Structure, Variation, and Reproducibility: Bayesian inference in problems arising from the study of RNA and an RNA-binding protein

by

Lauren Alpert Sugden

B.A., Wesleyan University, CT, 2008

M.S., Brown University, 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

This dissertation by Lauren Alpert Sugden 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_____ _____

Robert Reenan, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Peter M. Weber, Dean of the Graduate School

Vitae

Lauren Alpert Sugden

Applied Mathematics

Brown University

lauren_alpert@brown.edu

914-400-4704

Education

Ph.D *May 2014*

Applied Mathematics, Brown University

- Advisor: Charles Lawrence
- Dissertation: "Structure, Variation, and Reproducibility: Bayesian inference in problems arising from the study of RNA and an RNA-binding protein"

M.S. *May 2010*

Applied Mathematics, Brown University

B.A. *May 2008*

Mathematics & Physics, Wesleyan University

- Advisor: Ed Taylor

- Senior Project: "A Legendrian Unknotting Move"
- Graduated Phi Beta Kappa

Talks

- *Using a Bayesian Hierarchical Model to Assess Scientific Reproducibility:* Center for Computational Molecular Biology, Brown University, February 2014
- *Variation and Reproducibility: Bayesian Inference in the Investigation of an RNA-binding Protein:* Ursinus College, March 2014

Conference Posters

- *Computational Identification of Bacterial RNA Regulatory Motifs Using a Gibbs-Sampling Algorithm:* American Society for Microbiology Conference on Regulating with RNA in Bacteria, March 2011

Publications

(As Lauren A Sugden):

- Sugden, Lauren A., *et al.* "Assessing the validity and reproducibility of genome-scale predictions." *Bioinformatics* 29.22 (2013): 2844-2851.

(As Lauren V Alpert):

- Savva, Yiannis A., *et al.* "Auto-regulatory RNA editing fine-tunes mRNA re-coding and complex behaviour in Drosophila." *Nature communications* 3 (2012): 790.

- Wei, Donglai, Lauren V. Alpert, and Charles E. Lawrence. "RNAG: a new Gibbs sampler for predicting RNA secondary structure for unaligned sequences." *Bioinformatics* 27.18 (2011): 2486-2493.

- Reid, Daniel C., *et al.* "Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence." *RNA* 15.12 (2009): 2385-2397.

Honors

- Honorable Mention, NSF GRFP (2010)

Teaching Experience

- Instructor: Methods of Applied Mathematics I, Brown University, *Spring 2013*

- Instructor: One-week introductory course for HHMI summer program, Brown University MCB Department, *Summer 2011 & 2012*

- Staff Member: Math Resource Center, Brown University, *Spring 2011-Fall 2012, Fall 2013*

- Guest Lecturer: Essential Statistics, Brown University, 2 lectures, *Spring 2011*

- Teaching Assistant: Essential Statistics, Brown University, *Spring 2010*

- Teaching Assistant: Statistical Inference I, Brown University, *Fall 2009*
- Private Tutor: Information Theory, HS Algebra, Calculus B/C, Calculus I, Actuarial Exam Preparation, Introductory Biostatistics, Statistical Inference I

Certifications

- Sheridan Center Certificate I (reflective teaching practices)

Acknowledgements

I have received a wealth of support from many people over the past six years. I would like to thank the following people in particular:

Chip Lawrence: It's been a real pleasure learning from you and working with you. Your curiosity is contagious, and your support and encouragement have helped me grow as an independent scientist in ways I never expected.

Bill Thompson: Thank you for answering all of the questions I was too embarrassed to ask anyone else. Your patience is immeasurable, and I am so grateful that you've been there helping me every step of the way.

Rob Reenan and Yiannis Savva: Working with you has been a blast. I have so looked forward to all of our meetings, always curious about what exciting new ideas are in store.

The Lawrence lab: Thank you for offering thoughtful suggestions on my projects, and for sharing your projects with me. It has been really neat to watch all of them develop and intertwine in interesting ways. In particular, thank you to Matt Parks. It was so much nicer going through this process alongside someone else who understood the various ups and downs.

Mom, Dad, and Naomi: What can I say? I am so grateful for a lifetime of love

and support, with a healthy dose of silliness. You have kept me grounded throughout all of this, and there is no way that I would be here without you.

Donata and Bill: I am so appreciative for your thoughtful advice on navigating the world of academia, as well as the encouragement that you have given me over the last few years. I feel so lucky to be able to call you family.

Arthur: There are no words, but I'll try anyway. Since we met, you have been my #1 advocate, and that has meant everything to me. You have worked actively over that last six years to help me find confidence in myself as a scientist and as a teacher, and in the meantime, you have made the journey so much fun. Thank you.

Abstract of " Structure, Variation, and Reproducibility: Bayesian inference in problems arising from the study of RNA and an RNA-binding protein " by Lauren Alpert Sugden, Ph.D., Brown University, May 2014

Far from being solely a passive messenger between DNA and protein, RNA is a complex molecule involved in regulation at many levels. While the best-known RNAs, messenger RNAs (mRNA), do exist to carry information, there is also a large population of non-coding RNAs that can take on complicated structures that enable them to perform functions throughout the cell. Even mRNAs are not static molecules, however. Mechanisms for altering mRNA sequences result in transcripts that are no longer faithful representations of the information in the genome. One such mechanism is RNA editing by ADAR, which binds double-stranded RNA and targets a particular adenosine for conversion to inosine, which is recognized by the cell as guanosine. A major consequence of editing is amino-acid recoding, resulting in protein diversification.

In this thesis, we look at three problems motivated by the study of RNA and ADAR. First, we propose a method for assessing the reproducibility of genome-scale studies that make a large number of predictions, using the prediction of ADAR binding sites as a motivating example. We then address the problem of avoiding false positives when identifying subtle signals such as ADAR modifications in high-throughput sequencing data in the presence of genetic polymorphisms specific to laboratory populations. Finally, we turn to structural prediction of RNA, inferring the common structural and sequence characteristics of a set of related transcripts.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER **One**

Introduction

1.1 Regulatory roles of RNA

RNA, once thought to be primarily a passive messenger between the genes encoded in our DNA and the proteins that determine our phenotype, is now understood to be a complex class of molecules involved in the regulation of gene expression at many levels. RNA is intimately involved in the determination of which versions of which genes will be expressed in a given cell, and in what quantities. At times, RNA molecules are the targets of this regulation, and at other times they are active participants in the regulation of either themselves or other transcripts.

RNA molecules destined to become templates for protein synthesis, called mRNA, can be modified in ways that alter the downstream protein sequences and increase protein diversity. One major example of this is alternative splicing, a process by

which non-coding sequences are removed and coding sequences are joined together in a combinatorial fashion [8]. Another mechanism, RNA editing, described in Section 1.3, alters targeted nucleotides. mRNAs can also contain functional RNA elements within their own transcripts, including the group II catalytic intron [150], an RNA element that catalyzes its own splicing, and the riboswitch [155], a structure located in the untranslated region immediately upstream of some genes that, in the presence of a ligand, can sequester the ribosome-binding site and prevent translation of the mRNA into protein.

Non-coding RNAs (ncRNAs), those that do not become protein templates, are less well understood than mRNAs, but fall into a number of classes and perform a wide range of important functions. RNA splicing is carried out by the spliceosome, a complex containing both proteins and small nuclear RNAs [122]. Ribosomal RNAs and transfer RNAs play major roles in the translation of mRNAs into amino acid sequences. Through the process of RNA interference [34], microRNAs and endogenous small interfering RNAs (endo-siRNAs) can target mRNAs with complementary sequences, causing them to be silenced either through destruction of the mRNA or by the prevention of translation. In addition, endo-siRNAs help regulate the activity of transposons, DNA elements that can "jump" throughout the genome and potentially interrupt important genes [170]. The endo-siRNAs involved in this process are themselves derived from members of the particular transposon family that they regulate, and in addition to silencing RNA transcripts, they are also involved in transcriptional silencing— preventing transcription through heterochromatic silencing of genomic regions containing the transposons.

The ability of RNA to carry genetic information and engage in complex regulation as well as enzymatic activity has led scientists to propose the "RNA world" hypothesis, which gives us a possible glimpse into our past. The hypothesis holds

that the original building blocks of life as we know it were not the DNA, RNA, and protein that we are made of today, but solely self-replicating RNAs that carried out all necessary functions [120]. Some of these functions would have eventually been taken over by DNA (for information storage) and proteins (for enzymatic activities), but clearly remnants of the RNA world, if it existed, are still around today.

1.2 RNA secondary structure

The primary sequence of RNA is made up of adenine (A), cytosine (C), guanine (G), and uracil (U) nucleotides in a single string. Each of these nucleotides has binding preferences; As and Us form strong bonds, and Gs and Cs form bonds that are even stronger. These are called canonical base pairs because they are the base pairs found in the DNA double helix, substituting T (thymine) for U. Gs and Us can form non-canonical, or "wobble" base pairs, which are less energetically favorable than the canonical base pairs, but occur somewhat frequently in RNA base-pairing interactions. These binding preferences lead the RNA to fold on itself, creating what is referred to as its "secondary structure," in which some bases are involved in pairings and others are left single-stranded. The most common type of sub-structure is called a stem-loop, formed from two complementary regions in the RNA sequence, as illustrated in Figure 1.1. More complex structures can be built up from stem loops, including the cloverleaf structure of tRNA shown in figure 1.2. The 3-dimensional shape that the RNA forms in the cell is called its tertiary structure. While computational models, including the one described in this thesis, can fairly successfully predict secondary structure, the prediction of tertiary structure by computational means is still an extremely difficult problem [137].

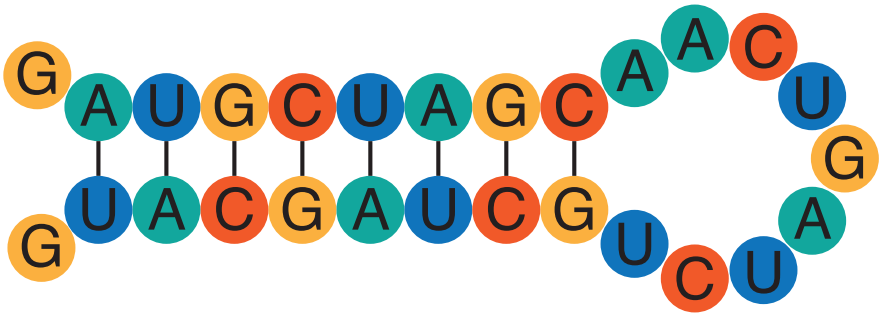

Figure 1.1: A stem loop

Figure 1.2: Cloverleaf structure of tRNA

These structures may be local within longer, mostly unstructured transcripts, as is the case for the riboswitch, or they can be global, as is the case for tRNAs and the RNAs involved in splicing. For most of the ncRNAs mentioned above, their proper function is dependent on their structure [177, 81, 148, 83]. Meanwhile, the role structure plays in the case of miRNAs and endo-siRNAs is slightly different. In both cases, the RNA molecules are cut by a protein called Dicer from precursor RNAs that form perfect stems [49]. For endo-siRNAs targeting transposons, this precursor molecule can be formed from two members of the target transposon family arranged in opposite orientation, called "inverted repeats," which from long double-stranded regions upon transcription [116].

1.3 RNA editing with ADAR

Structure can also be extremely important for the recognition of targets by RNA-binding proteins. One such protein is ADAR (adenosine deaminase acting on RNA), which carries out the most common type of RNA editing mentioned above. ADAR binds to double-stranded regions of RNA [71], targets adenosine (A) residues, catalyzing their conversion to inosine (I) through hydrolytic deamination [5]. Inosine is recognized by cellular machinery as guanosine, so editing essentially results in an A→G substitution. In imperfect stems, ADAR can target specific As through a combination of sequence and structural cues that are not completely understood, although a G immediately upstream of the targeted A is known to be a deterrent to ADAR binding [29] and it is thought that the location of bulges and internal loops within the stem serve to guide ADAR to its targets [90, 138]. This type of ADAR activity is referred to as "site-selective" editing. For these sites, the imperfect pairing often occurs between the region surrounding the target adenosine and a typically

intronic downstream region called the editing-site complementary sequence (ECS). In perfect double-stranded regions, ADAR can edit "promiscuously," altering up to 50% of the adenosines in the stem [164]. Since I-U base pairs are less stable than A-U or G-U pairs, promiscuous editing results in the disruption of the perfect stem structure.

While some organisms have multiple ADAR proteins, *Drosophila* have only one, called dADAR. One consequence of ADAR-mediated RNA editing is amino acid recoding, which has been shown in *Drosophila* to affect proteins associated with vesicular trafficking, ion homeostasis, signal transduction, ion channels, and the cytoskeleton [37, 167]. This recoding is also crucial for normal adult nervous system function [84]. In addition, variability in RNA editing between flies has been suggested as a mechanism for individual differences in behavior and neuronal physiology [62]. dADAR has even been shown to edit its own transcript, resulting in a recoding of Serine to Glycine in a subset of transcripts, creating two ADAR isoforms [85]. The consequences of this recoding were shown by Savva *et al.* [115] to alter editing in a site-specific manner, and to influence the location of dADAR within the nucleus. While the un-edited isoform was observed mainly in the nucleolus, the edited isoform localized to a nuclear punctus outside of the nucleolus. The consequence of this localization is as of yet unknown.

Mutations in the human ADAR genes have been shown to cause Aicardi Gutieres Syndrome [52], an autoimmune disorder that primarily affects the brain and skin, Bilateral Striatus Necrosis [66], a movement disorder associated with brain abnormalities and developmental regression, and Dyschromatosis Symmetrica Hereditaria [114], a pigmentation disorder. In addition, ADAR has been shown to be downregulated in motor neurons of ALS patients [108], and seizure activity has been shown to correlate with the loss of editing of GluR-2, a neurotransmitter receptor [45]. Ab-

normal patterns of editing have also been demonstrated in Schizophrenia [106] and Depression [58].

