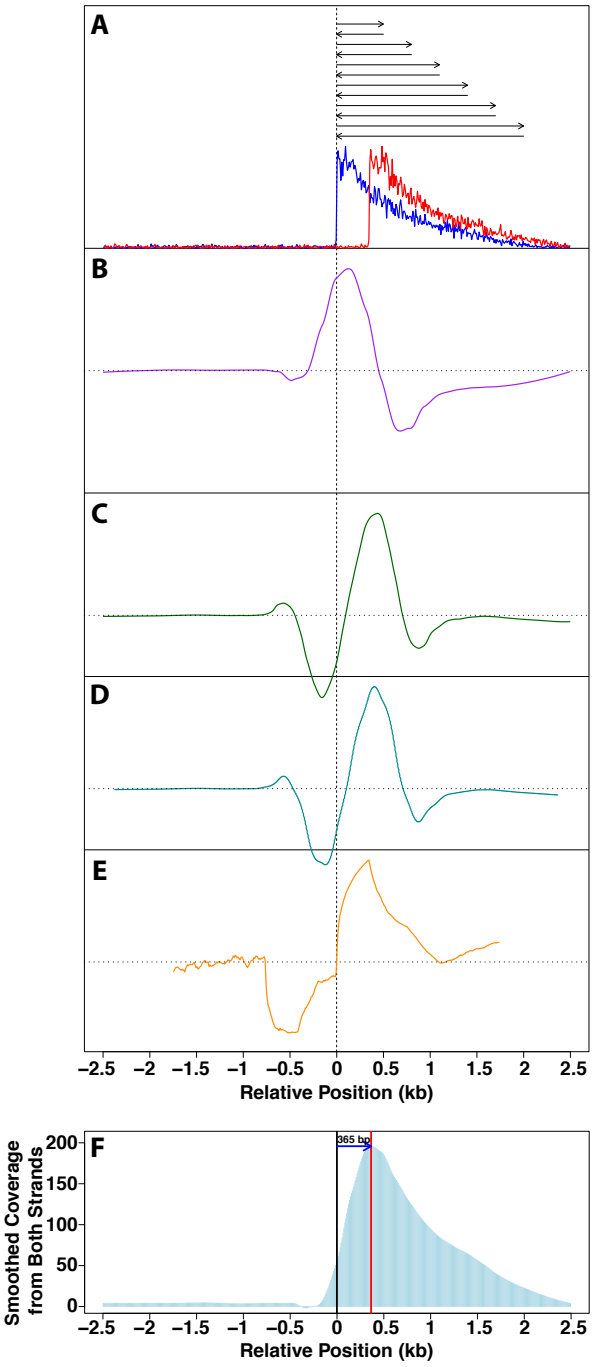


Figure 6.13: Simulated NS-seq data from a G-quadruplex structure.
(A-E) Same as in figure 6.10 except the arrows in (A) illustrate an example of size-selected G4-protected DNA strands (non-G4 strand is represented after fragmentation and PCR) from a point source G4 and simulation was conducted accordingly. **(F)** The coverage from summing 5' ends from both strands gives a shape and offset from the G4 that is very familiar in the literature for NS-seq and λ-exo enrichment studies.

6.8.4 Mapping yeast origins with NS-seq.

With the help of a new laboratory member, Miiko Sokka, we have begun improving and validating NS-seq by applying it to the genome of budding yeast (Saccharomyces cerevisiae) where the ORIs have been mapped by multiple orthogonal approaches, including early firing ORIs (Raghuraman et al 2001; Yabuki et al 2002), ORC binding (Wyrick et al 2001; Xu et al 2006; Eaton et al. 2010), mapping single-stranded DNA after hydroxyurea treatment (Feng et al. 2006), phylogenetic footprinting (Nieduszynski et al 2006), and marker frequency (Mller et al 2014). Many origins have been confirmed by 2D gels and autonomously repicating sequence assays. These data on yeast ORIs are catalogued in the OriDB database (Nieduszynski et al 2007; Siow et al 2012). The richness of origin mapping data in budding yeast makes it uniquely suitable for developing origin mapping methods for eukaryotes. To date, λ-exo-based NS-seq has not been tested on yeast, nor have other deep sequencing origin discovery techniques for metazoans including bubble-trap [Mesner et al., 2013], ini-seq [Langley et al., 2016], nascent strand capture and release [Kunnev et al., 2015a,Kunnev et al., 2015b]. Demonstrating that a method works on yeast does not guarantee the method will work as well on metazoans. Indeed gold standards such as the autonomously replicating sequence assay and 2D gels are examples of methods that worked great in yeast and less so in metazoans. Yeast origins are markedly different from metazoan origins. Moreover, the yeast genome is relatively homogenous in nucleotide content whereas metazoan genome tend to have a lot of variation in nucleotide content. Thus, in the case of λ-exo, there may be less influence of nucleotide content on digestion from bin to bin across the genome. Nonetheless, one would minimally like to have confidence that these techniques do what they claim in a well-characterized system. Thus, we want to ensure that NS-seq holds up to the yeast origin test. Is it sensitive? That is, does it detect most origins? Is it specific? That is, does it detect anything other than origins in yeast? To ensure it is both sensitive and specific, we plan to adopt a paired-end strand-specific approach that will allow for better detection and removal of potential biases as discussed in our review [Urban et al., 2015a]. We will also test alternate buffers, such as those that contain Na^+ ions rather than K^+ ions as discussed in this chapter. Alternative controls for lambda exonuclease biases other than using a non-replicating (G0) control will be explored, such as removing the RNA primer from the control set of replicating cells before kinase treatment. It has also been reported that removing the RNA primer and skipping kinase treatment leaves the unphosphorylated nascent strands even more protected from digestion than RNA primers [Yang et al., 2013]. One way to control biases in both the kinase and RNase reactions while ontaining a nascent strand test sample and a control sample would be split a sample of nascent strands into two equal portions. For one portion, first perform kinase treatment followed by RNase treatment, leaving unphosphorylated nascent DNA and phosphorylated parental DNA. For the other portion, the control, perform the RNase treatment first followed by kinase treatment, leaving all nascent strands and parental DNA phosphorylated. Treat both with λ-exo prepared from the same pooled reaction for the same amount of time. In the future, MinION single-molecule long-read strategies can be performed as well, which would encompass the same benefits as the paired-end, strand-specific Illumina approach. For the time being though, the

throughput of the MinION is too low for this particular method where very deep sequencing is important. Nonetheless, the MinION can be exploited in other ways for origin mapping, which we proposed to Oxford Nanopore in 2013 and is discussed elsewhere in this thesis.

CHAPTER 7

Development of single-molecule, genome-wide origin mapping methods.

7.1 Prologue

We have been interested in mapping replication origins with new methods. In particular we favor single-molecule methods that would allow the discrimination of single replication events. Current high throughput sequencing methods strictly give a population-level view of replication events, inevitably favoring high frequency events and hiding low-frequency events. It is the equivalent of being given an average without knowing anything else about the distribution. In biology, there are many interesting observations to be made in the variance and the shape of a distribution. Single-molecule techniques allow us to look at single events and compile them together to also get the population-level view. This chapter describes my ongoing work in collaboration with others to develop single-molecule genome-wide methods for studying DNA replication.

7.2 DNA combing

In 1966, Huberman and Riggs [Huberman and Riggs, 1966, Huberman and Riggs, 1968] presented a technique called DNA fiber autoradiography that pulsed replicating cells with ^{3}H-thymidine for it to be incorporated by replication forks moving along the genome in vivo. DNA fibers were then gently extracted from the cells and stretched out to visualize the radioactive replication tracts. It could be seen in pulse-chase experiments that two replication forks moved away from each other, where fork directionality was inferred by observing high amounts of radiation where the pulse began

to lower amounts as it moved along during the chase. These experiments confirmed the bidirectional nature of DNA replication in eukaryotic chromosomes. The modern adaptation of DNA fiber autoradiography comes in the form of DNA combing [Bensimon et al., 1994] and SMARD (single molecule analysis of replicating DNA) [Norio and Schildkraut, 2001], typically differentiated by how DNA is stretched out on microscope slides (machine or free-hand respectively). Henceforth, I will simply refer to these techniques as DNA combing experiments. In DNA combing experiments, replicating cells are incubated with nucleotide analogs (BrdU, CldU, EdU, etc) that are incorporated into the nascent DNA by moving replication forks [Norio and Schildkraut, 2001, Bensimon et al., 1994, Michalet et al., 1997, Herrick and Bensimon, 1999, Pasero et al., 2002, Anglana et al., 2003, Patel et al., 2006, Lebofsky et al., 2006, Herrick and Bensimon, 2009, Bianco et al., 2012, Técher et al., 2013, De Carli et al., 2016]. Labeled ultra long DNA molecules >100 kb are stretched out on glass microscope slides. The nucleotide analogs can then be visualized various ways under a microscope. Instead of ^{3}H-thymidine silver grain gradients in the overlying photographic emulsion to infer directionality, when two analogs are used, one after another, the nascent DNA tracts containing each can be visualized as different colors, red and green for example. If the 'red analog' is used first, followed by the 'green analog', then one can tell the fork was moving in the direction of red to green (Figure 7.1 A-B). When two replication forks are moving away from an origin of replication, a double fork pattern can be observed (Figure 7.1 C-D). Similarly, when two forks are converging on each other in an oncoming termination event, the inverse fork pattern can be observed (Figure 7.1 E-F). The length of the labeled segments can give information on fork speed. The distance between double fork patterns from replication origins can give inter-origin distance information for origins that fire on the same molecule in the same cell in the same window of time. DNA molecules representing the entire genome can be visualized in a single experiment. However, it is not possible by pulse-labeling alone to know where in the genome each replication fork pattern originates. This has been overcome by combining Fluorescence in situ hybridization (FISH) with DNA combing to interrogate a single locus (Figure 7.1 G). This is both low throughput and laborious. In a more ambitious approach, initiation sites across megabase regions have been inferred by creating a Morse code pattern of FISH probes along the given genomic region [Lebofsky et al., 2006] (Figure 7.1 H). The pattern allows one to infer where within the megabase stretch of sequence initiation events occur, though with low resolution. This latter approach could be powerful if applied genome-wide.

7.3 Mapping replication origins on single molecules with genome-wide Morse code

The Morse Code idea [Lebofsky et al., 2006], where the patterns from specific labeled sequences or sites along single DNA molecules act as a genomic address, has seen genome-wide equivalents [Schwartz et al., 1993, Lin et al., 1999, Neely et al., 2011, Lam et al., 2012, Mendelowitz and

Pop, 2014]. However, these optical mapping methods have largely been used for scaffolding genome assemblies and inferring structural variation. The first was the Argus technology from OpGen featuring the first generation of optical mapping technology [Schwartz et al., 1993, Lin et al., 1999] where DNA molecules that are stretched out on slides are subject to an in situ restriction enzyme digest. The stretched out DNA is stained and the restriction fragment lengths between the dark unstained cut sites are measured to derive the Morse code pattern of restriction sites. When one obtains many such ordered restriction maps from randomly selected DNA molecules from a given genome, overlapping portions of the restriction maps can be detected between DNA molecules. Ultimately, overlapping restriction patterns can be assembled to represent full chromosome restriction site maps similar to assembling long reads by detecting overlaps [Lin et al., 1999]. BioNano Genomics (BNG) came out with a higher throughput recognition sequence mapping method where DNA molecules are nicked with a nicking-endonuclease instead of being fully digested. The nicked sites are then briefly nick-translated in the presence of fluorescent nucleotides to allow visualization of the recognition sites [Lam et al., 2012]. Keeping the DNA molecules intact allows BNG to keep DNA in solution and iterate over a procedure of pulling DNA molecules into nano-channels where they can be visualized and ejected for new ones to take their place. Nonetheless, obtaining ordered Morse code patterns for many randomly selected DNA molecules to ultimately determine genome-scale maps is still the goal of this method. NabSys is another company that works with DNA in solution for recognition site mapping. Similar to BioNano, they nick recognition sites for subsequent labeling. However, optics are not involved in label detection. Instead, NabSys detects the recognition sites as they are driven through semiconductor-based nanodetectors at over 1 million bases per second (http://www.nabsys.com), and refer to this as electronic mapping. Genomic Vision has performed locus-specific Morse Code analyses visualized on microscope slides, but is yet to have a genome-scale equivalent. Any of these technologies, some more than others, present possibilities of being coupled with labeling replication tracts for a Morse Code approach to mapping replication origins in the genome as was done for a single 1.5 Mb locus [Lebofsky et al., 2006]. We proposed this idea to BNG in 2013. A limitation in the approach using BNG is that they can only visualize 3 colors: one for staining the DNA, one for visualizing the Morse code pattern of recognition sites, and one for replication tracts. Nonetheless, there are ways to use 3-colors that we discussed with BNG. Unfortunately, they were unwilling to provide support unless we bought a BioNano Irys machine, which we were unable to do. Although other groups will likely get this to work, we have heard from colleagues who have been trying that it has not been without frustration. Fortunately, Susan and I proposed this technique to another company who was willing to support its development. However, though I made the proposal, it is Yutaka Yamamoto in the laboratory who has been carrying out the bench experiments whereas I will carry out the analyses. We are testing multiple approaches that include nicking endonucleases, CRISPR, and FISH. Our goal is to test it on the yeast genome where all or most origins are well characterized. One group has recently used the nick-translation approach with Lambda DNA molecules in replicating frog egg extracts [De Carli et al., 2016]. More of my time has been given to a related technique, which I ultimately think is more promising and

discuss in the next section.

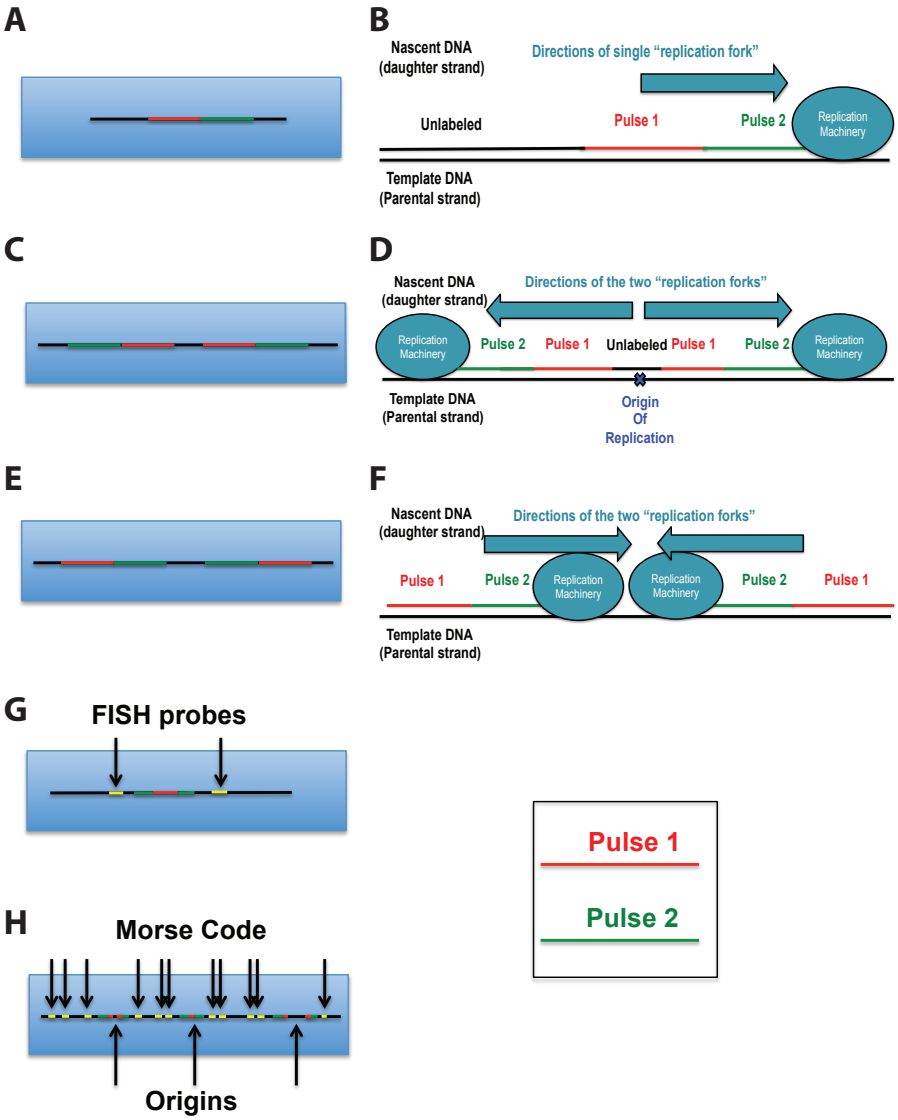


Figure 7.1: DNA combing experiments.

(A) Example of single replication fork tract, a "single fork" pattern. **(B)** Depiction of replication machinery direction leading to that pattern given the sequence of pulses. **(C)** Example of a double fork pattern emanating from an origin. **(D)** Depiction of replication fork directions moving away from origin. **(E)** Example double fork pattern of converging replication forks. **(F)** Depiction of replication forks converging in termination event. **(G)** Example of using two FISH probes to study a locus-specific origin. **(H)** Example of Morse Code pattern across megabase range molecule and double fork patterns from origins therein. Morse code patterns made with FISH probes were used this way for a 1.5 Mb region previously [Lebofsky et al., 2006], and more recently nick translation was used to do this with the 48.5 kb Lambda phage genome [De Carli et al., 2016]. Companies such as OpGen, BioNano Genomics, NabSys, and Genomic Vision use Morse Code patterns for genomic analyses. We have been working on ways to couple genome-wide Morse Code patterns with replication tract labeling to identify origins genome-wide.

7.4 Using the MinION to detect nucleotide analogs incorporated by DNA polymerase for single-molecule genome-wide studies of DNA replication.

7.4.1 Demonstrating a feasible approach to identifying replication tracts in MinION data

Over the past three decades, nanopore sequencing was shown to be able to discriminate DNA and RNA bases from each other, including methylated cytosine from cytosine [Branton et al., 2008, Deamer et al., 2016]. Oxford Nanopore Technologies (ONT) was founded and began developing commercial nanopore sequencing technology in 2005. In 2012, Clive Brown, the CTO of Oxford Nanopore, announced that ONT was able to sequence the entire lambda phage genome that consists of a single 48.5 kb molecule in a single read. At that point it struck me that combining the long read lengths and the ability to discriminate different bases would allow for high throughput sequencing version of the older DNA replication mapping technique called DNA combing. In this version of the technique, instead of visualizing the red-to-green tracts using fluorescence microscopy, the nucleotide analogs would be distinguished as part of the DNA sequence during base-calling (after learning how to differentiate them from the canonical bases). Red and green replication tracts could thereafter be distinguished at the sequence level from single molecule reads (Figure 7.2). This would allow us to map origins of replication on single DNA molecules, each one representing a single event (Figure 7.2 A-C). In some cases, the co-regulation of nearby origins on the same long molecule may be detected. Measuring the origin firing efficiency of each origin, could be accomplished just by counting the origins from a given site detected on single molecules. Initiation and termination zones could be analyzed, again at the single molecule level, potentially allowing the reconstruction of what is seen in ensemble data such as Okazaki fragment mapping [Smith and Whitehouse, 2012, McGuffee et al., 2013, Petryk et al., 2016]. The technique would allow us to map the statistical distributions of polymerase speed and acceleration over each base of the genome, and find polymerase pause sites, slow zones, and fast zones (Figure 7.2 D). This technique would allow us to understand the directionality of replication across the genome, letting us see the proportion of rightward moving forks and leftward moving forks over any given bp (Figure 7.2 D). This has been inferred by other techniques, but the MinION technique is singular in that it allows us to observe a collection of single replication events rather than signals that arise from an ensemble of replication events. Nonetheless, we can also look at the ensemble of single replication events to recapitulate what is seen in ensemble methods. For example, one could look at the population of molecules across the genome for population-level strand-switch signatures of replication origins (bulk changes in fork directionality) in addition to those detected on single molecules from double fork patterns. The population-level view should look similar to Okazaki fragment sequencing results. The single-molecule view would allow us to test some hypotheses that are raised from population-level Okazaki fragment data. For example, the strand-switch regions showing origins in Okazaki-fragment data are likely only areas

that are replicated by initiation more frequently than being passively replicated whereas initiation from areas that are passively replicated more often than not are masked at the population level. Since we could look at single-molecule double fork patterns in addition to population level strand-switches, we could show if that is true (or not).Overall, using the MinION to update the older DNA combing technique would be extremely powerful for studying multiple facets of the DNA replication program. Once the technique is established, using it in various mutant backgrounds makes it even more appealing. However, one drawback of this technique that may be persistent is the need to work with cells and tissues that can take up nucleotide analogs.

In 2013, Oxford Nanopore announced a program to beta test the Minion called the MinION Access Program (MAP). I wrote an application for MAP proposing to use their technology to try to develop an application of the MinION for studying DNA replication. This application won us a spot in the first round of MAP inductees. Subsequently, this study received funding from a Brown University seed grant and a NSF grant.

For the DNA replication application of the MinION, I saw six problems to solve when I embarked on this project: (i) getting very long reads, (ii) creating software to generally work with the MinION data, (iii) creating a base-caller that can discriminate nucleotide analogs X and Y from A, C, G, and T, (iv) identifying tracts of X and Y along DNA molecules, (v) identifying single fork and double fork patterns, (vi) mapping this information to the genome to perform single-molecule and population-level analyses. Below I describe progress we have made toward solving these problems.

7.4.2 Extending MinION Read Lengths

In order to ensure a high probability of capturing double fork patterns (e.g. from forks moving away from origins) and multiple double fork patterns on the same molecule, one needs to sequence long molecules. Toward this end, I developed modifications to the standard ONT protocols to increase the read lengths. Early developments were featured in a biorxiv preprint (`http://biorxiv.org/content/early/2015/06/22/019281`). I have since refined them and adapted them to subsequent kits and protocols released by ONT. Overall, these modified protocols more than double the read length N50 and maximum read lengths, drastically increase the amount of data from molecules >10 kb, and require no special reagents in addition to what is needed for the standard protocols. I have obtained many >100 kb reads with these protocols for a genome assembly project for the fungus fly, Sciara coprophila. They align with >80% identity to assemblies created by using PacBio data alone, thereby validating both the structure of the assembly in those regions and the validity of the ultra long reads we obtained. In addition to these protocols, I attempted to develop other methods to ensure that the majority of DNA molecules in a library are >100 kb. One method was in collaboration with Intact Genomics, a company that specializes in working with megabase DNA,

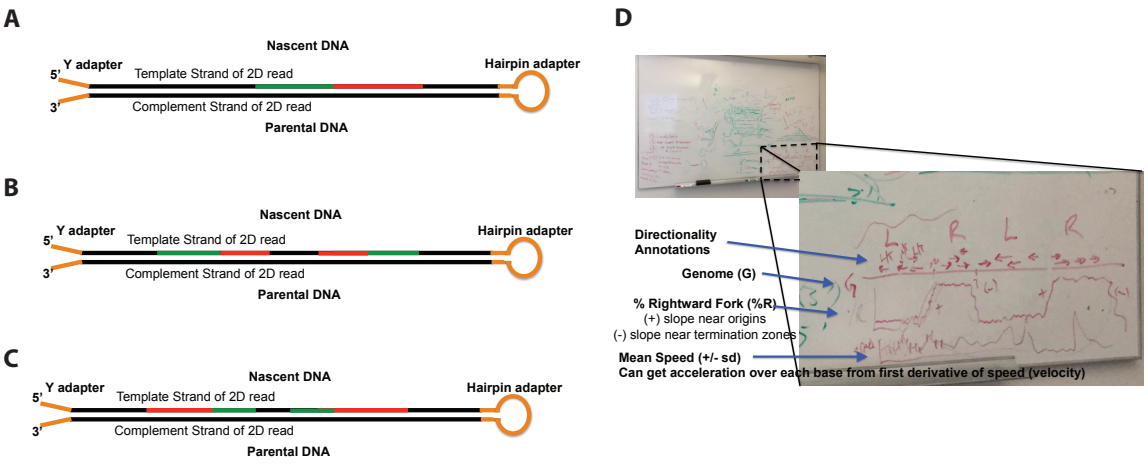


Figure 7.2: Studying DNA replication with the MinION.

As in DNA combining, cells are pulsed with a "red" analog, followed by a "green" analog. Instead of stretching the DNA molecules out on slides for visualization, MinION libraries are prepared with care to keep the DNA molecules as long as possible. DNA molecules that can be sequenced by the MinION can have a lead adapter and a hairpin adapter that allows the complementary strand to be sequenced as well. With information from both the nascent and parental strands of a replicating molecule, higher accuracy inferences could be drawn for identifying tracts with nucleotide analogs. However, it is likely not necessary. The illustrations here demonstrate what the combing patterns would look like on DNA molecules that have a lead adapter and a hairpin. **(A)** Example of single replication fork tract, a "single fork" pattern. **(B)** Example of a double fork pattern emanating from an origin. **(C)** Example double fork pattern of converging replication forks. **(D)** Not only can replication events be studied on single DNA molecules, but replication events from the population of molecules can be used to map statistical distributions on fork directionality and fork speed to the genome. Population-level techniques can be employed to reconstruct ensemble datasets whereas the information from single molecules can tell us about things that are hidden at the ensemble level. This image was a picture I took a couple years ago when telling Taehee Lee about these ideas.

and involved cutting >100 kb DNA out of Pulsed-Field Gels. Another method I worked on was performing library preparations completely in agarose microbeads [Koob and Szybalski, 1992, Zhang et al., 2012]. However, the libraries from both methods have not yet been successfully sequenced. In both cases, it seems that even tiny bits of left over agarose in the sequencing library immediately kills the MinION flow cells. Nonetheless, the protocols that I have successfully working are more than sufficient for this technique.

7.4.3 Working with MinION data

I have developed a suites of software for working with nanopore data called poreminion (`https://github.com/JohnUrban/poreminion`) and fast5tools (`https://github.com/JohnUrban/fast5tools`). These tools are able to work with the HDF5 files and extract and/or summarize all relevant information for working with the base-called sequences and signal events data. They create fasta/fastq files with all information needed to do quality and other filtering in their headers, and optionally create headers compatible with certain assemblers and assembly tools, such as Falcon and DAligner.

7.4.4 Developing a local base-caller

To design a base-caller, I used a hidden Markov model (HMM) approach [Durbin et al., 1998]. Base-calling nanopore data using HMMs was already demonstrated previously [Timp et al., 2012] and was the approach ONT used in their own base-caller, but they were unwilling to share the code at the time.

To understand how one can model and base-call indirect signals that arise from DNA sequences, it is worthwhile understanding first how DNA sequences can be modeled. One way DNA sequences can be modeled is by assuming complete independence of neighboring nucleotides. In this model the probability of seeing an A, C, G, or T corresponds to the frequency the given letter shows up in some sequence if interest. If 'A' occurs 35% of the time, then the chance of seeing an A is thought to be 35%, independent of neighboring bases. If you were generating a sequence of length L, you could think of it as rolling a single, possibly biased, 4-sided die L times. However, DNA sequences are usually better modeled by Markov chains. A Markov chain models the dependencies on the previous base(s) seen. For example, the probability of seeing a G differs if the previous base was an A, C, or T. In terms of generating a sequence from a Markov chain, one can think of it as rolling 4 different 4-sided dice one at a time, where the base that shows up in one roll specifies which die to use in the subsequent roll. For example, if 'A' is rolled, the next die that is used is the 'A' die that specifies the conditional probabilities of seeing A, C, G, or T next given the last base was an 'A'. If 'C' is rolled next, the die corresponding to the conditional probabilities from 'C' is used subsequently. For a sequence of length L, this repeats for L rolls, switching to the specified die for each new roll. The die used for the first roll is specified by a set of initial probabilities. Sometimes we cannot observe

the Markov chain directly, but instead observe a sequence of emissions that correlate with the underlying Markov chain. In this scenario, the Markov chain becomes a sequence of hidden states that we are interested in. Thus, given a sequence of emissions, one wants to know the states that gave rise to those emissions. Hidden Markov models allow us to infer those hidden states. In nanopore sequencing, the observable emissions are ionic current measurements, and we are interested in the underlying sequence of bases that gave rise to them.

A simple example application of HMMs is named the "Occasionally Dishonest Casino" [Durbin et al., 1998] where we observe a series of dice rolls over time from two dice, one fair and one biased, though we do not know when each is being used. Only one is used at a time, but each is often used multiple times in a row. Given the sequence of values from the dice rolls, we want to know which die was used for each roll. Thus, there are two underlying hidden states: fair and biased. To infer the hidden states, the first thing to consider is the initial probability of starting in either state. Does the person rolling the dice choose to start with either uniformly at random, or do they start with one more often then the other? These are called the initial probabilities. The second thing to consider is how often the person rolling the dice switches between the fair and biased dice. When the fair die was rolled last time, how likely is it to be used again for the next roll? How likely is it to switch to the biased die? The same questions apply when the biased die was rolled last. These are called the transition probabilities. Each state has a probability of staying in the same state or switching to the other state. Some times we know the transition probabilities and other times we need to learn them. The third thing to consider is the emission probabilities of each die, the set of probabilities for seeing each value. We know one die is fair. Therefore, for that die, the probability of seeing each of the 6 values (1-6) is equal: 1/6. However, the other is biased and the probabilities of seeing each value are not equal. In some situations, we may know the emission probabilities already. In others, we have to learn them. A similar problem to the dishonest casino, is that of CpG islands in genome sequences. One can think of a given DNA sequence as a series of rolls from two 4-sided dice (2 states) with different sets of emission and transition probabilities. In this problem, the emissions make up the observable DNA sequence and the hidden states are CpG islands and non-CpG islands. Using HMM tools, the underlying states can be inferred in order to define locations of CpG islands in the given genome sequence.

A state path is a hypothetical sequence of underlying states that could potentially explain the observed emissions (e.g. dice rolls). If one knows the three sets of probabilities discussed above, then one could simply calculate the probability of all possible state paths and choose the one with the highest probability. However, as the sequence of emissions gets longer and longer, the number of possible state paths grows exponentially. For example, in a 2-state model, there are 2^L possible state paths for an emission sequence of length L. This becomes even more problematic in scenarios with higher numbers of possible states. Fortunately, there is a class of algorithms called Dynamic Programming algorithms that allows us to avoid this brute force approach [Durbin et al., 1998].

There are three dynamic programming algorithms commonly associated with HMMs: the Forward, the Backward, and the Viterbi. If one already knows the initial, transition, and emission probabilities, then the Viterbi can be used straight away to identify the maximum likelihood state path. Alternatively, one can use the Forward and Backward algorithms to identify the Posterior Decoded state path, which identifies the state for each roll that has the highest posterior probability given all of the data. Some times this can result in adjacent states in the state path that are extremely unlikely or impossible to occur next to each other. In many scenarios, the Viterbi state path and the Posterior Decoded state path are the same. If any of the aforementioned probabilities need to be learned, the Baum-Welch algorithm, which is in the class of Expectation Maximization algorithms, is often employed [Durbin et al., 1998]. This allows one to start with random guesses at the parameters (the probabilities) to calculate an initial log likelihood of the model. It then uses the Forward and Backward algorithms with the given parameters, from which one can calculate expected counts of transitions and emissions given the data (e.g. the sequence of values from dice rolls) for each state to update the parameters. The log likelihood of the model with the updated parameters can then be calculated. This is iterated until the difference between the log likelihoods converges toward zero, typically stopping when the difference is smaller than some threshold for some number of iterations. The log likelihood of the model typically gets better with each iteration and the goal is to stop the algorithm when the log likelihood stops getting better. However, this does not guarantee the global optimum as it can get stuck in local optima. One way to get closer to the global optimum is to run the Baum-Welch multiple times, starting each time with a different set of random guesses at the parameters. Then one can take the set of parameters with the highest log likelihood across multiple instances of the Baum Welch.

The MinION samples data quickly such that there are shifting clouds of data points over time as DNA is pulled through the pore. It uses a segmentation algorithm to turn the shifting clouds of data points into events, each of which consists of a mean, a standard deviation, a start time, and a duration. We are given the sequence of events. Though it is likely all four pieces of information can be used, to simplify we can choose to only consider the event means. In base-calling the MinION data, the observable emissions are the nanopore event means and the hidden states are the 1024 5-mers that gave rise to them. We can think of the sequence of event means as a sequence of rolls from 1024 possible dice. In the occasionally dishonest casino and CpG island problems above, there are high probabilities of staying in the same state (i.e. using the same die as the last roll). However, in this problem the transition probability of moving to the next overlapping 5-mer (i.e. switching to another die) is higher than staying in the current 5-mer. Moreover, in the 2-state CpG island HMM, the DNA sequence is envisioned as series of rolls from two 4-sided dice, where each die had a set of 4 discrete probabilities. For the base-calling problem, instead of discrete probabilities, each 5-mer gives rise to a continuous distribution of possible emission values that can be characterized by a mean (μ) and standard deviation (σ). Here the event means are envisioned as a series of rolls from 1024 dice, where each die (state) emits values according to its distribution parameters. ONT

had already learned these parameters for all 5-mers containing the canonical bases: A, C, G, and T. Moreover, they provided those parameters for users to work with. Thus, most of the work for using HMMs was already done.

For more insight into how base-calling the event means works, it is easier to pretend we are base-calling individual bases, or 1-mers, rather than 5-mers. In this scenario, there are only 4 hidden states: A, C, G, T. Each state emits values according to a distribution defined by μ and σ. For example:

	μ	σ
A	40	2.0
C	50	2.5
G	60	2.2
T	70	1.8

Each state transitions to itself and the other 3 states defined by a transition probability matrix. For this example, we can assume uniform transitions. However, transitions need not be uniform, and we could easily learn better guesses by looking at the underlying genome sequence that MinION data is expected to come from:

	A	C	G	T
A	0.25	0.25	0.25	0.25
C	0.25	0.25	0.25	0.25
G	0.25	0.25	0.25	0.25
T	0.25	0.25	0.25	0.25

Then of course there is an initial probability of starting in any state. Again, we can assume uniform probabilities here for simplicity, but it would be trivial to get better estimates by looking at the underlying genome sequence and using the proportions of each base:

	initial
A	0.25
C	0.25
G	0.25
T	0.25

Event means are our observed emissions. We could also use both the sequence of emitted means and emitted standard deviations in an HMM where each state has two emissions. However, for simplicity of explanation we will look only at the emitted mean values.

Suppose the following sequence of event means is observed, generated using the above model:
61.33837, 36.95254, 73.61558, 72.06623, 49.38346, 41.01925, 61.82503, 72.51779, 54.14916, 61.73285

Because these emissions were generated using the model and a given state path, the state path is known. Let the states be numbered 1, 2, 3, and 4 where 1=A, 2=C, 3=G, 4=T. The state path used was:
3 1 4 4 2 1 3 4 2 3

This state path can easily be translated into a base sequence since we are using 1-mers:
GATTCAGTCG

The Viterbi algorithm or the Forward/Backward to do Posterior Decoding can be used to find a path through the states A, C, G, T that "best" explains the emissions. In this overly simplified simulation, both get it exactly correct. Base-calling gets slightly more complicated with dimers where there are 16 states with emission probabilities defined by μ and σ. For simplicity, uniform initial probabilities can be assumed, though better estimates can easily be gained by looking at the frequencies that each dimer occurs in the underlying genome sequence where the data is coming from. Each of the 16 states also has transition probabilities. Importantly, each dimer in a DNA sequence can only transition to 4 other dimers that start with the base that the previous overlapping dimer ended with. However, recall that a segmentation algorithm employed by ONT summarizes the raw stream of data into events. Often segmentation is correct, but sometimes over-segmentation can occur leading to multiple events for the same k-mer. Other times, under-segmentation can occur leading to one event for multiple adjacent k-mers. Thus, in addition to modeling transitions to the next overlapping dimer in a sequence, a real world model needs to allow transitions back to self (stay probabilities) and transitions to a non-overlapping dimer (skip probabilities). Nonetheless, in the simplest model without skips and stays, each dimer can only transition to 4 other dimers where their prefix is the same as the current dimers suffix (e.g. AT can switch to TA,TC,TG,TT). We can use uniform transition probabilities for these for demonstration purposes (see matrix below), although they can easily be learned by frequencies from the underlying genome. Everything else is set to 0 when assuming no skips and stays:

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	0.25	0.25	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0
AC	0	0	0	0	0.25	0.25	0.25	0.25	0	0	0	0	0	0	0	0
AG	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0.25	0	0	0	0
AT	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0.25
CA	0.25	0.25	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0
CC	0	0	0	0	0.25	0.25	0.25	0.25	0	0	0	0	0	0	0	0
CG	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0.25	0	0	0	0
CT	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0.25
GA	0.25	0.25	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0
GC	0	0	0	0	0.25	0.25	0.25	0.25	0	0	0	0	0	0	0	0
GG	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0.25	0	0	0	0
GT	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0.25
TA	0.25	0.25	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0
TC	0	0	0	0	0.25	0.25	0.25	0.25	0	0	0	0	0	0	0	0
TG	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0.25	0	0	0	0
TT	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0.25

Below are example emission parameters for the 16 dimers in alphabetical order:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
73.77	68.39	61.57	54.92	59.8	65.17	67.8	65.76	60.42	63.33	53.97	76.2	69.6	71.07	72.84	62.73
1.45	0.78	0.81	0.62	0.93	2.16	1.17	1	1.04	1.04	1.41	0.9	1.26	0.93	0.53	1.1

As illustrated in this example, many of the dimers have emission parameters that specify similar distributions (e.g. AC and CG or GC and TT in this randomized example). This is also seen in Figure 7.3 A. If one were to treat the emissions from a sequence of event means independently of the rest of the data, then assigning hidden states would be much more error-prone, as simulated in Figure 7.3 B-C. The dimer that gives the highest probability of seeing an emission will frequently be wrong when treating the emissions independently. Moreover, the probability of seeing a given emission can be very similar for multiple dimers. However, when considering all of the data and the entire path through the states, the transition probabilities given the previous dimer change the probabilistic landscape drastically. Whereas probabilities of seeing an emission may be similar for multiple dimers when assuming independence, there is often a much clearer winner amongst those dimers when modeling the dependence on the previous dimer. In the simple model where all but four transitions have probabilities of zero, most possibilities will be non-existent. This is illustrative of why HMMs, that model dependence on previous states, are a good choice for this problem. One could again use the Viterbi algorithm or Posterior Decoding to find a state path that best explains the data. In simulations, this gave sequences with $> 90\%$ accuracy (Fig. 7.3 B, D, E).

In the simplified dimer model above, translating the state path into a nucleotide sequence is straight forward. To initiate the sequence, start with the dimer from the first state in the path. For all subsequent states in the path that consist entirely of overlapping dimers, only add the second letter in the dimer to the growing sequence. In a more complicated model, where skips and stays are modeled by replacing the zeros in the transition matrix with non-zero probabilities, translating

state paths into nucleotide sequences becomes more complicated. When a dimer transitions to one of its neighboring dimers (those that share its suffix as their prefix), it is called a move of 1. If a dimer transitions to any other dimers aside from the 4 that share its suffix as a prefix, then it is a move > 1. It is not possible to know for sure if it was a move of 2, where a single nucleotide is skipped, or greater. However, without knowing the underlying sequence, we would have to assume it was a move of 2. If a dimer transitions back to itself (e.g. CA to CA), it is called a move of 0 or a "stay". However, note that when a homo-dimer (e.g. AA) transitions to itself, this could have come from a homopolymer (e.g. AAA). This is why detecting the true length of homopolymers is still a weakness of this technology. In summary, for dimers, the transition moves can be 0, 1, and 2. When translating the state path into a nucleotide sequence, one needs to account for the sequence of moves associated with the state path. Start the sequence as an empty string and add the dimer from the first state. For all subsequent states: (i) if the move was 0, then add nothing to the base sequence and move to the next state, (ii) if the move was 1, then add the second letter of that state's dimer to the base sequence and move to the next state, (iii) if the move was 2, add the entire dimer from the current state and move on to the next state.

For longer k-mers there are more possible moves in base-calling in that longer skips are detectable. For example, for 3-mers, the base-calling moves can be 0, 1, 2 and 3. For 4-mers, a move of 4 is introduced. Then there are 5-mers, which is what is used for base-calling MinION data (at the time), where moves of 0, 1, 2, 3, 4, and 5 are possible. These moves are modeled in the transition probability matrix. The Viterbi algorithm and/or Posterior Decoding are employed to find the best state paths explaining the sequence of events, as discussed for shorter k-mers. Translating the state path into a nucleotide sequence follows similar logic to that given for dimers. One starts the sequence with the 5-mer from the first state. For all subsequent states, to generalize for all k-mers, let 'm' be the move value for the current state and add the last m letters of the current k-mer (i.e. the suffix of length m) to the growing sequence and move on to the next k-mer. Thus, for 5-mers:
- For a move of 0, add 0 letters from the current 5-mer and move on.
- For a move of 1, add the 1 letter suffix of the current 5-mer.
- For a move of 2, add the 2 letter suffix from the current 5-mer.
- For a move of 3, add the 3 letter suffix from the current 5-mer.
- For a move of 4, add the 4 letter suffix from the current 5-mer.
- For a move of 5, add the entire 5-mer.

On simulated data, this works well achieving $> 90\%$ accuracy even in this much larger model of 1024 states (Fig. 7.4). Nonetheless, even on a large set of simulated data, these algorithms can be quite computationally demanding. Moreover, though I vectorized the code as much as possible and tried both R and Python, both were too slow. In addition, real world MinION data offered much bigger hurdles. Applying the above algorithms as stated is not totally sufficient for base-calling. Though it gave sequences with up to 70% accuracy on a few Lambda DNA reads, for most

Lambda DNA molecules it was much worse. This problem arose because each individual read has signal correction parameters that need to be estimated in order to first correct the event means for factors such as shift, drift, and scale. After signal correction, the HMM algorithms can be applied. Moreover, though one can start off with a likely set of transition parameters, it is beneficial and sometimes crucial to run the Baum-Welch to update the parameters given the data for each individual read. This makes the computational needs for base-calling a set of tens to hundreds of thousands of MinION reads even more demanding, and also requires a great deal of parallelization of operations. For these reasons, we eventually recruited Charles Lawrence and his student Taehee Lee who specialize in HMMs and probabilistic modeling more generally as well as Mark Howison and his associate David Berenbaum who specialize in High Performance Computing. Moreover, to help generate DNA sequences with nucleotide analogs for training and for yeast labeling experiments, we began collaborating with Nick Rhind and his student Victor Liu. However, before we could make much more progress on updating our base-caller, a local base-caller was published by Jared Simpson [David et al., 2016]. Around the same time, another group published a recurrent neural network (RNN) base-caller [Boža et al., 2016]. Soon thereafter, ONT released their own RNN base-caller. Therefore, it behooves us to use the optimized code already made available. Nonetheless, working through the HMM logic and writing code that simulated and decode nanopore data was an extremely important exercise in my graduate career.

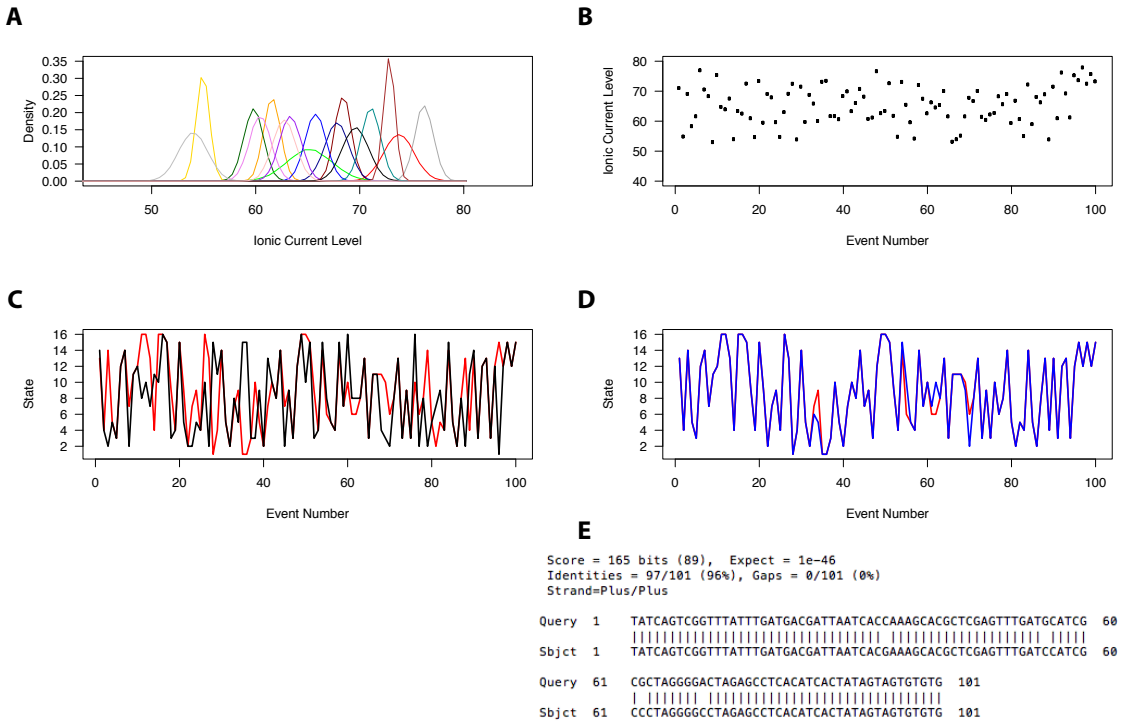


Figure 7.3: Base-calling simulated nanopore data where ionic current corresponds to dimers.

(A) The simulated emissions distributions of the 16 dimers, demonstrating a high degree of overlap. **(B)** Sequence of observed emissions generated given simulated initial, transition, and emission probabilities for the 16 dimer states. **(C)** The known state path that gave rise to the observed emissions in (B) in red, and the state path from treating all emissions independently in black. Due to the high overlap of the emissions distributions shown in (A), this approach only gets the correct states 55% of the time. This is why much more of the red line is seen compared to (D). The independent approach produces a sequence that cannot be aligned by BLAST. **(D)** The known state path is shown in red and the state paths from Viterbi algorithm and Posterior-Decoding that give identical answers here are represented in blue. They get the state correct 92% of the time in this example. This is why the majority of the red line is hidden. **(E)** BLAST [Altschul et al., 1990] alignment of the known sequence (subject) and the Viterbi state path translated sequence (query) has 96% identity. This is slightly higher identity than the 92% accurately predicted states since an additional percentage of wrong states can still provide the correct nucleotides to append by chance. Overall, (D) and (E) demonstrate the superiority of using a hidden Markov model approach for base-calling nanopore data since the sequence of underlying states is a Markov chain of overlapping dimers.

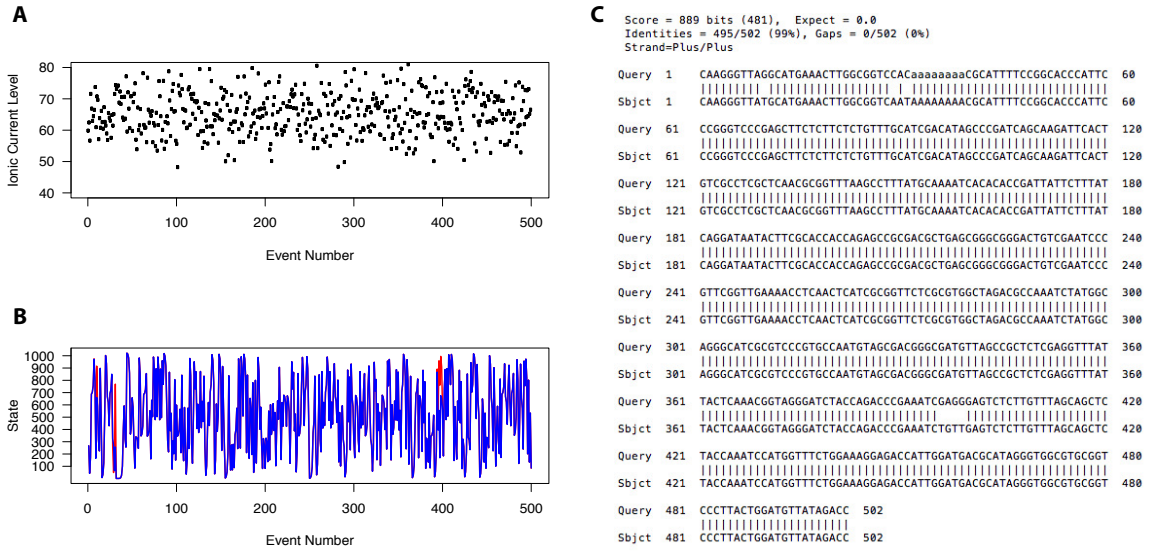


Figure 7.4: Base-calling simulated nanopore data where ionic current corresponds to 5-mers.

(A) Sequence of observed emissions generated from an underlying DNA sequence and given simulated emissions parameters for all 5-mers. **(B)** The known state path is shown in red and the state paths from Viterbi algorithm and Posterior-Decoding that give identical answers here are represented in blue. They get the state correct 95.6% of the time in this example. This is why the majority of the red line is hidden. **(C)** BLAST [Altschul et al., 1990] alignment of the known sequence (subject) and the Viterbi state path translated sequence (query) has 99% identity. This is slightly higher identity than the 95.6% accurately predicted states since an additional percentage of wrong states can still provide the correct nucleotides to append by chance.

7.4.5 Using a Hidden Markov Model to find DNA replication tracts on DNA molecules.

Ultimately, this experiment calls for labeling DNA molecules in vivo by incubating cells with one nucleotide analog, followed by a second nucleotide analog. Labeling is followed by extracting genomic DNA, preparing long read MinION libraries, and inferring what pieces of each DNA molecule, if any, are labeled with each analog. When I originally envisioned the solution to this problem, I saw it in two steps: (i) base-calling with an expanded model where the emissions parameters for 5mers (or 6mers) with analogs has been learned, and (ii) interrogating the base-called sequence to probabilistically identify tracts of incorporated analog. Using generative models as described above, I generated simulated emissions that included randomly generated emissions parameters for analog-containing 5mers. Using my HMM base-caller I was able to translate the simulated events into sequences of A, C, G, T, X, and Y, where X and Y represented nucleotide analogs that replaced T. The next problem was to identify high confidence stretches of X-incorporation and Y-incorporation. To do this I used yet another HMM approach described in Durbin et al [Durbin et al., 1998]. In that book, as described in a previous section, a 2-state HMM is used to classify stretches of CpG islands and non-CpG islands in a genomic sequence. The idea is that CpG islands have different frequencies of A, C, G, and T than other regions in the genome. To define the boundaries of CpG islands, HMM algorithms such as the Viterbi or Posterior Decoding can be used to define the most likely state path or most likely state over each nucleotide in the sequence given the set of emission and transition parameters for each state. Identifying analog-incorporated replication tracts in base-called DNA molecules is similar, but they contains three possible states: unlabeled, label-X, and label-Y. The emission parameters are the probabilities of seeing A, C, G, T, X, and Y in each state. The emissions for A, C, and G are constant for each state as DNA molecules can come from anywhere in the genome for all states. The emissions for T, X, and Y are different for each state though. For example, the unlabeled state has low probabilities of seeing X and Y, which are present from error in base-calling whereas the label-X state is expected to have the highest frequency of X and label-Y should have the highest frequency of Y. The emission parameters for T, X, and Y can be approximated for each state by understanding the incorporation frequency of the nucleotide analogs. As mentioned, the unlabeled state should be given small non-zero probabilities of seeing X or Y to account for errors in base-calling. Finally, a 4th state is also possible called "unlabeled-2" where seeing X and Y might be slightly higher than the background error rate from bleed through of the labeling procedure, but lower than label-1 and label-2 states. The initial probabilities and transition probabilities for each state can be approximated by knowing the duration of time used for each label in the experiment and how that correlates with expected label lengths from expected replication fork speeds. This also allows you to estimate what proportion of DNA molecules is expected to have label. When running simulations using this consecutive approach of HMM base-calling followed by a 3-state HMM to segment the DNA sequences into unlabeled and labeled states were successful, I was able to compare the segmentation boundaries identified by the 3-state HMM in each simulated DNA molecule to the known state path that generated them. Overall, it successfully identified the replication tracts in

these simulations. Nonetheless, when we began collaborating with Charles Lawrence he pointed out that this approach puts too much emphasis on the single base-calling solution. Instead, he proposed it would be better to do segmentation simultaneous with base-calling using a 2-layered HMM where the probabilities of all base-calling solutions can be propagated into the segmentation layer. We have also explored ideas of abandoning base-calling per se, and just probabilistically aligning the events to the genome sequence and segmenting into the labeled states in parallel. This work is ongoing, and is described in a little more detail in the following section.

7.4.6 Learning emission parameters for sequences with T-analogs

In this section I describe other progress we have made.

(i) **Designing a sequence to learn emission parameters for T-analogs:** We need to learn parameters for all possible 6-mers that contain at least one T-analog. To do this we need a sequence that contains all possible 6mers (originally 5mers). The shortest sequence containing all 6mers can be obtained by creating a deBruijn graph by connecting all 6-mers that overlap by 5 bases, then following an Eulerian cycle through that graph. This is called a "deBruijn Sequence". There are many Eulerian cycles in this graph. I chose one. Nick and Victor had it synthesized in 3 pieces that were cloned in vectors and that Victor subsequently used for PCR to create products with either regular T or a T analog (e.g. BrdU). However, we found that one of the pieces was extremely GC-rich with plenty of G-quadruplex (G4) motifs and was resistant to PCR. Therefore, I iterated over Eulerian cycles in the deBruijn graph until I obtained a sequence that was homogenous with respect to GC-content (i.e., it was distributed uniformly across the sequence and not concentrated in any on area). I then identified G4 motifs and broke them up by adding AT-rich 6mers while also making sure to retain all possible 6mers in the sequence. Nick and Victor ordered this final sequence for synthesis in 4 pieces. We received them cloned in vectors and have subsequently used them for PCR and MinION sequencing successfully.

(ii) **Multiplexing and de-multiplexing MinION flow cells and data:** To learn emissions of sequences with T analogs such as BrdU, we need to perform PCR with only A, C, G and the analog, X. We have to do this for all 4 PCR products, for multiple possible analogs, as well as with regular T. Thus, there are numerous conditions. Since the ONT base-caller cannot make sense out of the analog-incorporated sequences, we would not be able to put all these conditions on the same flow cell and separate them later at the base-sequence level. Moreover, ONTs barcoding and de-multiplexing approach only works for data the base-caller deems as high quality. However, the analog-incorporated sequences are seen as low quality by the base-caller and are therefore be put in the "fail" folder without de-multiplexing. Therefore, with ONT's provided software, putting all these conditions on a single flow cell would result in a lot of un-interpretable data. However, doing one condition per flow cell is prohibitively expensive. Therefore, we needed to devise a strategy to

barcode PCR products and to separate these conditions using the raw signal events ourselves. We have successfully designed multiplexing and de-multiplexing procedures for this purpose.

(iii) **Barcoding PCR molecules to multiplex MinION flow cells:** Victor and I have designed a way to add barcodes to the PCR primers that we are able to subsequently identify in the raw event signal for de-multiplexing.

(iv) **De-mutiplexing:** Mark Howison, David Berenbaum, and I have developed a software package, called BarDecoder, which uses a hidden Markov model to estimate the probability that a Fast5 read starts with a given adapter and barcode sequence. Using the means and standard deviations of events in a Fast5 read, the HMM can estimate the probability that those events align with known lead adapter and barcode sequences. The lead adapter sequence is not always aligned with the first event in the read, so the software is also designed to find the most likely starting event for the lead adapter sequence. Given multiple reads which all start with the same lead adapter and one of several known barcode sequences, the software can group the reads by barcode sequence. By being able to interpret the barcodes in signal space, this method lets us know what DNA sequence to expect, what analog was used, and which strand comes first.

