Bayesian statistical inference of non-allelic homologous recombination in the human genome using high-throughput sequencing data

by

Matthew Parks

B.Sc., Lehigh University; Bethlehem, PA, 2009

M.Sc., Brown University; Providence, RI, 2010

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2014

© Copyright 2014 by Matthew Parks

This dissertation by Matthew Parks is accepted in its present form
by The Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Charles Lawrence, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

William Thompson, Ph.D., Reader

Date_____ _____

Benjamin Raphael, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Peter Weber, Dean of the Graduate School

Vitae

Matthew Parks was born on June 1, 1987 in Carmel, NY. He graduated from Lehigh University in May 2009 with a Bachelor of Science degree in Mathematics, with honors. In August 2009, he began studying at the Division of Applied Mathematics at Brown University. Therein, Matthew received a Master of Science degree in Applied Mathematics in May 2010, and defended his Ph.D. dissertation on April 25, 2014.

Acknowledgements

First and foremost, I would like to thank my advisor Chip Lawrence for his guidance. Chip promoted true freedom of thought and was always willing to listen to my ideas, even when an idea was flawed or only partially developed. When I was stuck or did not know how to proceed, Chip pointed me in the right direction without solving the problem for me. On the personal side, Chip's encouragement, support, and positive attitude were consistent throughout the entirety of my research with him. Knowing that Chip would *always* be smiling whenever we had a meeting made the most important part of my graduate career truly enjoyable.

I would like to thank Bill Thompson for his help, advice, and support on several topics in my dissertation. Bill was always available for helpful discussion and developing ideas. And when I was stuck on a certain part of my project after having exhausted every approach, Bill was happy to dig into the details (including my poorly written, frantically debugged code) and helped me solve the problem.

I would like to thank Ben Raphael for his guidance through the complicated world of bioinformatics. Having no previous knowledge or experience in computational biology, Ben was very helpful for getting my bearings and learning the state-of-the-art approaches and technologies in the realm of structural variation.

I am grateful to Björn Sandstede for advocating on my behalf in securing the Brown/Tougaloo Faculty Fellowship. The experience was well worth the adminis-

trative hurdles, and would likely not have been possible without Björn's persistent support on my behalf.

I would like to thank my current and former colleagues in the Lawrence and Raphael lab groups and in the Division of Applied Mathematics for their friendship and camaraderie through my years at Brown. Graduate school is a challenging time, but it is much less stressful when your peers are friendly, encouraging, and "we are all in this together".

Finally, I would like to thank the other applied math professors and the Division of Applied Mathematics as a whole. The Division of Applied Mathematics is unique in the kind of environment it provides to its graduate students: healthy, supportive, stimulating, and understanding. Somebody told me during my very early days at Brown that "once you are admitted to this department, you are expected to graduate". That kind of attitude is priceless.

Abstract of " Bayesian statistical inference of non-allelic homologous recombination in the human genome using high-throughput sequencing data " by Matthew Parks, Ph.D., Brown University, May 2014

Non-allelic homologous recombination (NAHR) plays a major role in genome rearrangement and is implicated in numerous genetic disorders. But detection of NAHR poses a serious technical challenge because its breakpoints occur in nearly identical regions of highly homologous repeats. While a few structural variation algorithms identify rearrangements in repeat regions, reliable detection of NAHR remains out of reach. We present a probabilistic model of NAHR and demonstrate its ability to find previously-undetected NAHR rearrangements from low coverage sequencing data. We identify a reliable subset of calls and discuss their significance: segregation of NAHR in different populations, effects on highly studied genes such as GBA and CYP2E1, and associated features of NAHR.

Contents

List of Tables

List of Figures

Chapter One

Introduction

Non-Allelic Homologous Recombination (NAHR) is a biological mechanism for repairing broken chromosomes that results in gross genome rearrangements. A significant portion ($\sim 20\% - 50\%$, conservatively) of all genome rearrangement in humans, also called structural variation, is thought to be the result of NAHR [38, 37, 3, 57, 18]. Understanding and detecting NAHR in individuals provides valuable insight for a wide variety of genomic disorders, disease susceptibilities, and cancers [29, 13, 27, 74, 66, 60, 12, 36, 83].

Despite its importance and prevalence, NAHR is challenging to detect with either computational or biological techniques. The difficulty stems from four crucial properties: 1) NAHR is mediated by highly homologous repeats, 2) there are thousands of repeats across the human genome [8], 3) the breakpoints of an NAHR-mediated rearrangement are *always* at homologous positions of homologous repeats, 4) NAHR is capable of producing inversions, deletions, duplications, and translocations. Detection of NAHR thus requires a careful treatment of repetitive regions from large areas of the human genome. Currently, repetitive regions are a major weakness of biological and computational techniques for detection and verification[6].

Biological techniques for discovering instances of NAHR are frustrated by these properties. Array-based approaches, namely array CGH and SNP micro-arrays, fail in repeat regions because they assume diploid copies of each locus in the reference genome [6]. The FISH technique is limited to very large ($\geq$ 500 kb) rearrangements and is low-resolution [6]. Design of reliable primers for PCR-based techniques of genotyping rearrangement breakpoints is difficult because of the existence of other highly homologous copies of the repeat. Also, PCR-based techniques are only practical for a relatively few number of rearrangements [6], yet there are many repeats in the human genome and hence many potential sites of NAHR. While validations of NAHR have been done using these techniques [60, 38, 37], they are not used on

nearly the same scale as high-throughput sequencing.

High-throughput genome sequencing is a popular, powerful, and cost-efficient source of data for learning about genomes and inferring the variation between individuals. Typically, structural variations in an individual with respect to the reference are inferred from sequencing data by the mappings of paired-end reads to the reference genome. Nearly all of the most popular structural variation algorithms identify candidate structural variations using *only* "discordantly" [80] mapped paired-end reads, including VariationHunter, VariationHunter-CR, HYDRA, GASV, GASV-pro, CNVer, BreakDancer, PEMer, and Pindel[1] [30, 31, 65, 70, 71, 56, 82, 14, 56]. But the biology of the NAHR mechanism *requires* that NAHR breakpoints occur at homologous positions of homologous repeats, making it highly likely that paired-end reads generated from NAHR breakpoints are mapped *concordantly* to the original repeats. Thus, the very biology of NAHR implies that NAHR will very often go undetected by algorithms that rely on discordant reads, i.e. NAHR will be largely undetectable by most existing structural variation detection algorithms.

Directly modeling NAHR offers major advantages over naïve analysis of discordant reads via comparison of their mapping locations. The "rules of NAHR"[54, 44, 60] are studied and well-characterized [29, 13, 27, 53], and provide a structured framework for detecting NAHR from short-reads that is consistent with the biological mechanism. They determine *where* NAHR rearrangements may occur, *what* types of rearrangements are possible, and the exact location and sequence composition of the breakpoints. This has major implications for the analysis of sequencing data: the information provided by the rules of NAHR allow us to fully and exactly construct hypothetically rearranged genomes from which all reads were theoretically

[1]Pindel actually requires that one mate be mapped uniquely and the other mate to be unmapped. This is idea is sufficiently similar to discordantly mapped reads for inclusion here.

generated concordantly, thus bypassing entirely the notion of "discordant" read mappings. Further, the characteristics of NAHR and repeats indicate a "natural" way to evaluate read-depth, freeing our model from relying on arbitrary bins and sliding windows. Thus, founding our model on the "rules of NAHR" provides a completely new approach to high-throughput sequencing analysis for structural variation.

Data generated from repetitive regions requires extremely careful analysis, since, by definition, there is very little signal to distinguish highly homologous repeats. Any computational model must be sufficiently sensitive to recognize such subtle differences in signal, and further, accumulate these differences and make inference informed by the entirety of the competing, often conflicting signals. Probabilistic models offer a natural way to capture such subtleties, and Bayesian models, in particular, use probability theory to weigh competing signals against each other to draw inference.

We have constructed generative probabilistic models of both the NAHR mechanism and the high-throughput sequencing process. We use these models to perform a Bayesian statistical inference on the occurrence of NAHR-mediated rearrangements in human genomes, focusing on deletions and duplications. Our model makes three important contributions. First, it provides the first method that incorporates the biology of the NAHR mechanism explicitly and its implications for high-throughput sequencing. Second, it provides a mathematically rigorous framework for analyzing high-throughput sequencing data in repetitive regions of the genome. Lastly, it detects hitherto unreachable rearrangements in the human genome.

We analyzed publicly available, low coverage, short-read paired-end sequencing data for 44 individuals from the 1000 Genomes Project using our model. Due to the repetitive nature of these regions and the limitations of experimental validation

technology mentioned above [6], we restricted our calls to a reliable subset using a separate statistical test in lieu of experimental validation. Most of our results were novel when compared against several recent structural variation validation studies. Using our reliable subset of calls, we discuss important implications for several highly studied genes, and we draw additional inference on general characteristics of NAHR.

Chapter Two

Biology of NAHR

Here we briefly review the NAHR mechanism in humans. More detailed reviews can be found in [29, 13].

Allelic homologous recombination (AHR) repairs double-stranded breaks (DSBs) in chromosomes by using the allele on the sister chromatid as a template. This mechanism is highly faithful because the allelic region of the sister chromatid is a nearly-exact (up to polymorphism) copy of the DNA lost in the DSB. The crucial step in the AHR repair process is the homology search for locating the allele, which is to serve as a template for repair of the DSB via PCR. If the DSB occurs in a "unique" region of the genome, then the homology search will almost certainly find the allelic position on the sister chromatid, and AHR will proceed as usual. But if the DSB occurs in or near a repeat, then the homology search may instead find a paralog of the repeat, making the ensuing repair *non-allelic*. If the ensuing double Holliday junction is resolved via crossover, then non-allelic homologous recombination (NAHR) has occurred. The class of repeats that we will examine for NAHR in this study have been termed *low-copy repeats* (LCRs) and segmental duplications[8, 24].

The type of rearrangement depends on the location of the paralog with respect to the chromatid with the DSB, and on the orientation of the two mediating repeats (positive orientation if the genomic index of both sequences increases along the alignment profile of the pair of repeats). Intra-chromatid NAHR results in deletions if the repeats are positively-oriented, and inversions if negatively-oriented. Inter-chromatid NAHR between positively-oriented repeats results in deletions and duplications. Translocations result from positively-oriented inter-chromosomal NAHR or negatively-oriented repeats on different arms of the same chromosome.

The catch with NAHR is that since the DSB region and the template for repair are highly homologous repeats, then the breakpoint region "looks" almost exactly

the same before and after the rearrangement. Luckily, for any pair of repeats, there is often a small set of SNPs and short indels that distinguish the repeats; we refer to them as *variational positions* (VPs). Being conscious of the variational position pattern, we see that since NAHR involves a crossover, then at the breakpoint the VP pattern switches from one repeat's VP pattern to the other's. Indeed, the repeat containing the breakpoint is actually a new repeat; it is a *hybrid* LCR, composed of part of each of the two repeats that mediated the NAHR rearrangement, joined at the breakpoint. See Figure 2.1a-b contains a schematic illustration of variational positions and the resulting hybrid LCR for an NAHR duplication.

We can further characterize each NAHR rearrangement according to the VP pattern(s) exhibited in the resulting hybrid repeat(s). Of the two repeats that mediate an NAHR rearrangement, denote the one with a smaller genomic index as A and the other as B. Deletions result in a hybrid repeat with VP pattern $A \to B$, with the original A and B repeats deleted. Duplications preserve both A and B, and additionally create a hybrid with VP pattern $B \to A$. Inversions and translocations replace repeat A by a hybrid with pattern $A \to B$, and replace repeat B by a hybrid with pattern $B \to A$. Altogether, the resulting hybrid VP patterns and the types of rearrangements give us the "rules of NAHR".

Lastly, it is important to mention the closely-related gene conversion mechanism. Gene conversion follows exactly the same pathway as NAHR, but the double Holliday junctions are resolved via the non-crossover outcome. This does not produce a large-scale rearrangement as in NAHR; instead, it merely replaces a short tract of the genome by a copy of a donor tract. If B donates to A, then A is replaced by the hybrid $A \to B \to A$, and vice-versa. Notice that gene conversion events have *two* breakpoints, or rather, two switches in VP pattern. Although we do not focus on gene conversion in this study, it is necessary to include gene conversion in any model of

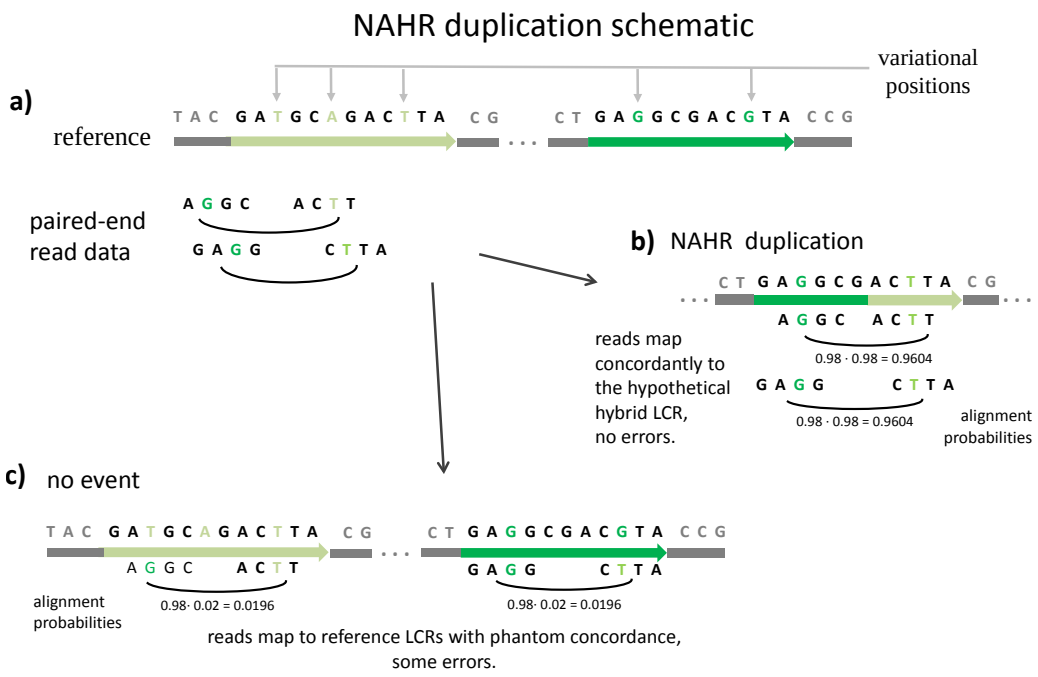


Figure 2.1: Schematic example of an NAHR duplication with paired-end read data. **(a)** A schematic reference genome and paired-end read data. The dark green and light green regions are homologous LCRs which form a potential NAHR event locus. Green nucleotides are the variational positions that distinguish the two LCRs. We consider two possible outcomes: no event and an NAHR duplication. **(b)** The hybrid LCR formed from the NAHR duplication event is shown with aligned paired-end reads. The hybrid LCR is novel to the individual - it does not exist in the original reference genome. Notice that the variational positions switch from "dark green" to "light green" after the breakpoint in the hybrid LCR. For simplicity in this schematic, we calculate the probability of a paired-end read's alignment according to the agreement between its mates' bases at the variational positions, although in the algorithm a full alignment probability is calculated. Suppose the probability of a read-error is 2%. Then the likelihood that the paired-end reads came from the hybrid LCR is $0.96^2 = 0.9604$ for each read. **(c)** If no NAHR event occurs at this locus, then this locus of the individual's genome is the same as in the reference. Notice that the paired-end reads are also aligned concordantly to these LCRs, albeit with a small number of errors at the variational positions. Here, the likelihood of the mediating LCR is $0.98 \cdot 0.02 = 0.0196$ for each paired-end read.

NAHR since the breakpoint signals of a gene conversion event mimic those of NAHR events. For example, a gene conversion event producing the hybrid $A \to B \to A$ could be easily mistaken as an NAHR deletion with hybrid $A \to B$ or an NAHR duplication with hybrid $B \to A$.

Chapter Three

A Bayesian statistical model for NAHR

The challenge in modeling NAHR is the repeats. Standard structural variation approaches encounter great difficulty due to the substantial uncertainty involved in mapping reads to repetitive regions. The problem is made even more challenging by the fact that the breakpoints of NAHR rearrangements are known to often occur at corresponding homologous positions deep inside of the highly homologous mediating repeats.

These challenges necessitate a novel framework for analyzing repeats and a new approach to evaluating reads in repetitive regions. That is, we must develop *two* new approaches: first, we abandon the traditional notion of a "repeat" family and instead define a new kind of relationships between repeats; and second, we do not look for anomalies among mapped reads to indicate a breakpoint, but rather calculate the probability of *all* of the data given a hypothetical rearranged genome. These two approaches are discussed in detail below.

3.1 A framework for repeats

Data analysis with repetitive sequences is a notoriously difficult task in computational biology. Because the sequences are so highly homologous, many biological techniques are unreliable or useless when applied to repeats [6]. Similarly, computational strategies are frustrated because data involving sequences cannot be confidently mapped to a single repeat out of many paralogous alternatives.

The relationships between homologous repeats can be rather complex. For example, a repeat may be homologous to a substring of another repeat, may be homologous to another repeat except for an insertion of nontrivial length, may be a mosaic

composition of several repeats, or may show self-similarity (substrings of the same repeat are homologous). Figure 3.1 illustrates several complicated relationships. We may also imagine scenarios in which, say, repeat A_1 is 95% homologous to A_2, which is 95% homologous to repeat A_3, etc., yet A_1 and A_{10}, say, are not very homologous at all!

Figure 3.1: A schematic example of complicated relationships encountered between repeats. For comparison, the nine sequences are shown in a multiple alignment. Note the long insertion in the first repeat (top) relative to the second and third repeats. The bottom six repeats are homologous to different non-disjoint substrings of the three longer repeats.

Due to these complexities, many classic bioinformatic techniques for characterizing groups of sequences fail when dealing with repeats. Specifically, using multiple sequences to construct a profile or perform a multiple alignment is not well-defined because extensive self-similarity within a sequence possibly results in non-linear multiple alignments or profiles [43]. Thus, trying to find "repeat families" (a complete set of mutually homologous repeats), though conceptually convenient, is problematic in theory and in practice.

We abandon the notion of repeat families and sequence profiles as methods for characterizing repeats. Instead, we resort to the most fundamental relationship between sequences: the pairwise alignment[1]. Rather than attempt to characterize whole groups of sequences at once, we instead consider only two sequences at a time, defining their relationship by their pairwise alignment. Because we only consider

[1]For consistency, whenever we speak of alignments of LCRs, we orient and complement as necessary the LCR with greater reference genome coordinates so that it can be properly aligned to the LCR with smaller reference genome coordinates.

pairs of highly homologous repeats at a time, we avoid the issues faced by profiles and multiple alignments described above. Relying on pairwise alignments alone, we associate regions of the genome according to homology and construct well-defined homology maps between sequences. As we will see below, this also gives rise to very natural probabilistic graphical model and a principled approach to evaluating the data.

3.1.1 The Human Segmental Duplication database

Following the completion of the early drafts of the human reference genome, Bailey, et. al. sought to identify segmental duplications in the human genome, where they defined a segmental duplication as a sequence of length ≥ 1 kb which shares $\geq 90\%$ sequence identity with another (disjoint) ≥ 1 kb sequence, i.e. long and highly homologous repeats [8]. The length criteria distinguishes segmental duplications from the many short, high-copy repeats (e.g. *Alu*s, etc.) also present in the human genome. Thus, the segmental duplications found and studied by Bailey, et. al. are also called *low-copy repeats* (LCRs) [24]; we will use these terms interchangeably. Short, high-copy repeats frustrate local alignment searches for long, highly homologous repeats in the human genome, as the many alignments between highly homologous short high-copy repeats obfuscate relationships between longer sequences. Bailey, et. al. accounted for short, high-copy repeats by employing a technique they called "fuguization", wherein they removed short, high-copy repeats and then proceeded to perform local alignments on the remaining sequence, enabling them to find longer stretches of homologous sequence. After identifying these longer repeats, they re-inserted the high-copy repeats to obtain the full sequences and pairwise relationships.

To account for long insertions between otherwise homologous sequences, they also relaxed gap extension penalties when performing the local alignments. Finally, Bailey, et. al. employed custom procedures for determining and refining the ends of highly homologous sequences they discovered [8].

The result of Bailey, et. al.'s procedure is the Human Segmental Duplication Database (HSDD), which lists all pairs of sequences in the human reference genome that are ≥ 1 kb and $\geq 90\%$ identity and gives their coordinates[2]. We assume that the HSDD is indeed complete (i.e. contains all LCRs and all possible pairs of homologous LCRs), and so we will refer to the repeats listed in the HSDD as *the* LCRs. We refer to positions in the reference genome that are not inside of any of the LCRs appearing in the HSDD as *unique*. Following the above discussion regarding the ambiguities and difficulties in defining relationships between groups or "families" of repeats, we take the HSDD as our fundamental set of relationships between LCRs. Identifying NAHR between shorter repeats is left for later work.

When attempting to construct a probabilistic model that evaluates sequencing data in repetitive regions, several issues arise:

1. If a read mapping algorithm encounters a read with several high quality candidate mapping locations, it will, by some method, choose one of the possible locations to place the read. If we are looking for rearrangements in repeats, then how do we find all reads homologous to a given repeat?

2. If we suspect a breakpoint has occurred at a specific location with an LCR, how do we know exactly which reads contain information relevant to this specific location?

[2]available at http://humanparalogy.gs.washington.edu/

3. If we are going to evaluate the read-depth of a particular region to determine if there has been an NAHR deletion or duplication, how should we choose which intervals to analyze and what sizes they should be?

The HSDD can be used to solve all of these problems, as detailed below, Overall, the entries of the HSDD tell us where to look to collect all of the reads mapped to different paralogs of a given LCR; performing pairwise alignments between the LCRs listed in the HSDD gives us exact coordinate mappings between homologous positions of a pair of LCRs; and equivalence classes of positions based on pairwise alignment mappings give rise to a natural, well-defined set of intervals for evaluating the data.

3.1.2 Graph construction

We begin with the human reference genome, over which the HSDD and all our work is defined:

`Definition.` The human reference genome $\mathcal{F}$ consists of both sex chromosomes and one copy of each autosome. Each sequence is assumed to be completely specified.[3]

We will use subscripts (e.g. $\mathcal{F}_i$) to refer to collections of intervals on $\mathcal{F}$, and function notation $\mathcal{F}(x)$ to refer to the nucleotide at position x in the reference sequence.

For every entry in the HSDD (i.e. a pair of homologous sequences in the human genome $\geq$ 1kb and $\geq$ 90% identity), we perform a standard global pairwise alignment[4]. The pairwise alignment implicitly defines a map between matched positions

[3] In this study we use GRCh37.

[4] we used Kalign[42] with parameters: `-f fasta -s 80.0 -e 3.0 -t 10.0 -m 0.0`.

of the two LCRs:

`Definition.` The pairwise alignment of two LCRs whose intervals are A and B induces a *homology map* m defined by: $m(x) = y$ for $x \in A, y \in B$ s.t. x and y are emitted together from the match state (even if the nucleotides mismatch) in the pairwise alignment of LCRs A and B.

A schematic example is shown in Figure 3.2. Due to indels, a homology map is injective but in general not surjective, and its domain is in general not an entire LCR, but rather some subsequence (i.e. only those positions not aligned in an indel). Thus, the HSDD implicitly defines a set of mappings $\{m_1, \dots\}$ on the reference genome $\mathcal{F}$.

Figure 3.2: A schematic example of the mapping m_1 induced from the pairwise alignment of two LCRs. Note that the map is injective, but due to indels it is not surjective nor is its domain the entire LCR denoted by x. Of course, by reversing the roles of the x and y LCRs, we would get the corresponding inverse map as well.

Using homology maps, we can identify corresponding "homologous positions" between LCRs.

`Definition.` For two positions $x, y \in \mathcal{F}$, then $x \overset{m}{=} y$ if $\exists i \ s.t. \ m_i(x) = y$.

We assume that $\overset{m}{=}$ is an equivalence relation. Specifically, we assume that the mappings m are transitive: if $m_i(x) = y$ and $m_j(y) = z$, then $\exists k : m_k(x) = z$. While this may not be true due to the criteria for the HSDD, assuming it is true is conservative: it will only serve to increase the computational complexity of our model (seen later).

We then use $\overset{m}{=}$ to create equivalence classes of homologous positions across LCRs:

`Definition.` For $x \in \mathcal{F}$, the equivalence class of positions homologous to x (i.e. m-equivalence class) is $M_x := \{x\} \cup \{y : x \overset{m}{=} y\}$.

A schematic example of the $\overset{m}{=}$ relation is given in Figure 3.3. For a helpful analogy, if we perform a multiple alignment between sequences, then all positions in the same column of the multiple alignment are m-equivalent.

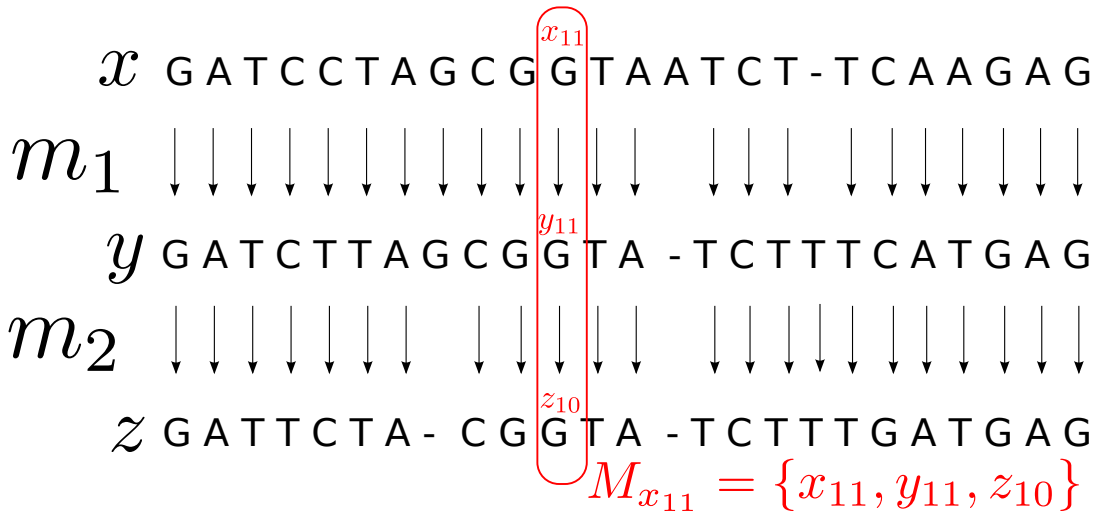


Figure 3.3: A schematic example of the construction of the equivalence class $M_{x_{11}}$. Thus, x_{11}, y_{11}, z_{10} are m-equivalent, i.e. $x_{11} \overset{m}{=} z_{10}$, etc..

The practical use of $\overset{m}{=}$ comes to light when analyzing a specific read inside of an LCR. Suppose read R was mapped to position x on reference $\mathcal{F}$ inside of an LCR. Then M_x tells us the *exact* locations (inside of LCRs) in $\mathcal{F}$ that R could have been generated from. Further, if we suspect a certain rearrangement and breakpoint among the LCRs, then we can determine exactly how this will affect the possible generating locations of R.

While $\overset{m}{=}$ is useful for analyzing individual reads, we are also interested in evaluating the number of reads over some given interval, i.e. read-depth. For a principled approach to read-depth inside of repeats, we introduce another relation on top of $\overset{m}{=}$.

Definition. Define $h_x := \{i : \exists y \in M_x \text{ s.t. } y \in \text{ran}(m_i) \text{ or } y \in \text{dom}(m_i)\}$.

In terms of the multiple alignment analogy, h labels each column of positions in the multiple alignment (an $\overset{m}{=}$ equivalence class) by the collection of map indices which relate the positions in that column. Figure 3.4 contains a schematic example of the h relation.

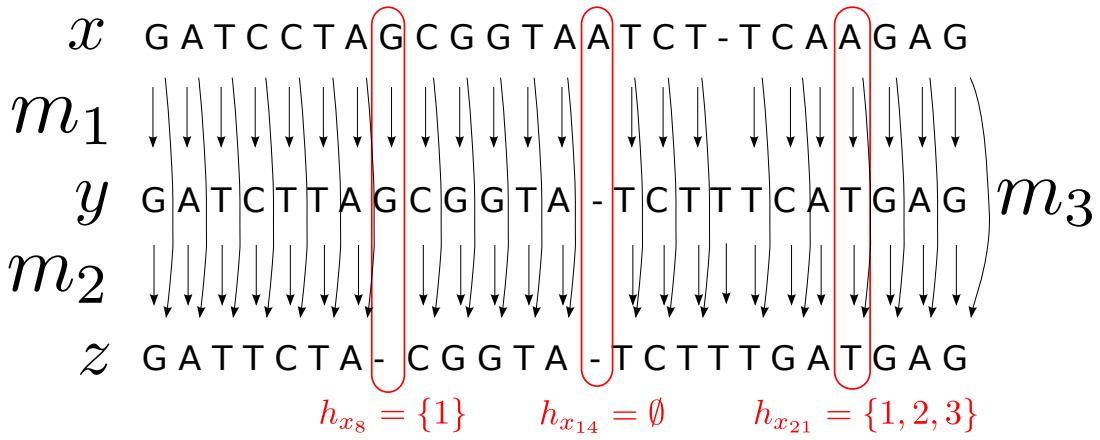


Figure 3.4: Schematic example of the h relation. Technically, there are more homology maps that correspond to $m_1^{-1}, m_2^{-1}, m_3^{-1}$, but they are not included here for simplicity. For x_8: note that $M_{x_8} = \{x_8, y_8\}$. Since x_8 and y_8 are related only by m_1, i.e. $m_1(x_{21}) = y_{21}$ as shown, then $h_{x_8} = \{1\}$. For x_{14}, note that it is not part of any homology map to any of the other sequences because it is an insertion; hence $M_{x_{14}} = \{x_{14}\}$ and so $h_{x_{14}} = \emptyset$. For x_{21}, note that $M_{x_{21}} = \{x_{21}, y_{21}, z_{20}\}$. Since $m_1(x_{21}) = y_{21}$ then $1 \in h_{x_{21}}$. Similarly, since $m_2(y_{21}) = z_{20}$, then $2 \in h_{x_{21}}$. Finally, $m_3(x_{21}) = z_{20}$, and so $3 \in h_{x_{21}}$.