Amino-acid recoding is not the only function of ADAR. A fascinating aspect of RNA editing is its interaction with the RNA interference pathway mentioned above, by which endo-siRNAs derived from inverted members of a transposon family target other members for silencing. This interaction arises because the double-stranded RNA molecules from which Dicer cuts the endo-siRNAs are the perfect target for promiscuous editing by ADAR. Since promiscuous editing disrupts the perfect double stem recognized by Dicer, it also inhibits the production of endo-siRNAs [164]. This interference has even been shown to have an effect on heterochromatic silencing [116].

1.4 Overview

In Chapter 2, we use a study of ADAR binding sites throughout the *Drosophila* genome to motivate an investigation into the reproducibility of genome-scale studies making thousands of predictions. We propose the use of a hierarchical model to assess reproducibility, and describe a scheme for designing experiments that will yield optimally reproducible results. In Chapter 3, we address the problem of identifying targets of ADAR from high-throughput sequencing data in the presence of genetic polymorphisms. We describe an expectation-maximization algorithm for inferring polymorphic sites while simultaneously learning the error rates of the sequencing technology and we extend this to the simultaneous inference of polymorphisms and editing sites. In Chapter 4, we turn to structural prediction of RNA. We introduce an algorithm, RNAG, that infers a common structure for multiple related sequences, and we describe an extension to RNA motif finding.

CHAPTER **Two**

Assessing the Reproducibility of Validation Experiments in Genome-scale Studies

2.1 Background and Motivation

2.1.1 The need for reproducibility

The need for validation and reproducibility in scientific research is clear. As a recent New York Times article about the recent spotlight on scientific irreproducibility put it, "If a result appears only under the full moon with Venus in retrograde, is it truly an advance in human knowledge?" [129] With the discovery that results of biological experiments have been shown to vary due to such seemingly benign factors as laboratory conditions, reagent lots, cell generations, and individual experimenter technique [89, 64, 102], this issue becomes all the more crucial.

For our purposes, we define validation and reproducibility broadly as follows:

Validation: Can the original investigator confirm his or her findings by an independent method?

Reproducibility: Given the necessary materials and information, can an independent investigator confirm the findings?

Criticism of many recent embarrassing incidents has brought a good deal of attention to both of these issues with claims of insufficient validation of bold research findings, or an inability to replicate such findings. In one study, a group of scientists set out to reproduce the results of several pre-clinical studies in cancer research [6]. They selected 53 papers that were deemed to be particularly influential in the field, and found that they were able to replicate only six, for an underwhelming success rate of about 11%. This was the case even after the original investigators were invited to provide their own reagents and even supervise the replicate experiments. Many of these irreproducible results led to secondary publications and even clinical trials, and in a disturbing trend, papers that could not be reproduced had a much higher citation rate, especially when published in high-impact journals.

Genomics has had its fair share of recent embarrassment as well; one article published in Science claimed to have found over 10,000 sites of RNA editing in human B cells, spanning all 12 possible nucleotide changes [141]. Since only two major types of RNA editing are currently known (A to I editing by ADAR, and C to U editing by proteins called APOBECs), other investigators were keen to dig deeper. Within a year, three independent labs had found that the majority of those findings were false positives, suggesting a combination of contributing factors including sequencing error, genetic variation, and mapping issues [134, 158, 111].

In another Science article looking at genomic imprinting, an epigenetic mechanism for preferentially expressing either the maternal or paternal allele of a particular gene, investigators claimed to have found the phenomenon at 1,300 loci [41]. This came as a surprise to the community, since only ∼100 imprinted sites had been seen until that point. Two years later, an independent lab tried to replicate the results, and found that they could only reproduce about 13% of them [17].

These two studies are part of a growing body of genomics literature that uses high-throughput sequencing technologies to produce long lists of sites. Other examples of such studies include the identification of 2,000 new gene promotors [99], thousands of targets of six transcription factors involved in regulation of the anterior-posterior axis in the embryo [182], 14,000 binding sites of proteins associated with insulators (DNA sequences that block the spread of chromatin) [91], and 1,600 sites of polymerase II stalling [68]. For the two genomics studies described above, the problem likely lies in a combination of validation and reproducibility. In the first study, they drew a random sample of 12 sites to validate with Sanger sequencing, but only 4 out of 12 were in previously-unknown nucleotide change categories (i.e. not A to G or C to U). Of these, three successfully validated. In the second study, the authors validate a "select few," but there is little additional information. In fact, it is alarmingly difficult to find a genomics study in which a random sample is drawn, and an estimate of the mean is given, along with confidence limits. And it is nearly impossible to find a study that employs replicates for validation studies.

Of course, these incidents are not specific to genomics or cancer research. Similar problems have arisen in genetics [72, 56], neuroscience [73], pharmacology [159], proteomics [36], and psychology [44, 181], and they crop up in widely-used technologies such as microarrays [67], siRNA-based screens [92], and mass spectrometry [36].

The response to all of these issues has been enormous. There have been many subsequent reports on the prevalence of irreproducibility in research [1, 126, 159], calls for transparency in publishing, including open-source code [128] and a publication platform for neutral studies [159], calls for increased statistical rigor, especially in big-data settings [26, 173, 146, 28], and editorials discussing the need for methods for assessing reproducibility [162, 27]. Scientific journals are taking note too; in April of 2013, the editors of *Nature* announced new editorial measures including more careful examination of statistical methods, encouraging the submission of raw data, and abolishing space restrictions on methods sections [2]. In addition, they now provide a checklist for authors that is designed to both increase transparency and prompt a more rigorous thought process for future submissions.

In a large-scale effort to address these issues, an organization called the Reproducibility Initiative was founded in 2013 with the express purpose of identifying and rewarding reproducible research [4]. The organization matches voluntary submissions with independent labs that attempt to reproduce the experiments, awarding a badge of reproducibility if the attempt is successful. This is an extremely worthwhile effort, but there are a few downsides. Recreating experiments in a new lab can be costly and time-intensive, but more importantly, the badge awarded provides only a binary "yes" or "no" answer to the question of reproducibility. This can be problematic, since in the presence of biological and sample preparation variability, the correct answer may well be "sometimes." Put another way, the scientific community needs methods for quantitatively assessing reproducibility. We provide such a method through the use of biological and sample preparation replicates.

2.1.2 Use of replicates

The use of replicates, whether technical, biological, or simulated, has been shown to be useful in many contexts. McShane *et al.* [80] and Kerr and Churchill [132] simulate microarray replicates to determine the stability of clusters of genes that exhibit similar expression patterns. In the search for differentially expressed genes, technical replicates provide additional power for microarrays [157], and biological replicates reduce false positives in conclusions drawn from serial analysis of gene expression (SAGE) data [74, 97] and improve accuracy in calls made from RNAseq data [121]. Xia *et al.* [78] use replicate time series datasets to capture time-delayed associations between microbes. Taking advantage of replicates on a larger scale, meta-analyses of genome-wide association studies (GWAS) like those described in Zeggini and Ioannidis [183] combine datasets from multiple laboratories to gain enough power to detect associations between particular genes and diseases such as type II diabetes [48] and Crohn's disease [63].

In some cases, replicates have been incorporated into optimal study design. Many articles have been written on the number of replicates required to detect a certain fold-change in gene expression via microarray studies [9, 157, 171, 175]. For GWAS, Moonesinghe *et al.* [95] find the required number of samples to replicate an association across studies with a certain level of between-study heterogeneity, and Pahl *et al.* [156] propose multistage designs which, for a given budget, maximize the power to find associations. Auer and Doerge [3] advocate careful design of RNA-seq experiments, including sampling, randomization, replication, and blocking. To our knowledge, no one in the biological community has used replicates to assess the reproducibility of validation studies, or has proposed a method for optimal design of such experiments with respect to reproducibility, as we do here.

There has, however, been considerable work on assessing the reproducibility of high-throughput experiments, especially in the context of ranked lists of putative sites [10]. In this context, reproducibility is most closely related to precision (or stability) in that the relevant issue is the similarity of two ranked lists generated from biological replicates (or different high-throughput platforms, different ranking algorithms, etc...). There are many different measures used to assess the similarity of two or more ranked lists, from Spearman's rank correlation [112, 77] to overlap counts for the top k sites [87], to weighted overlap counts that emphasize correlation between high ranking sites over low ranking sites [113]. Li *et al.* [142] improve on these measures with a mixture model consisting of reproducible and irreproducible sites, which assigns each signal a reproducibility index based on its consistency across replicates, which approximates its probability of being reproducible. They define the "irreproducible discovery rate" (IDR), an analog of the false discovery rate for multiple hypothesis testing [168], which determines the "expected rate of irreproducible discoveries" for sites whose probability of being irreproducible is below some threshold γ. Their methods provide a principled method for selecting sites for further study and for evaluating ranking algorithms. Although here we also address the issue of reproducibility, our focus is different. We are not concerned with the precision of high-throughput technologies or ranking algorithms, but rather with the reproducibility of independent validation experiments seeking to verify findings of such high-throughput experiments. The validation experiments taken individually give us information about the accuracy of the findings, whereas our model of biological replicates allows us to assess the reproducibility of the given validation scheme in the face of biological and sample preparation variation.

2.1.3 Search for and validation of ADAR editing sites

The model described in this chapter was published in *Bioinformatics* in 2013 [75], concurrent with the publication of a companion study by our collaborators that sought to identify the targets of the ADAR protein on a genome-wide scale[51]. Through five iterative cycles of prediction and training using single molecule sequencing and a random forest machine learning algorithm, and validation using Sanger sequencing, they produced a list of 1,799 validated sites and 1,782 "top tier" site predictions.

They carried out a validation experiment in a single pool of ten Canton-S *Drosophila* raised at a constant 25°C on standard molasses food and under 12 hour day/night cycles. RNA was extracted from whole body 1-2 day-old males using TRIzol reagent (Invitrogen), cDNA was made and amplified with reverse PCR using gene-specific primers. Since ADAR catalyzes an adenosine to inosine change, and inosine is recognized by the sequencing machinery as guanosine, Sanger sequencing was used to identify sites in the RNA containing both A and G where only an A was expected. An idealized example of an editing site represented in a Sanger trace is given in Figure 2.1. For this first validation experiment, they set out to validate 298 predicted sites, with the purpose of investigating editing levels in exons, introns, and intergenic regions. The PCR reaction failed for about 30% of the sites, leaving 205 sites, of which 151 successfully validated.

Since we were interested not only in the validation results for this experiment, but also in the effect of biological and sample preparation variation on the outcome of such an experiment, through the course of our collaboration, five additional validation experiments were carried out under the same experimental conditions. Four were performed in RNA pools created by a skilled investigator well versed in the

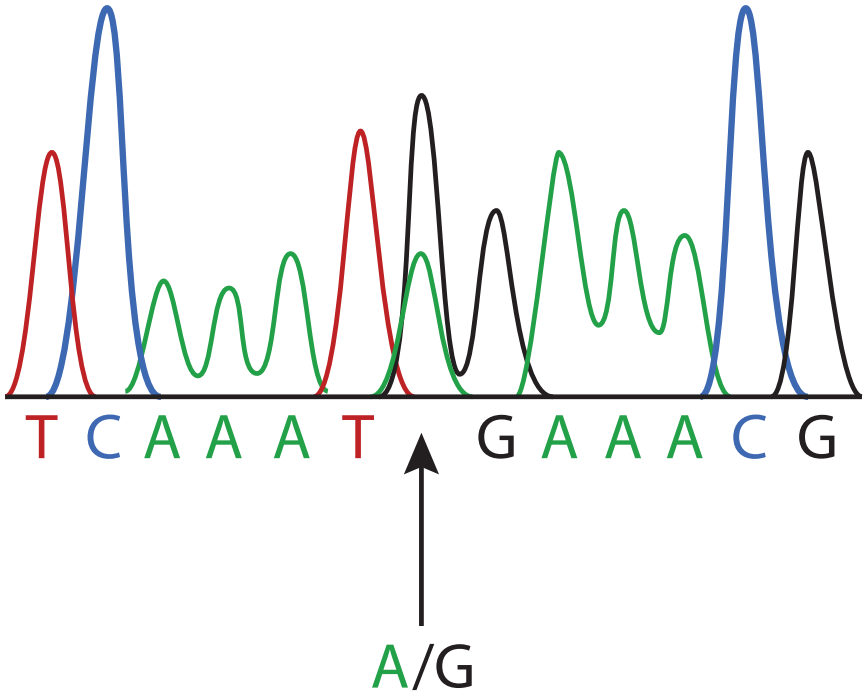

Figure 2.1: Identification of edit sites by Sanger sequencing. Editing is represented by overlapping peaks in green (A) and black (G).

protocol, and one was performed in a pool created by an investigator with much less experience, which was discarded upon being found to be of much lower quality than the others. The details of this can be found in Appendix A.

2.2 The model

2.2.1 Single sample validation

In the absence of biological and sample preparation variability, validation experiments provide an audience with necessary information for building on a set of findings. In particular, follow-up studies can be pursued with some level of confidence

in the results of the original study. A simple, yet rigorous way to do this is to draw a random sample of predictions to validate.

If we draw a sample of size n from a population of size N which contains K valid sites and $N - K$ invalid sites, the number that turn out to be valid in our sample follows a hypergeometric distribution with parameters N, K, and n. Thus the number of tested predictions k that turn out to be valid is given by the following distribution:

$$P(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{2.1}$$

When N is sufficiently large, however (as is our case), we can estimate this distribution with a binomial distribution with parameters $n = n$ and $p = \frac{K}{N}$. Then the probability of k valid predictions out of n is given by:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2.2}$$

We can calculate the maximum likelihood estimate (MLE) of p as $\hat{p} = \frac{k}{n}$, and use the central limit theorem to compute a confidence interval around the MLE for the true proportion valid in the top tier list, with a $(1 - \alpha)$ confidence limit given by the following expression:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.3}$$

where z_A is used to denote the point on the standard normal curve below which the area under the curve is A. For example, for a 95% confidence interval, α is set to 0.05, with $z_{1-\frac{0.05}{2}} = z_{0.975} = 1.96$. For the ADAR top-tier list in the first RNA pool

in which 151 out of 205 predictions validated, a 95% confidence interval for the true proportion valid is given by 0.74 ± 0.06, or (0.68,0.80). The rigorous interpretation of this interval is as follows: given the time and funds, had the same scientist run the same validation protocol on all top-tier predictions in this first RNA pool, under the same experimental conditions, then we have 95% confidence that the proportion valid out of all 1,782 predictions would be between 68% and 80%. This is a useful calculation, but of course, it does not account for any variation on the biological or sample preparation level.