(v) **Learning emission parameters of sequences with nucleotide analogs:** In collaboration with Charles Lawrence and Taehee Lee, we designed and implemented an EM algorithm for learning the emission parameters given the known sequences and corresponding observed ONT ionic current signals (events). After considering the problem as a probabilistic graphical model, we ran an algorithm for aligning the sequence with the corresponding signal given the emission parameters. Given those alignments we are able to update the emission parameters for re-alignment and iterate until the emission parameters converge. The aligning algorithm takes into consideration skip and stay probabilities necessary for modeling this data. In general, EM algorithms return several different results based on different local optima. In this problem, we need to learn the emission parameters for the 6-mers containing analogs. The parameters for 6mers containing only A, C, G, and T are already known from ONT. Therefore, we can fix the known parameters for the natural 6-mers during this process. We are able to get unique and reliable results with this approach in simulations using generative models and have begun applying this approach to real data.

(vi) **Simulations on identifying tracts of nucleotide analogs in otherwise normal ACGT data:** Learning new emission parameters will help us infer tracts of two types of label (X and Y) and their orders and orientations along a DNA molecule given the corresponding ionic current signals. To this end, in collaboration with Charles Lawrence and Taehee Lee, we have been developing a variant of 2-layer HMM. We have run several simulations under the assumption that each labeled region contains some shorter sub-regions consisting of A, C, G, and either T, X, or Y in each sequence. We are able to show that this model works well in inferring the existences and orders of labels. We are

also considering an alternative approach of using a neural network to catch the labeled regions.

7.5 Epilogue

All in all, we plan to pilot our methods in yeast where most origins have been identified and confirmed by multiple techniques. Moreover, a lot of other information about the yeast replication program is known. For example, there are data on origin efficiency, fork speeds, and termination regions that we can compare to our MinION or Morse code results. Since yeast chromosomes range from approximately 230 kb to 1.5 Mb, my long read protocols for the MinION could potentially capture the entirety of the smaller yeast chromosomes and large sections of the longer ones. It would be interesting to detect all of the replication activity of a chromosome on single MinION reads. Our longer-term goal is to use these techniques in metazoan systems where origin mapping has been less successful and where there is even less information on other aspects, such as origin efficiency.

Part IV:

Epilogue

This thesis sets the stage for future studies in Sciara, ushering this model system into the genomics era. Prior to this work, the sequence for only one re-replication locus in *Sciara coprophila* was known and little other DNA sequence information for *Sciara* was available. Now the genome sequence is available and at least 14 re-replication loci within it have been identified. I used an array of genomics technologies to put together the genome sequence (Chapter 3), and the strategy I used will certainly be successful when applied to other genomes. In particular, I developed protocols to obtain extremely long nanopore reads (Chapters 2–3). For example, a 131 kb read aligned to a PacBio-only genome assembly with over 91% identity across its full length. From the perspective of a nanopore, this is like sequencing 260,000 bases. That is like threading 12 stories of thread through the eye of a needle. Reads of this length will be extremely valuable for putting genomes together. In fact, 260,000 bases exceeds the length of some yeast chromosomes. It is inevitable that in the near future, megabase read lengths will be reported. At that point assembly becomes a simplified problem. Ultra long reads will also be invaluable to structural variation studies.

I used the signal level data from both PacBio and the MinION to weigh in on the presence of base-modifications in the embryonic genome (Chapter 2). Base modifications were aplenty as detected by algorithms that exploit kinetic variations in the PacBio data. The motifs that arose were the same that arose when analyzing 6mers in the nanopore data that caused ionic current distributions that were different than expected by the model. This work sets the stage for studies into imprinting in Sciara. In addition to mapping base modifications, I mapped many sites of DNA amplification in the salivary gland genome using high throughput Illumina sequencing (Chapter 4).

This was the central reason for assembling the genome. With the genome sequence, future studies in Sciara will advance to system-wide analyses, including ChIP-seq and RNA-seq, to obtain a global view of the DNA amplification program. Moreover, we can use nascent strand sequencing (NS-seq) now to identify initiation sites and zones in the Sciara genome during amplification and in other stages. In particular, it will be good to use the NS-seq conditions that we found to minimize the obstruction to λ-exo digestion that G4 structures cause when formed in vitro and to use some type of non-replicating or otherwise nascent strand negative control (Chapter 6). However, salivary glands can take up the nucleotide analogs necessary to use the MinION to detect replication tracts (Chapter 7) during amplification. The amplicons in Sciara would be a great system for using this technique, particularly for studying elongation. All in all, the tools are now in place to deeply explore the unique biology of Sciara coprophila.

Appendices

Supplementary Information: Sequencing Ultra Long DNA Molecules with the Oxford Nanopore MinION

A.1 Supplementary Methods

A.1.1 DNA Extraction

Genomic DNA (gDNA) was extracted from 51 mg (sample 1, Run A) and 53 mg (sample 2, Runs B and C) of synchronized 2 day old male fungus fly (Sciara) embryos from 30 mass matings each that were extensively washed with sterile TE-wash-buffer (10 mM Tris-Cl, 50 mM EDTA), pH 8.5. There were 5–8 adult males and 10 male-producing adult females per mass mating (note that Sciara females have either only male or only female offspring, which we can control). The washed male embryos were homogenized with a blue pestle inside a 1.5 ml microfuge tube in 200 μl DNAzol (Life Technologies): 5-10 gentle strokes for sample 1; 20 strokes for sample 2. Then 800 μl more DNAzol was added to the tube for 1 ml DNAzol total. The homogenate from sample 2 (not sample 1) was vortexed (full speed, 30 seconds) to further facilitate homogenization and lysis and to help ensure that most DNA breaks occurred prior to end repair (after which DNA breaks affect sequencing output and proportion of 2D reads). The DNAzol homogenate was (for both samples 1 and 2) incubated with 5 μl RNase A Solution (Qiagen) for 10 minutes at 37°C, followed by 5 μl Puregene Proteinase K (Qiagen) for 10 minutes at 37°C. The DNAzol homogenate was then centrifuged at 10,000xg for 10 minutes to pellet debris. The supernatant was transferred to a new 1.5 ml tube using a wide-bore tip (Axygen: Max Recovery, DNase/RNase-free, Sterile, Wide-bore tips) for sample 1 or a regular tip (Olympus: Low-binding, RNase/DNase-free, Sterile, Barrier tips) for sample 2. Then 500 μl 100% ethanol was added. The tube was capped and slowly inverted 50 times, incubated at room temperature for 2 minutes, then on ice for 2 minutes. Precipitated DNA was pelleted at 18,000xg for 10 minutes. The DNA pellet was washed twice with 80% ethanol, air-dried for 30 seconds, and re-suspended in 1xTE, pH 8.0. Sample 2 (not sample 1) was vortexed (full speed, 30 seconds) to help facilitate re-suspension and to help ensure that most DNA breaks occurred prior to end repair (after which DNA breaks affect sequencing output and proportion of 2D reads). Furthermore, to

help re-suspension of high molecular weight DNA, both samples were incubated at 37°C for 1 hour before an overnight incubation at room temperature and storage at 4°C in 1X TE (10 mM Tris-Cl, 1 mM EDTA), pH 8.0, for up to 2 weeks before use. Importantly, for all runs, the Covaris shearing step specified in the ONT protocol was skipped in order to maintain high molecular weight DNA.

A.1.2 AMPure XP beads clean-up #1

After DNA extraction, but before beginning the MinION library preparation, an appropriate volume of the DNAzol-extracted DNA was used to obtain 3, 3.6, and 4 μg of gDNA for Runs A, B, and C respectively. For all runs (A, B, and C), the DNA was cleaned with 1.0x AMPure XP beads (Beckman Coulter Agencourt) with the following specifications. An equal volume (1.0x) of AMPure beads was added to DNA in 1X TE (pH 8.0), incubated for 15 minutes at room temperature (RT), then pelleted on a magnetic rack for 5 minutes (RT) before removing the supernatant. The beads were then washed with 500 μl buffered 80% ethanol (always buffered with 10 mM Tris-Cl, pH 8.0) followed by a second wash with 200 μl buffered 80% ethanol, a brief gentle spin in a microfuge to collect any remaining 80% ethanol at the bottom of the tube, 1 minute re-pelleting on a magnet, removal of the remaining buffered 80% ethanol collected at the bottom of the tube without disturbing beads, air drying for 7 minutes, re-suspending in 175 μl Ultra Pure Water (UPW, Life Technologies UltraPure DNase/RNase-Free Distilled Water) at 37°C for 20 minutes to help ensure long DNA elutes off the beads, pelleting the beads on a magnet for 5 minutes, and then transferring 174 μl supernatant with DNA to a new tube (for PreCR). For Run A, wide-bore tips were used with gentle pipetting. For Runs B and C, normal tips and normal pipetting were used. For Run C, directly before eluting the DNA off the beads, a "rinse" of 200 μl 10 mM Tris (pH 8.0) was gently added to the tube wall opposite the magnetically pelleted beads (in order to avoid directly disturbing the beads), incubated with the beads at room temperature for 60 seconds, and gently removed. This is an additional rinse step used to help deplete DNA <10 kb (see Supplementary Fig. A.3B, Fig. 2.3D, and Supplementary Fig. A.4).

A.1.3 PreCR DNA Repair

Since 3-4x the ONT-recommended amount of starting material was used, PreCR (New England BioLabs, NEB) was performed in double the volume, with double the reagents, for double the time (all relative to the ONT protocol): 200 μl total volume with 174 μl AMPure cleaned DNA, 20 μl 10x ThermoPol buffer, 2 μl 100x NAD+, 2 μl 10 mM dNTPs, and 2 μl PreCR Repair Mix. The reaction was incubated at 37°C for $\geq$ 60 minutes.

A.1.4 AMPure XP beads clean-up #2

For Runs A and B, we proceeded as in clean-up #1, only with a 0.4x AMPure beads ratio (i.e. 80 μl AMPure beads were added to the 200 μl PreCR reaction) and eluting/transferring 85 μl to a new tube at the end (for End Repair). For Run C, sequential 0.4x clean-ups were performed in

the following way. In the first 0.4x AMPure clean-up, the beads were air-dried for only 2 minutes (after the buffered 80% ethanol washes) before adding 140 μl 0.4x AMPure solution (100 μl UPW, 40 μl AMPure beads), gently re-suspending the beads in the second 0.4x solution, and proceeding as in clean-up #1 for Run C. Importantly, a "rinse" was again performed before eluting DNA off the beads. Specifically, 100 μl 10 mM Tris-Cl (pH 8.0) was added to the tube wall opposite the pelleted beads, and incubated with the beads for 30 seconds at room temperature while the tube remained on the magnet before gently removing the rinse. Again (for all runs), elution off the beads (into 85 μl UPW) was performed for a longer time and a higher temperature than manufacturer recommendations to facilitate the elution of long DNA (>20 minutes, 37°C).

A.1.5 End Repair

The NEBNext End Repair Module (NEB) was used: 85 μl of DNA from the previous AMPure step, 10 μl NEBNext End Repair Reaction Buffer (10X), and 5 μl NEBNext End Repair Enzyme Mix. The reaction was incubated for 30 minutes at 22°C.

A.1.6 AMPure XP beads clean-up #3

For Run A, we proceeded as in #2, except that the elution was in 30 μl. For Run B, two sequential 0.4x AMPure clean-ups were performed as done for Run C in #2 (but with no rinse). For Run C, we proceeded as Run C in #2 with 2 sequential 0.4x washes and the rinse step. Specifically, 100 μl 10 mM Tris-Cl (pH 8.0) was added to the tube wall opposite the pelleted beads, then incubated with the beads for 30 seconds at room temperature while the tube remained on the magnet before gently removing the rinse and eluting for >20 minutes at 37°C. For all runs, DNA was eluted into 30 μl 10 mM Tris-Cl (pH 8.0).

A.1.7 dA-tailing

The NEBNext dA-Tailing Module (NEB) was used: 25 μl DNA from the previous AMPure step, 3 μl NEBNext dA-Tailing Reaction Buffer (10X), and 2 μl Klenow Fragment (3– 5 exo–). The reaction was incubated for 30 minutes at 37°C.

A.1.8 Adapter Ligation

The following were combined: 30 μl dA-tail reaction, 8 μl UPW, 10 μl ONT SQK-MAP004 Adapter Mix, 2 μl ONT SQK-MAP004 HP adapter, 50 μl 2X Blunt/TA Ligase Master Mix (NEB). The reaction was incubated for 30 minutes at 22°C.

A.1.9 Enrichment of HP-ligated DNA with His-Beads

550 μL UPW was added to 550 μL SQK-MAP004 2X Wash Buffer (this is called "1X Wash Buffer"), then mixed by inverting 10 times, and briefly spun down in a microfuge. "His-beads" (Dynabeads

His-tag Isolation and Pulldown; Life Technologies) were re-suspended by vortexing for 30 seconds. Then 10 μl of re-suspended beads was transferred to a 1.5 ml Protein LoBind tube (Eppendorf), and combined with 250 μl 1X Wash Buffer. The tube was placed on a magnet for 2 minutes before aspirating off the supernatant. The his-beads were re-suspended in another 250 μl 1X Wash Buffer and placed on the magnet for another 2 minutes before removing the supernatant. The twice-washed and pelleted his-beads were re-suspended in 100 μl undiluted (2X) Wash Buffer and are referred to as the "washed his-beads". 100 μl "washed his-beads" was added to the 100 μl adapter ligation reaction, mixed by gentle pipetting (wide-bore tips), and incubated at 22°C for 5 minutes. The his-beads were then pelleted using a magnetic rack for 2 minutes before removing the supernatant, then washed twice with 250 μl 1X Wash Buffer (with 30 second incubations). The tube was briefly spun in a microfuge to collect excess Wash Buffer at the bottom of the tube for removal after re-pelleting the beads on the magnetic rack for 2 minutes. The pelleted his-beads were re-suspended in 30 μl of Elution Buffer with gentle pipetting using a wide-bore tip, adding the buffer close to the his-beads (to avoid any residual wash buffer on the sides of the tube). The elution was incubated for 10 minutes at 22°C before pelleting with a magnetic rack for 2 minutes. The eluted supernatant (called the "Pre-Sequencing Mix" (PSM)) was transferred to a new Protein LoBind tube.

A.1.10 Filtering base-called fast5 files

Metrichor (ONT base-caller) returns updated fast5 files into two folders: "pass" and "fail". "Pass" contains only fast5 files where 2D base-calling was successful and the mean quality score (Q) of the 2D read is >9. Everything else (including other fast5s containing 2D reads with Q<9, fast5s with only 1D base-calling, and fast5s that failed base-calling) goes into the "fail" folder. To analyze all successfully base-called molecules, we filtered the contents of the fail folder to remove files that completely failed base-calling ("un-basecalled") and further filtered to remove any base-called files that contained the "time error" where a block of events is repeated. Both were accomplished by using our toolset for working with MinION data, called "poreminion" ((`https://github.com/JohnUrban/poreminion`)) which, for some of its functionality, sources the Fast5File classes from poretools [Loman and Quinlan, 2014]:

```
$ poreminion uncalled -m -o fail-filter fail/
$ poreminion timetest -m -o fail-filter fail/
```

The syntax for both subcommands is:
poreminion subcommand options (-m -o outprefix) target-directory (to search).

The "uncalled" subcommand identifies all fast5s that were not base-called due to either: (1) too many events, (2) too few events, or (3) no template found. The flag "-o" gives a prefix to poreminion, which writes text files containing the names of the fast5s in each of the above categories in addition to a summary statistics file describing how many fast5s were searched, how many were assigned to

each category, as well as the minimum, maximum, median, and mean number of events found in files of each category. It also reports the number of events for all files with too many events to base-call. The "-m" flag tells poreminion to not only report the names of the un-basecalled files, but to also move them into their own folders. The "timetest" subcommand searches for files with repeated blocks of events, which are identified by looking at the start times of all events in a fast5 file. If a fast5 file contains an event start time that is earlier than the event that preceded it, then the fast5 file contains this error and is reported ("-o") and moved to a folder for files with this error ("-m").

A.1.11 Obtaining molecule size, mean quality score (Q), other statistics, and plotting

Each base-called fast5 file from a MinION run describes data from a single molecule, yet there can be up to 3 reads per file: template, complement, and 2D. We define the molecule size as the length of the 2D read if present, the length of the template read if there is only a template read present, and the longer length between the template and complement reads when both are present in the absence of a 2D read. Thus, the only time there is a choice is in the latter situation. The majority (68-86%, Supplementary Table A.2, Supplementary Fig. 2.3) of files with a template and complement sequence have a 2D read, all of which have a template:complement sequence length ratio between 0.5 and 2. Moreover, most files that contain both template and complement sequences with a sequence length ratio between 0.5 and 2 have 2D reads (Supplementary Fig. 2.3). This means that when a choice needs to be made between the template and complement, they are vastly different sizes. The complement can be much smaller than the template, for example, when there is a nick in the complement strand. The template can be much shorter than the complement, for example, when the motor protein on the Y-adapter falls off allowing the template to zip through until it is caught by the hairpin motor protein. In these situations, the longer read better represents the size of the molecule that was sequenced. Importantly, molecule size allows for a single/non-redundant length from each base-called fast5 file (i.e. single molecule) for computing statistics, such as the summed length of sequenced molecules and molecule N50, in addition to statistics computed on all read types (such as 2D read length N50). This molecule size estimate as well as many descriptive metrics of the fast5 file were obtained with the poreminion subcommand "fragstats" after combining all base-called files from the pass and fail folders.

```
$ poreminion fragstats all-basecalled-files/ > fragstats.txt
```

The "fragstats" subcommand gives a tab-delimited output where each line describes an individual fast5 file (or single molecule) with the following columns (for version 0.4.3).
1 = read name

2 = estimated molecule/fragment size
3 = number input events
4 = if complement detected
5 = if 2D detected
6 = number of template events
7 = number of complement events
8 = length of 2D sequence
9 = length of template sequence
10 = length of complement sequence
11 = mean quality score of 2D sequence
12 = mean quality score of template sequence
13 = mean quality score of complement
14 = ratio of number template events to number complement events
15 = channel number molecule traversed
16 = heat sink temperature while molecule traversed
17 = number of called template events (after events pruned during base-calling)
18 = number of called complement events (after events pruned during base-calling)
19 = number of skips in template (number 0)
20 = number of skips in complement (number 0 moves)
21 = number of stays in template
22 = number of stays in complement
23 = strand score template
24 = strand score complement
25 = number of stutters in template
26 = number of stutters in complement

The tab-delimited fragstats.txt file was then brought into R to make most plots (Figures 2.3, 2.4, 2.5, and Supplementary Figures A.1, A.2, A.4, A.5, A.6). The fragstats file was also summarized (generating many of the statistics reported such as in Table 2.1 and Supplementary Tables A.1-A.6) using:

```
$ poreminion fragsummary -f fragstats.txt
```

A.1.12 Analyzing files with too many events (>1 million) to base-call

Time Error:

Poreminion was used to extract the event start times from the multi-million event fast5 files into

text files (columns of poreminion ouput = event mean, event standard deviation, event start time, event duration), which were brought into R for visualization.

$ poreminion events -f5 target.fast5 | cut -f 3 > start.times.txt

In the directory with the fast5 files that had too many events to base-call (this directory was made by "uncalled" filtering above), the poreminion "timetest" subcommand was used to further filter these multi-million event fast5s to keep only those without the 'time error'. Time errors (and lack thereof) were visualized in R from events text files extracted with the above poreminion command.

Lead adapter:
The first 50 events of the remaining multi-million event fast5 files were obtained with poreminion (command: poreminion events -f5 target.f5) and searched for evidence of the lead adapter event mean profile by comparing to 150 randomly sampled (50 from each run A,B,C) pass fast5s (containing high quality 2D reads). Since there were so few multi-million event files, it was sufficient to manually separate ones with the lead adapter profile from ones that did not. However, we also found that the simple rule of requiring that there be 2 or more events within the first 15 events that have means >80 was sufficient to automatically separate the files this way for visualization.

A.1.13 Looking at the number of "0 moves" (stays) vs. length and/or quality

The fragstats.txt file produced above was brought into R for the various plots comparing the number of stays ("0 moves") with other features such as read length, number of events, number of called events, and mean quality scores (Q).

A.1.14 Identifying G4 motif positions in template and complement reads

The poreminion subcommand "g4" was used in the following way:

$ poreminion g4 --minG 3 --maxN 7 --numtracts --noreverse -f5 all/ > g4s.bed

This subcommand uses the quadparser [Huppert and Balasubramanian, 2005,Huppert, 2010] regular expression, $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ (regular expression in Python: '([gG]{3,}\w{1,7}){3,}[gG]{3,}') to search the sequences inside fast5 files for G4 motifs (G_{3+} is specified by "--minG 3" and N_{1-7} is specified by "--maxN 7"). The "--noreverse" option specifies to only search the sequence given (not its complement), or in other words it specifies to NOT also search for the "C4" motif: $C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}$. The "--numtracts" option reports the number of poly-G tracts inside

a given G4 motif as the Python regular expression (above) searches for 4 or more adjacent poly-G tracts separated by 1-7 nucleotides. The more poly-G tracts, the more possible ways a G4 structure could form. For example, observing five poly-G tracts does not simply indicate two overlapping G4 motifs that can form only two G4 structures, but rather "5 choose 4" (i.e. 5) ways to choose four of those five poly-G tracts multiplied by the number of ways 4 poly-G tracts can fold together into a G4 structure (e.g. there are parallel and anti-parallel arrangements) as well as the multiplicative possibilities when varying the number and position of consecutive Gs (>3) used in the chosen poly-G tracts. Keeping track of the number of poly-G tracts inside each G4 motif allowed us to separate the G4 motifs into two groups: those with 4 poly-G tracts and those with >4 for the subsequent statistical analysis testing which group has more "0 moves" associated with it on average.

A.1.15 Identifying positions of stays ("0 moves") in template and complement reads

The poreminion subcommand "staypos" was used as follows:

$ poreminion staypos all/ > stays.bed

This subcommand goes through the base-called events of template and complement strands while keeping track of the index of each event relative to the output sequence (by accounting for base-caller "moves" of 0-5) and reporting the positions in the sequence that correspond to "0 moves" in the base-called events. Specifically, it reports the coordinates of the 5mer corresponding to the "0 move" in BED format (for example, if a 5mer starts at position 0, its end position is 5).

A.1.16 Comparing G4 and Stay positions in template and complement reads

For plotting, windowBed from BEDtools(5) was used to obtain all stay positions within 500 nucleotides of a G4 motif (done independently for each of the three runs):

$ windowBed -a g4s.bed -b stays.bed -w 500 > g4s.stays.500.windowbed

The resulting file contains lines with pairs of entries for the G4 motif position and stay ("0 move") position for each pair that is within 500 nucleotides of each other. This file was brought into R where distances between G4 centers and 'stay' centers (i.e. the middle nucleotide of a 5mer) from

G4-stay pairs were calculated as the distance between their centers. Centers were found by subtracting 1 from the end position of each BED entry (to account for BED format), then taking the mean of the start and resulting end positions. Histogram information was obtained by, for example, hist(distances, breaks=seq(from=-650.5, to=650.5, by=1), which results in the histogram midpoints being integers from -650 to 650. Histogram counts were lightly loess smoothed: loess(hist.counts ~ hist.mids, span=0.05)

Four null distributions were considered. For each G4 motif, a site of the same length was selected uniformly at random from: (null 1) any template or complement read with no G4 motifs, (null 2) any template or complement read, (null 3) anywhere within the same read the G4 motif was on, (null 4) anywhere within the same read the G4 motif was that did not overlap the G4 motif coordinates. These coordinates were selected with BEDtools(5):

$ shuffleBed -noOverlapping -excl readswithg4.bed -i g4s.bed -g template-and-complement-reads > g4s.shuffled.nonG4reads.bed

$ shuffleBed -noOverlapping -i g4s.bed -g template-and-complement-reads > g4s.shuffled.allreads.bed

$ shuffleBed -noOverlapping -chrom -i g4s.bed -g template-and-complement-reads > g4s.shuffled.sameread.bed

$ shuffleBed -allowBeyondChromEnd -noOverlapping -chrom -excl g4s.bed -i g4s.bed -g template-and-complement-reads > g4s.shuffled.sameread.notoverG4.bed

windowBed from BEDtools was used as above to collect pairs of randomly selected locations and "0 moves" that were within 500 nucleotides of each other. Histogram counts and smoothing was same as above. These null distributions were plotted on same plots as above.

For statistical analyses, we compared the number of "0 moves" within 50 nucleotides of the G4 motifs with their matched null positions from the four different null distributions. These counts were obtained using windowBed from BEDtools(5). For example:
windowBed -c -a g4s.bed -b stays.bed -l 50 -r 50 > g4s.stays.50.counts.txt

$ windowBed -c -a nulls.bed -b stays.bed -l 50 -r 50 > nulls.stays.50.counts.txt

For each of the four nulls described above, the pairs of counts from G4 motifs and the null were used as input to the "sign test" as well as the Wilcoxon signed rank test in R. Note that there seems to be more "0 moves" on reads with G4 motifs in general. The fourth null (null 4: selecting a random position on the same read as the G4 motif that does not overlap the G4 motif) serves as the best

matched pairs, controlling for any read-specific effects and still all p-values were significant (Supplementary Table A.9). To test the hypothesis that G4 motifs on the complement strand associated with more "0 moves" than G4 motifs on the template strand, the counts for each of these groups was used as input to the Wilcoxon rank sum test in R (Supplementary Table A.10). To test the hypothesis that G4 motifs with >4 poly-G tracts were associated with higher "0 move" counts than G4 motifs with only 4 poly-G tracts, the counts for each of these groups was used as input to the Wilcoxon rank sum test in R (Supplementary Table A.11).

A.2 Supplementary Figures

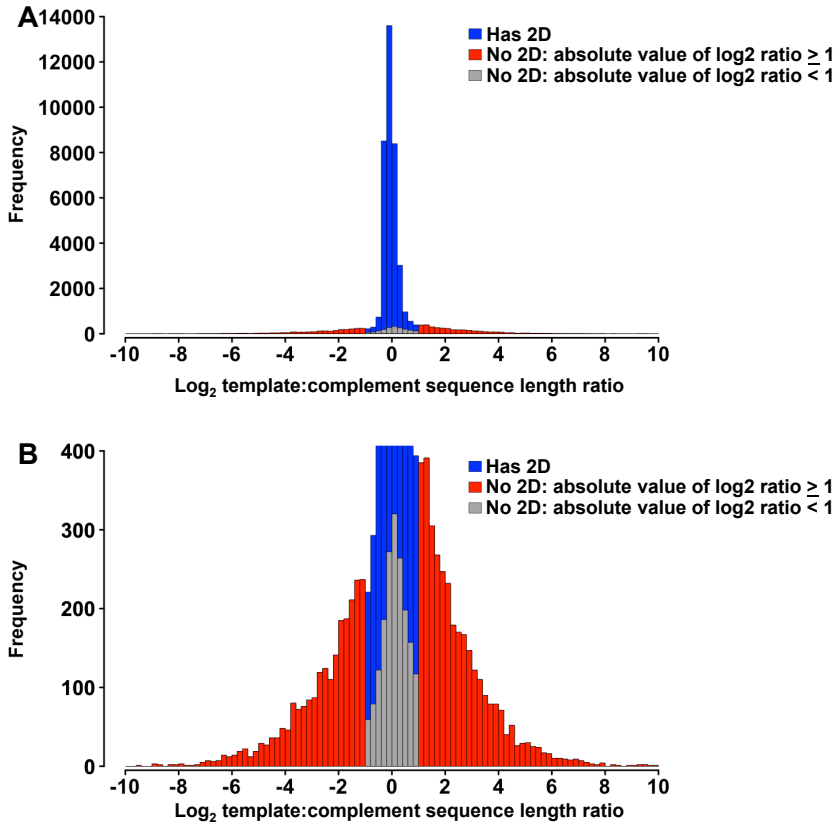


Figure A.1: Distribution of Log2(template:complement) for base-called fast5 files that contain both template and complement reads.
(A) shows a zoomed-out view of a histogram of Log2(ratio of template read length to complement read length) for all base-called fast5 files that have both template and complement reads. The histogram shows the number (frequency) of fast5 files as a function of Log2(ratio). **(B)** shows a zoomed-in view of the bottom of the same histogram as in A. Most fast5 files that have both template and complement reads also have 2D reads (blue). For those that do not, most have template-to-complement read length ratios that are either too big ($>$ 2; i.e. $>$ 1 in log2) or too small ($<$ 0.5; i.e. $<$ –1 in log2) for initiating 2D base-calling (red, absolute value of log2 ratio $\geq$ 1). However, there are also base-called fast5 files with both template and complement reads where the ratio is within range for 2D base-calling and where 2D base-calling fails (grey).

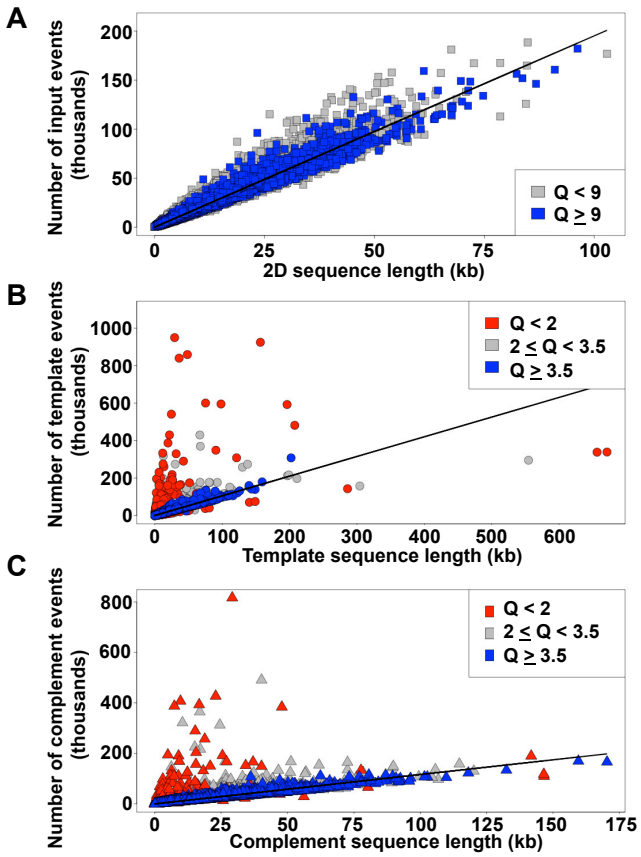


Figure A.2: number of pre-base-calling events vs. post-base-calling sequence length. **(A)** number of total input events vs. 2D read length for reads with 2D base-calls. The slope of the best fit line is approximately 2 (1.95) representing approximately 2 events per base on average for fast5 files that contain a 2D read. Blue represents the high quality ($Q \geq 9$) 2D reads. **(B)** number of template events versus template sequence length. The slope (1.051) of ~1 indicates ~1 event per base on average. Blue represents higher quality 1D reads ($Q \geq 3.5$) while red represents lower quality 1D reads ($Q < 2$). **(C)** number of complement events versus complement sequence lengths shows a slope of ~1 (1.166) representing ~1 event per base on average. Blue and red are as in B. Prior to base-calling, one can estimate molecule size using 1-2 events per base (depending on whether the fast5 file contains only template events, if there are also complement events, and if so, how many). A minority of exceptions have higher event:base ratios meaning that 1-2 events per base is often an upper limit estimate for molecule size. Importantly, template and complement sequences with mean quality scores ≥ 3.5 (blue in B and C) stay tightly packed around the approximately 1 event:base line whereas those with $Q < 3.5$ (grey and red in B and C) often fall off the line (especially when $Q < 2$; red). This suggests that using $Q \geq 3.5$ as a cut-off for 1D reads has a practical interpretation that may be useful, though the direct interpretation means it has a predicted accuracy of only about 45% ($10^{3.5/-10} = 0.447$). The utility of such a read may rely on whether or not the low level of accuracy is uniformly distributed across the read or if there are a high quality stretches broken up by low quality ones, which could still be taken advantage of in some applications [Warren et al., 2015a], and perhaps more so in signal space.

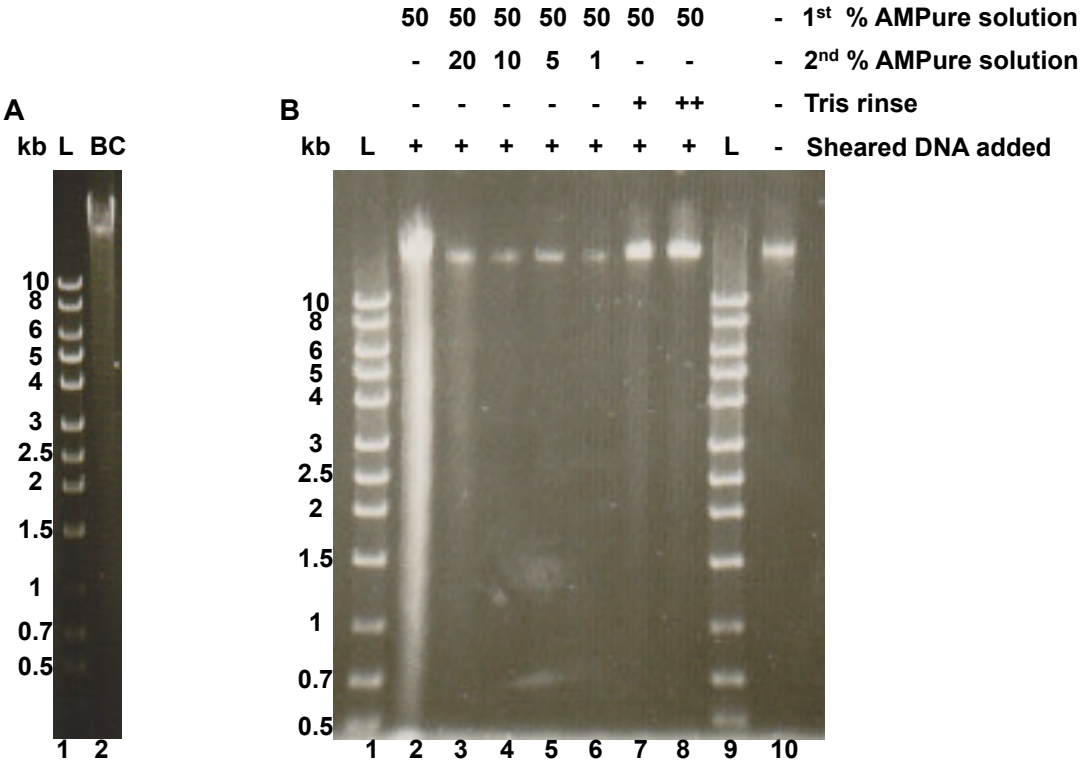


Figure A.3: Most DNA remains >10 kb with vortexing and simple modifications to the AMPure beads procedure that deplete DNA <10 kb
(A) Gel showing DNA source for runs B and C. "L" is the DNA ladder and " kb" shows the sizes of the bands in the ladder. "BC" is the genomic DNA used for both Run B and Run C where vortexing was used during and after DNA extraction. This gel shows that a substantial proportion remains above 10 kb after vortexing. **(B)** Gel demonstrating results of various approaches to deplete DNA shorter than 10 kb. "L" and " kb" same as in A.

Further description of Figure A.3B:

Genomic DNA was gently extracted and re-suspended. Eight 15 μl aliquots were set aside. The rest was lightly sonicated/sheared using the BioRuptor (Diagenode) to a very broad range with a lower limit near 0.5 kb and an upper limit at the size of unsheared DNA (unsheared DNA shown in lane 10). Equal volumes of sheared DNA were added to 7 of the 8 aliquots (lanes 2-8, not lane 10) in order to test several strategies to deplete the shorter, sheared DNA (<10 kb), all using AMPure beads (Lanes 3-8). Lane 2 shows the sheared+unsheared DNA without depletion of smaller DNA. Each of the aliquots with sheared DNA added (Lanes 2-8) was brought to 100 μl volume with Ultra Pure Water (UPW, Life Technologies UltraPure DNase/RNase-free, distilled water). 100 μl AMPure beads were then added to each to create a 50% AMPure mixture (or 1.0x ratio). The DNA was incubated in the 50% AMPure solution for 5 minutes before pelleting the beads on a magnet for 5

minutes and removing the supernatant. The beads were then washed twice with 80% ethanol while on the magnet. After the second 80% ethanol wash, the tubes with the beads were lightly spun and placed on the magnetic rack for 1 minute before removing the remaining 80% ethanol collected at the bottom from the light spin. The previous steps were performed to remove any buffers or salts associated with the DNA to have finer control over the AMPure ratio (for lanes 3-6) in subsequent steps. The beads from lane 2 were air-dried and eluted at this point. For lanes 3-8, instead of eluting, the beads were re-suspended in a second AMPure solution or "rinsed" once (lane 7) or twice (lane 8) before eluting. For lanes 3-6, AMPure solutions were premade by combining 40 μl AMPure with 160 μl UPW (20%, lane 3), 20 μl AMPure with 180 μl UPW (10%, lane 4), 10 μl AMPure with 190 μl UPW (5%, lane 5), and 2 μl AMPure with 198 μl UPW (1%, lane 6). In all cases (lanes 3-6), the second AMPure solution was added to the beads from the first AMPure step (after the 80% ethanol washes described above), and the beads were gently re-suspended, incubated for 10 minutes, and put on a magnetic rack for 5 minutes before the supernatant was removed and two 80% ethanol washes were performed while the beads remained on the rack. After the second 80% ethanol wash, the tubes with the beads were lightly spun and placed on the magnetic rack for 1 minute before removing the remaining 80% ethanol collected at the bottom from the light spin. The tubes were then allowed to air dry for 2-3 minutes. For lanes 7-8, after the first set of 80% ethanol washes, instead of proceeding to a second AMPure step, the tubes with the beads were lightly spun and placed on the magnetic rack for 1 minute before removing the remaining 80% ethanol collected at the bottom from the light spin. The tubes were then allowed to air dry on the magnetic rack for 2 minutes before adding 200 μl UPW very gently to the tube-wall opposite the beads while they remained on the magnetic rack. This is the "rinse". The beads were allowed to incubate in the rinse for 1 minute before very gently removing it. Lane 8 was subject to a second identical rinse, which seems to offer minimal additional benefits beyond the first rinse (compare lanes 7 and 8). For all elutions (lanes 2-8), DNA was eluted off the AMPure beads into 15 μl UPW at 37°C for 20 minutes to facilitate elution of long DNA. Loading buffer was added to each sample (lanes 2-8, 10). The DNA samples were then loaded onto the gel, electrophoresed, and stained with ethidium bromide for visualization. Although all conditions eliminated most of the smaller DNA (e.g. <10 kb), it was possible to tell where the tails of the smears in each lane ended using longer exposure times (data not shown). Lane 3 (20% AMPure) ended around 1 kb, Lane 4 (10% AMPure) ended above 2 kb, Lane 5 (5% AMPure) ended above 3 kb, Lane 6 (1% AMPure) ended above 5 kb, Lane 7 (1 rinse) ended above 1 kb, and Lane 8 (2 rinses) ended above 1.5 kb. While some of the lower percentage AMPure solutions performed better than the "rinse" as judged by where the tail ended, there were also comparable losses of the large DNA (compare lanes 3-6 to lane 2). In contrast, the rinses largely removed the small DNA while retaining the majority of the large DNA (compare lanes 7-8 with lane 2).

These results were used to inform us how to modify the AMPure steps for Run C. To deplete DNA <10 kb in Run C, we chose to (1) keep the strategy of sequential AMPure washes as done here (Lanes 3-6), but with the ONT recommended 0.4x AMPure ratio (~28-29% AMPure solution) in

the first and second sequential steps instead of the 50% (1x) AMPure solution used in the first step and lower percent (1-20%) solutions used in the second step here and (2) add a rinse step after the second sequential AMPure step. This was successful as determined by comparing the read lengths from Runs B and C, which came from the same source of DNA (Fig. 2.3d, Supplementary Fig. A.4).

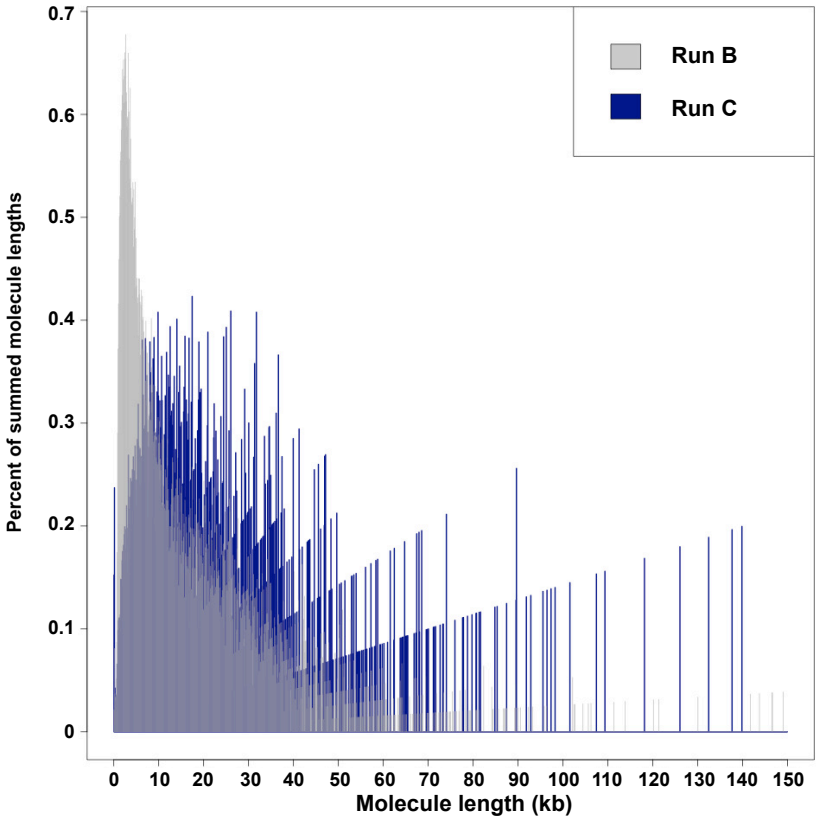


Figure A.4: The proportion of total summed molecule length as a function of molecule length for Run B (transparent grey) and Run C (dark blue).
Note that the different shades of grey is a result of the dark blue behind the transparent grey. Run B and Run C used the same source of DNA, but differed in library preparation (see Fig. 2.2). Run C used a new rinse step during all AMPure clean-ups as well as sequential 0.4x AMPure rounds in each clean-up step after PreCR.

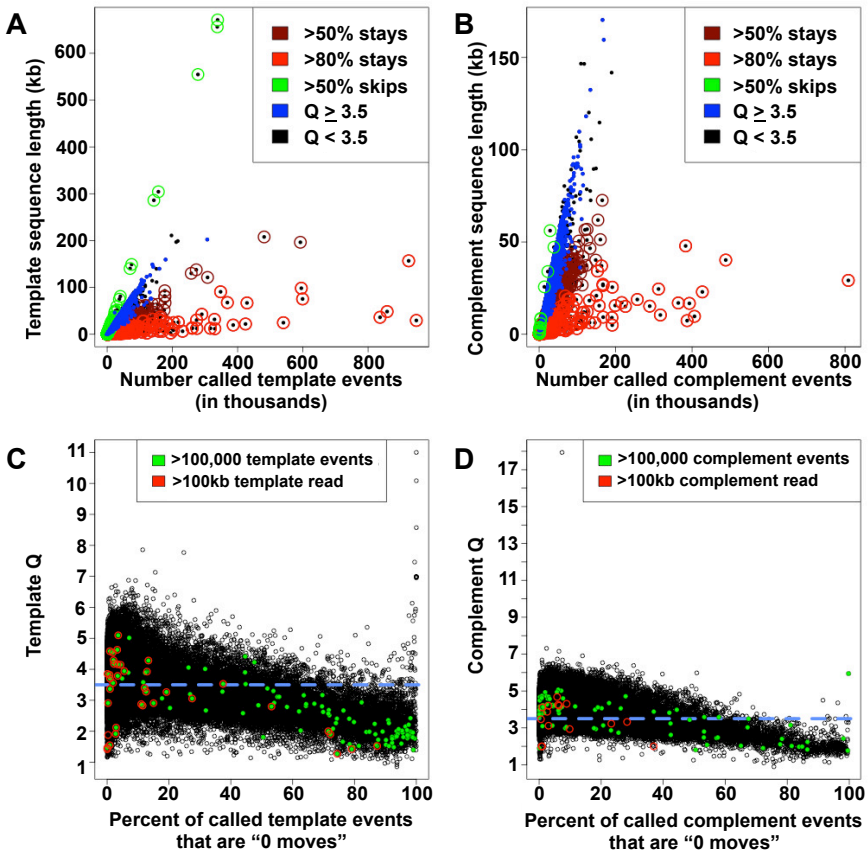


Figure A.5: number of base-called events, sequence lengths, percent of base-called events assigned "0 moves", and mean quality scores.
(A) number of base-called template events vs. template sequence length with information on mean quality scores and percent of base-called template events assigned a "0 move" ("stay") or a "skip" (move=2). **(B)** number of base-called complement events vs. complement sequence length with information on mean quality scores and percent of base-called complement events assigned a "0 move" ("stay") or a skip (move=2). In **(A)** and **(B)**, higher quality 1D reads (Q $\geq$ 3.5) are blue dots and those with Q $<$ 3.5 are black dots. Maroon circles are around dots with $\geq$50% of base-called events assigned a "0 move", red circles are around dots with $\geq$80% of base-called events assigned a "0 move", and green circles are around dots where $\geq$50% of the base-called events were "skips" (move=2). Importantly, template and complement reads with Q $\geq$ 3.5 (blue) rarely are encircled. **(C)** Percent of called template events assigned a "0 move" (stay) vs. mean quality score (Q). Slope of best fit line (lm(y ~ x) in R) is –0.01713 meaning for every 1 percent (or 10 percent) increase in "0 moves" there is an average decrement of 0.01713 (or 0.1713) in Q. **(D)** Percent of called complement events assigned a "0 move" (stay) vs. mean quality score (Q). Slope of best fit line (lm(y ~ x) in R) is -0.02363 meaning for every 1 percent (or 10 percent) increase in "0 moves" there is an average decrement of 0.02363 (or 0.2363) in Q. In **(C)** and **(D)**, the blue dashed line is at Q = 3.5, large green dots represent files with >100,000 called template **(C)** or complement **(D)** events and red circles encompass dots for files that had $\geq$100 kb template **(C)** or complement **(D)** sequences.

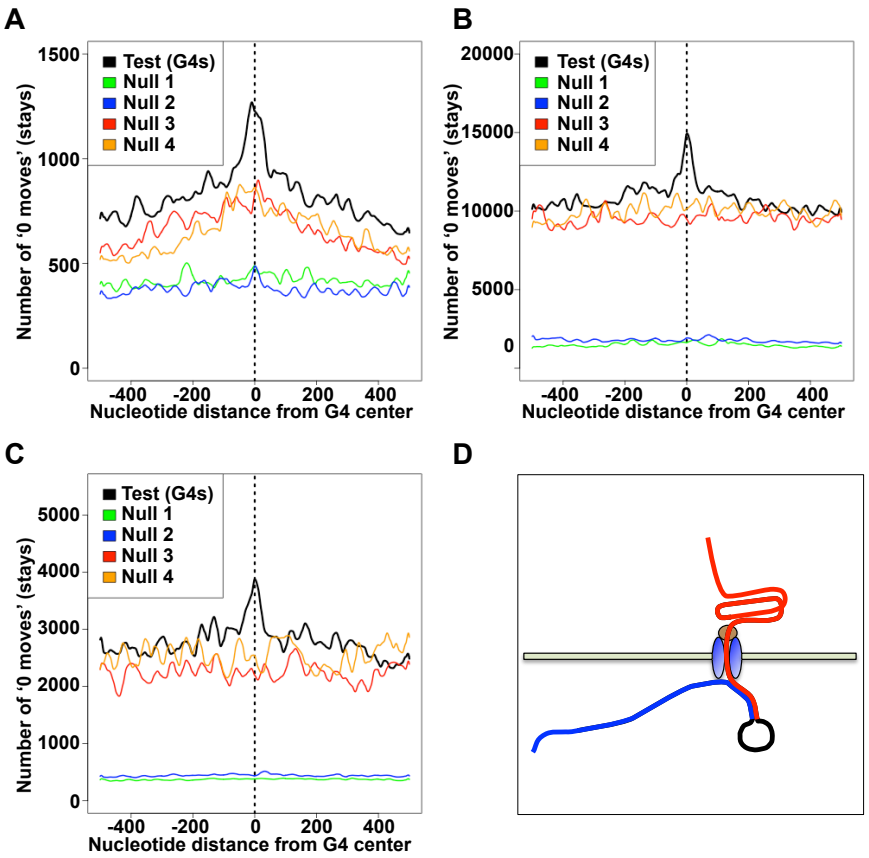


Figure A.6: Aggregate analyses of the distribution of "0 moves" around G4 motif centers.

The number of "0 moves" as a function of proximity to G4 motif centers (test condition) or the centers of randomly selected positions (null 1–4 conditions) for **(A)** Run A, **(B)** Run B, and **(C)** Run C. **(D)** Depicts the orientation (5' – 3') of a DNA molecule going through a pore from the top chamber to the bottom chamber and how a G4 might cause a DNA molecule to stall resulting in an accumulation of measurements of a 5mer or set of 5mers slightly upstream of the G4 (the adjacent upstream sequence is pulled through the pore before the downstream G4 blocks further translocation). The template strand is blue, the hairpin adapter is black, and the complement strand is red.

A.3 Supplementary Tables

Table A.1: Statistics on Mean Quality Scores (Q) for 2D and 1D reads from Runs A, B, and C.

Run A	2D	1D	Template	Complement
Median Q	9.23666606763	3.54781658091	3.41144557734	4.0100659241
Mean Q	9.13260594983	3.553805971	3.43220450158	3.93551548089
Standard Deviation of Q	1.08997018784	0.916461481548	0.871848558243	0.947880924972
Minimum Q	5.79424190524	0.965373861239	0.965373861239	1.12938538308
Maximum Q	12.2263387187	10.0869987652	10.0869987652	6.43558322934

Run B	2D	1D	Template	Complement
Median Q	8.59116706044	3.77054238697	3.58182241161	4.21229580534
Mean Q	8.61912781671	3.80409144806	3.59890226341	4.1571515184
Standard Deviation of Q	0.988906885917	0.782651340952	0.691010548999	0.804879951209
Minimum Q	4.51060190561	0.855017497526	0.855017497526	0.892928322743
Maximum Q	12.9319853199	11.0	11.0	7.31138333836

Run C	2D	1D	Template	Complement
Median Q	9.07294962204	3.79287980744	3.62309975937	4.13185844371
Mean Q	8.96810671735	3.73547257167	3.60595627604	4.05290820742
Standard Deviation of Q	0.976642865308	0.899676295249	0.872084862604	0.887443696552
Minimum Q	4.95929992076	1.11272829123	1.11272829123	1.16860148363
Maximum Q	12.9534046975	17.9360409474	7.46183460838	17.9360409474

The groups of 1D reads contains all template and complement reads for given run.
Top: Run A
Middle: Run B
Bottom: Run C

Table A.2: Read types in base-called fast5 files.

	Run A	Run B	Run C
Number of molecules (Number of successfully base-called files)	9,935	64,282	8,941
Number template only	6,770	26,923	5,293
Percent template only	68.14	41.88	59.20
Number that have a complement read	3,165	37,359	3,648
Percent that have a complement read	31.86	58.12	40.80
Number that have a 2D read	2,172	31,999	2,523
Percent that have a 2D read	21.86	49.78	28.22
Number that do NOT have a 2D read	7,763	32,283	6,418
Percent that does NOT have a 2D read	78.14	50.22	71.78
Percent of molecules that do NOT have a 2D read because they only contain a template read	87.21	83.40	82.47
Percent of molecules that do NOT have a 2D read that DO have a complement read	12.79	16.60	17.53
Number with a complement read that DO have a 2D read	2,172	31,999	2,523
Percent of molecules with a complement read that DO have a 2D read	68.63	85.65	69.16
Number with a complement read that do NOT have a 2D read	993	5,360	1,125
Percent of molecules with a complement read that do NOT have a 2D read	31.37	14.35	30.84

Table A.3: Molecule Length Statistics.

	Run A	Run B	Run C
Sum of all molecule lengths	49,778,211	386,880,692	70,086,870
Molecule N25	46,973	28,006	34,939
Molecule N50	25,238	13,553	20,824
Molecule N75	11,969	5,279	10,749
Mean molecule size	5,010.4	6,018.5	7,838.8
Median molecule size	152	2,817	3,116
Max molecule size	304,309	671,219	139,864
Min molecule size	5	5	5
Number of molecules >10 kb	1,433	10,145	2,318
Percent of molecules >10 kb	14.42	15.78	25.93
Sum of molecule lengths >10 kb	39,415,998	226,507,295	54,080,719
Percent of all summed molecule lengths from molecules >10 kb	79.18	58.55	77.16
Number of molecules >50 kb	139	396	119
Percent of molecules >50 kb	1.40	0.62	1.33
Sum of molecule lengths >50 kb	10,999,924	27,504,214	8,168,033
Percent of all summed molecule lengths from molecules >50 kb	22.10	7.11	11.65
Number of molecules >100 kb	21	24	8
Percent of molecules >100 kb	0.21	0.04	0.09
Sum of molecule lengths >100 kb	3,140,720	4,779,702	972,486
Percent of all summed molecule lengths from molecules >100 kb	6.31	1.24	1.39

Molecule size is determined as (1) the length of the 2D read if one is available, (2) the length of the template if only the template is available, or (3) the longer of the template and complement reads when both are available in the absence of a 2D read (See Supplementary Methods). Thus, the statistics are performed on a set of unique/non-redundant molecule length estimates (a single read for each molecule as opposed to 1–3 reads per molecule).

Table A.4: 2D Read Length Statistics.

	Run A	Run B	Run C
Sum of HQ (Q≥9) 2D lengths	13,220,898	74,311,821	15,790,614
HQ 2D N25	34,893	23,192	30,278
HQ 2D N50	18,145	11,620	18,008
HQ 2D N75	9,002	5,194	10,013
Mean HQ 2D length	10,722.5	6,519.72	11,593.7
Median HQ 2D length	6,393	3,580	8,061.5
Sum ALL 2D lengths	24,017,633	188,595,516	27,734,513
ALL 2D N25	37,314	23,647	31,130
ALL 2D N50	19,460	11,200	18,426
ALL 2D N75	9,614	4,681	9,801
Mean 2D length	11,057.8	5,893.8	10,992.7
Median 2D length	6,356	3,051	7,210
Min 2D length (Q)	94 (6.33)	81 (6.51)	82 (5.97)
Max 2D length (Q)	102,935 (8.74)	86,797 (9.01)	84,898 (8.87)
Max 2D length Q ≥ 9	96,237 (9.61)	86,797 (9.01)	71,380 (9.04)
Max 2D length Q ≥ 8.5	102,935 (8.74)	86,797 (9.01)	84,898 (8.87)
Max 2D length Q ≥ 8	102,935 (8.74)	86,797 (9.01)	84,898 (8.87)
Max 2D length Q ≥ 7.5	102,935 (8.74)	86,797 (9.01)	84,898 (8.87)

HQ = High Quality. 2D reads with mean quality scores (Q) > 9 are considered high quality reads by Oxford Nanopore. Note: The numbers of 2D reads >50 kb in Runs A, B, and C were 44, 79, and 22, respectively.

Table A.5: 1D Read Length Statistics.