We must introduce a technicality to proceed. Define the relation $< \cdot, \cdot >$ between m-equivalence classes as follows:

Definition. $< M_x, M_y >$ if $\exists k$ s.t. $\forall z \in M_x, \exists w \in M_y$ *s.t.* z, w are on the same chromosome, and $z + k = w$.

Conceptually, if, in a multiple alignment, the only state transitions between the column containing position x and the column containing position y are gap extension or match-to-match, then $< M_x, M_y >$.

Now we define a relation that extends $\stackrel{m}{=}$. The technical description is followed by a schematic figure and descriptive explanation.

`Definition.` Define the relationship $\stackrel{h}{=}$ between pairs of reference genome coordinates as follows:

1. If x, y are both unique, then $x \stackrel{h}{=} y$ if they are both on the same chromosome and both have the same bounding LCRs[5].

2. If x, y are both inside of (possibly different) LCRs and $< M_x, M_y >$ and $h_x = h_y \neq \emptyset$, then $x \stackrel{h}{=} y$.

3. If x, y are both inside of the same LCR and $h_x, h_y = \emptyset$ and $\forall z \in [x, y] : h_z = \emptyset$, then $x \stackrel{h}{=} y$.

Finally, we form equivalence classes using $\stackrel{h}{=}$:

`Definition.` $H_x := \{x\} \cup \{y : x \stackrel{h}{=} y\}$.

Figure 3.5 contains a schematic example of H-equivalence classes in the representation of a multiple alignment.

Appealing to the multiple alignment analogy, we elaborate the first two cases in the definition of $\stackrel{h}{=}$ in words. 2) consecutive positions (inside of LCRs) which have homologous positions in exactly the same set of LCRs are H-equivalent; 3) consecutive positions inside of the same insertion[6] inside of an LCR are H-equivalent. The H-equivalence classes essentially decompose the LCRs into elementary subunits

[5] i.e. if LCR A is the LCR with the greatest reference coordinates that are $< x$ and LCR B is the LCR with the smallest reference coordinates $> x$, then the same is true of LCRs A, B for y.

[6] Here, by "insertion" we mean a sequence which appears in one sequence but no others in the alignment.

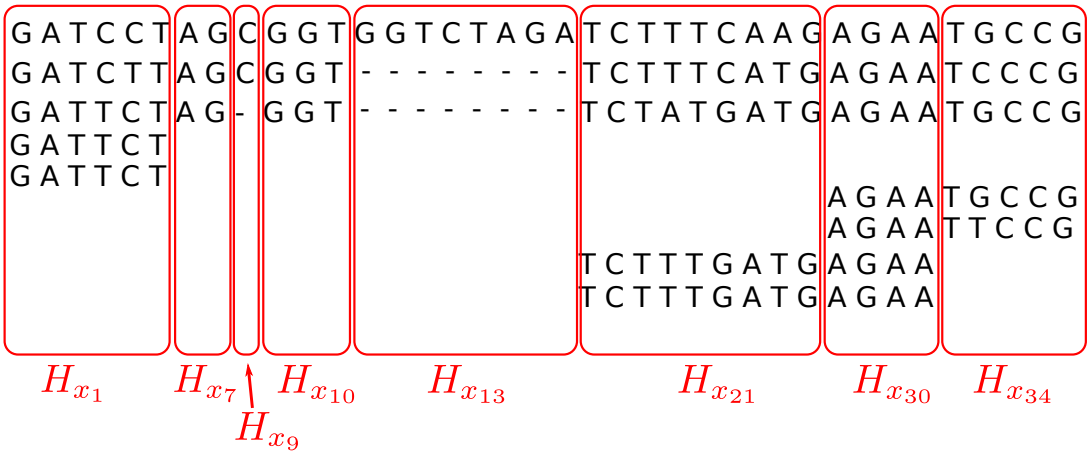


Figure 3.5: Schematic example of the H-equivalence classes. Note that the existence of $H_{x_{13}}$ follows from case 3) of Definition 3.1.2. The rest of the equivalence classes shown from case 2 of Definition 3.1.2. In particular, H_{x_7} and $H_{x_{10}}$ are distinct H-equivalence classes because of the $< \cdot, \cdot >$ relation.

of maximum length, similar in spirit to the concept of "ancestral duplicons", thought to be the foundational building blocks of modern segmental duplications [2]. From another perspective, $\overset{h}{=}$ is basically a generalization of the notion of a repeat family: if $A_1, A_2, \ldots$ were LCRs and $\forall x, y \in \cup_i A_i : M_x = M_y$, then LCRs $A_1, A_2, \ldots$ would be a "repeat family" in some sense. Of course, this is not the case in general, which is why we resort to this formalism.

We illustrate the application of $\overset{m}{=}$ and $\overset{h}{=}$ on a schematic example genome in Figures 3.6 and 3.7. First, we associate positions together according to the equivalence relation $\overset{m}{=}$. Visually, this amounts to "folding" the reference genome according to homology between LCRs. Once the genome is folded, groups of intervals are identified together via $\overset{h}{=}$, thus partitioning the genome $\mathcal{F}$ into $\mathcal{F}_1, \ldots$. This process is reminiscent of the de Bruijn graph and A-Bruijn graph constructions based on k-mer repeats, as described in [20, 63, 62].

The $\overset{m}{=}$ relation tells us where to look for reads homologous to certain regions of the genome; this is crucial for finding reads which may contain evidence of an

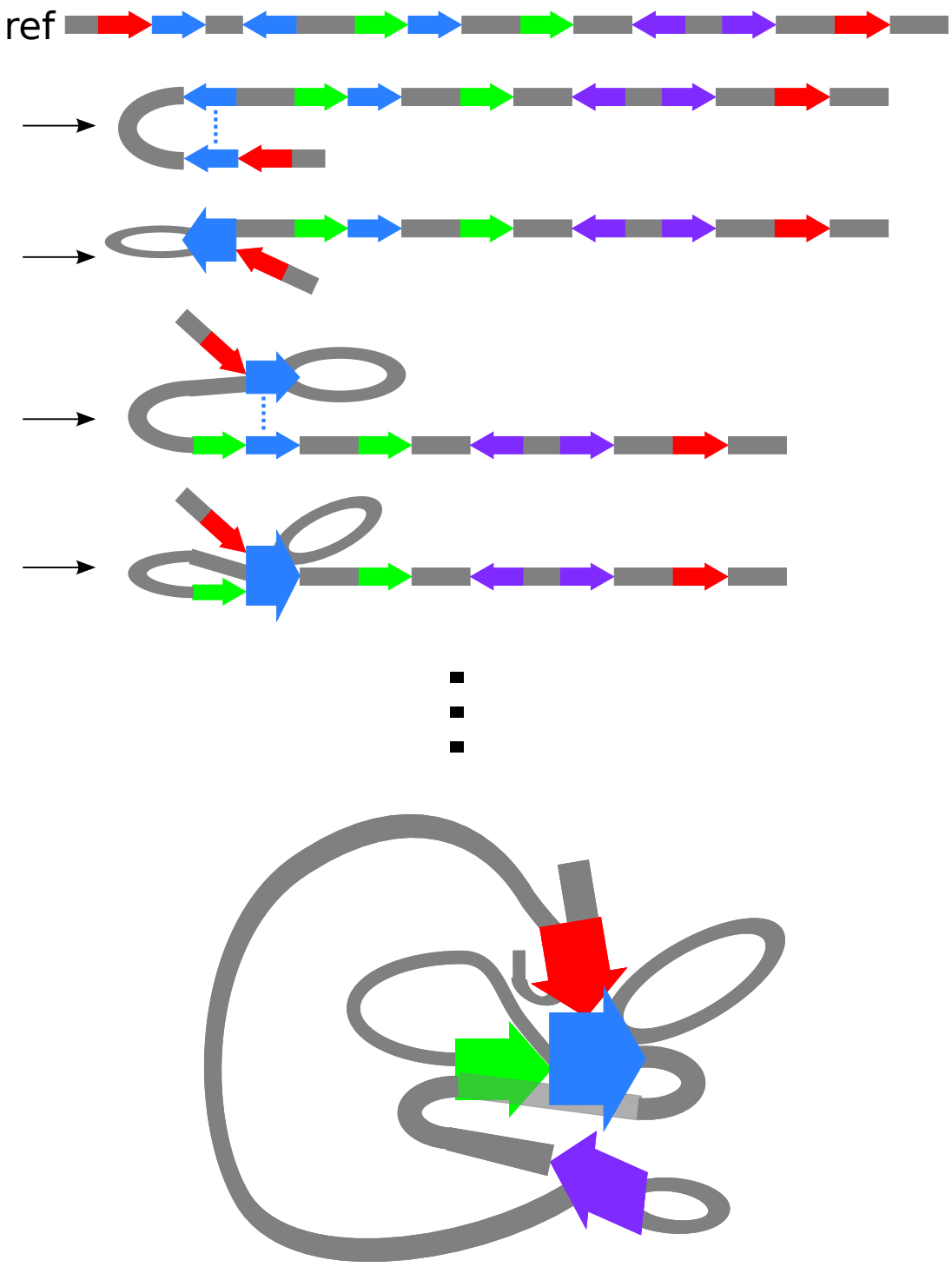


Figure 3.6: A schematic example of applying the $\overset{m}{=}$ equivalence relation to "fold" a genome according to homology and identify collections of homologous intervals. Homologous repeats are the same color. Grey regions are unique sequence.

NAHR event's breakpoint, but have been mapped to other homologous locations of the reference genome by an aligner. In light of the $\overset{m}{=}$ relation, it is helpful to modify

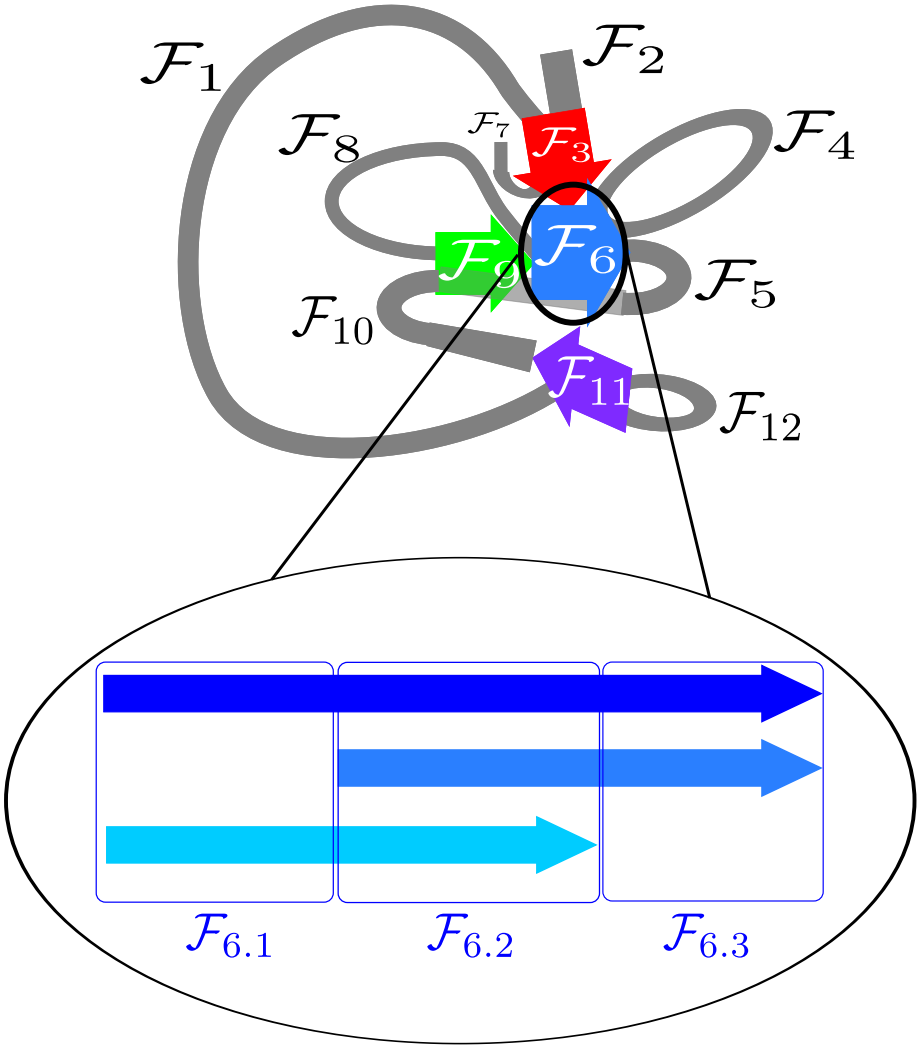


Figure 3.7: After "folding" a genome using the equivalence relation $\overset{m}{=}$, we apply $\overset{h}{=}$. The H-equivalence classes induce a partition of the genome into $\mathcal{F}_1, \ldots$. For a more detailed understanding, we magnify the "blue" repeat part of the folded genome. Upon closer inspection, the three blue repeats, represented by different shades of blue, are seen to have complicated homology relationships with respect to each other. As such, the blue region is itself split into several equivalence classes, according to $\overset{h}{=}$.

our thinking about aligners and instead consider them to be "homology-finders" rather than "read mappers". That is, read aligners are excellent at finding *some* location in the reference genome which has high homology to a given read, but are not so rigorous when *choosing* a single location among many to map a read to. With this understanding, we can use $\overset{m}{=}$ to evaluate reads in repetitive regions after a read aligner has "found" a location which has high homology with a read's sequence.

The importance of $\stackrel{h}{=}$ will become clear once we formalize the notion of an NAHR event below.

3.1.3 Variational positions

We come back to emphasize an important point: LCRs are in general *not* perfect copies of one another. For almost every pair of highly homologous LCRs we consider, there is a set of positions at which they disagree, either via mismatch or indel.

`Definition.` The *variational positions* (VPs) between LCRs A and B are the positions $v_1, \ldots$ of the pairwise alignment of A and B wherein there is a mismatch or indel.

Of course, since we are dealing with highly homologous sequences, there are relatively few VPs with respect to the length of either LCR. An example of variational positions taken from real LCRs in the human reference genome is shown in Figure 3.8.

A pair of homologous LCRs can be distinguished by the values taken at each of the variational positions $v_1, \ldots$. We will refer to the values taken at the VPs by an LCR as that LCR's "variational position pattern".

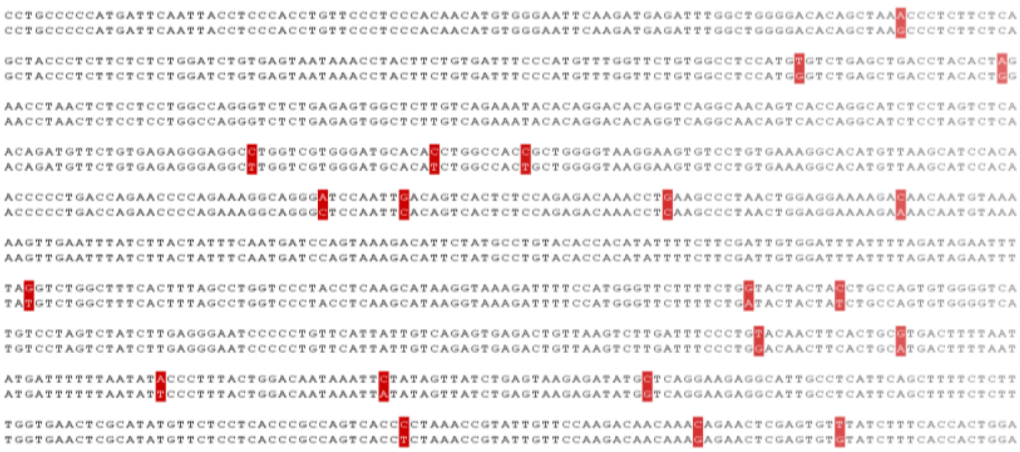


Figure 3.8: An example of variational positions from part of the human reference genome. Part of the pairwise alignment of two LCRS (chr 1 : 149015492−149030910 and chr 9 : 70427260−70411827) is shown. Variational positions are highlighted in red. Here, all variational positions are mismatches, but variational positions can be indels as well.

3.2 Events and breakpoints

3.2.1 NAHR and gene conversion events

According to the biology of NAHR, any pair of homologous repeats may mediate an NAHR event. For this study, we consider the pairs of LCRs listed in the HSDD to be all possible pairs of repeats that may mediate an NAHR event. Thus, each pair of LCRs listed in the HSDD defines a potential NAHR event. Although we are primarily interested in NAHR events, we recognize that the gene conversion mechanism is almost exactly the same as the NAHR mechanism and has very similar resulting breakpoints, and so each pair of LCRs also represents a possible gene conversion event. We capture these two mechanisms in the following random variable:

`Definition.` A potential NAHR or gene conversion *event* E is a random variable defined by two distinct, fixed, LCRs in $\mathcal{F}$ of length $\geq$ 1kb and $\geq$ 90% identity (i.e. paired in the HSDD), whose sample space is categorical and is determined according to the orientation and complementarity of the LCRs as discussed in 2.

We describe the sample space for a hypothetical event E in detail, whose LCRs we denote as A and B. Of course, a pair of LCRs may not experience an NAHR or gene conversion event, and hence *no event* is always a possible outcome. Also, any pair of LCRs may experience gene conversion (non-crossover homologous chromosome repair), in which there are two cases: A is donor or B is donor. Thus, *gene conversion - A donor* and *gene conversion - B donor* are also always in the sample space. As for NAHR: if the two LCRs lie on the same chromosome and are reverse complements, then the sample space contains *inversion*; if the LCRs lie on the same chromosome and have the same orientation and are not complements of each other, then both *deletion* and *duplication* are in the sample space. Again, since we are primarily concerned with NAHR (rather than gene conversion) in this study, we will refer to E simply as "event E" or "NAHR event E" instead of the proper "potential NAHR or gene conversion event E".

In this study, we ignore potential translocations (LCRs on different chromosomes or different arms of the same chromosome), and restrict the space of potential events to those on the same arm of the same chromosome wherein the length of sequence inbetween and including both LCRs is ≤ 250 kb. We employ these constraints because we will apply this model to data obtained from putatively healthy individuals, in whom very large (> 250 kb) rearrangements are thought to be very unlikely (since they are healthy), as are translocations.

3.2.2 Dependencies between events

Immediately we must ask: what is the relationship between different potential NAHR events? How do we determine the dependency between different potential NAHR events E? The answer to these questions highlights the usefulness of our homology-based graph construction in 3.1.2. Intuitively, two events E_1, E_2 are related if the intervals of the genome they potentially affect overlap, or if they affect homologous sequences so that reads from these homologous sequences *a priori* may have come from either sequence. We declare the region (potentially) affected by event E to be the contiguous interval of the reference genome inbetween and including the pair of homologous LCRs that mediate E.

Consider the partitioned reference genome obtained after applying relations $\overset{m}{=}$ and $\overset{h}{=}$. Define $A_i := \{j : \exists x \in \mathcal{F}_j, x \text{ is affected by } E_i\}$, i.e. A_i denotes the partitions of the reference genome potentially affected by an NAHR or gene conversion event at E_i. We declare two events E_i, E_j to be dependent if $A_i \cap A_j \neq \emptyset$. Figure 3.9 gives a schematic example of this process.

Looking closer, there are in fact two different kinds of dependency, both demonstrated in Figure 3.9. First, it may be the case that the intervals potentially affected by E_i and E_j actually overlap, as do E_2, E_5 in Figure 3.9. This is of great importance, as one event's outcome may constrain another event's space of possible outcomes. For example, if the intervals of E_i, E_j overlap, and E_i occurs as an NAHR deletion, then E_j *cannot* occur (as any non-null outcome) as one of the would-be mediating LCRs of an E_j event has been deleted! We refer to this type of dependency as an *exclusivity constraint*.

`Definition.` Suppose the regions potentially affected by events E_i and E_j overlap.

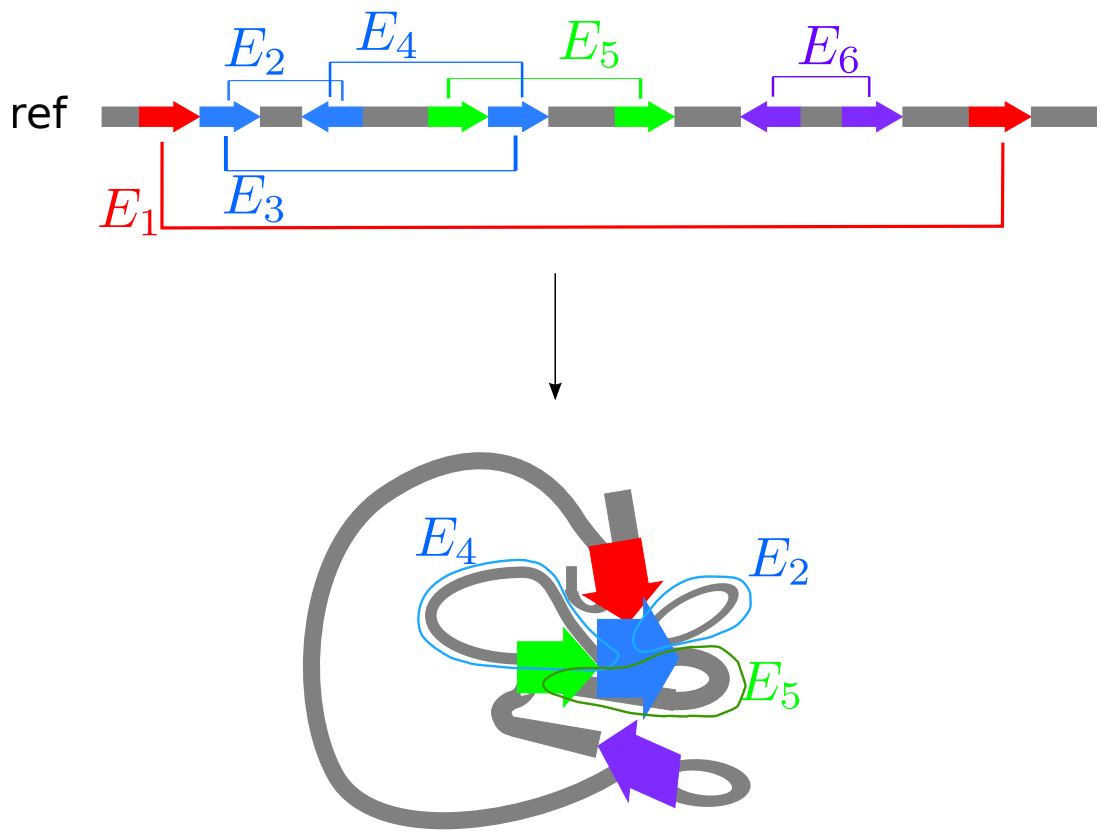


Figure 3.9: A schematic example of the space of potential NAHR events $E_1, \ldots$ and their mapping onto the folded reference genome. For clarity, only the mapping of E_2, E_4, E_5 are shown on the folded reference genome. Note that A_2, A_4, A_5 are pairwise non-disjoint, and therefore E_2, E_4, E_5 are not pairwise independent.

Then E_i and E_j are said to be related by an *exclusivity constraint*, written $E_i \perp E_j$. If one of of the events occurs as an NAHR event (deletion, duplication, or inversion), then the other event *must* occur as *no event*. The two events may, however, both occur as gene conversion events. If the joint outcomes (e_i, e_j) for (E_i, E_j) are forbidden as such, then we will write $e_i \perp e_j$.

Note that not all outcomes of the hypothetical event E_i would actually constrain E_j. Indeed, E_i may delete only (a small) part of E_j's mediating LCRs, or E_i may be a duplication which would actually create more LCRs like those of E_j! We ignore these possibilities for this study, and declare that if E_i results in an NAHR event, then E_j must take value *no event*. This is not the case for gene conversion events,

however: E_i and E_j could each occur as a gene conversion.

The other type of dependency is when the H-equivalence classes affected by E_i, E_j intersect, i.e. $A_i \cap A_j \neq \emptyset$, but the regions affected by E_i, E_j do not overlap. This is the case when there is some sequence potentially affected by E_i which is highly homologous ($\geq$ 1 kb, $\geq$ 90% identity) to another sequence potentially affected by E_j. That is, E_i, E_j affect different but homologous LCRs. We refer to this type of dependency as a *homology dependency*.

`Definition.` Suppose the regions potentially affected by events E_i and E_j do not overlap, but that $A_i \cap A_j \neq \emptyset$.. Then E_i and E_j are said to be related by a *homology dependency*.

Figure 3.10 gives an example of the two types of dependencies discussed here. Applying these relationships to the schematic reference genome in Figure 3.9, we arrive at the graphical model shown in Figure 3.11.

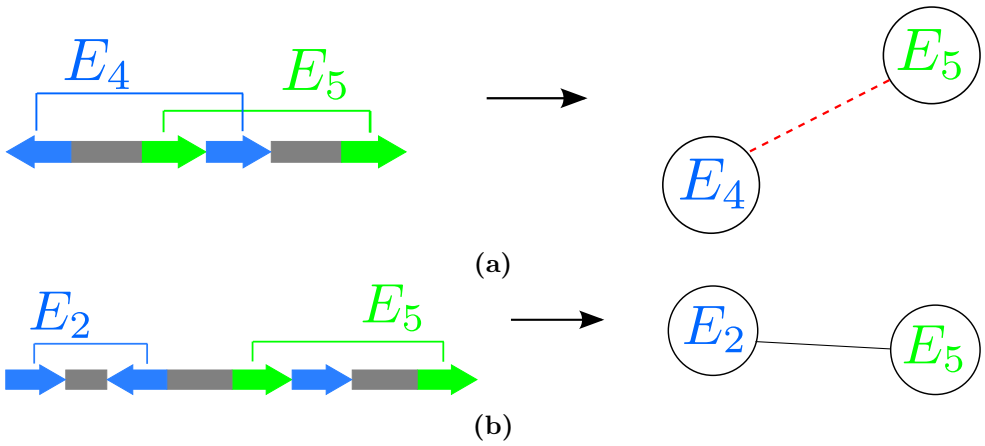


Figure 3.10: Schematic examples of an exclusivity constraint and a homology dependency between potential NAHR events. **(a)** The regions of the genome potentially affected by E_4 and E_5 overlap, and therefore E_4, E_5 are related by an exclusivity constraint, meaning that at most one of E_4, E_5 can result in an NAHR event. Exclusivity constraints are denoted by dotted red lines. **(b)** E_2 and E_4 are related by a homology dependency because there is homology between the regions of the genome affected by each event. Here, E_5 potentially affects a "blue" repeat, which is homologous to the "blue" repeats that mediate E_2.

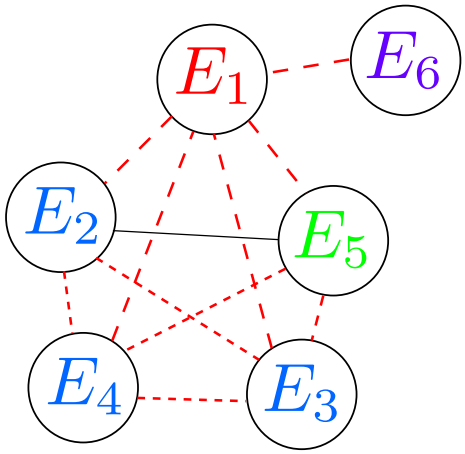


Figure 3.11: The graphical model on schematic potential NAHR events $E_1, \ldots, E_6$ induced by exclusivity constraints and homology dependencies. Exclusivity constraints are represented by dotted red lines, homology dependencies are solid black lines.

3.2.3 Graphical model

The actual dependencies between all potential NAHR and gene conversion events $E_1, \ldots, E_n$ on the reference genome $\mathcal{F}$ is shown in Figure 3.12.

Note that the partition of $\mathcal{F}$ induces a partition of E through the dependencies discussed above. Indeed, each connected component is an element of the partition (a collection of events).

Figure 3.11 is essentially a graphical model for NAHR. Although the graph as shown is missing several random variables which we introduce later (the data and breakpoint random variables), it does show the basic relationships between $E_1, \ldots, E_n$ and satisfies all of the usual rules for a graphical model with respect to summation and conditioning.

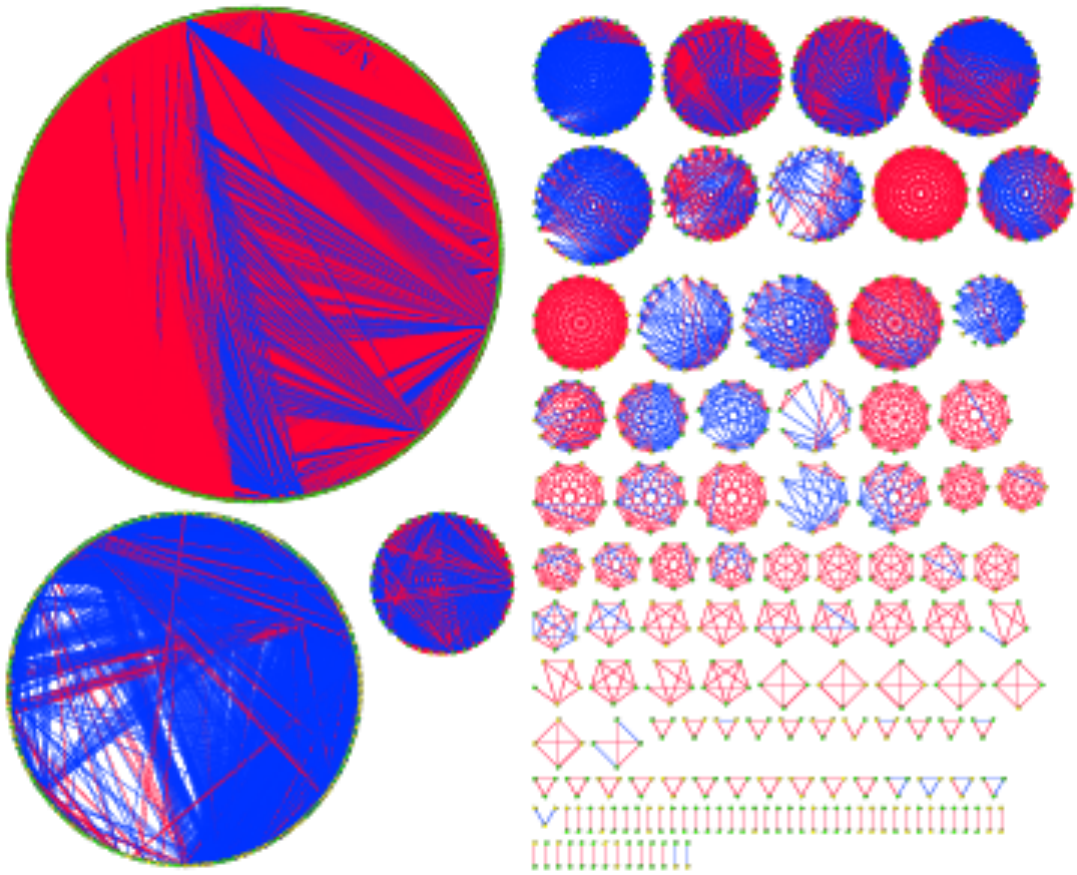


Figure 3.12: The graphical model on potential NAHR events $E_1, \ldots, E_n$ from the entire human genome induced by exclusivity constraints and homology dependencies. Each node is a potential NAHR event E_i, and each edge represents one of the dependency relations between events discussed above. Notice that the various potential NAHR events break into distinct connected components according to the dependencies. Nodes are arranged in a circular fashion, on the perimeter. Here $n = 1768$. Green nodes are potential NAHR deletion/duplications, while yellow nodes are potential NAHR inversions. Exclusivity constraints are shown in red, while homology dependencies are shown in blue. Not shown are 261 connected components of size 1, i.e. events with no dependencies on other events.