2.2.2 Hierarchical modeling of multiple samples

In order to address the effect of such variation, we employed biological (thus necessarily sample-preparation) replicates. In the context of the ADAR study, these were pools of amplified cDNA generated from different groups of flies, and prepared independently. In each replicate, a random sample of predictions is chosen for validation, resulting in some number of successes and failures. We model the number of successes in each pool with a binomial distribution with parameters N and p, where N is the size of the random sample in the pool, and p is the proportion of the total predictions that would be found to validate in that particular pool, given the time and funds. We then model the individual ps as draws from a common beta distribution. This hierarchical model allows us to draw power from all replicates at once, while still allowing the individual proportions to vary, accounting for possible variation. Finally, we place a prior on the parameters of the beta distribution, because in general we don't expect to have enough replicates to be truly confident in a point estimate. This model is illustrated in Figure 2.2.

Our task is now to infer the distribution of α and β from the existing replicates

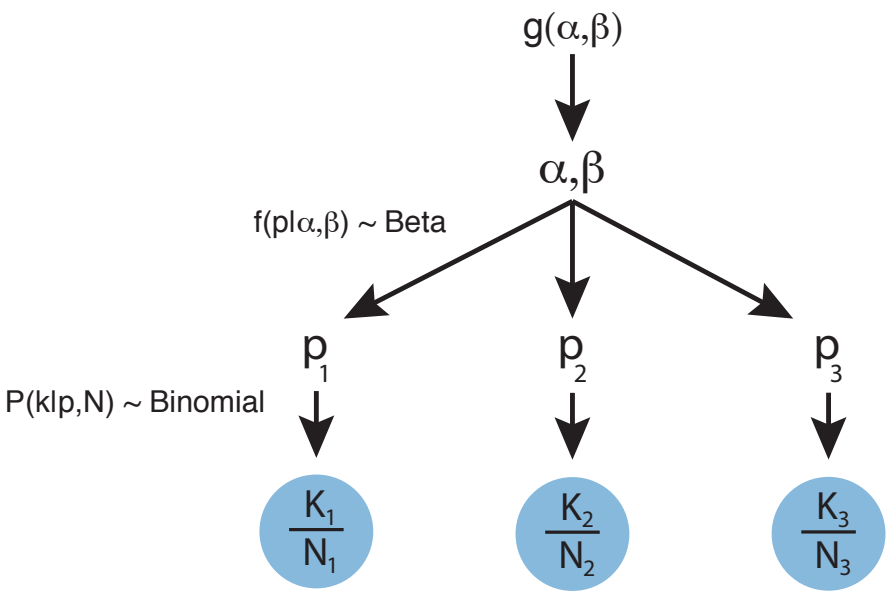

Figure 2.2: Illustration of the hierarchical model for three replicates

and use that distribution to understand what the results of a new replicate might look like. In this way, we can quantitatively assess reproducibility. To infer this distribution, we first write out the likelihood of the parameters α and β. For a single replicate, we have a beta-binomial likelihood:

$$\begin{aligned} P(k|\alpha,\beta,N) &= \int_p P(k|p,N)f(p|\alpha,\beta)dp \\ &= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)\Gamma(k+\alpha)\Gamma(N-k+\beta)}{\Gamma(k+1)\Gamma(N-k+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+N)} \\ &\propto \frac{\Gamma(\alpha+\beta)\Gamma(k+\alpha)\Gamma(N-k+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+N)} \end{aligned} \tag{2.4}$$

We have dropped the terms involving only N and k, since they are constant with respect to the distribution that we are interested in (the distribution of α and β. The complete likelihood term is given by the product of the individual likelihoods

for each replicate:

$$P(k_1, ..., k_M|\alpha, \beta, N_1, ..., N_M) \propto \prod_{m=1}^{M} \frac{\Gamma(\alpha+\beta)\Gamma(k_m+\alpha)\Gamma(N_m-k_m+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+N_m)} \tag{2.5}$$

Next we turn to the prior distribution. We wanted an uninformative, or diffuse prior to avoid imposing prior assumptions about the amount of variation that actually exists among replicates. We chose the diffuse prior recommended by Gelman *et al.* [35], a uniform distribution on transformed axes:

$$g(\alpha, \beta) \sim \text{Uniform}(\frac{\alpha}{\alpha+\beta}, \frac{1}{\sqrt{\alpha+\beta}}) \tag{2.6}$$

Note that the first term is the mean of the beta distribution, indicating that we do not place any weight on one proportion over another. To transform this prior into a distribution over variables α and β, we first take the determinant of the Jacobian of the function from α and β to the transformed axes. Let's call this function $h(\alpha, \beta)$, with components h_1 and h_2:

$$h_1(\alpha, \beta) = \frac{\alpha}{\alpha+\beta} \qquad h_2(\alpha, \beta) = \frac{1}{\sqrt{\alpha+\beta}} \tag{2.7}$$

Then the Jacobian of h is given by the following:

$$J_h = \begin{bmatrix} \frac{\partial h_1}{\partial \alpha} & \frac{\partial h_1}{\partial \beta} \\ \frac{\partial h_2}{\partial \alpha} & \frac{\partial h_2}{\partial \beta} \end{bmatrix} = \begin{bmatrix} \frac{\beta}{(\alpha+\beta)^2} & \frac{-\alpha}{(\alpha+\beta)^2} \\ \frac{-1}{2(\alpha+\beta)^{\frac{3}{2}}} & \frac{-1}{2(\alpha+\beta)^{\frac{3}{2}}} \end{bmatrix} \tag{2.8}$$

whose determinant is $(\alpha+\beta)^{-\frac{5}{2}}$. Since the original distribution is uniform, this is

also the expression for the prior in terms of the variables α and β:

$$f(\alpha, \beta) = (\alpha + \beta)^{-\frac{5}{2}} \tag{2.9}$$

The posterior distribution over the parameters of the beta distribution given the data in M pools is therefore given by

$$f(\alpha, \beta | k_1, ..., k_M, N_1, ..., N_M) \propto (\alpha + \beta)^{-\frac{5}{2}} \prod_{m=1}^{M} \frac{\Gamma(\alpha + \beta)\Gamma(k_m + \alpha)\Gamma(N_m - k_m + \beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + N_m)} \tag{2.10}$$

To visualize the distribution, we plot these unnormalized posterior probabilities over a 200x200 grid in the $(\log \frac{\alpha}{\beta}, \log(\alpha + \beta))$ plane, beginning with method of moments estimators, then shifting and rescaling the grid as needed until the contour plot occupies 80% of the width and height of the grid window, ensuring that we have a complete picture with enough resolution to capture important elements of the distribution. An example contour plot is shown in Figure 2.3.

We can draw samples from this distribution using a procedure outlined in Gelman *et al.* [35]. Briefly, we draw a row of the grid in proportion to its marginal distribution, then draw a column in proportion to its conditional probability given the chosen row. We then apply independent uniform horizontal and vertical jitters centered at the grid point and with width equal to the grid spacing. This ensures that we can cover the entire space, and amounts to approximating the posterior distribution by interpolating between grid points with step functions. Further details can be found in Appendix B.

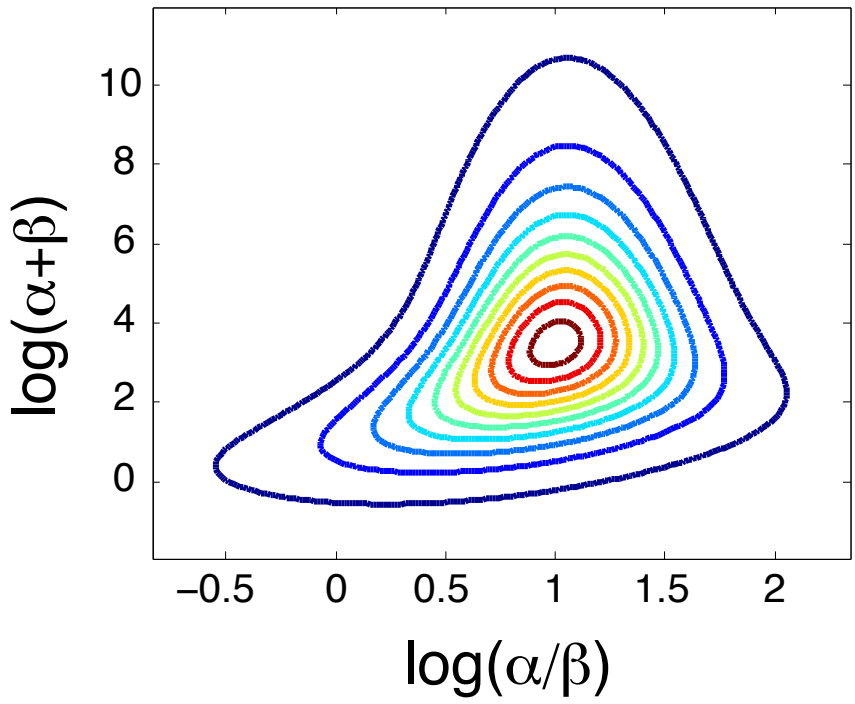

Figure 2.3: Example of contour plot representing the posterior distribution of α and β. Distribution is for three replicates of 20 experiments each, with 14, 15, and 16 successful validation experiments. Contour lines represent 10%,20%,...,90%,100% of the maximum value computed over the grid.

2.2.3 The predictive distribution

The idea of predictive inference is to use observed data to predict a new data point in the future. From a Bayesian perspective, this can be accomplished by computing the posterior distribution of the new data point. This task is often accomplished by constructing a model of all possible future data points, then marginalizing over all of the model parameters. In our case, the distribution we are interested in is $P(k_{\text{new}}|N_{\text{new}}, k_1^M, N_1^M)$, which we obtain by marginalizing over α, β, and p:

$$\begin{aligned} &P(k_{\text{new}}|N_{\text{new}}, k_1^M, N_1^M) \\ &\quad = \int_\alpha \int_\beta \int_p P(k_{\text{new}}|p, N_{\text{new}}) f(p|\alpha, \beta) f(\alpha, \beta|k_1^M, N_1^M) dp d\beta d\alpha \end{aligned} \tag{2.11}$$

where $P(k_{\text{new}}|p, N_{\text{new}})$ follows a binomial distribution, $f(p|\alpha, \beta)$ follows a beta distribution, and $f(\alpha, \beta|k_1^M, N_1^M)$ is the posterior distribution given in Equation 2.10.

We cannot compute this integral directly, but we can approximate it by drawing representative samples from each component. This will allow us to say something quantitative about the reproducibility of the validation experiments that have been performed. We implement the following procedure:

- for $j = 1$ to $1,000$:
    1. Draw (α_j, β_j) from the posterior distribution on α and β
    2. Draw the proportion valid in a new replicate, p_j, from a beta distribution with parameters α_j and β_j
    3. For a given sample size N, draw the number of successful validations k_j from a binomial distribution with parameters N and p_j

From these samples, we can estimate the distribution of successful validations in a new replicate, and we can extract summary statistics such as mean, variance and percentiles. We typically focus on the distribution of proportions instead of the number of successes, since the distribution over k is dependent on the choice of N which will vary by situation. We call this the "predictive distribution."

2.3 Application to the ADAR dataset

The results of the five validation experiments for the ADAR study are shown in Table 2.1. The initial sample sizes were all slightly larger at the start of the experiments, but for some predictions primers failed to amplify, so they could not be found either valid or invalid. The relatively large size of pool 1 was to allow our collaborators to investigate sequence categories including exons, introns, and intergenic regions. All validation experiments were carried out with Sanger sequencing, and the five RNA pools were prepared independently from individual biological samples.

We employed the grid-based inference procedure described above to infer the posterior distribution of the parameters of the beta distribution for this dataset. This distribution is illustrated in Figure 2.4 on transformed axes to display the mean and log standard deviation of the beta distribution itself. We observe moderate uncertainty in our knowledge of these values based on the data, supporting our decision to avoid point estimates in favor of using a diffuse prior to estimate the posterior distribution.

Replicate Pool	1	2	3	4	5
Number of Experiments	205	32	30	32	29
Successful Validations	151	17	21	24	18
Percent Successful	74%	53%	70%	75%	62%

Table 2.1: Results of 5 validation studies for the ADAR top-tier list.