	Run A	Run B	Run C
Sum 1D lengths	75,870,350	565,275,221	102,225,612
1D N25	41,287	25,353	33,308
1D N50	21,927	12,037	19,812
1D N75	10,432	4,776	10,349
Mean 1D length	5,791.6	5,561.5	8,120.2
Median 1D length	231	2,661	3,783
Sum template lengths	43,984,348	34,9320,608	66,858,648
Template N25	43,176	25,699	33,509
Template N50	22,324	12,198	20,001
Template N75	10,344	4,752	10,421
Mean template length	4,427.2	5,434.2	7,477.8
Median template length	151	2,565.5	2,892
Min template length (Q)	5 (10.09)	5 (6.95)	5 (6.97)
Max template length	304,309 (2.12)	671,219 (1.55)	139,864 (4.28)
Max template length Q > 4	122,689 (4.63)	143,763 (4.17)	139,864 (4.28)
Max template length Q > 3.5	202,293 (3.53)	143,763 (4.17)	139,864 (4.28)
Max template length Q > 3	210,931 (3.36)	143,763 (4.17)	139,864 (4.28)
Max template length Q > 2.5	210,931 (3.36)	554,720 (2.91)	139,864 (4.28)
Sum complement lengths	31,886,002	215,954,613	35,366,964
Complement N25	39,391	24,784	33,054
Complement N50	21,394	11,821	19,287
Complement N75	10,506	4,825	10,163
Mean complement length	10,074.6	5,780.5	9,694.9
Median complement length	5,049	2,836	5,700
Min complement length	17 (1.13)	14 (3.63)	10 (4.36)
Max complement length	170,334 (3.87)	146,631 (2.00)	132,439 (4.20)
Max complement length Q > 4	159,563 (4.30)	102,671 (4.20)	132,439 (4.20)
Max complement length Q > 3.5	170,334 (3.87)	102,671 (4.20)	132,439 (4.20)
Max complement length Q > 3	170,334 (3.87)	105,683 (3.23)	132,439 (4.20)
Max complement length Q > 2.5	170,334 (3.87)	120,149 (2.94)	132,439 (4.20)

Note: Statistics on "1D reads" are computed on all template and complement reads together (rather than treating the two read types separately).

Table A.6: Top 10 Read Lengths for given categories.

Top ten 2D lengths across all three runs

Rank	Length	Q	Run
1	102,935	8.74	A
2	96,237	9.61	A
3	91,062	9.93	B
4	86,797	9.01	C
5	84,898	8.87	A
6	84,716	8.28	B
7	84,477	7.75	A
8	83,697	9.34	B
9	82,373	9.73	A
10	78,658	8.24	A

Top ten 2D lengths with mean quality score (Q) > 9 across all three runs

Rank	Length	Q	Run
1	96,237	9.61	A
2	91,062	9.93	A
3	86,797	9.01	B
4	83,697	9.33	A
5	82,373	9.73	B
6	74,769	9.68	A
7	71,849	10.06	A
8	71,380	9.04	C
9	71,330	10.49	A
10	69,971	9.24	B

Top ten 1D lengths across all three runs

Rank	Length	Q	Run	Read Type
1	671,219	1.55	B	Template
2	656,409	1.92	B	Template
3	554,720	2.91	B	Template
4	304,309	2.11	A	Template
5	286,113	1.47	B	Template
6	210,931	3.37	A	Template
7	207,759	1.27	B	Template
8	202,293	3.53	A	Template
9	199,014	3.39	A	Template
10	196,898	2.88	A	Template

Top ten 1D lengths with mean quality Q > 3 across all three runs

Rank	Length	Q	Run	Read Type
1	210,931	3.37	A	Template
2	202,293	3.53	A	Template
3	199,014	3.40	A	Template
4	170,334	3.87	A	Complement
5	159,563	4.30	A	Complement
6	159,397	3.92	A	Template
7	148,912	3.70	A	Template
8	143,763	4.17	B	Template
9	139,864	4.28	C	Template
10	132,439	4.20	C	Complement

Table A.7: number of channels (out of 512) available for sequencing.

	Run A	Run B	Run C
Estimated number of channels available at QC	191	491	448
Number of channels that had at least 1 read throughout run	214	466	349

Table A.8: Filtering post-base-calling fast5 files.

	Run A	Run B	Run C
Number event files (submitted to Metrichor base-caller)	27,667	82,040	12,536
Number files filtered: no template found (NTF)	10,306 (37.3%)	6,820 (8.31%)	1,997 (15.9%)
Number files filtered: too few events (TFE)	7,414 (26.8%)	10,878 (13.26%)	1,583 (12.6%)
Number of files filtered: too many events (TME)	9 (0.03%)	46 (0.056%)	14 (0.11%)
Number of files filtered: time error (TE)	2 (0.007%)	14 (0.017%)	0 (0 %)
Number of files with <100,000 events with time error (percent of all files with <100,000 events)	0 (0%)	0 (0%)	0 (0%)
Number of files with 100,000 to 1,000,000 events with time error (percent of all files with 100,000 to 1,000,000 events)	2 (1.57%)	14 (5.51%)	0 (0%)
Number of TME files (>1,0000,000 events) that also had time errors	5 of 9 (55.6%)	30 of 46 (65.2%)	4 of 14 (28.6%)
Number of TME files (>1,0000,000 events) that did NOT have time errors	4 of 9 (44.4%)	16 of 46 (34.8%)	10 of 14 (71.4%)
Number of TME files (>1,0000,000 events) without time errors that did NOT have lead adapter profile	0 of 4 (0%)	6 of 16 (37.5%)	5 of 10 (50%)
Number of TME files (>1,0000,000 events) without time errors that had lead adapter profile	4 of 4 (100%)	10 of 16 (62.5)	5 of 10 (50%)
Number/percent of TME files (>1,0000,000 events) that passed time-error and lead adapter filtering	4 of 9 (44.4%)	10 of 46 (21.7%)	5 of 14 (35.7%)

Not all files submitted to the Metrichor base-caller are returned successfully base-called. These files come back mixed in with other fast5 files that were successfully base-called. Therefore, we filter them out and note how many while doing so. Moreover, a tiny percent of fast5 files at the time of these experiments contained an error where blocks of events were artificially repeated. We also filter and note how many of these exist as well. See methods and supplementary methods for more details on filtering.

Table A.9: Are there more "0 moves" near G4 motifs than sites selected at random?

		Null 1	Null 2	Null 3	Null 4
Run A	**Sign**	7.554388e-26	4.832767e-38	1.193051e-137	5.480738e-122
	Signed Rank	3.090888e-27	2.244648e-41	4.612996e-131	2.224541e-124
Run B	**Sign**	0	3.482397e-315	0	0
	Signed Rank	0	0	0	0
Run C	**Sign**	1.29939e-41	1.698042e-36	6.549949e-48	3.301627e-55
	Signed Rank	2.194097e-75	1.639146e-65	2.938907e-50	2.956933e-50
p-value product	**Sign**	0	0	0	0
	Signed Rank	0	0	0	0

"Sign" indicates non-parametric sign test p-value and "Signed Rank" indicates the non-parametric Wilcoxon Signed-Rank test p-value. 0 indicates that p<1e-324.

G4s and stays ("0 moves"): G4 motifs were paired with four different randomly sampled null distributions for the above statistical tests.
Null 1 = randomly selected positions on template and complement reads that do not have G4 motifs
Null 2 = randomly selected positions on all template and complement reads
Null 3 = randomly selected positions on the same template or complement read containing the G4 motif
Null 4 = randomly selected positions on the same template or complement read containing the G4 motif excluding positions that overlap the G4 motif

Note: 7.95%, 7.84%, and 7.24% of Template and Complement reads contain G4 motifs in Run A, Run B, and Run C, respectively.

Table A.10: Do G4 motifs on complement strand associate with more "0 moves" than G4 motifs on template?

	Rank Sum
Run A	1.033993e-30
Run B	3.952016e-26
Run C	0.004113
p-value product	1.680719e-58

"Rank Sum" indicates non-parametric Wilcoxon Rank Sum test p-value.

Table A.11: Do G4 motifs with >4 tracts associate with more "0 moves"?

	Rank Sum
Run A	0.002023
Run B	2.139787e-92
Run C	4.449e-14
p-value product	1.925878e-108

"Rank Sum" indicates non-parametric Wilcoxon Rank Sum test p-value.

Table A.12: Q score distributions for all reads, specific read types with >1 G4 motif, any read type with > 1 G4 motif

			Median Q	Mean Q	Std Dev of Q	Min Q	Max Q
Run A	All reads	2D	9.24	9.13	1.09	5.79	12.23
		Template	3.41	3.43	0.87	0.97	10.09
		Complement	4.01	3.94	0.95	1.13	6.44
	>1 G4 in specific read type	2D	8.33	8.51	0.95	6.62	11.27
		Template	3.85	3.79	0.71	1.18	5.57
		Complement	3.81	3.85	0.77	1.40	5.47
	≥1 G4 in any read type	2D	8.87	8.88	1.01	6.62	11.49
		Template	3.92	3.87	0.72	1.18	10.09
		Complement	3.92	3.91	0.81	1.40	5.72
Run B	All reads	2D	8.59	8.62	0.99	4.51	12.93
		Template	3.58	3.60	0.69	0.86	11.0
		Complement	4.21	4.16	0.80	0.89	7.31
	≥1 G4 in specific read type	2D	8.13	8.26	0.94	5.72	12.06
		Template	3.59	3.60	0.67	1.17	6.28
		Complement	3.95	3.97	0.76	1.36	6.29
	≥1 G4 in any read type	2D	8.46	8.50	0.97	5.72	12.06
		Template	3.72	3.71	0.65	1.17	6.28
		Complement	3.99	4.00	0.79	1.10	6.91
Run C	All reads	2D	9.07	8.97	0.98	4.96	12.95
		Template	3.62	3.61	0.87	1.11	7.46
		Complement	4.13	4.05	0.89	1.17	17.9
	>1 G4 in specific read type	2D	8.84	8.66	1.00	6.69	11.09
		Template	3.81	3.70	0.83	1.43	5.45
		Complement	3.93	3.83	0.81	1.18	5.43
	≥1 G4 in any read type	2D	8.97	8.83	0.94	5.66	11.09
		Template	4.11	3.91	0.81	1.43	5.94
		Complement	4.04	3.96	0.82	1.18	5.87

Table A.13: Schedule and recipes for adding sequencing mixes to flow cells during MinION sequencing.

	Run A				Run B				Run C			
	PSM	**EP**	**Fuel**	**HAP**	**PSM**	**EP**	**Fuel**	**HAP**	**PSM**	**EP**	**Fuel**	**HAP**
SM1	6	140	4	-	3	144	3	-	6	140	4	-
SM2	6	140	4	10	4	142	3	1	6	140	4	1
SM3	3	143	4	6	4	143	3	8	3	144	3	6
SM4	6	140	4	8	4	174	4	15	2	145	3	12
SM5	3	143	4	13	3	144	3	5.5	3	143	4	5
SM6	R	143	4	5	4	142	4	4	R	145	3	7
SM7	-	-	-	-	R	144	4	10	-	-	-	-

HAP = Hours After Previous (i.e. time after adding previous SM)
PSM = Pre-Sequencing Mix
SM = Sequencing Mix
EP = EP Buffer (ONT)
Fuel = "Fuel Mix" provided by ONT.
R = Remaining PSM
Sequencing Mix (SM) was prepared fresh immediately before adding to the flow cell at each time point.
All volumes reported for PSM, EP, and Fuel are in μl.

Supplementary Information: Single-molecule sequencing of long DNA molecules allows high contiguity de novo genome assembly and detection of DNA modification signatures for the fungus fly, Sciara coprophila

B.1 Supplementary Methods

DNA Extraction

PacBio sequencing details

Two DNA libraries were prepared and sequenced according to the manufacturers instructions and reflects the P5-C3 sequencing enzyme and chemistry, respectively. For each library, 6 μg of extracted, high-quality, genomic DNA isolated from Sciara coprophila was diluted in Qiagen elution buffer to 150 μL. The 150 μL aliquots were individually pipetted into the top chambers of Covaris G-tube spin columns and sheared gently for 60 seconds at 4500 rpm using an Eppendorf 5424 benchtop centrifuge and repeated at 60 seconds at 4500 rpm to further shear the DNA and place the aliquot back into the upper chamber, resulting in a 20,000 bp DNA shear, verified using a DNA 12000 Agilent Bioanalyzer gel chip. The sheared DNA was then re-purified using a 0.45X AMPure XP purification step (0.45X AMPure beads added, by volume, to each DNA sample dissolved in 200 μL EB, vortexed for 10 minutes at 2,000 rpm, followed by two washes with 70% alcohol and finally diluted in EB).

After purification, 2.7 μg of purified and sheared sample was taken into the DNA damage and end-repair steps. Briefly, the DNA fragments were repaired using DNA Damage Repair solution (1X DNA Damage Repair Buffer, 1X NAD+, 1 mM ATP high, 0.1 mM dNTP, and 1X DNA Damage Repair Mix) with a volume of 21.1 μL and incubated at 37°C for 20 minutes. DNA ends were repaired next by adding 1X End Repair Mix to the solution, which was incubated at 25°C for 5 minutes followed by the second 0.45X Ampure XP purification step. Next, 0.75 μM of Blunt Adapter was added to the DNA followed by 1X template Prep Buffer, 0.05 mM ATP low and 0.75 U/μL T4 ligase

to ligate (final volume of 47.5 μL) the SMRTbell adapters to the DNA fragments. This solution was incubated at 25°C overnight followed by a 65°C 10-minute ligase denaturation step. After ligation, the library was treated with an exonuclease cocktail to remove un-ligated DNA fragments using a solution of 1.81 U/μL Exo III 18 and 0.18 U/μL Exo VII, then incubated at 37°C for 1 hour. Two additional 0.45X Ampure XP purifications steps were performed to remove <2000 bp molecular weight DNA and organic contaminant.

Upon completion of library construction, the library was validated as 20 kb using another Agilent DNA 12000 gel chip. All libraries were sufficient for additional size selection to remove any library molecules < 7,000 bp. This step was conducted using Sage Science Blue Pippin 0.75% agarose cassettes to select library in the range of 7,000-50,000 bp. 16% of the input library eluted from the agarose cassette and was available for sequencing. This yield was sufficient to proceed to primer annealing and DNA sequencing on the PacBio RSII machine. Size-selection was confirmed by BioAnalysis and the mass was quantified using Qubit. Primer was then annealed to the size-selected SMRTbell with the full-length libraries (80°C for 2 minute 30 seconds followed by decreasing the temperature by 0.1° to 25C°). The polymerase-template complex was then bound to the P5 enzyme using a ratio of 10:1 polymerase to SMRTbell at 0.5 nM for 4 hours at 30°C and then held at 4°C until ready for magbead loading, prior to sequencing. The magnetic bead-loading step was conducted at 4°C for 60-minutes per manufacturers guidelines. The magbead-loaded, polymerase-bound, SMRTbell libraries were placed onto the RSII machine at a sequencing concentration of 75-100 pM across 24 SMRTcells and configured for 180-minute continuous sequencing runs on each SMRTcell. Sequencing was conducted to ample coverage across 24 SMRTcells.

Software versions used

ABruijn v0.3b
`https://github.com/fenderglass/ABruijn`
Git commit: 250761bac5589d83eeb1e00e19ba7fe78cbc1738

ABYSS v1.9.0
`https://github.com/bcgsc/abyss.git`
Git commit: 21106a4d0ad4b550da6f5b77b00620e530ca5037

ALE (C) 2010 Scott Clark

BEDtools v2.26.0
`https://github.com/arq5x/bedtools2.git`

BLAST v 2.2.30+
http://blast.ncbi.nlm.nih.gov/ v 2.2.30+

BlobTools v0.9.17
https://github.com/DRL/blobtools.git
383552447b2015620917aab8120699ae2d9d44be

Bowtie2 v2.2.9
https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.2.9/bowtie2-2.2.9-linux-x86_64.zip

BWA v 0.7.14-r1136
https://github.com/lh3/bwa.git
ee730f3832314fb70d369526275f2aadab064470

Canu v1.0, 1.1, 1.2 and several commits after v1.3 released
https://github.com/marbl/canu.git
4d07faca1d54852966e8ca9f4bbd4bf607b521d3

DBG2OLC
https://github.com/yechengxi/DBG2OLC.git
3e577397c0c2ad2a4b61da29be59572cef6e3086 I forked DBG2OLC and made/used a branch – obtain by:
git clone -b split_reads_by_backbone_minimize_open_files https://github.com/JohnUrban/DBG2OLC.git

Falcon v 0.7.3
git://github.com/PacificBiosciences/FALCON-integrate.git
c275aaac952ee1fd830dbe32ddc3bfb4e929fcb0

Fast5Tools
https://github.com/JohnUrban/fast5tools
c6e6026769457cbde57636f7044e288302d22946

FRCbam v1.3.0
https://github.com/vezzi/FRC_align.git
5b3f53e01cb539c857fd4230ec9410d76220fe22

HINGE

https://github.com/fxia22/HINGE.git
86db7bc630d2956031c4116739a7998f824738b9

LAP v1.1
http://www.cbcb.umd.edu/~cmhill/files/lap_release_1.1.zip

MaSuRCA v3.1.3
http://www.genome.umd.edu/masurca.html

Miniasm 0.2-r137-dirty
https://github.com/lh3/miniasm
17d5bd12290e0e8a48a5df5afaeaef4d171aa133

Minimap 0.2-r124-dirty
https://github.com/lh3/minimap 1cd6ae3bc7c7a6f9e7c03c0b7a93a12647bba244

Megahit v1.0.5
https://github.com/voutcn/megahit.git a851582f8a6e4ca9cbe41b946dc522336033adeb

Platanus v1.2.4
http://platanus.bio.titech.ac.jp

Pbh5tools v 0.8.0
https://github.com/PacificBiosciences/pbh5tools
2679c17687a868690595ae0d9e856c2fead28ff9

Poreminion v 0.4.4
https://github.com/JohnUrban/poreminion
c6a1bc8d7cfc7da675fdd0979622c4e200a16657

Quiver – GenomicConsensus_v1.1.0 (pbcore-1.2.4, ConcensusCore 1.0.1)
https://github.com/PacificBiosciences/GenomicConsensus

RaCon
https://github.com/isovic/racon
c342fc65f6fd686e7975238c207fb6f3aea671d1

REAPR v1.0.18
ftp://ftp.sanger.ac.uk/pub/resources/software/reapr/Reapr_1.0.18.test_data.tar.gz

SAMtools v1.3
https://github.com/samtools/samtools/releases/download/1.3/samtools-1.3.tar.bz2

SAMtool (HTSlib) v1.3
https://github.com/samtools/htslib/releases/download/1.3/htslib-1.3.tar.bz2

SGA v0.10.14
https://github.com/jts/sga.git 2ab3eae4d3d58b8d0dea3d60f2684c1181049a8f

Soapdenovo2 v2.04
https://github.com/aquaskyline/SOAPdenovo2.git dd6a98ba19bb21c3513a46ad5047d08e57583ab0

Velvet 1.2.08
https://www.ebi.ac.uk/~zerbino/velvet/

Illumina Assemblies

For most assemblers, we tried using all reads, all reads after BayesHammer [Nikolenko et al., 2013] from SPAdes [Bankevich et al., 2012] correction, filtered reads using Trimmomatic [Bolger et al., 2014], and filtered reads that were subsequently corrected with BayesHammer. Note that the BayesHammer corrected reads were produced as part of running SPAdes.

For trimming, we used Trimmomatic version 0.32:

IN="MSF0007_TGACCA_L004_R1_001.fastq.gz MSF0007_TGACCA_L004_R2_001.fastq.gz"
OUTPRE=maleSciaraTrimmed
OUT="${OUTPRE}_forward_paired.fq.gz ${OUTPRE}_forward_unpaired.fq.gz ${OUTPRE}_reverse_paired.fq.gz ${OUTPRE}_reverse_unpaired.fq.gz"
PRIMERS=neb_primers.fa
MINQUAL=5
MINLEN=100
java -jar /Path/To/Trimmomatic-0.32/trimmomatic-0.32.jar PE -trimlog trim.log $IN $OUT ILLUMINACLIP:${PRIMERS}:2:30:10 LEADING:$MINQUAL TRAILING:$MINQUAL SLIDINGWINDOW:4:$MINQUAL MINLEN:$MINLEN

Trimming produced 3 files: one file each for the forward and reverse reads that remained paired and one file containing the orphaned reads. In addition to the paired reads, the orphaned reads were

used with most assemblers.

Where relevant below:
- let the set of all reads (no quality filtering) be referred to as R1.fastq and R2.fastq
- let the quality-filtered reads be referred to as R1.q5.fastq, R2.q5.fastq, and SE.q5.fastq
- let the BayesHammer-corrected reads (using all read, no prior quality filtering) be referred to as R1.bh.fastq, R2.bh.fastq, and SE.bh.fastq (SE for those orphaned in the process)
- let the BayesHammer-corrected quality-filtered reads be referred to as R1.q5.bh.fastq, R2.q5.bh.fastq, and SE.q5.bh.fastq (quality-filtering and bayeshammer orphans all in same SE file)
- let corresponding sets of contamination-filtered (cf) reads after iteration "i" of contamination filtering be referred to as above, but with "cf$i" inserted inside the name. For example after iteration #1 (i=1):
– R1.cf1.fastq and R2.cf1.fastq
– R1.q5.cf1.fastq, R2.q5.cf1.fastq, and SE.q5.cf1.fastq
– Noting that reads are quality-filtered before filtered for contamination
– R1.cf1.bh.fastq, R2. cf1.bh.fastq, and SE.cf1.bh.fastq
– Noting the contamination-filtering comes before error-correction.
– R1.q5.cf1.bh.fastq, R2.q5.cf1.bh.fastq, and SE.q5.cf1.bh.fastq
– Noting quality-filtering precedes contamination-filtering, which precedes error-correction.

Where relevant below, mean (432), min (83) and deviation of insert sizes were obtained using Picard-Tools CollectInsertSizeMetrics after mapping the paired-end reads to an early long read assembly with Bowtie2.

ABYSS

Abyss – unfiltered reads:
K=k #where k was either 55 or 77
abyss-pe name=sciara j=16 k=$K in='R1.fastq R2.fastq'

Abyss – filtered reads:
K=k #where k was either 55 or 77
abyss-pe name=sciara j=16 k=$K lib='pe432' pe432='R1.q5.fastq R2.q5.fastq' se='SE.q5.fastq'

Megahit

Megahit – unfiltered reads:
megahit -1 R1.fastq -2 R2.fastq

Megahit – filtered reads:
megahit -1 R1.q5.fastq-2 R2.q5.fastq -r SE.q5.fastq

Platanus

Platanus – unfiltered reads:
platanus assemble -f R1.fastq R2.fastq -o out -t 16 -m 45
platanus scaffold -c out_contig.fa -b out_contigBubble.fa -IP1 R1.fastq R2.fastq -n1 83 -a1 432 -d1 21 -t 16 -o out
platanus gap_close -o out -c out_scaffold.fa -IP1 R1.fastq R2.fastq -t 16

Platanus – filtered reads:
platanus assemble -f R1.q5.fastq R2.q5.fastq SE.q5.fastq -o out -t 16 -m 45
platanus scaffold -c out_contig.fa -b out_contigBubble.fa -IP1 R1.fastq R2.fastq -n1 83 -a1 432 -d1 21 -t 16 -o out
platanus gap_close -o out -c out_scaffold.fa -IP1 R1.fastq R2.fastq -t 16

SGA

SGA – unfiltered:
#SAMtools 0.1.19 and ABYSS 1.9.0 in PATH.
PE1=R1.fastq
PE2=R2.fastq
See SGA recipe below.

SGA – filtered:
#SAMtools 0.1.19 and ABYSS 1.9.0 in PATH.
PE1=R1.q5.fastq
PE2=R2.q5.fastq
See SGA recipe below.
Note: SE.q5.fastq not used.

SGA recipe – same for unfiltered and filtered reads:

```
# Largely drawn from example here: https://github.com/jts/sga/blob/master/src/examples/sga-celegans.sh
# SGA has its own read correction, so did not use with BayesHammer corrected reads
NAME=sciara-male
OL=75
T=16
D=4000000
CK=51
COV_FILTER=2
MOL=55
R=10
MIN_PAIRS=5
MIN_LENGTH=200
SCAFFOLD_TOLERANCE=1
MAX_GAP_DIFF=0
CTGS=assemble.m$OL-contigs.fa
GRAPH=assemble.m$OL-graph.asqg.gz
SGA_ALIGN= /software/sga/sga/src/bin/sga-align
sga preprocess –pe-mode 1 $PE1 $PE2 > $NAME.fastq
sga index -a ropebwt –no-reverse -t $T $NAME.fastq
sga preqc -t $T ${NAME}.fastq > $NAME.preqc
sga correct -k $CK –discard –learn -t $T -o ${NAME}.ec.k$CK.fastq ${NAME}.fastq
sga index -a ropebwt -t $T ${NAME}.ec.k$CK.fastq
sga filter -x $COV_FILTER -t $T –homopolymer-check –low-complexity-check ${NAME}.ec.k$CK.fastq
sga fm-merge -m $MOL -t $T -o ${NAME}.merged.k$CK.fa ${NAME}.ec.k$CK.filter.pass.fa
sga index -d 1000000 -t $T ${NAME}.merged.k$CK.fa
sga rmdup -t $T ${NAME}.merged.k$CK.fa
sga overlap -m $MOL -t $T ${NAME}.merged.k$CK.rmdup.fa
sga assemble -m $OL -g $MAX_GAP_DIFF -r $R -o assemble.m$OL ${NAME}.merged.k$CK.rmdup.asqg.gz
$SGA_ALIGN –name ${NAME}.pe $CTGS $PE1 $PE2
sga-bam2de.pl -n $MIN_PAIRS –prefix libPE ${NAME}.pe.bam
sga-astat.py -m $MIN_LENGTH ${NAME}.pe.refsort.bam > libPE.astat
sga scaffold -m $MIN_LENGTH –pe libPE.de -a libPE.astat -o scaffolds.n$MIN_PAIRS.scaf $CTGS
sga scaffold2fasta -m $MIN_LENGTH -a $GRAPH -o scaffolds.n$MIN_PAIRS.fa -d $SCAFFOLD_TOLERANCE
–use-overlap –write-unplaced scaffolds.n$MIN_PAIRS.scaf
```

SOAPdenovo2

Config file for all:
#For example config file: http://soap.genomics.org.cn/soapdenovo.html
#maximal read length
max_rd_len=100

[LIB]
#average insert size
avg_ins=432
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used
asm_flags=3
#use only first 100 bps of each read
rd_len_cutoff=100
#in which order the reads are used while scaffolding
rank=1
# cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=3
#minimum aligned length to contigs for a reliable read location (at least 32 for short insert size)
map_len=32

#a pair of fastq file, read 1 file should always be followed by read 2 file
q1=R1.fastq ## dependent on which set of reads being used
q2=R2.fastq ## dependent on which set of reads being used

#fastq file for single reads
q=SE.fastq ## dependent on which set of reads being used (some do not have)

SPAdes

[square brackets only applicable when single-end read file used]
spades.py –pe1-1 $PE1 –pe1-2 $PE2 [–pe1-s $SE] -o default -t 64 –only-assembler –cov-cutoff auto
spades.py -k 21,33,55,77 –pe1-1 $PE1 –pe1-2 $PE2 [–pe1-s $SE] -o k2177 -t 64
spades.py -k 21,33,55,77 –pe1-1 $PE1 –pe1-2 $PE2 [–pe1-s $SE] -o k2177auto -t 64 –cov-cutoff auto

Velvet

Velvet – unfiltered reads:
[square brackets only applicable when single-end read file used]
K=K (55 or 77)
DIR=hash${K}
velveth ${DIR} ${K} -fmtAuto -shortPaired -separate $R1 $R2 [-short $SE]
velvetg ${DIR} -clean yes -exp_cov auto -cov_cutoff auto

Velvet – unfiltered reads:
#Ran with VelvetOptimiser 2.2.5 that came with Velvet, contributed to Velvet package by:
torsten.seemann@unimelb.edu.au, simon.gladman@unimelb.edu.au
VelvetOptimiser.pl –v –s 33 –e 77 –x 22 –f '-fmtAuto -shortPaired -separate R1.q5.fastq R2.q5.fastq –short2 SE.q5.fastq' –t 8

Evaluations

LAP

With sampleFastq.py from `https://github.com/JohnUrban/sciara-project-tools`
R1=R1.fastq
R2=R2.fastq
FILE=ASM.name.fasta
BASE=ASM.name
P=16
PROPORTION=p
#where p was either 0.0001, 0.001, and 0.01 for approximately 15k, 150k, and 1.5 M sampled paired reads respectively
#sampleFastq.py ouputs downsampled.1.fastq and downsampled.2.fastq

sampleFastq.py -1 R1.fastq -2 R2.fastq -wo -p $PROPORTION -o downsampled
bowtie2-build $FILE $BASE
calc_prob.py -p $P -a $FILE -q -1 $R1 -2 $R2 -X 800 -I 0 -o fr -m 432 -t 75 -b $BASE > ${BASE}.prob
sum_prob.py -i ${BASE}.prob > ${BASE}.lapscore

ALE and FRCbam

With: Bowtie 2 version 2.2.9, samtools v1.3, boost 1.55, ALE (Scott Clark 2010)
FILE=ASM.name.fasta
BASE=ASM.name
bowtie2-build $FILE $BASE
bowtie2 -p $P -q –very-sensitive -N 1 -x $BASE -1 $R1 -2 $R2 2> $base.err | samtools sort -o reads.bam
samtools index reads.bam
ALE reads.bam $FILE ${BASE}.ALE.txt >> $BASE.err
FRC –pe-sam reads.bam –pe-max-insert 800 –genome-size 292000000 –output ${BASE}

Note: Abyss outputs contigs with non-ACGT IUPAC letters, which was breaking ALE. Therefore, each non-ACGT IUPAC letter was converted to one of the ACGT bases it represented at random with respect to the frequency ACGT occurs in the assembly. This was done using IUPAC-to-ACGT.py from: `https://github.com/JohnUrban/sciara-project-tools`
IUPAC-to-ACGT.py --f abyss.asm.fasta > abyss.asm-acgt.fasta

REAPR

reapr facheck asm.fasta asm_renamed
reapr perfectmap asm_renamed.fa $R1 $R2 432 perfect
reapr smaltmap asm_renamed.fa $R1 $R2 mapped.bam -n $P
reapr pipeline asm_renamed.fa mapped.bam output_directory perfect

BlobTools and contamination analyses

BlobTools, contamination analyses, iterating Platanus assemblies:
Obtaining taxonomy info:
Platanus contig fasta (from assembly of quality filtered, error-corrected reads) was subdivided into 1000 different fasta files and submitted to SLURM as a batch script array that ran the following on each:
blastn -task megablast -query $Q -db $NT \
-outfmt '6 qseqid staxids bitscore std sscinames sskingdoms stitle' \
-culling_limit 5 \
-num_threads $P \
-evalue 1e-25 \

-out ${BLASTDIR}/${PRE}.${SLURM_ARRAY_TASK_ID}.blastout

After ensuring all jobs completed successfully (and re-running if not), final blast results were all combined into a single file.
Looking at blobplots using platanus kmer coverage:
asm=platanus.q5.bh.contig
blobtools create -i ${asm}.fa -y platanus –nodes tax/nodes.dmp –names tax/names.dmp -o ${asm} -t ${asm}.blast.results
blobtools blobplot -i ${asm}.BlobDB.json
blobtools blobplot -i ${asm}.BlobDB.json -r superkingdom
blobtools blobplot -i ${asm}.BlobDB.json -r order -p 10
blobtools blobplot -i ${asm}.BlobDB.json -r family -p 12

Blob plots using pre-amplification salivary gland coverage:
reads=pre-amp-sal-gland-reads
bowtie2 -p 8 -q –very-sensitive -N 1 -x ${asm} -U ${reads}.fq | samtools sort –threads 8 -T ${reads}.tmp -o ${reads}.bam
blobtools bam2cov -i ${asm}.fa -b ${reads}.bam -o bam2cov –mq 0
grep -v ^# bam2cov.{reads}.bam.cov | cut -f 1,3 > ${reads}.cov
blobtools create -i ${asm}.fa –cov ${reads}.cov –nodes tax/nodes.dmp –names tax/names.dmp -o ${reads}.${asm} -t $B
DB=${reads}.${asm}.BlobDB.json
##Then plots made similar to above
Getting table with taxon summaries for each contig:
blobtools view -o ${reads}.${asm}.table -i $DB -r all

Constructing BED file of contigs to keep:
grep -v ^# ${reads}.${asm}.table | awk '$6 != "Bacteria" && $5 >= 0.1' | awk 'OFS="\t" {print $1,0,$2}' | sortBed -i - > contigs.to.keep.bed

grep -v ^# ${reads}.${asm}.table | awk '$6 == "Bacteria" || $5 < 0.1' | grep Arthropoda | awk 'OFS="\t" {print $1,0,$2}' | sortBed -i - >> contigs.to.keep.bed

Get read pairs where at least 1 read maps inside contigs.to.keep:
bedtools pairtobed -abam $BAM -b contigs.to.keep.bed | samtools sort -n | samtools fastq -1 ${PRE}.m.R1.fq -2 ${PRE}.m.R2.fq -n -

Get reads where neither maps to anything in the assembly:
samtools view -bS -f12 $BAM | samtools sort -n | samtools fastq -1 ${PRE}.u_u.R1.fastq -2 ${PRE}.u_u.R2.fastq

-n -

Then 4 new platanus assemblies were made from:
(i) contamination-filtered reads ("contfilt1.noqualfilt")
platanus assemble -f $PE1 $PE2 -o $PREFIX -t $T -m $M
platanus scaffold -c $contig -b $bubble -IP1 $PE1 $PE2 -n1 83 -a1 432 -d1 21 -t $T -o $PREFIX
platanus gap_close -o $PREFIX -c $scaffold -IP1 $PE1 $PE2 -t $T

(ii) contamination-filtered reads that were then error-corrected with bayeshammer ("contfilt1.noqualfilt.bh"):
spades.py –pe1-1 $PE1 –pe1-2 $PE2 -o assembly -t 16 –only-error-correction
platanus assemble -f $PE1 $PE2 $SE -o $PREFIX -t $T -m $M
platanus scaffold -c $contig -b $bubble -IP1 $PE1 $PE2 -n1 83 -a1 432 -d1 21 -t $T -o $PREFIX
platanus gap_close -o $PREFIX -c $scaffold -IP1 $PE1 $PE2 -t $T

(iii) contamination-filtered reads that were quality filtered ("contfilt1.q5")
java -jar $TRIMMOMATIC_BASE/Trimmomatic-0.32/trimmomatic-0.32.jar PE -trimlog trim.log $IN $OUT ILLUMINACLIP:${PRIMERS}:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:5 MINLEN:100
platanus assemble -f $PE1 $PE2 $SE -o $PREFIX -t $T -m $M
platanus scaffold -c $contig -b $bubble -IP1 $PE1 $PE2 -n1 83 -a1 432 -d1 21 -t $T -o $PREFIX
platanus gap_close -o $PREFIX -c $scaffold -IP1 $PE1 $PE2 -t $T

(iv) contamination-filtered reads that were quality filtered and error-corrected, starting with quality filtered reads above ("contfilt1.q5.bh"):
spades.py –pe1-1 $PE1 –pe1-2 $PE2 –pe1-s $SE -o assembly -t 16 –only-error-correction
#SE read files were combined into 1
platanus assemble -f $PE1 $PE2 $SE -o $PREFIX -t $T -m $M
platanus scaffold -c $contig -b $bubble -IP1 $PE1 $PE2 -n1 83 -a1 432 -d1 21 -t $T -o $PREFIX
platanus gap_close -o $PREFIX -c $scaffold -IP1 $PE1 $PE2 -t $T

Bowtie2 indexes were made for each assembly:
bowtie2-build $FILE $base

For ALE and FRCbam, reads were mapped with bowtie2:
bowtie2 -p $P -q –very-sensitive -N 1 -x $base –maxins 800 -1 $R1 -2 $R2 2> $base.err | samtools sort -o reads.bam
samtools index reads.bam

Evaluations of contamination-filtered re-assemblies:
Assemblies (gap-closed scaffolds) were evaluated with the following tools. The set of contamination filtered reads (used to assemble "contfilt1.noqualfilt") were used for these evaluations. First, Bowtie2 indexes were made for each assembly:
bowtie2-build $FILE $base

For LAP, 145,000 reads were sampled using our sampling tool (https://github.com/JohnUrban/sciara-project-tools):
sampleFastq.py -1 $R1 -2 $R2 -wo -p 0.001 -o downsampled

For ALE and FRCbam, reads were mapped with bowtie2:
bowtie2 -p $P -q –very-sensitive -N 1 -x $base –maxins 800 -1 $R1 -2 $R2 2> $base.err | samtools sort -o reads.bam
samtools index reads.bam

(i) LAP
calc_prob.py -p $P -a $FILE -q -1 $R1 -2 $R2 -X 800 -I 0 -o fr -m 432 -t 75 -b $base > ${pre}.prob
sum_prob.py -i ${pre}.prob > ${pre}.lapscore

(ii) ALE
ALE reads.bam $FILE ${base}.ALE.txt >> $base.err

(iii) FRCbam
FRC –pe-sam reads.bam –pe-max-insert 800 –genome-size 292000000 –output ${base}.frc

(iv) REAPR
reapr facheck asm.fasta asm_renamed
reapr perfectmap asm_renamed.fa $R1 $R2 432 perfect
reapr smaltmap asm_renamed.fa $R1 $R2 mapped.bam -n $P
reapr pipeline asm_renamed.fa mapped.bam output_directory perfect

Assembly scores were similar. "contfilt1.q5" received best LAP and ALE scores. "confilt1.noqualfilt" had fewest number of features detected by FRCbam. "contfilt1.q5.bh" had the highest percentage of error-free bases as determined by REAPR. Since "contfilt.q5.bh" had the highest REAPR score and since q5.bh won out in the previous set of evaluations before contamination filtering, we chose this assembly to move forward with. The "contfilt1.q5.bh" contigs and scaffolds were separately BLAST for taxonomy information as above. BlobTools databases and plots were constructed as above for kmer coverage as well as coverage from pre-amplification salivary gland reads.

Removing scaffolds that have taxonomy information suggesting it is not arthropod sequence:
We used scripts (https://github.com/JohnUrban/sciara-project-tools/tree/master/taxon) that segregate sequence names based on the closest taxonomy level (species, genus, family, order, class, phylum, kingdom, superkingdom, othersuperkingdoms) to our target species (Bradysia coprophila, Bradysia, Sciaridae, Diptera, Insecta, Arthropoda, Metazoa, Eukaryota, Bacteria/Archaea) that is found in the BLAST hits for each sequence. We required only 1 BLAST hit for a given level for it to be considered the closest level (i.e. if 1 BLAST hit says "Bradysia coprophila", then it will be assigned to the species level as the closest level even if other levels have more hits).

```
NODES=tax/nodes.dmp
NAMES=tax/names.dmp
BFILE=platanus.blobfilt1.q5.bh.contig.blast.results
TAX=$BFILE.taxonomy.out
PRE=platanus.blobfilt1.q5.bh.contig.taxsum
ALL=q5-bh/platanus.blobfilt1.q5.bh_contig.names
taxonomyFromTaxID.py -i $BFILE -nc 1 -tc 2 -no $NODES -na $NAMES > $TAX
taxonomy-summarizer.py -i $TAX -o $PRE -a $ALL
```

The same contigs and scaffolds came back as Bacterial as with the BlobTools approach. Thirteen contigs (26 scaffolds) that were not labeled as Arthropod in the BlobTools table were labeled as Arthropod or closer to Bradysia coprophila with this more conservative approach. However, we checked a few contigs (and scaffolds) with the closest hits from the Metazoan kingdom with blastn (as opposed to megablast), and they came back with Drosophila and Anopheles genuses. Therefore, we used the more sensitive blastn to gather taxonomy information before filtering.

```
blastn -task blastn -query $FASTA -db $NT \
-outfmt '6 qseqid staxids bitscore std sscinames sskingdoms stitle' \
-num_threads $P  textbackslash
-evalue 1e-25  textbackslash
-out ${BLASTDIR}/${PRE}.blastout
```

Our taxonomy scripts were run:

```
taxonomyFromTaxID.py -i blastn.results -no nodes.dmp -na names.dmp > blastn.results.taxsum
taxonomy-summarizer.py -i blastn.results.taxsum -o taxsummary -a all-scaf-names.txt
```

Any contig/scaffold that had taxonomy information, but did not have at least one hit in the Arthropod phylum with blastn was removed.

```
cat taxsummary.kingdom.Metazoa.out taxsummary.othersuperkingdom1.Bacteria.out taxsummary.othersuperkingdo
taxsummary.superkingdom.Eukaryota.out | cut -f 1 > exclude.txt
```

extractFastxEntries.py -n exclude.txt -f plat.q5.bh.penultimate.scaffolds.fa -e > final.scaffolds.fa

This final scaffolds were used for further testing.

Long Read Assemblies

Obtaining fasta/fastq files from PacBio .bax.h5 files

Bash5tools.py from pbh5tools was used.
Either all PacBio sub-reads were used, or quality filtered subreads with a cutoff of 0.75.

bash5tools.py –readType subreads –outType fasta –outFilePrefix $PREFIX file.bax.h5

bash5tools.py –readType subreads –outType fasta –minReadScore 0.75 –outFilePrefix $PREFIX file.bax.h5

Obtaining fasta/fastq files from MinION fast5 files

Fast5 files were base-called with Oxford Nanopores Metrichor cloud base-caller at the time the data was produced. See information on libraries elsewhere for dates and MAP kit versions. Metrichor (ONT base-caller) returns updated fast5 files into two folders: “pass” and “fail”. “Pass” contains only fast5 files where 2D base-calling was successful and the mean quality of the 2D read is >9. Everything else (including other fast5s containing 2D reads with Q<9, fast5s with only 1D base-calling, and fast5s that failed base-calling) goes into the “fail” folder. For libraries 01, 02, 03, 04, 05, 06, 07, 08, and 09, our toolset called poreminion (https://github.com/JohnUrban/poreminion) was used as in Urban et al 2015 to filter fast5s in the fail/ that were not base-called due to various reasons: no template, too few events, too many events, or contained the event time error where blocks of events are repeated that was present for a small proportion of fast5 files in the early days of the MinION Access Program.
$ poreminion uncalled -m -o fail-filter fail/
$ poreminion timetest -m -o fail-filter fail/

The error of repeated blocks of events in a small proportion of fast5 files was fixed by ONT before constrction of Library09, which did not have any as expected. For all libraries, the pass/ and fail/ folders were individually archived and compressed with “tar -xzf pass.tar.gz pass/” or “tar -xzf fail.tar.gz fail/”. Since the innards of the HDF5 fast5 files changed many times, we created another toolset called Fast5Tools (https://github.com/JohnUrban/fast5tools) that is able to deal with all of the different fast5 versions used in this study. It was created to be more robust to changes and to easily allow modifications to deal with new versions. Fast5Tools maintains most of the functionality

of poreminion, although pre-filtering is no longer necessary.

For all libraries, Fast5Tools was used to extract fasta and fastq information from the pass.tar.gz and fail.tar.gz files. Using the --tarlite option, it is able to extract one fast5 file at a time from the tarchive, use it as necessary, and delete it before moving on. This prevents issues of disk space that may arise from full extracting the contents of the giant tar files before performing the needed operations. First to extract either all fasta or all fastq reads from every fast5 file, fast5tofastx.py was used:

$ fast5tofastx.py -r all -o fasta –tarlite pass.tar.gz > pass.all.fa
$ fast5tofastx.py -r all -o fasta –tarlite fail.tar.gz > fail.all.fa

The fast5 files can be filtered for length and quality information at the stage of using fast5tofastx.py. However, the headers of the resulting fasta/fastq entries contain all the information necessary to filter for reads using length, mean quality score, and other information. If the "-r all" option is used as above, then filtering can be done on the fasta/fastq files exactly as from fast5 files. This way one can keep the smaller fasta/fastq file and put the larger tarchives in longer term storage if necessary. The basic header structure looks like this:
>readtype|len:LEN|Q:Q|channel:CHANNEL|Read:READNUMBER|asic:ASIC_ID|run:RUN_ID|device:DEVICE_ID|

An example of the 3 read types (template, complement, and 2d) from a passing molecule:
>template|len:2781|Q:4.45259884791|channel:508|Read:80|asic:3372918206|run:609a1b4408f2200edefabc0d97b202bca
>complement|len:2783|Q:4.43810484237|channel:508|Read:80|asic:3372918206|run:609a1b4408f2200edefabc0d97b202
>2d|len:3248|Q:10.8914446375|channel:508|Read:80|asic:3372918206|run:609a1b4408f2200edefabc0d97b202bcaccf94d

The fasta/fastq derived from the fast5 files with Fast5Tools can then be filtered using filterFast5DerivedFastx.py. We obtained all 2D reads from the fasta/fastq file by:

$ filterFast5DerivedFastx.py -r 2d libX.pass.all.fa > libX.pass.2d.fa
$ filterFast5DerivedFastx.py -r 2d libX.fail.all.fa > libX.fail.2d.fa

High quality pass 2D reads can be obtained by filtering just the pass.all.fa file for each library (using the above command) or if all reads are combined in a single fasta, doing:

$ filterFast5DerivedFastx.py -r 2d libX.all.fa --minq 9 > libX.pass.2d.fa

We select 1 read per molecule using the "molecule" definition from Urban et al. 2015. This takes the 2D read if one is present, the longer of template or complement if 2D is not present (but complement is), or the template read if it is the only one present.

$ filterFast5DerivedFastx.py -r molecule libX.pass.all.fa > libX.pass.molecule.fa
$ filterFast5DerivedFastx.py -r molecule libX.fail.all.fa > libX.fail.molecule.fa

One could also use the "MoleQual" approach (-r MoleQual) where template vs. complement is chosen based on which one has a higher mean quality score instead of which is longer. However, we sought to use information gained from length in this study. In either case, one can also filter for some minimum mean quality score as well with --minq.

To obtain statistics from each fast5 file on the reads present, Fast5Tools can work directly with the fast5 files (or tarchive of fast5 files) as well as from the fast5-derived fasta files (particularly when all reads are dumped out from each file as above). Both approaches give the same results concerning length and mean quality scores (more information about events can be obtained directly from the fast5 file):

$ fast5stats.py –standard –errfile libX.errfiles.txt –tarlite libX.pass.tar.gz libX.fail.tar.gz > stats.txt
$ fast5DerivedFastxMoleculeStats.py libX.pass.all.fa libX.fail.all.fa > stats.txt

Note that "–errfile errfiles.txt" is populated with information on what files had errors or were not base-called for various reasons (as stated above). There is no option for this when grabbing the statistics from fasta files as any erroneous or un-basecalled fast5 is ignored during extraction with fast5tofastx.py. The output of these commands share the first 11 columns (fast5stats.py can produce more) and there is one line per molecule – i.e. one line per original fast5 file (even if the information is coming from fasta, which unites different reads from the same molecule from shared molecule name derived from the channel, the read number, the asic ID, the run ID, and device ID):
1 = molecule name
2 = molecule length
3 = molecule mean q score
4 = has complement
5 = has 2d
6 = 2d seq len
7 = template seq len
8 = complement seq len
9 = 2d mean q score
10 = template mean q score
11 = complement mean q score

The first 11 columns of the output table was summarized with fast5standardSummary.py:

$ fast5standardSummary.py -V -f libX.stats.txt

At the moment, this outputs 264 pieces of information including length statistics (minimum, maximum, mean, median, N25, N50, N75, expected length) on molecules (as defined in Urban et al 2015) as well as the various read types. One can simply combine all stats.txt tables from each library to compute the summary on all molecules from all libraries.

The headers for fast5-derived fasta files are not appropriate for use with falcon. To use our MinION data with the Falcon assembler, we used:

$ filterFast5DerivedFastx.py -o falcon file.fa > file.forfalcon.fa

ABRUIJN

abruijn.py longreads.fasta sciara $COV -t 32 –iterations 1

where longreads.fasta was either:
- all PacBio subreads
- All PacBio subreads and Oxford Nanopore “molecule” reads (one read per molecule)
- PacBio reads filtered to be > 0.75 and Oxford Nanopore 2D reads

For all PB reads, COV=42.
For all PB reads with 1 read per nanopore fast5 (“molecule” reads), COV=54.
For quality filtered PB reads with 2D nanopore reads, COV=47.

ABruijn runs required allocating 512 GB RAM and 32 CPUs, and 7-8 days of running time. Thus, it was impractical to try exhaustive combinations of reads, filtering, etc.

SMARTdenovo

smartdenovo.pl -c 1 longreads.fasta > wtasm.mak
make -f wtasm.mak

where longreads.fasta was either:
- all PacBio subreads
- PacBio reads filtered to be > 0.75
- All PacBio subreads and Oxford Nanopore “molecule” reads (one read per molecule)

- All PacBio subreads and Oxford Nanopore 2D reads
- PacBio reads filtered to be > 0.75 and Oxford Nanopore 2D reads

SMARTdenovo required <48 GB RAM, 8 threads, and <15 hours to complete. It was therefore practical to try more combinations of reads, filtering, etc.

Canu

PB=all_subreads.fasta
PBFILT=subreads.gt0.75.fasta
ONT2d=ont2d.fasta
ONTmol=ontmolecule.fasta
PBONT2D=allPBsubreads_ont2d.fasta
PBFILTONT2D=pbsubreads.gt0.75_ont2d.fasta
PBMOL=allPBsubreads_ontmolecule.fasta
PBMOL_MULTICORR= pball.ontmol.corrected.iter4.fasta #see below
COROUTCOV=500 ## excessively high to include all (not used in all commands though)
T=24:00:00
G=292m

One set of reads containing all PacBio subreads and Oxford Nanopore molecule reads (1 read per molecule) was corrected with Canu in an iterative fashion as follows:
(i) canu -correct corOutCoverage=500 corMinCoverage=0 corMhapSensitivity=high -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB -nanopore-raw $ONTmol "gridOptions=–time $T" oeaMemory=8 corMemory=40 minReadLength=500

This produced pball.ontmol.corrected.iter1.fasta

(ii) canu -correct corOutCoverage=500 corMinCoverage=0 corMhapSensitivity=high -p $NAME -d $NAME genomeSize=$G -nanopore-raw corrected.iter1.fasta "gridOptions=–time $T" oeaMemory=8 corMemory=40 minReadLength=500

(iii) Same as (ii) but with "nanopore-raw pball.ontmol.corrected.iter2.fasta"

(iv) Same as (iii) but with "nanopore-raw pball.ontmol.corrected.iter3.fasta"

The last iteration produced pball.ontmol.corrected.iter4.fasta. This was used as input to various assemblers. When input into Canu or Falcon, a 5th round of correction was performed accordingly

(i.e. no steps in either pipeline were skipped).

NAME=g292-default-all
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB “gridOptions=–time $T” oeaMemory=8

NAME=g292-default-all-min500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB “gridOptions=–time $T” minReadLength=500 oeaMemory=8

NAME=g292-default-all-corcov500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB “gridOptions=–time $T” oeaMemory=8 corOutCoverage=$COROUTCOV

NAME=g292-default-all-corcov500-min500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB “gridOptions=–time $T” minReadLength=500 oeaMemory=8 corOutCoverage=$COROUTCOV

NAME=g292-default-q0.75-corcov500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PBFILT “gridOptions=–time $T” oeaMemory=8 corOutCoverage=$COROUTCOV

NAME=g292-default-q0.75-corcov500-min500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PBFILT “gridOptions=–time $T” minReadLength=500 oeaMemory=8 corOutCoverage=$COROUTCOV

NAME=g292-default-pball-ont2d
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB -nanopore-raw $ONT2d “gridOptions=–time $T” oeaMemory=8 corMemory=30

NAME=g292-default-pball-ont2d-corcov500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB -nanopore-raw $ONT2d “gridOptions=–time $T” oeaMemory=8 corOutCoverage=$COROUTCOV corMemory=30

NAME=g292-default-pball-ont2d-corcov500-min500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB -nanopore-raw $ONT2d “gridOptions=–time $T” minReadLength=500 oeaMemory=8 corOutCoverage=$COROUTCOV corMemory=30

NAME=g292-default-pball-ont2d-corcov500-min500-aspb
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PBONT2D “gridOptions=–time $T” minReadLength=500 oeaMemory=8 corOutCoverage=$COROUTCOV corMemory=30

NAME=g292-default-pball-ontmol
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB -nanopore-raw $ONTmol “gridOptions=–time $T” oeaMemory=8 corMemory=40

NAME=g292-default-pball-ontmol-corcov500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB -nanopore-raw $ONTmol “gridOptions=–time $T” oeaMemory=8 corOutCoverage=$COROUTCOV corMemory=40

NAME=g292-default-pball-ontmol-corcov500-min500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PB -nanopore-raw $ONTmol “gridOptions=–time $T” minReadLength=500 oeaMemory=8 corOutCoverage=$COROUTCOV corMemory=40

NAME=g292-default-pball-ontmol-corcov500-min500-aspb
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PBMOL “gridOptions=–time $T ” minReadLength=500 oeaMemory=8 corOutCoverage=$COROUTCOV corMemory=40

NAME=g292-default-pbq75-ont2d
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PBFILT -nanopore-raw $ONT2d “gridOptions=–time $T” oeaMemory=8 corMemory=30

NAME=g292-default-pbq75-ont2d-corcov500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PBFILT -nanopore-raw $ONT2d “gridOptions=–time $T” oeaMemory=8 corOutCoverage=$COROUTCOV corMemory=30

NAME=g292-default-pbq75-ont2d-corcov500-min500
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PBFILT -nanopore-raw $ONT2d “gridOptions=–time $T” minReadLength=500 oeaMemory=8 corOutCoverage=$COROUTCOV corMemory=30

NAME=g292-default-pbq75-ont2d-corcov500-min500-aspb
canu -p $NAME -d $NAME genomeSize=$G -pacbio-raw $PBFILTONT2D “gridOptions=–time $T” minReadLength=500 oeaMemory=8 corOutCoverage=$COROUTCOV corMemory=30

Canu is very sophisticated in automatically requesting appropriate resources and distributing/parallelizing its jobs across as many CPUS/nodes available on our compute cluster. It therefore

finishes within a day or 2, and it was possible to test many different combinations of reads, filtering, parameters, etc. Indeed, we have also tested many earlier versions of Canu (not reported here) on many aforementioned combinations above.

Miniasm and RaCon

minimap -Sw5 -L100 -m0 -t8 longreads.fasta longreads.fasta | gzip -1 > reads.paf.gz
miniasm -f longreads.fasta reads.paf.gz > reads.gfa
awk '/^S/{print ">"$2"\n"$3}' reads.gfa > asm.fasta
minimap asm.fasta longreads.fastq > overlaps-for-racon.paf
racon -t 8 reads.fastq overlaps-for-racon.paf asm.fasta asm.racon.fasta

Where longreads.fasta was either:
- all PacBio subreads
- PacBio reads filtered to be > 0.75
- All PacBio subreads and All Oxford Nanopore reads (up to 3 reads per molecule)
- All PacBio subreads and Oxford Nanopore "molecule" reads (one read per molecule)
- All PacBio subreads and Oxford Nanopore 2D reads
- PacBio reads filtered to be > 0.75 and All Oxford Nanopore reads (up to 3 reads per molecule)
- PacBio reads filtered to be > 0.75 and Oxford Nanopore "molecule" reads (one read per molecule)
- PacBio reads filtered to be > 0.75 and Oxford Nanopore 2D reads

And where longreads.fastq in the minimap step for RaCon was always either:
- all PacBio subreads for PacBio only assemblies
- all PacBio subreads and all Oxford Nanopore reads (up to 3 per fast5 file) for assemblies using both PacBio and Oxford Nanopore reads.

Miniasm required < 64 GB RAM, 8 threads, and < 3 hours to complete. It was therefore practical to try many combinations of reads, filtering, etc. Since Miniasm is the only assembler without its own consensus step, RaCon was used. RaCon required < 150 GB RAM, 8 threads, and < 4.5 hours to finish. In total, Miniasm+RaCon took < 7.5 hours and is by far the fastest combination of assembly and consensus steps we tested.