3.2.4 Breakpoints

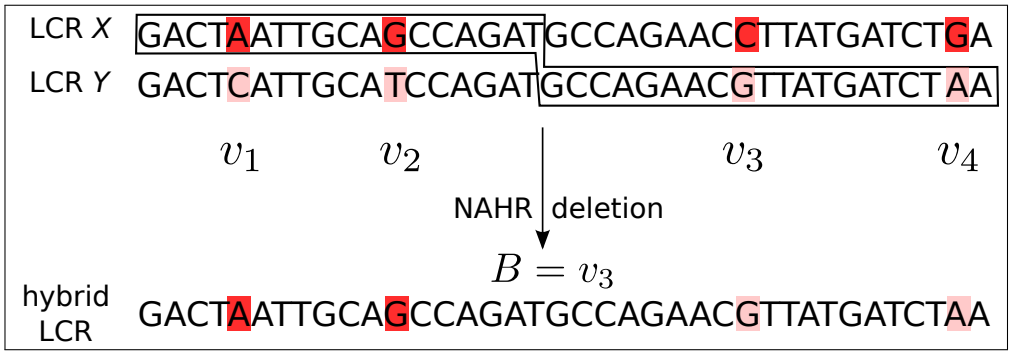
When an NAHR event occurs, it creates a new hybrid LCR(s) that have some breakpoint(s). We assume that the breakpoint occurs somewhere inside of the two mediating LCRs, excluding the case in which branch migration of the D-loop of the double Holliday Junction pushes the breakpoint outside of the pair of annotated LCRs. According to the "rules of NAHR" [54, 44, 60], the Rad51-mediated single-stranded DNA homology search mechanism chooses a long stretch of near-perfect homology

to initiate the PCR that repairs the double-stranded break in one of the chromatids [15, 67].

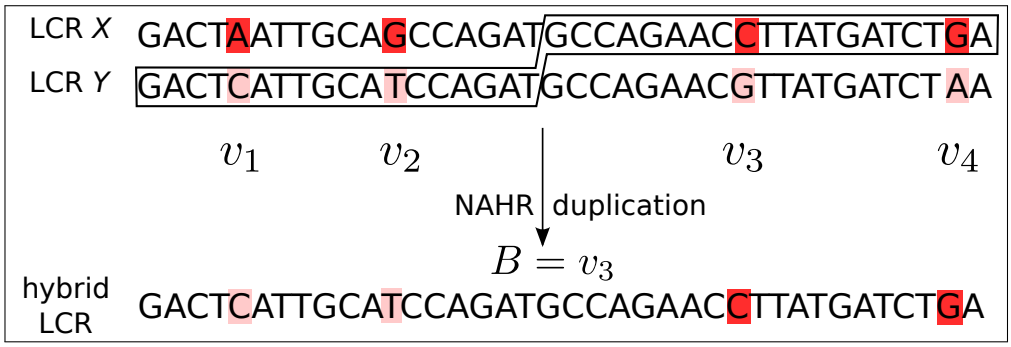
Crucially, this means that the breakpoints of the NAHR event are actually at homologous positions of the two LCRs. In other words, the breakpoint will occur at a pair of positions on the LCRs which are *aligned together in the pairwise alignment of the two LCRs*. Thus, we imagine that we could actually pinpoint the breakpoint of an NAHR event as a single position along the pairwise alignment of the two mediating LCRs (which represents a pair of points - one on each mediating LCR, at corresponding homologous positions).

However, in general we *cannot* resolve the breakpoint of an NAHR event down to a single position because we are dealing with highly homologous repeats. Indeed, the homology length and quality stringency imposed during the Rad51 homology search essentially guarantees that the actual breakpoint (i.e. point of initiation of PCR) of an NAHR event will be at some point within a stretch of near-perfect homology between the two mediating repeats. Thus, we can only hope to localize an NAHR breakpoint to the region between two variational positions. Thus, technically, we aim to identify a "breakpoint region" between two VPs, but we will continue to use just the term "breakpoint" for ease.

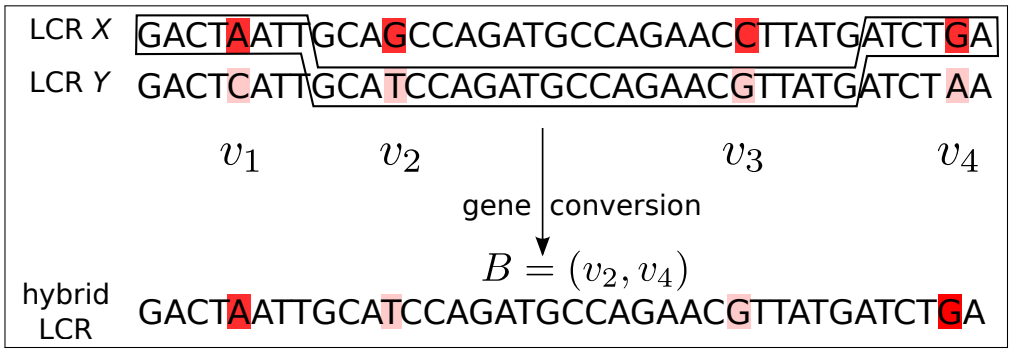
Finding the breakpoint of an NAHR event is therefore equivalent to finding the point in the newly-created hybrid LCR at which the variational position pattern "switches" from one LCR to another. This means breakpoints actually occurring in the interim region between two VPs will result in *exactly* the same hybrid LCRs, thus reducing the space of all possible breakpoints to the space of variational positions. We can therefore restrict the space of potential breakpoints to be the set of variational positions that distinguish a given pair of repeats.



(a)



(b)



(c)

Figure 3.13: A schematic example of the breakpoint random variable B for several homologous recombination events. In each case, the sequences of two distinct LCRs are shown aligned, labeled X and Y. We assume that LCR X has smaller genomic indices than LCR Y. Variational positions (VPs) are shown in shades of red - LCR X has dark red VPs, LCR Y has light red VPS. The VPs are labeled $v_1, \ldots, v_4$. **(a)** We suppose that LCRs X and Y mediate an NAHR deletion with breakpoint $B = v_3$, resulting in the hybrid LCR shown. The hybrid LCR contains both shades of variational positions - at v_1, v_2, the hybrid LCR matches the VPs of LCR X, and at v_3, v_4, the hybrid LCR matches the VPs of LCR Y. Note that the exact position of crossover was at a position inbetween v_2 and v_3, but since the LCRs are identical in $[v_2 + 1, v_3 - 1]$, then the breakpoint is essentially v_3; hence $B = v_3$. **(b)** Similar to part (a), but for an NAHR duplication. **(c)** We suppose a gene conversion event occurs, where LCR Y is the donor.

`Definition.` Consider an NAHR event E mediated by LCRs X, Y. The *breakpoint* B of NAHR event E is defined as follows. If E = *no event*, then $B = \emptyset$. If E results in an NAHR event, then B is the smallest variational position of the pairwise alignment profile between LCRs X and Y after the switch in variational position pattern between A and B in the newly created hybrid LCR. If E results in a gene conversion event, then B is an ordered pair of variational positions, each one being the smallest VP after a switch in VP pattern between the two mediating LCRs.

In particular, suppose event E has LCRs X, Y and LCR X has smaller reference genome coordinates than Y. A schematic example of the following cases are shown in Figure 3.13. If E results in an NAHR deletion, then B is the first variational position in the alignment profile which shows the Y VP pattern in the resulting hybrid LCR. If E results in an NAHR duplication, then B is the first variational position which shows the X VP pattern in the resulting hybrid LCR. If E is an inversion, then B is the first VP showing Y's VP pattern in the hybrid LCR which replaces LCR X.[7] If E is a gene conversion wherein X is the donor, then $B = (B_1, B_2)$ where B_1 is the smallest VP on the YXY hybrid that shows the X VP pattern, and B_2 is the smallest VP on the YXY hybrid that shows the Y VP pattern and is $> B_1$. The analogous holds for gene conversion events wherein Y is the donor.

In light of the complexity and uncertainty of branch migration and mismatch repair processes, our above definition of a breakpoint is inadequate. As observed in experimental validation studies, there can be complicated switching between the variational patterns of both LCRs [38]. A more sophisticated breakpoint model would allow for oscillations between informative patterns, and the breakpoint would be the entire region in which the complicated variational pattern switching occurs.

[7] Recall that NAHR inversions convert LCR X into an XY hybrid and LCR Y into a YX hybrid with respect to the orientation assumed during the pairwise alignment of X and Y.

Here we make a simplifying assumption:

`Assumption 1.` A recombinant LCR displays exactly one transition between the informative patterns of its mediating LCRs.

We could add the breakpoint random variables B_i for each event E_i to the graphical model in Figure 3.11 by creating nodes B_i and drawing dependencies for each B_i to its corresponding E_i and all other events connected to E_i (including their breakpoints), but this would clutter the visualization and provide no useful new information, and so we omit this addition to the graphical model.

3.2.5 Test genomes

Ultimately, we are interested in identifying NAHR events that occurred in some individual's genome G with respect to the reference genome $\mathcal{F}$. We refer to G as a test genome. Since we are focusing on NAHR, we assume,

`Assumption 2.` A test genome G differs from $\mathcal{F}$ *only* by NAHR and gene conversion events occurring at $E_1, \ldots$ with some breakpoints $B_1, \ldots$. That is, we exclude the possibility of SNPs and indels, structural variation by other mechanisms, and NAHR mediated by sequences other than those from the HSDD represented by $E_1, \ldots$.

Since $(E, B)_1^n$ completely specifies G, then we will use G and $(E, B)_1^n$ interchangeably.

Of course Assumption 2 is false, but it allows us to analyze data in which the only possible explanation for observed deviations from that expected in $\mathcal{F}$ is that some NAHR or gene conversion events occurred at certain E_i with breakpoints B_i. This simplification is reasonable since NAHR is one of the more common rearrangement

mechanisms and is restricted to repetitive regions, as mentioned in chapter 1. If there are instances in which our assumption is wrong, i.e. there is a non-NAHR rearrangement which deletes or duplicates a non-trivial amount of sequence at the locus specified by one of the potential NAHR events E_i, then our model is likely to nonetheless detect that there is a rearrangement (due to the strength of the read-depth signal alone), but will of course falsely attribute it to NAHR.

We now describe how to apply an NAHR event to $\mathcal{F}$ to obtain G.

Consider the partition $\mathcal{F}_1, \ldots$ of $\mathcal{F}$ obtained from the procedure in 3.1.2, and an NAHR event and breakpoint (E, B). We obtain a partition $G_1, \ldots$ of G from the partition of $\mathcal{F}$ simply by modifying the intervals in the appropriate partitions of $\mathcal{F}$ according to the changes implicated by (E, B). Thus, the partition $G_1, \ldots$ of G correspond to the partition $\mathcal{F}_1, \ldots$ of $\mathcal{F}$, wherein the intervals contained in G_i are either deleted, duplicated, replaced by (part of) a new hybrid LCR, or have an additional (part of a) hybrid LCR - according to $E(, B)$. The M- and H- equivalence classes are transformed in similar fashion.

Of major importance, note that we can now construct *any* part of a hypothetical test genome G merely by applying the specified rearrangements. This is a major innovation in structural variation: rather than dealing with reads mapped as given, which contain impossible anomalies (e.g. mates of a paired-end read mapped to different chromosomes), we can evaluate realistic (i.e., concordant) alignments of the reads to a hypothetical genome. Specifically, we align reads against hybrid LCRs in the putatively rearranged genome.

3.3 High-throughput sequencing data

Now that we have a framework for working with repeats, we show how the data benefits from this framework.

3.3.1 Paired-end reads

The basic atoms of data available from high-throughput sequencing are reads. We will exclusively consider paired-end reads, but with minor adjustments all of the following can be applied to other types of reads.

During next-generation sequencing, the genome is cleaved into short *fragments* whose mean length is typically in the vicinity of $100 - 300$ bp. In paired-end sequencing, both ends of the fragment are sequenced, producing a pair of mate reads (where the pairs are known) of the same length, typically $36 - 100$ bp.

`Definition.` A *paired-end read* P is an ordered triple $P = (R^a, R^b, L)$. The random vectors R^a, R^b are the nucleotide sequences pertaining to each read in the pair, and the random variable L is the smallest genomic index on test genome G which generated the fragment from which R^a, R^b were sequenced (i.e. left-endpoint of the fragment). We will refer to a paired-end read as simply a "read" for brevity when it is clear that we are not discussing the individual mated reads R^a, R^b. The "fragment implied by P" is the hull of the genomic coordinates to which R^a, R^b are aligned.

See Figure 3.14 for a schematic example of a paired-end read.

We emphasize that we are defining paired-end reads in relation to the *test* genome G, not the reference $\mathcal{F}$. As such, we avoid entirely the notions of split-reads, dis-

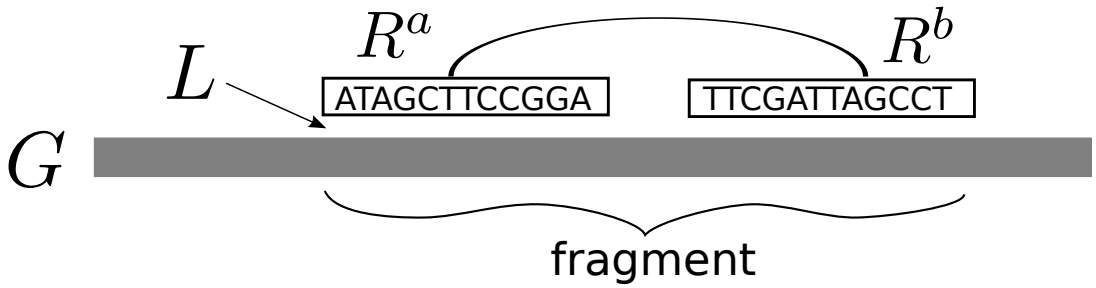


Figure 3.14: A schematic paired-end read.

cordant reads, or any other oddities that arise when comparing reads against the reference genome $\mathcal{F}$. Indeed, "discordant reads" and the like do not actually exist; rather, those notions are artifacts of aligning paired-end reads to genomes from which they were not generated. Instead, we know that every single paired-end read in a dataset was generated *concordantly* from G, and so it only remains to construct the genome G and align reads against it. But we know how to construct G, or any small part of G, *exactly* given a set of events $(E, B)_1^n$, as outline in 3.2.5. Thus, we can always construct any region of interest in G and *concordantly* align reads to it.

Note that we defined the generating location L of a paired-end read P to be *only* the left-endpoint and *not* the entire fragment implied by P. This will be of great convenience and will be crucial for helpful independence assumptions, as we will see later. Note also that we distinguish between a *generating* location and a *mapping* location. That is, the generating location is the location from which the paired-end read was actually produced, while the mapping location is where the read was placed by the read aligner and may not actually be the location that produced the read.

Although a paired-end read P is completely specified by its nucleotide sequences R^a, R^b and the location L of the test genome G from which it was generated, the location L is always unknown - sequencing machines only output (paired) nucleotide sequences, without any other indication of the generating location. Thus, our data

is partially hidden; we always deal with $(R^a, R^b, \cdot)$.

`Definition.` The data D pertaining to test genome G consists of some number C of paired-end reads P, whose read sequences are completely observed but whose generating locations are hidden: $D := \left(R^a, R^b, \cdot\right)_1^C$.

3.3.2 Partition of the data D

Although the generating location L of each paired-end read P is hidden, we can still partition the data D just like we partitioned G and $\mathcal{F}$. Read-error rates vary by machine and from run to run, but in general read-error rates generated by Illumina sequencing machines (which we consider exclusively) are quite low, ranging from 1% to 0.001%. Since *each* read is $\approx 36 - 100$ bp in length, then it is extremely unlikely that we will find reads with so many errors that they are significantly non-homologous to their generating location L. Thus, while technically a paired-end read P with hidden L could have come from anywhere in the genome G, in fact there are very few locations which could have generated P with non-negligible probability. For reads from unique regions, we assume there is only one possible location, i.e. the unique region where we found the read mapped to (ignoring issues of mappability[8]). For reads from repeat regions, there are several locations (paralogs of the repeat of the generating location) which realistically may have generated P.

`Assumption` **3.** Consider a paired-end read P which has been mapped to location ℓ in $\mathcal{F}$ by a read aligner (e.g. BWA). We consider three cases:

1. Suppose ℓ is unique and the fragment implied by p does not intersect any LCR.

[8]Mappability refers to the uniqueness of a certain k-mer in $\mathcal{F}$. For example, a 36-bp k-mer may be repeated many times throughout $\mathcal{F}$, though we don't consider it to be a repeat.

- If ℓ exists in G and was not duplicated, then we assume that ℓ generated P, i.e. $L = \ell$.
- If ℓ exists in G and was duplicated, then we assume that ℓ or its newly-created duplicate(s) are possible generating locations for P.

2. Suppose that the fragment implied by P has been mapped to a repetitive location ℓ in $\mathcal{F}$. Then we assume that all homologous positions on LCRs homologous to ℓ (including ℓ itself and any newly-created duplicates or hybrids) that exist in G are possible generating locations for P, i.e. $L \in M_\ell$.

3. For "edge" cases, where the fragment implied by P intersects an LCR but ℓ lies outside of the LCR, we find all positions on LCRs that are homologous to the positions in the fragment which do intersect the LCR, and then infer the positions that would correspond to ℓ.[9]

In summary, the space of possible generating locations for reads mapped to LCRs are homologous positions of homologous LCRs, and the space of possible generating locations for reads mapped to unique regions is simply the mapping location. In other words, the space of possible generating location is the M-equivalence class containing the mapping location, adjusted according to the NAHR events $(E, B)_1, \ldots$. For clarity of exposition we will ignore edge cases in further derivations below; the edge cases are always similar to the LCR cases as in 3.

We may associate each read P with a particular M-equivalence class of homologous positions, and thereby with a particular particular member $\mathcal{F}_i$ of the partition of $\mathcal{F}$. We then associate P with the corresponding member G_i of the partition G.

[9]We identify a slightly broader region about the corresponding end of homologous LCRs and perform a Viterbi alignment of P to the small, selected region, and take the leftmost endpoint of the Viterbi alignment as a generating location. We use this approximation, again, because read-errors are so low that while the exact alignment of a read may have uncertainty, the endpoints of the alignment should have little uncertainty.

This induces a partition $D_1, \ldots$ of the data D corresponding to the partitions of G and $\mathcal{F}$. We can add D to the graphical model in 3.11 by creating nodes $D_1, \ldots$ for the partition of D, and for each D_i drawing dependencies between D_i and all events which potentially affect $\mathcal{F}_i$. Nonetheless, we will continue to think of the graphical model shown in Figure 3.11 because the D_i nodes do not provide any new useful information (they are merely a technicality) and merely clutter the visualization.

3.4 Probabilistic model

Now that we have defined all of our random variables, we can form a probabilistic model for them. The probabilistic model breaks into two parts: the prior, consisting of the biological r.v.s E and B; and the likelihood, consisting of the data D given the priors (E, B).

3.4.1 The prior

Prior on events

For simplicity, we would immediately wish to assume *a priori* independence of the events E_1^n, and that each event E_i occurs as *no-event* with the same probability $1 - \gamma$. In this spirit, we define f on events E

`Definition.` The naïve p.m.f. of an event outcome $E = e$ is f, where:

$$f(E = e) = \begin{cases} 1 - \gamma & : e = \textit{ no event} \\ \gamma \cdot \alpha \cdot \frac{1}{2} & : e = \textit{gene conversion} \\ \gamma \cdot (1 - \alpha) \cdot \frac{1}{2} & : e = \textit{ deletion, duplication} \\ \gamma \cdot (1 - \alpha) & : e = \textit{ inversion} \end{cases}$$

where $\alpha \in (0, 1)$ reflects the preference for the non-crossover outcome (gene conversion) as opposed to the crossover outcome (NAHR) of the double Holliday junction repair mechanism. (Also, recall that if the mediating LCRs of an event have the same orientation, then the only possible NAHR outcomes for the event are deletions and duplications; whereas if the LCRs have opposite orientation, then the only possible NAHR outcome for the event is an inversion).

But in light of the discussion in 3.2.2, there are indeed dependencies between NAHR events. The homology dependencies do not imply dependency between E_1^n *in the prior*, since events related via a homology dependency do not physically impact each other. The exclusivity constraints, however, have important implications.

`Definition.` Λ is the set of pairs of NAHR event outcomes that are impossible according to the exclusivity constraints among the events E_1^n. That is,

$$\Lambda := \left\{(e_i, e_j) : E_i \perp E_j \text{ and } e_i \perp e_j\right\} \tag{3.4.1.1}$$

We modify the naïve independence assumption to account for the exclusivity constraints, arriving at the prior distribution on events:

$$\mathbb{P}\Big(E_1^n = e_1^n\Big) := \prod_{i=1}^{n} f(E_i = e_i) \cdot \prod_{\lambda \in \Lambda} (1 - \mathbf{1}_\lambda). \tag{3.4.1.2}$$

In our calculations (presented later), we set $\alpha = 1 - 10^{-3}$ and $\gamma = 1 - 10^{-6}$ for f.

Note that a more informed prior could have been constructed according to reported correlations between rates of NAHR and LCR features as studied in [52, 12, 73, 60, 51], but the likelihood will be dominant in our problem (due to large amounts of data and low read-error rates), so for simplicity we use the model presented here.

Also, note that the exclusivity constraints cause the prior on events to add up to some $\delta < 1$. As usual, since all of this goes into Bayes' Rules, the factor δ cancels out, and so we ignore it.

Prior on breakpoints

`Assumption` **4.** The breakpoints of events are *a priori* mutually conditionally independent given all the event outcomes. Also, the breakpoint B_i of event E_i is conditionally independent of all other events given E_i.

Intuitively, Assumption 4 states: when an event occurs, its choice of breakpoint does not depend on other events or other breakpoints.

`Assumption` **5.** For a given outcome of a given event, we assume that each possible breakpoint is *a priori* equally likely.

A more informed prior on breakpoints is possible, such as by using the literature on noted double-strand break sequence features [29, 27, 13] or by using the distance

between neighboring variational positions to determine probabilities. But, again, the likelihood is expected to be dominant due to the large amount of data and low error rates, so we use a simple model for now.

Let $\mathscr{B}_i$ be the set of all variational positions for event E_i. Then $\mathscr{B}_i$ is the space of possible NAHR breakpoints for event E_i, and $\mathscr{B}_i^2 := \big\{(b,c) : b,c \in \mathscr{B}_i, b < c\big\}$ is the space of possible gene conversion breakpoints, following section 3.2.4. Then we define

$$\mathbb{P}\Big(B_1^n = b_1^n \;\Big|\; E_1^n = e_1^n\Big) = \prod_{i=1}^{n} \mathbb{P}\Big(B_i = b_i \;\Big|\; E_i = e_i\Big) = \prod_{i=1}^{n} g(B_i = b_i \mid E_i = e_i)$$

$$\text{where } g(B_i = b_i \mid E_i = e_i) = \begin{cases} 1 & : b_i = \emptyset, e_i = \textit{ no event} \\ \frac{1}{\left|\mathscr{B}\right|} & : b_i \in \mathscr{B}, e_i \text{ is NAHR} \\ \frac{1}{\left|\mathscr{B}^2\right|} & : b_i \in \mathscr{B}^2, e_i \text{ is gene conversion} \end{cases}$$

i.e. $B \mid E = e \quad \sim \quad \textit{unif}\Big(\left|\mathscr{B}\right|\Big)$ when e is NAHR, and $B \mid E = e \quad \sim \quad \textit{unif}\Big(\left|\mathscr{B}^2\right|\Big)$ when e is gene conversion.

Despite high homology, long LCRs can have a large number of variational positions; for example, 95% similar LCRs of length 10 kb would have 500 variational positions. Appealing again to low read-error rates and the large amount of data, we restrict $\mathscr{B}$ in practice to keep the computation feasible. Details are in Appendix A.

3.4.2 The likelihood

We now turn to the likelihood $\mathbb{P}\Big(D \mid (E,B)_1^n = (e,b)_1^n\Big)$. Recall that D consists of three types of random variables: the number of reads C, the read sequences R^a, R^b, and the hidden generating locations L. First we treat the number of reads C.

Number of reads C

First, an important conditional independence assumption:

`Assumption` **6.** Fragments of the test genome G are generated independently given G.

It would be convenient to assume that fragments of G are generated uniformly over the length of G, i.e. that $C \sim$ Poisson. However, studies show this to be false; instead, the fragment distribution has been found to be dependent on GC-content [11, 16]. In particular, Benjamini & Speed 2012 found the dependency to be on the GC-content of the entire fragment, as opposed to only the GC-content of the reads or some region near the fragment.

Suppose the fragments input to a sequencing machine have mean length ξ; so a fragment of length ξ may have GC-content $0, \ldots, \xi$. We estimate the fragmentation rate for each possible GC-content k of a fragment of length ξ by dividing the number of observed paired-end reads whose implied fragment of length ξ has GC-content k by the number of positions in $\mathcal{F}$ which have GC-content k in their hypothetical fragment of length ξ. We formalize this below.

Define the function g on $\mathcal{F}$ by $g(x) := \sum_{k=0}^{\xi-1} \mathbf{1}_{\mathcal{F}(x+k)\in\{\mathtt{G},\mathtt{C}\}}$, i.e. $g(x)$ gives the GC-

content of the hypothetical fragment of length ξ whose left-endpoint in $\mathcal{F}$ is x. Let $\Lambda_0^{\mathcal{F}}, \ldots, \Lambda_\xi^{\mathcal{F}}$ be the partition of $\mathcal{F}$ induced by g; so for example, $\Lambda_k^{\mathcal{F}}$ is the set of positions $x \in \mathcal{F}$ such that the hypothetical fragment of length ξ with left-endpoint at x has GC-content k. We partition the data D by GC-content in a similar manner: for each $P \in D$, let ℓ_P be the location P was mapped to by the read aligner. Then for $k = 0, \ldots, \xi$, define $\rho_k := \left|\left\{P \in D : g(\ell_P) = k\right\}\right|$, i.e. the observed number of paired-end reads whose implied fragment of length ξ has GC-content $= k$. We can then approximate the rate of fragmentation for GC-content k by $\lambda_k := \frac{\rho_k}{\left|\Lambda_k^{\mathcal{F}}\right|}$. Note that this approximation assumes that all mutations (SNPs, indels, NAHR, etc.) have a negligible impact on overall GC-content of the genome.[10] We can now define a distribution on read-depth.

`Assumption 7.` The probability that a paired-end read is generated from position $x \in \mathcal{F}$ is $\lambda_{g(x)}$.

For a collection of points $H \subseteq G$, we will use the notation $\lambda_H := \sum_{y \in H} \lambda_{g(y)}$.

Benjamini & Speed 2012 also showed that even when taking into account GC-content bias, the distribution of fragments across the reference genome has variation slightly higher than the variation of the Poisson distribution, which is the distribution often chosen to model read-depth. We include this additional variation by allowing the Poisson distribution parameter to vary according to a Gamma distribution; i.e. a Gamma-Poisson mixture. The resulting distribution is the Negative Binomial distribution, here interpreted as an "overdispersed Poisson". We derive the parameters for the Negative Binomial distribution as follows.

Let $H \subseteq G$ be a portion of the genome G. By assumptions 6 and 7, we should

[10] Note that both studies of coverage bias [11, 16] also analyzed "mappability", i.e. the uniqueness of k-mers in the reference genome. Since our model considers all possible mapping locations, then it is unnecessary for us to correct according to mappability.

expect $\mu := \lambda_H$ fragments to be created from H on average. Thus, μ should be the mean of the Poisson distribution, and also the Gamma distribution. Intuitively, the additional variation is represented by how much the Poisson parameter is allowed to vary about μ, i.e. the variance of the Gamma distribution in the mixture. We chose to specify the variance of the Gamma distribution as some percentage of the mean; that is, we wanted the Gamma distribution to have variance $\sigma^2 := \Big(\alpha \cdot \lambda_H\Big)^2$ for some parameter α. We chose $\alpha = 0.05$ via manual inspection of several regions in test individuals. It is then easy to solve for the parameters of the Gamma distribution, obtaining: scale parameter $k = \frac{1}{\alpha^2}$ and shape parameter $\theta = \alpha^2 \cdot \lambda_H$. We then obtain the parameters of the Negative Binomial by calculating the integral of the Gamma-Poisson mixture.

`Assumption 8.` Let $H \subseteq G$ be a portion of the genome G. The probability distribution on the number of reads C generated from H by the sequencing machine has p.m.f. $\psi\Big(C \;\Big|\; H\Big)$, given by:

$$\begin{aligned}\psi\Big(C = c \;\Big|\; H\Big) &:= \mathit{NegBin}\left(c; \frac{1}{\alpha^2}, \frac{\alpha^2 \cdot \lambda_H}{1 + \alpha^2 \cdot \lambda_H}\right) \\ &= \binom{c + \frac{1}{\alpha^2} - 1}{c} \cdot \left(1 - \frac{\alpha^2 \cdot \lambda_H}{1 + \alpha^2 \cdot \lambda_H}\right)^{\frac{1}{\alpha^2}} \cdot \left(\frac{\alpha^2 \cdot \lambda_H}{1 + \alpha^2 \cdot \lambda_H}\right)^{c}. \quad (3.4.2.1)\end{aligned}$$

In terms of the NAHR variables, we now have a form for the probability of an observed read-depth given a set of NAHR rearrangements.

$$\mathbb{P}\Big(C = c \;\Big|\; (E, B)_1^n = (e, b)_1^n\Big) = \psi\Big(C = c \;\Big|\; (E, B)_1^n = (e, b)_1^n\Big). \quad (3.4.2.2)$$

Having conditioned on the number of reads C, we may now proceed to model

the reads themselves.