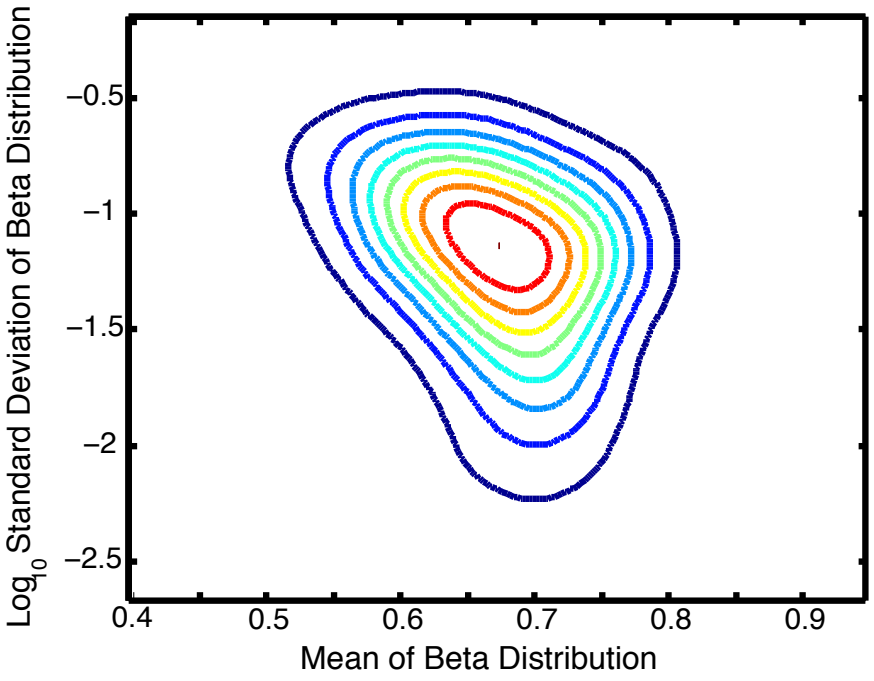

Figure 2.4: Illustration of the posterior variation in the parameters of the beta distribution for the ADAR dataset

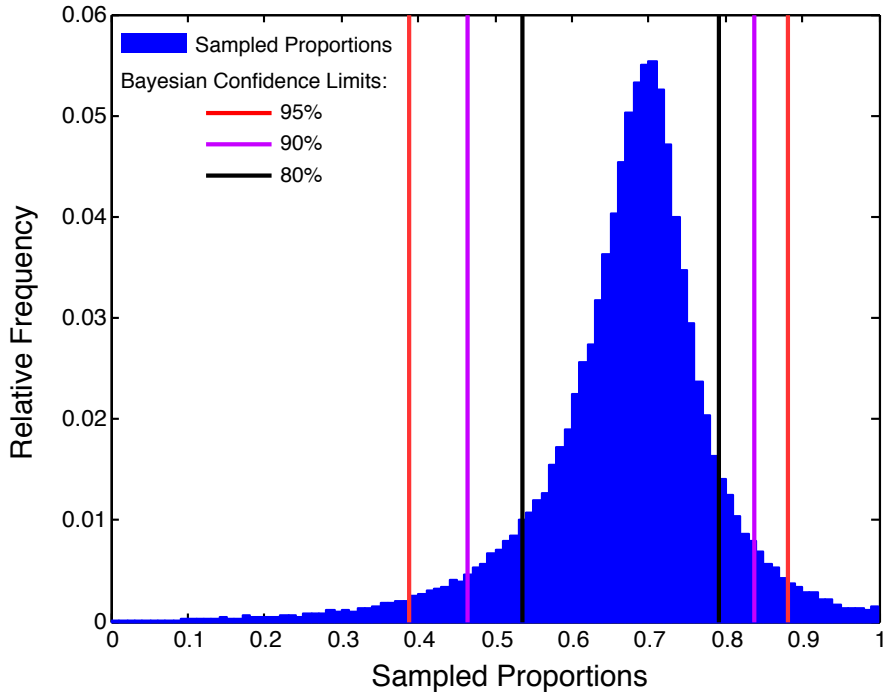

Figure 2.5: Predictive distribution of a new replicate from the ADAR dataset, with Bayesian confidence intervals

2.3.1 Reproducibility of the ADAR dataset

Using our predictive inference procedure to draw from the posterior distribution on the hyper-parameters, we estimated the distribution of proportions valid in a new replicate. The result is shown in Figure 2.5. The distribution has a mean of 67%, with the interpretation that in a new validation study following the same protocol under the same experimental conditions, we expect on average a 67% validation rate. We can also extract Bayesian credibility intervals; for our dataset, the 80%, 90%, and 95% credibility intervals are given by (0.54,0.79), (0.47,0.84), and (0.39,0.88) respectively. We expect that the lower bounds of the credibility intervals will be of particular interest to investigators. For example, using the lower bound of the 80% credibility interval, we can say that in a new validation study, the probability of validating fewer than 55% of predictions is less than 1 in 10.

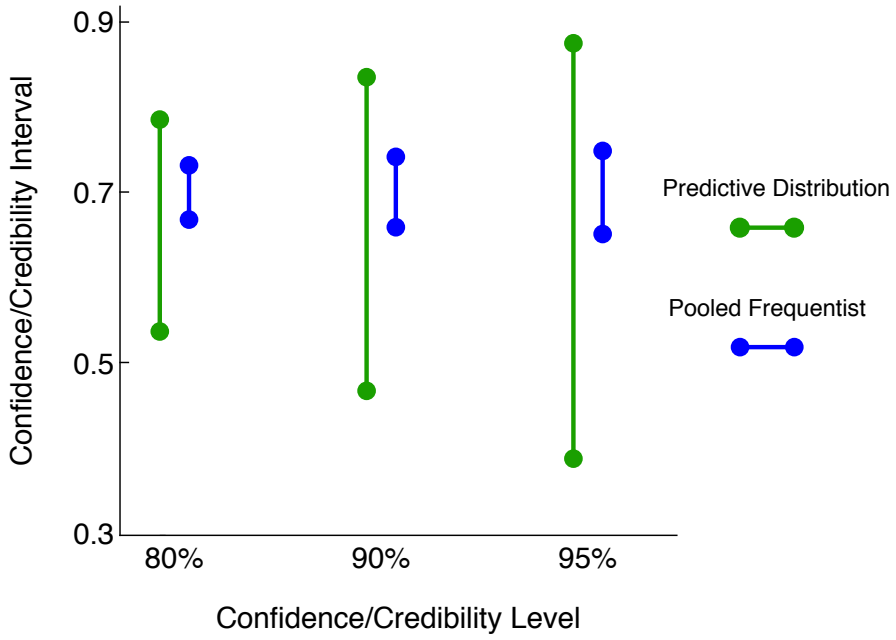

Figure 2.6: Comparison of predictive distribution credibility limits and pooled frequentist confidence limits

We can compare these limits to frequentist confidence limits that pool the experiments from all five replicates. In total, we have 231 successful validations out of 328 experiments, for a sample mean of 70.4%, and 80%, 90%, and 95% confidence intervals spanning (0.672,0.736), (0.663,0.746), and (0.655,0.753), respectively. These intervals are a great deal narrower than their predictive distribution counterparts, as shown in Figure 2.6, since they ignore the potential role of variation.

It should be emphasized here that the predictive distribution credibility limits assume strict adherence to the experimental protocol. Large variations from the protocol do not constitute legitimate replication experiments, and they could lead to results that vary more than these intervals would suggest. On the other hand, small unavoidable variations are exactly what this hierarchical model is designed to capture.

2.3.2 Comparison of Bayesian and frequentist confidence intervals for individual pools

Confidence limits are typically interpreted as an inherent characteristic of the population of predictions, but the presence of biological and sample preparation variation between replicates means that confidence limits in a single pool are actually a function not only of the population of predictions, but also of the particular characteristics of the replicate pool. Aside from providing information about reproducibility, our hierarchical model also allows us to balance the limited information that we get from each replicate with what we learn from all of the replicates at once, to learn the individual proportions valid in each pool.

To calculate hierarchical credibility limits in each replicate, we first compute the posterior distribution of α and β by the usual procedure. Because of the conjugate beta binomial relationship, it is straightforward to find the posterior distribution of p given data k and N, and a particular (α, β) prior:

$$f(p|\alpha, \beta, k, N) \sim \text{Beta}(\alpha + k, \beta + N - k) \tag{2.12}$$

To estimate the posterior distribution of p for each replicate, we draw 1,000 (α, β) pairs from the posterior distribution on α and β, then draw p from the corresponding posterior beta distribution with parameters $\alpha+k$ and $\beta+N-k$. From the estimated distribution over p, we extract the credibility limits of interest. Figure 2.7 shows these credibility limits side by side with the standard single-sample frequentist limits, each for 95% confidence. What we see is that the hierarchical credibility limits are drawn toward the posterior mean of the beta distribution, and that they are a bit narrower than their frequentist counterparts. These are both due to the fact that

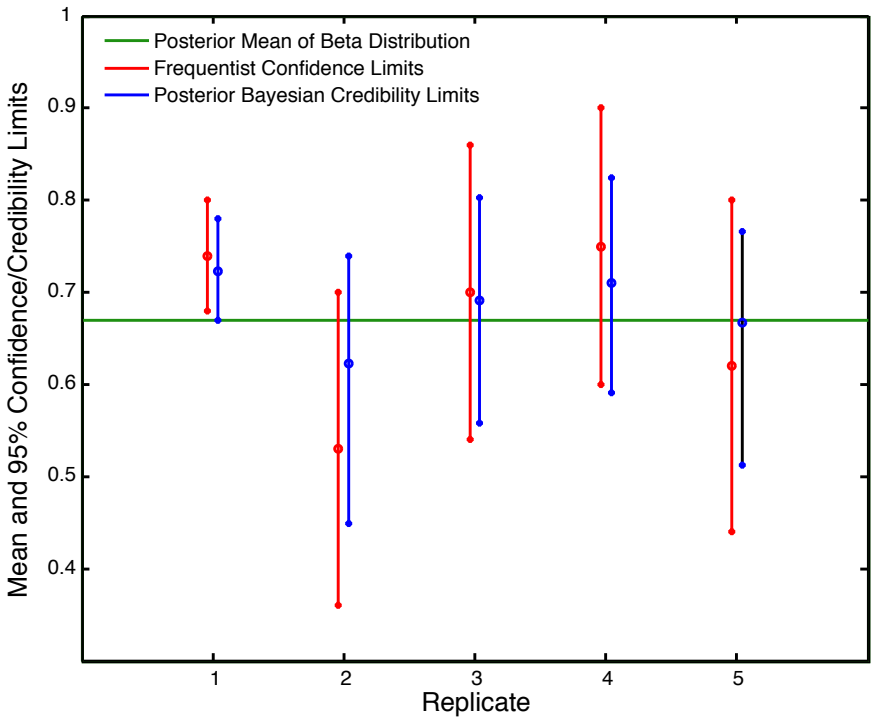

Figure 2.7: Comparison of frequentist and posterior Bayesian confidence limits for ADAR dataset

all proportions are drawn from a common distribution, and that the hierarchical model allows us to draw power from all replicates at once to describe the inherent characteristics of the population of predictions while combining this information with the replicate-specific success rates. In contrast, the frequentist limits only see what is in the particular pool, and cannot benefit from information from the other replicates.

2.3.3 Cross-validation

The number of replicate experiments required to obtain a sufficiently large sample to persuasively validate the confidence limits from our predictive distribution would have been time- and cost-prohibitive for our collaborators. We therefore undertook a cross-validation study with the 5 replicates already performed. Leaving out each

replicate in turn, we computed two sets of confidence/credibility intervals. For the first set, we pooled the remaining four replicates and used the standard normal approximation procedure to compute 80%, 90%, and 95% confidence intervals. For the second set, we used our predictive inference procedure on the remaining four replicates to infer the predictive distribution for a new replicate, then extracted 80%, 90%, and 95% credibility intervals from the predictive distribution. We then compared the proportion in the remaining replicate to these intervals. The data from this cross-validation study is in Table 2.2, and the abstract locations of each left-out proportion with respect to the cross-validation confidence intervals are illustrated in Figures 2.8 and 2.9. Using the standard frequentist procedure, the remaining proportion lands outside of the 90% confidence interval in 4 out of 5 cases, and outside of the 95% confidence interval in 3 out of 5 cases. In contrast, using the predictive inference procedure, the remaining proportion lands inside the 80% credibility interval 4 out of 5 times, which is precisely what we expect. Because of the interdependence of these intervals, we do not have statistical significance to reject the first model, but what is clear is that by accounting for potential variability, we do a better job of modeling the data than if we ignore it. We carried out a third cross-validation study in which we computed the intervals from the posterior distribution over p given pooled experiments and an uninformative prior. The credibility intervals generated were nearly identical to the frequentist pooled confidence intervals, indicating that we do a better job of modeling the data not because we use Bayesian methods per se, but because our hierarchical model allows for potential variation. Details of this third cross-validation study can be found in Appendix C.