DBG2OLC (hybrid approach)

We used the contigs from our "final" platanus assembly, which was assembled with a set of reads that went through one round of contamination-filtering, followed by quality filtering with Trimmomatic, and error-correction with BayesHammer from SPAdes.

./DBG2OLC k 17 KmerCovTh 2 MinOverlap 20 AdaptiveTh 0.002 LD1 0 MinLen 200 Contigs platanus.contigs.fa RemoveChimera 1 f longreads.fasta
cat platanus.contigs.fa longreads.fasta > ctg_pb.fasta
mkdir consensus_dir
split_and_run_pbdagcon.path.sh backbone_raw.fasta DBG2OLC_Consensus_info.txt ctg_pb.fasta consensus_dir > consensus_log.txt

where longreads.fasta was either:
- all PacBio subreads
- PacBio reads filtered to be > 0.75
- All PacBio subreads and Oxford Nanopore "molecule" reads (one read per molecule)
- All PacBio subreads and Oxford Nanopore 2D reads
- PacBio reads filtered to be > 0.75 and Oxford Nanopore 2D reads

Falcon

Assemblies either used:
- all PacBio reads (PBall),
- filtered PacBio reads (PBfilt),
- all PacBio reads and ONT 2D reads (PBall ONT2d),
- all PacBio reads and ONT molecules (one read per molecule) (PBall ONTmol), or
- filtered PacBio reads and ONT 2D reads (PBfilt ONT2d)

All assemblies were initiated via:
fc_run.py fc_run.cfg logging.ini

The logging file for all contained these lines:
[loggers]
keys=root,pypeflow,fc_run

[handlers]
keys=stream,file_pypeflow,file_fc_run,file_all

[formatters]
keys=form01,form02

[logger_root]

```
level=NOTSET
handlers=stream,file_all

[logger_pypeflow]
level=NOTSET
handlers=file_pypeflow
qualname=pypeflow
propagate=1

[logger_pwatcher]
level=NOTSET
handlers=file_pwatcher
qualname=pwatcher
propagate=1

[logger_fc_run]
level=NOTSET
handlers=file_fc_run
qualname=fc_run
propagate=1

[handler_stream]
class=StreamHandler
level=INFO
formatter=form02
args=(sys.stderr,)

    [handler_file_pypeflow]
class=FileHandler
level=DEBUG
formatter=form01
args=('pypeflow.log',)

[handler_file_pwatcher]
class=FileHandler
level=DEBUG
formatter=form01
args=('pwatcher.log',)
```

```
[handler_file_fc_run]
class=FileHandler
level=DEBUG
formatter=form01
args=('fc_run.log',)

[handler_file_all]
class=FileHandler
level=DEBUG
formatter=form01
args=('fc.log',)

[formatter_form01]
format=%(asctime)s - %(name)s:%(lineno)d - %(levelname)s - %(message)s

[formatter_form02]
format=[%(levelname)s] %(message)s
```

Configuration file (fc_run.cfg) for "Falcon Default PBall" and "Falcon Default PBfilt" assemblies:

```
[General]

use_tmpdir = false
job_type = slurm
jobqueue = production
#stop_all_jobs_on_failure = true

# list of files of the initial bas.h5 files
input_fofn = input.fofn
input_type = raw

# The length cutoff used for seed reads used for initial mapping
length_cutoff = -1
genome_size = 292000000

# The length cutoff used for seed reads usef for pre-assembly (¿0, was not able to do -1)
length_cutoff_pr = 500
```

#Pre-Assembly
sge_option_da = –cpus-per-task 8 –mem 30g –time 48:000:00 –qos=ccmb-condo
sge_option_la = –cpus-per-task 2 –mem 30g –time 48:00:00 –qos=ccmb-condo
pa_DBsplit_option = -a -x500 -s200

pa_HPCdaligner_option = -v -B128 -e0.70 -M24 -l1000 -s100
pa_concurrent_jobs = 1000

#consensus for error correction
sge_option_cns = –cpus-per-task 8 –mem 60g –time 48:00:00 –qos=ccmb-condo
falcon_sense_option = –output_multi –min_idt 0.70 –min_cov 4 –max_n_read 200 –n_core 8
cns_concurrent_jobs = 1000

# overlap detection for assembly
sge_option_pda = –cpus-per-task 8 –mem 30g –time 48:00:00 –qos=ccmb-condo
sge_option_pla = –cpus-per-task 2 –mem 30g –time 48:00:00 –qos=ccmb-condo
ovlp_concurrent_jobs = 1000
ovlp_DBsplit_option = -x500 -s200
ovlp_HPCdaligner_option = -v -B128 -e.96 -M16 -l500 -s100

# overlap filtering
overlap_filtering_setting = –max_diff 40 –max_cov 80 –min_cov 2 –n_core 12

sge_option_fc = –cpus-per-task 16 –mem 30g –time 48:00:00 –qos=ccmb-condo

Configuration file (fc_run.cfg) for "Falcon Seed=25 PBall", "Falcon Seed=25 PBfilt", "Falcon Seed=25 PBfilt ONT2d", "Falcon Seed=25 PBall ONT2d", "Falcon Seed=25 PBall ONTmol" assemblies:

# Same as "Default" above, but the following line added:
seed_coverage = 25

Configuration file (fc_run.cfg) for "Falcon Seed=30 PBall" and "Falcon Seed=30 PBfilt" assemblies:

Same as "Default" above, but the following line added:

seed_coverage = 30

Configuration file (fc_run.cfg) for "Falcon Seed=25 Relaxed PBfilt ONT2d", "Falcon Seed=25 Relaxed PBall ONT2d" and "Falcon Seed=25 Relaxed PBall ONTmol":

Same as "Falcon Seed=25 PBfilt ONT2d", "Falcon Seed=25 PBall ONT2d" and "Falcon Seed=25 PBall ONTmol" above, but the following lines added/changed:

# specified a length cut-off
length_cutoff = 7500

# In section titled "consensus for error correction", changed --min_idt from 0.70 to 0.65
falcon_sense_option = –output_multi –min_idt 0.65 –min_cov 4 –max_n_read 200 –n_core 8

# In section titled "overlap detection for assembly", changed -e from .96 to .7
ovlp_HPCdaligner_option = -v -B128 -e.70 -M16 -l500 -s100

Quiver

The basic approach was to align all raw PacBio reads using PBalign, merging and sorting the alignments, and using Quiver:

pbalign $BAX $REF $OUTPRE.cmp.h5 --forQuiver --tmpDir $TMPDIR --nproc $THREADS
cmph5tools.py merge --outFile $MERGEDCMP $CMPDIR/*.cmp.h5
cmph5tools.py sort --deep $IN_CMP --tmpDir $TMP_SORT
samtools faidx $REF
quiver -j$THREADS $INPUT -r $REF -o $OUTGFF -o $OUTFASTQ -o $OUTFASTA --noEvidenceConsensusCall=l
--verbose

Prior to all alignment/polishing steps, for Canu assemblies, tigs in the unassembled file that consisted of at least 2 reads were added back to the assembly, and for both Canu and Falcon assemblies, bubble tigs were added to the assembly. All tigs that were added back to the assemblies were given names to identify them after polishing for optional removal. For all assemblies that had spaces in the contig names, the first "word" before the first space was used as the tig name. All of these operations were accomplished with python scripts found at `https://github.com/JohnUrban/sciara-project-tools`: filterCanuFasta.py and fasta_name_changer.py. This pipeline was automated using shell scripts also found in that repository.

Pilon

The basic approach was to build a bowtie2 index for each assembly, map the paired-end illumina reads to each assembly with bowtie2, index the BAM alignment files with SAMtools, mark duplicates with Picard Tools, and use Pilon to polish.

bowtie2-build $ASM $BASE

bowtie2 -p $P --very-sensitive -N 1 --minins 0 --maxins 1000 -x $BT2 -1 $R1 -2 $R2 | samtools sort --threads $P -o $PRE.bam

samtools index ${PRE}.bam

java -Xmx${JX} -jar $JAR MarkDuplicates INPUT=${PRE}.bam OUTPUT=${PRE}.markdup.bam METRICS_FILE=${PRE}.metrics.txt REMOVE_DUPLICATES=false ASSUME_SORTED=true

samtools index ${PRE}.markdup.bam

java -Xmx${JX} -jar $PILONJAR --genome $ASM --output $PRE --changes --frags ${READS} --diploid --fix bases --nostrays

This pipeline was automated using shell scripts also found at `https://github.com/JohnUrban/sciara-project-tools`.

BUSCO, LAP, ALE, REAPR, FRC^{bam}

Same as for Illumina assemblies.
Percent Illumina reads mapped taken from Bowtie2 output.

Size statistics

Used asm-stats.py from `https://github.com/JohnUrban/sciara-project-tools`

Aligning BioNano Maps with Maligner

Merge all RawMolecule BNX files (all.bnx) using RefAligner utility from BioNano Genomics
./RefAligner -bnx -merge -i scia_copr_2013_031_1P_2016-04-13_15_45/Detect Molecules/RawMolecules.bnx -i Scia_copr_2013_032_1P_2016-04-12_10_39/Detect Molecules/RawMolecules.bnx -i Scia_copr_2013_032_1P_2016-04-12_15_55/Detect Molecules/RawMolecules.bnx -i Scia_copr_2013_032_1P_2016-04-13_11_27/Detect

Molecules/RawMolecules.bnx -i Scia_copr_2013_032_1P_2016-04-14_11_49/Detect Molecules/RawMolecules.bnx -o all -minSNR 2.75 -minlen 150 -minsites 8 -MaxIntensity 0.6

Convert BNX to Maligner input
bnx2maligner.py -b all.bnx ¿ bionano.maps

smooth the maps for maligner alignment using utilities from Maligner
smooth_maps_file -m 1000 bionano.maps > bionano.smoothed.maps

Split up maps to align them to assemblies in parallel (split into 48 files)
mkdir split
cd split
split -l 4079 -d ../bionano.smoothed.maps bionano.smoothed.maps

Convert assembly to smoothed map for Maligner using Maligner utilities (BssSI = CACGAG)
make_insilico_map -o $ASM_OUT_PFX $ASM_FASTA CACGAG smooth_maps_file -m 1000 ${ASM_OUT_PFX}.map ¿ ${ASM_OUT_PFX}.smoothed.maps

BWA and Sniffles

BWA was used to index assemblies and align all PacBio subreads and all ONT reads (up to 3 reads per fast5 file). Sniffles was used to detect structural variants (SVs) using either the PacBio reads, ONT reads, or both combined.

bwa index $ASM -p $BASE

bwa mem -t $MTHREADS -M -x $TYPE $BWAIDX $FASTQ | samtools sort -T $TYPE --threads $MTHREADS -o $TYPE.bam
where "$TYPE" was "pacbio" or "ont2d"

samtools merge –threads $P combined.bam $PBBAM $ONTBAM

sniffles -m $BAM -b $BEDPE.bedpe
where $BAM was either PacBio alignments, ONT alignments, or both combined.

Percent mapped and ratio of number of alignments to number of uniquely aligning reads also obtained from BWA output. This pipeline was automated using shell scripts also found at `https://github.com/JohnUrban/sciara-project-tools`.

MarginStats for percent identity of Nanopore Reads

The BAM file containing the alignments of all ONT reads mapped to the highest ranked PacBio-only assembly was filtered to remove unmapped reads using SAMtools. Then the resulting BAM files was partitioned into 400 smaller files using splitSAM.py from `https://github.com/JohnUrban/sciara-project-tools`. This enabled us to compute percent identities in parallel in the next step. MarginStats was then used to calculate the end-to-end percent identities of the reads in each file. Since reads were extracted using our set of tools for working with fast5 files from ONT called fast5tools (`https://github.com/JohnUrban/fast5tools`), read names contained mean quality scores, which could then be paired with the percent identities from MarginStats.:

samtools view -bh -F4 canu10.minrl500.dip3x.pball.quiverfinal1.pilon2x/ont2d.bam > reads/ont.bam
splitSAM.py –bam reads/ont.bam –nfiles 100 –nreads 1382636

On each BAM:
marginStats –noStats –printValuePerReadAlignment –identity $BAM $fq $ref > wd/${pre}.txt

On each MarginStats output text file:
awk '{sub(/\ /,"\n"); gsub(/\t/,"\n"); print}' wd/$PRE.txt | grep -v ^ValuesIdentity | paste -
<(grep -v ^@ <(samtools view splitfiles/${PRE}.bam)) | awk 'OFS="\t" {print $2,$1}' > pairs/$PRE.txt

Other analyses including plotting and visualization

All plotting and visualization done in R.

DNA modifications with PacBio data

The assembly was split up into individual contigs such that each contig could be processed separately in parallel. This also helped get through a bug in the code that caused ipdSummary to get stuck at the end of some contigs, preventing all others from being processed.

Kinetics Tools (https://github.com/PacificBiosciences/kineticsTools) was used to detect m6A, m4C, and m5C_TET - for each contig:
ipdSummary $INPUT –reference $REF –identify m6A,m4C,m5C_TET –methylFraction
–gff basemods.gff –csv kinetics.csv –pvalue $PVALUE –minCoverage 3
–methylMinCov 10 –identifyMinCov 5 -j $THREADS –maxAlignments 1000000
–ms_csv multisite.csv –bigwig ipd.bigWig –refContigs $contigName

AgIn (https://github.com/hacone/AgIn) was also used to detect CpG methylation - for each contig:

B=P5C3_HdrR
g=-0.88
L=35 #L=50 and L=35 were both performed
/path/to/AgIn/target/dist/bin/launch -i kinetics.csv -f $REF -o AgIn.$tig.$B.$g.$L -b $B -g -0.88 -l $L -c predict

DNA modifications with MinION data

Nanopolish was used to find 6mers in the MAP006 data that had different ionic current means than expected given the ONT model.:
G=assembly.fasta R=reads.fasta B=reads.bam samtools faidx $G
bwa index $G
for lib in lib14 lib15 lib16 lib20 lib21; do
    nanopolish extract $lib/pass/ -o $lib.pass.fa -t 2d
    nanopolish extract $lib/fail/ -o $lib.fail.fa -t 2d
done

cat lib*pass.fa > $R # Or cat lib*pass.fa lib*fail.fa > $R

bwa mem -M -x ont2d $G $R -t $T — samtools sort -T $lib.pass –threads $T — samtools view -F 4 -q 1 -bSh - > $B

samtools index $B

nanopolish methyltrain –reads $R –bam $B –genome $G -t $T –models-fofn=ont.alphabet_nucleotide.R7.fofn –train-kmers all –rounds=5 –progress

This was also performed on combined pass and fail 2D reads.

B.2 Supplementary Figures

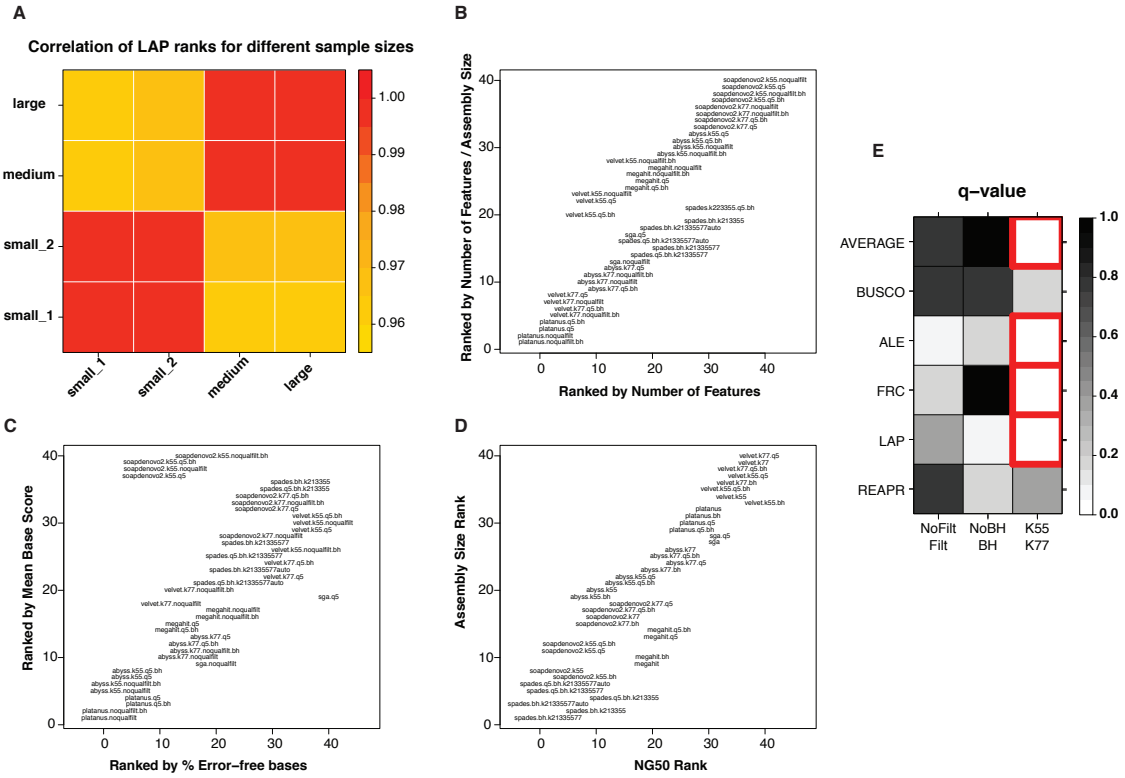


Figure B.1: Supplementary information on evaluating short read assemblies.
(A) LAP rank correlations for different sample sizes of paired-end reads. Two small samples of approximately 15,000 reads, one medium sample of approx. 150,000 reads, and a large sample of 1.5 million. **(B)** Scatter plot of rankings from two different FRC^{bam} scores: number of features and number of features normalized to assembly length. The correlation was 0.88. **(C)** Scatter plot of rankings from two different REAPR scores: % error-free bases and mean base score. The correlation was 0.62. **(D)** Scatter of NG50 ranks and assembly size ranks. **(E)** We tested whether various conditions made a difference in assembly rankings using the Wilcoxon signed-rank test. We looked at the effect of quality filtering reads (NoFilt vs Filt), the effect of error-correcting reads with BayesHammer [Nikolenko et al., 2013] (NoBH vs BH), and the effect of Kmer size (K55 vs K77). The matrix shows Benjamini-Hochberg-adjusted p-values (referred to as q-values) for controlling the False Discovery Rate [Benjamini and Hochberg, 1995] for these tests. The only parameter that made a difference in rankings was value of Kmer size. ALE, FRC, LAP and Mean ranks had $q < 0.05$ (highlighted in red).

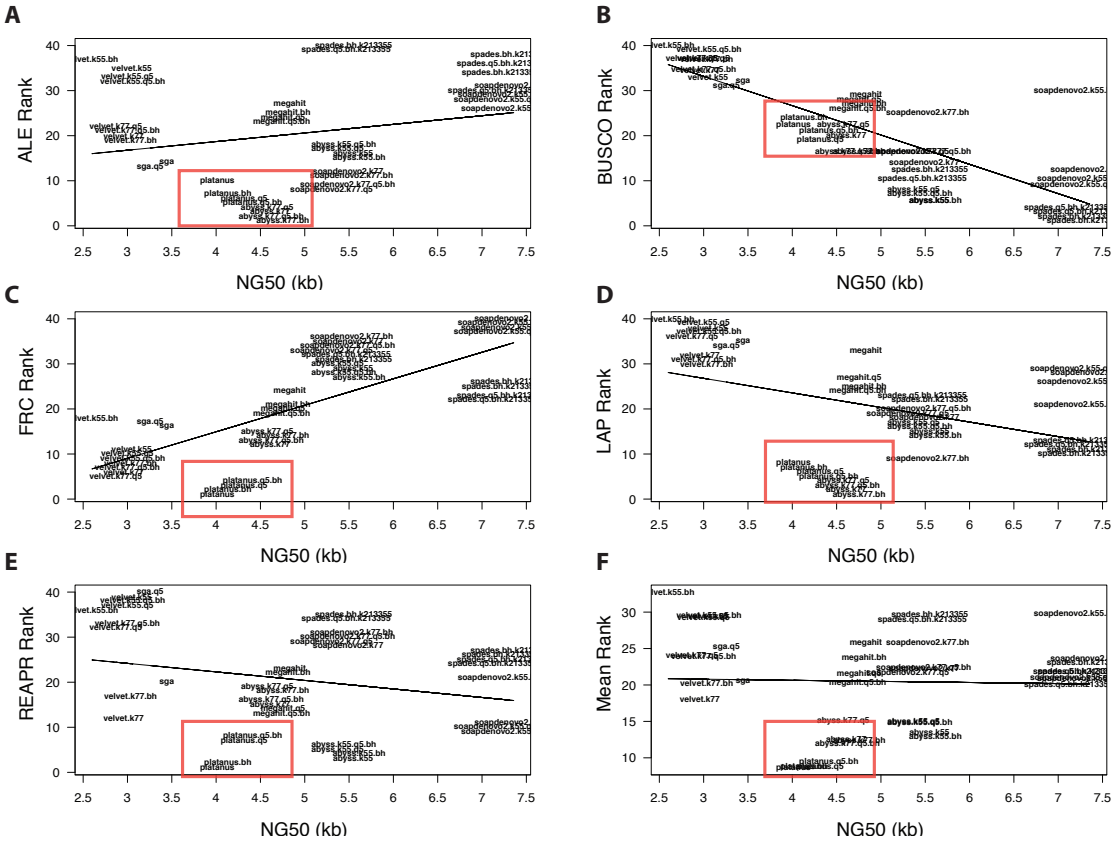


Figure B.2: Platanus does better than NG50 would predict.

(A) ALE ranks vs NG50. **(B)** BUSCO ranks vs NG50. **(C)** FRC^{bam} ranks vs. NG50. **(D)** LAP ranks vs NG50. **(E)** REAPR ranks vs. NG50. **(F)** Mean Ranks vs NG50 (no correlation). For all, Platanus clusters are outlined in red. Abyss K=77 often appears close by to Platanus. While NG50 tends be a fairly good predictor of the ranks of many assemblies (particularly for BUSCO, FRC, and LAP ranks), it fails to be a good predictor of all Platanus rankings except BUSCO. Overall, NG50 typically predicts Platanus to be moderately ranked whereas Platanus is usually very close rank=1. Plotting NG50 vs the actual metric scores yields the same conclusions.

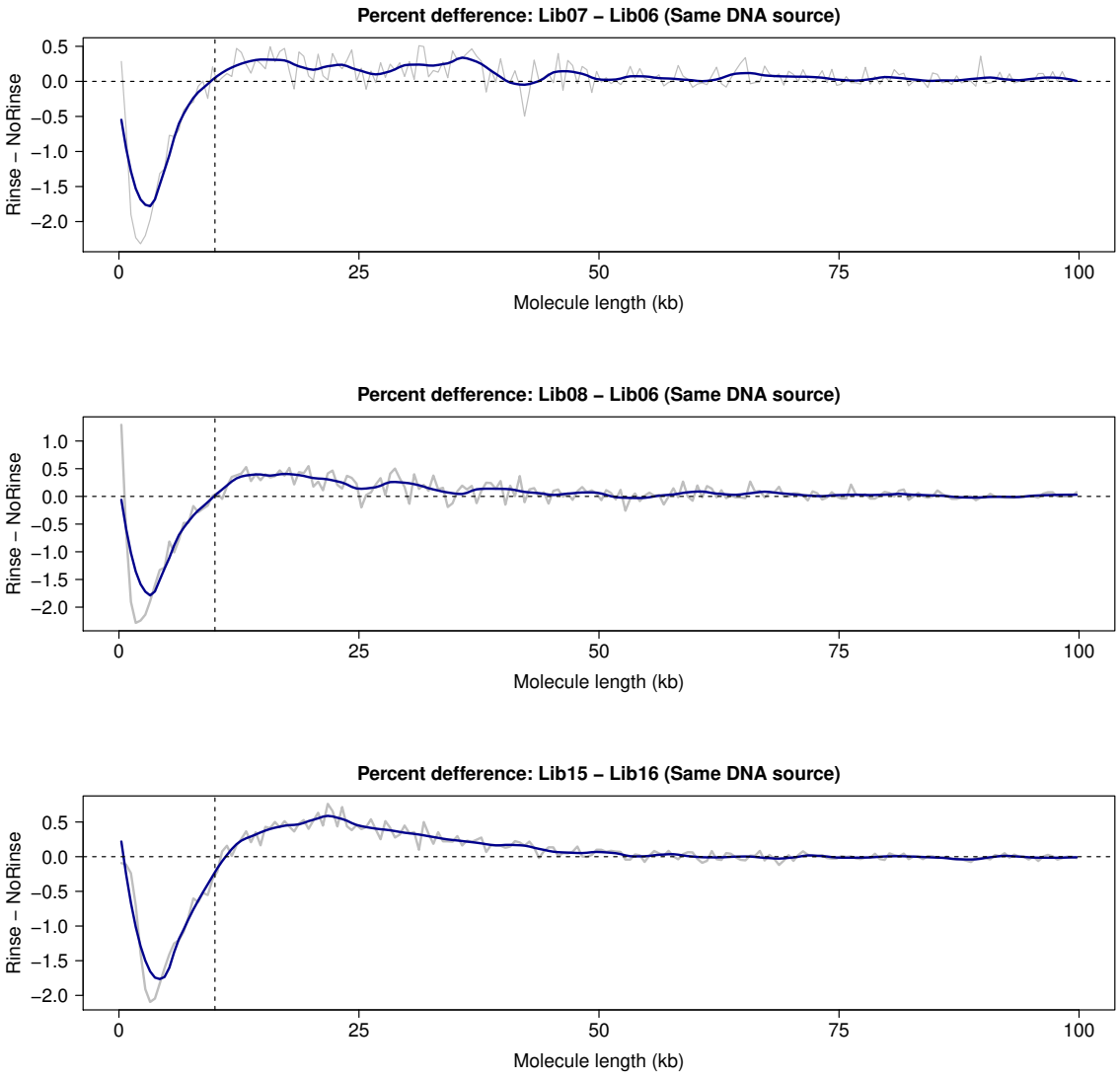


Figure B.3: Reproducibility of using rinses in AMPure steps to deplete DNA smaller than 10-12 kb.

Libraries 06, 07, and 08 were prepared from the same source. Library 6 did not have rinse steps where as libraries 07 and 08 did. Libraries 06 and 07 are Runs B and C from our preprint [Urban et al., 2015a]. These were all from MAP004 reagents and the original MinION. Libraries 15 and 16 are from MAP006 reagents and the MinION MkI, and are from the same starting source of DNA. Library 15 included rinses whereas library 16 did not. The 3 rinse libraries are depleted of DNA $<$ 10-12 kb compared to their respective no rinse counterparts. The grey lines in the background shows the values from 500 bp bins. The thicker dark blue line is a lightly Loess smoothed version.

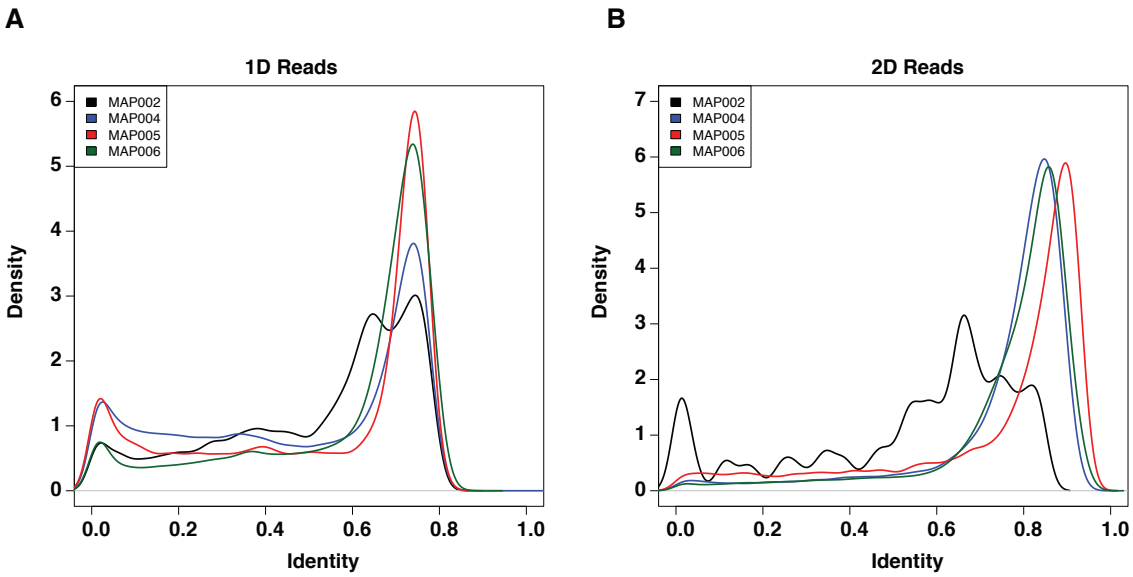


Figure B.4: Percent identity densities of nanopore reads from different MAP kits. **(A)** 1D reads. **(B)** 2D reads. MAP005 seems to have had the highest percent identity for 2D reads. Otherwise, MAP004-006 are comparable. MAP002 performed more poorly than the others. Though this is in part due to it being generated by an earlier kit and base-caller, it is also likely because it is represented by only 1 library with much fewer reads than in the other groups.

B.3 Supplementary Tables

A

Sex	Stage/tissue	Cell type	Mean (pg)	Stderr (pg)	N	Bp length* (Mb)
F	Larvae_4th_instar	Hemocytes	0.571	0.004	65	558.438
M	Larvae_4th_instar	Hemocytes	0.523	0.007	77	511.494
F	Pharate_pupa	Hemocytes	0.591	0.004	74	577.998
M	Pharate_pupa	Hemocytes	0.540	0.003	38	528.12
F	Adult	Hemocytes	0.554	0.003	74	541.812
F	Adult	Hemocytes	0.557	0.004	62	544.746
M	Adult	Hemocytes	0.503	0.002	100	491.934
M	Adult_44A_testis	Sperm	0.491	0.016	162	480.198
M	Adult_44A_duct	Sperm	0.494	0.017	60	483.132
M	Adult_43E_testis	Sperm	0.519	0.018	100	507.582
M	Adult_43E_duct	Sperm	0.521	0.013	50	509.538
M	Adult_45_testis	Sperm	0.535	0.014	65	523.23
M	Adult_45_duct	Sperm	0.484	0.019	50	473.352

B

Stage	F (pg)	M (pg)	F-M (pg)	bp length (Mb)
Larvae	0.571	0.523	0.048	46.944
Pupae	0.591	0.540	0.051	49.878
Adult*	0.555	0.503	0.052	50.856
Mean	0.572	0.522	0.05	**48.9**

C

Stage	Sex	Total (pg)	X (pg)	Total-X (pg)	bp length of diploid autosomes (Mb)	Bp length of haploid autosomes (Mb)	Bp length f haploid autosomes + X (Mb)
Larvae	F	0.571	0.05	0.521	509.538	254.769	303.669
Larvae	M	0.523	0.05	0.473	462.594	231.297	280.197
Pupae	F	0.591	0.05	0.541	529.098	264.549	313.449
Pupae	M	0.540	0.05	0.49	479.22	239.61	288.51
Adult	F	0.555	0.05	0.505	493.89	246.945	295.845
Adult	M	0.503	0.05	0.453	443.034	221.517	270.417
Mean	-	-	-	**0.497167**	**486.229**	**243.115**	**292.015**

D

DNA content in sperm	DNA length in sperm (Mb)	Haploid autosome (Mb)	Single X (Mb)	2X+A (Mb)	Total L length (sperm – 2X – A) (Mb)
0.505	494.235412731	243.115	48.9	340.915	153.32

E

L length given cn=1 (Mb)	L length given cn=2 (Mb)	L length given cn=3 (Mb)	L length given cn=4 (Mb)	Haploid germ line length, cn = 1 (Mb)	Haploid germ line length, cn = 2 (Mb)	Haploid germ line length, cn = 3 (Mb)	Haploid germ line length, cn = 4 (Mb)
153.32	**76.6602**	51.1068	38.3301	445.335	**368.675**	343.122	330.345

Table B.1: Calculating genome size from nuclear DNA content measurements. We wanted to estimate the approximate length of the haploid genome in the germ-line (ChrII, ChrIII, ChrIV, ChrX, ChrL) as well as the somatic haplod genome length (ChrII, ChrIII, ChrIV, ChrX). To do so, we used started with the values in Table 2 from Ellen Rasch [Rasch, 2006], adapted here in **(A)**. *The Mb length of all DNA in the nucleus given pg mean and Dolezel conversion [Dolezel et al., 2003]. **(B)** Since male somatic cells have a single X and female somatic cells have two, the pg weight and bp length of a single chromosome X can be inferred by subtracting the male value from the female value. *Since two female adult values are in previous table (A), the female adult value was obtained by: $(74 * 0.554 + 62 * 0.557)/(74 + 62) = 0.555$. **(C)** The diploid autosome DNA content and bp length can be inferred by subtracting the value of 1X for males or 2X for females, which in turn allows inferring the haploid autosome DNA content and length by dividing by 2, and in turn allows one to add the value of a single X for the expected somatic haploid genome (chrII, chrIII, chrIV, chrX) DNA content and length. Thus, for this study, our expected genome and assembly sizes should be around 292 Mb. **(D)** The expected length of the full germ-line genome can now be deduced by adding the weight of a single L chromosome. The data from sperm in the above table, can give us an idea of how much DNA content (and length) is contributed by L chromosome DNA. The sperm nucleus has 2 copies of chromosome X, but a single copy of each autosome. So the L content can be found by subtracting 2X+A from the total. We took the average DNA content in sperm via: $(162 * 0.491 + 60 * 0.494 + 100 * 0.519 + 50 * 0.521 + 65 * 0.535 + 50 * 0.484)/(162 + 60 + 100 + 50 + 65 + 50) = 0.505353182752$. **(E)** However, as discussed in Rasch (2006) and demonstrated by Rieffel and Crouse [Rieffel and Crouse, 1966], there can be between 0 and 4 copies of the L chromosome in a sperm nucleus, though the majority have 2 copies. This is reflected in Rasch (2006) data where 78% of sperm have 2 copies of the L, 20% have 1 or 3 copies, and 1% have 0 or 4 copies. Thus, L size and haploid germ-line genome size can be deduced by assuming an L chromosome copy number of 1-4, but we ultimately expect using a copy number of 2 gives the closest approximation. The predominance of 2 copies of chrL indicate that the haploid genome (ChrII, ChrIII, ChrIV, ChrX, ChrL) is likely 369 Mb, whereas the somatic haploid genome (ChrII, ChrIII, ChrIV, ChrX) is approximately 292 Mb. "cn" indicates copy number.

Supplementary Information: The DNA puffs of Sciara coprophila before, during, and after developmentally programmed intrachromosomal DNA amplification

C.1 Supplementary Methods

C.1.1 qPCR primer validation

C.1.2 Fluorescence in situ hybridization (FISH)

This protocol is adapted from "A simplified and efficient protocol for nonradioactive in situ hybridization to polytene chromosomes with a DIG-labeled DNA probe," by E.R. Schmidt.

Preparation of Slides and Coverslips

1. Sonicate out-of-the-box Corning microscope slides in Milli-Q ultrapure water and powdered labware detergent for 60 minutes (two rounds of 30 minutes). The sonicator can fit six slide beds (10 slides) at a time.
2. Rinse slides for at least 30 minutes under running hot tap water (rinse in autoclave overflow container propped up over drain with petri dish cover).
3. Rinse slides for at least 30 minutes under running deionized water.
4. Dunk slides in a staining dish of Milli-Q ultrapure water for final rinse. When dunking slides in staining dishes, refresh staining dish solution every three slide beds.
5. Dunk slides in a staining dish with 95% ethanol twice (two separate staining dishes).
6. Allow slides to air dry in the slide beds overnight.
7. Dunk slides in staining dishes with 1:10 dilution (a 1:1 mix of 1:10 dilutions in PBS and water) of poly-L-lysine (0.1% w/v stock) for around 15 minutes (up to an hour). Poly-lysine dilutions should be kept refrigerated and only the amount necessary should be poured out at a time. Warm the solution (30 seconds in the microwave) before use.

8. Allow slides to air dry in the slide beds covered (with tinfoil, etc.) to prevent dust from sticking. Drying may take as long as overnight.
9. Soak 22x22mm coverslips in RainX for at least 2 minutes.
10. Allow coverslips to air dry on wooden rack.

Chromosome Squash

1. In Roberts buffer on a siliconized microscope slide, dissect salivary glands from Sciara larva that are at least in the early eyespot stage under the dissecting microscope.
2. Place a siliconized coverslip on top of the siliconized dissection slide next to the salivary glands and add 40μl of 45% acetic acid on top of the coverslip.
3. Carefully transfer the salivary glands from the dissection site to the drop of acetic acid on the coverslip. Clear debris from the salivary glands as necessary with a kimwipe.
4. Let the salivary glands sit in the acetic acid for 3 minutes.
5. Place a prepared poly-lysine microscope slide on top of the salivary glands such that the two slides are parallel and only intersect over where the prep is. The coverslip should stick to the prepared slide when picking it up and flipping the slide over such that the coverslip is again facing upwards.
6. Blot excess liquid that escapes from under the coverslip with a kimwipe.
7. With a pencil eraser, gently nudge (not really pressing down, but should feel very slight resistance) the coverslip to move it around the slide. The goal of this movement is to move the chromosomes out of the nuclear envelope and outside the cellular membrane. Check under the phase microscope to see if the chromosomes have successfully escaped.
8. Tap the eraser lightly over the prep to squash. The first round of tapping should gentle, trying to separate chromosomes but not stretching them out. The second round can tap more vigorously (still not moving the coverslip at all). Periodically check the progress of the squash using the phase microscope. Tapping in a spiral in from the edge can help chromosomes move back towards the middle. Also can tap in specific locations identified through the microscope that particularly need more spreading.
8. When the chromosome have been satisfactorily spread (not balled up but not in pieces, banding pattern is viewable, etc.), place the prep in a folder of absorbent paper with the coverslip in the middle. Place a wooden block on top of the folder positioned over the slide and use the hand clamp on the block and the table to squash the prep.
9. To make permanent preps, submerge the prep-site on the slide in liquid nitrogen and hold a little past it stops bubbling.
10. Use a diamond pen to score the corners of the coverslip on the prep slide. Then use a razor blade to flip off the coverslip and dispose of it.
11. Put the slide with prep in 95% ethanol in a staining dish for 10 minutes and allow the slide to air dry afterwards (note: this ethanol solution should be made fresh every day).

12. Use the diamond pen to label the slide for your personal notes and also on the back of the slide, draw a circle around where the prep is. Store the slide in a covered slide box.
13. In notebook, note down characteristics of the prep including general eyespot stage, presence of puffs, quality of prep, etc.

Labeling Reaction

1. Add 1μg (maximum) template DNA to 15μL of sterile, double distilled water into a screw-cap tube. The final total volume needs to be 16μL, so adjust the volume of water added according to the amount of DNA.
2. Denature the DNA by putting the tube in a foam holder and let sit in a boiling water bath for 10-15 minutes.
3. Spin down the tube briefly at 18,000g (press stop when the centrifuge reaches maximum speed).
4. Place the tube on ice for 10 minutes to maintain denatured state.
5. Mix Fluorescein High-Prime (contains Klenow polymerase) and add 4μL to the denatured DNA tube. 0.5μL of additional Klenow could be added to ensure progression of reaction. Mix gently by pipetting up and down, but avoid air bubbles and shaking. Spin down to collect the solution.
6. Wrap the reaction in aluminum foil and incubate at 37°C for overnight or at least 1 hour.
7. Stop reaction by adding 2μL 0.2M EDTA, pH 8.0 to tube or heating the tube at 65°C or more for 10 minutes.
8. Store probe at 4°C.

Pre-hybridization Treatment and Denaturation

Slide Preparation

1. Preheat the slides (dry in a slide bed) overnight in the oven (60-65°C). Let cool a little at room temperature before use.
2. Place the slides in a staining dish through a hydration series of 70%, 50%, and 30% ethanol, all for 2 minutes minimum (pour out solution and reuse staining dish).
3. Incubate the slides in 0.1x SSC for 2 minutes.
4. Incubate the slides in 2x SSC for 5-10 minutes, twice.
5. Quickly rinse slides in 0.1M NaOH, then pour that solution out and replace for a 90 second wash, mixing gently.
6. Quickly rinse slides in 2x SSC, then replace solution for a 10 minute wash on the shake table.
7. Wash the slides through a dehydration series of 0.1x SSC, 30%, 50%, 70%, and 90% ethanol, all for 2 minutes minimum. Wash the slides in 90% ethanol twice total.

8. Leave slides in slide bed to air dry for 30 minutes at room temperature (can be moved to oven to facilitate drying).

Hybridization Buffer

1. In a screw-cap tube mix 54μl sterile distilled water to 20μl of labeled probe (kept in foil) and mix by flicking.
2. Denature the contents of the tube by putting it in boiling water in a foam float for 10 minutes.
3. Put the tube on ice to maintain denatured state.
4. Add 25μl of 20x SSC to the tube.
5. Add 1μl 10% SDS to the tube.
6. Measure the volume of the tube with a pipette to verify if it is 100μl . If it is not, add enough sterile distilled water to increase the volume to 100μl.
7. The buffer may be stored at -20°C for long periods of time, but must be boiled for 10 minutes and put on ice right before use.

Hybridization

1. Boil water then denature the hybridization solution in it for 10 minutes. Immediately put on ice after to stop the denaturation, then centrifuge quickly to collect the solution.
2. Pipette 11μl of the hybridization solution over where the chromosomes are located on the slide. Use a 22x22mm siliconized coverslip to spread the solution (polish with kimwipe before use).
3. Use in situ hybridization glue (SciGene cytobond) to seal the edges of the coverslip. Should use a generous amount of glue (dont be reserved).
4. Place slides in a humid chamber (water) and let sit at room temperature for 30 minutes.
5. Wrap the humid chamber in saran wrap (wrap horizontally and vertically in two separate pieces) and let incubate at 60°C overnight.

Signal Amplification – Antibody Application

Note: slides and fluorescent reactants should be covered with foil at all times.
1. Use forceps to remove the glue from the coverslips (should easily peel off).
2. Put the slides in a 2x SSC rinse to facilitate the removal of the coverslip.
3. Replace the solution and let the slides wash for 10 minutes on the shake table.
4. Rinse the slides in PBST then replace the solution and let the slides wash for 5 minutes on the

shake table.
5. Keep slides in PBST and take them out one by one as they are processed. Wipe the slides and shake (flick with wrist) dry. Then add 50μl of the primary antibody (which depends on which kit, fluorescein is 1/1000 Alea 488-conjugated rabbit anti-fluorescein Ab in PBST with 1mg/mL BSA) onto the chromosome prep. Cover with a 24x50mm coverslip and tack down the far side with a spot of glue. 6. Let slides incubate in humid chamber wrapped in saran wrap and foil for 2-3 hours at 37°C.
7. Remove the glue tack and use a PBST rinse to take off the coverslips. Replace the solution and wash for 4 minutes twice.
8. Keep slides in PBST and take them out one by one as they are processed. Wipe the slides and shake dry. Then add 50μl of the secondary antibody (fluorescein kit: 1/1000 Alexa 488-conjugated goat anti-rabbit Ab in PBST). Cover with a 24x50mm coverslip and tack down the far side with a spot of glue.
9. Let slides incubate in a humid chamber wrapped in foil for a minimum of 1 hour at 37°C.
10. Complete the same two washes in step 7.
11. Rinse in PBS then replace the solution and wash slides 3 times for 5 minutes. Wash one last time in PBS for 15 minutes.
12. Wipe and shake dry the slides. Then add 20μl of Vectorshield mounting media with DAPI on to the prep. Cover with a 22x22mm coverslip while avoiding air bubbles. Let media set overnight at 4°C. Tack the corners with nail polish after allowing slides to warm up to room temperature for 10 minutes. Wait 5 minutes before viewing.

Poly-L-Lysine Coating of Slides in PBS

1. Add 30mL poly-L-Lys solution (Sigma #P8920) and 30mL tissue culture PBS (pH range 7.2-7.6) to 250mL water to prepare 300mL of poly-L-Lys working solution. This should be done in a plastic beaker and graduated cylinder since poly-Lys sticks to glass.
2. Transfer super clean slides to poly-L-Lys solution and shake at least 15 minutes (can be up to an hour). Shaking in a plastic tray is best.
3. Transfer racks to new chambers filled with Milli-Q water. Plunge up and down 5 times to rinse.
4. Air dry at room temperature for overnight. Cover the slides with foil to prevent dust.
5. The poly-L-Lys solution can be stored at 4°C after sterilization, or -20°C without sterilization at least for at least 4 months.

SSC (Standard Saline Citrate)

Make large batches of 20x SSC to dilute into 2x SSC working solutions.
20x SSC: 175.32g NaCl (3.0M) and 88.23g sodium citrate (0.3M) in 1L distilled water, pH adjusted to 7.4.
20x SSC to 2x SSC is a ten-fold dilution (e.g. 100mL 20x SSC with 900mL distilled water).
1x SSC is 0.15M NaCl and 0.015M sodium citrate at pH 7.0.

PBS (Phosphate-buffered Saline)

Make large batches of 10x PBS to dilute into 1x PBS working solutions.
10x PBS: 80g NaCl, 2g KCl, 14.4g Na2HPO4, and 2.4g KH2PO4 in 1L distilled water, pH adjusted to 7.4
10x PBS to 1x PBS is a ten-fold dilution (e.g. 50 mL 10x PBS with 450mL sterile water).
1x PBS is 137mM NaCl, 2.7mM KCl, 10mM Na2HPO4, and 1.8mM KH2PO4 at pH 7.4.

PBST (Phosphate-buffered Saline + Tween)

Prepare working solution (1x) as needed.
1x PBST: 100mL 10x PBS, 899mL distilled water, and 1mL of Tween 20 (100%)

C.1.3 FISH probes

The following sequences were ordered for synthesis from Integrated DNA Technologies (IDT).

>Amplicon4
GTCGTAGACCTCTGCTACTTTATATTGACTCAATGAGTTTACCAGACTTTTTTATTTT
GCGTTTTCATTATTTATAAATCCAAAAATAAGAATCGACAAATGCGTCAAAACAAATT
GTTATTTTGTTTTGAAGTAGAATGTTTAAGTTCAGCTGGTGAATAACTATGCCTTGAT
GTTGTACGAAAAAAACCATATTATTAAGTACGTCACCAATACTCTCTCTAGCAAACAA
ACTCTCAGCTCTATGTGTCAAATTCACACCTTATGTCTTTGTATTCTACAAACACTCTT
CCACATTTTGAATGACTACGCTGTAATTTACATGTAAATTCCAAAGGCTACTTGAAAT
TTTCACCACCTTAATAAAATTATTTTGTTTGCTGGAGAGAGTATTGACGTCACACTTA
AACATTTTTTTTAAAACTTAACCCTTAGCCCTTAAGACGAATTAAAATCTCAATTCCA
ACGATACCAACCAAAAAAAAAACTAAAATGTCACTCATTTGTTTGTTGATTTTTCCAA
TTTTCAAAAATCATAAAATATAATTCTGCTGGATGAGTAGCAATCCGATTCCCTACTA

TGAAAAGAAAGAGCTAGAACCAAGCCGAGATCATTCGGAACCAGCAAAAAAGCGAGT
GTAACTTGAATATTCAGAATTTGTACTTAATCAGTCTGAATGGTACGACTTCGGACCC
GAGGTGCAAAACAATGAATTAGAAAGCAATCTCGATCAGTCTCAATCTGGCTGGTATG
ACTTTGGACCGGACGAACTAGATTTTCTGGAACCATTAATGAAAAACAATGACTTCTT
GATTCCCATTGAGCAAAGCGATATTGTTAATTTCGAACTAGATCCGGAACCCTCTGAC
CATAATCAGAGCATGCAAGAAACTGAAGACTTTCCATCGTAATATCTACTTCAACTTT
TACCTTTTCATTGTTTATTCCTATGCATCCTAGGCCTAAAATGCGTGAAGAAATTCAA
GAATTGAAAGCTCA

>Amplicon5
TTACCTTGGTTTCTGGTGGGTTTTGATTTTTGTTTTTTAAACGTTGCCACACGTTATT
GACGATTTCAAGATATTTTAATTGAGAGGTTACGCTGATCGCTTGATCGCTGGCAAAA
GTTAAAGACCTTATACATGTGTTTCACATTGCAATGTGGGTGTAGTGGACGGTAAACG
AGGCACACACATACCAGTAAAGTCACGTCTCATATCAAACATATCAACGTAATTGCAA
TCAGTAGTTTTAGATTAGTCTACCACCAGCACTCGATCAAACAACATAAGAATAAATA
AACGTAGCCTAAAGTGCACCGACTCGTCTTATTATTTATAATAACTAATATGGGCAGC
ATAGAGAGACTGAATTTTAAGACCCGTGCTTCCATCTACATACAAAACACATACCGAA
CGAGCCAATCTTAATTTTGTTTATTTATTTGTGTTGCGCGCTCTTCGCCAGCGTGTAG
TACTAGATCATTTGAATGTGCGATATGACAGCACTGTATACAGTTGACATTTAACTCT
GTCAGATGTGCGCGATAATATTATACCAAAAGTATTGTTTGTAAACTCGAGTTTGTTT
TATTTTGGGTTTGGATTCTAAACCACCTAACAGACGAGAAAAAAAAACGTTTACGAATC
GTGTCCTGTTCAAACTACACAACAACATCATGGATCTACTAGGCTCAATTCTTAATTC
AATGGATAAACCACCCGAAGCGAATCAGAAGCAAAAGGAAATCATAAAAAGTTTGTCA
CCGAAAATTTTCACATTTTTTTTTCCTTTCGTCTATTGTCGCTTCTCTTTCGTGATTTT
GTTGTAGTTGAACGACGGAATGTTTTGAAGATAAAAATGATTTTTTTTTTTGAACTTAC
AGAACAGAACGAACACATTGAAAAGATGAGAAATCGTGAAAAAGAGGAGCTAAATCG
CTTCCGGCAGTTCGTTGAAGAGCGAATCGATCGAATATCCAAAGATGACAGCCGCAAA
TTCATTCAATTTCA

>Amplicon5-B
TTTTTTAAACGTTGCCACACGTTATTGACGATTTCAAGATATTTTAATTGAGAGGTTA
CGCTGATCGCTTGATCGCTGGCAAAAGTTAAAGACCTTATACATGTGTTTCACATTGC
AATGTGGGTGTAGTGGACGGTAAACGAGGCACACACATACCAGTAAAGTCACGTCTC
ATATCAAACATATCAACGTAATTGCAATCAGTAGTTTTAGATTAGTCTACCACCAGCA
CTCGATCAAACAACATAAGAATAAATAAACGTAGCCTAAAGTGCACCGACTCGTCTTA
TTATTTATAATAACTAATATGGGCAGCATAGAGAGACTGAATTTTAAGACCCGTGCTT
CCATCTACATACAAAACACATACCGAACGAGCCAATCTTAATTTTGTTTATTTATTTG
TGTTGCGCGCTCTTCGCCAGCGTGTAGTACTAGATCATTTGAATGTGCGATATGACAG
CACTGTATACAGTTGACATTTAACTCTGTCAGATGTGACACAACAACATCATGGATCT

ACTAGGCTCAATTCTTAATTCAATGGATAAACCACCCGAAGCGAATCAGAAGCAAAAG
GAAATCATAAAAAGTTTGTCACCGAAAATTTTCACATTTTTTTTTCCTTTCGTCTATT
GTCGCTTCTCTTTCGTGATTTTGTTGTAGTTGAACGACGGAATGTTTTGAAGATAAAA
ATGATTTTTTTTTTGAACTTACAGAACAGAACGAACACATTGAAAAGATGAGAAATCG
TGAAAAAGAGGAGCTAAATCGCTTCCGGCAGTTCGTTGAAGAGCGAATCGATCGAAT
ATCCAAAGATGACAGCCGCAAATTCATTCAATTTCAACCATTGGACAAGGTCTATCGG
AGTGTTGTGTGAGTAGCCTATCTATCTGTTATTGTTTTATCCTGACAGACCTCCGTTC
TCTTTTGAGAACTAGGTCTTCGTTGATCTTTATCGAATCTCACGAAGTTTAAGAATCG
AAGTCCTTAATTGTTA

>Amplicon6
TTCTTTCATTACTTTTATGGAACTGCTCTGATTACTCTGGGTAGAGTACAGATATCGG
TTCTAGTTTTCTTCAAAGCAGTTTTATAATTGAATGGTTATTCTAATACCACGTCAAA
TGGGGCTCGTAGTCAAAGCAGGTTCGTTCTTTTTGACCCTATGGCGGTTTCACCACTC
CAGCAACGAGCTGGAGCAGTCCTATTTCAATCCATCAATTCACCTAACTCCGAGCATA
TTCCACTTGAATATCACGAGATATTCCCGGCAATTTCACCGTCATCAAGTGTTTTTTT
GCTTCTGTGCTGTCCTCAACACTTTTGAACTCTACGAACACCATGTCCGATCGATTTG
GCACGTGGCGAATCTCGTTGAATCCAGGCAAATGCGTGAACATCTAGAGAGAGGAAA
AATATTTAATATGGGAAATAGTTAAAGGAAAAGTGGATAGAGTGTACTGCGGACAAC
TCGGATGGTGTCATGTGTTTCGGAAGTCGACGAACATACAACGCGTTGCTCGCTTCCG
GTTCCGGTCGCGACAATGAAACGACCGTTTCGGCCGGTGAATTTGATTTCTTCTTCTT
CGTTTGCTTAGCTTCGTTCCGTAAGCTTTCATTTTTAGCGAATCCAATTCGAATTGGA
TTGCCGTACACGTGTTGGAAGCCTATTATCAAAATATCAAATAAAGATTACTTAGTGT
CGCAATGAACAATGTAATGCTTTTACTTTGCAATTGAGTTTTCGCAGTAATTGCGTGC
TTCAATTGTTCAAAGCACACAAATGCCTGCGATCGCAAATGTCTCTTGTTCGAGGTAT
ATACGTCATACAGTGGACCGTACGACCTGAACACTGCTATCAATCTATCTTTCTCAAA
TCTGCCTTCGGAACATTGGTCGGCAAATTGCCGATGAATAACGTTTTGTTTGGTGGGT
AAACCATTTTCATCTATAATGAGAAAATTTACTTGCTTCTTGAGTCGAAATAGTTATT
TATGACATCAAGTGCG

>Amplicon7
CCTCAGTAACATACTTTTTGCGGTCAATAACAAGTGCAGTCTGGACCACTGCATTACT
ATGCCTTCCGAACCATCCCTAGAAGTATGTAAAGTTAGGTGTTCAATTGATATCTCGC
CTAAATGTAGGCTAGTCAAAAATGCACACAAATTCATCATAGTAAGACTTCGTACGGG
TTTTAAAAAATAAGTAACGAGTATGTCTCGTACGTTGGCCTTATTCTGCTGACAGAAT
TGCAATATATATACGTAGTTCAATAAAAGAAAAAGTTTGCACAACTGAAACCTCCCGA
CGTAAACACTTTTTGTATTTATGGCATGAATCTCCAATCATGTCACCTTGGCTTCATA
TAAATAGCAATCTACACATCTACCTAAATGAATCTCCTGTTGGGCAGGCAGAATCAGC
AAAACATCTACAACGCCTAGTGGATATTGATGATGCTGTGAGATGATATTAGTTCCAT