Read sequences R^a, R^b

For the moment, suppose we know the generating locations L_1^c of paired-end reads P_1^c. Then we assume that the nucleotides of each read were produced independently of all other reads.

`Assumption` **9.** The sequencing machine generates the bases of each read in a paired-end read independently of all other paired-end reads, and also independently of its mate read, given the test genome G and the generating locations of every paired-end read. That is, the collection of reads $\left\{R_1^a, R_1^b, R_2^a, R_2^b, \ldots R_c^a, R_c^b\right\}$ is a mutually conditionally independent set given G and $L_1, \ldots, L_c$.

Thus we may focus on one read at a time.

The probability of observing a certain read (a string of nucleotides) given its generating fragment (reference sequence) is a product of error-rates for generating each base from the given fragment and for transitioning along the fragment (skipping or lagging along the bases of the fragment). Since we are assuming a specific fragment, then we are looking for a kind of global alignment likelihood for a read and a fragment.

We developed a new theory and algorithm for aligning a paired-end read P to an exact fragment that putatively generated P. Our new pairwise sequence aligner, called a conditional HMM aligner, models the case of one sequence directly "generating" another, rather than considering the two sequences to be evolutionary descendants from some common hypothetical ancestral sequence, as is the case in

conventional sequence alignment theory. This new theory permits the incorporation of biases and error-rates that depend on the features of the reference sequence and are specific to given sequencing technology.

For example, studies have shown that Illumina sequencing machines have context-dependent read-error rates, where the context is the triplet consisting of the base currently being read by the machine and the two previous bases read by the machine [55, 58, 59, 39, 5, 4]. Conventional HMM sequence alignment theory is unable to allow such biases because neither sequence is assumed given, and so a conventional HMM cannot condition on any features (such as contexts) in one of the sequences. Note that this additional fine level of detail is absolutely vital to the study of NAHR: LCRs only differ by a sparse set of variational positions to begin with, and so every read-error matters! In our case, we tailored our aligner to account for the biases and error-rates of Illumina sequencing machines as studied in [55, 58, 59, 39, 5, 4]. Specifically, a conservatively high read-error rate of 2% was assumed, which was increased (liberally) in the presence of contexts and features specifically mentioned in [55, 58, 59, 39, 5, 4]. While we did not properly learn the parameters from the data, we did use higher error-rates than suggested in the literature, which makes our model conservative. Chapter 5 covers this new theory in greater detail. Figure 3.15 gives a schematic example of the conditional HMM aligner and the features affecting error-rates.

While it is theoretically clear how to align a paired-end read P to a suspected generating location $L \in \mathcal{F}$, there are nuisance details that must be addressed in practice. Appendix B discusses the actual procedure of aligning P to L.

In summary,

`Definition.` $\varphi\Big((R^a, R^b) = (r^a, r^b) \;\Big|\; L = \ell\Big)$ gives the marginal likelihood of the

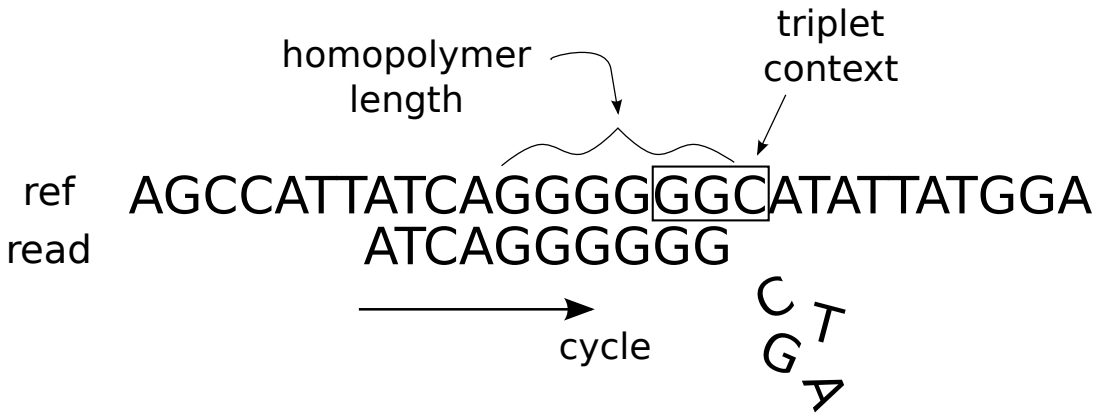


Figure 3.15: A representation of the conditional HMM sequence aligner. As presented in chapter 5, an adjustment in conventional HMM sequence alignment theory allows incorporation of features of the reference genome that are believed to be associated with error-rates and biases in sequencing. Here, the read is being elongated to the right, and the homopolymer length, triplet context, and read cycle are taken into account to determine a read error probability.

paired-end read $P = (R^a, R^b, L)$ given that P was generated from the fragment at location L according to the conditional HMM aligner.

It remains to address the generating location of paired-end reads.

Generating location L

Technically, we never observe the generating location L for any paired-end read P. However, following Assumption 3, we can identify a small set of homologous locations from across the genome which realistically could have generated P. These positions are the M-equivalence class from section 3.1.2, i.e. the set of homologous positions on homologous repeats. As usual when dealing with hidden variables, we marginalize over the (small) set of possible mapping locations for each read P. Thanks to the independence assumptions 6 and 9, we are able to consider each read separately from all other reads. We present this intuition formally below.

For read P_i, define $\mathcal{L}_i\Big((e,b)_1^n\Big)$ to be the set of possible generating locations ℓ in

the test genome G obtained by applying NAHR events $(e,b)_1^n$ to $\mathcal{F}$. The independence assumptions 6 and 9 imply the following derivation:

$$\begin{aligned}
&\mathbb{P}\Big(D = (r^a, r^b, \cdot)_1^c \;\Big|\; C = c, (E,B)_1^n\Big) \\
&\quad = \sum_{\ell_1^c \in \mathcal{L}_1^c\left((e,b)_1^n\right)} \mathbb{P}\Big((r^a, r^b)_1^c \;\Big|\; L_1^c = \ell_1^c\Big) \cdot \mathbb{P}\Big(L_1^c = \ell_1^c \;\Big|\; (E,B)_1^n = (e,b)_1^n\Big) \\
&\quad = \sum_{\ell_1^c \in \mathcal{L}_1^c\left((e,b)_1^n\right)} \prod_{i=1}^{c} \mathbb{P}\Big((r^a, r^b)_i \;\Big|\; L_i = \ell_i\Big) \cdot \mathbb{P}\Big(L_i = \ell_i \;\Big|\; (E,B)_1^n = (e,b)_1^n\Big) \\
&\quad = \prod_{i=1}^{c} \sum_{\ell \in \mathcal{L}_i\left((e,b)_1^n\right)} \varphi\Big((r^a, r^b)_i \;\Big|\; L_i = \ell_i\Big) \cdot \mathbb{P}\Big(L_i = \ell_i \;\Big|\; (E,B)_1^n = (e,b)_1^n c\Big)
\end{aligned} \tag{3.4.2.3}$$

Thus, we may indeed consider the generating location L_i of each read P_i independently.

Before examining the nucleotide sequence of the read, each possible homologous generating location L is *a priori* equally likely because the LCRs are so highly homologous. Thus, we define the distribution on the generating location L of a paired-end read to be

$$L \;\Big|\; (E,B)_1^n = (e,b)_1^n \quad \sim \quad \mathit{unif}\Big(\Big|\mathcal{L}\big((e,b)_1^n\big)\Big|\Big). \tag{3.4.2.4}$$

The joint distribution

Now that we have defined the likelihood and prior distributions and the dependence relationships between all biological and technological random variables, we collect them into one expression for the joint distribution of the data D and the events and breakpoints $(E, B)_1^n$.

Thanks to the partitions of $\mathcal{F}, G, D$, and E_1^n, we can split the likelihood into factors along the partitions. Recall that the partition of $\mathcal{F}$ is constructed from homology and LCRs; by definition of the partition of G, there is a bijection between the partitions of G and $\mathcal{F}$; similarly, there is a bijection between the partition of D and $\mathcal{F}$ (and thus G); and finally, by construction, the partition of E_1^n (i.e. connected components) is merely a subset of the partition of $\mathcal{F}$ (and G and D). Thus, we may collect elements of the partition of D into a coarser partition of D that corresponds to the partition of $(E, B)_1^n$. This could be formally presented, but that would be tedious.

Let $D_{(1)}, \ldots$ be the coarse partition of D that corresponds to the partition $E_{(1)}, \ldots$ of E_1^n. Then the joint distribution breaks into factors along the partition:

$$\mathbb{P}\Big(D = d, (E, B)_1^n = (e, b)_1^n\Big) = \prod_k \mathbb{P}\Big(D_{(k)}, (E, B)_{(k)}\Big) \tag{3.4.2.5}$$

Then, for a single element k of the corresponding partitions of D and E_1^n, we can write the joint as:

$$
\begin{aligned}
&\mathbb{P}\Big(D_{(k)} = d_{(k)}, (E,B)_{(k)} = (e,b)_{(k)}\Big) \\
&= \mathbb{P}\Big(D_{(k)} = d_{(k)} \;\Big|\; (E,B)_{(k)} = (e,b)_{(k)}\Big) \cdot \mathbb{P}\Big((E,B)_{(k)} = (e,b)_{(k)}\Big) \\
&= \mathbb{P}\Big((R^a, R^b, \cdot)_1^{C_{(k)}} = (r^a, r^b, \cdot)_1^{c_{(k)}} \;\Big|\; (E,B)_{(k)} = (e,b)_{(k)}\Big) \\
&\cdot \mathbb{P}\Big(B_{(k)} = b_{(k)} \;\Big|\; E_{(k)} = e_{(k)}\Big) \\
&= \prod_{i=1}^{c_{(k)}} \sum_{\ell \in \mathcal{L}_i\left((e,b)_{(k)}\right)} \varphi\Big((r^a, r^b)_i \;\Big|\; L_i = \ell_i\Big) \cdot \mathbb{P}\Big(L_i = \ell_i \;\Big|\; (E,B)_{(k)} = (e,b)_{(k)}\Big) \\
&\cdot \prod_{(E,B)_j \in (E,B)_{(k)}} g\big(B_j = b_j \;\big|\; E_j = e_j\big) \cdot \prod_{E_j \in E_{(k)}} f(E_i = e_i) \cdot \prod_{\lambda \in \Lambda} (1 - \mathbf{1}_\lambda) \qquad (3.4.2.6)
\end{aligned}
$$

Inference

We ran our model on the connected components of the graphical model in Figure 3.12 with smaller computational complexity. As such, we were able to complete the sums in Bayes' Rule and calculate the posterior distribution exactly.

To make a set of NAHR calls, we first marginalized the posterior space over breakpoints, i.e.

$$
\mathbb{P}\big(E_1^n \;\big|\; D\big) = \sum_{B_1^n = b_1^n} \mathbb{P}\big((E,B)_1^n \;\big|\; D\big). \qquad (3.4.2.7)
$$

This is the posterior space of the NAHR events (deletion, duplication, etc.) without regard to breakpoints. We took this step because, when an NAHR event occurs, there will always be a discernible change in read-depth; however, there may not be

strong evidence of a breakpoint. For example, it may be the case that the switch in variational positions at the breakpoint of a hybrid LCR was not captured by any paired-end reads because the paired-end reads are relatively short and the distance between consecutive variational positions is relatively long.

Due to the amount of data and the very clear distinction between NAHR event signals and *no event* signals, the posterior probability of NAHR event occurrences $\mathbb{P}(E \mid D)$ was always very near 0 or 1. Thus, we computed the *maximum a posteriori* (MAP) estimate for NAHR events E_1^n from $\mathbb{P}(E_1^n \mid D)$. Conditioning on the MAP NAHR event estimates, we proceeded to compute the MAP NAHR breakpoint estimates for each NAHR event from $\mathbb{P}(B_1^n \mid E_1^n, D)$.

In summary, we used a two-step MAP estimator to infer NAHR events and breakpoints, based on the following identity:

$$\mathbb{P}\big((E,B)_1^n \mid D\big) = \mathbb{P}\big(B_1^n \mid E_1^n, D\big) \cdot \mathbb{P}\big(E_1^n \mid D\big) \tag{3.4.2.8}$$

$$= \left[\prod_{j=1}^{n} \mathbb{P}\big(B_j \mid E_j, D\big)\right] \cdot \mathbb{P}\big(E_1^n \mid D\big) \tag{3.4.2.9}$$

3.4.3 Ploidy adjustment

The entire model has so far been developed with a haploid reference and test genome in mind, but humans are diploid. We sketch here how to make a diploid adjustment, but we do not fully specify it because that would be tedious and unenlightening.

We construct the diploid genome by using two copies of the haploid genome. For

terminology, let us say that the above model was developed for the maternal copy of each chromosome. For each event and associated breakpoint random variable E, B described above, we create E', B' that represent the corresponding events and breakpoints on the paternal copy of each chromosome. Each copied event E' has homology dependency with the corresponding E and with all other neighbors of E on the maternal chromosomes (regardless of the type of dependency between E and its neighbors). Also, each copy E' has the dependency relationships with the other events on the paternal chromosomes analogous to the relationships E has with other events on the maternal chromosomes.

This diploid adjustment increases the computational complexity of our model substantially, as it duplicates every node of the graphical model shown in Figure 3.12 and creates homology dependencies between each node and all other nodes in the connected component on the "other" copy of the chromosomes.

Chapter Four

Empirical FDR

4.1 Development of FDR and a new approach

Classically, a statistician takes a datapoint and asks, "If the null hypothesis is true, what is the probability that I will see a datapoint as unlikely as this one or more unlikely? If that probability p is less than α, then I will reject the null hypothesis". Later, the issue of multiple comparisons was raised: "If you examine a large number of datapoints, then you should expect some of them to be as unlikely or more-so than your critical value α by chance alone. So then how do you identify datapoints which are *actually* statistically significant?" Dunn addressed this problem with a simple, conservative suggestion: that the statistician adjusts the critical value α so that *less than one* datapoint from the null is expected to be statistically significant due to chance alone (using a Bonferroni inequality) [21]. A less stringent alternative was later proposed by Benjamini & Hochberg: the statistician should estimate the proportion of discoveries expected to be false, called the false discovery rate (FDR) [10]. Storey & Tibshirani extended the FDR concept to the q-value, for assigning an FDR-like probability to each datapoint [75].

Bradley Efron highlighted a practical complication, reminding us that all models are flawed and that this will have a major impact on calculating the FDR and identifying statistically significant datapoints. The p-values of samples drawn from the null distribution should be uniformly distributed on $(0, 1)$, and so the corresponding z-values should be $N(0, 1)$. But an even slightly flawed null likelihood distribution will produce z-values which are *not* $N(0, 1)$. Efron pointed out that even a slight deviation from $N(0, 1)$ has major ramifications for statistically significant datapoints. As a practical adjustment, acknowledging that all models are flawed, he suggested *learning* the distribution of z-values under the null using a carefully chosen subset of the datapoints when there is enough data available (i.e. in the big data scenario).

Further, Efron showed that the FDR framework is really just a 2-component finite mixture of transformed p-values, and the FDR is thereby calculated from Bayes' rule. Finally, Efron also pointed out that in the big data scenario, the distribution of the test statistic can be estimated non-parametrically, making his FDR procedure more robust [23].

But all of these FDR procedures take as given what is perhaps the most challenging step: that the investigator chose an *accurate* distribution for the data and derived statistic under the null hypothesis. This step is, of course, not trivial; it is the art of probabilistic modeling!

Efron illustrated that FDR is merely Bayes' rule performed on a 2-component mixture of z-values. Thus, FDR is merely a statistical test performed on the result of another statistical test - quite confusing. Why start there? Why not perform "FDR" directly on the data, rather than calculating an inherently flawed p-value and then trying to correct an FDR procedure using a non-parametric approach? Instead, we should perform an "FDR-like" test *immediately* on a statistics derived from the raw observed data.

We extend the insights of Efron and draw out the Bayesian nature of FDR; namely, that *FDR is merely finite mixtures in disguise*. Efron suggested that when enough data is available, the distribution of z-values under the null hypothesis can be estimated from a particular subset of the data, provided certain assumptions hold. We generalize this idea and apply it one step *before* the test for statistical significance. That is, in the big data scenario, we non-parametrically estimate the distribution of the *test statistic, not the* p-*value*, under the null. We then apply a procedure similar to Efron's to represent the probability of an interesting *datapoint*, rather than an interesting z-value. We name this procedure "empirical FDR".

4.1.1 Formulation of empirical FDR

Suppose we have "big data", in which we can identify a large amount of data that can be safely assumed to be from the null. We define the following random variables,

- X = observed data/statistic.
- Y = datapoint is from the null ($Y = 0$) or alternative ($Y = 1$).
- C = datapoint is from control data ($C = 0$) or from test data ($C = 1$).
- f = empirical distribution of X. (mixture of null and alternative)
- $\theta_{Y,C}$ = parameters of mixture distributions that compose f.
- π = mixing proportions for f.

We posit that the likelihood of the data X is given by a mixture of null and alternative distributions:

$$\mathbb{P}\big(X \mid \theta, \pi\big) = \sum_{i=0}^{1} \mathbb{P}\big(X \mid \theta_{i,1}, Y = i\big) \cdot \mathbb{P}\big(Y = i \mid \pi\big) \tag{4.1.1.1}$$

i.e.,

$$f(X; \theta, \pi) = \sum_{i=0}^{1} f_i(X; \theta_{i,1}) \cdot \pi_i \tag{4.1.1.2}$$

To perform FDR-like calculations in the manner of Efron (below), we need to estimate the mixture distribution f of the test statistic X, and also the distribution f_0

of X under the null hypothesis (i.e. we need to estimate $\theta_{0,1}$).

The test statistic's distribution f can be easily estimated non-parametrically, thanks to the big data set. As for the null distribution f_0, we will consider the situation in which a relevant control data set is available. For example, in many molecular genetics and genomic studies, data is obtained from both treated samples and control samples (e.g. knockout mice and wild-type mice). In other cases, such as the NAHR scenario addressed here, data is sampled from across the genome, but only a *predetermined* set of regions of the genome are of interest, and therefore data from the regions of the genome that are predetermined to *not* be of interest serve as control data.

We start by estimating the likelihood for X in the control data, which is given by,

$$\mathbb{P}\big(X \;\big|\; \theta_{0,0}, Y = 0, C = 0\big). \tag{4.1.1.3}$$

We estimate $\theta_{0,0}$ using the control data. In the big data setting, this can be done non-parametrically for robustness, given that some portion of the big data set can be identified and safely assumed to be drawn from the null.

We want a model for the data in the test region under the null hypothesis, i.e. we want to estimate $\theta_{0,1}$. We would like to use the form and parameters $\theta_{0,0}$ learned from the control data, but we recognize, as did Efron, that models are not perfect, and there are confounding effects etc., which necessitate small practical adjustments to get from $\theta_{0,0}$ to $\theta_{0,1}$. This can be done in a variety of ways. In the NAHR case below, we again draw from Efron and adjust $\theta_{0,0}$ to get to $\theta_{0,1}$ using location and

scale adjustment parameters α, β. We determine α, β by fitting the same truncated part of f as did Efron.

Below, we apply this procedure to the NAHR problem.

4.2 Application of empirical FDR to the Bayesian NAHR model

Ideally, we would be sufficiently confident in the Bayesian model described in Chapter 3 to present its raw results and have them experimentally validated in a biology lab. We did not do this for two reason. First, experts in the techniques of validating genome rearrangements confirmed, via personal correspondence, that current methods of experimental validation were unreliable in the highly homologous repetitive regions evaluated by our NAHR model, and so validation could not be done. Second, we were aware of several limitations and shortcomings of our model, and wished to restrict our initial calls to a more conservative, high quality set. To restrict our calls, we designed a novel statistical test based on the work of Efron in [23].

4.2.1 Limitations of the Bayesian NAHR model

Often in mathematics, bad ideas and failed approaches are discarded to the trash bin of history and never spoken of, and the solution is presented as a complete, pristine entity. But a short review of failed approaches can sometimes motivate the correct solution. We briefly discuss a sequence of events which lead to the development of an "empirical FDR" approach based on the work of Efron 2004 [23].

The first instance of the Bayesian model presented in Chapter 3 ignored gene conversion. This lead to many puzzling calls, wherein the same event was called as a deletion on one copy of a chromosome, a duplication on the other copy, and *both* calls had very strong evidence of breakpoints. Of course a deletion and duplication at the same location makes little sense; a much simpler explanation is that there was *not* an NAHR event at all at that location (they cancel each other out!). But then how do we explain the strong breakpoint evidence?

This paradox can be explained by the gene conversion mechanism, as it follows the exact same homologous recombination pathway as NAHR until the last step, except that the double Holliday junctions are resolved via non-crossover in gene conversions. Clearly, the breakpoint evidence of "switches" in variational position patterns was so strong that it forced the model to address it in the call. In some sense, a deletion and duplication at the same locus is the "closest point" in the space of NAHR events to a gene conversion event at that locus.

Thus, we rebuilt the Bayesian model to include gene conversion events, as presented in 3. While this lead to an improvement in our model's results, it did not solve the problem. From manual inspection of our calls (not shown), it seems that the "switches" in variational position patterns can often be far more complicated than we assumed in 1. This could be due to a number of causes, including a misspecified reference genome, or a complicated behavior of the mismatch repair mechanism.

The gene conversion mechanism imposed a substantial computational burden on our Bayesian model. Not only did it increase the sample space of each event on each chromosome by 2, but gene conversion events require a *pair* of breakpoints, thus squaring the size of the breakpoint space. As discussed in 3.4.1, we imposed a reasonable heuristic that glanced at the data and filtered out incredibly unlikely

breakpoints for NAHR or gene conversion events, but to keep the computation feasible, we restricted the possible breakpoints to an incredibly tiny set (four). This means that complicated breakpoint regions, of NAHR or gene conversion, could not be explained by our model due to our drastic restriction of allowable breakpoints and assumption of simple breakpoint (assumption 1). Thus, the Bayesian model in Chapter 3 still produced puzzling false positive calls, albeit at a smaller rate than previous instances.

To filter out false positive NAHR deletion and duplication calls which the Bayesian model called presumably because of an inadequate breakpoint model and gene conversion, we sought to develop a statistical test based on read-depth alone, believing that the read-depth signal in candidate NAHR regions was sufficiently strong to alone call a deletion or duplication. We extracted the read-depth distribution (section 3.4.2) from the Bayesian model and attempted several parametric and non-parametric statistical tests to evaluate our calls. None of these approaches worked, and it appeared to be because the model for read-depth which we obtained from the literature was substantially flawed (discussed later).

Thus, we sought a novel test statistic for read-depth and a method of evaluation which somehow accounted for the misspecifications of the original read-depth model, yet still included its insights as far as the fragment bias due to GC-content. We refer to the test statistic as the *empirical read-count ratio* and the evaluation procedure as *empirical FDR*. Both are discussed below.

4.2.2 The empirical read-count ratio γ

For a statistic, we choose to use the ratio γ of observed read-depth to GC-sensitive expected read-depth under the null hypothesis (*no event*). The expected number of reads is calculated according to the GC-sensitive fragmentation rate, as described in 3.4.2. Thus, the chosen statistic γ normalizes by the length of the region and is intended to account for GC-content bias.

In the Bayesian model, we evaluated read-depth over a partition of homologous regions. For our empirical FDR procedure, we wished to instead evaluate read-depth over single, contiguous intervals, regardless of homology to other regions of the genome. The traditional read-depth statistic is just the count of reads in an interval, but of course this must be reconsidered in the context of repeats. Thus, we had to develop another new way to obtain read-depth for a specific interval which contains repeats. Below we discuss how to calculate γ.

Calculating γ in unique regions

In unique regions, there is no ambiguity in the mapping of reads: we can safely assume that the read really was generated from the location it was mapped to with probability $= 1$, as discussed in 3. Thus, for unique region U, we calculate γ, in the traditional way: by simply counting. In the notation of 3.4.2, let P be a paired-end read, ℓ_P be the mapping location of P, $g(x)$ be the GC-content of the hypothetical fragment starting at x, and $\lambda_{g(x)}$ be the fragmentation rate at position x. Then,

$$\gamma = \frac{\left|\left\{P \;\middle|\; \ell_P \in U\right\}\right|}{\sum\limits_{x \in U} \lambda_{g(x)}} \tag{4.2.2.1}$$

Calculating γ in repeat regions

This is the non-trivial case. Our overall strategy is as follows. Since reads from repeats cannot be mapped to a single instance of the repeat with overwhelming probability (as they can be in unique regions, usually), then we calculate the likelihood that the read was generated from each repeat, and use Bayes' Rule to create a posterior distribution over repeats for a given read. We also use this method to calculate the expected read-depth in repeats when deletions or duplications are suspected to have occurred. A technical exposition follows.

Consider a paired-end read P mapped to position x inside of an LCR. In the notation of 3.1.2, M_x represents the set of all positions homologous to x, i.e. the space of possible generating locations of P. *A priori*, P is equally likely to have come from each $z \in M_x$, and so $\forall z \in M_x, \mathbb{P}\big(z\big) = \frac{1}{\left|M_x\right|}$. For each possible generating location $z \in M_x$, we align P using the conditional HMM aligner to obtain the likelihood $p_z := \mathbb{P}\big(P \;\big|\; z\big)$. Then we form the posterior distribution on possible generating locations for P as

$$\mathbb{P}\big(z \;\big|\; P\big) := \frac{\mathbb{P}\big(P \;\big|\; z\big) \cdot \mathbb{P}\big(z\big)}{\sum\limits_{w \in M_x} \mathbb{P}\big(P \;\big|\; w\big) \cdot \mathbb{P}\big(w\big)} \tag{4.2.2.2}$$

Finally, suppose X is a repetitive region. Let $P_1, \ldots, P_n$ be the set of all paired-

end reads who have a possible generating location inside of X, i.e. $\forall i$, if P_i has possible generating locations M_i, then $M_i \cap X \neq \emptyset$. We calculate γ for the region X as

$$\gamma = \frac{\sum_{i=1}^{n} \sum_{z \in M_i \cap X} \mathbb{P}\big(z \mid P_i\big)}{\sum_{x \in X} \lambda_{g(x)}} \tag{4.2.2.3}$$

Now that we know how to compute the test statistic γ, we show how to evaluate γ via a non-parametric adaptation of the local FDR procedure developed by Efron [23], i.e. "empirical FDR".

4.2.3 Developing empirical FDR by adapting Efron's local *fdr*

Review of Efron's local *fdr* and comparison to the NAHR scenario

Following Efron, we view the false discovery rate from the Bayesian perspective: the observed empirical distribution f of the test statistic γ is really a mixture of two distributions: f_0, the distribution of γ under the null hypothesis, and f_1 the distribution of γ under the alternative. We could then calculate the "local FDR", written *fdr*, as

$$fdr(\gamma) := \frac{f_0(\gamma)}{f(\gamma)} \tag{4.2.3.1}$$

Note that this is actually, an upper-bound for the FDR, since we ignore the mixing proportion coefficient from the numerator.

The important question is then: what is the distribution f_0 of γ under the null hypothesis? Assuming that our GC-sensitive read-depth model is quite good, we would naturally suppose that $\gamma \sim N(1, \sigma^2)$ for some σ^2. But this assumption is not necessarily true. Although substantial progress has been made in identifying biases in coverage in Illumina sequencing machines [11], not all sources of bias have been investigated, and no complete, verified model with all parameters learned is available.

Efron developed the above perspective of *fdr* in the context of normalized z-values for some test. In such a scenario, the z-values for data coming from the alternative hypothesis are thought to be generally far from 0, while the z-values for the data generated from the null hypothesis are thought to be generally close to 0. The distribution f_0 of the z-values coming from the null hypothesis in Efron's case would traditionally be assumed to be $N(0, 1)$. But Efron shows that if f_0 differs even somewhat from $N(0, 1)$, then it has important implications for the consequent FDR calculations and thus for identifying a multiple-comparison corrected statistically significant subset of the data. Efron proposed that when a large amount of data is available, f_0 can instead be estimated empirically from a subset of the data. Therefore, Efron suggested that if $<\sim 10\%$ of the data is believed to come from the alternative hypothesis, then the distribution of the z-values under the null can be estimated using only the data around the large, central peak near 0, as most datapoints producing z-values in this region near 0 are assumed to be from the null [23].

We adapt Efron's approach and apply a similar idea to our case for the ratio

γ. For regions that did not experience an NAHR deletion or duplication, we expect γ to generally be near 1. For regions that did experience an NAHR deletion or duplication, we expect γ to be much less than 1 (for deletions) or much greater than 1 (for duplications). Therefore, we may empirically estimate the distribution f_0 of γ under the null hypothesis by using the γ in a small interval around the central peak near 1, analogously to Efron.

But we have a considerable advantage in our big data scenario: an expansive control region of the genome unlikely to have experienced an NAHR deletion or duplication. Therefore, for each individual, we can estimate the distribution f_0 under the null using data from the "control region" of the genome: the unique regions, away from LCRs and potential NAHR events, where we may safely assume that there are relatively very few genome rearrangements.

For a given individual, let the observed distribution of γ among the potential NAHR events be f. As in Efron 2004, f can be obtained by smoothing the observed empirical distribution of all γ (we used lowess smoothing). Thus, to calculate the *fdr* as in 4.2.3.1, we only need an expression for f_0. Below, we detail how to construct the distribution f_0 of γ under the null distribution.

Constructing f_0

For each individual, and for each potential NAHR locus tested, we randomly selected a contiguous region of the same length as the potential NAHR locus from the "control region" of the reference genome (away from LCRs and potential NAHR loci, not near the centromere or telomere) and calculated its observed-to-expected read count ratio γ. For each individual, we thus empirically estimated $f_0^c :=$ the distribution of

γ under the null hypothesis in the control region of the genome.