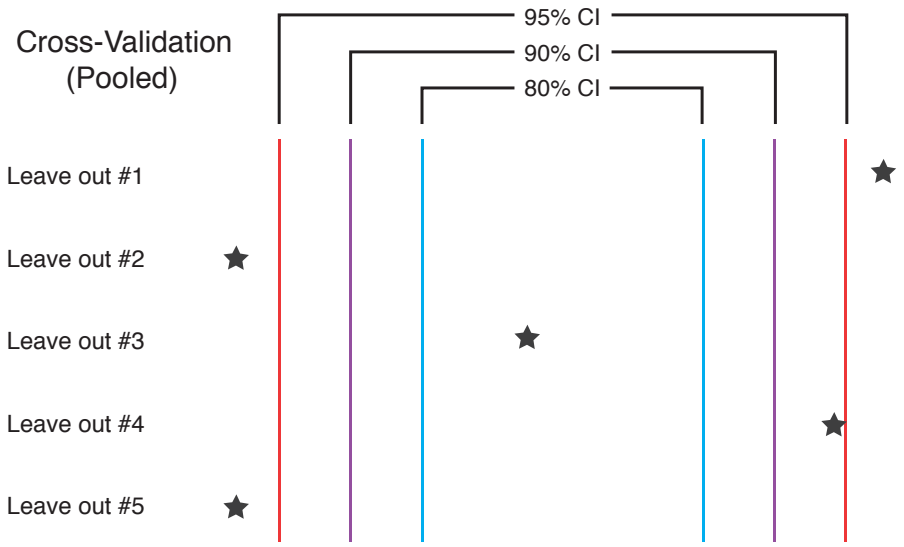

Figure 2.8: Illustration of cross-validation results using pooled frequentist confidence limits

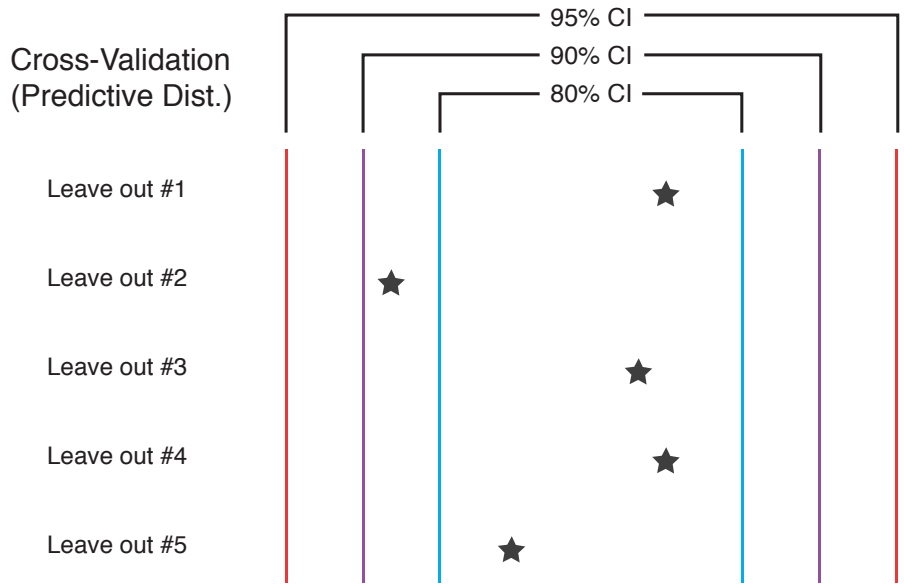

Figure 2.9: Illustration of cross-validation results using credibility limits from our hierarchical model

RNA Pool	Valid/Total	Cross Validation 80% CI		Cross Validation 90% CI		Cross Validation 95% CI	
		Predictive Dist.	Pooled	Predictive Dist.	Pooled	Predictive Dist.	Pooled
1	151/205 (74%)	(0.47,0.77)	(0.60,0.71)	(0.36,0.83)	(0.58,0.72)	(0.23,0.88)	(0.57,0.73)
2	17/32 (53%)	(0.60,0.80)	(0.69,0.76)	(0.52,0.84)	(0.68,0.77)	(0.42,0.89)	(0.67,0.77)
3	21/30 (70%)	(0.46,0.82)	(0.67,0.74)	(0.38,0.87)	(0.66,0.75)	(0.29,0.92)	(0.65,0.76)
4	24/32 (75%)	(0.47,0.82)	(0.67,0.73)	(0.39,0.87)	(0.66,0.74)	(0.26,0.92)	(0.65,0.75)
5	18/29 (62%)	(0.50,0.83)	(0.68,0.75)	(0.40,0.87)	(0.67,0.76)	(0.33,0.91)	(0.66,0.76)

Table 2.2: Cross-validation results for the ADAR dataset

2.4 Study design

If an investigator uses this model to assess the reproducibility of a set of validation experiments, there are a few favorable characteristics that he or she might hope for. A predictive distribution with a high mean, small variance, and high 10^{th} percentile would indicate that on average in a new replicate, many predictions would validate, with a high worst-case-scenario validation rate, and in general, the results of the new validation study will be relatively predictable. On the other hand, a distribution with low mean, low 10^{th} percentile, and high variance would indicate that on average the validation rate will be low, with relatively little predictability. Once a collection of predictions has been chosen, the investigator has little control over the mean of the predictive distribution, since the mean is most closely tied to the identity of the predictions in this list. However, the variance and 10^{th} percentile are variables that can be tuned by carefully designing the validation experiments.

2.4.1 Optimal design algorithm

In a study in which the investigator is willing to carry out a total of 100 validation experiments, there are many ways that this study could be designed. As was the case in our ADAR study, the time and expense involved in validation studies often come at the level of the validation experiments themselves, and comparatively little effort is needed for the creation of multiple replicate pools. In such a situation, our investigator has a good deal of flexibility in deciding how to divide a fixed number of experiments into a variable number of replicates. He or she could carry out all 100 validation experiments in a single pool, one experiment each in 100 replicate pools, or any combination in between. The former would leave a great deal of uncertainty

about inter-pool variation since we only have data from a single replicate. The latter would leave considerable uncertainty about the proportions valid in each pool since they would each have sample size 1. Our goal is to find a middle ground where we can balance these two types of uncertainty, resulting in a predictive distribution with high 10^{th} percentile and low variance.

Where this balance will be struck depends upon the expectations of the investigator with regard to a few properties of the study: the average validation rate, the expected variation between replicates, and the fraction of experiments expected to work. The first two represent a guess for the mean and standard deviation of the beta distribution, and we call these expected parameters μ_E and σ_E. The third acknowledges the fact that occasionally validation experiments can fail before a result can be obtained. In the ADAR study, this was due to the failure of primers during PCR. In these cases, particular predictions are found neither valid nor invalid, so they are excluded from downstream analysis. We assume that whether or not an experiment works is independent of whether or not it would be found to be valid in the replicate, and we let q be the fraction of experiments that are expected to work. These expectations may be based on previous experience, intuition, or even a pilot study. Finally, the investigator must decide how many total validation experiments he or she is willing to carry out, which we'll call N. The question then is how many replicates to employ, assuming an approximately even division of experiments to replicates.

Given user-defined μ_E, σ_E, q, and N, our algorithm proceeds as follows:

1. Split N into a representative subset of the possible numbers of replicates N_{rep} with $N_{exp} = \frac{N}{N_{rep}}$ experiments per replicate. See Appendix D for details.

2. For each N_{rep} (and corresponding N_{exp}):

    a) Generate samples from the hierarchical model with user-defined parameters:

        i. Generate N_{rep} values of p_i from a beta distribution with mean μ_E and standard deviation σ_E

        ii. For i = 1,...,N_{rep}, simulate experiment failure by drawing $\hat{N}_{exp,i}$ from independent binomial distributions with parameters N_{exp} and q

        iii. For i = 1,...,N_{rep}, draw the number of positive validations k_i from a binomial distribution with parameters $\hat{N}_{exp,i}$ and p_i

    b) From generated samples k_i and $\hat{N}_{exp,i}$ for all i, use the predictive inference procedure to compute the predictive distribution for a new replicate, taking note of the standard deviation and 10^{th} percentile of the distribution.

    c) Repeat from a) 3000 times, averaging the standard deviations and 10^{th} percentiles.

3. Select the value for N_{rep} that returns the most favorable predictive distribution, i.e. one with the highest 10^{th} percentile and smallest standard deviation.

An example of the output from this algorithm is shown in Figure 2.10, displaying the standard deviation and 10^{th} percentile of the predictive distribution for a varying number of replicates. Here, the expected mean and standard deviation of the beta distribution were $\mu_E = 0.6$ and $\sigma_E = 0.09$, with expected fraction of working experiments $q = 0.6$, and a total of 96 validation experiments. Since an increase in uncertainty is associated with high standard deviation, the red curve gives us an idea of the uncertainty in the predictive distribution. What we see is that uncertainty initially decreases, reflecting the fact that we are learning about inter-pool variability, reaches a minimum, and then begins to increase again, reflecting the fact that

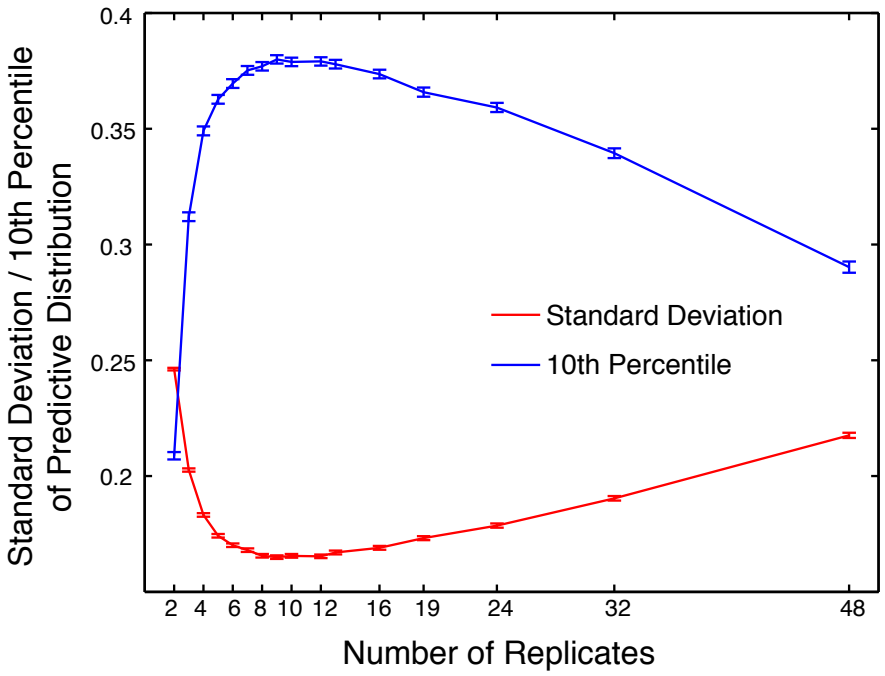

Figure 2.10: Optimal design output example.

the loss of knowledge about the proportions valid in each pool starts to outweigh our increase in knowledge about inter-pool variability. The complementary trend is embodied by the 10^{th} percentile curve in blue. Since we want a predictive distribution with low standard deviation and high 10^{th} percentile, we would recommend about 9 replicates for this study, with 10-11 experiments in each. Furthermore, we would expect that the predictive distribution inferred from the 9 replicates would have a standard deviation around 0.15, and a 10^{th} percentile around 0.38, indicating that the probability of validating fewer than 38% in a new biological replicate is only 1 in 10. In most cases, as in this one, there is a range of replicate numbers that return very similar reproducibility results.

2.4.2 Effect of input parameters

Of course, the shape of the standard deviation and 10^{th} percentile curves for the expected predictive distribution depend on the input variables: the expected mean and standard deviation of the beta distribution, the expected fraction of successful experiments, and the total number of experiments to be carried out. In Figure 2.11 we can see the effect of the input standard deviation (σ_E) on the choice of replicate number. We carried out five simulation experiments for varying σ_E keeping the other parameters fixed: $\mu_E = 0.85$, $q = 0.90$, and $N = 96$. We varied σ_E from 1% of the mean, or 0.0085 (in red) to 20% of the mean, or 0.17 (in purple). It is clear that an increase in the expected standard deviation leads to an increase in the suggested number of replicates, from about 12 for the narrowest spread to about 32 for the widest, for this set of parameters. Our interpretation is that in the presence of higher variability, more replicates are needed to ascertain the extent of that variability.

Since our procedure uses data generated from our model, another way to interpret these curves is that we want to find the number of replicates that comes closest to returning the input standard deviation. By design, our prior injects a certain amount of uncertainty into the predictive distribution, so we should see standard deviations that are somewhat higher than σ_E, an effect that will be strongest for small N. Indeed, this is what we see in these curves. For the optimal number of replicates for the red curve, we get a standard deviation around 0.12, while the input standard deviation was 0.0085. Likewise, for the optimal number of replicates in the purple curve, we get a standard deviation around 0.2, while the input was 0.17. By increasing N, we can force the predictive distribution standard deviation closer and closer to that of the distribution from which the replicates were generated. This is clear is Figure 2.12 which displays the standard deviation curves for experiments

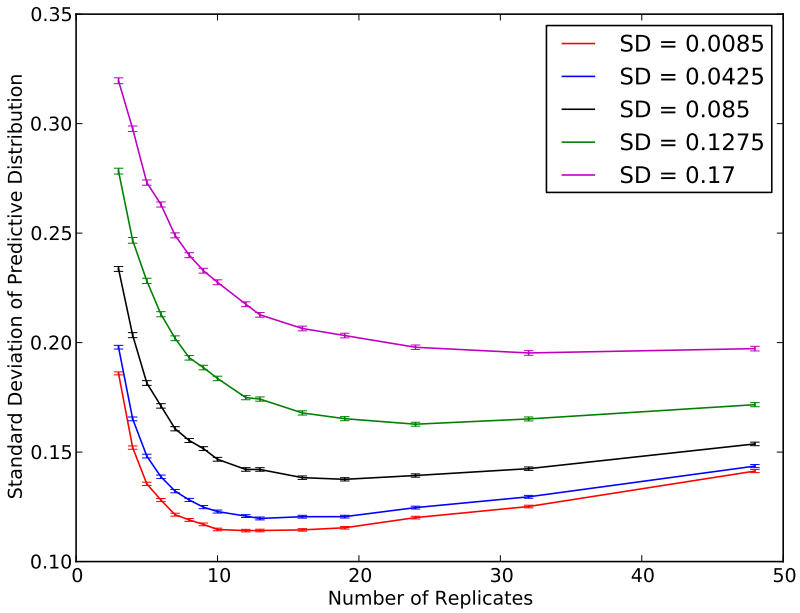

Figure 2.11: Effect of input standard deviation on optimal number of replicates for $N = 96$

with the same parameters, but with 288 total experiments instead of 96.

To the extent that the investigator can control various aspects of the validation experiments, he or she might be interested in what effect changes in μ_E, σ_E, q, and N will have on the predictive distribution, and thus on the level of reproducibility he or she will be able to report. Figure 2.13 shows qualitatively the effect that each of these parameters has on the 10^{th} percentile of the optimal predictive distribution. We varied each of the four parameters while keeping the other three constant, then for each four-tuple, we used our optimal design procedure to choose the optimal number of replicates, then extracted the predicted 10^{th} percentile for that number of replicates. Of course, higher 10^{th} percentiles are desirable, so what we can see from the figure is that a higher level of reproducibility can be generated from experiments that have large number of experiments (N), high validation rates (μ), low variation between replicates (σ), and a validation protocol that fails infrequently (q).