TTTGCTAGGCATTTACAGGTACGAAAATGAAAAAAAAAAGAGAAGAGTAGAATGACGC
TTTCGAAACCCTCAAAATTTTCCTTAAGAATATTTTCAGTGAATAGAAAATATTTTGA
CTCGTATGATCCTTCAGAAAGTATGAGTGATTTATTTTATTTATTTATTTATGGACAA
CAAATTAAACTGGAGACACCACGTCGACAAAAAGCGCGACCAATTAAACTTAAAATCA
GTCAAATGTACCAGTTAACCAACGTTGGTCGGATACAGCTCTAAACAAAATAAACTTG
TACTGTAAATGATTGATTTATAAATCAATTATCGCGCTATTGGACTAGCTGTATGGCT
CTGAGCTGTACTAAAAAATCCGACATCGATGCTACACAAACAAGTCAAAAGAACTCCC
TACGAATGATCACCAACTGACATACGACACTTAAATTCAATGGATAGCCGATCTAATT
CGAGAATACGTCATAAAACACGAGAAGAGACTGTTAAATCACATATTCTGTCATCTTA
TTACTAGACATTAGT

>Amplicon8
TCTCGGTACACTCTTTCTCTATACCAGATTCATTCACAAATTTTTACTTGTTTATTTGG
CATTTTTGCACAAATTTTACCTTTCATATTTATCATAAAATTTTCAACTAAGTGACAG
CCGACTGATTTGTGTCGGAATTCGTACACTTTTATTTTCGAATTGAAGAGATTCGTTC
GATATTGTATCATCACAGCCGTATATCCTTCAAAGCTTGTTCAGCGGACACGCTTAGC
AAGTGTGGTGCTTTCGTCACGTTTAAAAACAAGGTGCAGGCATTTGGTATGAATTAAG
AAGAAACTTGACAAAGATGATGATACTTCAATGCAAACAAGTAGGAGAATGTTACCAA
CGATTTTCAAGCATAGGTAGTACTCTAAATTGAGTACTGAAACGGGCCAGTGTATCAA
ACAAATTTTCATACAAAATAGAACTCTATTTGACAGGTGTCGCTCGAAGTACTTTTGC
TTGAAGTGCGTTGATATTACTTGTAATGTTAGTTATAAAGTAGTTGTAGTCAAAACGA
GAGTAAATGAATTATACTTTTACAAACGTCTTTATCCAACCAACGCACTTCAAGCATG
AGTGGTACTCTTAATAGAATACTGAAACTGGACAGTGAAGTGCTTTGCTTTATCACCT
TAATAATTGTGCTTTGTTTGTTAGAGGGAGTGAGAGCTACTATTCATGAGTCAGCAGT
AGTTGAACCTCTTCACAAATTATGAACCATTGAGAGTAAAAAACTTCTTTGTTGTCTT
GCGGCCAGAAAACAACCTCTTTACAGTCTCGCTGAAAACCAATTTTTCGCATTTTCTG
GTCGCAAGACAACAGAGTACTGCTGTACTGTCTTGCAGGAAATAGAAATACAAAGAG
GAAATGCAGCTACAGTTTTAAAGACTTTCGAACAGAACTCAGAAAATAATTCTAACGT
TTTCGACATTTTGTAATTTACTGAAACTATGAAATTGTAAAAGAGACAATAACAATAC
ATGTGTCTCCAGCC

>Amplicon9
ATTCTGTTTTCTATTCGTTTGCCATTTCACGGTTGAATAAGTTCTTGAGTTAACGGAA
TCGTGTTTGATGCAAATTTTGCAATTTACATGCACCATCCCCAGAAAAAATCACGAGT
GAAAAGTATAACAACAGCAGAGTAACACAAAAAAAAACAAAGAGTTGGAAGTATAAG
AAATTTGGTACAAGACAAGCAAAGCATAGAAGGAAGAAAAAAAAAAACGGAGAAGAA
AAATGTCGCATGTGATGTGAACCCAAAAAGTTAGGGAAAAACTGAGAAATCTTGATA
AAAATAATGGCGTTGACTTTACAATTTTCTGTACCCCAGTAACAATAACAAAGCGACA
TAAATTTGCTGTTCTGAAAACGAATTTAAATAGTTTATTGCTTTCTTTTTGATTTAAT

TTATGGATTTTCTTTGCTTATTCAGCGAATGGAGTGGAACGAATATAATAGAAGGACT
TCCCCGTTTCCACCGTACACTACAAAATGTTCTACGATCTTGAGGTATACATTTTACG
ATAAAATCAATCTAACTCGTGTGTTTTTTGAGCTACAAGATTCGTTAACTGTCAACTT
CGATAAGATAAAGGTTTAGATAACAGCGCTGGAGATGGGTTATGTATCTATGCCATAT
GTATCTGGCAAAGTGCCGGCAACCTTAAAAAAATCACTCACTAAATTTCTGCCATTTG
TCATTAAAATCTATTACTTCGTTTGTTCAAGCAGACACCAGGTAGAGTTCTTGGTTTA
ATTTTCTCGCCGTTTTCGACAACAGGTGCCTGGTTATATATCGTTCCTTGAACTTCCC
TCGTCAACATATCATATTAACGTCAAGCAGTCTCTTGTTTCTATTTTCATTCACTCATT
TTTCGGATATTTGCTTTTTGTTCTTTTCCTTTTTGTACTGAAAACAATTTGTTTCTATA
ACTTCGAAAAAGAAACACCAATCCACCGCTGTGTACCCCCTACCCACTTTTCCTCTAA
TGTATTCTAAACTC

>Amplicon10
ACAATTTATTTTCGGATATTTACTTATTTACTCAAGTGGTCGATCTGTGCGTCAACTC
TCATTATAAAGAATCCAAAAAAATAAACTCACAAGTAGCAAATTGTGATACAGTTCCA
AAATGAAATTACTTCAATGAATTGTTGGCGTCGTTATTCAAGCGTTACATATGGTATA
GCAATGCTAAGTCCCGGGTTTACTTACAATTGATTAATGGTAGCTGATAATTATATCC
ATATCTCAAAGGAAGATAATGTTCGATATGATCCGCTTACTCTGAGCGCTAGTCCTAT
AGGACCTAGAGGAGGTGTAAACTCCAATGAGATTGTATAAAAACTATATTTTATGAAG
CATAGAAGGTCTTTCAGAGATAGACTGAGATGTACACTAGGCCTGCCCTCGTTCATAA
GCACCCTTCTAGCACTCAAAAACAAAAATGTGAGAACAAATATTTTTCCAACTCTTAC
AAGGATTGCTGCGATTTTCTGAGAATTATTTCTAGAATGTTTCCGTAAAATAGGCCCT
GCTGCTCCCTATAATGAAAATGGCTTATTCCTACAATATGAAAATCGTTATTTACATA
TTTTCTGTGGGTAGTAGATTTTAGCCCAAAGACAAGTGAAAAAGTCTAAAATCAACCG
TCCACAAAGGACGTATACAACGTTTTATGAAAGGGATCAAATTACGAGAAAAATTTGT
AATTCTGTCCCGTGGTAAGAAAATATTCGTGTGAATTATTCATTTTAATCCCGAGATT
TGCAATTGATTTTACATGTTGAGGACGGGGAAGGGCCTATTTTAACGGAATATCTTGA
AAATTTTTACATTTTACAAGCGATTTTCACTGTGTTCGCTCCTCAGATCGCTTATCGA
AATTGGTTTTCATGTACTCGAGCCAACGAGCTAACTTTCACTTTCATATTTTTTGTCA
TTTTAACAAAAAAATCAAACAAAATTTCAGAATTTGGACGCAAATGGCATGTAAATCT
AAGAGTAATTATAC

C.1.4 Ecdysone Receptor isoform validation

BLAST was used with the two known EcR isoforms from Sciara (EcR-A and EcR-B) [Foulk et al., 2013] to pull out transcripts in the Trinity assembly from all combined salivary gland samples.

blastn -db salglandtranscripts -query EcR-isoforms.fa -outfmt 6 -dust no -culling_limit 1 -max_target_seqs

1 -task blastn > ecr.blast

This resulted in identifying 5 transcripts that Trinity marked as coming from the same gene (c20107_g1), but different isoforms (i1, i2, i3, i4, i5). The transcripts were pulled out of the Trinity assembly using a custom python tool (`https://github.com/JohnUrban/sciara-project-tools`):

extractFastxEntries.py –fastx Trinity.fasta –fa -c c20107_g1_i1,c20107_g1_i2,c20107_g1_i3,c20107_g1_i4,c20107_g1_i5 > ecr-transcripts.fa

The transcripts were then aligned to the genome assembly with BLAST:

blastn -db genome -query ecr-transcripts.fa -outfmt 6 -dust no -culling_limit 1 -max_target_seqs 1 -task blastn > ecr-transcripts-aln-to-assembly.blast

These BLAST results were converted to BED format and each entry was put in a separate file to view in separate IGV tracks. The alignments were manually studied to understand the exon/intron structure. Isoform i5 corresponds to EcR-A and i1 (or i2) corresponds to EcR-B. Isofroms i3 and i4 share EcR-A-specific exons, though they lack its 2 5' exons, and each have one of their own 5' exons. Isoform i1 and i2 are EcR-B-like and differ only a little in the 5' exon boundaries.

Primers were designed with NCBI Primer BLAST. Specificity of primer pairs to both the transcriptome and genome assemblies using analyzePrimerPairs.py (`https://github.com/JohnUrban/sciara-project-tools`).
Universal primers designed for shared 3' end of EcR transcripts - these were used as positive controls for the reverse transcription PCR prcoedure:
o-EcR-U-fwd: GCTTTTTCCGGCGTAGTGTC
o-EcR-U-rev: GTGTAGGTGCGATTGTTGGC

The following primers were used to amplify specific isoforms. For instance, using the i2 forward primer and the universal (U) reverse primer should amplify a large segment of only the 2nd isoform:
o-EcR-A-l-fwd: CAGCTCACCAACAGCAATCG
o-EcR-B-l-fwd: GTTCTTGAAACGACCGCCTG
o-EcR-i1-l-fwd: GGACCCACATTTTTGTGTATGGG
o-EcR-i2-l-fwd: TGTGAAGAGACTTTGAAGAGATTTT
o-EcR-i3-l-fwd: GTATCGACAAAACTGAGCAGCA
o-EcR-i4-l-fwd: GCGGCGGTACAATGCTCTAT
o-EcR-i5-l-fwd: CAGACTCAGTGGATAATTTTGTTGG
o-EcR-U-l-rev: AAGTCACAGCCAAACGATAAGG

Below are the primers used for Sanger sequencing (Genewiz). The products produced from PCR reactions using the primers above were the substrates for sequencing. Sequencing was done to confirm whether the unique splice sites in the Trinity assembled transcripts were real or not. The letter indicates which isoform types they bind to [A-like or B-like isoforms] and the number indicates how many bases they are away from the splice site of interest:
o-EcR-A-seq-110-rev: ACTGTTGCACTGACACAATGTT
o-EcR-A-seq-360-rev: ACTCAACACGCTACCCAACG
o-EcR-B-seq-120-rev: TGTCAAGTCCTTTCTAAATCACATC
o-EcR-B-seq-310-rev: AATCCACCGGGTCCAAGAAC

The following is the protocol used for reverse transcriptase PCR:
Total RNA was extracted from 30 female, mixed eyespot stage, Holo2 larvae using TRIzol (ThermoFisher), following the manufacturer's instructions. Reverse transcription was carried out using Superscript III (ThermoFisher). Up to 5 μg total RNA was combined with 1 ul 50 ng/μl random primers, and 1 μl 10 mM dNTP mix in a final volume of 13 μl, and incubated in a thermocycler at 65°C for 5 minutes before transferring to ice. Then the following mix was added to the reaction: 4 μl 5X FSS buffer, 1 μl 0.1 M DTT, 1 μl Murine RNase Inhibitor, and 1 μl Superscript III. Reverse transcription was carried out in a thermocycler with the following procedure: 25°C for 10 minutes, 50°C for 50 minutes, 85°C for 5 minutes, 4° hold.

PCR reactions were carried out using Q5 HotStart Polymerase (NEB) according to the manufacturer's instructions. For a single reaction, the recipe was (25 μl volume): 5 μl Q5 buffer, 0.5 μl 10 mM dNTP mix, 1.25 μl Forward primer, 1.25 μl Reverse primer, 1 μl template, 0.25 μl Q5 HotStart Polymerase, 15.75 μl UltraPure Water (UPW, Invitrogen). The reaction was run in a thermocycler with the following procedure: 98°C for 30 seconds, then 40 cycles of 98°C for 10 seconds, 66°C 30 seconds, 72°C for 2 minutes. The procedure ended with 72°C for 2 minutes and holding at 4°C. The PCR products were run on 1% agarose gels at 150V for 30 minutes.

Supplementary Information: Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins

D.1 Supplementary Figures

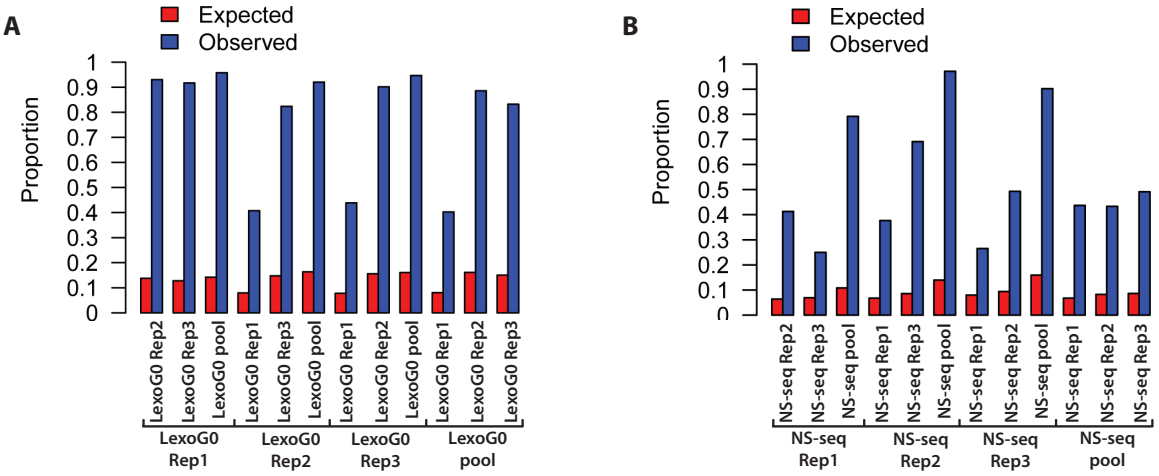


Figure D.1: Observed vs Expected overlap for replicates.

Barplot visualizations of the observed proportion of peak overlaps compared to the expected proportion of peak overlaps between replicates with each other and between replicates and the peak set resulting from pooling all reads. The proportions are of the peak set written in horizontal words that brackets three other peak sets. For example, the first three pairs of expected and observed proportions are of Rep1 that overlap the sets labeled under each expected and observed pair of bars (Rep2, Rep3, and Pool). This figure is related to Tables D.4 and D.5 where the expected and observed proportion values can be found along with p-values and other information.

(A) LexoG0$_{G0gDNA}$ peaks were called as described in Supplementary Methods. We identified 110,704 peaks, 194,025 peaks and 183,622 peaks in LexoG0 Reps1-3, respectively, and 196,851 peaks in the LexoG0 pooled data set. We observed significantly higher overlap of the peaks in each replicate with those in the LexoG0$_{G0gDNA}$ peak set from pooled reads (95.7%, 92.0% and 94.7%, respectively) than would be expected at random ($p<10^{-323}$). These results strongly support the conclusion that we were able to reproducibly identify peaks derived from λ-exonuclease (λ-exo) digested non-replicating DNA genome-wide. Moreover, the peak set from the pooled LexoG0 reads is representative of all the biological replicates, and most analyses were performed using this set of peaks from pooled reads.

(B) NS$_{G0gDNA}$ peaks were called as described in the Supplementary Methods. We identified 100,594 peaks, 95,030 peaks and 87,013 peaks in NS-seq Rep1-3, respectively, and 162,098 peaks from the NS-seq pooled read data set. The replicates all had significantly higher overlap than expected at random (also see Table D.5). All of the replicates showed significant overlap ($p<10^{-323}$) with the NS$_{G0gDNA}$ peak set called from pooled reads (79.1%, 97.1% and 90.2%, respectively). These results suggest that we were able to reproducibly detect peaks enriched by λ-exo digestion of replicating DNA and that the NS$_{G0gDNA}$ peak set from pooled reads is representative of the individual replicates, so most analyses were performed with peaks resulting from the pooled set of reads.

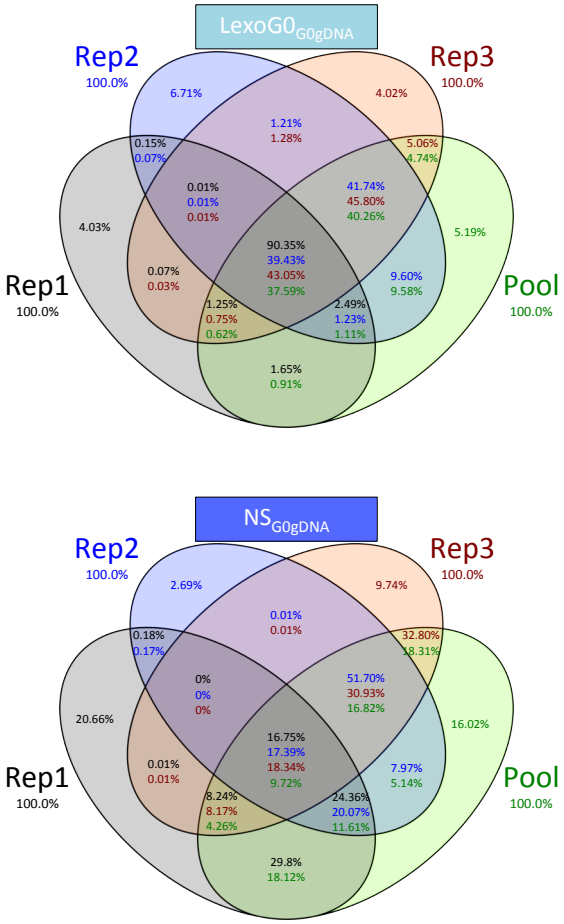


Figure D.2: Venn diagrams of overlaps of replicate peak sets and peak set from pooled reads.

Venn diagrams of the number of overlaps between all LexoG0$_{G0gDNA}$ peak sets (Reps 1-3 and set from pooled reads) and all NS$_{G0gDNA}$ peak sets (Reps 1-3 and set from pooled reads). For both, black text is replicate 1, blue text is replicate 2, brick red text is replicate 3, and green text is the set of peaks from pooled reads. Text and ellipsis color correspond for a given set. The four-way Venn diagram graphic was downloaded from `http://www.math.cornell.edu/~numb3rs/lipa/imgs/venn4.png`. Venn diagram values were obtained with a custom Python script that employed pybedtools [Dale et al., 2011] and were used to annotate the four-way Venn diagram graphic. Note that the area (size) of each section in the four-way Venn diagram does not correspond to the values within it. For both LexoG0$_{G0gDNA}$ and NS$_{G0gDNA}$, most of the mass (sum of percentages) of each replicate is within the pooled data set (green ellipsis), showing that it represents each well. Summing all percentages of a given color gives 100% (i.e. the full dataset represented by that color).

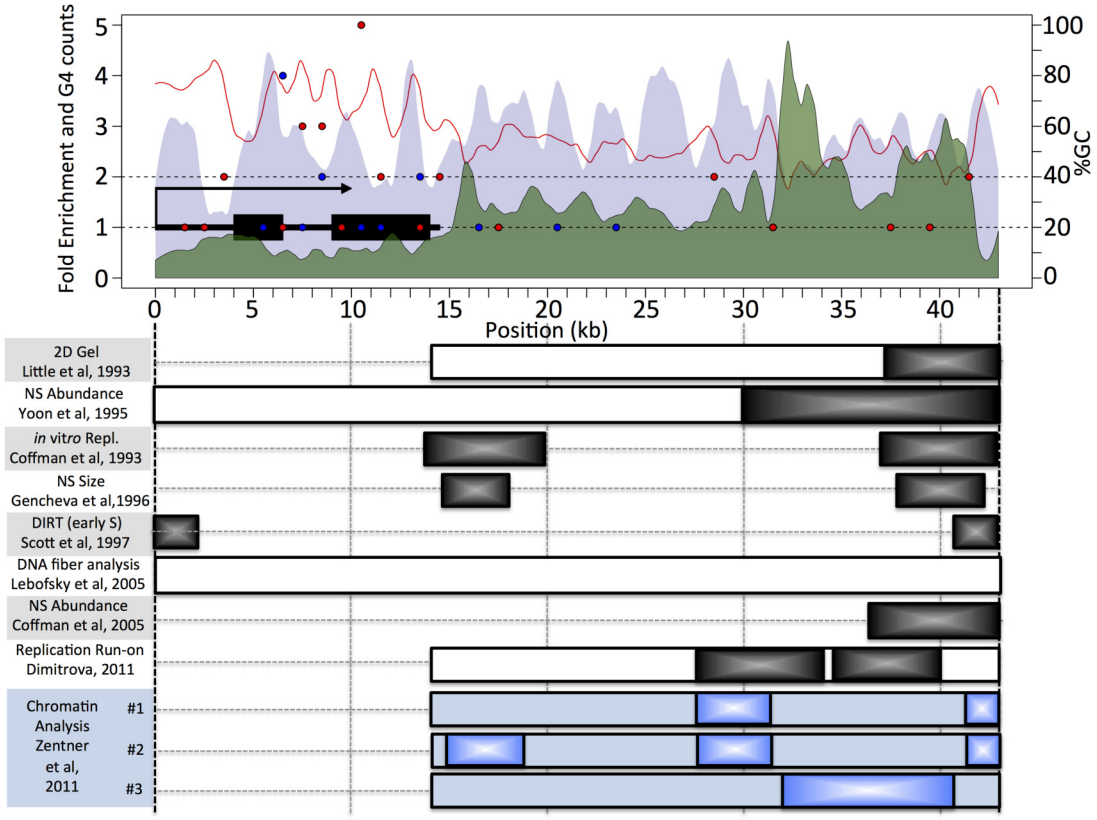


Figure D.3: Integrative look at origin activity and chromatin marks in human rDNA repeats.

At the top, NS-seq fold-enrichment signal when using G0gDNA (light blue-grey) and LexoG0 (green) as the control is shown the same way as in Figure 6.5B. Fold enrichment values are on the left Y-axis. The blue and red circles represent G4 counts in 1 kb bins on the positive and negative strand, respectively. G4 counts are also on the left Y-axis. The %GC signal is shown as a red line and is measured on the right Y-axis. The rRNA gene is depicted inside the plot with an arrow representing the transcription start site and direction of transcription. Note that rDNA repeats are typically tandem repeats and that the positions from 30-43 kb represent the upstream region, including the promoter, for the next rDNA repeat.

Legend for Supp. Fig. D.3 continued:
Below the fold enrichment plot are bars representing sites where previous studies found rDNA replication initiation activity. In general, the white bar (black outline) represents the entire area where initiation was detected, while the black bars represent sites of most frequent initiation activity. Although two studies detect initiation events everywhere across the rDNA repeat, most studies find replication activity restricted, or most frequent, in the intergenic spacer (also known as the 'non-transcribed spacer', NTS). Moreover, the most initiation activity across all studies appears to be between positions 30-43 kb and/or around positions 14-20 kb depending on the study. These areas are also the highest areas in the fold enrichment signal from our data, which suggest there are 3 preferred areas for initiation (near 15.5-16.5 kb, 31.5-34 kb, and 38-41.5 kb). Below the replication initiation bars are bars that represent three groups (#1, #2, and #3) of chromatin marks from a recent study [Zentner et al., 2011]. Group #1 represents H3K4me1, H3K4me2, H3K4me3, and H3K9ac. Group #2 represents H3K27ac only. Group #3 represents both H3K27me3 and H4K20me1. Similar to the representation of initiation activity above, the light blue bars (black outline) represent the area where a given chromatin mark is detected and the blue bars represent where the given mark was most enriched. H3K27ac marks the initiation zone that seems to occur near 15-20 kb and a pair of marks, H3K27me3 and H4K20me1, seem to coincide with most of the initiation zone between 30-43 kb. All relevant rDNA references are listed in the figure [Little et al., 1993, Yoon et al., 1995, Coffman et al., 1993, Gencheva et al., 1996, Scott et al., 1997, Lebofsky and Bensimon, 2005, Coffman et al., 2005, Dimitrova, 2011, Zentner et al., 2011].

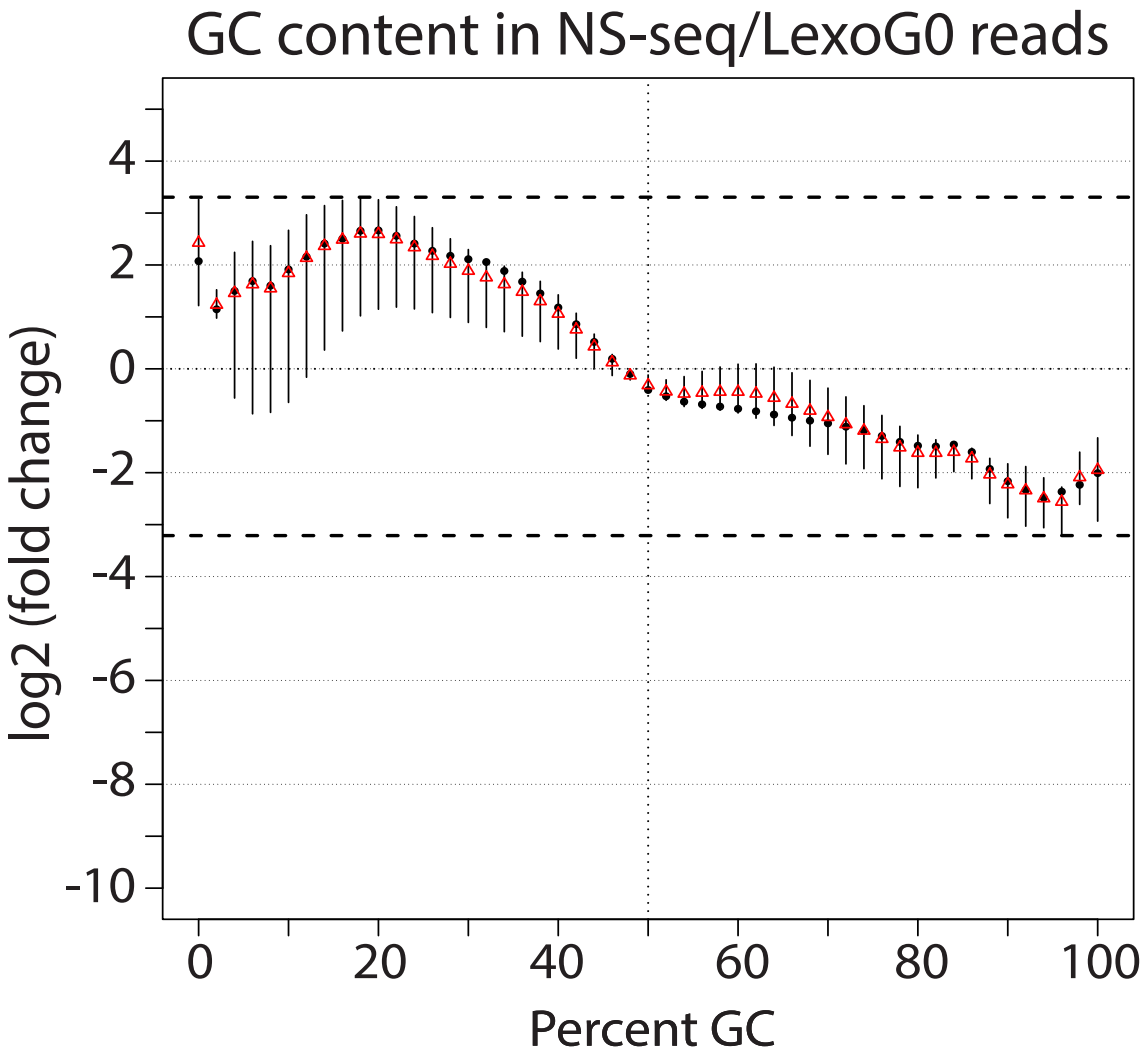


Figure D.4: Comparison of GC content in NS-seq vs. LexoG0 reads.
The log2(fold change) of the distribution of GC content in the three replicates of NS-seq reads relative to the pooled LexoG0 reads (i.e. log2(NS-seq/LexoG0)). This figure is supplementary to Figures 6.2A and 6.2B in the paper, is purposefully plotted on the same Y-axis scale (for direct comparison), and shows directly that NS-seq reads are enriched in AT-rich reads and depleted in GC-rich reads relative to LexoG0 reads. Over each GC% the minimum to maximum (line segment), median (black dot), and mean (red triangle) values for the NS-seq replicates (relative to the pooled LexoG0 reads) are shown.

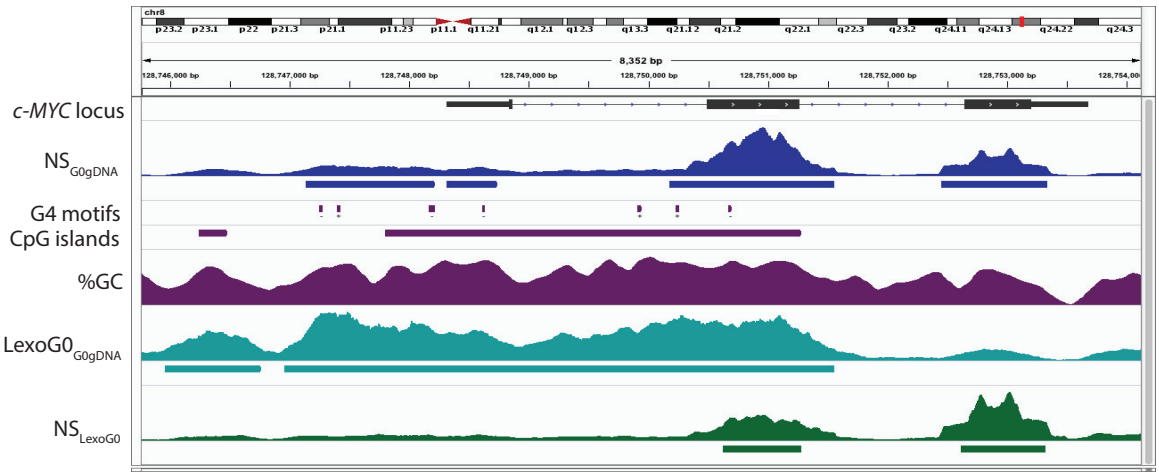


Figure D.5: MYC locus.

The NS_{G0gDNA}, LexoG0_{G0gDNA} and NS_{LexoG0} peak sets are illustrated at the MYC locus where origin activity in the promoter region and in exon two has been well characterized in HeLa cells [Tao et al., 2000]. NS_{G0gDNA} identified three peaks across this locus (blue), overlapping (i) the first MYC exon and upstream promoter region, (ii) the second MYC exon, and (iii) the last MYC exon. The locations of CpG islands, G4 motifs and GC content across the locus are indicated in purple. The LexoG0_{G0gDNA} peaks (cyan) overlap the CpG islands and G4 motifs as well as the first two NS_{G0gDNA} peaks. NS_{LexoG0} (green) lacks the upstream peak that overlaps with CpG islands, G4 motifs, and a LexoG0_{G0gDNA} peak, but contains the second exon peak that also overlaps these features. The third exon peak, which does not overlap these features, was preserved as expected. The first exon peak had the weakest fold-enrichment in NS_{G0gDNA}, suggesting that the preferred initiation sites in MCF7 cells are over the second and third exons in contrast to HeLa cells [Tao et al., 2000]. Given that the first exon peak is absent in NS_{LexoG0}, there may be some loss of sensitivity to weakly enriched origins in strongly λ-exo-biased regions. However, the presence of the second exon peak in NS_{LexoG0} demonstrates the advantage of controlling NS-seq with LexoG0 over the alternative procedure of discarding all NS_{G0gDNA} peaks that overlap LexoG0_{G0gDNA} peaks.

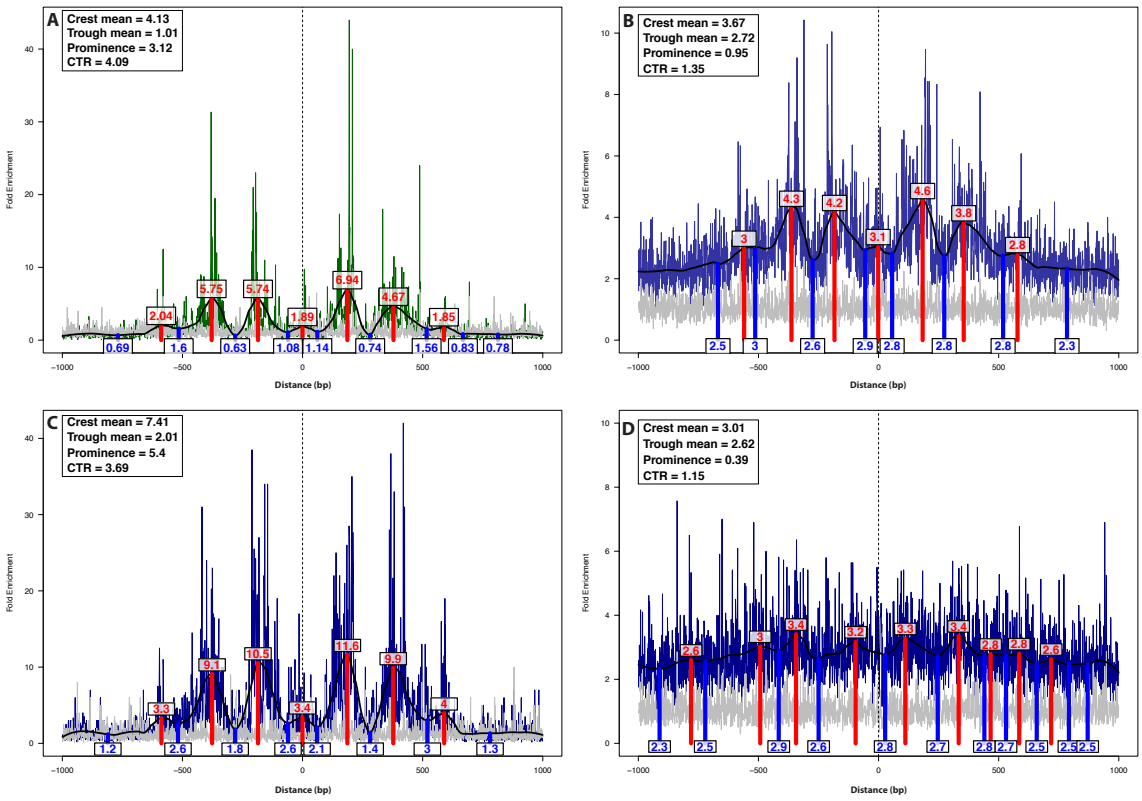


Figure D.6: Partitioning NS_{G0gDNA} into peaks that are and are not represented after controlling λ-exo biases decomposes G4 phasing into a stronger phased signal and a less phased signal.

This shows images from Fig. 6.6 at scales customized to each panel instead of displaying all on the same scale. See legend from Fig. 6.6 for more information. **(A)** all NS_{LexoG0} summits, **(B)** all NS_{G0gDNA} summits **(C)** NS_{G0gDNA} summits represented in NS_{LexoG0}, and **(D)** NS_{G0gDNA} summits not represented in NS_{LexoG0}.

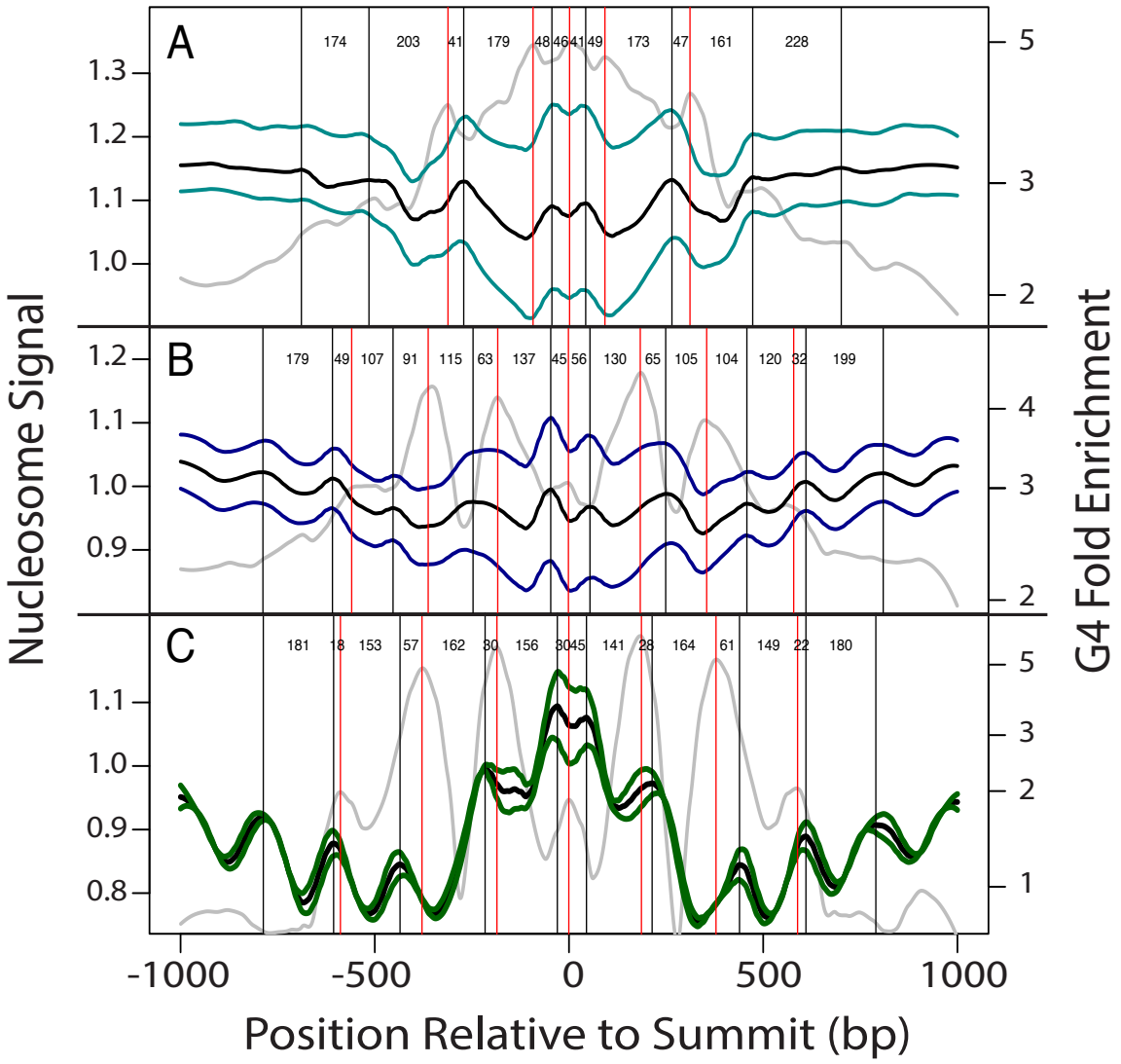


Figure D.7: Phasing of G4s is offset from phasing of nucleosomes.
This figure is similar to Figure 6.7A-C in the paper, but shows both the crest positions of nucleosomes (vertical black lines) and of G4 enrichments (vertical red lines) with distances between adjacent crests (regardless of crest-type) for the three subsets of peak summits within 1 kb of G4s: **(A)** LexoG0$_{G0gDNA}$ (cyan), **(B)** NS$_{G0gDNA}$ (blue), **(C)** NS$_{LexoG0}$ (green). This allows easy comparison of where G4 enrichment crests are with respect to nucleosome signal crests. In all cases they are offset from each other. The colored lines show the mean nucleosome signal over each position around summits for K562 and GM12878 cell lines while the central black lines show the mean of the two cell line signals. The light grey line shows the log transformed G4 enrichment profile. The right Y-axis shows G4 enrichment values (not log transformed), which are not uniformly spaced since they are mapped onto a log scale. Also see Table D.10, which is related to this figure. The distances from left to right in bp are: LexoG0$_{G0gDNA}$ = 174, 203, 41, 179, 48, 46, 41, 49, 173, 47, 161, 228; NS$_{G0gDNA}$ = 179, 49, 107, 91, 115, 63, 137, 45, 56, 130, 65, 105, 104, 32, 199; NS$_{LexoG0}$ = 181, 18, 153, 57, 162, 30, 156, 30, 45, 141, 28, 164, 61, 149, 22, 180.

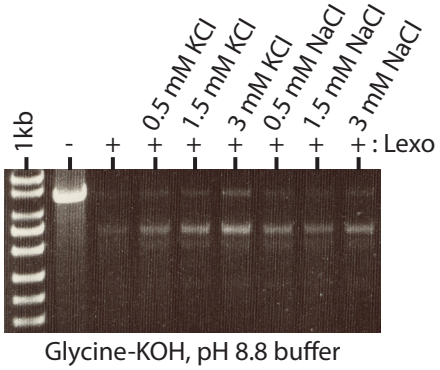


Figure D.8: Effects of K^+ and Na^+ concentrations on λ-exo digestion.
Increasing the concentration of potassium present in the plasmid experiments (glycine-KOH buffer, pH 8.8) by titrating KCl resulted in stronger bands signifying that more G4 structures were stabilized, thwarting λ-exo digestion. Increasing the concentration of sodium ions present in the glycine-KOH buffer (pH 8.8) by titrating NaCl did not result in stronger bands signifying that Na^+ ions did not contribute to the stabilization of more G4 structures. It is likely that sodium ions did not contribute much to G4 stability compared to K^+ ions because G4 folding and unfolding kinetics differ in the presence of each [Shim et al., 2009]. G4s fold fast and unfold slowly in the presence of K^+ while both folding and unfolding are fast in the presence of Na^+ . In addition, the melting temperature for G4s stabilized by Na^+ ions is much lower than that for G4s stabilized by K^+ ions [Kankia and Marky, 2001]. At 37°C (the temperature of λ-exo digestion in our experiments and those of others), most G4s are not likely stabilized by Na^+ , but they are very likely stabilized in the presence of K^+.

D.2 Supplementary Tables

Mapping Statistics				
Sample	Read Length	Total # Reads	Mappable Reads	% Mappable
LexoG0 Rep1	50	162102273	115150124	71.1
LexoG0 Rep2	50	196524435	174625028	88.9
LexoG0 Rep3	50	174860103	153657189	87.9
LexoG0 pool	50	533486811	443432341	83.1
NS-seq Rep1	50	128247879	57397008	44.7
NS-seq Rep2	50	139320824	89354586	64.1
NS-seq Rep3	50	136227515	96762425	71
NS-seq pool	50	403796218	243514019	60.3
G0gDNA	50	193565007	181911420	94

Table D.1: Read mapping statistics.
Shows numbers of reads obtained from Illumina HiSeq 2000 and numbers of reads that were mapped to hg19.

Number of Peaks		
Sample	Background Control	# of Peaks Called
$LexoG0_{G0gDNA}$ Rep1	G0gDNA	110704
$LexoG0_{G0gDNA}$ Rep2	G0gDNA	194025
$LexoG0_{G0gDNA}$ Rep3	G0gDNA	183622
$LexoG0_{G0gDNA}$ pool	G0gDNA	196851
NS_{G0gDNA} Rep1	G0gDNA	100594
NS_{G0gDNA} Rep2	G0gDNA	95030
NS_{G0gDNA} Rep3	G0gDNA	87013
NS_{G0gDNA} pool	G0gDNA	162098
NS_{LexoG0} pool	LexoG0	66831

Table D.2: Peak statistics.
Shows numbers of peaks in the peak sets discussed in the paper.

Fold Enrichment Correlation of Replicates		
Sample1	**Sample2**	**FE**
$LexoG0_{G0gDNA}$ Rep1	$LexoG0_{G0gDNA}$ Rep2	0.789778
$LexoG0_{G0gDNA}$ Rep1	$LexoG0_{G0gDNA}$ Rep3	0.781041
$LexoG0_{G0gDNA}$ Rep2	$LexoG0_{G0gDNA}$ Rep3	0.950533
$LexoG0_{G0gDNA}$ Pool	$LexoG0_{G0gDNA}$ Rep1	0.84553
$LexoG0_{G0gDNA}$ Pool	$LexoG0_{G0gDNA}$ Rep2	0.975653
$LexoG0_{G0gDNA}$ Pool	$LexoG0_{G0gDNA}$ Rep3	0.971312
NS_{G0gDNA} Rep1	NS_{G0gDNA} Rep2	0.675875
NS_{G0gDNA} Rep1	NS_{G0gDNA} Rep3	0.571822
NS_{G0gDNA} Rep2	NS_{G0gDNA} Rep3	0.806056
NS_{G0gDNA} Pool	NS_{G0gDNA} Rep1	0.783419
NS_{G0gDNA} Pool	NS_{G0gDNA} Rep2	0.927668
NS_{G0gDNA} Pool	NS_{G0gDNA} Rep3	0.924786

Table D.3: Correlations of Fold Enrichment Between Replicates.
Shows Pearson's product-moment correlation coefficients (Pearson's r) of genome-wide fold enrichment (FE) signals (wigCorrelate). When comparing replicates to each other, it is a measure of reproducibility. When comparing a replicate to the fold enrichment signal resulting from the pooled reads, it is a measure of how well the pooled data represent the replicate.

$LexoG0_{G0gDNA}$ replicates and pooled set:
The genome-wide fold enrichment signals from the three replicates were highly correlated with each other showing Pearson's r ranging from 0.78 to 0.95. All three replicates were highly correlated with the fold enrichment signal obtained from pooled reads as well (Pearson's r = 0.84, 0.97 and 0.97, for Rep1-3 respectively). As all replicate fold enrichment signals were highly correlated (and peak sets significantly overlapped; Table D.4) and since the pooled data set was a balanced representation of each (determined by FE signal correlation and peak overlaps; Table D.4), the peaks resulting from the pooled data set were used for most analyses and the pooled set of LexoG0 reads was used as the LexoG0 control for NS_{LexoG0}.

NS_{G0gDNA} replicates and pooled set:
The fold enrichment signals of NS_{G0gDNA} replicates were highly correlated with each other displaying Pearson's r ranging from 0.67 to 0.81. Additionally, each of the replicates was highly correlated with peaks called from the pooled set of reads (Pearson's r = 0.78, 0.93 and 0.92, for Rep1-3 respectively). As all replicates were highly correlated and reproducible and since the pooled data set was representative of each (both also determined by peak overlap; Table D.5), the peaks resulting from the pooled data set were used for most analyses and the pooled set of NS-seq reads was used as the NS treatment file for NS_{LexoG0} MACS2 peak calling.

fileA	fileB	expNum	obsNum	expProportion	obsProportion	pVal *	obsToExpRatio
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ Rep1	7383.991861	110704	0.066700317	1	0	14.99243256
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ Rep2	15100.44786	102960	0.136403814	0.930047695	0	6.818340817
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ Rep3	14015.89049	101496	0.126606902	0.91682324	0	7.241494936
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ pool	15586.88905	105998	0.140797885	0.957490244	0	6.800459005
LexoG0$_{G0gDNA}$ Rep2	LexoG0$_{G0gDNA}$ Rep1	15100.83794	79055	0.077829341	0.407447494	0	5.235139953
LexoG0$_{G0gDNA}$ Rep2	LexoG0$_{G0gDNA}$ Rep2	30250.35358	194025	0.155909566	1	0	6.413974616
LexoG0$_{G0gDNA}$ Rep2	LexoG0$_{G0gDNA}$ Rep3	28146.5769	159841	0.145066754	0.823816518	0	5.67887884
LexoG0$_{G0gDNA}$ Rep2	LexoG0$_{G0gDNA}$ pool	31158.04845	178511	0.160587803	0.920041232	0	5.729209912
LexoG0$_{G0gDNA}$ Rep3	LexoG0$_{G0gDNA}$ Rep1	14016.20383	80497	0.076331833	0.438384289	0	5.743138511
LexoG0$_{G0gDNA}$ Rep3	LexoG0$_{G0gDNA}$ Rep2	28146.47907	165512	0.153284895	0.901373474	0	5.880380263
LexoG0$_{G0gDNA}$ Rep3	LexoG0$_{G0gDNA}$ Rep3	26181.34223	183622	0.142582818	1	0	7.013467774
LexoG0$_{G0gDNA}$ Rep3	LexoG0$_{G0gDNA}$ pool	28998.48508	173809	0.157924895	0.946558691	0	5.993726897
LexoG0$_{G0gDNA}$ pool	LexoG0$_{G0gDNA}$ Rep1	15587.34068	79185	0.079183447	0.402258561	0	5.080083999
LexoG0$_{G0gDNA}$ pool	LexoG0$_{G0gDNA}$ Rep2	31158.14638	174283	0.158282896	0.885354913	0	5.593497055
LexoG0$_{G0gDNA}$ pool	LexoG0$_{G0gDNA}$ Rep3	28998.67702	163796	0.147312826	0.832081117	0	5.648395612
LexoG0$_{G0gDNA}$ pool	LexoG0$_{G0gDNA}$ pool	32085.86815	196851	0.162995708	1	0	6.135130864

∗ pVal = 0 corresponds to a p value < 10e-323

Table D.4: Overlap statistics between replicate peak sets and the peak set resulting from pooled reads for LexoG0$_{G0gDNA}$.

The observed and expected proportions are visualized in Figure D.1. The observed and expected number of overlaps is the observed and expected number of peaks in fileA that overlap peaks in fileB. Note that in the tables, a p-value of 0 simply means it was so small that R considered it 0, which occurs around 2.5e^{-324}. Thus, elsewhere when discussing p-values, if it was 0 in R, we say p < 10^{-323}. The expected proportion and p-values were obtained through the binomial model described in the Supplementary Methods. Significant overlap between replicates is a measure of reproducibility while significant overlap between replicates and the peak set resulting from pooled reads is a measure of how well the pooled set represents the replicate.

fileA	fileB	expNum	obsNum	expProportion	obsProportion	pVal *	obsToExpRatio
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ Rep1	4765.535408	100594	0.047373953	1	0	21.10864602
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ Rep2	6336.469076	41533	0.062990527	0.412877508	0	6.554596811
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ Rep3	6876.176742	25147	0.068355734	0.249985089	0	3.657119493
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ pool	10775.47803	79622	0.107118496	0.791518381	0	7.389184939
NS$_{G0gDNA}$ Rep2	NS$_{G0gDNA}$ Rep1	6336.781606	35764	0.066681907	0.376344312	0	5.643874481
NS$_{G0gDNA}$ Rep2	NS$_{G0gDNA}$ Rep2	7719.422757	95030	0.08123143	1	0	12.31050598
NS$_{G0gDNA}$ Rep2	NS$_{G0gDNA}$ Rep3	8083.091148	65670	0.085058309	0.691044933	0	8.124367126
NS$_{G0gDNA}$ Rep2	NS$_{G0gDNA}$ pool	13136.28301	92302	0.138233011	0.971293276	0	7.026492952
NS$_{G0gDNA}$ Rep3	NS$_{G0gDNA}$ Rep1	6876.732809	23076	0.079031097	0.265201751	0	3.355663313
NS$_{G0gDNA}$ Rep3	NS$_{G0gDNA}$ Rep2	8083.346126	42878	0.092898143	0.492776941	0	5.30448645
NS$_{G0gDNA}$ Rep3	NS$_{G0gDNA}$ Rep3	8330.722321	87013	0.095741123	1	0	10.44483259
NS$_{G0gDNA}$ Rep3	NS$_{G0gDNA}$ pool	13759.67614	78512	0.158133568	0.902301955	0	5.705948252
NS$_{G0gDNA}$ pool	NS$_{G0gDNA}$ Rep1	10776.0039	70852	0.066478327	0.43709361	0	6.574979058
NS$_{G0gDNA}$ pool	NS$_{G0gDNA}$ Rep2	13136.27618	70182	0.081039101	0.432960308	0	5.342609963
NS$_{G0gDNA}$ pool	NS$_{G0gDNA}$ Rep3	13759.23495	79603	0.084882201	0.49107947	0	5.78542341
NS$_{G0gDNA}$ pool	NS$_{G0gDNA}$ pool	22354.11578	162098	0.137904945	1	0	7.251371585

∗ pVal = 0 corresponds to a p value < 10e-323

Table D.5: Overlap statistics between replicate peak sets and the peak set resulting from pooled reads for NS$_{G0gDNA}$.

See Table D.4 description for more information.

A. Peak density vs. feature density

	G4 (100 kb bins)		CpG Islands (1 Mb bins)	
q < 0.001 peak sets	**Pearson**	**Spearman**	**Pearson**	**Spearman**
$LexoG0_{G0gDNA}$ pool	0.704	0.704	0.646	0.746
NS_{G0gDNA} pool	0.692	0.363	0.802	0.490
NS_{LexoG0} pool	-0.248	-0.260	-0.364	-0.472

B. Average Fold Enrichment Signal vs. G4 density

	G4 (100 kb bins)	
Fold Enrichment Signal	**Pearson**	**Spearman**
$LexoG0_{G0gDNA}$ pool	0.862	0.776
NS_{G0gDNA} pool	0.692	0.564
NS_{LexoG0} pool	-0.124	0.004

Table D.6: Correlation of peak densities and fold enrichment signals with G4 and CpG Island densities.

Correlations of peaks or average fold enrichment with given feature in 100 kb or 1 Mb bins. Both Pearson's r and Spearman's rank-order correlation are given.

file A	file B	expNum	obsNum	expProportion	obsProportion	pVal	obsToExpRatio
$LexoG0_{G0gDNA}$	$LexoG0_{G0gDNA}$	32085.86815	196851	0.162995708	1	0	6.135130864
$LexoG0_{G0gDNA}$	NS_{G0gDNA}	26783.96267	62230	0.136062111	0.316127426	0	2.323405269
$LexoG0_{G0gDNA}$	NS_{LexoG0}	12895.44162	12513	0.065508642	0.063565844	0.999762207	0.970342883
$LexoG0_{G0gDNA}$	G4	29934.79818	72554	0.152068306	0.368573185	0	2.423734396
$LexoG0_{G0gDNA}$	CpG	3830.944083	23865	0.019461136	0.121233827	0	6.229534935
file A	**file B**	**expNum**	**obsNum**	**expProportion**	**obsProportion**	**pVal**	**obsToExpRatio**
NS_{G0gDNA}	$LexoG0_{G0gDNA}$	26784.04091	76265	0.16523363	0.470486989	0	2.847404552
NS_{G0gDNA}	NS_{G0gDNA}	22354.11578	162098	0.137904945	1	0	7.251371585
NS_{G0gDNA}	NS_{LexoG0}	10741.98268	60434	0.066268447	0.372823847	0	5.625963271
NS_{G0gDNA}	G4	25311.32205	56600	0.156148269	0.349171489	0	2.236153445
NS_{G0gDNA}	CpG	3207.148297	13405	0.019785243	0.082696887	0	4.17972565
file A	**file B**	**expNum**	**obsNum**	**expProportion**	**obsProportion**	**pVal**	**obsToExpRatio**
NS_{LexoG0}	$LexoG0_{G0gDNA}$	12895.94605	13492	0.192963536	0.20188236	3.0316E-09	1.046220258
NS_{LexoG0}	NS_{G0gDNA}	10742.37149	62357	0.16073935	0.933055019	0	5.804770398
NS_{LexoG0}	NS_{LexoG0}	5057.979259	66831	0.07568313	1	0	13.21298419
NS_{LexoG0}	G4	13814.15162	23718	0.206702752	0.354895183	0	1.71693497
NS_{LexoG0}	CpG	1590.659661	39	0.023801225	0.000583562	1	0.02451813
file A	**file B**	**expNum**	**obsNum**	**expProportion**	**obsProportion**	**pVal**	**obsToExpRatio**
G4	$LexoG0_{G0gDNA}$	29931.68791	174050	0.083394456	0.484931056	0	5.814907617
G4	NS_{G0gDNA}	25308.61823	93270	0.070513846	0.259865094	0	3.685305897
G4	NS_{LexoG0}	13812.17603	24383	0.038482925	0.067934926	0	1.765326474
CpG	$LexoG0_{G0gDNA}$	3830.800926	26189	0.134366921	0.918589968	0	6.83642938
CpG	NS_{G0gDNA}	3207.019083	12600	0.112487516	0.441950193	0	3.928882141
CpG	NS_{LexoG0}	1590.538004	33	0.055788776	0.001157489	1	0.020747697

Table D.7: Overlap statistics for peaks and other genomic features.
Overlap analysis of peak sets (resulting from pooled read sets) with each other and with other features (CpG islands and G4s). The observed number of overlaps is the number of peaks in file A that overlap peaks in file B. When p-value is 0, interpret it as $p < 10^{-323}$.