It turns out that the means of the empirically estimated f_0^c for each individual vary considerably, as shown in Figure 4.2. They also deviated substantially from Normal, as seen in the QQ-plot in figure 4.1. This serves as confirmation that the GC-sensitive coverage model learned from the literature is somewhat problematic. We also concluded from our exploratory data analysis 1) The *fdr* for each individual should be calculated separately, i.e. *not* by pooling observations across individuals; 2) We should *not* assume that γ has mean 1 under the null hypothesis for each individual; and instead 3) That the test region and control region read-depth is reasonably approximated by a separate 2-component Gaussian mixture f_0^c to each individual, denoted as

$$f_0^c(\gamma) := p_0 \cdot g(\gamma; \mu_0, \sigma_0^2) + p_1 \cdot g(\gamma; \mu_1, \sigma_1^2), \tag{4.2.3.2}$$

where g is the Gaussian density.[1]

Here we observed that the central peaks of f_0^c and the observed distribution f for potential NAHR loci differed considerably. Thus, to derive a distribution for γ for potential NAHR loci leveraging the information from the control region, we enforced a series of constraints, inspired by the similar adjustment made by Efron, to transform f_0^c into an appropriate f_0.

For a given individual, let the observed distribution of γ in the potential NAHR loci be f. As in Efron 2004, f can be obtained by smoothing (we used lowess smoothing). To estimate f_0, Efron suggested using the center and half-width of the central peak in f to define a Normal distribution. In this spirit, we transform f_0^c

[1] In future work, we could explore a hierarchical model for the distributions across different individuals, rather than having each individual be completely independent.

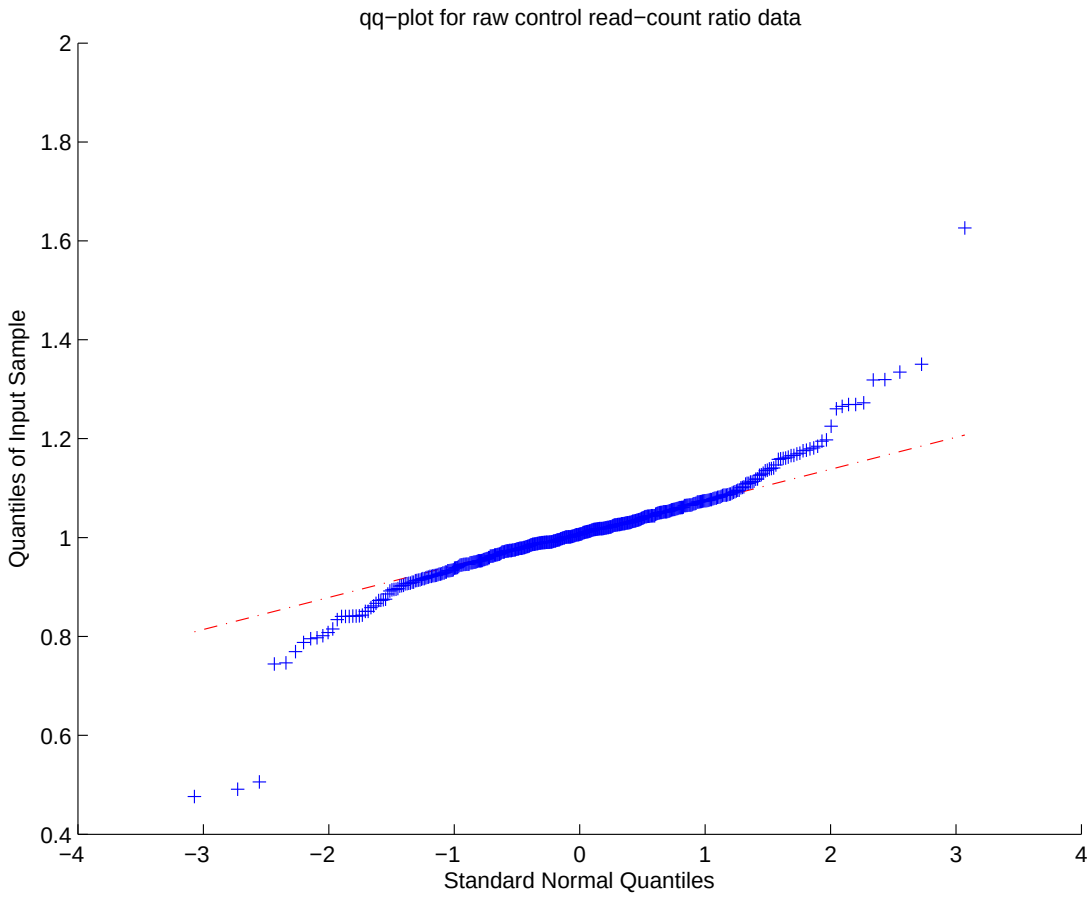


Figure 4.1: QQ-plot of the empirical read-count ratio γ in control regions for individual NA19818. Notice that the empirical read-count ratio γ deviates substantially from Normal.

into f_0 by minimizing the Euclidean distance between their half-widths and central peaks, as follows.

For some function h, let $P_c^h := (x_c^h, y_c^h)$ be the geometric location of the central peak of h, i.e. $h(x_c) = y_c$ and $y_c = \text{argmax}_y\, h(y)$. Let $P_\ell^h := (x_\ell^h, y_\ell^h)$ be the geometric location of the lower half-width, i.e. $h(x_\ell^h) = y_\ell^h = \frac{y_c^h}{2}$ with $x_\ell^h < x_c^h$. Similarly, $P_u^h := (x_u^h, y_u^h)$ be the geometric location of the upper half-width, i.e. $h(x_u^h) = y_u^h = \frac{y_c^h}{2}$ with $x_u^h > x_c^h$. Also, let $d(P_1, P_2)$ be the Euclidean distance between points P_1, P_2. We define an adjustment to the control Gaussian mixture in equation 4.2.3.2 as

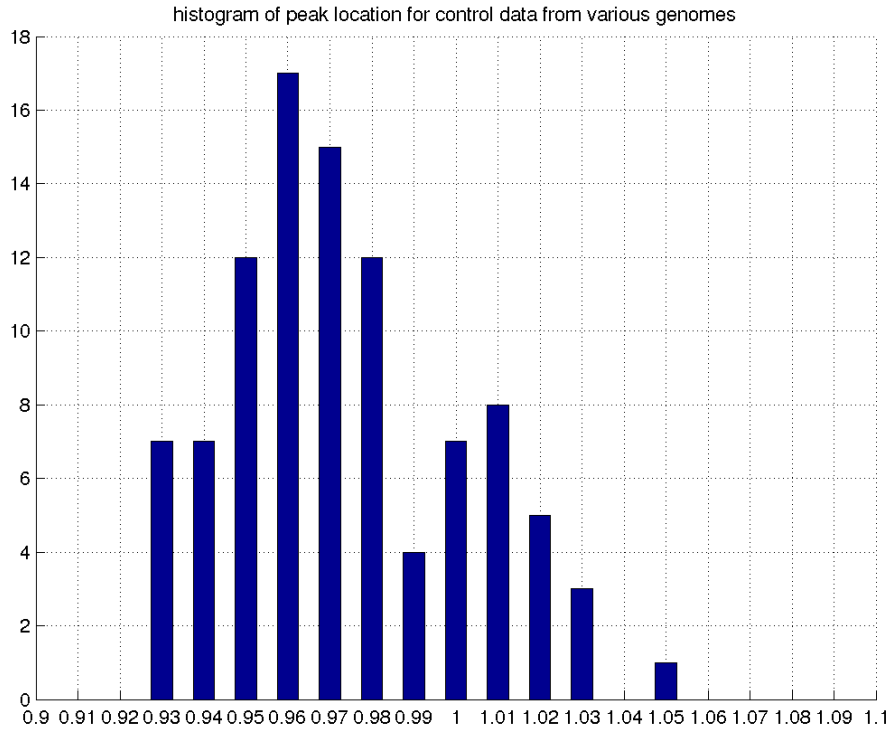


Figure 4.2: Histogram of the central peak location for the observed-to-expected read-depth ratios for data from the control regions of the 44 individuals tested. Notice the considerable variation in central peak location - this suggests that we should *not* simply assume that the observed-to-expected read-depth ratio γ has mean 1. Indeed, it seems that the GC-sensitive fragmentation rate calculated according to the literature actually overestimates the expected read-depth.

$$f_{\alpha,\beta}(\gamma) := p_0 \cdot g\big(\gamma; \mu_0 + \alpha, (\sigma_0 \cdot \beta)^2\big) + p_1 \cdot g\big(\gamma; \mu_1 + \alpha, (\sigma_1 \cdot \beta)^2\big), \tag{4.2.3.3}$$

where, again, g is the Gaussian density.

Finally, we set $f_0 := f_{\alpha^*,\beta^*}$ where

$$(\alpha^*, \beta^*) = \underset{(\alpha,\beta)}{\operatorname{argmin}}\, d(P_\ell^f, P_\ell^{f_{\alpha,\beta}}) + d(P_c^f, P_c^{f_{\alpha,\beta}}) + d(P_u^f, P_u^{f_{\alpha,\beta}}) \tag{4.2.3.4}$$

We are now able to calculate the empirical FDR as in 4.2.3.1.

4.2.4 Differences from Efron

Our empirical FDR method deviates substantially from Efron's local FDR, even though we drew heavily from his procedure. Efron **takes the p-values as given**, and proceeds to empirically estimate the distribution of corresponding z-values by a mixture. This rests on several crucial assumptions:

Statistical assumptions in Efron's analysis

1. The likelihood that produced the p-values well-approximated the true distribution of the data under the null.
   (a) i.e. the investigator had an appropriate probability distribution for the test statistic, and so the p-values are reasonably reflective of reality.
   (b) *But coming up with a good model for the test statistic is the hardest part!*
2. Efron knows approximately where most of the null data should lie, where the alternative data should lie, and that they are separated.
   (a) Since the p-values were calculated correctly (1), he knows that the z-values for the null data should be around 0.
3. He assumes that $\leq 10\%$ of the data is from the alternative.
4. Assuming that the p-values are appropriate, then Efron knows that the null distribution for z-values should be approximately Normal, and approximately $N(0, 1)$.

   (a) The p-values for data under the null should have distribution $unif(0,1)$. Hence z-values for null data have distribution $N(0,1)$.

5. Knowing that the null distribution of z-values is approximately $N(0,1)$, Efron can estimate the parameters of f_0 *using only the data around the central peak* (i.e. he truncates the data).

   (a) That is, he does not worry about what the tails of f_0 are like because he assumes it is Normal!

   (b) Thus, when he fits an empirical f_0, he can ignore the null data in the tails.

But our situation differs on several major points. Efron starts one step ahead of us: suppose you have good p-values, what now? We are one stage early: we do not have a model good enough to get reliable p-values, so then how can we empirically calculate FDR directly from the data, without recourse to a model which gives p-values?

Compared to Efron's assumptions listed above, our scenario differs substantially:

Statistical realities in the NAHR problem

1. Unlike 1, we do not have a good model for the test statistic, and so we cannot get reliable p-values.

   (a) The literature on the model of the test statistic (read-depth) is inadequate.

   (b) The data is extremely messy and complicated - flawed lab preparation, artifacts, post-processing of the data, aligner heuristics, etc.

2. Unlike 2, we do not know where the null data should lie. We *think* it should be around 1, but we don't know for sure.

   (a) Samples from the control region of the genome confirm that it is true, our null should be around 1.

3. We also assume that $\leq 10\%$ of the data is from the alternative via appeal to *a priori* knowledge of biological rarity of NAHR.

4. Unlike 4, we have no idea what the null distribution f_0 should be like.

   (a) We need control region data to estimate f_0.

5. Unlike 5, we cannot use *only* the central peak to estimate f_0 because we have no idea what the tails of f_0 should be like.

   (a) Efron's assumptions that model that gave p-values is good $\implies$ f_0 has Normal form $\implies$ he knows what the tails are like.

   (b) We need control region data to get an idea of the tails of f_0. (QQ-plots confirmed that Normal is insufficient in our case, so we turned to a mixture of Normals).

Chapter Five

A context-dependent conditional HMM

5.1 Motivation

A common approach for analyzing short-read data from high-throughput sequencing is to align them to some known sequence, usually a reference genome for the organism in question. Many algorithms and programs have been developed and optimized for this purpose: to quickly map the millions of short-reads that are output from sequencing machines onto a known genome [46, 7, 28, 77, 41, 47, 49]. To make such an operation computationally feasible, general heuristics are employed, such as allowing only a maximum edit distance in an alignment, disallowing gaps, or using seed-and-extend strategies. These methods serve their purpose very well, and these computational tools are widely used in the bioinformatics community.

Each sequencing technology has its own biases and error-rates, both in genome coverage and in generating base-calls. However, the aforementioned computational tools are indifferent to these critical shortcomings of sequencing technology. While certain parameters of the large-scale aligners may be tunable by the users, in general there is no correspondence to any specific biases in the read data, and are thus uninformed of any systematic biases or errors in the sequencing technology.

In many applications of sequencing, the basic alignment method in these large-scale aligners is sufficient. For example, for analysis of copy-number variation, small discrepancies in alignments are mostly irrelevant. However, in studies concerned with specific positions in the genome, such as RNA editing or the genome-wide association study (GWAS), then tiny adjustments in an alignment have important ramifications for analysis and interpretation. In the present study, we are forced to rely on the sparse set of SNP and indel differences that distinguish highly homologous repeats to detect the occurrence of rearrangements resulting from NAHR in the human genome.

Since LCRs are already so highly homologous, determining a meaningful likelihood over the possible generating locations of a given read requires an alignment algorithm that is carefully informed by the sequencing biases and error rates.

We have developed an hidden Markov model (HMM) specifically for read generation, which we refer to as a *conditional HMM*. The theoretical foundation of our conditional HMM is fundamentally different from conventional sequence alignment theory. We consider "conventional" sequence alignment theory to be that appearing in Durbin, et. al. 1998 [22].

Most high-throughput sequencing technologies construct each individual read according to a linear, base-by-base generating mechanism. This linear process is decidedly *not* captured by conventional HMM sequence alignment theory. By constructing a new theoretical formulation that *does* appropriately model the actual sequencing process, we are able to consider characteristics of the reference genome when determining the probability that a given read was generated from a certain region of the reference. This new model allows sufficient flexibility for the inclusion of specific biases and error-rates of a given sequencing technology, and is sufficiently general to be applicable to any sequencing technology that reads bases in a sequential fashion.

Illumina sequencing machines are the most popular in the sequencing community. Several studies have analyzed various aspects of Illumina's sequencing process and characterized its biases and reported error profiles [55, 58, 59, 39, 5, 4, 11, 17]. We illustrate the application of our new HMM to high-throughput reads generated by Illumina sequencing machines.

5.2 Conditional HMM theory

Conventional HMM sequence alignment theory supposes a fixed but unknown ancestral sequence, from which *pairs* of bases (one possibly empty, i.e. a "gap") are sequentially emitted according to a Markov chain. The state sequence of the Markov chain is hidden, and the only observations are the final sequences A and B. Since the ancestral sequence is completely unknown and the sequences A and B are only ever considered *jointly*, then features of the ancestral sequence that may impact the transition and emission probabilities (e.g. certain motifs, homopolymers, etc.) *cannot* be considered at any point in the generation of the two sequences A, B. In short, standard sequence alignment theory is uninformed by contexts of the generating, ancestral sequence which may have an impact on the sequences being generated.

In analysis of high-throughput sequencing data, a reference genome sequence is both fixed and known, and analyses generally assume that the reads were generated directly from that fixed, known reference sequence. This stands in direct contrast to the conventional HMM theory. In the case of high-throughput sequencing data, the reference sequence is completely fixed and known, and the read is generated *conditionally* from that reference; it is *not* the case that both the read and reference are generated jointly, as is the foundation of conventional HMM sequence alignment theory.

Conditioning on the reference genome which generated the reads thus provides a crucial benefit that standard sequence alignment theory does not. Specifically, a conditional distribution of a read given a reference can be made sensitive to features of the reference sequence.

The transition and emission random variables in sequence alignment theory have

new, more intuitive, physical interpretations in conditional HMMs compared to conventional "ancestral sequence" HMMs. Now, for conditional HMMs: "insert" means an insertion in the read with respect to the reference, and occurs when the image processing step falsely detected a new base during read generation which did not exist in the fragment serving as a template; "delete" means a deletion in the read with respect to the reference, and occurs when the image processing step failed to call a base during read generation; "match" means that a certain base in the read was read from a certain base in the reference, and occurs when the image processing step detected the presence of a base corresponding to a certain position in the template; and "mismatch" means that a certain base in the read is matched to a certain base in the reference, but they disagree, and occurs when the image processing step detected the wrong base.[1] A mismatch is usually thought of as a "read error".

Despite these changes, the probabilistic forms of the conditional HMM and related alignment algorithms are essentially the same as in the conventional "ancestral sequence" HMM case; we only need to generalize certain transition and emission expressions to allow incorporation parameters representing the various types of errors and biases discussed above.

First, some definitions and notation.

- $\mathcal{F} = \mathcal{F}_1\mathcal{F}_2 \ldots \mathcal{F}_n$ is the reference sequence of the genome from which reads were generated. $\mathcal{F}$ is a completely defined, deterministic, fixed sequence of nucleotides, in $5'$ to $3'$ order.

- Let the transition state space be $\mathcal{T} := \{\texttt{M,I,D}\}$ where `M,I,D` stands for match,

[1] Of course, it could have in fact detected the correct base, but there was a SNP in $\mathcal{F}$ which makes it appear as "mismatch". We ignore the SNP possibility because we condition on $\mathcal{F}$, i.e. assume that $\mathcal{F}$, as is, generated R.

insert, delete, respectively.

- Let the space of emissions be the set of nucleotides; $\mathcal{N} := \{\texttt{A},\texttt{C},\texttt{G},\texttt{T}\}$.
- A read R is a random vector of some fixed length m, whose components $R_1, \ldots, R_m$ are random variables that take values in $\mathcal{N}$.
- A is the alignment of R to $\mathcal{F}$. Here, $\forall t \in \mathcal{T} : A_t$ is a binary $m \times n$ matrix, where m is the length of R and n is the length of $\mathcal{F}$, and $A_t(i,j) = 1$ if the state relating R_i and $\mathcal{F}_j$ is t.
- $\theta_{\mathcal{T}}$ are the transition probability parameters; $\theta_{\mathcal{N}}$ are the emission probability parameters.
- $\forall s \in \mathcal{T} : \mathbf{1}_{s \to \texttt{M}}(i,j)$ is the indicator function for transition to the match state, i.e. $\mathbf{1}_{s \to \texttt{M}(}i,j) = 1$ when $A_s(i-1, j-1) = 1$ and $A_{\texttt{M}}(i,j) = 1$.
- $\forall s \in \mathcal{T} : \mathbf{1}_{s \to \texttt{I}}(i,j)$ is the indicator function for transition to the insert state, i.e. $\mathbf{1}_{s \to \texttt{I}}(i,j) = 1$ when $A_s(i-1, j) = 1$ and $A_{\texttt{I}}(i,j) = 1$.
- $\forall s \in \mathcal{T} : \mathbf{1}_{s \to \texttt{D}}(i,j)$ is the indicator function for transition to the delete state, i.e. $\mathbf{1}_{s \to \texttt{D}}(i,j) = 1$ when $A_s(i, j-1) = 1$ and $A_{\texttt{D}}(i,j) = 1$.

The joint distribution of a read R and an alignment A is:

$$\mathbb{P}\big(R, A \mid \mathcal{F}\big) = \prod_{j=1}^{n}\prod_{i=1}^{m}$$
$$\prod_{t\in\mathcal{T}}\Big[\mathbb{P}\big(R_i \mid A_{\mathtt{M}}(i,j)=1, \mathcal{F}, \theta_{\mathcal{N}}\big)$$
$$\cdot\,\mathbb{P}\big(\mathbf{1}_{t\to\mathtt{M}}(i,j)=1 \mid \forall t\in\mathcal{T}: A_t(i-1,j-1), \mathcal{F}, \theta_{\mathcal{T}}\big)\Big]^{\mathbf{1}_{t\to\mathtt{M}}(i,j)}$$
$$\cdot\prod_{t\in\{\mathtt{M},\mathtt{I}\}}\Big[\mathbb{P}\big(R_i \mid A_{\mathtt{I}}(i,j)=1, \mathcal{F}, \theta_{\mathcal{N}}\big)$$
$$\cdot\,\mathbb{P}\big(\mathbf{1}_{t\to\mathtt{I}}(i,j)=1 \mid \forall t\in\mathcal{T}: A_t(i-1,j), \mathcal{F}, \theta_{\mathcal{T}}\big)\Big]^{\mathbf{1}_{t\to\mathtt{I}}(i,j)}$$
$$\cdot\prod_{t\in\{\mathtt{M},\mathtt{D}\}}\Big[\mathbb{P}\big(R_i \mid A_{\mathtt{D}}(i,j)=1, \mathcal{F}, \theta_{\mathcal{N}}\big)$$
$$\cdot\,\mathbb{P}\big(\mathbf{1}_{t\to\mathtt{D}}(i,j)=1 \mid \forall t\in\mathcal{T}: A_t(i,j-1), \mathcal{F}, \theta_{\mathcal{T}}\big)\Big]^{\mathbf{1}_{t\to\mathtt{D}}(i,j)} \tag{5.2.0.1}$$

5.3 Biases and errors of Illumina sequencing machines

Several studies of Illumina sequencer biases and error profiles have been made since the advent of Illumina's Genome Analyzer II machine [55, 58, 59, 39, 5, 4, 11, 17]. Each study reports slightly different characterizations of biases and errors, sometimes contradicting other studies. Only two analyses, Meacham et. al. 2011 and Abnizova et. al. 2012, performed statistical analyses to support their claims [55, 4]. As such, our conditional HMM error parameters are grounded in the findings of these two studies, but we we also consider select findings from the studies that do not perform any statistical analyses to make our error model more robust.

Different Illumina machines were used among the studies we reviewed for identi-

fying biases and error-profiles, and all were at least as recent as the Genome Analyzer II model. As for the error model we include in our conditional HMM, we will not include parameters specific to machines, versions of machines, or runs on a machine, even though preliminary analysis shows that error-rates vary non-trivially in all three cases.

Below we discuss the specific features that other studies found to be associated with systematic biases and elevated error-rates. There are four basic features associated with biases and changes in error-rates: contexts, quality scores, homopolymers, and cycle. See Figure 3.15 for a schematic. When the machine is reading position i in the reference genome $\mathcal{F}_j$, the context of j is considered to be the nucleotide at position j of $\mathcal{F}$ and the nucleotides at some number of positions in the reference (the length of the context) that were read *prior* to position j; see Appendix D for a technical definition of the notion of "prior" bases read. The cycle of a read base is the number of bases which have been read thus far in the creation of the read; so the cycle of R_i is i, where the first base in R that was read is denoted R_1. Each base in a read has an associated Phred quality score $Q_{\text{phred}} = -10 \cdot \log_{10}(p_{err})$, i.e. the log-probability that the base-call is wrong, produced by post-processing base-callers from measurements of the light intensity of each fluorescent nucleotide. Finally, a homopolymer is a string of repeated nucleotides in the reference; here, we define a homopolymer to be ≥ 3 consecutive identical nucleotides in the reference.

Rather than include *only* the error-prone sequence features discussed below, we include generalizations of the four types of features to obtain a more comprehensive, detailed model of errors for the conditional HMM. The specific features highlighted by other studies and the generalizations we will include in the conditional HMM are discussed below.

5.3.1 Quality scores

Documentation for Illumina's companion software CASAVA states that the quality scores are not always reliable, and at times not even independent per base [1, p.32] [58, p.2]. Meacham et. al. 2011 performed a multiple-comparison test and concluded that quality scores do *not* account for the elevated error-rates at locations of "systematic error" that they studied [55, p.5]. This contradicts other findings to the contrary [58, p. 11], and rigorously rebukes the assertion that incorporating quality scores sufficiently accounts for errors [58, p.8]. Note that only Meacham et. al. 2011's finding is backed by a statistical test. While Meacham, et. al.'s statistical analysis of quality score-sufficiency was only performed for a particular subset of substitution errors, we conservatively assume that for all biases and error-profiles considered below, the associated quality scores do not sufficiently reflect the uncertainty in the base-call and hence warrant the inclusion of additional uncertainty.

We assume independence in quality scores, and we incorporate the quality score Q_{phred} as parameter for the emissions probability in our conditional HMM.

5.3.2 Contexts

Meacham et. al. 2011 characterized positions with statistically significant elevated error rates as preceded by `G`, and found the triplet-motif `GGT`, with an error at the `T`, to be the majority of these significant locations [55, p.4]. In general, their results also show that such positions are well-characterized as `GG`X, with the error at X and $X \in \{$`A,C,T`$\}$. Abnizova et. al. found the duples `G`X ($X \in \{$`A,C,G,T`$\}$) and TX, $X \in \{A, C, T, G\}$; and the triplets AGT and `GGT` to have a statistical significant elevated error-rates [4, p.10,11]. Nakamura et. al. 2011 reported `GGC` to be a common

motif among error positions [59, p.3].

Nakamura et. al. 2011 reported a phenomenon which we dub "stuttering": "the mismatched base was often similar to a preceding reference base" [59, p.5]. Abnizova et. al. 2012 also reported stuttering, though suggested only in specific instances (e.g. they warn that AAC may be incorrectly read as AAA, but they do not suggest similar problems for AAG or AAT)[4, p.16].

Clearly, a number of contexts, of varying lengths, have been implicated in elevated error-rates for base calls. As triplet contexts are the most general case of the contexts mentioned, we will include the full set of all 64 triplet contexts into emissions error model of the conditional HMM.

5.3.3 Homopolymers

Minoche et al. 2011 found indel error rate to increase with homopolymer length [58, p.11]. This is important because "Illumina sequencing is considered to be robust against homopolymer errors"[58, p.11]. We therefore distinguish between gap openings/extensions in homopolymer versus non-homopolymer regions in the conditional HMM, and include the length and type (the nucleotide repeated) of homopolymer into our transition error model.

5.3.4 Cycle

Systematic errors have also been noted to increase with cycle[59, 58], presumably because of accumulating light "pollution" over time which degrades the quality of later

calls. While quality scores have been noted to decrease with cycle, we include cycle as a separate parameter because it is unknown if increase in errors due to increase in cycle is sufficiently explained by a corresponding decrease in quality scores.

5.4 Multivariate multinomial logistic regression error model

To relate the various features of DNA sequences and the high-throughput sequencing machines listed in 5.3 to read error rates, we employed multivariate multinomial logistic regression (MMLR). Note that we did not test if the regression assumptions actually hold for our problem, and while logistic regression is quite robust, an examination of the consistency of our results with its assumptions would likely be a useful exercise in future work.

d this should certainly be done in later work. Below we sketch the general form for MMLR, and then relate it to our problem.

Consider a multinomial variable Y' which has possible outcomes labeled $y_1, \ldots, y_m$. Let the random column vector $Y = [Y_1 \ldots Y_m]$ be the corresponding dependent variable for the regression, where $Y_i = 1, Y_{j \neq i} = 0$ when $Y' = y_i$. The independent variable is represented by the column vector $\boldsymbol{x} = [1 x_1 \ldots x_n]^T$, where 1 is included to allow for a constant term. Multivariate multinomial logistic regression assumes that the probabilities of the outcomes of the dependent variable Y are related to a linear combination of the independent variable $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$ via the multinomial logit. The coefficients of the linear combination are denoted by $\beta_{i,j}$. Specifically, for each multinomial outcome y_i there is a parameter $\beta_{i,0}$ for the constant term, and

for each pair y_i, x_j, there is an associated parameter $\beta_{i,j}$. The parameters of the regression are then the $n+1 \times m$ matrix B. Column i of B, written β_i, is the vector $[\beta_{i,0} \dots \beta_{i,n}]^T$ of regression parameters for the constant term and $x_1, \dots .x_n$ associated with multinomial outcome y_i.

Typically, multivariate multinomial logistic regression is expressed as

$$\mathbb{P}\big(Y_i = 1 \bigm| B^T \boldsymbol{x}\big) = \frac{\exp \beta_i^T \boldsymbol{x}}{\sum_{j=1}^{m} \exp \beta_j^T \boldsymbol{x}} \tag{5.4.0.1}$$

For our conditional HMM, we actually used four MMLR models: one for emissions, and one for transitioning *from* each possible state (`match, insert, delete`). All MMLR models used the same independent variables $\boldsymbol{x}$, listed in 5.4.1 and parameterized below. For the emissions MMLR model, the dependent variable Y represents a base-call at a certain position in a read, as made by the sequencing machine; i.e. $Y \in \mathcal{N}$. For the transitions MMLR models, Y represents the state just transitioned *to* some state. Note that each transition MMLR model has a different space of possible outcomes for the dependent variable: when coming from `match`, any state may come next; when coming from `insert`, the next state may only be `match` or `insert`; when coming from `delete`, the next state may only be `match` or `delete`.

5.4.1 Parameterization of features for the independent variable

Here we define random variables that represent the general features of DNA sequences and the high-throughput sequencing machine process that are believed to

be associated with sequencing biases and changes in error rates.

For each context $b_1b_2b_3 \in \mathcal{N}^3$ we include the random variable $I_{b_1b_2b_3}(j)$ which is the indicator function and is $= 1$ when $\mathcal{F}_{j-2}\mathcal{F}j-1\mathcal{F}_j = b_1b_2b_3$. As a convention, the rightmost base written in the context is the base which the machine was attempting to read; thus b_3 is the base being read by the machine in context $b_1b_2b_3$. $C = 1, \ldots, m$ is the random variable representing cycle along the read R of length m. $H_\ell(j)$ is the random variable representing the length of the homopolymer at position j of $\mathcal{F}$, where $H_\ell(j) = 0$ if there is not a homopolymer as defined in 5.3 at j. $H_\mathcal{N}(j)$ is the homopolymer type (the repeated nucleotide) at position j in $\mathcal{F}$, where $H_\mathcal{N}(j) = 0$ if there is not a homopolymer as defined in 5.3 at j. Finally, q_i is the quality score at position i of read R.

5.5 Learning multivariate multinomial logistic regression parameters via expectation maximization

Learning the parameters of a multivariate multinomial logistic regression model using a dataset is nothing new. Importantly, classic approaches assume that the data is completely observed. But our problem includes an important deviation: the dataset is *not* completely observed, i.e. the dataset contains hidden variables.