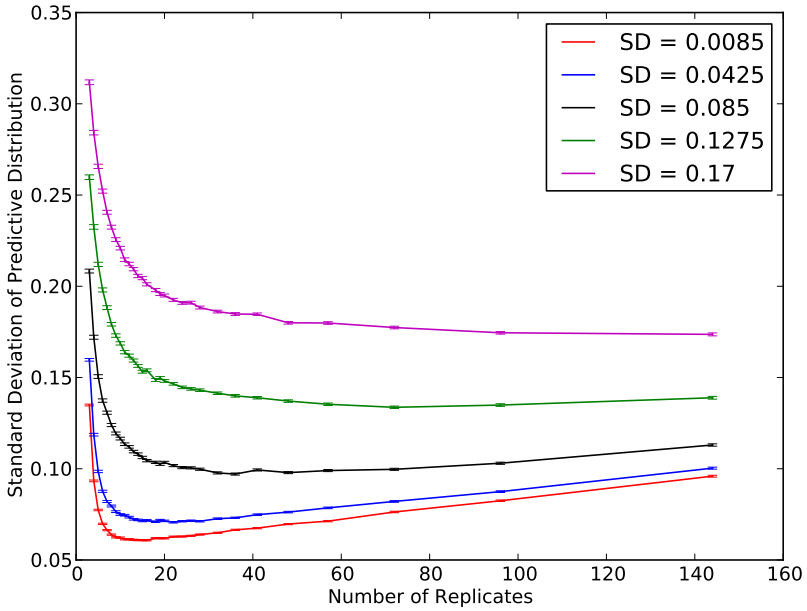

Figure 2.12: Effect of input standard deviation on optimal number of replicates for $N = 288$

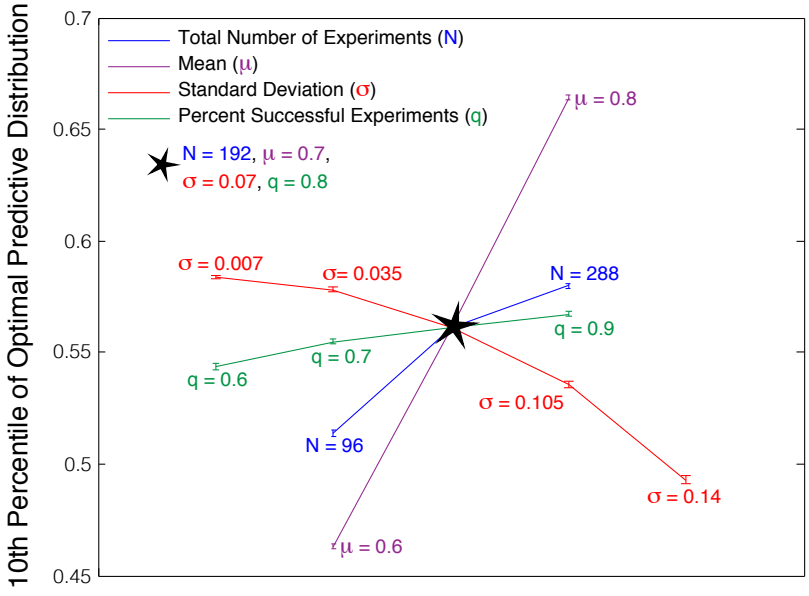

Figure 2.13: Qualitative effect of input parameters on 10^{th} percentile of optimal predictive distribution. X-axis labels are given individually for each parameter.

2.5 Model robustness

Our model makes a few assumptions. The first is that we can model the number of predictions valid in a given pool with a binomial distribution. As discussed above, the number valid actually follows a hypergeometric distribution, but since the entire pool of predictions is so large, a binomial distribution is a reasonable assumption. The second assumption we make is that we can model the proportions valid in each pool with a beta distribution. The reason for using the beta distribution is in a sense less principled, as we do not know and cannot model all of the possible sources of variation between replicates. However the beta distribution has two major benefits. First, it is the conjugate prior to the binomial distribution, which makes computation and inference through our model easier. Second, it is flexible; compared to many other distributions, the beta distribution can take on a wider range of shapes. In the end, the beta distribution will be sufficient if it is flexible enough to approximate the distributions over proportions that might exist in actual biological experiments.

To test this, we generated data from our model, replacing the beta distribution with other distributions over the interval $[0, 1]$. Given this simulated data, we used our predictive inference procedure (with the beta distribution back in place) to compute the predictive distribution for the dataset, noting its mean, standard deviation, and 10^{th} percentile. A diagram of this procedure can be found in Figure 2.14. For each probability distribution $f(p)$, we simulated data from the $f(p)$-binomial model for increasing numbers of replicates, each with 100 validation experiments. While this quickly starts to exceed the number of experiments that an investigator would likely do, we can get an idea of whether the statistics of interest from the predictive distribution converge to the true values for the distribution $f(p)$.

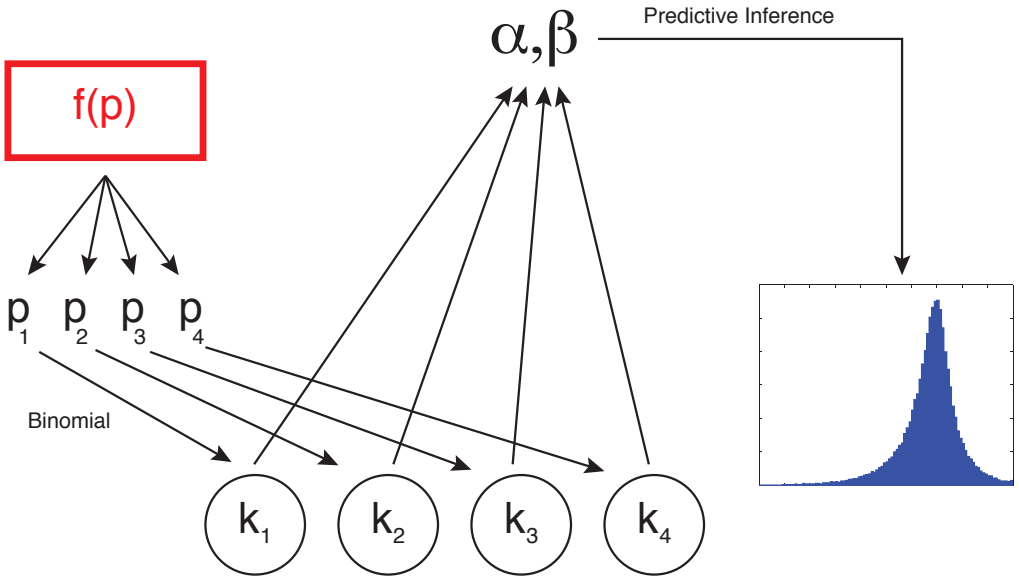

Figure 2.14: Diagram of test for model robustness

For comparison, we first generated data from a beta-binomial distribution. Since this is the distribution we use in our inference procedure, we expect the mean, standard deviation, and 10^{th} percentile to converge fairly quickly (recall that we do not expect immediate convergence, since our prior imposes a dose of uncertainty). Including the beta distribution, we considered the following distributions for $f(p)$:

Model 0 p drawn from a beta distribution with $\alpha = 3$, $\beta = 2$

Model 1 p drawn from a normal distribution with $\mu = 0.7$, $\sigma = 0.08$ (truncated to the interval $[0, 1]$)

Model 2 $1 - p$ drawn from an exponential distribution with $\mu = 1$ (truncated to the interval $[0, 1]$)

Model 3 p drawn uniformly from $[0.5, 1]$

Model 4 p drawn from a triangular distribution with min = 0.5, mode = 0.8, max = 0.95

Model 5 p drawn from a Gaussian mixture, 60% $N(\mu = 0.8, \sigma = 0.05)$, 40% $N(\mu = 0.5, \sigma = 0.15)$

For each of these models, we compared the true population mean, standard deviation, and 10^{th} percentile to their predictive distribution counterparts averaged over 100 simulated datasets for each number of replicates. The results are displayed in Figure 2.15. For model 0, the beta distribution, we do indeed see what we expect; each statistic of the predictive distribution (represented with solid lines) converges very quickly to the population statistics of the beta distribution (represented with dotted lines). For our other unimodal distributions (models 1, 2, and 4), the predictive distribution quickly becomes quite accurate as the data accumulates, and also seems to be converging to the true characteristics of each distribution $f(p)$. The predictive distribution is slightly less accurate when $f(p)$ is either a bimodal or piecewise distribution (models 3 and 5), reflecting the inability of the beta distribution to approximate these shapes. Even so, our inference procedure does a respectable job of estimating the mean and standard deviation in these two cases. We believe that in most biological experiments, a bimodal or multimodal distribution is unlikely, especially for experiments using highly inbred model organisms. Thus, the assumptions of our model should be appropriate for most applications.

2.6 Discussion

The need for reproducibility in scientific research has always been central, but has only recently become a major focus of the greater scientific community. Here, we

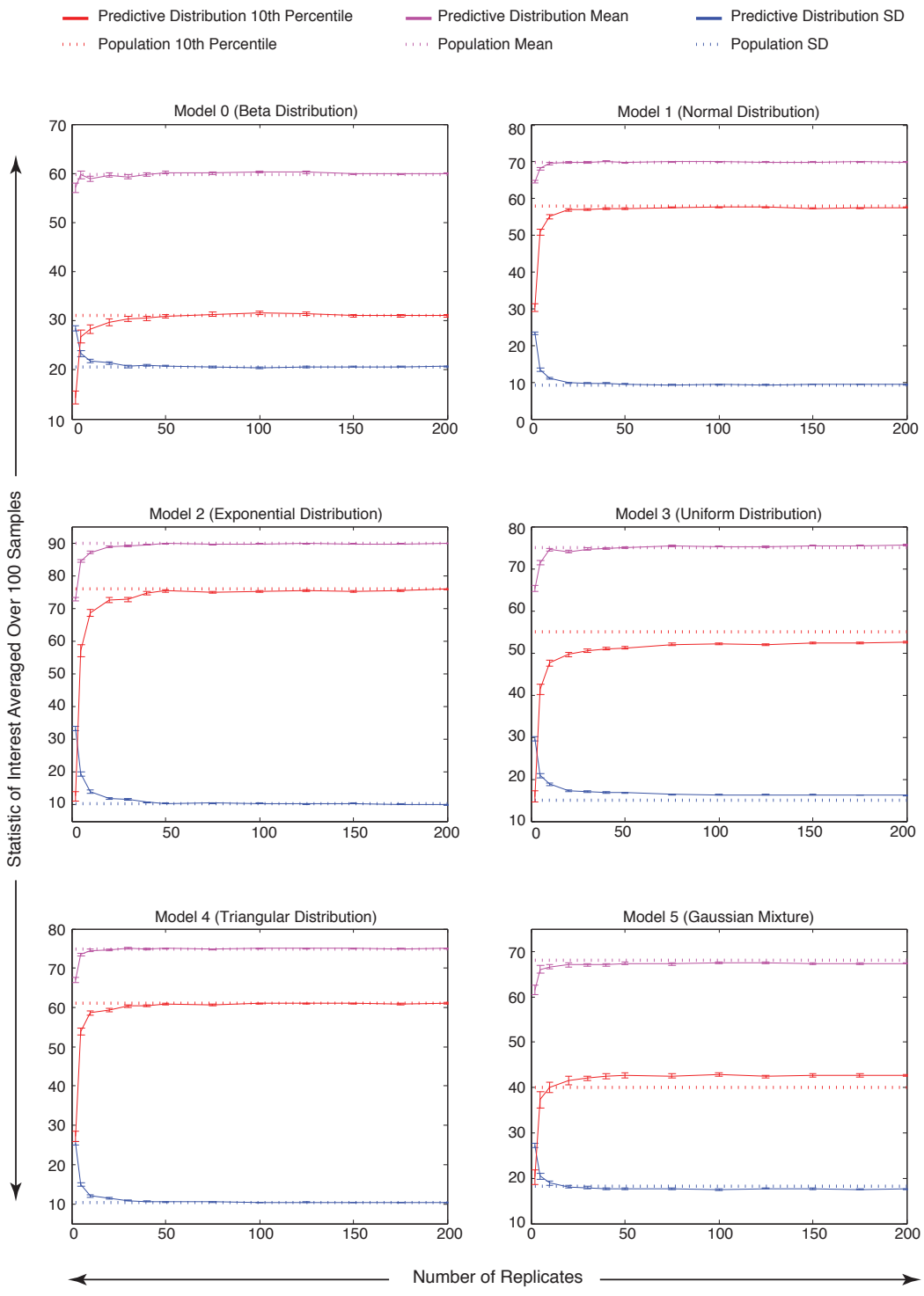

Figure 2.15: Results of test for model robustness

present a procedure that addresses these issues in the context of high-throughput studies like the ADAR study described here, where thousands of predictions are made, and only a relatively small fraction can be validated. We studied closely the example of ADAR-mediated editing sites, but the method is generalizable and could just as easily apply to any of the high-throughput site prediction studies described earlier. Studies of this kind have become more frequent with the advent of high-throughput technologies like Illumina and large collaborations like ENCODE, modENCODE, and the 1000 Genomes Project. Most studies already have their own schemes for validation, which they carry out with varying levels of statistical rigor, typically validating a select few, and almost never providing confidence limits or utilizing replicates. The authors and any investigators hoping to follow up on these studies would all benefit from a carefully designed validation study using statistically random samples in biological replicates to assess reproducibility.

As the accuracy of technology inevitably grows, it may be tempting for investigators to assume a single replicate is sufficient to address reproducibility, implicitly assuming that technical variation is the only kind of variation that matters. However, even with perfect technology, multiple biological replicates are still necessary to assess the reproducibility of a set of results. It is worth noting that each of our validation replicates was performed in a pool of 10 flies each. We expect that biological replicates will be even more crucial in studies where replicates consist of individual organisms.

In our analysis of ADAR-mediated editing data, we found that the 95% confidence intervals for individual replicates showed substantial variation. This was not unexpected, given the documented variation in ADAR activity in individual flies [62] and observations about the effects of experimental conditions described earlier, and it underscored the need for statistical tools that can address the effect of such varia-

tion on the reproducibility of results. Our software predicted that a new experiment using the same protocol under the same experimental conditions would validate on average 67% of sites, and that 90% of the time, the fraction validated would be at least 55%. We emphasize the 10^{th} percentile because it represents a worst-case scenario for a future experiment. Of course, because these results are affected by our choice of the diffuse prior on α and β, the intervals we generate may be too conservative in the case where more is known about these parameters *a priori*.