NSG0gDNA and LexoG0gDNA Correlation		
Sample1	**Sample2**	**FE**
NS_{G0gDNA} Pool	$LexoG0_{G0gDNA}$ Pool	0.557402
NS_{G0gDNA} Pool	$LexoG0_{G0gDNA}$ Rep1	0.682616
NS_{G0gDNA} Pool	$LexoG0_{G0gDNA}$ Rep2	0.511703
NS_{G0gDNA} Pool	$LexoG0_{G0gDNA}$ Rep3	0.507415
$LexoG0_{G0gDNA}$ Pool	NS_{G0gDNA} Pool	0.557402
$LexoG0_{G0gDNA}$ Pool	NS_{G0gDNA} Rep1	0.714478
$LexoG0_{G0gDNA}$ Pool	NS_{G0gDNA} Rep2	0.569375
$LexoG0_{G0gDNA}$ Pool	NS_{G0gDNA} Rep3	0.325473

NS_{LexoG0} Correlation		
Sample1	**Sample2**	**FE**
NS_{LexoG0} Pool	$LexoG0_{G0gDNA}$ Pool	0.171399
NS_{LexoG0} Pool	$LexoG0_{G0gDNA}$ Rep1	0.327537
NS_{LexoG0} Pool	$LexoG0_{G0gDNA}$ Rep2	0.133805
NS_{LexoG0} Pool	$LexoG0_{G0gDNA}$ Rep3	0.13193
NS_{LexoG0} Pool	NS_{G0gDNA} Pool	0.781847
NS_{LexoG0} Pool	NS_{G0gDNA} Rep1	0.422503
NS_{LexoG0} Pool	NS_{G0gDNA} Rep2	0.703584
NS_{LexoG0} Pool	NS_{G0gDNA} Rep3	0.849206

Table D.8: Correlations Between NS_{G0gDNA}, $LexoG0_{G0gDNA}$, and NS_{LexoG0} Fold Enrichment signals.

Sample Name	Total num raw Reads	num Mappable Reads to hg19+rDNA	num reads mapped to rDNA repeat in context of hg19	num reads (mapq >= 2) mapped to rDNA repeat in context of hg19	number non-redundant* reads with mapq >= 2
LexoG0 Rep1	162102273	116086307	1901110	1615177	1203114
LexoG0 Rep2	196524435	176323067	3291863	2767654	1893121
LexoG0 Rep3	174860103	155299989	3097505	2588199	1781096
LexoG0 Pool	533486811	447709363	8290478	6971030	4877331
NS-seq Rep1	128247879	57840335	848305	725977	605176
NS-seq Rep2	139320824	89873195	954705	809750	702167
NS-seq Rep3	136227515	97416920	1209455	1028929	882839
NS-seq pool	403796218	245130450	3012465	2564656	2190182
G0 gDNA	193565007	181619332	696286	614554	610713

Table D.9: Read mapping statistics for rDNA analyses.
Numbers of Illumina HiSeq 2000 reads for each sample that mapped to hg19+rDNA, a modified hg19 genome that contained a copy of the 43 kb rDNA repeat (`http://www.ncbi.nlm.nih.gov/nuccore/555853?report=fasta`) as an additional "chromosome", and how many reads mapped to the rDNA repeat itself.

Summits +/- 1kb	G4s	expNum	obsNum	expProportion	obsProportion	pVal	obsToExpRatio
$LexoG0_{G0gDNA}$	G4 centers	49865.5276	90810	0.2533161	0.4613134	0	1.821098
NS_{G0gDNA}	G4 centers	41062.03318	70772	0.2533161	0.4366001	0	1.723539
NS_{LexoG0}	G4 centers	16929.36828	23248	0.2533161	0.3478625	0	1.373235

G4s	Summits +/- 1kb	expNum	obsNum	expProportion	obsProportion	pVal	obsToExpRatio
G4 centers	$LexoG0_{G0gDNA}$	49856.47853	156537	0.1389081	0.436137	0	3.139752
G4 centers	NS_{G0gDNA}	41054.57748	115280	0.1143846	0.3211885	0	2.807969
G4 centers	NS_{LexoG0}	16926.2996	26481	0.04715937	0.07378029	0	1.564489

	$LexoG0_{G0gDNA}$	NS_{G0gDNA}	NS_{LexoG0}
numSummits +/- 1kb that overlap G4s	90810	70772	23248
% of all summits +/- 1kb that overlap >= 1 G4	46.13134	43.66001	34.78625
Of those that overlap >= 1 G4, % that overlaps 1 G4:	56.77128	60.97044	91.63369
Of those that overlap >= 1 G4, % that overlaps 2 G4s:	23.2067	18.25157	5.785444
Of those that overlap >= 1 G4, % that overlaps 3 G4s:	9.767647	9.399197	0.8817963
Of those that overlap >= 1 G4, % that overlaps 4 G4s:	4.740667	5.269033	0.5204749
Of those that overlap >= 1 G4, % that overlaps 5 G4s:	2.469992	2.797717	0.4129387
Of those that overlap >= 1 G4, % that overlaps 6 G4s:	1.342363	1.514723	0.2623882
Of those that overlap >= 1 G4, % that overlaps >= 7 G4s:	1.701351	1.79732	0.5032679
On average, of those that overlap >= 1 G4, overlaps this many G4s:	1.865907	1.853035	1.163068

Table D.10: G4s within 1 kb of peak summits.

Since the G4 enrichment signal around NS_{G0gDNA} and NS_{LexoG0} was phased, with inter-crest distances reminiscent of nucleosome spacing, it suggested that there was a relationship between G4s and nucleosomes. Thus, nucleosomal signal was assayed around the subset of summits that were proximal to G4s, defined as summits that have $\geq$ 1 G4 motif within 1 kb in either direction. A summit window is defined here as a summit +/- 1 kb. This table provides the statistics on how many summit windows overlap G4s and vice versa. Moreover, for summit windows that overlap $\geq$ 1 G4, how many overlap 1 G4, 2 G4s, 3 G4s (etc) is shown. The G4 enrichment signal that summarizes all NS_{LexoG0} summits is highly prominent and phased around the summit position (Figures 6.4E, 6.6A, D.6A). Nonetheless, most (91.6%) of the 34.7% of NS_{LexoG0} summits that are proximal to G4s have just a single G4 nearby, which is typically 3' to the summit when strand information is considered (Figure 6.3F) and is typically spaced 1-3 nucleosomal distance units (185-210 bp) away from the NS summits (Figure 6.3C).

Experiment	A	B	Pearson	Spearman
$LexoG0_{G0gDNA}$	mean nucleosome smoothed	G4	-0.84944833	-0.9062846
$LexoG0_{G0gDNA}$	K562 smoothed nucleosome	G4	-0.95448538	-0.96137874
$LexoG0_{G0gDNA}$	GM12878 smoothed nucleosome	G4	-0.00048484	-0.20675595
$LexoG0_{G0gDNA}$	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.1891064	0.3402856
$LexoG0_{G0gDNA}$	K562 smoothed nucleosome	mean nucleosome smoothed	0.9499651	0.976812
$LexoG0_{G0gDNA}$	GM12878 smoothed nucleosome	mean nucleosome smoothed	0.4863646	0.5113621
$LexoG0_{G0gDNA}$	K562 raw nucleosome	GM12878 raw nucleosome	0.2135404	0.366179
NS_{G0gDNA}	mean nucleosome smoothed	G4	-0.80415049	-0.87139479
NS_{G0gDNA}	K562 smoothed nucleosome	G4	-0.79586782	-0.86944983
NS_{G0gDNA}	GM12878 smoothed nucleosome	G4	-0.42899402	-0.48135622
NS_{G0gDNA}	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.308657032	0.355864991
NS_{G0gDNA}	K562 smoothed nucleosome	mean nucleosome smoothed	0.909581044	0.912055128
NS_{G0gDNA}	GM12878 smoothed nucleosome	mean nucleosome smoothed	0.675986399	0.680323131
NS_{G0gDNA}	K562 raw nucleosome	GM12878 raw nucleosome	0.340282529	0.413766276
NS_{LexoG0}	mean nucleosome smoothed	G4	-0.00690628	-0.04702974
NS_{LexoG0}	K562 smoothed nucleosome	G4	-0.08227209	-0.08133883
NS_{LexoG0}	GM12878 smoothed nucleosome	G4	0.047354732	-0.03519783
NS_{LexoG0}	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.950578073	0.968156115
NS_{LexoG0}	K562 smoothed nucleosome	mean nucleosome smoothed	0.983128609	0.9888404
NS_{LexoG0}	GM12878 smoothed nucleosome	mean nucleosome smoothed	0.991333183	0.993413977
NS_{LexoG0}	K562 raw nucleosome	GM12878 raw nucleosome	0.946049507	0.961888765

Table D.11: Correlation of nucleosome positioning between cell lines.

Nucleosome signal was plotted around the subset of peak summits that were proximal to G4s (had G4s within 1 kb; Table D.10). The consistency of nucleosomal positioning relative to these peak summits was tested by correlation as well as how much variation there was between the 2 cell line signals. Note that the **raw** nucleosome signal for a given cell line is the fold enrichment of the mean nucleosome score over each relative position from the summit (for all specified summits) divided by the "genomic mean score at random" over each position (from shuffling the specified peak summits). The **smoothed** nucleosomal signal for a given cell line results from lightly loess smoothing the raw signal defined above to round out jagged edges. The **"mean nucleosome smoothed"** signal results from taking the overall mean from the 2 cell line raw nucleosomal signal means (defined above) at each position and lightly loess smoothing it to round out jagged edges.

Shown in table, correlations of aggregated nucleosome signals between cell lines. The correlations between aggregated G4 and nucleosome signals are also provided.

Experiment	A	B	Pearson	Spearman
NS_{G0gDNA} (in NS_{LexoG0})	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.9105249	0.926113
NS_{G0gDNA} (in NS_{LexoG0})	K562 raw nucleosome	GM12878 raw nucleosome	0.9051878	0.9230251
NS_{G0gDNA} (not in NS_{LexoG0})	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.901602	0.938893
NS_{G0gDNA} (not in NS_{LexoG0})	K562 raw nucleosome	GM12878 raw nucleosome	0.903426	0.933372

Table D.12: Correlation of nucleosome positioning around NS_{G0gDNA} summits between cell lines after decomposition.
See description of Table D.10 for more information.
Pearson's r and Spearman's rank-order correlation between the nucleosome signal from each cell line after partitioning the NS_{G0gDNA} summits into NS_{G0gDNA} summits that overlap with NS_{LexoG0} summit windows and those that do not overlap with NS_{LexoG0} summit windows.

Sample	Sum of deviations^2 from mean over each position
Before Decomposition	
$LexoG0_{G0gDNA}$	28.86581618
NS_{G0gDNA}	17.95427157
NS_{LexoG0}	1.43678066
After Decomposition of NS_{G0gDNA}	
NS_{G0gDNA} (peaks in NS_{LexoG0})	1.674045
NS_{G0gDNA} (peaks NOT in NS_{LexoG0})	28.96305

Table D.13: How much the nucleosome signal around summits differs between cell lines.
See description of Table D.10 for more information.
The amount of variation between the two cell line signals was measured by taking the sum of squared deviations of the "smoothed cell line signals" (defined above) from the "mean nucleosome smoothed signal" (defined above) over each position.

D.3 Supplementary Methods

D.3.1 λ-exonuclease (λ-exo) digestion of plasmid DNA.

pFRT.myc6xERE is a 7180 bp plasmid that contains a 2.4 kb genomic fragment from the promoter region of the MYC gene [Malott and Leffak, 1999] with a region shown to be unnecessary for origin activity (Δ11; [Liu et al., 2003]) replaced by a 6x estrogen response element (6xERE) cassette. This construct contains the NHE III_1 element of the MYC promoter. The purine-rich strand of this element has been shown to form a G4 structure (Pu27; reviewed by Brooks and Hurley 2010). Plasmid DNA was linearized with BglII (New England Biolabs (NEB)), purified using Ampure beads (Beckman Coulter) and labeled at the 3' end using terminal transferase (New England Biolabs) and α^{32}P-dCTP (Perkin Elmer) under conditions that add 3-6 nucleotides (approx. 1:400 ratio of 3' ends to α^{32}P-CTP; 37°C for 1 hr). Labeled fragments were purified over a Sephadex G-50 column (Sigma) and 200 ng was digested overnight (16-18 hours) with λ-exo in the buffer indicated; ten units of a custom, high concentration preparation of λ-exo from Fermentas (20 units/μl) were used per reaction (enzyme:DNA = 50 units/μg). Four buffer conditions were used: 67 mM glycine-KOH pH 8.8 and pH 9.4 or 67 mM glycine-NaOH pH 8.8 and pH 9.4; all with 2.5 mM $MgCl_2$ and 50 μg/ml bovine serum albumin. The linearized plasmid was made single stranded, as necessary, by boiling for 5 minutes and transferring directly to ice. Reaction products were run out on 0.8% agarose, dried on a gel dryer and exposed to a phospho-imaging plate. In unlabeled plasmid experiments, about 700 ng of single stranded plasmid DNA and 20 units of λ-exo were used per reaction (enzyme:DNA = 28.6 units/μg). The digestion products were run on 0.8% agarose and stained with ethidium bromide. G4 deletion mutants of pFRT.myc6xERE were generated with the Q5 Site-directed Mutagenesis Kit (New England Biolabs) following the manufacturer's directions. Pu27 was deleted and replaced with a HindIII restriction site using the following primers:
oMycG4Pu27for: 5'-CTTATAAGCGCCCCTCCCGGG-3';
oMycG4Pu27rev: 5'-CTTGAGGAGACTCAGCCGGGC-3'.

Pu30 was deleted and replaced with a BamHI restriction site:
oMycG4Pu30for: 5'-TCCGTACAGACTGGCAGAGAG-3'
oMycG4Pu30rev: 5'-TCCACACGGAGTTCCCAATTTC-3'

D.3.2 Predicting G4s in the plasmid sequence.

The QGRS mapper [Kikin et al., 2006] (`http://bioinformatics.ramapo.edu/QGRS`) was used with default parameters to predict another G4 sequence (Pu30) in the pFRT.myc6xERE sequence. QGRS was used for this analysis as it offers the advantage of providing "G scores" for each G4 candidate, with higher scores belonging to candidates that are more likely to actually form G4s.

D.3.3 λ-exonuclease (λ-exo) digestion and sequencing of non-replicating DNA (LexoG0).

Three biological replicates were performed. For each, genomic DNA was purified from serum starved cells (9.6% S-phase) with 15 ml of DNAzol (Invitrogen) following the manufacturer's directions and resuspended in DNA hydration buffer (Qiagen). 150 μg of DNA was sonicated to a size range of 200 bp to 10 kb in a Biorupter Standard (Diagenode) and purified with Agencourt Ampure XP beads (Beckman Coulter). In order to investigate the genome-wide, nascent strand independent λ-exo biases in the genomic DNA it is important to avoid enriching the small amount of contaminating S-phase DNA that may be present. Fragmentation of the DNA by sonication breaks any long, RNA-primed nascent strands associated with replication forks into smaller fragments, ensuring that short RNA-protected fragments (if present) are distributed throughout the genome rather than only near origins, thus preventing origin sequences from being accidentally enriched. The fragmented DNA was made single stranded by boiling for 10 minutes, and transferring to ice. The 5' ends were phosphorylated with 50 units of T4 Polynucleotide Kinase (T4 PNK; New England Biolabs) for 1 hour at 37°C. The reaction was stopped by incubating 15 minutes at 75°C. Following phenol:chloroform extraction and ethanol precipitation, the phosphorylated fragments were digested with 100 units of λ-exo (Fermentas) in glycine-KOH pH 9.4 buffer in a total volume of 100 μl. The reaction was stopped by incubating 15 minutes at 75°C before the samples were phenol:chloroform extracted and ethanol precipitated. λ-exo digested fragments were electrophoresed on a 1.5% UltraPure LMP agarose (Invitrogen) gel. Fragments in the range of 500-1500 nt were then purified by melting at 65°C for 10 min before sequential extraction with phenol, phenol-chloroform, and chloroform, followed by resuspension in 10 μl elution buffer (Qiagen). The concentration was determined by NanoDrop (Thermo Scientific); the starting DNA samples were depleted approx. 1000-fold. The purified single stranded fragments were made double stranded with random hexamers and Klenow (New England Biolabs), then sonicated to a size of 100-600 bp. Illumina libraries were prepared using the NEBNext kit (New England Biolabs) following the manufacturer's directions. 200-500 bp library fragments were size selected on 2% NuSieve agarose (Lonza) and were gel-purified (Qiagen). Libraries were sequenced on the Illumina HiSeq 2000 platform.

D.3.4 Sequencing of undigested non-replicating genomic DNA (G0gDNA).

For the G0gDNA control, undigested genomic DNA from serum starved MCF7 cells (6.8% S-phase) was sonicated to a size range of 100-600 bp, and Illumina libraries were prepared using the NEBNext kit (New England Biolabs) following the manufacturer's directions. 200-500 bp library fragments were size selected on 2% NuSieve agarose (Lonza) and were gel purified (Qiagen). Libraries were sequenced on the Illumina HiSeq 2000 platform.

D.3.5 λ-exonuclease (λ-exo) digestion and sequencing of replicating DNA: Nascent-strand sequencing (NS-seq).

Three biological replicates were performed. For each, genomic DNA was purified from asynchronously growing MCF7 cells (35-40% S-phase) with 15 ml of DNAzol (Invitrogen) following the manufacturer's directions and resuspended in DNA hydration buffer (Qiagen). Nascent strands were prepared by adapting the protocol developed for replication initiation point mapping [Gerbi and Bielinsky, 1997] for NS-seq. DNA was handled gently to prevent breakage of long RNA-primed nascent DNA throughout the entire preparation in order to keep short RNA-primed nascent DNA close and specific to origins (in contrast to LexoG0 where purposeful fragmentation was performed). Replicative Intermediate (RI) DNA was enriched from 150 μg of genomic DNA by BND-cellulose chromatography (Sigma). Typically, 40 to 50 μg (approx. 25-30%) of starting material was recovered. The RI DNA was made single stranded by boiling for 10 minutes and transferring to ice. The 5' ends were phosphorylated with 50 units of T4 PNK for 1 hour at 37°C and the reaction was stopped by incubating 15 minutes at 75°C. The fragments were digested with 100 units of λ-exo in glycine-KOH pH 8.8 buffer in a total volume of 100 μl. The enzyme:DNA ratio was kept low (2-2.5 units/μg DNA) to preserve the nascent strands because λ-exo can lose specificity at high enzyme:DNA ratios and digest the RNA primer at the 5' end of DNA [Yang et al., 2013]. λ-exo digested fragments were electrophoresed on a 1.5% UltraPure LMP agarose gel and fragments in the range of 500-1500 nt were purified and resuspended in Qiagen elution buffer. The concentration was determined by Nanodrop; 21-96 ng of DNA was recovered for the replicates reported here, representing an approximately 500-2500 fold depletion of the starting DNA. Nascent strand enrichment was determined by qPCR at the MYC locus using the following primers:

Control locus primers:
oMyc RT set 1-2 fwd, 5'-TTGCCAATTGCCTCTGGTTGAGAC-3';
oMyc RT set 1-2 rev, 5'-GACTTTGCTGTTTGCTGTCAGGCT-3';

Test locus primers:
oMyc RT set 16-2 fwd, 5'- TGAACCAGAGTTTCATCTGCGACC-3';
oMyc RT set 16-2 rev, 5'- AGAAGCCGCTCCACATACAGTCCT-3'.

Sequencing libraries were made from nascent strand preparations where the MYC origin was >60-fold enriched. Single stranded nascent strands were made double stranded with random hexamers and Klenow and sonicated to a size of 100-600 bp. Illumina libraries were prepared using the NEBNext kit following the manufacturer's directions. 200-500 bp library fragments were size selected on 2% NuSieve agarose, gel purified and sequenced on the Illumina HiSeq 2000.

Although other studies used higher enzyme:DNA ratios, we kept the ratio lower to preserve RNA-primed DNA [Yang et al., 2013]. Nonetheless, that there was only 21-96 ng of enriched DNA at the end of the preparations (up to 2500-fold depletion of the starting DNA) and that the MYC

origin was enriched >60-fold in each replicate indicates that the amount of λ-exo was sufficient.

D.3.6 Mapping and manipulating reads.

Fasta files of human genome build hg19 were downloaded from the UCSC Genome Browser (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz`) [Lander et al., 2001, Kent et al., 2002,Karolchik et al., 2004,Kent et al., 2010]. A Bowtie2 index was made with 'bowtie2-build -f hg19.fa hg19' [Langmead and Salzberg, 2012]. Illumina reads in fastq format were mapped to the hg19 Bowtie2 index, using the parameters "--very-sensitive -N 1". The SAM format output of Bowtie2 was piped into SAMtools [Li et al., 2009] to retain only reads that mapped to hg19 and converted to BAM format with "samtools view -F 4 -bS". "samtools sort" was used to sort the BAM files. In cases where BAM files of reads from replicates needed to be merged, "samtools merge" was used. See Table D.1 for hg19 mapping statistics.

D.3.7 GC content in mappable reads.

GC content in mappable reads was obtained with a custom Python script (`https://github.com/JohnUrban/LexoNSseq2015`) that collects this information from SAM files. Only mappable reads with 50 unambiguous bases were used (no N content) for calculating GC content of reads. Briefly, the Python script counted the number of G and C bases in each 50 bp read and reported how many reads had each GC count from 0-50 (i.e. a histogram of numberGC vs. numberReads). This histogram information was brought into R where GC counts (0-50) were turned into percents (0-100) by 100*GCcount/50 (where 50 is the read length) and the number of reads with each GC count in a given dataset was normalized by the total number of reads summed over all GC counts in that dataset (i.e. percentGC vs. proportionOfReads): numberReadsWithGCcount/totalNumberReads. The normalized distributions of GC content in LexoG0 (Figure 6.2A) and NS-seq (Figure 6.2B) reads for each replicate were plotted in R as the log2(fold change) compared to the normalized distribution of GC content in G0 genomic DNA reads – i.e. log2(NS-seq/gDNA) and log2(LexoG0/gDNA). This was also done for NS-seq reads relative to LexoG0 reads (Figure S2).

D.3.8 FRiT scores.

For FRiT scores (Fraction of Reads in Telomeres), mappable reads in BAM files were converted back to fastq files with SamToFastq.jar from Picard Tools (`http://picard.sourceforge.net`). The fastq files of mappable reads were re-mapped (using the same Bowtie2 parameters as above) to a model telomere sequence composed of 1000 human telomere repeats (TTAGGG). The number of telomere-mappable reads was then divided by the total number of input mappable reads for normalization and multiplied by one million to get the number of hits per million reads (i.e. the FRiT score).

D.3.9 G4-CPMR and G4-Start-Site-CPMR.

G4 motifs in hg19 mappable reads were identified and counted (for G4 CPMR where CPMR is counts per million reads) using a Python script modified from Dario Beraldi's quadparser.py (`http://bioinformatics-misc.googlecode.com/svnhistory/r16/trunk/quadparser.py`) to analyze only the forward strand of reads in fastq files (`https://github.com/JohnUrban/LexoNSseq2015`). For both G4 CPMR and G4-start-site CPMR scores, we only considered the original read sequences (forward strands), not their reverse complements, as the read sequences represent 5' ends of fragments that λ-exo may have encountered (whereas reverse complements represent 3' ends of fragments). This is accomplished by searching only for "([gG]{3,}\w{1,7}){3,}[gG]{3,}" (the Python regular expression for $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$) and not for "([cC]3,\w{1,7}){3,}[cC]{3,}" ($C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}$), which identifies G4s on the opposite strand. Hg19 mappable reads were converted back to fastq format with SamToFastq.jar from Picard Tools (`http://picard.sourceforge.net`) with the specification to return the original forward strand sequences for all reads. The number of G4 motifs in each fastq file of mappable reads was counted with the Python script (G4 counts), then divided by the total number of input mappable reads (for the given sample) to normalize and multiplied by one million to get the G4-CPMRs. The Python script was also used to keep track of which position each G4 motif started on in order to get G4 start site counts over each position of the reads (which represent the 5' ends of fragments). To get G4-start-site-CPMRs, the start site count for each position was divided by the total number of input mappable reads to normalize and multiplied by one million. Note that when the G4-start-site-CPMR is summed up over all positions, it is equal to the G4-CPMR.

D.3.10 rDNA locus profiling.

For profiling signals over the rDNA repeat, the fastq files of raw reads from the HiSeq2000 were mapped with Bowtie2 [Langmead and Salzberg, 2012] using the same parameters as above to a modified version of hg19, referred to here as hg19+rDNA, that contained a copy of the 43 kb rDNA repeat (`http://www.ncbi.nlm.nih.gov/nuccore/555853?report=fasta`) as an additional "chromosome". Only mappable reads were retained in BAM format by piping the Bowtie2 output into "samtools view -F 4 bS -" [Li et al., 2009]. Mapping reads to the rDNA repeat in the context of hg19 was done to ensure that reads that would map elsewhere in the genome with higher alignment scores were not forced to map to the rDNA, as was performed in a recent paper studying the chromatin landscape of the rDNA repeat [Zentner et al., 2011]. Reads that mapped to the rDNA repeat with higher alignment scores than elsewhere in hg19 were extracted with SAMtools specifying '-q 2' and the name of the rDNA chromosome. Since the human genome contains >400 copies of the rDNA repeat, there was very high read depth coverage over each bp. To reduce possible spurious effects of PCR biases, "macs2 filterdup" was used with the 'auto' option on the extracted rDNA reads, which allowed the binomial distribution to determine how many reads can pileup at the same position on the same strand given the length of the repeat (approx. 43 kb) and number of reads

mapped to it with a p-value of 0.00001. The BED format results of macs2 filterdup were piped into BEDTools "sortBed" [Quinlan and Hall, 2010] to sort and then into BEDTools bedToBam to convert back into BAM format. It is noteworthy that the rDNA fold enrichment results (in Figure 6.5B) were extremely robust and similar with and without any read filtering steps. The depth over each bp of the 43 kb rDNA repeat was obtained using BEDTools "genomeCoverageBed" with "-d" set and was normalized by the number of reads (in millions) that mapped to hg19+rDNA for the given sample to give the signal per million mapped reads (SPMR) over the rDNA locus for that sample. The depth files containing SPMR information were taken into R and plotted. Fold enrichments were taken over each position and fold enrichment trends were obtained by loess smoothing the fold enrichment signal (span=0.05). G4 motifs were mapped strand-specifically across the rDNA locus using our customized quadparser Python script (`https://github.com/JohnUrban/LexoNSseq2015`). The position information was taken into R, and strand-specific G4 counts were taken in 1 kb bins across the locus for visualization with the FE plots. For %GC signal across the rDNA repeat, BEDTools "makewindows" was used with "-w 5 -s 1" to create 5 bp sliding windows (incremented by 1 bp) across the rDNA locus. BEDTools "nucBed" was used to obtain the %GC in each window and this score was assigned to the middle bp in the 5 bp window. This raw %GC signal was brought into R and loess smoothed (span=0.05) before plotting with the fold enrichment signals.

D.3.11 Genome-wide Peak Calling.

Genomic regions that were significantly enriched over a background control (called "peaks") were identified with MACS2 [Zhang et al., 2008]. To avoid calling low complexity peaks (e.g. regions with only one or a few positions with numerous reads), before peak calling each replicate of mapped reads was further filtered for redundant reads that mapped to the same location on the same strand (potential PCR artifacts) by keeping only one read per position with 'macs2 filterdup'. Each replicate of mappable reads was filtered individually before pooling instead of filtering the pooled set to avoid eliminating reads from separate replicates that independently align to the same position and, therefore, should not be treated as PCR artifacts in the pooled set. For peak calling, 'macs2 callpeak' was used with '--nomodel', which turns off the ChIP-seq specific model builder, '--keep-dup all' since redundant read filtering was already performed as a pre-processing step, and '--extsize=350', which MACS2 uses as an estimate of the average Illumina library fragment size and for smoothing. Peak set names below are in treatment$_{control}$ format consistent with the nomenclature in the paper. LexoG0$_{G0gDNA}$ and NS$_{G0gDNA}$ peaks were called relative to the undigested G0gDNA reads to control for amplicons or deletions present in the MCF7 genome and to control for any biases introduced during library construction and sequencing. Thus, LexoG0 or NS-seq was set as the treatment (-t) and G0gDNA as the control (-c). NS$_{LexoG0}$ peaks were called with NS-seq set as the treatment (-t) and LexoG0 as the control (-c) to control for nascent strand independent biases of λ-exo, such as its %GC and G4 biases, in addition to any amplicons or deletions and biases introduced in the sequencing process. All peaks were called with '--downsample' to use equivalent numbers of

reads between the treatment and control and to avoid the assumption of linearity introduced in downscaling. The local windows used to estimate local biases in the controls ('dynamic lambda') while scanning the genome for peaks were 5000 (--slocal) and 50000 (--llocal). These window sizes were chosen to cover the local region around a source of nascent strands (or G4-protected fragments), which we size selected up to 1500 bp, and to cover a region spanning the typical width of replication initiation zones. For all peak calling, we set a high stringency cutoff of $q < 0.001$ corresponding to a false discovery rate (FDR) of 0.1%. Since MCF7 is a female cell line, chrY data were excluded from all subsequent analyses. chrM (mitochondrial chromosome) was also removed from consideration. The output from MACS2 contains both the peak regions and peak summits (the bp of highest coverage inside a given peak region), which were each used for various analyses. It should be noted that the LexoG0 samples were designed first and foremost to characterize λ-exo biases in non-replicating cells (with undigested gDNA as the control). LexoG0 data were subsequently used to control these biases in NS-seq. It is possible that the LexoG0 control does not control for biases introduced by BND, if any BND biases exist and if they remain after the λ-exo digestion step. However, since the BND step only reduces the input DNA approximately 3-fold and the λ-exo-digestion step reduces the input up to 2500-fold, it is likely that BND biases are lost and overwritten by the strong λ-exo biases characterized in this paper. An alternative approach to LexoG0 could be to pass nonreplicating DNA through BND before λ-exo digestion. However, BND enrichment of non-replicating DNA recovers a much smaller amount of DNA leading to larger relative enrichments of select sites in the genome specific to non-replicating gDNA. Since this would create BND biases not in proportion to those in replicating DNA, it raises additional BND issues rather than alleviating the potential BND bias issue and therefore this alternative is not necessarily an improvement upon LexoG0. LexoG0 side steps these issues by improving upon the standard undigested G0gDNA control, which only corrects for copy number and biases introduced in library construction and sequencing. In contrast, controlling with λ-exo-digested G0gDNA (LexoG0) corrects for nascent strand independent λ-exo biases while also controlling for copy number and biases introduced during library construction and sequencing.

D.3.12 Shuffling peaks/features and computing %GC of peak sequences.

For analyses where peaks, peak summits, or other genomic features (e.g. G4 motifs) required shuffling throughout the genome, shuffleBed from BEDTools [Quinlan and Hall, 2010] was employed with the constraints that the peaks stay on the chromosome they start out on (-chrom), do not overlap after the shuffle (-noOverlapping), and were not shuffled into hg19 gap regions nor onto chromosomes Y and M (-excl). The shuffled features were piped into sortBed to sort before being written to file. Hg19 gap locations were obtained from the UCSC Table Browser [Kent et al., 2002, Karolchik et al., 2004, Kent et al., 2010]. The '.genome' file needed for this and some other BEDTools analyses was made with UCSC Kent Utilities Tool 'faSize -detailed' (`http://hgdownload.cse.ucsc.edu/admin/exe/`; [Kent et al., 2002, Karolchik et al., 2004, Kent et al., 2010]) on our copy of hg19.fa. For analyses interrogating the %GC in peaks and shuffled peaks (Figure 6.2C), "nucBed" from BEDTools was

used to obtain the %GC information for each feature in a BED file and those results were brought into R for visualization. In R, a histogram of the %GC scores (which range from 0-100) for a given BED file was made with breaks=seq(0,100,0.5). The resulting bin counts were then loess smoothed (span=0.075) over the bin midpoints before plotting to lightly smooth out jagged edges.

D.3.13 Overlap analyses.

For overlap analyses, 'intersectBed' from BEDTools was used [Quinlan and Hall, 2010]. To obtain the number of features in BED file A that overlapped features in BEDfile B, '-u' was set, file A was set to '-a', file B set to '-b', and the output was piped into 'wc -l'. Any feature in A that overlapped at least 1 feature in B by at least 1 bp was counted. To test for significance, we used a binomial model that conservatively estimates the upper-tailed p-value obtained if one did permutation tests into infinity. Briefly, the number of distinct positions that a feature from A can be shuffled onto in the genome is estimated as:

$$|\text{Total positions}| = G - C * (\mu_A - 1)$$

Where G is the size of the mappable genome, which for hg19 is 2.835679040e9, C is the number of contiguous sequence components (i.e. regions separated by gaps, of which there are 257 in hg19 when considering only chr 1-22 and chrX), and μ_A is the mean interval size of features in file A. The number of distinct "successful positions" a feature in A can be shuffled to (where success indicates overlap with a feature in B), the probability of success, the probability of seeing x overlaps, and the upper tailed p value were estimated as:

$$|\text{Successful positions}| = \min((\mu_A+\mu_B-1)*|B|, |\text{Total positions}|)$$

$$\text{Probability of success} = p = |\text{Successful positions}| \ / \ |\text{Total positions}|$$

$$P(X = x) = \binom{|A|}{x} p^x * (1-p)^{|A|-x}$$

$$Pvalue = \sum_{x=|obs|}^{|A|} P(X = x)$$

Where μ_A and μ_B are the mean interval sizes of features in file A and B respectively, |A| and |B| are the number of features in file A and B respectively, and |obs| is the observed number of overlaps of features in A with features in B. The expected number of overlaps, |exp|, is obtained by p*|A|. This estimate of the number of successful positions results in a conservative p-value estimate because it assumes that all peaks in B are capable of forming disjoint sets of successful positions. In other words, it assumes that a successful position as determined by an arbitrary feature b_i is not also a successful position as determined by another arbitrary feature, b_k. Often this assumption is true. In cases when it is not true, the probability of success (p) and, therefore, |exp| are both overestimated, which is conservative with respect to |obs|.

D.3.14 Features and feature densities across genome.

Feature density correlation analysis casts a wide net to see if the density of feature A in a genomic neighborhood is able to predict the density of feature B in that genomic neighborhood. When comparing feature set A to feature set B (eg. NS-seq peaks and CpGs), we chose to make bin sizes big enough such that feature counts in the bins for both A and B have a dynamic range and are not mostly zero counts. If one feature is numerous and the other is not, small bin sizes would result in mostly zeros for the rare feature and a range of counts for the other. This means the detectable correlation, if any, will be low at that level of resolution (bin size) due to the prevalence of zero counts for the rare feature. Using larger bin sizes allows both features to have a dynamic range of counts and allows the possibility to detect higher correlations if they exist, despite the lower resolution. Generally our analyses used 100 kb bins (for example, when comparing NS-seq peak counts with G4 motif counts; Figure 6.3 B-D) as was used for many similar analyses in a previously published NS-seq paper [Besnard et al., 2012], but it was more appropriate to use 1 Mb bins to explore correlations of peaks with CpG islands. There are relatively very few CpG islands compared to the size of the genome. When 100 kb bins are used, only approximately 40% of the bins have CpG islands in them and 60% have zero counts. In contrast, approximately 88% of the 1 Mb bins contain CpG islands and 100% contain NS-seq peaks, both with a dynamic range of counts. Thus, 1 Mb is an appropriate bin size for this particular analysis of peaks and CpG islands, despite the lower resolution. How close the CpG islands and NS-seq peaks (or other pairs of features) are to each other is the subject of other analyses such as direct overlap and proximity distributions.

To obtain feature (peaks, G4 motifs, CpG islands, etc) densities, defined as counts in 100 kb or 1 Mb bins, first BEDTools 'make windows' was used to partition hg19 into 100kb or 1Mb bins [Quinlan and Hall, 2010]. To eliminate noise from the analysis, the following bins were discarded: any bin smaller than the specified size, any bin that overlapped a gap, and any bin on chrM or chrY. To get the feature counts inside each retained bin, BEDTools "coverageBed" was used with the feature BED file as '-a', the genomic windows as '-b', and '-counts' set. Each resulting bedGraph file was sorted with sortBed to ensure that the counts in the same bins for different features were all in the same order and brought into R where they were subject to both Pearson and Spearman correlation tests (using cor()) and, in some cases, scatter plotted (peak sets vs G4 motifs in Figure 6.4D). For predicted G4 motif densities, G4 motifs were predicted with our Python implementation of quadparser (searching for $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ and $C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}$ to predict G4s on both strands). We also downloaded the predicted G4 motifs from the Non-B DataBase [Cer et al., 2013] (`http://nonb.abcc.ncifcrf.gov/apps/QueryGFF/feature/`) to compare to our set and found that it was identical. RefSeq genes and CpG island locations were downloaded from the UCSC Table Browser [Kent et al., 2002, Karolchik et al., 2004, Kent et al., 2010]). All peaks, density signals, fold enrichment signals, and –log10(p) signals across the genome or genomic stretches were visualized in the Integrative Genomics Viewer (IGV) [Robinson et al., 2011, Thorvaldsdóttir et al., 2013]. For example, Figure 6.4 B-C shows G4 density and peak density in 100 kb bins across chromosomes 3 and 6, respectively.

D.3.15 Profiling G4s within 1 kb around peak summits.

G4 positions were defined as the center position of each predicted G4 motif. The peak summits were identified by MACS2 [Zhang et al., 2008] as the bp of highest coverage inside each peak. "slopBed" from BEDTools [Quinlan and Hall, 2010] was used to extend the peak summits equal lengths (e.g. 1kb or 2kb) in each direction. The slopBed output was piped into "intersectBed -wb -a G4centers.bed -b -". The '-wb' flag instructs BEDTools to return the pair of entries that overlapped. Here that means that both the G4 center that overlapped a windowed peak summit and the windowed peak summit that was overlapped are returned on the same line. The windowed peak summits in the paired-entry BEDTools output were then converted back to single bp summit positions such that the paired information contained a peak summit and a G4 center within the window size. The resulting file was loaded into R for further analysis. In R, the start sites of the G4 centers were subtracted from the start sites of their corresponding peak summits. This returns G4 center distances from the peak summit between -1*windowSize to windowSize, with 0 representing the peak summit position. When not considering what strand the G4 is on: if a G4 center start site is to the right of the peak summit, then subtraction results in a positive distance between 1 and windowSize; if the G4 center is to the left of the peak summit, then subtraction results in a negative distance between -1*windowSize and -1; if the G4 center start site is the same position as the peak summit start site, it returns 0. To incorporate information about which strand the G4 was on such that any G4 5' to the peak summit produces a negative distance and any G4 3' to the peak summit produces a positive distance: distances for G4 motifs on the positive strand need no further correction, but the distances for G4 motifs on the negative strand need to be multiplied by -1. Thus, for G4 motifs that occur on the negative strand of the genome sequence: those that are to the right of the peak summit incur a negative distance; those to the left incur a positive distance; those that share the summit position remain as a distance of 0. The distances of G4 centers to peak summits were then counted and plotted. To test what G4 motif centers around peak summits would look like at random, the G4 motif locations were shuffled with shuffleBed (using parameters established above) and the same process was applied to the shuffled G4 motif centers. It was then possible to calculate the fold enrichment of the G4 counts near peak summits over the random distribution at each position (Test/Control1). The fold enrichment signal was loess smoothed to more clearly show the trend (span=0.1). As an additional control for the calculation of fold enrichment with the random distribution, G4 motif locations were shuffled with shuffleBed a second time for "control #2". Both randomized controls were then used together to calculate the fold enrichment at random (Control2/Control1), which is centered around 1-fold so long as the procedure works correctly. The crest-to-crest distances of the wave-like G4 enrichment signal around peak summits, were calculated by first using a custom R script to objectively identify crests with a standardized definition. Specifically, using the loess smoothed G4 count data around the peak summits, we required a G4 enrichment crest to have a higher smoothed count than the counts of at least 55 bp to each side and for the crest count to have a fold enrichment ≥ 1.5 over the count in that position when shuffled at random. A range of other window size values gives the same results

for NS_{G0gDNA} and NS_{LexoG0}. With crest positions identified, distances between crests could then be calculated. All plotting was done in R.

D.3.16 Prominence, CTR, and decomposition of the G4 enrichment signal around NS_{G0gDNA}.

Trough positions in the G4 enrichment signal around peak summits were identified in each G4 enrichment signal in a similar fashion to how crests were identified (above). Troughs for G4 enrichment signal around all NS_{LexoG0} summits, all NS_{G0gDNA} summits, and the subset of NS_{G0gDNA} summits that overlapped NS_{LexoG0} summit windows were identified by requiring that the smoothed count in a trough position be lower than the counts of at least 55 bp to each side, but lower than > 144 surrounding positions total, and have a fold enrichment of < 3. The G4 Fold Enrichment scores over crest and trough positions were collected and the means for each (crest_{mean}, trough_{mean}) were computed. Prominence of the crests, qualitatively defined as the amount that crests jut out above troughs was quantified by: $\text{crest}_{mean} - \text{trough}_{mean}$. The phasing of the G4 enrichment at the crests, qualitatively defined as how concentrated the signal is at crests (relative to troughs) was quantified as the crest-to-trough ratio (CTR): $\text{crest}_{mean}/\text{trough}_{mean}$. The NS_{G0gDNA} summits were partitioned into two subsets: one subset containing summits that overlapped (i.e. were inside of) NS_{LexoG0} summit windows (summit +/- 1 kb) and the other subset containing summits that did not overlap NS_{LexoG0} summit windows. These subsets are described as NS_{G0gDNA} summits represented in NS_{LexoG0} and NS_{G0gDNA} summits not represented in NS_{LexoG0}, respectively. The subsets were then treated individually as described in the section titled, "Profiling G4s within 1 kb around peak summits". Decomposition of the G4 enrichment signal around NS_{G0gDNA} summits into a stronger wave-like component (for those represented in NS_{LexoG0}) and a roughly uniform component (for those <u>not</u> represented in NS_{LexoG0}) was the result of analyzing the partition this way. All plotting was done in R.

D.3.17 Profiling nucleosome signal around peak summits.

Since the spacing of the crests of the waves of G4 enrichment around our peak summits was suggestive of nucleosome spacing, we also looked at the nuclesome signal around those summits for which G4s were nearby (within 1 kb to either side, see Table D.10). The available nucleosome data at UCSC [Kent et al., 2002, Karolchik et al., 2004, Kent et al., 2010] was downloaded (K562 and GM12878 cells) [Kundaje et al., 2012]. Peak summits were extended 1 kb to each side to produce 2001 bp summit windows (same as for the G4 analysis above). For each summit window, the nucleosome signal over each individual bp of the 2001 bp was obtained. Then the mean over each individual relative position (-1000 to 1000) around all the summits was calculated. Genome positions for which nucleosome signal was not available, represented as ".", were treated as missing data. In other words, means over each position were calculated only from the sum of available scores divided by the number of available scores (in contrast to treating all missing data as 0, which is

an invalid assumption). The same procedure was done after shuffling the peaks (shuffleBed) to obtain the genome-wide mean nucleosome scores over each position expected at random. In R, for each raw cell line signal (see Tables D.11,D.12,D.13), the ratio of the two means (the mean score for the test sample, μ_{test}, and the mean score for the shuffled sample, $\mu_{shuffle}$) at each position, j, in the 2001 bp window was plotted ($\mu_{test,j}/\mu_{shuffle,j}$). The cell line signals were lightly loess smoothed (span = 0.075) for plotting (colored lines in Figures 6.7 and D.7; smoothed cell line nucleosome signal in Tables D.11,D.12,D.13). For the mean signal between the 2 cell lines, the mean between the two raw cell line signals at each position was taken, $((\mu_{test,j,K562}/\mu_{shuffle,j,K562}) + (\mu_{test,j,GM12878}/\mu_{shuffle,j,GM12878}))/2$, before light loess smoothing (span = 0.075) (black lines in Figures 6.7 and D.7). The crests in the wave-like mean nucleosome signal between the two cell lines were identified similar to how crests were identified in the G4 signal around summits. Specifically, we required crest positions of the mean nucleosome signal between cell lines to have higher scores than > 50 bp to each side, but at least higher than 130 surrounding positions total, and to have minimum height difference (or greater) between the potential crest position and the lowest point within the left or right window in order to ignore positions that are arbitrarily higher than surrounding area. The subset of NS_{G0gDNA} summits with > 1 G4 within 1 kb, were further partitioned, as in the G4 analysis above, to two subsets with one containing all NS_{G0gDNA} summits that are represented in NS_{LexoG0} and the other containing all NS_{G0gDNA} summits not represented in NS_{LexoG0}. The raw and smoothed cell line nucleosome signals as well as the raw and smoothed mean nucleosome signal between cell lines around these two subsets of NS_{G0gDNA} summits were computed as described above. The decomposition of the nucleosome signal around NS_{G0gDNA} summits into a stronger wave-like component resembling the nucleosome signal around NS_{LexoG0} summits and a less wave-like component resembling the nucleosome signal around LexoG0_{G0gDNA} summits was the result of this partitioning process. All plotting, correlations, and "divergence" calculations were done in R.

Bibliography

[Abbot and Gerbi, 1981] Abbot, A. and Gerbi, S. A. (1981). Spermatogenesis in Sciara coprophila. II. Precocious chromosome orientation in meiosis II. *Chromosoma*, 83(19-27).

[Abdurashidova et al., 2000] Abdurashidova, G., Deganuto, M., Klima, R., Riva, S., Biamonti, G., Giacca, M., and Falaschi, A. (2000). Start sites of bidirectional DNA synthesis at the human lamin B2 origin. *Science (New York, N.Y.)*, 287(5460):2023–6.

[Adams et al., 2000] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V.,

Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). The genome sequence of Drosophila melanogaster. *Science (New York, N.Y.)*, 287(5461):2185–95.

[Aggarwal and Calvi, 2004] Aggarwal, B. D. and Calvi, B. R. (2004). Chromatin regulates origin activity in Drosophila follicle cells. *Nature*, 430(6997):372–6.

[Aladjem, 2007] Aladjem, M. I. (2007). Replication in context: dynamic regulation of DNA replication patterns in metazoans. *Nature reviews. Genetics*, 8(8):588–600.

[Aladjem and Fanning, 2004] Aladjem, M. I. and Fanning, E. (2004). The replicon revisited: an old model learns new tricks in metazoan chromosomes. *EMBO reports*, 5(7):686–91.

[Aladjem et al., 1998] Aladjem, M. I., Rodewald, L. W., Kolman, J. L., and Wahl, G. M. (1998). Genetic dissection of a mammalian replicator in the human beta-globin locus. *Science (New York, N.Y.)*, 281(5379):1005–9.

[Alexander et al., 2015] Alexander, J. L., Barrasa, M. I., and Orr-Weaver, T. L. (2015). Replication Fork Progression during Re-replication Requires the DNA Damage Checkpoint and Double-Strand Break Repair. *Current Biology*, 25(12):1654–1660.

[Altman and Fanning, 2001] Altman, A. L. and Fanning, E. (2001). The Chinese hamster dihydrofolate reductase replication origin beta is active at multiple ectopic chromosomal locations and requires specific DNA sequence elements for activity. *Molecular and cellular biology*, 21(4):1098–110.

[Altman and Fanning, 2004] Altman, A. L. and Fanning, E. (2004). Defined sequence modules and an architectural element cooperate to promote initiation at an ectopic mammalian chromosomal replication origin. *Molecular and cellular biology*, 24(10):4138–50.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

[Amabis and Amabis, 1984a] Amabis, D. C. and Amabis, J. (1984a). Effects of ecdysterone in polytene chromosomes of Trichosia pubescens. *Developmental Biology*, 102(1):1–9.

[Amabis and Amabis, 1984b] Amabis, D. C. and Amabis, J. (1984b). Hormonal control of gene amplification and transcription in the salivary gland chromosomes of Trichosia pubescens. *Developmental Biology*, 102(1):10–20.

[Amabis and Cabral, 1970] Amabis, J. M. and Cabral, D. (1970). RNA and DNA Puffs in Polytene Chromosomes of Rhynchosciara: Inhibition by Extirpation of Prothorax. *Science*, 169(3946).

[Amabis and Simrs, 1971] Amabis, J. M. and Simrs, L. C. G. (1971). Puff induction and regression in Rhynchosciara angelae by the method of salivary gland implantation. *Genetica*, 42(4):404–413.

[Ammar et al., 2015] Ammar, R., Paton, T. A., Torti, D., Shlien, A., and Bader, G. D. (2015). Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Research*, 4:17.

[Anglana et al., 2003] Anglana, M., Apiou, F., Bensimon, A., and Debatisse, M. (2003). Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell*, 114(3):385–94.

[Arias and Walter, 2005] Arias, E. E. and Walter, J. C. (2005). Replication-dependent destruction of Cdt1 limits DNA replication to a single round per cell cycle in Xenopus egg extracts. *Genes & Development*, 19(1):114–126.

[Asano and Wharton, 1999] Asano, M. and Wharton, R. P. (1999). E2F mediates developmental and cell cycle regulation of ORC1 in Drosophila. *The EMBO journal*, 18(9):2435–48.

[Ashburner, 1971] Ashburner, M. (1971). Induction of Puffs in Polytene Chromosomes of in vitro Cultured Salivary Glands of Drosophila melanogaster by Ecdysone and Ecdysone Analogues. *Nature, Published online: 14 April 1971; — doi:10.1038/10.1038/newbio230222a0*, 230(15):222.

[Ashburner, 1972] Ashburner, M. (1972). Patterns of puffing activity in the salivary gland chromosomes of Drosophila. *Chromosoma*, 38(3):255–281.

[Ashburner, 1973] Ashburner, M. (1973). Sequential gene activation by ecdysone in polytene chromosomes of Drosophila melanogaster: I. Dependence upon ecdysone concentration. *Developmental Biology*, 35(1):47–61.

[Ashburner, 1974] Ashburner, M. (1974). Sequential gene activation by ecdysone in polytene chromosomes of Drosophila melanogaster: II. The effects of inhibitors of protein synthesis. *Developmental Biology*, 39(1):141–157.

[Ashburner et al., 1990] Ashburner, M., Ashburner, M., Ashburner, M., Chihara, C., Meltzer, P., Richards, G., Becker, H.-J., Beermann, W., Burtis, K., Thummel, C., Jones, C., Karim, F., Hogness, D., Clever, U., Clever, U., Karlson, P., Peck, A., Segraves, W., Hogness, D., Thummel, C., Thummel, C., Burtis, K., and Hogness, D. (1990). Puffs, genes, and hormones revisited. *Cell*, 61(1):1–3.

[Ashburner and Richards, 1976] Ashburner, M. and Richards, G. (1976). Sequential gene activation by ecdysone in polytene chromosomes of Drosophila melanogaster,: III. Consequences of ecdysone withdrawal. *Developmental Biology*, 54(2):241–255.

[Ashton et al., 2014] Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., and O'Grady, J. (2014). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*, 33:296–300.

[Austin et al., 1999] Austin, R. J., Orr-Weaver, T. L., and Bell, S. P. (1999). Drosophila ORC specifically binds to ACE3, an origin of DNA replication control element. *Genes & development*, 13(20):2639–49.

[Aves, 2009] Aves, S. J. (2009). DNA Replication Initiation. In Vengrova, S. and Dalgaard, J. Z., editors, *DNA Replication: Methods and Protocols*, pages 1–16. Humana Press, New York.

[Baker et al., 2012] Baker, A., Audit, B., Chen, C.-L., Moindrot, B., Leleu, A., Guilbaud, G., Rappailles, A., Vaillant, C., Goldar, A., Mongelard, F., D'Aubenton-Carafa, Y., Hyrien, O., Thermes, C., and Arneodo, A. (2012). Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *PLoS computational biology*, 8(4):e1002443.

[Balbiani, 1881] Balbiani, E. (1881). Sur la structure du noyau des cellules salivaires chez les larves de Chironomus. *Zool. Anz.*, 4:637–641.

[Bandura et al., 2005] Bandura, J. L., Beall, E. L., Bell, M., Silver, H. R., Botchan, M. R., and Calvi, B. R. (2005). humpty dumpty is required for developmental DNA amplification and cell proliferation in Drosophila. *Current biology : CB*, 15(8):755–9.

[Bankevich et al., 2012] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477.

[Bartholdy et al., 2015] Bartholdy, B., Mukhopadhyay, R., Lajugie, J., Aladjem, M. I., and Bouhassira, E. E. (2015). Allele-specific analysis of DNA replication origins in mammalian cells. *Nature Communications*, 6:7051.

[Bashir et al., 2012] Bashir, A., Klammer, A. A., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., Sebra, R., Sorenson, J., Bullard, J., Yen, J., Valdovino, M., Mollova, E., Luong, K., Lin, S., LaMay, B., Joshi, A., Rowe, L., Frace, M., Tarr, C. L., Turnsek, M., Davis, B. M., Kasarskis, A., Mekalanos, J. J., Waldor, M. K., and Schadt, E. E. (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology*, 30(7):701–707.

[Basso et al., 2002] Basso, L., Vasconcelos, C., Fontes, A., Hartfelder, K., Silva, J., Coelho, P., Monesi, N., and Paçó-Larson, M. (2002). The induction of DNA puff BhC4-1 gene is a late response to the increase in 20-hydroxyecdysone titers in last instar dipteran larvae. *Mechanisms of Development*, 110(1):15–26.

[Beall et al., 2004] Beall, E. L., Bell, M., Georlette, D., and Botchan, M. R. (2004). Dm-myb mutant lethality in Drosophila is dependent upon mip130: positive and negative regulation of DNA replication. *Genes & development*, 18(14):1667–80.

[Beall et al., 2002] Beall, E. L., Manak, J. R., Zhou, S., Bell, M., Lipsick, J. S., and Botchan, M. R. (2002). Role for a Drosophila Myb-containing protein complex in site-specific DNA replication. *Nature*, 420(6917):833–837.

[Been and Rasch, 1972] Been, A. C. and Rasch, E. M. (1972). Cellular and secretory proteins of the salivary glands of Sciara coprophila during the larval-pupal transformation. *The Journal of Cell Biology*, 55(2).

[Bell and Dutta, 2002] Bell, S. P. and Dutta, A. (2002). DNA replication in eukaryotic cells. *Annual review of biochemistry*, 71:333–74.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (.*

[Bensimon et al., 1994] Bensimon, A., Simon, A., Chiffaudel, A., Croquette, V., Heslot, F., and Bensimon, D. (1994). Alignment and sensitive detection of DNA by a moving interface. *Science (New York, N.Y.)*, 265(5181):2096–8.

[Berbenetz et al., 2010] Berbenetz, N. M., Nislow, C., and Brown, G. W. (2010). Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS genetics*, 6(9):e1001092.

[Berlin et al., 2015] Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., and Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, 33(6):623–630.

[Besnard et al., 2012] Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., and Lemaitre, J.-M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature structural & molecular biology*, 19(8):837–44.

[Bianco et al., 2012] Bianco, J. N., Poli, J., Saksouk, J., Bacal, J., Silva, M. J., Yoshida, K., Lin, Y.-L., Tourrière, H., Lengronne, A., and Pasero, P. (2012). Analysis of DNA replication profiles in budding yeast and mammalian cells using DNA combing. *Methods (San Diego, Calif.)*, 57(2):149–57.