We want to learn the parameters for generating the bases of a read from a given sequence. But from the sequencing machine we are only given reads (known strings of nucleotides); we are *not* told from where in the genome these reads were generated,

nor what was the actual sequence of states of generation. If we restrict to reads from unique sequences (following the same thinking as in 3.3.2), then we can confidently assume that we know the location which generated each read, but still not the state sequence taken during base generation. In terms of the above formulation, the sequence of transition states `match, insert, delete` taken during read generation is hidden.

This is of major importance. It means that our dataset does not include even a single observation! Rather, we only have a set of reads and associated mapping locations, but no direct observation of which reference position generated which read position.

Expectation maximization (EM) is the classic approach for estimating parameters for a model using a dataset in which some of the variables are hidden. Briefly, EM has two steps. In the E-step, we calculate an expression for the expected value of the likelihood of the data (hidden and observed variables), where the expectation is over the hidden variables and uses the current parameter estimates. In the M-step, we find the parameter values which maximize the expression calculated in the E-step. This two step procedure is iterated until the estimated parameters converge. Intuitively, the E-step produces a set of *weighted* observations as a substitute for fully observed observations, where the weight of an observation is the probability associated with the hidden variable's state in the observation.

In our problem, the E-step amounts to calculating the marginal posterior probability of each possible state relationship between every position of the two sequences, i.e. $\forall t \in \mathcal{T}, \forall i = 1..m, \forall j = 1..n : \mathbb{P}\big(A_t(i,j) = 1 \;\big|\; R, \mathcal{F}\big)$. We compute this via the usual sum-forward and sum-backward algorithms, as presented in [22] adapted for our conditional HMM aligner in section 5.2. In the dataset, we have *zero* observa-

tions of a specific read position being generated from a specific reference position. But from the E-step, we now have an observation for every possible combination of read position, reference position, and state relating them.

Chapter Six

NAHR results on real data

We analyzed low coverage Illumina paired-end read data for 44 individuals obtained from the publicly available database of the 1000 Genomes Project, focusing on the detection of NAHR deletions and duplications. We chose the 44 low-coverage individuals with the largest datasets.[1] We analyzed the same 324 potential NAHR deletion/duplication loci on each individual. This subset of all possible NAHR loci was chosen strictly based on computational constraints - following the description in section 3.2, we chose the 324 loci with the smallest computational complexity, as determined by the number of potential NAHR loci that must be simultaneously considered during probability calculations. Note analyzing the same 324 potential NAHR events across 44 individuals gives a total space of $44 \times 324 = 14,256$ possible NAHR event calls. To isolate a reliable set of NAHR deletion/duplication calls, we required a candidate call to have $fdr \leq 0.01$ according to the repeat-sensitive read-depth test of chapter 4.

Our results are summarized in Table 6.1. 1043 of our NAHR event calls passed the *fdr* threshold, which were called at 109 distinct potential NAHR loci when collapsed across the 44 genomes. Of the 1043 calls, 722 were duplications and 321 were deletions. Notice that the total number of distinct loci with a positive NAHR event call in some individual is *not* the sum of the number of such distinct loci for deletions and duplications separately - this is because some loci were called as NAHR deletions in some individuals, while called as NAHR duplications in others. The median number of positive NAHR calls per individual is 24 (7.41% of all 324 tested loci). Comparing against structural variation calls and experimentally validated rearrangements reported in [57, 37, 38], we found that only 106 of our 1043 calls (21 of the 109 distinct loci with a positive NAHR event call) were previously reported, and of the 106 previously reported calls, 59 were positively experimentally validated.

[1] i.e. we ranked the low-coverage individuals by size of their `.bam` file (in GB), and then selected the 44 individuals with the largest `.bam` file size.

Repetitive regions pose difficulties not only for detecting rearrangements, but in constructing the reference genome as well. As such, an immediate concern would be that putative NAHR rearrangements in fact reflect anomalies in the reference genome rather than genuine rearrangements in the individual; that an NAHR "signal" is just an artifact of a poorly constructed region of the reference. In such cases, we would expect that all (or nearly all) of the individuals tested would display such a signal. Further, the erroneous signal displayed by each individual would perhaps be of slightly different magnitude, and a criticism could be that our model merely chooses *some* of the individuals to make a call on according to some arbitrary threshold imposed on a signal that does not actually well-separate the data.

Figures 6.1 and 6.2 address both aspects of this concern of a reference error. Figure 6.1 shows that, in general, a given potential NAHR locus has NAHR event calls in only a subset of the 44 testes individuals. Among the loci in which an NAHR event was called positive in at least one individual, the number of individuals with some NAHR event at a particular locus has median 5 (11.4%); far from all 44. Further, only 7 (6.4%) loci had a positive NAHR event call in 34 (77.3%) of the tested individuals.

Figure 6.2 addresses the concern regarding the separation between positive and negative read-depth signals, as one might argue that these there is a region of high ambiguity between positive and negative read-depth signals, in which case our *fdr* threshold (0.01) is just an arbitrary cutoff that does not delineate any separation in signal. Of the 109 distinct loci with a positive NAHR event call in some individual, we selected the 31 loci for which between 9 and 35 individuals contained a positive NAHR event call of some kind at the given locus and compared their *fdr* values. The criteria that an NAHR event was called at the same locus in 9 to 35 of the individuals was used so that both sets of $\log fdr$ values, for positive calls and for

negative calls, were nontrivial. For each of these 31 loci, we formed a boxplot of the $\log fdr$ values for those individuals with a positively called NAHR event at the locus and a boxplot of the $\log fdr$ values for those individuals with a negative NAHR call. It is clear from Figure 6.2 that these two populations are very well separated among the 31 loci. Notice that, in Figure 6.2, whenever there is an occassional overlap in whiskers, it is always the case that the negative calls' $\log fdr$ extends into the very low fdr region. These are presumably merely a small number of false negatives. To illustrate the signal difference with a specific example, we selected potential NAHR locus chr 19 : $[8336858, 8366563]$. We called a homozygous deletion for Japanese individual NA18973 at this locus, and we called no NAHR deletions or duplications for Yoruban individual NA18523 at this locus. These calls are credible from looking at the read-depth graphs alone. Further, NA18973 has $fdr = 1.2 \times 10^{-3}$ and NA18523 has $fdr > 0.99$.

Since our model evaluates both read-depth and read alignments to make NAHR event calls, it is possible for our model to make a high-confidence NAHR *event* call at a locus due to a strong read-depth signal, and yet have much lower confidence about the location of the breakpoint of that NAHR event. We can identify the subset of our positive NAHR event calls that have strong evidence of a breakpoint by imposing a threshold of 6 on the breakpoint log-odds (section C.1) for each called NAHR event. Of the 1043 positive NAHR event calls across the 44 individuals, 512 calls had breakpoints with log-odds ≥ 6. Below, we analyze in detail the impact of our positive NAHR calls on genes using the 512 NAHR calls with high-confidence breakpoints.

We present a more intuitive summary of breakpoint support in Figures 6.3 to supplement the more rigorous probabilistic calculations that gave the breakpoint log-odds. For each of the 512 NAHR event calls with a high-confidence breakpoint

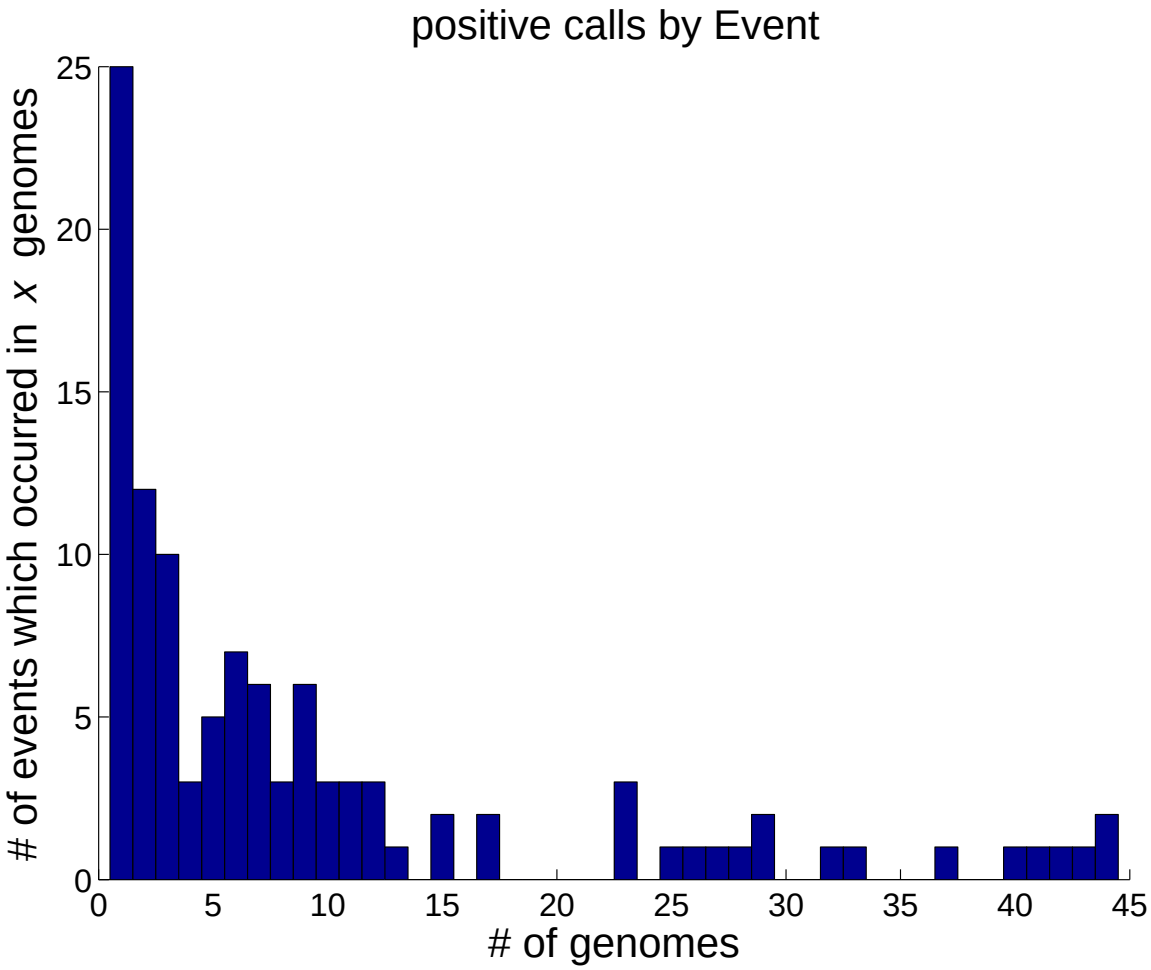


Figure 6.1: Distribution of positive NAHR event calls by number of individuals. Note that out of 109 distinct loci with a positive NAHR event call, only 7 (6.4%) loci were called as positive in $\geq$ 34 (77.3%) of the tested individuals. Among loci with a positive NAHR event call, the median number of individuals with a positive call at any particular locus was 5 (11.4%).

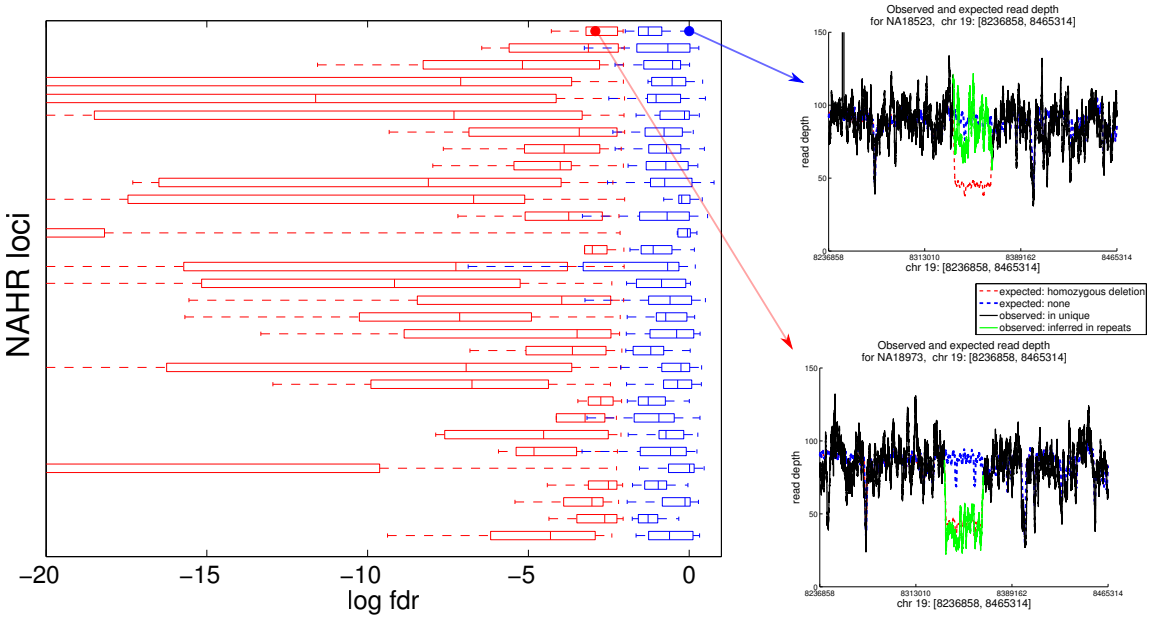


Figure 6.2: Boxplots of the log *fdr* for 31 loci with a positive NAHR event call. Of the 109 distinct loci with a positive NAHR event call in some individual, we selected the 31 loci for which between 9 and 35 (out of 44) individuals had a positive NAHR event call at the locus. For each of the 31 loci, we created a boxplot (red) of log *fdr* for those individuals with a positive event call at the locus, and a separate log *fdr* boxplot (blue) for those individuals with a negative event call at the locus. Within a locus, the positive and negative boxplots are adjacent. Whisker lengths cover up observations within up to 1.5 of the interquartile range. For simplicity, outliers are not shown. For the potential NAHR locus chr 19 : $[8336858, 8366563]$, we plot the read-depth for Japanese individual NA18973 for whom we called a homozygous deletion at this locus, and Yoruban individual NA18523 for whom we did not call an NAHR deletion or duplication at this locus. Read-depth inferred in repeats as in chapter 4 is shown in green, observed read-depth in unique regions is shown in black, expected read-depth given no NAHR deletions or duplications is shown in blue, and the expected read-depth given a homozygous NAHR deletion is shown in red. Read-depth is plotted as 1250 bp-wide moving sums. NA18973 has $fdr = 1.2 \times 10^{-3}$ and NA18523 has $fdr > 0.99$.

(i.e. breakpoint log-odds ≥ 6), we counted the number of paired-end reads which had a greater probability of being generated from the breakpoint region of the novel hybrid LCR formed by the called NAHR event as opposed to anywhere else in the reference genome.[2] These are the very paired-end reads which display the "switch" in variational position patterns, as illustrated in Figure 2.1. Notice that most events had ≥ 5 paired-end reads supporting the called NAHR breakpoint.

To understand the *kinds* of reads that supported our 512 high-confidence NAHR breakpoint calls, Figure 6.4 contains a histogram of the number of *discordant* paired-

[2]see 4.2.2 for how we calculated the probability of each possible generating location of a read.

end reads supporting each such call. Recall that nearly all other structural variation algorithms detect structural variation using *only* discordantly mapped reads. But of the 512 NAHR event calls with a high-confidence breakpoint, 425 (83%) were supported by *zero* discordant paired-end reads, guaranteeing them to be undetectable by other algorithms, *by definition*. Another 10.4% would be extremely unlikely to be detected by other algorithms since so few ($\leq$ 2)) discordant reads support them. Thus, most of the support for our high-confidence NAHR breakpoints comes from paired-end reads that were mapped with phantom concordance, i.e. mapped concordantly to a highly homologous region of an LCR from which they were *not* actually generated.

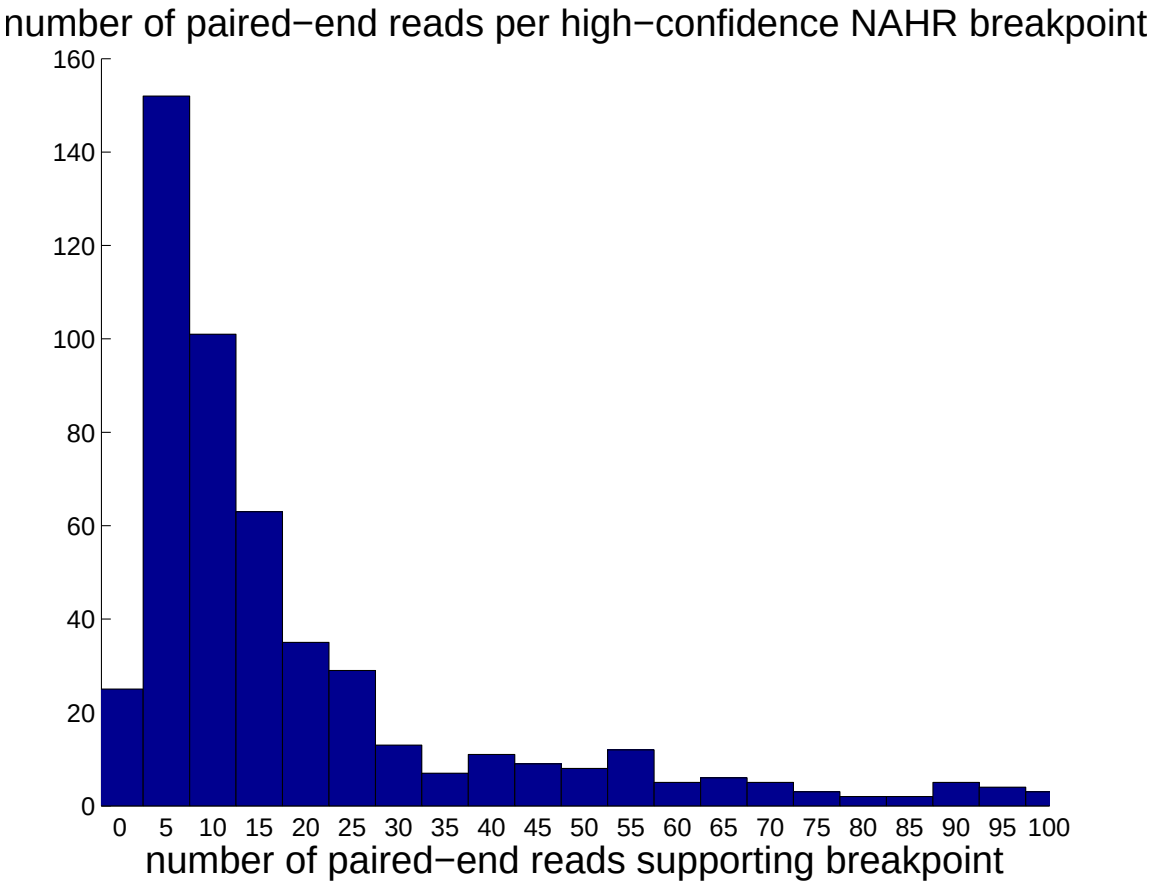


Figure 6.3: Histogram of the number of paired-end reads supporting high-confidence NAHR breakpoint calls. In total, 512 NAHR event calls had a high-confidence breakpoint. 12 outliers with $\geq$ 103 paired-end reads not shown.

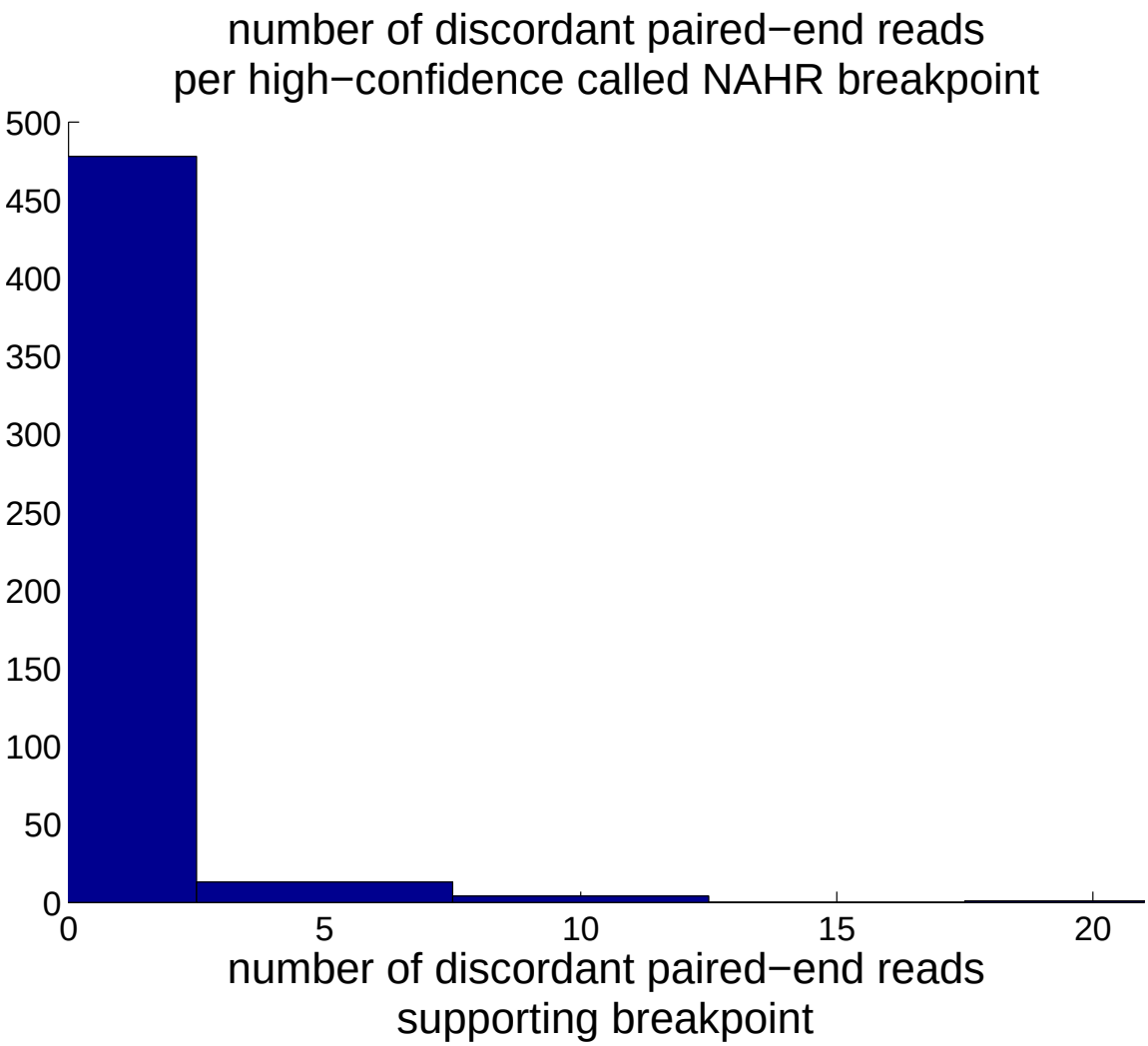


Figure 6.4: Histogram of the number of discordant paired-end reads supporting high-confidence NAHR breakpoint calls. In particular, 425 (83%) of the 512 NAHR event calls with a high-confidence breakpoint were supported by *zero* discordant paired-end reads, and instead supported only by reads originally mapped with phantom concordance. Further, 478 (93.4%) were supported by ≤ 2 discordant paired-end reads. Since most other structural variation detection algorithms ignore phantom concordant reads and *only* use discordant reads to find structural variation breakpoints, then 425 (83%) of our high-confidence NAHR breakpoint calls are guaranteed to go undetected by other algorithms *by definition*, and another 10.4% would be extremely unlikely to be detected by other algorirthms since so few (≤ 2)) discordant reads support them.

6.1 Relations to ancestry

To explore the relationship of our detected NAHR events to ancestry, we tested the hypothesis that the occurrence of NAHR events was independent of ancestry. Of our 1043 calls passing the *fdr* threshold, 431 calls (41.3%) were in individuals of

African ancestry, 309 (29.6%) in individuals of Asian ancestry, and 303 (29.1%) in individuals of European ancestry. These numbers are reasonable, as the reference genome is European. Testing the relationship between ancestry (African, Asian, European) and NAHR loci (109 distinct loci with a positive NAHR event call), we find that ancestry and NAHR events are overall *not* independent ($\chi^2 = 302.8$, $df = 216$, p-value $= 8.8 \times 10^{-5}$). That is, the occurrence of NAHR deletions and duplications across the genome is *not* independent of ancestry.

Because of the limited sample size (44 individuals), we were unable to identify specific ancestry-related events that are statistically significant after correction for multiple comparisons.

6.2 Impact on genes

We searched a database of Ensembl genes on the reference genome and found that 216 genes were affected[3] by the 109 distinct loci with some NAHR event call passing the *fdr* threshold. The median number of affected genes per individual was 52. Checking against the COSMIC database of mutations found in various cancers, we found that 53 such sites were affected by our 109 distinct loci with a positive NAHR event call, with a median number of cancer-related genes listed in COSMIC affected per individual of 6. Both databases were obtained from BioMart [34]. The affected genes included several highly studied genes, such as those involved in hemoglobin (HBA1, HBA2, HBMA, HBZ), haptoglobin (HP, HPR), and drug metabolism (CYP2E1).

A more detailed analysis of the impact of an NAHR event on a gene depends on

[3]we considered a gene to be affected by a specific NAHR event if the gene intersects either of the mediating LCRs or the sequence inbetween, i.e. the gene is inside the potential NAHR locus at which the NAHR event was called.

the relative locations of the gene, the mediating LCRs, the NAHR breakpoints, and any pseudogenes. The simplest case is when a gene is contained in the region between the two mediating LCRs, but not intersecting either LCR; then an NAHR deletion or duplication will *completely* delete or duplicate the gene, resp.. If a gene is contained within one of the mediating LCRs then the exact locations of the breakpoints are very important. If the gene lies within the breakpoints, then it will be completely deleted or duplicated, as before. If the gene lies outside of the breakpoints, then it will be physically unaffected, although the distance to its promoter or regulatory elements may change. If the breakpoint intersects the gene, then a new fusion gene will arise. The composition of this fusion gene will depend on what was lying at the homologous position on the other LCR - another gene or a pseudogene. Finally, sometimes a gene actually *contains* a pair of LCRs (as does the haptoglobin gene *HP*, for example) - in which case, an NAHR deletion or duplication will cause a "contraction" or "expansion" of the gene, resp.. Clearly the exact location of the breakpoint within the gene will have major implications for transcription. Several instances of these scenarios are highlighted below in Section 6.3 and Table 6.3.

NAHR is an important evolutionary mechanism for the creation of pseudogenes and the creation of novel genes via fusion or contraction/expansion. When we are highly confident in a called NAHR breakpoint, we can examine the precise impact of the NAHR event on genes and pseudogenes. For each of our 512 NAHR event calls with a high-confidence breakpoint (we searched a database obtained from BioMart[34] of all Ensembl genes and pseudogenes to determine the impact of the called NAHR event and breakpoint. Table 6.2 contains the results. In particular, notice that 381 genes and 12 pseudogenes genes were duplicated completely, and 19 novel genes were formed via fusion.

6.3 Case study

We now demonstrate the various facets of our model by investigating a single NAHR event call in detail; a two-copy duplication of a 20.5 kb on chromosome 1 with breakpoints $155,184,704$ and $155,205,331$ on Yoruban individual NA19129. This rearrangement is novel; it was not reported in any of the previous validation studies [57, 37, 38]. The called breakpoints are deep inside the mediating LCRs: 4531 bp and 4564 bp inside of LCRs of lengths 10583 and 12491, respectively. This highlights the crucial role that variational positions play in detecting breakpoints of NAHR events inside repeats, as we describe in detail below. Indeed, the called breakpoints are nearly right in the middle of the mediating LCRs, far away from any flanking unique regions which could have been used to "anchor" a mates of any overlapping paired-end reads, as some algorithms attempt to do.

Note that we detected an NAHR event at this locus in exactly one other individual: NA19190, also Yoruban. The call for NA19190 was identical to that for NA19129: also a two-copy duplication, and with the same breakpoints.

6.3.1 Hybrid reads

For illustrative purposes, we collected all paired-end reads for NA19129 which displayed the switch in variational positions as implied by the called NAHR two-copy duplications breakpoints. Figure 6.5 shows a multiple alignment of these paired-end reads against each of the LCRs A and B that mediated the NAHR duplications, and against the hybrid LCR BA resulting from these duplications. There are two

variational positions v_1 and v_2 of interest at this locus of the mediating LCRs[4]. Our model called a two-copy NAHR duplication between LCRs A and B, with both duplications having a breakpoint in the region $[v_1+1, v_2]$.[5] When the reads are aligned against LCR A, then all of the reads agree with the reference at v_2, but disagree with the reference at v_1, and they all display *the same incorrect base* (`G` instead of `A`). On the other hand, when the reads are aligned against LCR B, the situations is reversed: they now agree with the reference at v_1, but disagree with the reference at v_2 and, again, they all display *the same incorrect base* (`C` instead of `T`). Finally, when the reads are aligned to hybrid BA that would hypothetically result from an NAHR duplication with breakpoints in the region $[v_1+1, v_2]$, then *all* of the reads agree with the reference at *both* v_1 and v_2. This is very strong evidence in favor of the hybrid over either mediating LCR; indeed, the log-odds ratio (Section C.1) of the reads aligning to the mediating LCRs vs. the hybrid LCR is -12.3.

Unaware of the "rules of NAHR", a naïve approach may dismiss the disagreements at v_1 and v_2 as SNPs or read-errors, or may ignore the reads for containing too little information. For example, of the 8 paired-end reads shown spanning the breakpoint, *all* of them were mapped concordantly by BWA[6]. Many structural variation algorithms *only* use *discordant* paired-end reads to find structural variations - including VariationHunter, VariationHunter-CR, HYDRA, GASV, GASV-pro, CNVer, BreakDancer, PEMer, and Pindel[7] [30, 31, 65, 70, 71, 56, 82, 14, 56] - and

[4] In reference genome coordinates, v_1 represents chr1 : 155184576 and chr1 : 155205203, and v_2 represents chr1 : 155184704 and chr1 : 155205331. Each is a pair of positions because the mediating LCRs are homologous, and mismatches in the parwise alignment of the mediating LCRs are considered variational positions.