There is a technical limitation to our model that arises from the statistical formulation. The predictive distribution we describe is well defined everywhere except in the precise circumstance in which every replicate pool has a validation rate of either 100% or 0%, in which case our improper prior leads to an improper posterior distribution. We consider such extremes to be very unlikely in most validation studies; however, this limitation occasionally comes to bear in experimental design, where we simulate thousands of replicates. This affects almost all simulations with μ_E above 95%, and most simulations with μ_E around 90% with large σ_E. The effect is otherwise negligible.

We rely on numerical approximations in three places: during predictive inference, we compute the posterior distribution of the hyperparameters over a 200x200 grid, as there is no closed-form expression for computing it directly, then we sample from the grid to approximate the predictive distribution. In our optimal study design procedure, we rely on a sampling approximation to find the average mean, standard deviation, and 10^{th} percentile of the population of predictive distributions resulting from a set of input parameters. For the first two approximations, computing the posterior distribution of α and β over a grid and sampling to obtain the predictive distribution, we follow the procedure outlined in Gelman *et al.* [35], computing over a grid dense enough to capture the important features of the distribution, then we

sample a large number of points (1,000) to approximate the predictive distribution. We found that neither varying the grid density nor increasing the number of samples noticeably changed the results. For study design, the default number of samples is 3,000, which tends to create very narrow error bars.

One aspect of the optimal design procedure that we have not investigated as of yet is the problem of model mis-specification. The recommended number of replicates is based on the expectations of the investigator, but if these expectations turn out to be inaccurate, then the study may proceed with a sub-optimal number of replicates leading to a less desirable predictive distribution than if the true optimal number of replicates had been employed. In these cases, it would be useful to have a measure of the effect of errors in each of the parameters μ_E, σ_E, and q. For example, it may turn out that it is safer to underestimate the mean and overestimate the standard deviation, or the opposite could be true. We also have not investigated the effect of dividing experiments unevenly between replicates in our optimal design procedure. Our intuition is that an even division will be optimal because the loss of information in a replicate with few experiments will be greater than information gained in a replicate with a larger number of experiments, but this is unproven.

Finally, it should be noted that just as the results of any study should only be generalized to the population from which the data are drawn, the results of any reproducibility analysis can only be generalized to the population to which the replicates belong. We assume that follow-up validation studies will follow the original protocol under the same experimental conditions as the original experiments. To the extent that this is not possible, the credibility intervals that we report may be too narrow to accurately reflect the population of follow-up validation studies performed by other investigators in other laboratories. Therefore, care must be exercised in how claims of reproducibility are made, and authors should be sure to specify the

population to which their results generalize. In some cases, large collaborations between laboratories (such as those associated with modENCODE) will be able to carry out replicates representing a larger portion of the possible variability, and will be able to make even stronger claims about the reproducibility of their findings.

Tools for the assessment of reproducibility given replicates and for the design of replicate validation experiments in the context of high-throughput studies with large number of predictions are available as downloadable code and as a web server at http://ccmbweb.ccv.brown.edu/reproducibility.html. The optimal design procedure is rather time-consuming, so we have also pre-computed optimal design curves and recommended replicate numbers for a wide range of input parameters.

CHAPTER **Three**

Identifying Genetic Variation in Laboratory Organisms

3.1 Background and Motivation

As next-generation sequencing technologies such as Illumina become less costly and more widely utilized, the potential for many applications including cancer genomics [50], miRNA identification [79], and gene expression analysis [82] are being realized. Many of these applications depend on accurate alignment of reads to a reference genome, and reliable SNP calling. In some cases, identifying SNPs may be the end goal, as for projects like dbSNP [105], and in others, it may be crucial for avoiding false-positives in later analysis. This is especially true in studies of discrete changes like RNA editing, which alters single nucleotides. It is likely that a failure to rigorously account for SNPs has led to false positives in many previous editing studies; in fact, a preliminary look at the results of the editing study described in Chapter 2 has revealed a false discovery rate of about 15% due to polymorphisms. The model

described here, with the accompanying algorithm, allows us to simultaneously infer the hidden DNA and RNA sequences of our flies while learning the model parameters. We first restrict the model to SNP discovery using genomic data, then extend it to the simultaneous discovery of both SNPs and editing sites, using both genomic and transcriptomic data.

3.1.1 Identification of SNPs

There has been considerable work on the problem of SNP-calling and the related problem of genotype-calling from alignments of high-throughput sequence data, in the context of humans or other individually sequenced organisms. Genotype calling differs from SNP calling in that the former aims to determine the underlying genotype of a particular individual at a given site, while the latter aims to identify the polymorphic sites in a population, however, the two problems go hand in hand when the data consist of sequences from individual organisms.

Some genotype-calling algorithms use simple cutoff methods to determine heterozygous and homozygous sites. In one method, if the proportion of mapped reads containing the alternative (non-reference) allele lies between 20% and 80%, the site is called heterozygous, while a proportion above 80% or below 20% indicates a homozygous genotype for the alternative and reference alleles respectively [93]. An alternative of this model varies these cutoffs based on the read depth at the position [43]. A more popular approach uses a Bayesian framework, in which the posterior probability of the possible genotypes is inferred, and the genotype with the highest probability is reported. This involves the computation of the likelihood of each genotype (homozygous reference, heterozygous, and homozygous alternative), based on a model of sequence errors, as well as prior distributions on the prevalence of SNPs.

Most algorithms derive error rates from quality scores reported by the sequencing technology, either directly [31, 65, 70] or through a recalibration [101]. One algorithm, SeqEM [47], uses an expectation maximization (EM) algorithm at each site to learn the site-specific error rate and genotype frequencies and infer a genotype. An EM algorithm is also used by Heng Li [139] to learn the allele frequency at a particular site in a polyploid organism, assuming a simple error model.

Our interest has a slightly different emphasis; we want to identify SNPs in a population of organisms sequenced in a single pool, including positions that differ from the reference for the whole population, as well as unfixed polymorphisms that are present in the population at some level between 0% and 100%. This pooled sequencing is more common when dealing with model organisms like *Drosophila melanogaster*, since the identity of individual genomes matters less than the characteristics of the population as a whole. We use a similar Bayesian framework to those described above for inferring polymorphic positions, with the added complication that we no longer have only three possible levels of polymorphism (0%,50%, and 100% for homozygous reference, heterozygous, and homozygous alternative, respectively), since polymorphisms can be present at any level, and we do not have individual-level data. This is an especially important aspect of our model, since unfixed polymorphisms have the potential to be even more problematic than fixed polymorphisms because they are trickier to deal with and harder to account for with heuristic filters. We also rigorously address errors by learning context-specific error rates simultaneously alongside the polymorphism rate and the probability distribution over polymorphism levels with an algorithm that takes advantage of the information from all genomic sites.

Finally, we note that a lot of work has gone into cataloging SNPs in *Drosophila melanogaster*, mainly for the purpose of facilitating forward genetic screens and ex-

amining the general architecture of variation [46, 33, 98, 14]. While these will be very useful for those purposes, they do not help an investigator interested in the actual genome sequences of his own flies. Unlike some other organisms where individuals or stocks can be frozen down to halt evolutionary time, no such possibility exists for flies. Time marches on, and with it come new mutations, carrying the laboratory stocks further and further from the reference genome. We present this algorithm as a way to discover the genomes of a batch of laboratory flies (or other organisms) that, through a combination of time, population bottlenecks, genetic crosses, and other factors, may vary substantially not only from the reference genome, but also from other batches in the same lab or in other labs.

3.1.2 Previous studies identifying editing sites

Considerable work, including the companion paper from the previous chapter, has focused on the identification of editing sites in various model organisms using high-throughput sequencing techniques. The issue of SNPs and the possibility for false positives is almost always acknowledged, and various attempts to control them are made in the process of calling editing sites.

A fairly common technique is to rely wholly on previous knowledge about genetic variation. In human-derived cell lines, Bahn *et al.* [61] only consider sites previously determined by an independent group [88] to be homozygous, and in a survey of 15 mouse strains, Danecek *et al.* also begin with "detailed knowledge of genomic variation" from a previous study [110]. In a comprehensive study of editing sites in a single male Han Chinese individual, Peng *et al.* [117] relies on the previously known genotype of the individual. In this last case, there should be little concern about missing genetic variation apart from somatic variation, but in the first two, there is

always a chance of polymorphisms between the reference organisms or cells and the ones used for seeking out editing sites.

Another approach is to rely only partly on known polymorphisms. In one study of editing sites in a human-derived cell line, the investigators used DNA and RNA data to predict editing sites by first discarding known SNPs, then requiring that the DNA data be homozygous [53]. In a second study by the same authors, RNA data from multiple individuals was used to call edits without DNA data via two different methods, each of which required the filtering of known SNPs, followed by a step to distinguish rare genetic variants from editing events by assuming that editing events would be present in multiple individuals, in contrast to rare SNPs, which would not [54].

Other studies do not rely at all on known SNPs. This is more often the case in studies of *Drosophila*, because of the relatively unregulated way in which fly stocks are obtained and propagated, along with the short generation time. A modENCODE study looking at various aspects of the fly transcriptome attempted to control for SNPs by requiring that for the RNA reads overlapping the site of interest, at least 100 exactly matched the genome [39]. While this will control for fixed polymorphisms well, it will likely miss-call unfixed polymorphisms as editing events. Another study, looking at co-transcriptional editing events, filtered out potential polymorphisms by sequencing the genome at 39X coverage, and requiring that all reads exactly match the genome [160].

When it comes to calling editing sites, most of these studies employ a series of heuristic filters on their data, including coverage requirements, minimum frequency requirements, and mapping filters. Only two studies include a rigorous mathematical formulation coupled with such filters. Bahn *et al.* [61] compute a log likelihood ratio

at each genomic position that compares the probability of the RNA data given an editing event to the probability of the data given no editing event. If this ratio is larger than 2, they predict an editing event at that position. Peng *et al.* [117] perform an initial "variant calling" step that uses techniques similar to the SOAPsnp software package [101], one of the Bayesian formulation approaches to SNP calling described in the previous section. Multiple studies report low overlap with others [160, 117, 51], likely in part because each study uses its own set of heuristics to control for genetic variation and call editing events.

We present here a mathematical model of the hidden DNA and RNA sequences of a pooled population of organisms (*Drosophila*, in our case), and an algorithm that allows us to simultaneously infer the presence of genetic polymorphisms and editing sites in a mathematically rigorous way that does not depend on heuristic filters to predict these events. This has the added benefit of producing predictions with clear probabilistic interpretations (i.e. "a 96% chance of a T→G polymorphism at position x").

3.1.3 Expectation Maximization Algorithm

Our algorithm uses the expectation maximization (EM) algorithm, an iterative method for obtaining maximum likelihood estimates of population parameters in the presence of missing data or hidden variables. This algorithm was introduced in 1977 by Dempster, Laird, and Rubin [16], and today is widely used in many fields including genomics, linguistics, and computer vision. Do and Batzoglou [21] give an excellent description of EM and its common uses in biology. Given a set of unknown population parameters θ, data X, and hidden (or "latent") variables Z, EM works iteratively to indirectly improve $\log L(\theta|X) = P(X|\theta)$ by choosing new parameters

$\theta^{(t+1)}$ to maximize the following:

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}}\{\log L(\theta|X,Z)\} = \mathbb{E}_{Z|X,\theta^{(t)}}\{P(X,Z|\theta)\} \tag{3.1}$$

where $\theta^{(t)}$ contains the current parameter estimates. The expected value is taken with respect to the posterior distribution over the hidden parameters Z given the current parameter estimates and the data, so in practice the algorithm proceeds by computing this posterior distribution, then using the distribution to compute expected sufficient statistics from which we can calculate estimates of θ. The log likelihood of the parameters given the data is guaranteed to increase with each iteration, and the algorithm iterates until the change in the log likelihood of the parameters given the data and hidden variables drops below some threshold ϵ (or until the parameter estimates themselves stop changing). If the surface we are optimizing over is convex, then we are guaranteed to converge to the maximum likelihood estimate for θ. In general, the surface may not be convex, so typically the algorithm is initialized at multiple starting points. A diagram of the EM algorithm can be found in Figure 3.1. Although the primary purpose of EM is to find the estimates for θ, it also allows us to infer the hidden states Z after convergence.

3.2 Illumina output and quality scores

3.2.1 The Data

We used the Illumina sequencing platform to generate single-end reads of length 50 from two laboratory fly stocks. The first line, LoxP, is a control line generated by homologous recombination, and the second, ADAR0, is a mutant fly stock that

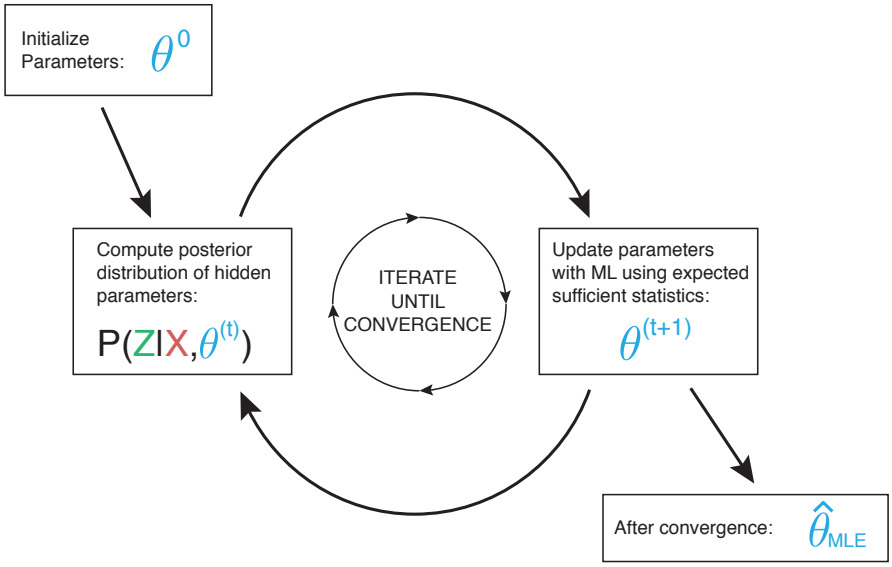

Figure 3.1: Diagram of the Expectation Maximization (EM) algorithm

cannot perform RNA editing, and was generated at the same time, in the same laboratory, and through the same methods as the LoxP line. For each experiment, DNA and RNA were isolated from the heads of 100 inbred flies, for two biological replicates each, run in the same sequencing lane. The output from Illumina varied by experiment, but the number of reads typically fell between 75 and 150 million reads for the biological replicates combined. We used BWA [140] to align the reads from Illumina to the *Drosophila melanogaster* reference sequence (April 2006 assembly) with default settings. For the purposes of this study, we restricted the alignment to unique regions of the genome. Repetitive regions complicate this mapping step, as there are often a large number of possible mappings for a read derived from one such region.