[Bielinsky et al., 2001] Bielinsky, A. K., Blitzblau, H., Beall, E. L., Ezrokhi, M., Smith, H. S., Botchan, M. R., and Gerbi, S. A. (2001). Origin recognition complex binding to a metazoan replication origin. *Current biology : CB*, 11(18):1427–31.

[Bielinsky and Gerbi, 1998] Bielinsky, A. K. and Gerbi, S. A. (1998). Discrete start sites for DNA synthesis in the yeast ARS1 origin. *Science (New York, N.Y.)*, 279(5347):95–8.

[Bielinsky and Gerbi, 1999] Bielinsky, A. K. and Gerbi, S. A. (1999). Chromosomal ARS1 has a single leading strand start site. *Molecular cell*, 3(4):477–86.

[Bielinsky and Gerbi, 2001] Bielinsky, A. K. and Gerbi, S. A. (2001). Where it all starts: eukaryotic origins of DNA replication. *Journal of cell science*, 114(Pt 4):643–51.

[Bienz-Tadmor et al., 1991] Bienz-Tadmor, B., Smith, H. S., and Gerbi, S. A. (1991). The promoter of DNA puff gene II/9-1 of Sciara coprophila is inducible by ecdysone in late prepupal salivary glands of Drosophila melanogaster. *Cell regulation*, 2(11):875–88.

[Biffi et al., 2013] Biffi, G., Tannahill, D., McCafferty, J., and Balasubramanian, S. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. *Nature chemistry*, 5(3):182–6.

[Biggar and Li, 2014] Biggar, K. K. and Li, S. S.-C. (2014). Non-histone protein methylation as a regulator of cellular signalling and function. *Nature Reviews Molecular Cell Biology*, 16(1):5–17.

[Blow et al., 2011] Blow, J. J., Ge, X. Q., and Jackson, D. A. (2011). How dormant origins promote complete genome replication. *Trends in biochemical sciences*, 36(8):405–14.

[Blow and Gillespie, 2008] Blow, J. J. and Gillespie, P. J. (2008). Replication licensing and cancer–a fatal entanglement? *Nature reviews. Cancer*, 8(10):799–806.

[Blow et al., 2001] Blow, J. J., Gillespie, P. J., Francis, D., and Jackson, D. A. (2001). Replication origins in Xenopus egg extract Are 5-15 kilobases apart and are activated in clusters that fire at different times. *The Journal of cell biology*, 152(1):15–25.

[Bochman et al., 2012] Bochman, M. L., Paeschke, K., and Zakian, V. A. (2012). DNA secondary structures: stability and function of G-quadruplex structures. *Nature reviews. Genetics*, 13(11):770–80.

[Bolger et al., 2014] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–20.

[Bolisetty et al., 2015] Bolisetty, M., Rajadinakaran, G., and Graveley, B. (2015). Determining Exon Connectivity in Complex mRNAs by Nanopore Sequencing. *bioRxiv*.

[Borowiec and Schildkraut, 2011] Borowiec, J. A. and Schildkraut, C. L. (2011). Open sesame: activating dormant replication origins in the mouse immunoglobulin heavy chain (Igh) locus. *Current opinion in cell biology*, 23(3):284–92.

[Bosco et al., 2001] Bosco, G., Du, W., and Orr-Weaver, T. L. (2001). DNA replication control through interaction of E2F-RB and the origin recognition complex. *Nature cell biology*, 3(3):289–95.

[Boža et al., 2016] Boža, V., Brejová, B., and Vina, T. (2016). DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads. *arXiv*.

[Bozeman and Metz, 1949] Bozeman, M. L. and Metz, C. W. (1949). Further Studies on Sensitivity of Chromosomes to Irradiation at Different Meiotic Stages in OOcytes of Sciara. *Genetics*, 34(3):285–314.

[Bradnam et al., 2013] Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. A., Ganapathy, G., Gibbs, R. A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T. D., Paten, B., Paulo, O. S., Phillippy, A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D. S., Ruby, J. G., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., and Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10.

[Branton et al., 2008] Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V., Ramsey, J. M., Riehn, R., Soni, G. V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., and Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26(10):1146–1153.

[Breier et al., 2004] Breier, A. M., Chatterji, S., and Cozzarelli, N. R. (2004). Prediction of Saccharomyces cerevisiae replication origins. *Genome biology*, 5(4):R22.

[Breuer and Pavan, 1955] Breuer, M. E. and Pavan, C. (1955). Behavior of polytene chromosomes of rhynchosciara angelae at different stages of larval development. *Chromosoma*, 7(1):371–386.

[Brewer and Fangman, 1987] Brewer, B. J. and Fangman, W. L. (1987). The localization of replication origins on ARS plasmids in S. cerevisiae. *Cell*, 51(3):463–71.

[Brewer and Fangman, 1993] Brewer, B. J. and Fangman, W. L. (1993). Initiation at closely spaced replication origins in a yeast chromosome. *Science (New York, N.Y.)*, 262(5140):1728–31.

[Brewer and Fangman, 1994] Brewer, B. J. and Fangman, W. L. (1994). Initiation preference at a yeast origin of replication. *Proceedings of the National Academy of Sciences of the United States of America*, 91(8):3418–22.

[Brewer et al., 2015] Brewer, B. J., Payen, C., Di Rienzi, S. C., Higgins, M. M., Ong, G., Dunham, M. J., and Raghuraman, M. K. (2015). Origin-Dependent Inverted-Repeat Amplification: Tests of a Model for Inverted DNA Amplification. *PLOS Genetics*, 11(12):e1005699.

[Brewer et al., 2011] Brewer, B. J., Payen, C., Raghuraman, M. K., and Dunham, M. J. (2011). Origin-Dependent Inverted-Repeat Amplification: A Replication-Based Model for Generating Palindromic Amplicons. *PLoS Genetics*, 7(3):e1002016.

[Brooks and Hurley, 2010] Brooks, T. A. and Hurley, L. H. (2010). Targeting MYC Expression through G-Quadruplexes. *Genes & cancer*, 1(6):641–649.

[Burhans et al., 1986] Burhans, W. C., Selegue, J. E., and Heintz, N. H. (1986). Isolation of the origin of replication associated with the amplified Chinese hamster dihydrofolate reductase domain. *Proceedings of the National Academy of Sciences of the United States of America*, 83(20):7790–4.

[Burhans et al., 1990] Burhans, W. C., Vassilev, L. T., Caddle, M. S., Heintz, N. H., and DePamphilis, M. L. (1990). Identification of an origin of bidirectional DNA replication in mammalian chromosomes. *Cell*, 62(5):955–65.

[Burhans et al., 1991] Burhans, W. C., Vassilev, L. T., Wu, J., Sogo, J. M., Nallaseth, F. S., and DePamphilis, M. L. (1991). Emetine allows identification of origins of mammalian DNA replication by imbalanced DNA synthesis, not through conservative nucleosome segregation. *The EMBO journal*, 10(13):4351–60.

[Burke et al., 2001] Burke, T. W., Cook, J. G., Asano, M., and Nevins, J. R. (2001). Replication factors MCM2 and ORC1 interact with the histone acetyltransferase HBO1. *The Journal of biological chemistry*, 276(18):15397–408.

[Burton et al., 2013] Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31(12):1119–1125.

[Butler et al., 2008] Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, 18(5):810–20.

[Cadoret et al., 2008] Cadoret, J.-C., Meisch, F., Hassan-Zadeh, V., Luyten, I., Guillet, C., Duret, L., Quesneville, H., and Prioleau, M.-N. (2008). Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(41):15837–42.

[Calvi et al., 1998] Calvi, B. R., Lilly, M. A., and Spradling, A. C. (1998). Cell cycle control of chorion gene amplification. *Genes & development*, 12(5):734–44.

[Candido-Silva et al., 2015] Candido-Silva, J. A., Machado, M. C. R., Hartfelder, K. H., de Almeida, J. C., Paçó-Larson, M. L., and Monesi, N. (2015). Amplification and expression of a salivary gland DNA puff gene in the prothoracic gland of Bradysia hygida (Diptera: Sciaridae). *Journal of Insect Physiology*, 74:30–37.

[Cao et al., 2015] Cao, M. D., Ganesamoorthy, D., Elliott, A., Zhang, H., Cooper, M., and Coin, L. (2015). Real-time strain typing and analysis of antibiotic resistance potential using Nanopore MinION sequencing. *bioRxiv*.

[Capuano et al., 2014] Capuano, F., Mülleder, M., Kok, R., Blom, H. J., and Ralser, M. (2014). Cytosine DNA Methylation Is Found in Drosophila melanogaster but Absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Other Yeast Species. *Analytical Chemistry*, 86(8):3697–3702.

[Carminati et al., 1992] Carminati, J. L., Johnston, C. G., and Orr-Weaver, T. L. (1992). The Drosophila ACE3 chorion element autonomously induces amplification. *Molecular and cellular biology*, 12(5):2444–53.

[Cayirlioglu et al., 2001] Cayirlioglu, P., Bonnette, P. C., Dickson, M. R., and Duronio, R. J. (2001). Drosophila E2f2 promotes the conversion from genomic DNA replication to gene amplification in ovarian follicle cells. *Development (Cambridge, England)*, 128(24):5085–98.

[Cayirlioglu et al., 2003] Cayirlioglu, P., Ward, W. O., Silver Key, S. C., and Duronio, R. J. (2003). Transcriptional repressor functions of Drosophila E2F1 and E2F2 cooperate to inhibit genomic DNA synthesis in ovarian follicle cells. *Molecular and cellular biology*, 23(6):2123–34.

[Cayrou et al., 2015] Cayrou, C., Ballester, B., Peiffer, I., Fenouil, R., Coulombe, P., Andrau, J.-C., van Helden, J., and Méchali, M. (2015). The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome research*, 25(12):1873–85.

[Cayrou et al., 2012a] Cayrou, C., Coulombe, P., Puy, A., Rialle, S., Kaplan, N., Segal, E., and Méchali, M. (2012a). New insights into replication origin characteristics in metazoans. *Cell cycle (Georgetown, Tex.)*, 11(4):658–67.

[Cayrou et al., 2011] Cayrou, C., Coulombe, P., Vigneron, A., Stanojcic, S., Ganier, O., Peiffer, I., Rivals, E., Puy, A., Laurent-Chabalier, S., Desprat, R., and Méchali, M. (2011). Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome research*, 21(9):1438–49.

[Cayrou et al., 2012b] Cayrou, C., Grégoire, D., Coulombe, P., Danis, E., and Méchali, M. (2012b). Genome-scale identification of active DNA replication origins. *Methods (San Diego, Calif.)*, 57(2):158–64.

[Cer et al., 2013] Cer, R. Z., Donohue, D. E., Mudunuri, U. S., Temiz, N. A., Loss, M. A., Starner, N. J., Halusa, G. N., Volfovsky, N., Yi, M., Luke, B. T., Bacolla, A., Collins, J. R., and Stephens,

R. M. (2013). Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic acids research*, 41(Database issue):D94–D100.

[Chakraborty et al., 2011] Chakraborty, A., Shen, Z., and Prasanth, S. G. (2011). "ORCanization" on heterochromatin: linking DNA replication initiation to chromatin organization. *Epigenetics : official journal of the DNA Methylation Society*, 6(6):665–70.

[Chakraborty et al., 2016] Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., and Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic acids research*, 44(19):e147.

[Chambers et al., 2015] Chambers, V. S., Marsico, G., Boutell, J. M., Di Antonio, M., Smith, G. P., and Balasubramanian, S. (2015). High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nature Biotechnology*, 33(8):877–881.

[Chapman and Johnston, 1989] Chapman, J. and Johnston, L. (1989). The yeast gene, DBF4, essential for entry into S phase is cell cycle regulated. *Experimental Cell Research*, 180(2):419–428.

[Check Hayden, 2012] Check Hayden, E. (2012). Nanopore genome sequencer makes its debut. *Nature*.

[Check Hayden, 2014] Check Hayden, E. (2014). Data from pocket-sized genome sequencer unveiled. *Nature*.

[Check Hayden, 2015] Check Hayden, E. (2015). Pint-sized DNA sequencer impresses first users. *Nature*, 521(7550):15–16.

[Chen et al., 2010] Chen, C.-L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., D'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O., and Thermes, C. (2010). Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome research*, 20(4):447–57.

[Chen and Bell, 2011] Chen, S. and Bell, S. P. (2011). CDK prevents Mcm2-7 helicase loading by inhibiting Cdt1 interaction with Orc6. *Genes & development*, 25(4):363–72.

[Chesnokov, 2007] Chesnokov, I. N. (2007). Multiple functions of the origin recognition complex. *International review of cytology*, 256:69–109.

[Chin et al., 2013] Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., and Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods*, 10(6):563–9.

[Chin et al., 2016] Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., and Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*.

[Clark et al., 2013] Clark, S. C., Egan, R., Frazier, P. I., and Wang, Z. (2013). ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics (Oxford, England)*, 29(4):435–43.

[Clark et al., 2012] Clark, T. A., Murray, I. A., Morgan, R. D., Kislyuk, A. O., Spittle, K. E., Boitano, M., Fomenkov, A., Roberts, R. J., and Korlach, J. (2012). Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic acids research*, 40(4):e29.

[Claycomb et al., 2004] Claycomb, J. M., Benasutti, M., Bosco, G., Fenger, D. D., and Orr-Weaver, T. L. (2004). Gene amplification as a developmental strategy: isolation of two developmental amplicons in Drosophila. *Developmental cell*, 6(1):145–55.

[Claycomb et al., 2002] Claycomb, J. M., MacAlpine, D. M., Evans, J. G., Bell, S. P., and Orr-Weaver, T. L. (2002). Visualization of replication initiation and elongation in Drosophila. *The Journal of cell biology*, 159(2):225–36.

[Claycomb and Orr-Weaver, 2005] Claycomb, J. M. and Orr-Weaver, T. L. (2005). Developmental gene amplification: insights into DNA replication and gene expression. *Trends in genetics : TIG*, 21(3):149–62.

[Clever and Karlson, 1960] Clever, U. and Karlson, P. (1960). Induktion von puff-veränderungen in den speicheldrüsenchromosomen von Chironomus tentans durch Ecdyson. *Experimental Cell Research*, 20(3):623–626.

[Coffman et al., 1993] Coffman, F. D., Georgoff, I., Fresa, K. L., Sylvester, J., Gonzalez, I., and Cohen, S. (1993). In vitro replication of plasmids containing human ribosomal gene sequences: origin localization and dependence on an aprotinin-binding cytosolic protein. *Experimental cell research*, 209(1):123–32.

[Coffman et al., 2005] Coffman, F. D., He, M., Diaz, M.-L., and Cohen, S. (2005). DNA replication initiates at different sites in early and late S phase within human ribosomal RNA genes. *Cell cycle (Georgetown, Tex.)*, 4(9):1223–6.

[Cohen et al., 2010] Cohen, S., Agmon, N., Sobol, O., and Segal, D. (2010). Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. *Mobile DNA*, 1(1):11.

[Cohen and Segal, 2009] Cohen, S. and Segal, D. (2009). Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. *Cytogenetic and genome research*, 124(3-4):327–38.

[Comoglio et al., 2015] Comoglio, F., Schlumpf, T., Schmid, V., Rohs, R., Beisel, C., and Paro, R. (2015). High-Resolution Profiling of Drosophila Replication Start Sites Reveals a DNA Shape and Chromatin Signature of Metazoan Origins. *Cell Reports*, 11(5):821–834.

[Conroy et al., 2010] Conroy, R. S., Koretsky, A. P., and Moreland, J. (2010). Lambda exonuclease digestion of CGG trinucleotide repeats. *European biophysics journal : EBJ*, 39(2):337–43.

[Costa et al., 2011] Costa, A., Ilves, I., Tamberg, N., Petojevic, T., Nogales, E., Botchan, M. R., and Berger, J. M. (2011). The structural basis for MCM2-7 helicase activation by GINS and Cdc45. *Nature structural & molecular biology*, 18(4):471–7.

[Crouse, 1943] Crouse, H. (1943). Translocations in Sciara: their bearing on chromosome behavior and sex determination. *Univ. Missouri Res. Bull.*, 379:1–75.

[Crouse, 1949] Crouse, H. (1949). The resistance of Sciara (Diptera) to the mutagenic effects of irradiation. *The Biological bulletin*, 97(3):311–4.

[Crouse, 1960a] Crouse, H. V. (1960a). The Controlling Element in Sex Chromosome Behavior in Sciara. *Genetics*, 45(10):1429–43.

[Crouse, 1960b] Crouse, H. V. (1960b). The nature of the influence of X-translocations on sex of progeny in Sciara coprophila. *Chromosoma*, 11(1):146–166.

[Crouse, 1968] Crouse, H. V. (1968). The role of ecdysone in DNA-puff formation and DNA synthesis in the polytene chromosomes of Sciara coprophila. *Proceedings of the National Academy of Sciences of the United States of America*, 61(3):971–8.

[Crouse et al., 1971] Crouse, H. V., Brown, A., and Mumford, B. C. (1971). –chromosome inheritance and the porblem of chromosome "imprinting" in Sciara (Sciaridae, Diptera). *Chromosoma*, 34(3):324–39 8.

[Crouse et al., 1977] Crouse, H. V., Gerbi, S. A., Liang, C. M., Magnus, L., and Mercer, I. M. (1977). Localization of ribosomal DNA within the proximal X heterochromatin of Sciara coprophila (Diptera, Sciaridae). *Chromosoma*, 64(4):305–318.

[Crouse and Keyl, 1968] Crouse, H. V. and Keyl, H. G. (1968). Extra replications in the "DNA-puffs" of Sciara coprophila. *Chromosoma*, 25(3):357–64.

[Cunningham et al., 2015] Cunningham, C. B., Ji, L., Wiberg, R. A. W., Shelton, J., McKinney, E. C., Parker, D. J., Meagher, R. B., Benowitz, K. M., Roy-Zokan, E. M., Ritchie, M. G., Brown, S. J., Schmitz, R. J., and Moore, A. J. (2015). The genome and methylome of a beetle with complex social behavior, Nicrophorus vespilloides (Coleoptera: Silphidae). *Genome biology and evolution*, pages evv194–.

[Dale et al., 2011] Dale, R. K., Pedersen, B. S., and Quinlan, A. R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics (Oxford, England)*, 27(24):3423–4.

[Danis et al., 2004] Danis, E., Brodolin, K., Menut, S., Maiorano, D., Girard-Reydet, C., and Méchali, M. (2004). Specification of a DNA replication origin by a transcription complex. *Nature cell biology*, 6(8):721–30.

[Das-Bradoo and Bielinsky, 2009] Das-Bradoo, S. and Bielinsky, A.-K. (2009). Replication initiation point mapping: approach and implications. *Methods in molecular biology (Clifton, N.J.)*, 521:105–20.

[David et al., 2016] David, M., Dursi, L. J., Yao, D., Boutros, P. C., and Simpson, J. T. (2016). Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics*, page btw569.

[Davidson et al., 2006] Davidson, I. F., Li, A., and Blow, J. J. (2006). Deregulated replication licensing causes DNA fragmentation consistent with head-to-tail fork collision. *Molecular cell*, 24(3):433–43.

[De Carli et al., 2016] De Carli, F., Gaggioli, V., Millot, G. A., and Hyrien, O. (2016). Single-molecule, antibody-free fluorescent visualisation of replication tracts along barcoded DNA molecules. *The International journal of developmental biology*.

[de Cicco and Spradling, 1984] de Cicco, D. V. and Spradling, A. C. (1984). Localization of a cis-acting element responsible for the developmentally regulated amplification of Drosophila chorion genes. *Cell*, 38(1):45–54.

[de Saint Phalle and Sullivan, 1996] de Saint Phalle, B. and Sullivan, W. (1996). Incomplete sister chromatid separation is the mechanism of programmed chromosome elimination during early Sciara coprophila embryogenesis. *Development (Cambridge, England)*, 122(12):3775–84.

[Deamer et al., 2016] Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524.

[Delgado et al., 1998] Delgado, S., Gómez, M., Bird, A., and Antequera, F. (1998). Initiation of DNA replication at CpG islands in mammalian chromosomes. *The EMBO journal*, 17(8):2426–35.

[Delidakis and Kafatos, 1989] Delidakis, C. and Kafatos, F. C. (1989). Amplification enhancers and replication origins in the autosomal chorion gene cluster of Drosophila. *The EMBO journal*, 8(3):891–901.

[Dellino et al., 2013] Dellino, G. I., Cittaro, D., Piccioni, R., Luzi, L., Banfi, S., Segalla, S., Cesaroni, M., Mendoza-Maldonado, R., Giacca, M., and Pelicci, P. G. (2013). Genome-wide mapping of

human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome research*, 23(1):1–11.

[DePamphilis and Bell, 2010] DePamphilis, M. and Bell, S. (2010). *Genome Duplication*. Garland Science, New York.

[DePamphilis, 1993] DePamphilis, M. L. (1993). Eukaryotic DNA replication: anatomy of an origin. *Annual review of biochemistry*, 62:29–63.

[DePamphilis, 1997] DePamphilis, M. L. (1997). The search for origins of DNA replication. *Methods (San Diego, Calif.)*, 13(3):211–9.

[DePamphilis et al., 2006] DePamphilis, M. L., Blow, J. J., Ghosh, S., Saha, T., Noguchi, K., and Vassilev, A. (2006). Regulating the licensing of DNA replication origins in metazoa. *Current Opinion in Cell Biology*, 18(3):231–239.

[DePamphilis ML, 2006] DePamphilis ML (2006). *DNA Replication and Human Disease*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

[Deshpande et al., 2013] Deshpande, V., Fung, E. D., Pham, S., and Bafna, V. (2013). Cerulean: A hybrid assembly using high throughput short and long reads. *arXiv*.

[Dhar et al., 2012] Dhar, M. K., Sehgal, S., and Kaul, S. (2012). Structure, replication efficiency and fragility of yeast ARS elements. *Research in microbiology*, 163(4):243–53.

[DiBartolomeis and Gerbi, 1989] DiBartolomeis, S. M. and Gerbi, S. A. (1989). Molecular characterization of DNA puff II/9A genes in Sciara coprophila. *Journal of molecular biology*, 210(3):531–40.

[Diffley, 2010] Diffley, J. F. X. (2010). The many faces of redundancy in DNA replication control. *Cold Spring Harbor symposia on quantitative biology*, 75:135–42.

[Dijkwel and Hamlin, 1995] Dijkwel, P. A. and Hamlin, J. L. (1995). The Chinese hamster dihydrofolate reductase origin consists of multiple potential nascent-strand start sites. *Molecular and cellular biology*, 15(6):3023–31.

[Dijkwel et al., 2002] Dijkwel, P. A., Wang, S., and Hamlin, J. L. (2002). Initiation sites are distributed at frequent intervals in the Chinese hamster dihydrofolate reductase origin of replication but are used with very different efficiencies. *Molecular and cellular biology*, 22(9):3053–65.

[Dimitrova, 2011] Dimitrova, D. S. (2011). DNA replication initiation patterns and spatial dynamics of the human ribosomal RNA gene loci. *Journal of cell science*, 124(Pt 16):2743–52.

[Dlaska et al., 2008] Dlaska, M., Anderl, C., Eisterer, W., and Bechter, O. E. (2008). Detection of circular telomeric DNA without 2D gel electrophoresis. *DNA and cell biology*, 27(9):489–96.

[Dolezel et al., 2003] Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, 51(2):127–8; author reply 129.

[Dorn et al., 2009] Dorn, E. S., Chastain, P. D., Hall, J. R., and Cook, J. G. (2009). Analysis of re-replication from deregulated origin licensing by DNA fiber spreading. *Nucleic acids research*, 37(1):60–9.

[Dorn and Cook, 2011] Dorn, E. S. and Cook, J. G. (2011). Nucleosomes in the neighborhood: new roles for chromatin modifications in replication origin control. *Epigenetics : official journal of the DNA Methylation Society*, 6(5):552–9.

[Doyle and Metz, 1935a] Doyle, W. L. and Metz, C. (1935a). Structure of the chromosomes in the salivary gland cells in Sciara (Diptera). *The Biological Bulletin*, 69(1):126–135.

[Doyle and Metz, 1935b] Doyle, W. L. and Metz, C. W. (1935b). Observations on the Structure of Living Salivary Gland Chromosomes in Sciara. *Proceedings of the National Academy of Sciences of the United States of America*, 21(2):75–8.

[Drosopoulos et al., 2012] Drosopoulos, W. C., Kosiyatrakul, S. T., Yan, Z., Calderano, S. G., and Schildkraut, C. L. (2012). Human telomeres replicate using chromosome-specific, rather than universal, replication programs. *The Journal of cell biology*, 197(2):253–66.

[DuBois, 1933] DuBois, A. (1933). Chromosome behavior during cleavage in the eggs of Sciara coprophila (Diptera) in the relation to the problem of sex determination. *Z. Wiss. Biol. Abt B - Z. Zellforsch. Mikrosk. Anat.*, 19:595–614.

[Dubois, 1932] Dubois, A. M. (1932). Elimination of Chromosomes during Cleavage in the Eggs of Sciara (Diptera). *Proceedings of the National Academy of Sciences of the United States of America*, 18(5):352–6.

[Duncker et al., 2009] Duncker, B. P., Chesnokov, I. N., and McConkey, B. J. (2009). The origin recognition complex protein family. *Genome biology*, 10(3):214.

[Durbin et al., 1998] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

[Dutta and Bell, 1997] Dutta, A. and Bell, S. P. (1997). Initiation of DNA replication in eukaryotic cells. *Annual review of cell and developmental biology*, 13:293–332.

[Eastman et al., 1980] Eastman, E. M., Goodman, R. M., Erlanger, B. F., and Miller, O. J. (1980). 5-Methylcytosine in the DNA of the polytene chromosomes of the dipteraSciara coprophila,Drosophila melanogaster andD. persimilis. *Chromosoma*, 79(2):225–239.

[Eaton et al., 2010] Eaton, M. L., Galani, K., Kang, S., Bell, S. P., and MacAlpine, D. M. (2010). Conserved nucleosome positioning defines replication origins. *Genes & development*, 24(8):748–53.

[Eid et al., 2009] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910).

[English et al., 2012] English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., and Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one*, 7(11):e47768.

[Ficq and Pavan, 1957] Ficq, A. and Pavan, C. (1957). Autoradiography of Polytene Chromosomes of Rhynchosciara angelae at Different Stages of Larval Development. *Nature*, 180(4593):983–984.

[Field et al., 2004] Field, L. M., Lyko, F., Mandrioli, M., and Prantera, G. (2004). DNA methylation in insects. *Insect Molecular Biology*, 13(2):109–115.

[Finn and Li, 2013] Finn, K. J. and Li, J. J. (2013). Single-Stranded Annealing Induced by Re-Initiation of Replication Origins Provides a Novel and Efficient Mechanism for Generating Copy Number Expansion via Non-Allelic Homologous Recombination. *PLoS Genetics*, 9(1):e1003192.

[Flusberg et al., 2010] Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., and Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 7(6):461–465.

[Foulk et al., 2006] Foulk, M. S., Liang, C., Wu, N., Blitzblau, H. G., Smith, H., Alam, D., Batra, M., and Gerbi, S. A. (2006). Ecdysone induces transcription and amplification in Sciara coprophila DNA puff II/9A. *Developmental biology*, 299(1):151–63.

[Foulk et al., 2015] Foulk, M. S., Urban, J. M., Casella, C., and Gerbi, S. A. (2015). Characterizing and controlling intrinsic biases of Lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. *Genome research*, pages gr.183848.114–.

[Foulk et al., 2013] Foulk, M. S., Waggener, J. M., Johnson, J. M., Yamamoto, Y., Liew, G. M., Urnov, F. D., Young, Y., Lee, G., Smith, H. S., and Gerbi, S. A. (2013). Isolation and characterization of the ecdysone receptor and its heterodimeric partner ultraspiracle through development in Sciara coprophila. *Chromosoma*, 122(1-2):103–19.

[Frum et al., 2008] Frum, R. A., Chastain, P. D., Qu, P., Cohen, S. M., and Kaufman, D. G. (2008). DNA replication in early S phase pauses near newly activated origins. *Cell cycle (Georgetown, Tex.)*, 7(10):1440–8.

[Fu et al., 2015] Fu, Y., Luo, G.-Z., Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X., Doré, L., Weng, X., Ji, Q., Mets, L., and He, C. (2015). N6-Methyldeoxyadenosine Marks Active Transcription Start Sites in Chlamydomonas. *Cell*, 161(4):879–892.

[Gabrusewycz-Garcia, 1968] Gabrusewycz-Garcia, N. (1968). RNA metabolism of polytene chromosomes of Sciara coprophila. *J. Cell Biol.*, 39:49.

[Gabrusewycz-Garcia, 1971] Gabrusewycz-Garcia, N. (1971). Studies in polytene chromosomes of sciarids. *Chromosoma*, 33(4):421–435.

[Gabrusewycz-Garcia and Kleinfeld, 1966] Gabrusewycz-Garcia, N. and Kleinfeld, R. G. (1966). A study of the nucleolar material in Sciara coprophila. *The Journal of Cell Biology*, 29(2).

[Gabrusewycz-Garcia and Mariano Garcia, 1974] Gabrusewycz-Garcia, N. and Mariano Garcia, A. (1974). Studies on the fine structure of puffs in Sciara coprophila. *Chromosoma*, 47(4):385–401.

[Gabrusewycz-Garica, 1964] Gabrusewycz-Garica, N. (1964). Cytological and autoradiographic studies in Sciara coprophila salivary gland chromosomes. *Chromosoma*, 15:312–44.

[Gambus et al., 2011] Gambus, A., Khoudoli, G. A., Jones, R. C., and Blow, J. J. (2011). MCM2-7 form double hexamers at licensed origins in Xenopus egg extract. *The Journal of biological chemistry*, 286(13):11855–64.

[Ge et al., 2015] Ge, W., Deng, Q., Guo, T., Hong, X., Kugler, J.-M., Yang, X., and Cohen, S. M. (2015). Regulation of Pattern Formation and Gene Amplification During Drosophila Oogenesis by the miR-318 microRNA. *Genetics*, 200(1).

[Ge et al., 2007] Ge, X. Q., Jackson, D. A., and Blow, J. J. (2007). Dormant origins licensed by excess Mcm2-7 are required for human cells to survive replicative stress. *Genes & development*, 21(24):3331–41.

[Gencheva et al., 1996] Gencheva, M., Anachkova, B., and Russev, G. (1996). Mapping the sites of initiation of DNA replication in rat and human rRNA genes. *The Journal of biological chemistry*, 271(5):2608–14.

[Genest et al., 2015] Genest, P.-A., Baugh, L., Taipale, A., Zhao, W., Jan, S., van Luenen, H. G. A. M., Korlach, J., Clark, T., Luong, K., Boitano, M., Turner, S., Myler, P. J., and Borst, P. (2015). Defining the sequence requirements for the positioning of base J in DNA using SMRT sequencing. *Nucleic acids research*, 43(4):2102–15.

[Gerbi and Urnov, 1996] Gerbi, S. and Urnov, F. (1996). Differential DNA replication in insects. *DNA replication in eukaryotic cells.*

[Gerbi, 1971] Gerbi, S. A. (1971). Localization and characterization of the ribosomal RNA cistrons in Sciara coprophila. *Journal of Molecular Biology*, 58(2):499–511.

[Gerbi, 1986] Gerbi, S. A. (1986). Unusual chromosome movements in sciarid flies. *Results and problems in cell differentiation*, 13:71–104.

[Gerbi, 2005] Gerbi, S. A. (2005). Mapping origins of DNA replication in eukaryotes. *Methods in molecular biology (Clifton, N.J.)*, 296:167–80.

[Gerbi and Bielinsky, 1997] Gerbi, S. A. and Bielinsky, A. K. (1997). Replication initiation point mapping. *Methods (San Diego, Calif.)*, 13(3):271–80.

[Gerbi and Bielinsky, 2002] Gerbi, S. A. and Bielinsky, A. K. (2002). DNA replication and chromatin. *Current opinion in genetics & development*, 12(2):243–8.

[Gerbi et al., 1999] Gerbi, S. A., Bielinsky, A. K., C, L., VV, L., and FD, U. (1999). Methods to map origins of replication in eukaryotes. In Oxford, C. S., editor, *Eukaryotic DNA Replication: a Practical Approach*, pages 1–42. Oxford University Press.

[Gerbi et al., 1993] Gerbi, S. A., Liang, C., Wu, N., DiBartolomeis, S. M., Bienz-Tadmor, B., Smith, H. S., and Urnov, F. D. (1993). DNA amplification in DNA puff II/9A of Sciara coprophila. *Cold Spring Harbor symposia on quantitative biology*, 58:487–94.

[Ghbeish et al., 2001] Ghbeish, N., Tsai, C.-C., Schubiger, M., Zhou, J. Y., Evans, R. M., and McKeown, M. (2001). The dual role of ultraspiracle, the Drosophila retinoid X receptor, in the ecdysone response. *Proceedings of the National Academy of Sciences*, 98(7):3867–3872.

[Ghodsi et al., 2013] Ghodsi, M., Hill, C. M., Astrovskaya, I., Lin, H., Sommer, D. D., Koren, S., and Pop, M. (2013). De novo likelihood-based measures for comparing genome assemblies. *BMC research notes*, 6:334.

[Giacca et al., 1994] Giacca, M., Zentilin, L., Norio, P., Diviacco, S., Dimitrova, D., Contreas, G., Biamonti, G., Perini, G., Weighardt, F., and Riva, S. (1994). Fine mapping of a replication origin of human DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 91(15):7119–23.

[Gilbert, 2005] Gilbert, D. M. (2005). Origins go plastic. *Molecular cell*, 20(5):657–8.

[Gilbert, 2010] Gilbert, D. M. (2010). Evaluating genome-scale approaches to eukaryotic DNA replication. *Nature reviews. Genetics*, 11(10):673–84.

[Gimenes et al., 2009] Gimenes, F., Assis, M. A., Fiorini, A., Mareze, V. A., Monesi, N., and Fernandez, M. A. (2009). Intrinsically bent DNA sites in the Drosophila melanogaster third chromosome amplified domain. *Molecular Genetics and Genomics*, 281(5):539–549.

[Glastad et al., 2011] Glastad, K. M., Hunt, B. G., Yi, S. V., and Goodisman, M. A. D. (2011). DNA methylation in insects: on the brink of the epigenomic era. *Insect Molecular Biology*, 20(5):553–565.

[Gnerre et al., 2011] Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1513–8.

[Goday and Esteban, 2001] Goday, C. and Esteban, M. R. (2001). Chromosome elimination in sciarid flies. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 23(3):242–50.

[Gómez and Antequera, 1999] Gómez, M. and Antequera, F. (1999). Organization of DNA replication origins in the fission yeast genome. *The EMBO journal*, 18(20):5683–90.

[Gómez and Antequera, 2008] Gómez, M. and Antequera, F. (2008). Overreplication of short DNA regions during S phase in human cells. *Genes & development*, 22(3):375–85.

[Gómez and Brockdorff, 2004] Gómez, M. and Brockdorff, N. (2004). Heterochromatin on the inactive X chromosome delays replication timing without affecting origin usage. *Proceedings of the National Academy of Sciences of the United States of America*, 101(18):6923–8.

[Goodman and Benjamin, 1973] Goodman, R. M. and Benjamin, W. B. (1973). Nucleoprotein methylation in salivary gland chromosomes of Sciara coprophila: Correlation with DNA synthesis. *Experimental Cell Research*, 77(1):63–72.

[Goodwin et al., 2015a] Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M., and McCombie, W. R. (2015a). Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome. *bioRxiv*.

[Goodwin et al., 2015b] Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., and McCombie, W. R. (2015b). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, pages gr.191395.115–.

[Gopalakrishnan et al., 2001] Gopalakrishnan, V., Simancek, P., Houchens, C., Snaith, H. A., Frattini, M. G., Sazer, S., and Kelly, T. J. (2001). Redundant control of rereplication in fission yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23):13114–9.

[Grabherr et al., 2011] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–52.

[Graessmann et al., 1973] Graessmann, A., Graessmann, M., and Larat, F. J. S. (1973). Involvement of RNA in the Process of Puff Induction in Polytene Chromosomes. In Hamkalo, B. and Papaconstantinou, J., editors, *Molecular Cytogenetics*, pages 209–215. Springer New York.

[Gray et al., 2007] Gray, S. J., Liu, G., Altman, A. L., Small, L. E., and Fanning, E. (2007). Discrete functional elements required for initiation activity of the Chinese hamster dihydrofolate reductase origin beta at ectopic chromosomal sites. *Experimental cell research*, 313(1):109–20.

[Greciano et al., 2009] Greciano, P. G., Ruiz, M. F., Kremer, L., and Goday, C. (2009). Two new chromodomain-containing proteins that associate with heterochromatin in Sciara coprophila chromosomes. *Chromosoma*, 118(3):361–376.

[Green et al., 2010] Green, B. M., Finn, K. J., and Li, J. J. (2010). Loss of DNA replication control is a potent inducer of gene amplification. *Science (New York, N.Y.)*, 329(5994):943–6.

[Green and Li, 2005] Green, B. M. and Li, J. J. (2005). Loss of rereplication control in Saccharomyces cerevisiae results in extensive DNA damage. *Molecular biology of the cell*, 16(1):421–32.

[Green et al., 2006] Green, B. M., Morreale, R. J., Ozaydin, B., Derisi, J. L., and Li, J. J. (2006). Genome-wide mapping of DNA synthesis in Saccharomyces cerevisiae reveals that mechanisms preventing reinitiation of DNA replication are not redundant. *Molecular biology of the cell*, 17(5):2401–14.

[Greer et al., 2015] Greer, E. L., Blanco, M. A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., Hsu, C.-H., Aravind, L., He, C., and Shi, Y. (2015). DNA Methylation on N6-Adenine in C. elegans. *Cell*, 161(4):868–878.

[Greninger et al., 2015] Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., Bouquet, J., Stryke, D., Somasekar, S., Linnen, J., Dodd, R., Mulembakani, P., Schneider, B., Muyembe, J.-J., Stramer, S., and Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *bioRxiv*.

[Gros et al., 2015] Gros, J., Kumar, C., Lynch, G., Yadav, T., Whitehouse, I., and Remus, D. (2015). Post-licensing Specification of Eukaryotic Replication Origins by Facilitated Mcm2-7 Sliding along DNA. *Molecular Cell*, 60(5):797–807.

[Guan et al., 2009] Guan, Z., Hughes, C. M., Kosiyatrakul, S., Norio, P., Sen, R., Fiering, S., Allis, C. D., Bouhassira, E. E., and Schildkraut, C. L. (2009). Decreased replication origin activity in temporal transition regions. *The Journal of cell biology*, 187(5):623–35.

[Haas et al., 2013] Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., and Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512.

[Hackney et al., 2007] Hackney, J. F., Pucci, C., Naes, E., and Dobens, L. (2007). Ras signaling modulates activity of the ecdysone receptor EcR during cell migration in the Drosophila ovary.

Developmental dynamics : an official publication of the American Association of Anatomists, 236(5):1213–26.

[Halder et al., 2009] Halder, K., Halder, R., and Chowdhury, S. (2009). Genome-wide analysis predicts DNA structural motifs as nucleosome exclusion signals. *Molecular bioSystems*, 5(12):1703–12.

[Hamamoto et al., 2015] Hamamoto, R., Saloura, V., and Nakamura, Y. (2015). Critical roles of non-histone protein lysine methylation in human tumorigenesis. *Nature Reviews Cancer*, 15(2):110–124.

[Hamlin et al., 2010] Hamlin, J. L., Mesner, L. D., and Dijkwel, P. A. (2010). A winding road to origin discovery. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 18(1):45–61.

[Hamlin et al., 2008] Hamlin, J. L., Mesner, L. D., Lar, O., Torres, R., Chodaparambil, S. V., and Wang, L. (2008). A revisionist replicon model for higher eukaryotic genomes. *Journal of cellular biochemistry*, 105(2):321–9.

[Han et al., 1999] Han, H., Hurley, L. H., and Salazar, M. (1999). A DNA polymerase stop assay for G-quadruplex-interactive compounds. *Nucleic acids research*, 27(2):537–42.

[Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674.

[Handeli et al., 1989] Handeli, S., Klar, A., Meuth, M., and Cedar, H. (1989). Mapping replication units in animal cells. *Cell*, 57(6):909–20.

[Hanlon and Li, 2015] Hanlon, S. L. and Li, J. J. (2015). Re-replication of a Centromere Induces Chromosomal Instability and Aneuploidy. *PLOS Genetics*, 11(4):e1005039.

[Hänsel-Hertsch et al., 2016] Hänsel-Hertsch, R., Beraldi, D., Lensing, S. V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M., Tannahill, D., and Balasubramanian, S. (2016). G-quadruplex structures mark human regulatory chromatin. *Nature Genetics*, 48(10):1267–1272.

[Hansen et al., 2010] Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., and Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):139–44.

[Harland and Laskey, 1980] Harland, R. M. and Laskey, R. A. (1980). Regulated replication of DNA microinjected into eggs of Xenopus laevis. *Cell*, 21(3):761–71.

[Hartl et al., 2007] Hartl, T., Boswell, C., Orr-Weaver, T. L., and Bosco, G. (2007). Developmentally regulated histone modifications in Drosophila follicle cells: initiation of gene amplification is

associated with histone H3 and H4 hyperacetylation and H1 phosphorylation. *Chromosoma*, 116(2):197–214.

[Hay and DePamphilis, 1982] Hay, R. T. and DePamphilis, M. L. (1982). Initiation of SV40 DNA replication in vivo: location and structure of 5' ends of DNA synthesized in the ori region. *Cell*, 28(4):767–79.

[Heck and Spradling, 1990] Heck, M. M. and Spradling, A. C. (1990). Multiple replication origins are used during Drosophila chorion gene amplification. *The Journal of cell biology*, 110(4):903–14.

[Heintz and Hamlin, 1982] Heintz, N. H. and Hamlin, J. L. (1982). An amplified chromosomal sequence that includes the gene for dihydrofolate reductase initiates replication within specific restriction fragments. *Proceedings of the National Academy of Sciences of the United States of America*, 79(13):4083–7.

[Heinzel et al., 1991] Heinzel, S. S., Krysan, P. J., Tran, C. T., and Calos, M. P. (1991). Autonomous DNA replication in human cells is affected by the size and the source of the DNA. *Molecular and cellular biology*, 11(4):2263–72.

[Heitz and Bauer, 1933] Heitz, E. and Bauer, H. (1933). Beweise fur die Chromosomennatur der Kernschleifen in den Knauelkernen von Bibio hortulanus. *Z. Zellforsch.*, 17:67.

[Heller et al., 2011] Heller, R. C., Kang, S., Lam, W. M., Chen, S., Chan, C. S., and Bell, S. P. (2011). Eukaryotic origin-dependent DNA replication in vitro reveals sequential action of DDK and S-CDK kinases. *Cell*, 146(1):80–91.

[Hemerly et al., 2009] Hemerly, A. S., Prasanth, S. G., Siddiqui, K., and Stillman, B. (2009). Orc1 controls centriole and centrosome copy number in human cells. *Science (New York, N.Y.)*, 323(5915):789–93.

[Hendrickson et al., 1987] Hendrickson, E. A., Fritze, C. E., Folk, W. R., and DePamphilis, M. L. (1987). The origin of bidirectional DNA replication in polyoma virus. *The EMBO journal*, 6(7):2011–8.

[Herrick and Bensimon, 1999] Herrick, J. and Bensimon, A. (1999). Single molecule analysis of DNA replication. *Biochimie*, 81(8-9):859–71.

[Herrick and Bensimon, 2009] Herrick, J. and Bensimon, A. (2009). Introduction to molecular combing: genomics, DNA replication, and cancer. *Methods in molecular biology (Clifton, N.J.)*, 521:71–101.

[Hershman et al., 2008] Hershman, S. G., Chen, Q., Lee, J. Y., Kozak, M. L., Yue, P., Wang, L.-S., and Johnson, F. B. (2008). Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic acids research*, 36(1):144–56.

[Heyn and Esteller, 2015] Heyn, H. and Esteller, M. (2015). An Adenine Code for DNA: A Second Life for N6-Methyladenine. *Cell*, 161(4):710–713.

[Hoshina et al., 2013] Hoshina, S., Yura, K., Teranishi, H., Kiyasu, N., Tominaga, A., Kadoma, H., Nakatsuka, A., Kunichika, T., Obuse, C., and Waga, S. (2013). Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA. *The Journal of biological chemistry*, 288(42):30161–71.

[Hua et al., 2014] Hua, B. L., Li, S., and Orr-Weaver, T. L. (2014). The role of transcription in the activation of a Drosophila amplification origin. *G3 (Bethesda, Md.)*, 4(12):2403–8.

[Huang et al., 1998] Huang, D. W., Fanti, L., Pak, D. T., Botchan, M. R., Pimpinelli, S., and Kellum, R. (1998). Distinct cytoplasmic and nuclear fractions of Drosophila heterochromatin protein 1: their phosphorylation levels and associations with origin recognition complex proteins. *The Journal of cell biology*, 142(2):307–18.

[Huang et al., 2013] Huang, Y.-C., Smith, L., Poulton, J., and Deng, W.-M. (2013). The microRNA miR-7 regulates Tramtrack69 in a developmental switch in Drosophila follicle cells. *Development*, 140(4).

[Huberman and Riggs, 1966] Huberman, J. A. and Riggs, A. D. (1966). Autoradiography of chromosomal DNA fibers from Chinese hamster cells. *Proceedings of the National Academy of Sciences of the United States of America*, 55(3):599–606.

[Huberman and Riggs, 1968] Huberman, J. A. and Riggs, A. D. (1968). On the mechanism of DNA replication in mammalian chromosomes. *Journal of molecular biology*, 32(2):327–41.

[Huberman et al., 1987] Huberman, J. A., Spotila, L. D., Nawotka, K. A., El-Assouli, S. M., and Davis, L. R. (1987). The in vivo replication origin of the yeast 2 microns plasmid. *Cell*, 51(3):473–81.

[Hunt et al., 2013] Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome biology*, 14(5):R47.

[Huppert, 2010] Huppert, J. L. (2010). Structure, location and interactions of G-quadruplexes. *The FEBS journal*, 277(17):3452–8.

[Huppert and Balasubramanian, 2005] Huppert, J. L. and Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. *Nucleic acids research*, 33(9):2908–16.

[Hyrien et al., 2003] Hyrien, O., Marheineke, K., and Goldar, A. (2003). Paradoxes of eukaryotic DNA replication: MCM proteins and the random completion problem. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 25(2):116–25.

[Hyrien et al., 1995] Hyrien, O., Maric, C., and Méchali, M. (1995). Transition in specification of embryonic metazoan DNA replication origins. *Science (New York, N.Y.)*, 270(5238):994–7.

[Hyrien and Méchali, 1993] Hyrien, O. and Méchali, M. (1993). Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of Xenopus early embryos. *The EMBO journal*, 12(12):4511–20.

[Hyrien et al., 2013] Hyrien, O., Rappailles, A., Guilbaud, G., Baker, A., Chen, C.-L., Goldar, A., Petryk, N., Kahli, M., Ma, E., D'Aubenton-Carafa, Y., Audit, B., Thermes, C., and Arneodo, A. (2013). From simple bacterial and archaeal replicons to replication N/U-domains. *Journal of molecular biology*, 425(23):4673–89.

[Iizuka and Stillman, 1999] Iizuka, M. and Stillman, B. (1999). Histone acetyltransferase HBO1 interacts with the ORC1 subunit of the human initiator protein. *The Journal of biological chemistry*, 274(33):23027–34.

[Ilves et al., 2010] Ilves, I., Petojevic, T., Pesavento, J. J., and Botchan, M. R. (2010). Activation of the MCM2-7 helicase by association with Cdc45 and GINS proteins. *Molecular cell*, 37(2):247–58.

[Ip et al., 2015] Ip, C. L., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., Leggett, R. M., Eccles, D. A., Zalunin, V., Urban, J. M., Piazza, P., Bowden, R. J., Paten, B., Mwaigwisya, S., Batty, E. M., Simpson, J. T., Snutch, T. P., Birney, E., Buck, D., Goodwin, S., Jansen, H. J., O'Grady, J., and Olsen, H. E. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*, 4.

[Iyer and Struhl, 1995] Iyer, V. and Struhl, K. (1995). Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *The EMBO journal*, 14(11):2570–9.

[Jain et al., 2015] Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, 12:351–356.

[Kajitani et al., 2014] Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., and Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*, 24(8):1384–95.

[Kalejta et al., 1998] Kalejta, R. F., Li, X., Mesner, L. D., Dijkwel, P. A., Lin, H. B., and Hamlin, J. L. (1998). Distal sequences, but not ori-beta/OBR-1, are essential for initiation of DNA replication in the Chinese hamster DHFR origin. *Molecular cell*, 2(6):797–806.

[Kamath et al., 2016] Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A., and Tse, D. N. (2016). HINGE: Long-Read Assembly Achieves Optimal Repeat Resolution. *bioRxiv*.

[Kankia and Marky, 2001] Kankia, B. I. and Marky, L. A. (2001). Folding of the thrombin aptamer into a G-quadruplex with Sr(2+): stability, heat, and hydration. *Journal of the American Chemical Society*, 123(44):10799–804.

[Kaplan and Dekker, 2013] Kaplan, N. and Dekker, J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nature biotechnology*, 31(12):1143–7.

[Karlsson et al., 2015] Karlsson, E., Lärkeryd, A., Sjödin, A., Forsman, M., and Stenberg, P. (2015). Scaffolding of a bacterial genome using MinION nanopore sequencing. *Scientific reports*, 5:11996.

[Karnani et al., 2010] Karnani, N., Taylor, C. M., Malhotra, A., and Dutta, A. (2010). Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. *Molecular biology of the cell*, 21(3):393–404.

[Karolchik et al., 2004] Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic acids research*, 32(Database issue):D493–6.

[Kent et al., 2002] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome research*, 12(6):996–1006.

[Kent et al., 2010] Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics (Oxford, England)*, 26(17):2204–7.

[Kiang et al., 2010] Kiang, L., Heichinger, C., Watt, S., Bahler, J., and Nurse, P. (2010). Specific replication origins promote DNA amplification in fission yeast. *Journal of Cell Science*, 123(18):3047–3051.

[Kikin et al., 2006] Kikin, O., D'Antonio, L., and Bagga, P. S. (2006). QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic acids research*, 34(Web Server issue):W676–82.

[Kilianski et al., 2015] Kilianski, A., Haas, J. L., Corriveau, E. J., Liem, A. T., Willis, K. L., Kadavy, D. R., Rosenzweig, C. N., and Minot, S. S. (2015). Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *GigaScience*, 4(1):12.

[Kim et al., 2015] Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360.

[Kim et al., 2011] Kim, J. C., Nordman, J., Xie, F., Kashevsky, H., Eng, T., Li, S., MacAlpine, D. M., and Orr-Weaver, T. L. (2011). Integrative analysis of gene amplification in Drosophila follicle cells: parameters of origin activation and repression. *Genes & development*, 25(13):1384–98.

[Kim and Orr-Weaver, 2011] Kim, J. C. and Orr-Weaver, T. L. (2011). Analysis of a Drosophila amplicon in follicle cells highlights the diversity of metazoan replication origins. *Proceedings of the National Academy of Sciences of the United States of America*, 108(40):16681–6.

[Kirilly et al., 2011] Kirilly, D., Wong, J., Lim, E., Wang, Y., Zhang, H., Wang, C., Liao, Q., Wang, H., Liou, Y.-C., Wang, H., and Yu, F. (2011). Intrinsic Epigenetic Factors Cooperate with the Steroid Hormone Ecdysone to Govern Dendrite Pruning in Drosophila. *Neuron*, 72(1):86–100.

[Kobayashi et al., 1998] Kobayashi, T., Rein, T., and DePamphilis, M. L. (1998). Identification of primary initiation sites for DNA replication in the hamster dihydrofolate reductase gene initiation zone. *Molecular and cellular biology*, 18(6):3266–77.

[Koboldt et al., 2012] Koboldt, D. C., Larson, D. E., Chen, K., Ding, L., and Wilson, R. K. (2012). Massively parallel sequencing approaches for characterization of structural variation. *Methods in molecular biology (Clifton, N.J.)*, 838:369–84.

[Kolmogorov et al., 2016] Kolmogorov, M., Armstrong, J., Raney, B. J., Streeter, I., Dunn, M., Yang, F., Odom, D., Flicek, P., Keane, T., Thybert, D., Paten, B., and Pham, S. (2016). Chromosome assembly of large and complex genomes using multiple references. *bioRxiv*.

[Komitopoulou et al., 1988] Komitopoulou, K., Margaritis, L. H., and Kafatos, F. C. (1988). Structural and biochemical studies on four sex-linked chorion mutants ofDrosophila melanogaster. *Developmental Genetics*, 9(1):37–48.

[Koob and Szybalski, 1992] Koob, M. and Szybalski, W. (1992). Preparing and using agarose microbeads. *Methods in enzymology*, 216:13–20.

[Koren and McCarroll, 2014] Koren, A. and McCarroll, S. A. (2014). Random replication of the inactive X chromosome. *Genome research*, 24(1):64–9.

[Koren et al., 2013] Koren, S., Harhay, G. P., Smith, T. P. L., Bono, J. L., Harhay, D. M., Mcvey, S. D., Radune, D., Bergman, N. H., and Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome biology*, 14(9):R101.

[Koren and Phillippy, 2015] Koren, S. and Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, 23:110–120.

[Koren et al., 2012] Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., and Adam M Phillippy (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7):693–700.

[Koren et al., 2016] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., and Phillippy, A. M. (2016). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv*.

[Koutsovoulos et al., 2016] Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A. A., and Blaxter, M. (2016). No evidence for

extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. *Proceedings of the National Academy of Sciences of the United States of America*, 113(18):5053–8.

[Krysan et al., 1989] Krysan, P. J., Haase, S. B., and Calos, M. P. (1989). Isolation of human sequences that replicate autonomously in human cells. *Molecular and cellular biology*, 9(3):1026–33.

[Kumar et al., 2013] Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., and Blaxter, M. (2013). Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in genetics*, 4:237.

[Kundaje et al., 2012] Kundaje, A., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M., Smith, C. L., Raha, D., Winters, E. E., Johnson, S. M., Snyder, M., Batzoglou, S., and Sidow, A. (2012). Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome research*, 22(9):1735–47.

[Kunnev et al., 2015a] Kunnev, D., Freeland, A., Qin, M., Leach, R. W., Wang, J., Shenoy, R. M., and Pruitt, S. C. (2015a). Effect of minichromosome maintenance protein 2 deficiency on the locations of DNA replication origins. *Genome Research*.

[Kunnev et al., 2015b] Kunnev, D., Freeland, A., Qin, M., Wang, J., and Pruitt, S. C. (2015b). Isolation and sequencing of active origins of DNA replication by nascent strand capture and release (NSCR). *Journal of Biological Methods*, 2(4):33.

[Laetsch et al., 2016] Laetsch, D. R., Koutsovoulos, G., and Stajich, J. (2016). blobtools: blobtools v0.9.19.4. *zenodo*.

[Lam et al., 2012] Lam, E. T., Hastie, A., Lin, C., Ehrlich, D., Das, S. K., Austin, M. D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M., and Kwok, P.-Y. (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, 30(8):771–776.

[Lam et al., 2015] Lam, K.-K., LaButti, K., Khalak, A., and Tse, D. (2015). FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. *Bioinformatics (Oxford, England)*, 31(19):3207–9.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray,

A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Landis et al., 1997] Landis, G., Kelley, R., Spradling, A. C., and Tower, J. (1997). The k43 gene, required for chorion gene amplification and diploid cell chromosome replication, encodes the Drosophila homolog of yeast origin recognition complex subunit 2. *Proceedings of the National Academy of Sciences of the United States of America*, 94(8):3888–92.