[5] This is theoretically the smallest possible region in which a breakpoint can be called for such an NAHR event. Since there is no other variational position between v_1 and v_2, then the sequence spanning $[v_1+1, v_2-1]$ on LCR A is identical to the corresponding sequence on LCR B. Hence, all NAHR duplications whose breakpoint lies somewhere in $[v_1+1, v_2]$ will have identical resulting hybrid LCRs.

[6] and designated as "properly paired".

[7] Pindel actually requires that one mate be mapped uniquely and the other mate to be unmapped. This is idea is sufficiently similar to discordantly mapped reads for inclusion here.

would thus ignore these reads. Further, 7 of the 8 paired-end reads have a mate with mapping quality 0. Mates that are given mapping quality 0 are considered to contain too little information to be confidently mapped to a unique location, and are ignored by many structural variation algorithms, including Breakdancer, MAQ, and PEMer [14, 48, 40]. But when we precisely construct the hypothetical hybrid LCR that results from an NAHR duplication at this locus according to the "rules of NAHR", we see that in fact *both* mates of all of these paired-end reads contribute a significant amount of power: 7 out of 8 of the reads have a posterior probability > 0.96 of mapping to the hybrid LCR as opposed to either mediating LCR. Thus, although the difference in signal between the hybrid LCR and the mediating LCRs is slight (literally a few SNPs), there is much discriminative power to be gained from the low read-error rate ($\approx 2\%$) and from multiple reads displaying the signal.

While the paired-end reads in Figure 6.5 strongly support the called breakpoint, there are substantially fewer reads than one might expect given the coverage of 30.8x. In unique regions, virtually every single fragment overlapping the breakpoint of a rearrangement is informative of the rearrangement, and since we called a two-copy NAHR duplication event with the same breakpoint, we would naïvely expect approximately 30.8 reads that display the breakpoint. But in the NAHR setting, we can only localize the breakpoint to the region between two variational positions of a pair of repeats. We would then expect 15.3 paired-end reads to span $[v_1, v_2]$.[8]. Additionally adjusting for GC-bias in fragment distribution, we would expect 13.8 paired-end reads spanning $[v_1, v_2]$.[8] This shows the dramatic degree to which breakpoint-finding in repetitive regions differs from unique regions. Further, we note that although a paired-end read may span the breakpoint region, it may not necessarily be *informative* of the breakpoint; that is, it may not capture, within the mate reads, the

[8] calculation in section C.6

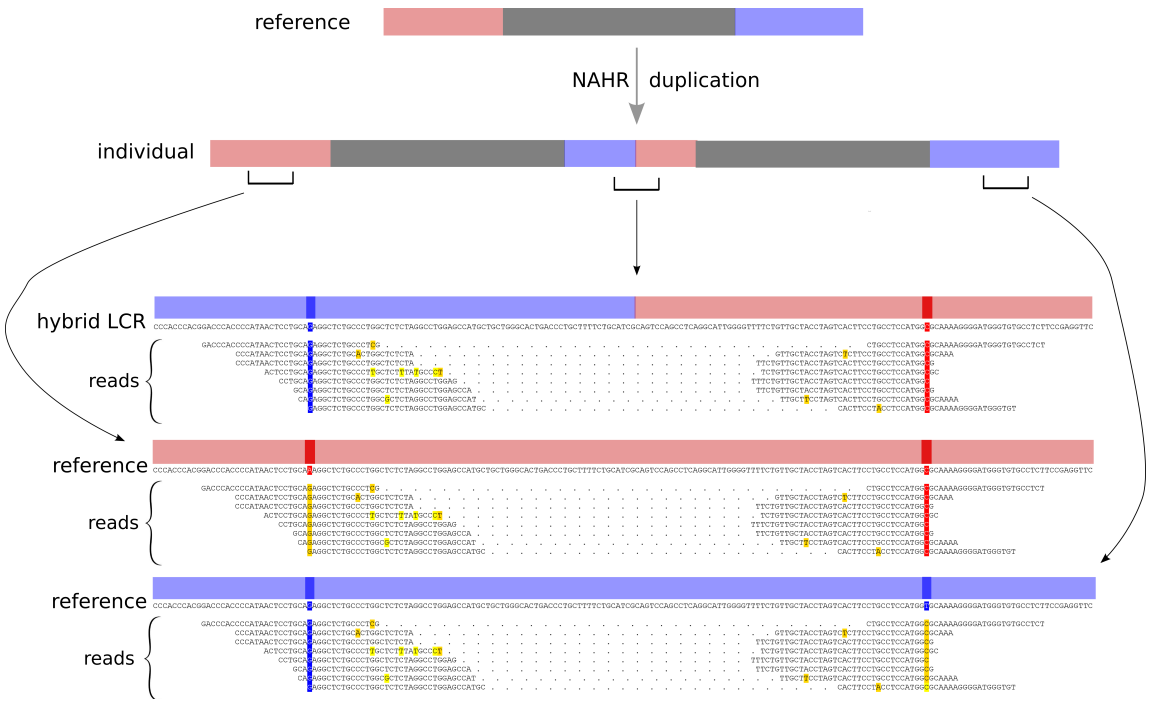


Figure 6.5: Multiple alignments of the paired-end reads which display the expected switch in variational position patterns at the breakpoint of a duplication on chromosome 1 with breakpoints $155,184,704$ and $155,205,331$ on Yoruban individual NA19129. The called NAHR event was mediated by LCR A (red) with coordinates $[155180173, 155190755]$ and LCR B (blue) with coordinates $[155200767, 155213257]$. The same set of paired-end reads are aligned in each case. At the top is a schematic representation of the reference genome (not to scale) in the region chr $1 : [155180173, 155213257]$. The homologous LCRs are red and blue. These LCRs mediate an NAHR duplication to result in the individual's genome (NA19129). We then collected all paired-end reads which display the expected switch in variational positions. This same collection of paired-end reads are shown in three different multiple alignments: against the hybrid LCR constructed from the reference according to the breakpoints called, against LCR A, and against LCR B. Variational positions are colored according to which LCR's variational position pattern they agree with - hence, all variational positions on LCR A are red, all variational positions on LCR B are blue, and the part of the hybrid LCR before the breakpoint came from LCR B has blue variational positions, while the part of the hybrid LCR coming from LCR A has red variational positions. and hence its variational positions are blue. Positions in the reads that disagree with the reference/hybrid LCR at variational positions are colored yellow. Positions in the reads that agree with the reference/hybrid LCR at variational positions are colored the same as in the reference/hybrid LCR. Dots represent the unsequenced part of the fragment inbetween the two mates of a paired-end read. When aligning the reads to LCR A, notice that the reads perfectly agree at the variational positions on the right-hand side of the alignment, but completely disagree with the variational positons of the left-hand side of the alignment. But when aligning the reads to LCR B, the situation is reversed: the reads now completely disagree with the right-hand side, but perfectly agree on the left-hand side. Finally, when aligning to the hybrid LCR, all disagreements between the reads and the reference are now resolved. The log-odds ratio of the probability of the reads given that there was no NAHR event (i.e. the hybrid does not exist) vs. the probability of the reads given that the two-copy duplication indeed occurred (i.e. the hybrid LCR does exist) is -13; strong support for the two-copy duplication and the specific hybrid LCR.

variational positions on either side of the breakpoint region that indicate a switch in the variational postions pattern between the two mediating LCRs. If we imposed the requirement that a hypothetical paired-end read also be informative of the switch in variational positions, then we would expect even fewer than 13.8 reads. Indeed, we observed 8 informative paired-end reads at this breakpoint.

For contrast, we considered the paired-end reads in the vicinity of the same called breakpoint, but in European individual NA07051 and Yoruban individual NA18501, for whom we did not call any NAHR events at this locus. To summarize the above, in NA19129, we found 8 paired-end reads informative of the called breakpoint which displayed a "switch" in variational position; that is, 8 reads simultaneously correctly matched variational positions from *both* mediating LCRs. In NA07051, there were 8 paired-end reads potentially informative of a breakpoint at the same location as called in NA19129. However, zero of them correctly matched variational positions from *both* LCRs. Instead, 5 paired-end reads correctly matched variational positions from the first mediating LCR but not the second, while 3 paired-end reads correctly matched the second mediating LCR but not the first. Similarly, NA18501 contained 6 reads informative of the breakpoint called in NA19129. Again, zero paired-end reads showed a switch in variational position; 2 correctly matched the first LCR's variational positions but not the second LCR's, and vice-verse for the other 4. The corresponding multiple alignments for NA07051 and NA18501 of all paired-end reads in the region of the breakpoint called in NA19129 are in section C.4.

6.3.2 Read depth

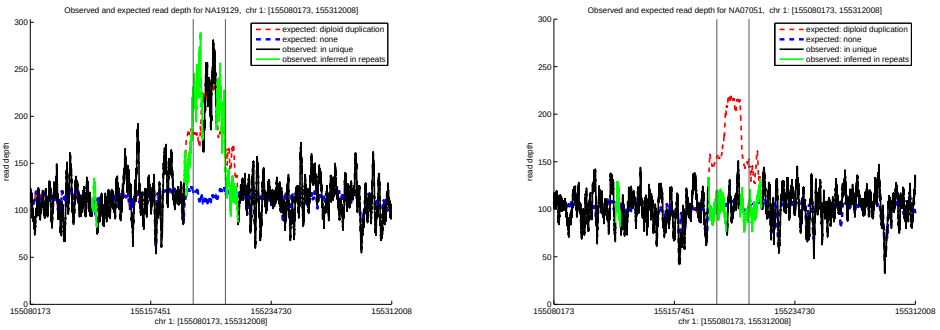
Following chapter 4, we infer the read-depth in the repetitive regions and plot its signal in Figure 6.6a alongside the unique read-depth signal. Note that the signal

in repetitive regions "transitions" from following the expected signal for no event to the expected signal under the called two-copy duplication, as we expect. Further, the "transition" in read-depth occurs approximately right at the called breakpoints - this shows the significant amount of information contained even in reads coming from highly repetitive regions. Including inferred read-depth from repeat regions together with the unique region, we calculate the *fdr* to be 1.3×10^{-3}. This again indicates that a two-copy duplication is much more likely than a null event at this locus. Indeed, we were able to detect NAHR events involving only repetitive regions (e.g. NAHR between tandem LCRs), as shown in Table 6.3.

For perspective, Figure 6.6b shows the unique and inferred read-depth signal at the same locus for European individual NA07051. Our model determined that this individual did *not* have an NAHR event at this locus. The *fdr* for this locus in NA07051 individuals is > 0.99. It is already obvious from the read-depth signal graph alone that, indeed, individual NA07051 did not experience an NAHR event at this locus, and the *fdr* reinforces this conclusion. This also serves as another a particular example of the strong difference in signal between positively called NAHR events and negatively called ones.

6.3.3 Relations to disease

The two-copy NAHR duplication studied here affects the genes GBA and $MTX1$ and their respective pseudogenes $GBAP1$ and $MTX1P1$. Mutations in GBA cause Gaucher's Disease and are strongly associated with Parkinson's Disease in populations worldwide [69, 26]. Mutations in $MTX1$ have also been linked to Parkinson's disease [26]. Somatic mutations in GBA have also been linked to lung cancer, and somatic mutations in $GBAP$ and $MTX1$ have also been linked to endometrium



(a) **(b)**

Figure 6.6: Observed and expected read-depth signals for a two-copy duplication on chromosome 1 with breakpoints 155, 184, 704 and 155, 205, 331 for two individuals. The mediating LCRs have coordinates [155180173, 155190755] and [155200767, 155213257]. For perspective, the observed and expected read-depths are shown for additional 100kb on each side of the breakpoints. The blue dotted line is the expected read-depth signal if there was no NAHR event at this locus. The red dotted line is the expected read-depth signal given the rearrangement we called at this locus (two-copy duplication). The solid green line is the inferred observed read-depth signal in repetitive regions, as described in chapter 4. The solid black line is the observed read-depth signal in the unique regions of this locus. Vertical thin black lines mark the called breakpoints of the two-copy duplication. Read-depth curves are calculated as sliding 1250 bp window sums for presentation. The expected read-depth signals are highly non-uniform due to the GC-bias in fragment distribution, mentioned in Section 3.4.2. **(a)** Read-depth in unique and repeat regions for NA19129. Notice that the observed read-depth signal for the unique sequence inbetween the mediating LCRs follows the expected read-depth signal of a two-copy NAHR duplication (red dotted line) much closer than the expected read-depth signal if there was no NAHR event (blue dotted line). We also see the inferred observed read-depth signal "transition" from closely following the expected no-event signal (blue) to the expected duplication signal (red) and back again at approximately the location of the breakpoints. Together, the observed read-depth signal in the unique regions and the inferred observed read-depth signal in the repetitive regions give an $fdr = 1.3 \times 10^{-3}$; strong support for the proposed two-copy NAHR duplication. **(b)** Unique read-depth and inferred repeat read-depth is shown at the same locus on European individual NA07051. Our model determined there was *not* an NAHR event for this inidivual at this locus. The *fdr* at this locus for NA07051 is > 0.99.

cancer [25].

The gene context of our two-copy NAHR duplication is shown in Figure 6.7. According to our called breakpoints, two identical fusion GBA genes are created from the called two-copy NAHR duplication. The fusion genes consist of the first 1.1 kbp of GBA followed by the last 12.5 kbp of pseudogene $GBAP1$. The breakpoint occurs 1093 bp inside of GBA, and is 901 bp inside of the coding region of GBA. Two additional complete copies of the pseudogene $MTX1P1$ are also formed.

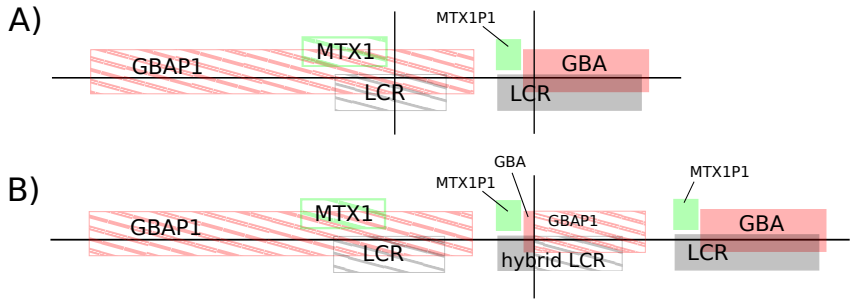


Figure 6.7: Effect of the called NAHR duplication in NA19129 on the *GBA* gene. **A)** Gene context of the locus of the studied NAHR two-copy duplication as it appears in the reference genome (i.e. before the called NAHR two-copy duplication). The mediating LCRs are grey. The first LCR is a 10.6 kbp segment of the *GBAP*1 pseudogene and contains the latter 3.5 kbp of *MTX*1. The second LCR consists of pseudogene *MTX*1*P*1 and all but the last 1.4 kbp of *GBA*. The breakpoints of the called NAHR duplication are vertical black lines. **B)** The same locus, after the called NAHR duplication. The hybrid LCR formed from the duplication contains an extra copy of peudogene *MTX*1*P*1 and a novel fusion gene, consisting of the first 1.1 kbp of *GBA* followed by the last 12.5 kbp of pseudogene *GBAP*1. The breakpoint region (containing the switch in variational positions) is marked by the vertical black line.

6.4 Other important examples of NAHR

To highlight the biological impact of NAHR, we briefly present four more positive NAHR event calls and their impact on several highly studied genes, including RNASE2, RNASE3, FLG, CYP2E1, SPRN, SYCE1, HP, HPR, and TXNL4B. Table 6.3 contains basic information for each called NAHR event, as well as figures similar to those above, demonstrating the hybrid read alignments, read-depth signal, and genome context. Table 6.3 also describes the genes affected by each call and their functions or associated genomic disorders. All calls presented in Table 6.3 are novel.

6.5 Discussion

We have developed a Bayesian probabilistic model for detecting non-allelic homologous recombination using high-throughput sequencing data. To our knowledge, our model is the first to utilize the specific features of the molecular mechanisms involved

in NAHR. We also modelled the generation of high-throughput sequencing data, including biases in fragment distribution and error-rates during base generation as presented in recent literature.

To obtain a set of highly reliable NAHR event calls, we applied a read-depth-based *fdr* analysis to our initial calls and retained only those with $fdr \leq 0.01$. The result is a set of 1043 highly reliable calls across the 44 tested genomes, composed of 321 deletions and 722 duplications. Collapsing these 1043 across individuals, we arrive at a set of 109 distinct NAHR loci with a positive NAHR event call. We selected one particular two-copy NAHR duplication for in-depth discussion and illustration above.

6.6 Impact on the understanding of NAHR

6.6.1 No "hotspots"

Since 109 out of 324 distinct NAHR loci experienced an NAHR deletion or duplication event in some individual, then approximately 33.6% of all tested potential NAHR loci were found to be "active". Further, among the 109 loci with a positive NAHR event call, the median number of individuals with an NAHR event call at any locus is 5(11.4%) (see Table 6.1). This suggests that NAHR activity is fairly common and dispersed across the human genome - it is not concentrated in a small fraction of "hotspots" out of all potential NAHR loci. The distribution of positive NAHR calls across individuals and loci can be seen in Figure 6.1 and Figure C.1.

6.6.2 Frequency of NAHR

Our conservative discoveries suggest that NAHR occurs at a much higher frequency than some contemporary estimates in the literature. Studying a pair of related genomic disorders that can arise from NAHR, Liu, et. al. 2011 found the rate of NAHR deletions and duplications during male meiosis to be $\approx 10^{-5}$ to 10^{-7} [51]. Turner, et. al. 2008 developed sperm-based assays to measure the *de novo* rate of NAHR deletions and duplications at four NAHR "hotspots" in the human genome, also finding the meiotic rates to be $\approx 10^{-5}$ to 10^{-7} [79]. If we assume that the *de novo* rate of NAHR is 10^{-5} genome-wide, then for the 324 potential deletion/duplication NAHR loci we analyzed, we expect $10^{-5} \cdot 324 = 0.00324$ total *de novo* deletions and duplications *per* generation. If the difference between an individual and the reference could be quantified as a number of generations, then we would expect 0.324 (0.1%) NAHR deletions and duplications to be present in an individual who is separated from the reference by 100 generations. Our analysis found that the median number of NAHR deletion/duplication calls passing the *fdr* threshold is 24 (7.41%) per individual; much higher than previously thought.

While our observations are much higher than the rates from other studies would suggest, we are not alarmed by the discrepancy. Liu et. al. 2011 were concerned with a particular region of chromosome 17 in which structural variation can cause serious genetic disorders. Such a sensitive region may not be representative of NAHR rates genome-wide; indeed, it may be more highly conserved due to the demonstrated severe consequences of mutation. Similarly, Turner, et. al. 2008 studied only 4 NAHR "hotspots", and all of them were associated with severe genetic disorders. Thus, the current experimentally estimated rates of NAHR are derived quite a small sample of regions wherein NAHR causes severe genetic disorders, and thus not necessarily

representative of the frequency of NAHR in general.

6.6.3 Features correlated with occurrence of NAHR

A number of genomic architectural features have been hypothesized to play a role in the occurrence of NAHR in the human genome, including LCR length, distance between LCRs, percent identity of LCRs, distance to telomere or centromere, length of MEPS, distance of breakpoint to MEPS, and several breakpoint motifs. Some studies have empirically calculated the correlation coefficients between the rate of NAHR and some of these features. We tested each of these features to see if any of them were over- or under-represented in our reliable call set compared to the space of potential NAHR loci we examined.

For each feature of interest, we computed its CDF for the 324 potential NAHR duplication/deletion loci we examined. We then computed the empirical CDF of the feature for the subset of the 109 distinct NAHR loci in our reliable call set, and tested the distributions for equality using a χ^2 goodness-of-fit test. Table 6.4 contains the results. Details of the various calculations can be found in section C.3.

Length of LCRs Several studies have hypothesized a relationship between the length of the mediating LCRs and the rate of NAHR [52, 12, 73, 60]. Liu et. al. 2011 [51] found a positive correlation between NAHR rate and LCR length. We did *not* find the length of mediating LCRs to be statistically significantly associated with the occurrence of NAHR (p-value = 0.65).

Distance between LCRs The distance between LCRs has also been thought to play a role in the rate of NAHR [52, 12, 73, 60]. Liu et. al. 2011 [51] found a negative correlation between NAHR rate and inter-LCR distance, as hypothesized. We found that the distance between the mediating LCRs was statistically significantly shorter than expected (p-value $= 3.9 \times 10^{-9}$, mean background inter-LCR distance $= 52$ kbp, mean observed inter-LCR distance $= 17$ kbp). This result is additionally believable since our model has stronger power to detect NAHR events that affect longer stretches of the genome (Section 6.7).

LCR length over inter-LCR distance Liu et. al. 2011 [51] found the strongest correlation to be between rate of NAHR and the ratio of LCR length over inter-LCR distance. Our results agree: we found the ratio to be statistically significantly higher among occurring NAHR loci compared to background (p-value $= 7.75 \times 10^{-13}$).

Sequence identity The overall degree of sequence identity between mediating LCRs has also been thought to play a role as well [73]. We found that sequence identity is statistically significantly higher among occurring NAHR events compared to background (p-value $= 0.0108$).

Proximity to centromere/telomere The subtelomeric and pericentromeric regions of chromosome have been noted to be hotspots of recombination and enriched with segmental duplications [50, 68, 60]. We tested whether NAHR occurred in these regions at a rate higher than would be expected due to the layout of LCRs across the genome. We found that NAHR is statistically significantly overrepresented near the telomere (p-value $= 0.041$), but does not deviate significantly from the background for proximity to the centromere.

6.7 Limitations of the model

1. Due to computational complexity, we restricted our analysis to a subset (324) of potential NAHR loci among all 1769 potential NAHR loci involving ≤ 250 kb implied by the Human Segmentla Duplication Database. More informed conclusions about the features of the NAHR mechanism (Section 6.6) could be drawn if we could feasibly analyze a larger set of potential NAHR loci.

2. Naturally our model has stronger power to detect NAHR events that affect longer stretches of the genome simply due to the larger amount of data available in such cases.

3. The *fdr* test we applied is more conservative for deletions relative to duplications: observed-to-expected read-depth ratios are necessarily bounded below by 0, but unbounded above - yet our empirical null distributions were nearly symmetric.

4. For computational feasability, we assumed a very simple breakpoint model, where NAHR events have a single switch in variational position patterns. Experimental studies show that sometimes the breakpoint region is more complex, with multiple switches in breakpoint patterns [38, 79]. Thus, when we call a breakpoint, it may be the case that we have found "one of" the switches, or "one of" the breakpoints.

5. As described in Section 3.2, the gene conversion mechanism is almost identical to the NAHR (crossover) mechanism, and results in hybrid LCRs with breakpoints that mimic those resulting from NAHR. Thus, when we encounter paired-end reads which strongly support a particular breakpoint, it is important to determine if it is an NAHR breakpoint or a gene conversion breakpoint.

This decision is further complicated by the fact that, as has been experimentally observed and validated in [38, 79] and mentioned in [52], the breakpoint regions of NAHR and gene conversion events may "switch" between the two mediating LCRs before finally crossing over (NAHR) or not crossing over (gene conversion). Because gene conversion inherently involves pairs of breakpoints, it imposes a severe computational burden on our model. Our inference of gene conversio nwas therefore limited because we considered (for computational concerns) only a small number of possible breakpoints for each gene conversion event. As a result, our power to distinguish NAHR from gene conversion was hindered and complicated.Indeed, identifying hybrid breakpoint signals as the product of gene conversion or NAHR is especially important for inversions. While we did model inversions, we do not report any inversion results here due to large numbers false-positives; presumably gene conversion events mistakenly called as NAHR inversions. Restricting to a set of final calls according to an *fdr* threshold therefore served a secondary purpose: filtering out false-positive NAHR deletion and duplication calls due to gene conversion complications. In future work, we plan to implement a more sophisticated model for the breakpoint region to more clearly distinguish gene conversion events from NAHR events.

Table 6.1: summary statistics

	# loci with event call [a]	# distinct loci with an NAHR event call [b]	median calls / person [b]	median # people / locus with call	# previously reported [c]	# positively validated [d]	median affected genes / person	# in Africans [c]	# in Asians [c]	# in Europeans [c]
deletion	321 (2.25%)	65 (20.1%)	5.5 (1.7%)	4 (9.09%)	106(33.0%)	59(18.4%)	7.5	120 (37.4%)	123 (38.3%)	78 (24.3%)
duplication	722 (5.06%)	64 (19.8%)	16 (4.94%)	3 (6.82%)	0	0	42.5	311 (43.1%)	186 (25.8%)	225 (31.2%)
total	1043 (7.32%)	109 (33.6%)	24 (7.41%)	5 (11.4%)	106(10.2%)	59(5.7%)	52	431 (41.3%)	309 (29.6%)	303 (29.1%)

[a] out of all $44 \times 324 = 14,256$ potential NAHR event loci
[b] out of all possible 324 distinct loci
[c] out of number of positive event calls
[d] subset of # previously reported. Non-positively validated calls were *not* negatively validated.

Table 6.2: Breakpoint impact on genes

fusions: gene-gene	16
fusions: gene-pseudogene	3
fusions: pseudogene-pseudogene	1
deleted genes	39
deleted pseudogenes	14
duplicated genes	381
duplicated pseudogenes	12
expanded genes	251
expanded pseudogenes	46
contracted genes	52
contracted pseudogenes	3
unaffected genes on LCRs	403
unaffected pseudogenes on LCRs	10

Of our 1043 positive NAHR calls across the 44 individuals, 512 had highly confident breakpoints (log-odds ratio ≥ 6). Checking these 512 breakpoints against databases of genes and pseudogenes, we determined the impact of each breakpoint on various genes and pseudogenes. Note that one NAHR event may affect multiple genes (see Table 6.3 and Figure 6.7 for examples), and so the total number of genes affected may not be equal to the number of called breakpoints passing the log-odds ratio threshold.

Table 6.3: Example novel detections

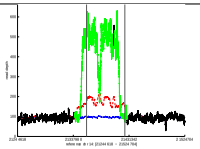
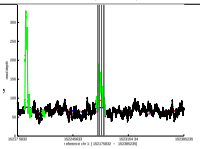
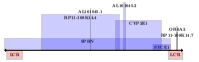
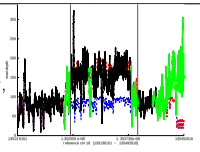
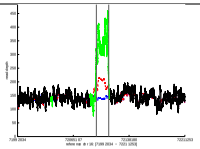
affected genes of interest	molecular functions & associated diseases of selected genes	reference context	resultant genes	call	breakpoint coordinates in reference	read-depth	breakpoint log-odds (sec. C.1)	individual
RNASE2, RNASE3	RNASE3: asthma, immune host defenses [78, 64]		fusion gene: RNASE2→RNASE3	duplication	chr 14: [21360223, 21424308]		13.2	NA18949
FLG	FLG: Ichthyosis vulgaris, atopic dermatitis, asthma, allergic rhinitis, food allergy [32]		expanded FLG	duplication	chr 1: [152276781, 152281644]		28.2	NA19204
CYP2E1, SPRN, SYCE1	CYP2E1: drug metabolism, diabetes, obesity, fasting, alcohol & non-alcohol liver disease [84]. SPRN: Creutzfeldt-Jakob disease [9]. SYCE1: synaptonemal complex of meiosis [19]		duplicated CYP2E1, SPRN, SYCE1	duplication	chr 10: [135249727, 135393855]		5.5	NA18501
HP, HPR, TXNL4B	HP: haptoglobin, diabetic cardiovascular disease [45], Crohn's Disease [61], & others [35]. HPR: protects against Trypanosomiasis [72]. TXNL4B: ribosome assembly [81], pre-mRNA splicing, interacts with Prp6 [76, 33]		duplicated part of HPR, expanded TXNL4B	duplication	chr 16: [72095319, 72111510]		33.9	NA19108

Table 6.4: Features of LCRs and the rate of NAHR

feature	p-value	mean theoretical	mean significant	change w.r.t. theoretical mean
LCR length	0.654053	8872.42	6528.66	-26%
inter-LCR distance	**1.83706e-08**	51896.59	17523.47	-66%
LCR length over inter-LCR distance	**8.75584e-13**	0.52	0.87	+66%
log 1 minus percent identity	0.0165497	-3.30	-3.51	+6.5%
distance from telomere	0.0859108	69248964.17	64900872.21	-6.3%
distance from centromere	0.596767	1721505473.66	1829686573.29	+6.3%

Using the 109 distinct positive NAHR loci, we performed a χ^2 goodness-of-fit test for several features of LCRs that other studies reported to be correlated with rates of NAHR. Only the distributions of inter-LCR distance and ratio of LCR length to inter-LCR distance were found to be statistically significantly different for the 109 positively called NAHR loci compared to the background distribution composed of all 324 potential NAHRloci.

Chapter Seven

Conclusion

Specifically modeling NAHR fills an important gap in contemporary analysis of structural variation, and provides a new, biology-inspired computational approach that is nearly orthogonal to existing algorithms. Studies routinely exclude repetitive regions from analysis due to computational and experimental difficulties, and ignore so-called "concordantly mapped" paired-end reads that contain crucial information of NAHR. As such, our model addresses largely unstudied (from the computational perspective) regions of the genome. Nonetheless, repetitive regions and corresponding NAHR rearrangements play important but still mysterious roles in a range of genomic disorders [12, 83, 66, 74, 13]. In our analysis, a median of 52 Ensembl genes and 6 genes associated with cancer were affected by NAHR deletions or duplications per individual. Over the 44 individuals studied here, 216 distinct Ensembl genes and 53 cancer-associated genes were affected by NAHR events that we called.

Appendix A

Possible NAHR breakpoint restriction

Despite high homology, long LCRs can have a large number of variational positions; for example, 95% similar LCRs of length 10 kb would have 500 variational positions. Appealing to low read-error rates and the large amount of data, we restrict $\mathscr{B}$ in practice to keep the computation feasible. In practice, we first construct every possible NAHR hybrid LCR and align each read to all relevant breakpoint regions. For each breakpoint region, we count how many paired-end reads had a posterior probability ≥ 0.95 of being generated from the breakpoint as opposed to any possible generating location appearing in $\mathcal{F}$, where the posterior probability was calculated as $\mathbb{P}\Big(\ell \;\Big|\; P\Big) = \frac{\mathbb{P}\Big(P \;\Big|\; \ell\Big)}{\sum_{\ell'} \mathbb{P}\Big(P \;\Big|\; \ell'\Big)}$ using the conditional HMM aligner discussed in chapter 5. For each event separately, we then ranked the potential breakpoints by the number of reads which passed the posterior probability threshold, and selected the 4 potential breakpoints to be $\mathscr{B}$. If all breakpoints had zero reads passing the posterior probability threshold, then we selected the 2 sparsest VP to be $\mathscr{B}$, where by "sparseness of a VP" we mean the distance to the nearest VP or end of the LCR. We applied the same procedure to identify $\mathscr{B}^2$, except that we made separate $\mathscr{B}^2$ for each "switch" in VP pattern (first LCR pattern to second LCR pattern, and vice-versa) specific to each outcome e.