[Landis and Tower, 1999] Landis, G. and Tower, J. (1999). The Drosophila chiffon gene is required for chorion gene amplification, and is related to the yeast Dbf4 regulator of DNA replication and cell cycle. *Development (Cambridge, England)*, 126(19):4281–93.

[Landt et al., 2012] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein,

C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–31.

[Langley et al., 2016] Langley, A. R., Gräf, S., Smith, J. C., and Krude, T. (2016). Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic acids research*, page gkw760.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9.

[Lebofsky and Bensimon, 2005] Lebofsky, R. and Bensimon, A. (2005). DNA replication origin plasticity and perturbed fork progression in human inverted repeats. *Molecular and cellular biology*, 25(15):6789–97.

[Lebofsky et al., 2006] Lebofsky, R., Heilig, R., Sonnleitner, M., Weissenbach, J., and Bensimon, A. (2006). DNA replication origin interference increases the spacing between initiation events in human cells. *Molecular biology of the cell*, 17(12):5337–45.

[Lemarteleur et al., 2004] Lemarteleur, T., Gomez, D., Paterski, R., Mandine, E., Mailliet, P., and Riou, J.-F. (2004). Stabilization of the c-myc gene promoter quadruplex by specific ligands' inhibitors of telomerase. *Biochemical and biophysical research communications*, 323(3):802–8.

[Leonard and Méchali, 2013] Leonard, A. C. and Méchali, M. (2013). DNA replication origins. *Cold Spring Harbor perspectives in biology*, 5(10):a010116.

[Li et al., 2015] Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676.

[Li, 2016] Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics (Oxford, England)*, 32(14):2103–10.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9.

[Li and Breaker, 1999] Li, Y. and Breaker, R. R. (1999). Kinetics of RNA Degradation by Specific Base Catalysis of Transesterification Involving the 2-Hydroxyl Group. *Journal of the American Chemical Society*, 121(23):5364–5372.

[Liachko et al., 2014] Liachko, I., Youngblood, R. A., Tsui, K., Bubb, K. L., Queitsch, C., Raghuraman, M. K., Nislow, C., Brewer, B. J., and Dunham, M. J. (2014). GC-rich DNA elements enable replication origin activity in the methylotrophic yeast Pichia pastoris. *PLoS genetics*, 10(3):e1004169.

[Liang and Gerbi, 1994] Liang, C. and Gerbi, S. A. (1994). Analysis of an origin of DNA amplification in Sciara coprophila by a novel three-dimensional gel method. *Molecular and cellular biology*, 14(2):1520–9.

[Liang et al., 1993] Liang, C., Spitzer, J. D., Smith, H. S., and Gerbi, S. A. (1993). Replication initiates at a confined region during DNA amplification in Sciara DNA puff II/9A. *Genes & development*, 7(6):1072–84.

[Liew et al., 2013] Liew, G. M., Foulk, M. S., and Gerbi, S. A. (2013). The ecdysone receptor (ScEcR-A) binds DNA puffs at the start of DNA amplification in Sciara coprophila. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 21(4):345–60.

[Lin et al., 1999] Lin, J., Qi, R., Aston, C., Jing, J., Anantharaman, T. S., Mishra, B., White, O., Daly, M. J., Minton, K. W., Venter, J. C., and Schwartz, D. C. (1999). Whole-Genome Shotgun Optical Mapping of Deinococcus radiodurans. *Science*, 285(5433).

[Lin et al., 2016] Lin, Y., Yuan, J., Kolmogorov, M., Shen, M. W., and Pevzner, P. A. (2016). Assembly of Long Error-Prone Reads Using de Bruijn Graphs. *bioRxiv*.

[Lipford and Bell, 2001] Lipford, J. R. and Bell, S. P. (2001). Nucleosomes positioned by ORC facilitate the initiation of DNA replication. *Molecular cell*, 7(1):21–30.

[Little, 1967] Little, J. W. (1967). An exonuclease induced by bacteriophage lambda. II. Nature of the enzymatic reaction. *The Journal of biological chemistry*, 242(4):679–86.

[Little et al., 1993] Little, R. D., Platt, T. H., and Schildkraut, C. L. (1993). Initiation and termination of DNA replication in human rRNA genes. *Molecular and cellular biology*, 13(10):6600–13.

[Liu et al., 2003] Liu, G., Malott, M., and Leffak, M. (2003). Multiple functional elements comprise a Mammalian chromosomal replicator. *Molecular and cellular biology*, 23(5):1832–42.

[Liu et al., 2012] Liu, J., McConnell, K., Dixon, M., and Calvi, B. R. (2012). Analysis of model replication origins in Drosophila reveals new aspects of the chromatin landscape and its relationship to origin activity and the prereplicative complex. *Molecular biology of the cell*, 23(1):200–12.

[Liu et al., 2015] Liu, J., Zimmer, K., Rusch, D. B., Paranjape, N., Podicheti, R., Tang, H., and Calvi, B. R. (2015). DNA sequence templates adjacent nucleosome and ORC sites at gene amplification origins in Drosophila. *Nucleic acids research*, 43(18):8746–61.

[Livak and Schmittgen, 2001] Livak, K. J. and Schmittgen, T. D. (2001). Analysis of Relative Gene Expression Data Using Real- Time Quantitative PCR and the 2 C T Method. *METHODS*, 25:402–408.

[Loman et al., 2015] Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, advance on.

[Loman and Quinlan, 2014] Loman, N. J. and Quinlan, A. R. (2014). Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, pages btu555–.

[Lombraña et al., 2016] Lombraña, R., Álvarez, A., Fernández-Justel, J., Almeida, R., Poza-Carrión, C., Gomes, F., Calzada, A., Requena, J., and Gómez, M. (2016). Transcriptionally Driven DNA Replication Program of the Human Parasite Leishmania major. *Cell Reports*, 16(6):1774–1786.

[Lu and Tower, 1997] Lu, L. and Tower, J. (1997). A transcriptional insulator element, the su(Hw) binding site, protects a chromosomal DNA replication origin from position effects. *Molecular and cellular biology*, 17(4):2202–6.

[Lu et al., 2001] Lu, L., Zhang, H., and Tower, J. (2001). Functionally distinct, sequence-specific replicator and origin elements are required for Drosophila chorion gene amplification. *Genes & development*, 15(2):134–46.

[Lubelsky et al., 2012] Lubelsky, Y., MacAlpine, H. K., and MacAlpine, D. M. (2012). Genome-wide localization of replication factors. *Methods (San Diego, Calif.)*, 57(2):187–95.

[Lubelsky et al., 2014] Lubelsky, Y., Prinz, J. A., DeNapoli, L., Li, Y., Belsky, J. A., and MacAlpine, D. M. (2014). DNA replication and transcription programs respond to the same chromatin cues. *Genome research*, 24(7):1102–14.

[Lubelsky et al., 2011] Lubelsky, Y., Sasaki, T., Kuipers, M. A., Lucas, I., Le Beau, M. M., Carignon, S., Debatisse, M., Prinz, J. A., Dennis, J. H., and Gilbert, D. M. (2011). Pre-replication complex proteins assemble at regions of low nucleosome occupancy within the Chinese hamster dihydrofolate reductase initiation zone. *Nucleic acids research*, 39(8):3141–55.

[Lucas et al., 2007] Lucas, I., Palakodeti, A., Jiang, Y., Young, D. J., Jiang, N., Fernald, A. A., and Le Beau, M. M. (2007). High-throughput mapping of origins of replication in human cells. *EMBO reports*, 8(8):770–7.

[Lunyak et al., 2002] Lunyak, V. V., Ezrokhi, M., Smith, H. S., and Gerbi, S. A. (2002). Developmental changes in the Sciara II/9A initiation zone for DNA replication. *Molecular and cellular biology*, 22(24):8426–37.

[Luo et al., 2012] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W., and Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18.

[Lyko and Maleszka, 2011] Lyko, F. and Maleszka, R. (2011). Insects as innovative models for functional studies of DNA methylation. *Trends in Genetics*, 27(4):127–131.

[MacAlpine and Bell, 2005] MacAlpine, D. M. and Bell, S. P. (2005). A genomic view of eukaryotic DNA replication. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 13(3):309–26.

[MacAlpine et al., 2004] MacAlpine, D. M., Rodríguez, H. K., and Bell, S. P. (2004). Coordination of replication and transcription along a Drosophila chromosome. *Genes & development*, 18(24):3094–105.

[MacAlpine et al., 2010] MacAlpine, H. K., Gordân, R., Powell, S. K., Hartemink, A. J., and MacAlpine, D. M. (2010). Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading. *Genome research*, 20(2):201–11.

[Maccallum et al., 2009] Maccallum, I., Przybylski, D., Gnerre, S., Burton, J., Shlyakhter, I., Gnirke, A., Malek, J., McKernan, K., Ranade, S., Shea, T. P., Williams, L., Young, S., Nusbaum, C., and Jaffe, D. B. (2009). ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome biology*, 10(10):R103.

[Macheret and Halazonetis, 2015] Macheret, M. and Halazonetis, T. D. (2015). DNA Replication Stress as a Hallmark of Cancer. *Annual Review of Pathology: Mechanisms of Disease*, 10(1):425–448.

[Machida et al., 2005] Machida, Y. J., Hamlin, J. L., and Dutta, A. (2005). Right place, right time, and only once: replication initiation in metazoans. *Cell*, 123(1):13–24.

[Madoui et al., 2015] Madoui, M.-A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., Lemainque, A., Wincker, P., and Aury, J.-M. (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics*, 16(1):327.

[Malott and Leffak, 1999] Malott, M. and Leffak, M. (1999). Activity of the c-myc replicator at an ectopic chromosomal location. *Molecular and cellular biology*, 19(8):5685–95.

[Marahrens and Stillman, 1994] Marahrens, Y. and Stillman, B. (1994). Replicator dominance in a eukaryotic chromosome. *The EMBO journal*, 13(14):3395–400.

[Maric and Prioleau, 2010] Maric, C. and Prioleau, M.-N. (2010). Interplay between DNA replication and gene expression: a harmonious coexistence. *Current opinion in cell biology*, 22(3):277–83.

[Marie-Nelly et al., 2014] Marie-Nelly, H., Marbouty, M., Cournac, A., Flot, J.-F., Liti, G., Parodi, D. P., Syan, S., Guillén, N., Margeot, A., Zimmer, C., and Koszul, R. (2014). High-quality genome (re)assembly using chromosomal contact data. *Nature communications*, 5:5695.

[Martin and Wang, 2011] Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nature reviews. Genetics*, 12(10):671–82.

[Masai et al., 2010] Masai, H., Matsumoto, S., You, Z., Yoshizawa-Sugata, N., and Oda, M. (2010). Eukaryotic chromosome DNA replication: where, when, and how? *Annual review of biochemistry*, 79:89–130.

[Mayan, 2013] Mayan, M. D. (2013). RNAP-II Molecules Participate in the Anchoring of the ORC to rDNA Replication Origins. *PLoS ONE*, 8(1):e53405.

[McConnell et al., 2012] McConnell, K. H., Dixon, M., and Calvi, B. R. (2012). The histone acetyltransferases CBP and Chameau integrate developmental and DNA replication programs in Drosophila ovarian follicle cells. *Development*, 139(20):3880–3890.

[McGarry and Kirschner, 1998] McGarry, T. J. and Kirschner, M. W. (1998). Geminin, an inhibitor of DNA replication, is degraded during mitosis. *Cell*, 93(6):1043–53.

[McGuffee et al., 2013] McGuffee, S. R., Smith, D. J., and Whitehouse, I. (2013). Quantitative, genome-wide analysis of eukaryotic replication initiation and termination. *Molecular cell*, 50(1):123–35.

[McIntosh and Blow, 2012] McIntosh, D. and Blow, J. J. (2012). Dormant origins, the licensing checkpoint, and the response to replicative stresses. *Cold Spring Harbor perspectives in biology*, 4(10).

[Méchali, 2010] Méchali, M. (2010). Eukaryotic DNA replication origins: many choices for appropriate answers. *Nature reviews. Molecular cell biology*, 11(10):728–38.

[Méchali et al., 2013] Méchali, M., Yoshida, K., Coulombe, P., and Pasero, P. (2013). Genetic and epigenetic determinants of DNA replication origins, position and activation. *Current opinion in genetics & development*, 23(2):124–31.

[Mendelowitz and Pop, 2014] Mendelowitz, L. and Pop, M. (2014). Computational methods for optical mapping. *GigaScience*, 3(1):33.

[Mendelowitz et al., 2015] Mendelowitz, L. M., Schwartz, D. C., and Pop, M. (2015). Maligner: a fast ordered restriction map aligner. *Bioinformatics (Oxford, England)*, 32(7):1016–1022.

[Mesner et al., 2006] Mesner, L. D., Crawford, E. L., and Hamlin, J. L. (2006). Isolating apparently pure libraries of replication origins from complex genomes. *Molecular cell*, 21(5):719–26.

[Mesner et al., 2009] Mesner, L. D., Dijkwel, P. A., and Hamlin, J. L. (2009). Purification of restriction fragments containing replication intermediates from complex genomes for 2-D gel analysis. *Methods in molecular biology (Clifton, N.J.)*, 521:121–37.

[Mesner et al., 2013] Mesner, L. D., Valsakumar, V., Cieslik, M., Pickin, R., Hamlin, J. L., and Bekiranov, S. (2013). Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome research*, 23(11):1774–88.

[Mesner et al., 2011] Mesner, L. D., Valsakumar, V., Karnani, N., Dutta, A., Hamlin, J. L., and Bekiranov, S. (2011). Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription. *Genome research*, 21(3):377–89.

[Metz, 1931] Metz, C. (1931). Unisexual progenies and sex determination in Sciara. *Quart. Rev. Biol.*, 6:306–312.

[Metz and Gay, 1934a] Metz, C. and Gay, E. H. (1934a). Organization of Salivary Gland Chromosomes in Sciara in Relation to Genes on JSTOR. *Proceedings of the National Academy of Sciences of the United States of America*, 20(12):617–621.

[Metz and Schmuck, 1929] Metz, C. and Schmuck, M. (1929). Unisexual progenies and the sex chromosome mechanism in Sciara. *Proc. Nat. Acad. Sci.*, 15:863–866.

[Metz, 1925] Metz, C. W. (1925). Chromosome behavior in Sciara (Diptera). *Anat. Rec.*, 31:346–347.

[Metz, 1930] Metz, C. W. (1930). A possible alternative to the hypothesis of selective fertilization in Sciara. *Am. Nat.*, 64:380–382.

[Metz, 1934] Metz, C. W. (1934). Evidence Indicating that in Sciara the Sperm Regularly Transmits Two Sister Sex Chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 20(1):31–6.

[Metz, 1938] Metz, C. W. (1938). Chromosome behavior, inheritance and sex determination in Sciara. *Am. Nat.*, 72:485–520.

[Metz and Boche, 1939] Metz, C. W. and Boche, R. D. (1939). Observations on the Mechanism of Induced Chromosome Rearrangements in Sciara. *Proceedings of the National Academy of Sciences of the United States of America*, 25(6):280–4.

[Metz and Gay, 1934b] Metz, C. W. and Gay, E. H. (1934b). CHROMOSOME STRUCTURE IN THE SALIVARY GLANDS OF SCIARA. *Science*, 80(2086).

[Michalet et al., 1997] Michalet, X., Ekong, R., Fougerousse, F., Rousseaux, S., Schurra, C., Hornigold, N., van Slegtenhorst, M., Wolfe, J., Povey, S., Beckmann, J. S., and Bensimon, A. (1997). Dynamic molecular combing: stretching the whole human genome for high-resolution studies. *Science (New York, N.Y.)*, 277(5331):1518–23.

[Mikheyev and Tin, 2014] Mikheyev, A. S. and Tin, M. M. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, 14(6):1097–1102.

[Miotto et al., 2016] Miotto, B., Ji, Z., and Struhl, K. (2016). Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 113(33):E4810–9.

[Mohammad et al., 2007] Mohammad, M. M., Donti, T. R., Sebastian Yakisich, J., Smith, A. G., and Kapler, G. M. (2007). Tetrahymena ORC contains a ribosomal RNA fragment that participates in rDNA origin recognition. *The EMBO journal*, 26(24):5048–60.

[Mok et al., 2001] Mok, E. H., Smith, H. S., DiBartolomeis, S. M., Kerrebrock, A. W., Rothschild, L. J., Lange, T. S., and Gerbi, S. A. (2001). Maintenance of the DNA puff expanded state is independent of active replication and transcription. *Chromosoma*, 110(3):186–96.

[Monesi et al., 2003] Monesi, N., Basso, L. R., and Paco-Larson, M. L. (2003). Identification of regulatory regions in the DNA puff BhC4-1 promoter. *Insect Molecular Biology*, 12(3):247–254.

[Monesi et al., 1998] Monesi, N., Jacobs-Lorena, M., and Paçó-Larson, M. L. (1998). The DNA puff gene BhC4-1 of Bradysia hygida is specifically transcribed in early prepupal salivary glands of Drosophila melanogaster. *Chromosoma*, 107(8):559–569.

[Monesi et al., 2004] Monesi, N., Silva, J., Martins, P., Teixeira, A., Dornelas, E., Moreira, J., and Paçó Larson, M. (2004). Immunocharacterization of the DNA puff BhC4-1 protein of Bradysia hygida (Diptera: Sciaridae). *Insect Biochemistry and Molecular Biology*, 34(6):531–542.

[Monesi et al., 2001] Monesi, N., Sousa, J., and Paçó-Larson, M. (2001). The DNA puff BhB10-1 gene is differentially expressed in various tissues of Bradysia hygida late larvae and constitutively transcribed in transgenic Drosophila. *Brazilian Journal of Medical and Biological Research*, 34(7):851–859.

[Mostovoy et al., 2016] Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., Lee, J., Chu, C., Lin, C., Džakula, Ž., Cao, H., Schlebusch, S. A., Giorda, K., Schnall-Levin, M., Wall, J. D., and Kwok, P.-Y. (2016). A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, 13(7):587–590.

[Moyer et al., 2006] Moyer, S. E., Lewis, P. W., and Botchan, M. R. (2006). Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proceedings of the National Academy of Sciences*, 103(27):10236–10241.

[Mukhopadhyay et al., 2014] Mukhopadhyay, R., Lajugie, J., Fourel, N., Selzer, A., Schizas, M., Bartholdy, B., Mar, J., Lin, C. M., Martin, M. M., Ryan, M., Aladjem, M. I., and Bouhassira, E. E. (2014). Allele-specific genome-wide profiling in human primary erythroblasts reveal replication program organization. *PLoS genetics*, 10(5):e1004319.

[Myers et al., 2000] Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C. (2000). A whole-genome assembly of Drosophila. *Science (New York, N.Y.)*, 287(5461):2196–204.

[Nawotka and Huberman, 1988] Nawotka, K. A. and Huberman, J. A. (1988). Two-dimensional gel electrophoretic method for mapping DNA replicons. *Molecular and cellular biology*, 8(4):1408–13.

[Neely et al., 2011] Neely, R. K., Deen, J., and Hofkens, J. (2011). Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers*, 95(5):298–311.

[Negrini et al., 2010] Negrini, S., Gorgoulis, V. G., and Halazonetis, T. D. (2010). Genomic instability an evolving hallmark of cancer. *Nature Reviews Molecular Cell Biology*, 11(3):220–228.

[Newlon and Theis, 1993] Newlon, C. S. and Theis, J. F. (1993). The structure and function of yeast ARS elements. *Current opinion in genetics & development*, 3(5):752–8.

[Nguyen et al., 2001] Nguyen, V. Q., Co, C., and Li, J. J. (2001). Cyclin-dependent kinases prevent DNA re-replication through multiple mechanisms. *Nature*, 411(6841):1068–73.

[Nikolenko et al., 2013] Nikolenko, S. I., Korobeynikov, A. I., and Alekseyev, M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC genomics*, 14 Suppl 1(Suppl 1):S7.

[Nordman and Orr-Weaver, 2012] Nordman, J. and Orr-Weaver, T. L. (2012). Regulation of DNA replication during development. *Development*, 139(3):455–464.

[Nordman et al., 2014] Nordman, J. T., Kozhevnikova, E. N., Verrijzer, C. P., Pindyurin, A. V., Andreyeva, E. N., Shloma, V. V., Zhimulev, I. F., and Orr-Weaver, T. L. (2014). DNA copy-number control through inhibition of replication fork progression. *Cell reports*, 9(3):841–9.

[Norio and Schildkraut, 2001] Norio, P. and Schildkraut, C. L. (2001). Visualization of DNA replication on individual Epstein-Barr virus episomes. *Science (New York, N.Y.)*, 294(5550):2361–4.

[Orr et al., 1984] Orr, W., Komitopoulou, K., and Kafatos, F. C. (1984). Mutants suppressing in trans chorion gene amplification in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 81(12):3773–7.

[Orr-Weaver et al., 1989] Orr-Weaver, T. L., Johnston, C. G., and Spradling, A. C. (1989). The role of ACE3 in Drosophila chorion gene amplification. *The EMBO journal*, 8(13):4153–62.

[Orr-Weaver and Spradling, 1986] Orr-Weaver, T. L. and Spradling, A. C. (1986). Drosophila chorion gene amplification requires an upstream region regulating s18 transcription. *Molecular and cellular biology*, 6(12):4624–33.

[Osheim and Beyer, 1991] Osheim, Y. N. and Beyer, A. L. (1991). EM analysis of Drosophila chorion genes: amplification, transcription termination and RNA splicing. *Electron microscopy reviews*, 4(1):111–28.

[Osheim and Miller, 1983] Osheim, Y. N. and Miller, O. L. (1983). Novel amplification and transcriptional activity of chorion genes in Drosophila melanogaster follicle cells. *Cell*, 33(2):543–53.

[Osheim et al., 1988] Osheim, Y. N., Miller, O. L., and Beyer, A. L. (1988). Visualization of Drosophila melanogaster chorion genes undergoing amplification. *Molecular and cellular biology*, 8(7):2811–21.

[Pacek et al., 2006] Pacek, M., Tutter, A. V., Kubota, Y., Takisawa, H., and Walter, J. C. (2006). Localization of MCM2-7, Cdc45, and GINS to the Site of DNA Unwinding during Eukaryotic DNA Replication. *Molecular Cell*, 21(4):581–587.

[Pacek and Walter, 2004] Pacek, M. and Walter, J. C. (2004). A requirement for MCM7 and Cdc45 in chromosome unwinding during eukaryotic DNA replication. *The EMBO journal*, 23(18):3667–76.

[Paeschke et al., 2011] Paeschke, K., Capra, J. A., and Zakian, V. A. (2011). DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell*, 145(5):678–91.

[Painter, 1933] Painter, T. (1933). A new method for the study of chromosome rearrangements and the plotting of chromosome maps. *Science*, 78:585–586.

[Paixão et al., 2004] Paixão, S., Colaluca, I. N., Cubells, M., Peverali, F. A., Destro, A., Giadrossi, S., Giacca, M., Falaschi, A., Riva, S., and Biamonti, G. (2004). Modular structure of the human lamin B2 replicator. *Molecular and cellular biology*, 24(7):2958–67.

[Pak et al., 1997] Pak, D. T., Pflumm, M., Chesnokov, I., Huang, D. W., Kellum, R., Marr, J., Romanowski, P., and Botchan, M. R. (1997). Association of the origin recognition complex with heterochromatin and HP1 in higher eukaryotes. *Cell*, 91(3):311–23.

[Paranjape and Calvi, 2016] Paranjape, N. P. and Calvi, B. R. (2016). The Histone Variant H3.3 Is Enriched at Drosophila Amplicon Origins but Does Not Mark Them for Activation. *G3 (Bethesda, Md.)*, 6(6):1661–71.

[Pardue et al., 1970] Pardue, M. L., Gerbi, S. A., Eckhardt, R. A., and Gall, J. G. (1970). Cytological localization of DNA complementary to ribosomal RNA in polytene chromosomes of Diptera. *Chromosoma*, 29(3):268–290.

[Park et al., 2007] Park, E. A., MacAlpine, D. M., and Orr-Weaver, T. L. (2007). Drosophila follicle cell amplicons as models for metazoan DNA replication: A cyclinE mutant exhibits increased replication fork elongation. *Proceedings of the National Academy of Sciences*, 104(43):16739–16746.

[Park, 2009] Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80.

[Park and Asano, 2008] Park, S. Y. and Asano, M. (2008). The origin recognition complex is dispensable for endoreplication in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 105(34):12343–8.

[Pasero et al., 2002] Pasero, P., Bensimon, A., and Schwob, E. (2002). Single-molecule analysis reveals clustering and epigenetic regulation of replication origins at the yeast rDNA locus. *Genes & development*, 16(19):2479–84.

[Patel et al., 2006] Patel, P. K., Arcangioli, B., Baker, S. P., Bensimon, A., and Rhind, N. (2006). DNA replication origins fire stochastically in fission yeast. *Molecular biology of the cell*, 17(1):308–16.

[Pelizon et al., 1996] Pelizon, C., Diviacco, S., Falaschi, A., and Giacca, M. (1996). High-resolution mapping of the origin of DNA replication in the hamster dihydrofolate reductase gene domain by competitive PCR. *Molecular and cellular biology*, 16(10):5358–64.

[Pendleton et al., 2015] Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M. H.-Y., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C.-S., Guo, Y., Paxinos, E. E., Korbel, J. O., Darnell, R. B., McCombie, W. R., Kwok, P.-Y., Mason, C. E., Schadt, E. E., and Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, 12(8):780–786.

[Perkins et al., 2003] Perkins, T. T., Dalal, R. V., Mitsis, P. G., and Block, S. M. (2003). Sequence-dependent pausing of single lambda exonuclease molecules. *Science (New York, N.Y.)*, 301(5641):1914–8.

[Petryk et al., 2016] Petryk, N., Kahli, M., D'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C.-L., and Hyrien, O. (2016). Replication landscape of the human genome. *Nature Communications*, 7:10208.

[Picard et al., 2014] Picard, F., Cadoret, J.-C., Audit, B., Arneodo, A., Alberti, A., Battail, C., Duret, L., and Prioleau, M.-N. (2014). The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells. *PLoS genetics*, 10(5):e1004282.

[Potenski and Klein, 2014] Potenski, C. J. and Klein, H. L. (2014). How the misincorporation of ribonucleotides into genomic DNA can be both harmful and helpful to cells. *Nucleic Acids Research*, 42(16):10226–34.

[Poulson and Metz, 1938] Poulson, D. and Metz, C. (1938). Studies on the structure of nucleolus-forming regions and related structures in the giant salivary gland chromosomes of Diptera. *J. Morph.*, 63:363–395.

[Powell et al., 2015] Powell, S. K., MacAlpine, H. K., Prinz, J. A., Li, Y., Belsky, J. A., and MacAlpine, D. M. (2015). Dynamic loading and redistribution of the Mcm2-7 helicase complex through the cell cycle. *The EMBO Journal*, 34(4):531–543.

[Prasanth et al., 2010] Prasanth, S. G., Shen, Z., Prasanth, K. V., and Stillman, B. (2010). Human origin recognition complex is essential for HP1 binding to chromatin and heterochromatin organization. *Proceedings of the National Academy of Sciences of the United States of America*, 107(34):15093–8.

[Prioleau et al., 2003] Prioleau, M.-N., Gendron, M.-C., and Hyrien, O. (2003). Replication of the chicken beta-globin locus: early-firing origins at the 5' HS4 insulator and the rho- and betaA-globin genes show opposite epigenetic modifications. *Molecular and cellular biology*, 23(10):3536–49.

[Putnam et al., 2016] Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., Haussler, D., Rokhsar, D. S., and Green, R. E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research*, 26(3):342–50.

[Quick et al., 2015] Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Neal, K., Nye, K., Peters, T., De Pinna, E., Robinson, E., Struthers, K., Webber, M., Catto, A., Dallman, T. J., Hawkey, P., and Loman, N. J. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome biology*, 16(1):114.

[Quick et al., 2014] Quick, J., Quinlan, A. R., and Loman, N. J. (2014). A reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer. *GigaScience*, 3(1):22.

[Quinlan and Hall, 2010] Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2.

[Quinn et al., 2001] Quinn, L. M., Herr, A., McGarry, T. J., and Richardson, H. (2001). The Drosophila Geminin homolog: roles for Geminin in limiting DNA replication, in anaphase and in neurogenesis. *Genes & development*, 15(20):2741–54.

[Radding, 1966] Radding, C. M. (1966). Regulation of lambda exonuclease. I. Properties of lambda exonuclease purified from lysogens of lambda T11 and wild type. *Journal of molecular biology*, 18(2):235–50.

[Rand et al., 2016] Rand, A. C., Jain, M., Eizenga, J., Musselman-Brown, A., Olsen, H. E., Akeson, M., and Paten, B. (2016). Cytosine Variant Calling with High-throughput Nanopore Sequencing. *bioRxiv*.

[Rasch, 1970a] Rasch, E. M. (1970a). DNA cytophotometry of salivary gland nuclei and other tissue systems in dipteran larvae. In Wied, G. and Bahr, G., editors, *In Introduction to Quantitative Cytochemistry*, pages 357–397. Academic Press, New York.

[Rasch, 1970b] Rasch, E. M. (1970b). Two-wavelength cytophotometry of Sciara salivary gland chromosomes. In Wied, G. and Bahr, G., editors, *Introduction to Quantitative Cytochemistry*, volume 2, pages 335–355. Academic Press, New York.

[Rasch, 2006] Rasch, E. M. (2006). Genome size and determination of DNA content of the X chromosomes, autosomes, and germ line-limited chromosomes of Sciara coprophila. *Journal of morphology*, 267(11):1316–25.

[Rasmussen et al., 2016] Rasmussen, E. M., Vågbø, C. B., Münch, D., Krokan, H. E., Klungland, A., Amdam, G. V., and Dahl, J. A. (2016). DNA base modifications in honey bee and fruit fly genomes suggest an active demethylation machinery with species- and tissue-specific turnover rates. *Biochemistry and Biophysics Reports*, 6:9–15.

[Remus et al., 2004] Remus, D., Beall, E. L., and Botchan, M. R. (2004). DNA topology, not DNA sequence, is a critical determinant for Drosophila ORC-DNA binding. *The EMBO journal*, 23(4):897–907.

[Ribeiro et al., 2012] Ribeiro, F. J., Przybylski, D., Yin, S., Sharpe, T., Gnerre, S., Abouelleil, A., Berlin, A. M., Montmayeur, A., Shea, T. P., Walker, B. J., Young, S. K., Russ, C., Nusbaum, C., MacCallum, I., and Jaffe, D. B. (2012). Finished bacterial genomes from shotgun sequence data. *Genome Research*, 22(11):2270–2277.

[Richardson and Li, 2014] Richardson, C. D. and Li, J. J. (2014). Regulatory Mechanisms That Prevent Re-initiation of DNA Replication Can Be Locally Modulated at Origins by Nearby Sequence Elements. *PLoS Genetics*, 10(6):e1004358.

[Rieffel and Crouse, 1966] Rieffel, S. M. and Crouse, H. V. (1966). The elimination and differentiation of chromosomes in the germ line of sciara. *Chromosoma*, 19(3):231–76.

[Risse et al., 2015] Risse, J., Thomson, M., Patrick, S., Blakely, G., Koutsovoulos, G., Blaxter, M., and Watson, M. (2015). A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. *GigaScience*, 4(1):60.

[Robinson et al., 2011] Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1):24–6.

[Romero and Lee, 2008] Romero, J. and Lee, H. (2008). Asymmetric bidirectional replication at the human DBF4 origin. *Nature structural & molecular biology*, 15(7):722–9.

[Rowntree and Lee, 2006] Rowntree, R. K. and Lee, J. T. (2006). Mapping of DNA replication origins to noncoding genes of the X-inactivation center. *Molecular and cellular biology*, 26(10):3707–17.

[Royzman et al., 1999] Royzman, I., Austin, R. J., Bosco, G., Bell, S. P., and Orr-Weaver, T. L. (1999). ORC localization in Drosophila follicle cells and the effects of mutations in dE2F and dDP. *Genes & development*, 13(7):827–40.

[Rudkin and Corlette, 1957] Rudkin, G. T. and Corlette, S. L. (1957). Disproportionate synthesis of DNA in a polytene chromosome region. *Proceedings of the National Academy of Sciences of the United States of America*, 43(11):964–8.

[Saha et al., 2004] Saha, S., Shan, Y., Mesner, L. D., and Hamlin, J. L. (2004). The promoter of the Chinese hamster ovary dihydrofolate reductase gene regulates the activity of the local origin and helps define its boundaries. *Genes & development*, 18(4):397–410.

[Salzberg et al., 2012] Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., Marçais, G., Pop, M., and Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research*, 22(3):557–67.

[Sánchez, 2008] Sánchez, L. (2008). Sex-determining mechanisms in insects. *The International Journal of Developmental Biology*, 52(7):837–856.

[Sánchez, 2014] Sánchez, L. (2014). Sex-Determining Mechanisms in Insects Based on Imprinting and Elimination of Chromosomes. *Sexual Development*, 8(1-3):83–103.

[Santocanale and Diffley, 1996] Santocanale, C. and Diffley, J. F. (1996). ORC- and Cdc6-dependent complexes at active and inactive chromosomal replication origins in Saccharomyces cerevisiae. *The EMBO journal*, 15(23):6671–9.

[Santocanale et al., 1999] Santocanale, C., Sharma, K., and Diffley, J. F. (1999). Activation of dormant origins of DNA replication in budding yeast. *Genes & development*, 13(18):2360–4.

[Sasaki et al., 1999] Sasaki, T., Sawado, T., Yamaguchi, M., and Shinomiya, T. (1999). Specification of regions of DNA replication initiation during embryogenesis in the 65-kilobase DNApolalpha-dE2F locus of Drosophila melanogaster. *Molecular and cellular biology*, 19(1):547–55.

[Sawaya et al., 2015] Sawaya, S., Boocock, J., Black, M. A., and Gemmell, N. J. (2015). Exploring possible DNA structures in real-time polymerase kinetics using Pacific Biosciences sequencer data. *BMC Bioinformatics*, 16(1):21.

[Saxena and Dutta, 2005] Saxena, S. and Dutta, A. (2005). GemininCdt1 balance is critical for genetic stability. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 569(1):111–121.

[Schaarschmidt et al., 2004] Schaarschmidt, D., Baltin, J., Stehle, I. M., Lipps, H. J., and Knippers, R. (2004). An episomal mammalian replicon: sequence-independent binding of the origin recognition complex. *The EMBO journal*, 23(1):191–201.

[Schepers and Papior, 2010] Schepers, A. and Papior, P. (2010). Why are we where we are? Understanding replication origins and initiation sites in eukaryotes using ChIP-approaches. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 18(1):63–77.

[Schimke et al., 1986] Schimke, R. T., Sherwood, S. W., Hill, A. B., and Johnston, R. N. (1986). Overreplication and recombination of DNA in higher eukaryotes: potential consequences and biological implications. *Proceedings of the National Academy of Sciences of the United States of America*, 83(7):2157–61.

[Schmittgen and Livak, 2008] Schmittgen, T. D. and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative CT method. *Nature Protocols*, 3(6):1101–1108.

[Schmuck, 1934] Schmuck, M. (1934). The male somatic chromosome group in Sciara pauciseta. *Biol. Bull.*, 66:224–227.

[Schneiderman and Gilbert, 1964] Schneiderman, H. A. and Gilbert, L. I. (1964). Control of Growth and Development in Insects. *Science*, 143(3604).

[Schwartz et al., 1993] Schwartz, D., Li, X., Hernandez, L., Ramnarain, S., Huff, E., and Wang, Y. (1993). Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 262(5130).

[Schwarzbauer et al., 2012] Schwarzbauer, K., Bodenhofer, U., and Hochreiter, S. (2012). Genome-wide chromatin remodeling identified at GC-rich long nucleosome-free regions. *PloS one*, 7(11):e47924.

[Schwed et al., 2002] Schwed, G., May, N., Pechersky, Y., and Calvi, B. R. (2002). Drosophila minichromosome maintenance 6 is required for chorion gene amplification and genomic replication. *Molecular biology of the cell*, 13(2):607–20.

[Scott et al., 1997] Scott, R. S., Truong, K. Y., and Vos, J. M. (1997). Replication initiation and elongation fork rates within a differentially expressed human multicopy locus in early S phase. *Nucleic acids research*, 25(22):4505–12.

[Sequeira-Mendes et al., 2009] Sequeira-Mendes, J., Díaz-Uriarte, R., Apedaile, A., Huntley, D., Brockdorff, N., and Gómez, M. (2009). Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS genetics*, 5(4):e1000446.

[Sher et al., 2012] Sher, N., Bell, G. W., Li, S., Nordman, J., Eng, T., Eaton, M. L., Macalpine, D. M., and Orr-Weaver, T. L. (2012). Developmental control of gene copy number by repression of replication initiation and fork progression. *Genome research*, 22(1):64–75.

[Sherstyuk et al., 2014] Sherstyuk, V. V., Shevchenko, A. I., and Zakian, S. M. (2014). Epigenetic landscape for initiation of DNA replication. *Chromosoma*, 123(3):183–99.

[Shibata et al., 2012] Shibata, Y., Kumar, P., Layer, R., Willcox, S., Gagan, J. R., Griffith, J. D., and Dutta, A. (2012). Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. *Science (New York, N.Y.)*, 336(6077):82–6.

[Shim and Gu, 2012] Shim, J. and Gu, L.-Q. (2012). Single-molecule investigation of G-quadruplex using a nanopore sensor. *Methods*, 57(1):40–46.

[Shim et al., 2009] Shim, J. W., Tan, Q., and Gu, L.-Q. (2009). Single-molecule detection of folding and unfolding of the G-quadruplex aptamer in a nanopore nanocavity. *Nucleic acids research*, 37(3):972–82.

[Simão et al., 2015] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, 31(19).

[Simon et al., 2016] Simon, C. R., Siviero, F., and Monesi, N. (2016). Beyond DNA puffs: What can we learn from studying sciarids? *genesis*, 54(7):361–378.

[Simpson and Durbin, 2010] Simpson, J. T. and Durbin, R. (2010). Efficient construction of an assembly string graph using the FM-index. *Bioinformatics (Oxford, England)*, 26(12):i367–73.

[Simpson et al., 2009] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–23.

[Simpson et al., 2016] Simpson, J. T., Workman, R., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2016). Detecting DNA Methylation using the Oxford Nanopore Technologies MinION sequencer. *bioRxiv*.

[Simpson, 1990] Simpson, R. T. (1990). Nucleosome positioning can affect the function of a cis-acting DNA element in vivo. *Nature*, 343(6256):387–9.

[Siow et al., 2012] Siow, C. C., Nieduszynska, S. R., Müller, C. A., and Nieduszynski, C. A. (2012). OriDB, the DNA replication origin database updated and extended. *Nucleic acids research*, 40(Database issue):D682–6.

[Smith and Whitehouse, 2012] Smith, D. J. and Whitehouse, I. (2012). Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature*, 483(7390):434–8.

[Soares et al., 2003] Soares, M. A. M., Monesi, N., Basso, L. R., Stocker, A. J., Pa-Larson, M. L., and Lara, F. J. S. (2003). Analysis of the amplification and transcription of the C3-22 gene of Rhynchosciara americana (Diptera: Sciaridae) in transgenic lines of Drosophila melanogaster. *Chromosoma*, 112(3):144–151.

[Sovic et al., 2015] Sovic, I., Sikic, M., Wilm, A., Fenlon, S. N., Chen, S., and Nagarajan, N. (2015). Fast and sensitive mapping of error-prone nanopore sequencing reads with GraphMap. *bioRxiv*, 10.1101/02.

[Spradling, 1981] Spradling, A. C. (1981). The organization and amplification of two chromosomal domains containing Drosophila chorion genes. *Cell*, 27(1 Pt 2):193–201.

[Spradling and Mahowald, 1980] Spradling, A. C. and Mahowald, A. P. (1980). Amplification of genes for chorion proteins during oogenesis in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 77(2):1096–100.

[Spradling and Mahowald, 1981] Spradling, A. C. and Mahowald, A. P. (1981). A chromosome inversion alters the pattern of specific DNA replication in Drosophila follicle cells. *Cell*, 27(1 Pt 2):203–9.

[Stephenson et al., 2015] Stephenson, R., Hosler, M. R., Gavande, N. S., Ghosh, A. K., and Weake, V. M. (2015). Characterization of a Drosophila Ortholog of the Cdc7 Kinase. *Journal of Biological Chemistry*, 290(3):1332–1347.

[Stocker and Pavan, 1974] Stocker, A. and Pavan, C. (1974). The influence of ecdysterone on gene amplification, DNA synthesis, and puff formation in the salivary gland chromosomes of Rhynchosciara hollaenderi. *Chromosoma*, 45(3):295–319.

[Sun et al., 2008] Sun, J., Smith, L., Armento, A., and Deng, W.-M. (2008). Regulation of the endocycle/gene amplification switch by Notch and ecdysone signaling. *The Journal of cell biology*, 182(5):885–96.

[Suzuki et al., 2016] Suzuki, Y., Korlach, J., Turner, S. W., Tsukahara, T., Taniguchi, J., Qu, W., Ichikawa, K., Yoshimura, J., Yurino, H., Takahashi, Y., Mitsui, J., Ishiura, H., Tsuji, S., Takeda, H., and Morishita, S. (2016). AgIn: measuring the landscape of CpG methylation of individual repetitive elements. *Bioinformatics (Oxford, England)*, 32(19):2911–9.

[Swift, 1962] Swift, H. (1962). Nucleic acids and cell morphology in dipteran salivary glands. In Allen, J., editor, *Molecular Control of Cellular Activity*, pages 73–125. McGraw-Hill, New York.

[Swimmer et al., 1989] Swimmer, C., Delidakis, C., and Kafatos, F. C. (1989). Amplification-control element ACE-3 is important but not essential for autosomal chorion gene amplification. *Proceedings of the National Academy of Sciences of the United States of America*, 86(22):8823–7.

[Szalay and Golovchenko, 2015] Szalay, T. and Golovchenko, J. A. (2015). A de novo DNA Sequencing and Variant Calling Algorithm for Nanopores. *bioRxiv*.

[Szüts and Krude, 2004] Szüts, D. and Krude, T. (2004). Cell cycle arrest at the initiation step of human chromosomal DNA replication causes DNA damage. *Journal of Cell Science*, 117(21).

[Takahashi et al., 2004] Takahashi, T. S., Yiu, P., Chou, M. F., Gygi, S., and Walter, J. C. (2004). Recruitment of Xenopus Scc2 and cohesin to chromatin requires the pre-replication complex. *Nature cell biology*, 6(10):991–6.

[Takawa et al., 2012] Takawa, M., Cho, H.-S., Hayami, S., Toyokawa, G., Kogure, M., Yamane, Y., Iwai, Y., Maejima, K., Ueda, K., Masuda, A., Dohmae, N., Field, H. I., Tsunoda, T., Kobayashi, T., Akasu, T., Sugiyama, M., Ohnuma, S.-i., Atomi, Y., Ponder, B. A. J., Nakamura, Y., and Hamamoto, R. (2012). Histone Lysine Methyltransferase SETD8 Promotes Carcinogenesis by Deregulating PCNA Expression. *Cancer Research*, 72(13):3217–3227.

[Takayama et al., 2014] Takayama, S., Dhahbi, J., Roberts, A., Mao, G., Heo, S.-J., Pachter, L., Martin, D. I. K., and Boffelli, D. (2014). Genome methylation in D. melanogaster is found at specific short motifs and is independent of DNMT2 activity. *Genome Research*, 24(5):821–830.

[Takayama et al., 2003] Takayama, Y., Kamimura, Y., Okawa, M., Muramatsu, S., Sugino, A., and Araki, H. (2003). GINS, a novel multiprotein complex required for chromosomal DNA replication in budding yeast. *Genes & Development*, 17(9):1153–1165.

[Tao et al., 2000] Tao, L., Dong, Z., Leffak, M., Zannis-Hadjopoulos, M., and Price, G. (2000). Major DNA replication initiation sites in the c-myc locus in human cells. *Journal of cellular biochemistry*, 78(3):442–57.

[Técher et al., 2013] Técher, H., Koundrioukoff, S., Azar, D., Wilhelm, T., Carignon, S., Brison, O., Debatisse, M., and Le Tallec, B. (2013). Replication dynamics: biases and robustness of DNA fiber analysis. *Journal of molecular biology*, 425(23):4845–55.

[Theis and Newlon, 1997] Theis, J. F. and Newlon, C. S. (1997). The ARS309 chromosomal replicator of Saccharomyces cerevisiae depends on an exceptional ARS consensus sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 94(20):10786–91.

[Thomer et al., 2004] Thomer, M., May, N. R., Aggarwal, B. D., Kwok, G., and Calvi, B. R. (2004). Drosophila double-parked is sufficient to induce re-replication during development and is regulated by cyclin E/CDK2. *Development*, 131(19).

[Thorvaldsdóttir et al., 2013] Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–92.

[Timp et al., 2012] Timp, W., Comer, J., and Aksimentiev, A. (2012). DNA base-calling from a nanopore using a Viterbi algorithm. *Biophysical journal*, 102(10):L37–9.

[Tower, 2004] Tower, J. (2004). Developmental gene amplification and origin regulation. *Annual review of genetics*, 38:273–304.

[Underwood et al., 1990] Underwood, E. M., Briot, A. S., Doll, K. Z., Ludwiczak, R. L., Otteson, D. C., Tower, J., Vessey, K. B., and Yu, K. (1990). Genetics of 51D-52A, a region containing several maternal-effect genes and two maternal-specific transcripts in Drosophila. *Genetics*, 126(3):639–50.

[Urban et al., 2015a] Urban, J. M., Bliss, J., Lawrence, C. E., and Gerbi, S. A. (2015a). Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. *bioRxiv*, page 019281.

[Urban et al., 2015b] Urban, J. M., Foulk, M. S., Casella, C., and Gerbi, S. A. (2015b). The hunt for origins of DNA replication in multicellular eukaryotes. *F1000Prime Reports*, 7(30).

[Urban et al., 2016] Urban, J. M., Yamamoto, Y., Kadota, L., Lee, A., Bliss, J. E., Smith, H. S., DiBartolomeis, S. M., and Gerbi, S. A. (2016). The DNA puffs of Sciara coprophila before, during, and after developmentally programmed intrachromosomal DNA amplification. In Urban, J. M. and Gerbi, S. A., editors, *The genome and DNA puff sequences of the fungus fly, Sciara coprophila, and genome-wide methods for studying DNA replication.* Brown University, Providence.

[Urnov et al., 2002] Urnov, F. D., Liang, C., Blitzblau, H. G., Smith, H. S., and Gerbi, S. A. (2002). A DNase I hypersensitive site flanks an origin of DNA replication and amplification in Sciara. *Chromosoma*, 111(5):291–303.

[Valenzuela et al., 2011] Valenzuela, M. S., Chen, Y., Davis, S., Yang, F., Walker, R. L., Bilke, S., Lueders, J., Martin, M. M., Aladjem, M. I., Massion, P. P., and Meltzer, P. S. (2011). Preferential localization of human origins of DNA replication at the 5'-ends of expressed genes and at evolutionarily conserved DNA sequences. *PloS one*, 6(5):e17308.

[Valton et al., 2014] Valton, A.-L., Hassan-Zadeh, V., Lema, I., Boggetto, N., Alberti, P., Saintomé, C., Riou, J.-F., and Prioleau, M.-N. (2014). G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *The EMBO journal*, 33(7):732–46.

[van Oijen et al., 2003] van Oijen, A. M., Blainey, P. C., Crampton, D. J., Richardson, C. C., Ellenberger, T., and Xie, X. S. (2003). Single-molecule kinetics of lambda exonuclease reveal base dependence and dynamic disorder. *Science (New York, N.Y.)*, 301(5637):1235–8.

[VanBuren et al., 2015] VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., Spittle, K., Hall, R., Gu, J., Lyons, E., Freeling, M., Bartels, D., Ten Hallers, B., Hastie, A., Michael, T. P., and Mockler, T. C. (2015). Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature*, 527(7579):508–511.

[Vaser et al., 2016] Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2016). Fast and accurate de novo genome assembly from long uncorrected reads. *bioRxiv*.

[Vashee et al., 2003] Vashee, S., Cvetic, C., Lu, W., Simancek, P., Kelly, T. J., and Walter, J. C. (2003). Sequence-independent DNA binding and replication initiation by the human origin recognition complex. *Genes & development*, 17(15):1894–908.

[Vassilev et al., 1990] Vassilev, L. T., Burhans, W. C., and DePamphilis, M. L. (1990). Mapping an origin of DNA replication at a single-copy locus in exponentially proliferating mammalian cells. *Molecular and cellular biology*, 10(9):4685–9.

[Vaughn et al., 1990] Vaughn, J. P., Dijkwel, P. A., and Hamlin, J. L. (1990). Replication initiates in a broad zone in the amplified CHO dihydrofolate reductase domain. *Cell*, 61(6):1075–87.

[Venter et al., 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel,

J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The Sequence of the Human Genome. *Science*, 291(5507).

[Vezzi et al., 2012] Vezzi, F., Narzisi, G., and Mishra, B. (2012). Feature-by-feature–evaluating de novo sequence assembly. *PloS one*, 7(2):e31002.

[Vujcic et al., 1999] Vujcic, M., Miller, C. A., and Kowalski, D. (1999). Activation of silent replication origins at autonomously replicating sequence elements near the HML locus in budding yeast. *Molecular and cellular biology*, 19(9):6098–109.

[Walker et al., 2014] Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963.

[Wang et al., 2004] Wang, L., Lin, C.-M., Brooks, S., Cimbora, D., Groudine, M., and Aladjem, M. I. (2004). The human beta-globin replication initiation region consists of two modular independent replicators. *Molecular and cellular biology*, 24(8):3373–86.

[Warren et al., 2015a] Warren, R. L., Vandervalk, B. P., Jones, S. J., and Birol, I. (2015a). LINKS: Scaffolding genome assemblies with kilobase-long nanopore reads. *bioRxiv*.

[Warren et al., 2015b] Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., and Birol, I. (2015b). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, 4(1):35.

[Weaver et al., 2010] Weaver, S., Dube, S., Mir, A., Qin, J., Sun, G., Ramakrishnan, R., Jones, R. C., and Livak, K. J. (2010). Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods*, 50:271–276.

[Weisenfeld et al., 2014] Weisenfeld, N. I., Yin, S., Sharpe, T., Lau, B., Hegarty, R., Holmes, L., Sogoloff, B., Tabbaa, D., Williams, L., Russ, C., Nusbaum, C., Lander, E. S., MacCallum, I., and Jaffe, D. B. (2014). Comprehensive variation discovery in single human genomes. *Nature Genetics*, 46(12):1350–1355.

[Whittaker et al., 2000] Whittaker, A. J., Royzman, I., and Orr-Weaver, T. L. (2000). Drosophila double parked: a conserved, essential replication protein that colocalizes with the origin recognition complex and links DNA replication with mitosis and the down-regulation of S phase transcripts. *Genes & development*, 14(14):1765–76.

[Wong and Huppert, 2009] Wong, H. M. and Huppert, J. L. (2009). Stable G-quadruplexes are found outside nucleosome-bound regions. *Molecular bioSystems*, 5(12):1713–9.

[Woodward et al., 2006] Woodward, A. M., Göhler, T., Luciani, M. G., Oehlmann, M., Ge, X., Gartner, A., Jackson, D. A., and Blow, J. J. (2006). Excess Mcm2-7 license dormant origins of replication that can be used under conditions of replicative stress. *The Journal of cell biology*, 173(5):673–83.

[Wu et al., 1993] Wu, N., Liang, C., DiBartolomeis, S. M., Smith, H. S., and Gerbi, S. A. (1993). Developmental progression of DNA puffs in Sciara coprophila: amplification and transcription. *Developmental biology*, 160(1):73–84.

[Wyrick et al., 2001] Wyrick, J. J., Aparicio, J. G., Chen, T., Barnett, J. D., Jennings, E. G., Young, R. A., Bell, S. P., and Aparicio, O. M. (2001). Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. *Science (New York, N.Y.)*, 294(5550):2357–60.

[Xie and Orr-Weaver, 2008] Xie, F. and Orr-Weaver, T. L. (2008). Isolation of a Drosophila amplification origin developmentally activated by transcription. *Proceedings of the National Academy of Sciences*, 105(28):9651–9656.

[Xu et al., 2012] Xu, J., Yanagisawa, Y., Tsankov, A. M., Hart, C., Aoki, K., Kommajosyula, N., Steinmann, K. E., Bochicchio, J., Russ, C., Regev, A., Rando, O. J., Nusbaum, C., Niki, H., Milos, P., Weng, Z., and Rhind, N. (2012). Genome-wide identification and characterization of replication origins by deep sequencing. *Genome biology*, 13(4):R27.

[Xu et al., 2006] Xu, W., Aparicio, J. G., Aparicio, O. M., and Tavaré, S. (2006). Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in S. cerevisiae. *BMC genomics*, 7:276.

[Yamamoto et al., 2015] Yamamoto, Y., Bliss, J., and Gerbi, S. A. (2015). Whole Organism Genome Editing: Targeted Large DNA Insertion via ObLiGaRe Nonhomologous End-Joining in Vivo Capture. *G3 (Bethesda, Md.)*, 5(9):1843–7.

[Yang et al., 2013] Yang, X., Chockalingam, S. P., and Aluru, S. (2013). A survey of error-correction methods for next-generation sequencing. *Briefings in bioinformatics*, 14(1):56–66.

[Yao et al., 2007] Yao, Y., Wang, Q., Hao, Y.-h., and Tan, Z. (2007). An exonuclease I hydrolysis assay for evaluating G-quadruplex stabilization by small molecules. *Nucleic acids research*, 35(9):e68.

[Yarosh and Spradling, 2014] Yarosh, W. and Spradling, A. C. (2014). Incomplete replication generates somatic DNA alterations within Drosophila polytene salivary gland cells. *Genes & development*, 28(16):1840–55.

[Ye et al., 2016] Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. S. (2016). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Scientific Reports*, 6:31900.

[Yoon et al., 1995] Yoon, Y., Sanchez, J. A., Brun, C., and Huberman, J. A. (1995). Mapping of replication initiation sites in human ribosomal DNA by nascent-strand abundance analysis. *Molecular and cellular biology*, 15(5):2482–9.

[Zentner et al., 2011] Zentner, G. E., Saiakhova, A., Manaenkov, P., Adams, M. D., and Scacheri, P. C. (2011). Integrative genomic analysis of human ribosomal DNA. *Nucleic acids research*, 39(12):4949–60.

[Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–9.

[Zhang et al., 2015] Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P., Liu, J., Li, C., Liu, B., Luo, Y., Zhu, Y., Zhang, N., He, S., He, C., Wang, H., and Chen, D. (2015). N6-Methyladenine DNA Modification in Drosophila. *Cell*, 161(4):893–906.

[Zhang and Tower, 2004] Zhang, H. and Tower, J. (2004). Sequence requirements for function of the Drosophila chorion gene locus ACE3 replicator and ori-beta origin elements. *Development (Cambridge, England)*, 131(9):2089–99.

[Zhang et al., 2012] Zhang, M., Zhang, Y., Scheuring, C. F., Wu, C.-C., Dong, J. J., and Zhang, H.-B. (2012). Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *Nature Protocols*, 7(3):467–478.

[Zhang et al., 2008] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137.

[Zhu et al., 2006] Zhu, J., Chen, L., Sun, G., and Raikhel, A. S. (2006). The Competence Factor Ftz-F1 Potentiates Ecdysone Receptor Activity via Recruiting a p160/SRC Coactivator. *Molecular and Cellular Biology*, 26(24):9402–9412.

[Zimin et al., 2013] Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics (Oxford, England)*, 29(21):2669–77.

[Zou and Stillman, 1998] Zou, L. and Stillman, B. (1998). Formation of a preinitiation complex by S-phase cyclin CDK-dependent loading of Cdc45p onto chromatin. *Science (New York, N.Y.)*, 280(5363):593–6.