Appendix B

Conditional alignment in practice

B.1 Aligning against a generating location, in practice

For practical reasons, we calculated the probability that a read R was generated from location L of genome G using a two-step procedure. First, we took a slightly larger region around L and performed a glocal Viterbi alignment (global for the read R, local for the region of G) using the conditional HMM aligner, and found the endpoints $[p_1, p_2]$ of the Viterbi alignment. Because of the very low error-rates in sequencing, especially with regards to indels, we assumed that the read R was generated from the region delineated by the endpoints $[p_1, p_2]$ of the Viterbi alignment with probability 1. Thus, we assumed that $[p_1, p_2]$ was the fragment from which read R was generated. Next, again using the conditional HMM aligner, we performed the sum-forward algorithm for the read R conditioning on the sequence at $[p_1, p_2]$. The sum-forward algorithm gives the probability of observing the nucleotides of R given that R was generated from the sequence at $[p_1, p_2]$, marginalized over the alignment (i.e. the actual "path" that sequencing took.)

This procedure was performed for each mate R^a, R^b of paired-end read P separately.

Appendix C

More details on NAHR results

C.1 Breakpoint odds

Modelling the mechanics of NAHR allowed our model to precisely construct every possible breakpoint region and align reads against them. We quantify the evidence of a called breakpoint by calculating an odds ratio of alignment probabilities for reads relevant to the breakpoint region. Given a precise breakpoint B, we may denote a small region around B and all paralogous regions as $\mathcal{L}$, and collect all reads mapped to a region in $\mathcal{L}$. The likelihood P_0 of the null hypothesis (there was no NAHR, and so B is not a breakpoint) can then be computed by aligning each read to every location in $\mathcal{L}$. The likelihood P_A of the alternative hypothesis (B is indeed the breakpoint) is calculated by including the the newly-formed hybrid breakpoint region in $\mathcal{L}$, removing from $\mathcal{L}$ the pair of regions which together form the hybrid, and aligning each read to each region in this modified set. The log-odds ratio $\log P_A - \log P_0$ represents how much more likely the existence of a specific breakpoint is compared to the null case (no breakpoint).

C.2 Calls per genome

Figure C.1 below shows the distribution of number of calls per individual.

C.3 Calculation of NAHR features

For each feature below, we calculated the distribution f of the feature using all 324 distinct loci tested, and the observed distribution f of the feature only for those 109

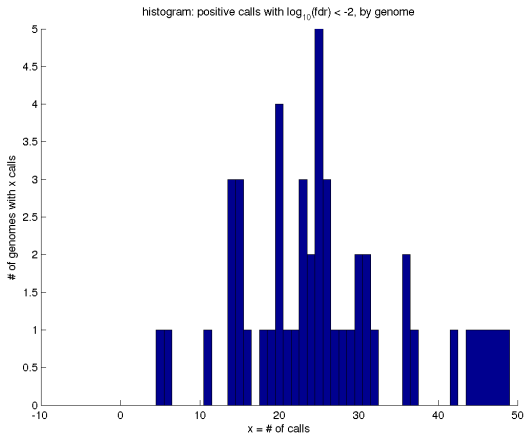


Figure C.1: Histogram of the number of positive NAHR calls appearing in an individual.

distinct loci with an NAHR deletion or duplication call in at least one individual. We then performed a χ^2 goodness-of-fit test to determine if the difference between f and g was statistically significant.

C.3.1 Length of LCRs

NAHR events are mediated by a pair of homologous LCRs, which are likely not the same length due to indels, but close. For a potential NAHR event, we calculate the length of its LCRs as the length of the pairwise alignment of the two mediating LCRs.

C.3.2 Distance between LCRs

Following Liu et. al., 2011, we calculate the inter-LCR distance as the "length of the segment in between LCRs plus the length of one LCR"[51].

C.3.3 LCR length over inter-LCR distance

This follows directly from above.

C.3.4 Sequency identity

For a potential NAHR event, we performed a pairwise alignment of its two mediating LCRs. Every position in the alignment with a mistmach or that is part of a small indel was considered a variational position. The percent divergence was calculated as $t = \frac{\text{number variational positions}}{lengthofalignment}$. We used $\log t$ to measure the association between sequency identity and occurrence of NAHR.

C.3.5 Proximity to centromere/telomere

We calculated the distance to the centromere as the distance from the closest index out of both mediating LCRs to the closest index to the centromere. The distance to the telomere was calculated analogously, using the telomere on the same chromosome arm as the two mediating LCRs.

C.4 Hybrid read alignments for negative examples NA07051 and NA18501

See Figures C.2 and C.3.

C.5 Previously called rearrangements

We checked our NAHR calls against experimentally validated structural variations reported in three previous studies: Mills, et. al. ; Turner et. al. and Kidd et. al.. Importantly, there were several errors and inconsistencies in the validation data reported in Mills et. al., as confirmed via personal correspondence with the author. As such, we imposed a criteria on the calls reported in Mills, et. al. to obtain a "highly reliable" subset of validations. Specifically, we required

C.6 Coverage calculations

Suppose we want to calculate the expected number of reads spanning some interval $[v_1, v_2]$ in the genome. Let $\alpha(x)$ denote the expected number of reads whose left endpoint is at position x in the genome, i.e. $\alpha(x)$ is the per-position fragmentation rate for position x. Let m be the length of a hypothetical read fragment (e.g. m is the median fragment length of a paired-end read library).

Then the expected number of reads spanning the genome interval $[v_1, v_2]$ is calculated as $\sum_{d=0}^{m-1} \alpha_{v_1-d} \cdot \mathbf{1}_{v_1-d+m-1 \geq v_2}$.

Using only the coverage C of an individual's dataset, we define $\alpha_x := \frac{C}{L}$, where $L =$ length of haploid reference genome. To account for the GC-content fragmentation bias, we would instead define $\alpha_x := \lambda(x)$, where $\lambda(x)$ is calculated as in section 3.4.2.

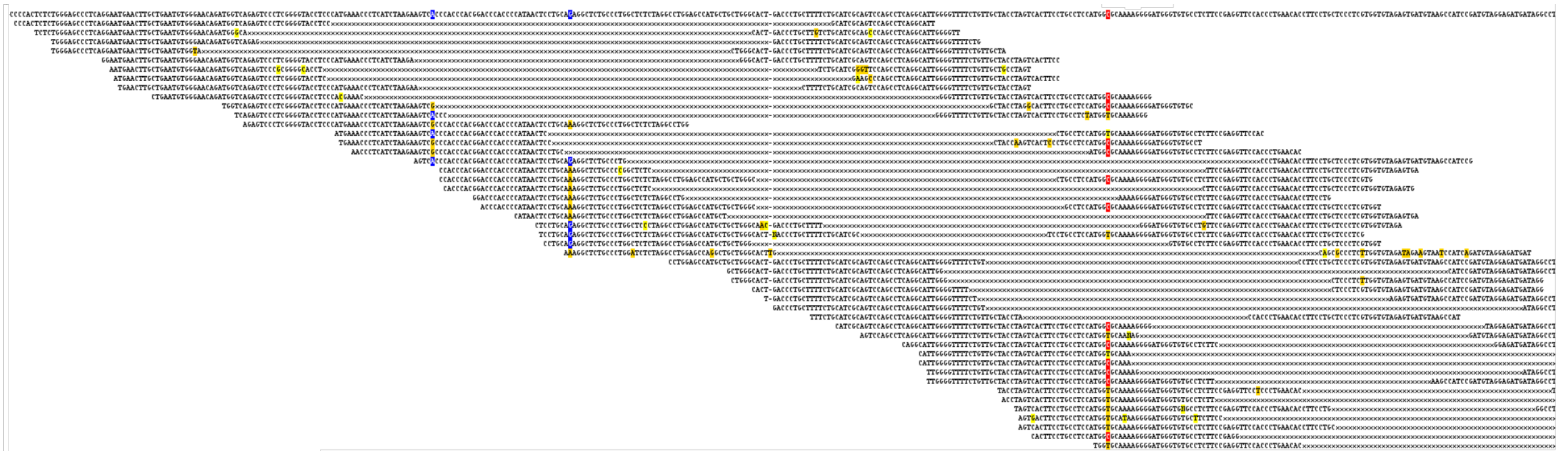


Figure C.2: Paired-end reads for individual NA07051 aligning to the breakpoint region from the case study in section 6.3. Notice that the reads do not show the switch in variational position consistent with a breakpoint for any kind of NAHR event at this location.

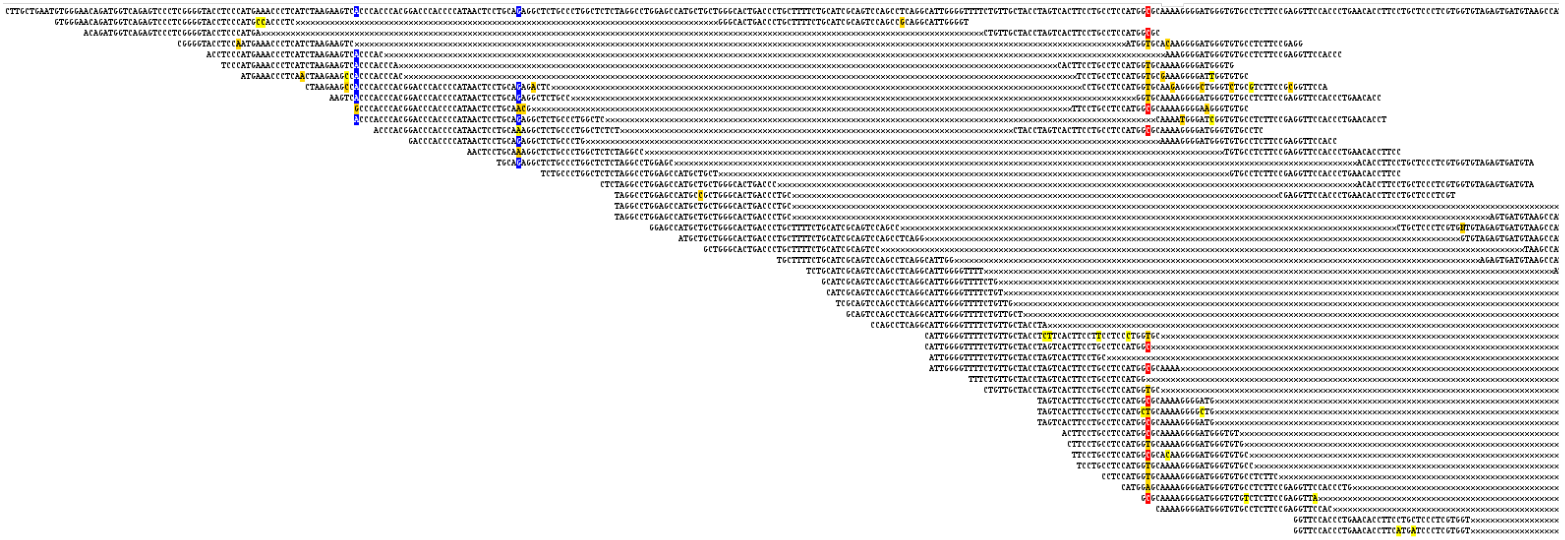


Figure C.3: Paired-end reads for individual NA18501 aligning to the breakpoint region from the case study in section 6.3. Notice that the reads do not show the switch in variational position consistent with a breakpoint for any kind of NAHR event at this location.

Appendix D

Read generation contexts

Read generation during next-generation sequencing is a nuanced process, and so defining the reference context of a base call can be confusing. As in PCR, a single strand of DNA serves as a template to which fluorescent nucleotides are sequentially added. Thus, the bases in the read are actually complementary to the strand of DNA which produced the read. Note also, the the *read* is created from $5'$ to $3'$, so that if R_1 was the first nucleotide created in R, then R_1 is the $5'$ end of R (thus, the template DNA is read from $3'$ to $5'$ for creation of the read R).

The raw bases of the read are therefore reverse complements of the template DNA strand which produced them, when the bases of template DNA strand are ordered from $5'$ to e' of the reference. A read aligner (e.g. BWA), will find the best alignment of the "raw" read *and the raw read's reverse complement* to the reference and reverse-complement the read as necessary so that it matches the reference where it is aligned. Note that the reference is single-stranded and ordered $5'$ to $3'$. Each mapped read has an associated indicator for whether it actually aligns to the reverse strand of the reference (i.e. the best alignment involved the reverse-complement of the read).

Thus if a read has been mapped to the reference and was *not* reverse-complemented, then the context at position j of the reference is $\mathcal{F}_{j-2}\mathcal{F}_{j-1}\mathcal{F}_j$. If the read was reverse-complemented, then the context at j in the reference is the complement of $\mathcal{F}_{j+2}\mathcal{F}_{j+1}\mathcal{F}_j$.

Bibliography

[1] *CASAVA Software Version 1.7 User Guide.*

[2] Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature*, 39(11):1361–1368, 2007.

[3] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

[4] Irina I. Abnizova, Steven Leonard, Tom Skelly, Andy Brown, David K. Jackson, Marina Gourtovaia, Guoying Qi, Rene te Boekhorst, Nadeem Faruque, Kevin Lewis, and Tony Cox. Analysis of context-dependent errors for illumina sequencing. *J. Bioinformatics and Computational Biology*, 10(2), 2012.

[5] Irina I. Abnizova, Tom Skelly, Fedor Naumenko, Nava Whiteford, Clive Brown, and Tony Cox. Statistical comparison of methods to estimate the error probability in short-read illumina sequencing. *J. Bioinformatics and Computational Biology*, 8(3):579–591, 2010.

[6] Can Alkan, Bradley P. Coe, and Evan E. Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376, March 2011.

[7] Can Alkan, Jeffrey M. Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O. Kitzman, Carl Baker, Maika Malig, Onur Mutlu, S. Cenk Sahinalp, Richard A. Gibbs, and Evan E. Eichler. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet*, 41(10):1061–1067, October 2009.

[8] J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research*, 11(6):1005–1017, June 2001.

[9] J. A. Beck, T. A. Campbell, G. Adamson, M. Poulter, J. B. Uphill, E. Molou, J. Collinge, and S. Mead. Association of a null allele of SPRN with variant Creutzfeldt-Jakob disease. *Journal of medical genetics*, 45(12):813–817, December 2008.

[10] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[11] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72, May 2012.

[12] Claudia M. Carvalho, Feng Zhang, and James R. Lupski. Structural variation of the human genome: mechanisms, assays, and role in male infertility. *Systems biology in reproductive medicine*, 57(1-2):3–16, February 2011.

[13] Jian-Min Chen, David N. Cooper, Claude Férec, Hildegard Kehrer-Sawatzki, and George P. Patrinos. Genomic rearrangements in inherited disease and cancer. *Seminars in Cancer Biology*, June 2010.

[14] Ken Chen, John W. Wallis, Michael D. McLellan, David E. Larson, Joelle M. Kalicki, Craig S. Pohl, Sean D. McGrath, Michael C. Wendl, Qunyuan Zhang, Devin P. Locke, Xiaoqi Shi, Robert S. Fulton, Timothy J. Ley, Richard K. Wilson, Li Ding, and Elaine R. Mardis. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–681, September 2009.

[15] Zhucheng Chen, Haijuan Yang, and Nikola P. Pavletich. Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature*, 453(7194):489–484, May 2008.

[16] Ming-Sin S. Cheung, Thomas A. Down, Isabel Latorre, and Julie Ahringer. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic acids research*, 39(15):e103, August 2011.

[17] Ming-Sin S. Cheung, Thomas A. Down, Isabel Latorre, and Julie Ahringer. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic acids research*, 39(15):e103, August 2011.

[18] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, October 2012.

[19] Yael Costa, Robert Speed, Rupert Öllinger, Manfred Alsheimer, Colin A. Semple, Philippe Gautier, Klio Maratou, Ivana Novak, Christer Höög, Ricardo Benavente, and Howard J. Cooke. Two novel proteins recruited by synaptonemal complex protein 1 (SYCP1) are at the centre of meiosis. *Journal of Cell Science*, 118(12):2755–2762, June 2005.

[20] N.G. de Bruijn and P. Erdos. A combinatorial problem. *Koninklijke Netherlands: Academe Van Wetenschappen*, 49:758–764, 1946.

[21] Olive J. Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, March 1961.

[22] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, July 1998.

[23] Bradley Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):pp. 96–104, 2004.

[24] Evan E. Eichler. Masquerading Repeats: Paralogous Pitfalls of the Human Genome. *Genome Research*, 8(8):758–762, August 1998.

[25] Simon A. Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Y. Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, Jon W. Teague, Peter J. Campbell, Michael R. Stratton, and P. Andrew Futreal. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research*, 39(Database issue):D945–D950, January 2011.

[26] Ziv Gan-Or, Anat Bar-Shira, Tanya Gurevich, Nir Giladi, and Avi Orr-Urtreger. Homozygosity for the MTX1 c.184T¿A (p.S63T) alteration modifies the age of onset in GBA-associated Parkinson's disease. *Neurogenetics*, 12(4):325–332, November 2011.

[27] Wenli Gu, Feng Zhang, and James Lupski. Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1):4+, 2008.

[28] Faraz Hach, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E. Eichler, and S. Cenk Sahinalp. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature methods*, 7(8):576–577, August 2010.

[29] P. J. Hastings, James R. Lupski, Susan M. Rosenberg, and Grzegorz Ira. Mechanisms of change in gene copy number. *Nature reviews. Genetics*, 10(8):551–564, August 2009.

[30] Fereydoun Hormozdiari, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, 19(7):1270–1278, July 2009.

[31] Fereydoun Hormozdiari, Iman Hajirasouliha, Phuong Dao, Faraz Hach, Deniz Yorukoglu, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, 26(12):i350–i357, June 2010.

[32] Alan D. Irvine, W. H. Irwin McLean, and Donald Y. M. Leung. Filaggrin Mutations Associated with Skin and Allergic Diseases. *N Engl J Med*, 365(14):1315–1327, October 2011.

[33] Tengchuan Jin, Feng Guo, Yang Wang, and Yuzhu Zhang. High-resolution crystal structure of human Dim2/TXNL4B. *Acta crystallographica. Section F, Structural biology and crystallization communications*, 69(Pt 3):223–227, March 2013.

[34] Arek Kasprzyk. BioMart: driving a paradigm change in biological data management. *Database*, 2011(0):bar049, January 2011.

[35] Ishmael Kasvosve, Marijn M. Speeckaert, Reinhart Speeckaert, Gwinyai Masukume, and Joris R. Delanghe. *Haptoglobin Polymorphism and Infection*, volume 50, pages 23–46. Elsevier, 2010.

[36] Wahab A. Khan, Joan Hm H. Knoll, and Peter K. Rogan. Context-based FISH localization of genomic rearrangements within chromosome 15q11.2q13 duplicons. *Molecular cytogenetics*, 4(1):15+, 2011.

[37] Jeffrey M. Kidd, Gregory M. Cooper, William F. Donahue, Hillary S. Hayden, Nick Sampas, Tina Graves, Nancy Hansen, Brian Teague, Can Alkan, Francesca Antonacci, Eric Haugen, Troy Zerr, N. Alice Yamada, Peter Tsang, Tera L. Newman, Eray Tüzün, Ze Cheng, Heather M. Ebling, Nadeem Tusneem, Robert David, Will Gillett, Karen A. Phelps, Molly Weaver, David Saranga, Adrianne Brand, Wei Tao, Erik Gustafson, Kevin McKernan, Lin Chen, Maika Malig, Joshua D. Smith, Joshua M. Korn, Steven A. McCarroll, David A. Altshuler, Daniel A. Peiffer, Michael Dorschner, John Stamatoyannopoulos, David Schwartz, Deborah A. Nickerson, James C. Mullikin, Richard K. Wilson, Laurakay Bruhn, Maynard V. Olson, Rajinder Kaul, Douglas R. Smith, and Evan E. Eichler. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, May 2008.

[38] Jeffrey M. Kidd, Tina Graves, Tera L. Newman, Robert Fulton, Hillary S. Hayden, Maika Malig, Joelle Kallicki, Rajinder Kaul, Richard K. Wilson, and Evan E. Eichler. A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms. *Cell*, 143(5):837–847, November 2010.

[39] Martin Kircher, Patricia Heyn, and Janet Kelso. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*, 12(1):382+, July 2011.

[40] Jan O. Korbel, Alexej Abyzov, Xinmeng Jasmine J. Mu, Nicholas Carriero, Philip Cayting, Zhengdong Zhang, Michael Snyder, and Mark B. Gerstein. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome biology*, 10(2):R23+, February 2009.

[41] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25–10, March 2009.

[42] Timo Lassmann and Erik Sonnhammer. Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6(1):298+, 2005.

[43] Christopher Lee, Catherine Grasso, and Mark F. Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics (Oxford, England)*, 18(3):452–464, March 2002.

[44] J. Lee and J. Lupski. Genomic Rearrangements and Gene Copy-Number Alterations as a Cause of Nervous System Disorders. *Neuron*, 52(1):103–121, October 2006.

[45] Andrew P. Levy, Irit Hochberg, Kathleen Jablonski, Helaine E. Resnick, Elisa T. Lee, Lyle Best, and Barbara V. Howard. Haptoglobin phenotype is an independent risk factor for cardiovascular disease in individuals with diabetes. *Journal of the American College of Cardiology*, 40(11):1984–1990, December 2002.

[46] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[47] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, November 2008.

[48] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, November 2008.

[49] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, January 2008.

[50] Elena V. Linardopoulou, Eleanor M. Williams, Yuxin Fan, Cynthia Friedman, Janet M. Young, and Barbara J. Trask. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, 437(7055):94–100, September 2005.

[51] Pengfei Liu, Melanie Lacaria, Feng Zhang, Marjorie Withers, P. J. Hastings, and James R. Lupski. Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. *American journal of human genetics*, 89(4):580–588, October 2011.

[52] J. R. Lupski. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in genetics : TIG*, 14(10):417–422, October 1998.

[53] James Lupski. Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. *Genome Biology*, 5(10):242+, 2004.

[54] James Lupski. Genomic disorders ten years on. *Genome Medicine*, 1(4):42+, 2009.

[55] Frazer Meacham, Dario Boffelli, Joseph Dhahbi, David Martin, Meromit Singer, and Lior Pachter. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 12(1):451+, November 2011.

[56] Paul Medvedev, Marc Fiume, Misko Dzamba, Tim Smith, and Michael Brudno. Detecting copy number variation with mated short reads. *Genome Research*, 20(11):1613–1622, November 2010.

[57] Ryan E. Mills, Klaudia Walter, Chip Stewart, Robert E. Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai C. Yoon, Kai Ye, R. Keira Cheetham, Asif Chinwalla, Donald F. Conrad, Yutao Fu, Fabian Grubert, Iman Hajirasouliha, Fereydoun Hormozdiari, Lilia M. Iakoucheva, Zamin Iqbal, Shuli Kang, Jeffrey M. Kidd, Miriam K. Konkel, Joshua Korn, Ekta Khurana, Deniz Kural,

Hugo Y. K. Lam, Jing Leng, Ruiqiang Li, Yingrui Li, Chang-Yun Lin, Ruibang Luo, Xinmeng J. Mu, James Nemesh, Heather E. Peckham, Tobias Rausch, Aylwyn Scally, Xinghua Shi, Michael P. Stromberg, Adrian M. Stutz, Alexander E. Urban, Jerilyn A. Walker, Jiantao Wu, Yujun Zhang, Zhengdong D. Zhang, Mark A. Batzer, Li Ding, Gabor T. Marth, Gil McVean, Jonathan Sebat, Michael Snyder, Jun Wang, Kenny Ye, Evan E. Eichler, Mark B. Gerstein, Matthew E. Hurles, Charles Lee, Steven A. McCarroll, and Jan O. Korbel. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, February 2011.

[58] Andre Minoche, Juliane Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, 12(11):R112+, November 2011.

[59] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, Md Altaf-Ul-Amin, Naotake Ogasawara, and Shigehiko Kanaya. Sequence-specific error profile of Illumina sequencers. *Nucleic acids research*, 39(13):e90, July 2011.

[60] Zhishuo Ou, Paweł Stankiewicz, Zhilian Xia, Amy M. Breman, Brian Dawson, Joanna Wiszniewska, Przemyslaw Szafranski, M. Lance Cooper, Mitchell Rao, Lina Shao, Sarah T. South, Karlene Coleman, Paul M. Fernhoff, Marcel J. Deray, Sally Rosengren, Elizabeth R. Roeder, Victoria B. Enciso, A. Craig Chinault, Ankita Patel, Sung-Hae H. Kang, Chad A. Shaw, James R. Lupski, and Sau W. Cheung. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome research*, 21(1):33–46, January 2011.

[61] Maria Papp, PeterLaszlo Lakatos, Karoly Palatka, Ildiko Foldi, Miklos Udvardy, Jolan Harsfalvi, Istvan Tornai, Zsuzsanna Vitalis, Tamas Dinya, Agota Kovacs, Tamas Molnar, Pal Demeter, Janos Papp, Laszlo Lakatos, and Istvan Altorjay. Haptoglobin Polymorphisms Are Associated with Crohn's Disease, Disease Behavior, and Extraintestinal Manifestations in Hungarian Patients. *Digestive Diseases and Sciences*, 52(5):1279–1284, May 2007.

[62] Paul A. Pevzner, Haixu Tang, and Glenn Tesler. De Novo Repeat Classification and Fragment Assembly. *Genome Research*, 14(9):1786–1796, September 2004.

[63] Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, August 2001.

[64] David Pulido, Marc Torrent, David Andreu, M. Victoria Nogués, and Ester Boix. Two human host defense ribonucleases against mycobacteria, the eosinophil cationic protein (RNase 3) and RNase 7. *Antimicrobial agents and chemotherapy*, 57(8):3797–3805, August 2013.

[65] Aaron R. Quinlan, Royden A. Clark, Svetlana Sokolova, Mitchell L. Leibowitz, Yujun Zhang, Matthew E. Hurles, Joshua C. Mell, and Ira M. Hall. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome research*, 20(5):623–635, May 2010.

[66] Mariko Sasaki, Julian Lange, and Scott Keeney. Genome destabilization by homologous recombination in the germ line. *Nature Reviews Molecular Cell Biology*, 11(3):182–195, February 2010.

[67] Yonatan Savir and Tsvi Tlusty. RecA-Mediated Homology Search as a Nearly Optimal Signal Detection System. *Molecular Cell*, 40(3):388 – 396, 2010.

[68] Xinwei She, Julie E. Horvath, Zhaoshi Jiang, Ge Liu, Terrence S. Furey, Laurie Christ, Royden Clark, Tina Graves, Cassy L. Gulden, Can Alkan, Jeff A. Bailey, Cenk Sahinalp, Mariano Rocchi, David Haussler, Richard K. Wilson, Webb Miller, Stuart Schwartz, and Evan E. Eichler. The structure and evolution of centromeric transition regions within the human genome. *Nature*, 430(7002):857–864, August 2004.

[69] Ellen Sidransky and Grisel Lopez. The link between the GBA gene and parkinsonism. *The Lancet Neurology*, 11(11):986–998, November 2012.

[70] Suzanne Sindi, Elena Helman, Ali Bashir, and Benjamin J. Raphael. A geometric approach for classification and comparison of structural variants. *Bioinformatics (Oxford, England)*, 25(12):i222–i230, June 2009.

[71] Suzanne Sindi, Selim Onal, Luke Peng, Hsin T. Wu, and Benjamin Raphael. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biology*, 13(3):R22+, 2012.

[72] A. B. Smith, J. D. Esko, and S. L. Hajduk. Killing of trypanosomes by the human haptoglobin-related protein. *Science (New York, N.Y.)*, 268(5208):284–286, April 1995.

[73] Paweł Stankiewicz and James R. Lupski. Genome architecture, rearrangements and genomic disorders. *Trends in genetics : TIG*, 18(2):74–82, February 2002.

[74] Paweł Stankiewicz and James R. Lupski. Structural variation in the human genome and its role in disease. *Annual review of medicine*, 61(1):437–455, 2010.

[75] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, August 2003.

[76] Xiaojing Sun, Hua Zhang, Dan Wang, Dalong Ma, Yan Shen, and Yongfeng Shang. DLP, a Novel Dim1 Family Protein Implicated in Pre-mRNA Splicing and Cell Cycle Progression. *Journal of Biological Chemistry*, 279(31):32839–32847, July 2004.

[77] Novocraft Technologies. Novoalign.

[78] Marc Torrent, M. Victòria Nogués, and Ester Boix. Eosinophil cationic protein (ECP) can bind heparin and other glycosaminoglycans through its RNase active site. *J. Mol. Recognit.*, 24(1):90–100, January 2011.

[79] Daniel J. Turner, Marcos Miretti, Diana Rajan, Heike Fiegler, Nigel P. Carter, Martyn L. Blayney, Stephan Beck, and Matthew E. Hurles. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet*, 40(1):90–95, January 2008.

[80] Eray Tuzun, Andrew J. Sharp, Jeffrey A. Bailey, Rajinder Kaul, V. Anne Morrison, Lisa M. Pertz, Eric Haugen, Hillary Hayden, Donna Albertson, Daniel Pinkel, Maynard V. Olson, and Evan E. Eichler. Fine-scale structural variation of the human genome. *Nat Genet*, 37(7):727–732, July 2005.

[81] Heather A. Woolls, Allison C. Lamanna, and Katrin Karbstein. Roles of Dim2 in ribosome assembly. *The Journal of biological chemistry*, 286(4):2578–2586, January 2011.

[82] Kai Ye, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, November 2009.

[83] Maisa Yoshimoto, Olga Ludkovski, Dave DeGrace, Julia L. Williams, Andrew Evans, Kanishka Sircar, Tarek A. Bismar, Paulo Nuin, and Jeremy A. Squire. PTEN genomic deletions that characterize aggressive prostate cancer originate close to segmental duplications. *Genes, chromosomes & cancer*, 51(2):149–160, February 2012.

[84] Ulrich M. Zanger and Matthias Schwab. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*, 138(1):103–141, April 2013.