BWA aligns each read to either to the reference (or "Watson") strand, or the opposite complementary (or "Crick") strand, and it reports Illumina quality scores for each base call. The quality scores Q range from 0 to 41, and are meant to use characteristics of the raw signal to provide a measure of the confidence in the base

call. The scores are calibrated, using reads mapped to known references, to be a log transformation of the error rate e: $Q = -10 \log_{10} e$. The inverse function gives the estimated error rate in terms of the quality score: $e = 10^{-\frac{Q}{10}}$.

3.2.2 PhiX Preliminary Study

Many applications take as given the error rates derived from quality scores reported by Illumina, however errors in Illumina sequencing have been shown to be affected by more than just raw signal characteristics: in particular, it has been suggested that the sequence context immediately preceding the sequenced base has an effect on error rates [59]. We carried out a preliminary study using phiX, a single-stranded DNA virus run as a control alongside other samples on the Illumina machinery, to look at this context effect and build our own error model. PhiX has a very short genome (5386 nucleotides), making it easy to map reads to the reference and identify any polymorphisms. Furthermore, the read coverage at any given nucleotide is extremely high, allowing for precise estimates of read errors at each position. We found that the errors did indeed vary by sequence context, defined as the three-letter sequence ending in the nucleotide of interest. We also found that by restricting our dataset to reads in which no base call had a quality score below 30, the effects of other contributing factors from the literature were greatly reduced, as described in Section 3.7.5. Furthermore, we could still maintain about 70% of our dataset, as shown in Figure 3.2.

Through an investigation of multiple independent PhiX runs, we also noticed that the error rates varied widely. In some runs, average error rates were more than twice as high as in others, with error rates for some sequence contexts varying even more, as shown in Figure 3.3. This highlighted the need for a method that could

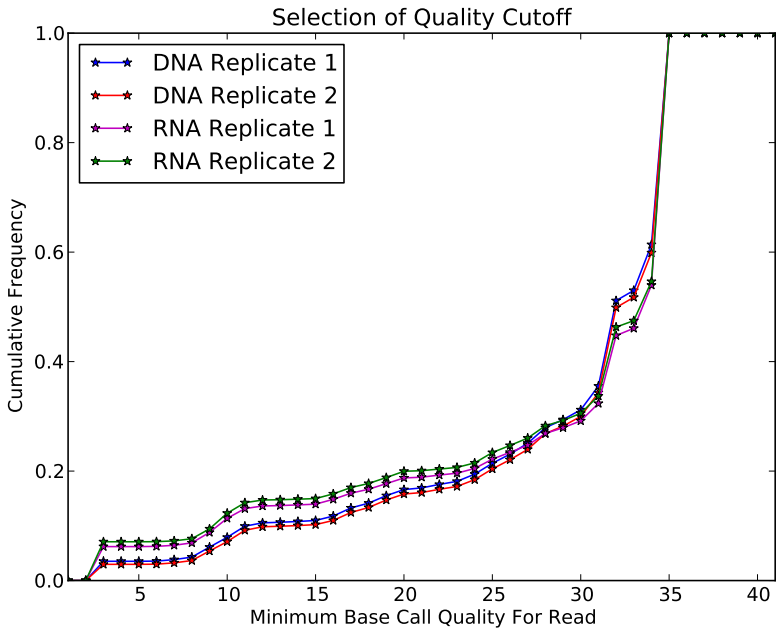

Figure 3.2: Cumulative distribution of minimum quality score for four samples in two runs

learn the run-specific error rates while searching for polymorphisms and editing sites.

3.3 The Model

Our goal is to identify the polymorphisms present between the reference sequence and our laboratory flies, and the set of editing events present between the DNA and RNA of our flies. We use the word "polymorphism" instead of "SNP" to highlight our focus on the population as a whole, and the fact that we are not concerned with the alleles present in any particular fly. We have two sets of hidden variables — the true genome and the true RNA of our flies. We also have two sets of data, the reads from the DNA and the RNA. Finally, we have the reference sequence, which is known. We have constructed a generative model, which is illustrated at the highest

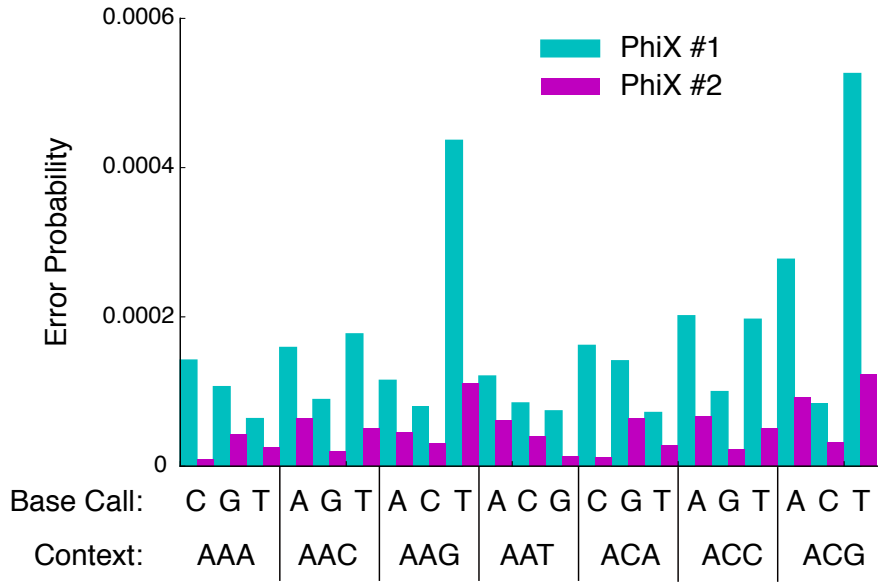

Figure 3.3: Variation in context-specific error rates across PhiX runs for a subset of contexts

level in Figure 3.4. The reference emits the hidden DNA sequence of our flies, with possible polymorphisms. The true DNA sequence then emits the true RNA sequence, with possible editing events. Finally, the true DNA and RNA sequences emit the DNA and RNA reads respectively, with possible machine errors.

For the purposes of this study, in which we focus on non-repetitive (unique) regions of the genome, we take the alignment of reads to the reference as given. Figure 3.5 illustrates this alignment step. In this example, there is an A in the reference genome at the position of interest, and we have three reads each mapping to the Watson and Crick strands of the reference. At most positions, the bases in the reads match the corresponding bases in the reference, but at the highlighted position we see a few departures; half of the base calls match the reference, and half do not. This could mean either that the sequencing machinery made a few mistakes, or that some of our flies actually differ at this position from the reference, indicating a polymorphism. If these reads are from RNA, this could be evidence of an editing

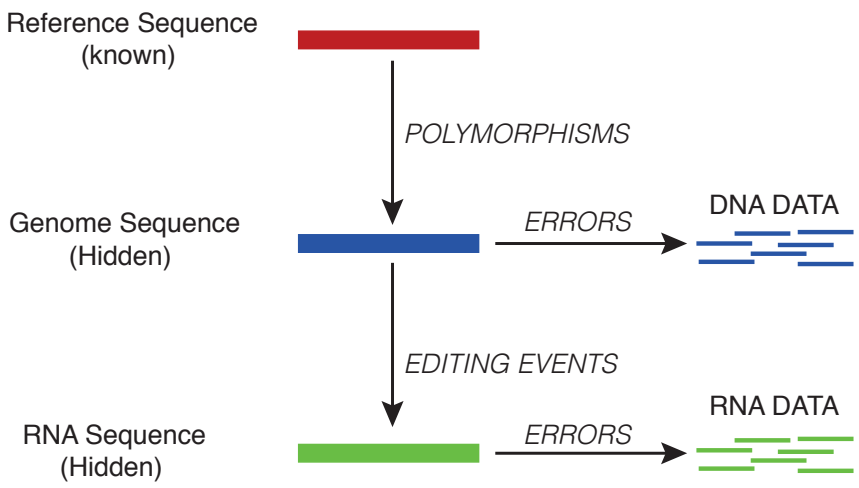

Figure 3.4: High-level model diagram

event. Our goal is to determine which of these explanations is the truth.

Figure 3.5 also illustrates an important aspect of our error model. We model and learn errors based on the three-letter sequence context ending in the position of interest, which depends on which strand the read maps to. For reads mapping to the forward (Watson) strand, the molecule was synthesized from left to right, resulting in sequence context GGA. For reads mapping to the reverse (Crick) strand, the molecule was synthesized from right to left, so the sequence context is AGT. This leads to a complicated dependency structure in our model, as illustrated in Figure 3.6. Since we may have nucleotides from both strands at any given position, the collection DNA bases called at position i depends not only on the DNA at position i, but also on the (hidden) DNA at positions $i-2$, $i-1$, $i+1$, and $i+2$. The same relationship exists for the RNA base calls. This bidirectional dependence makes inference quite difficult, so we impose an assumption to ease computation: that the first two nucleotides of the context for the context-specific errors are drawn from the reference nucleotides, which are known. This assumption, as illustrated in Figure 3.7, greatly reduces the computational complexity, since the base calls at each position are now conditionally

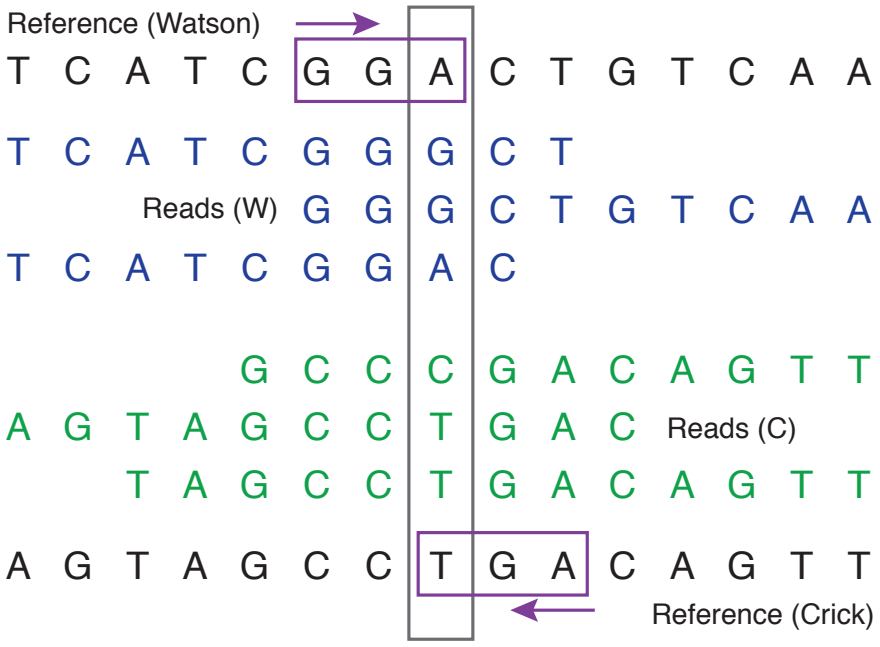

Figure 3.5: Illustration of read mapping and sequence context

independent from base calls at other positions given the reference. Furthermore, we do not expect this assumption to have a large effect since SNPs are relatively rare.

3.4 Restricted Model

We begin by restricting our model to the inference of polymorphisms only. This requires genomic data but no transcriptomic data, and we only need to infer one set of hidden variables— the true DNA sequences of our laboratory flies. The algorithm we describe here would be of great use as a preliminary step in a studies of precise, small-scale phenomena like RNA editing, in which the accuracy of downstream predictions depends on the thorough identification of genomic differences between the reference and the population being studied. We begin by defining the model variables and parameters and the structure of the generative model in this context,

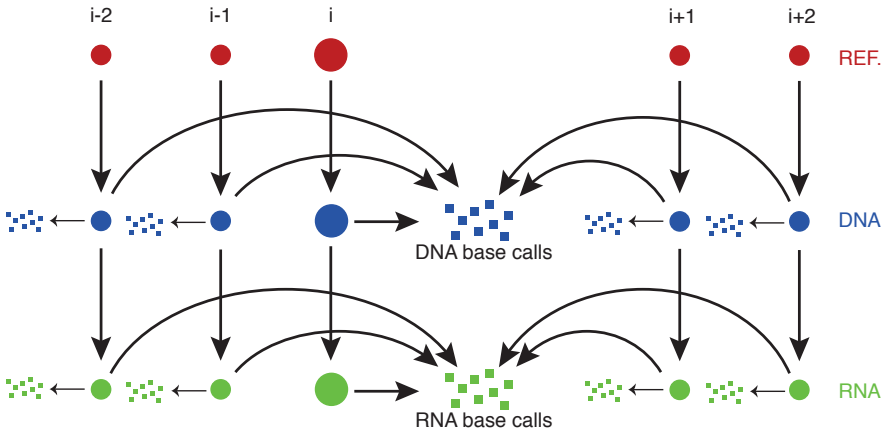

Figure 3.6: Interdependence in the model

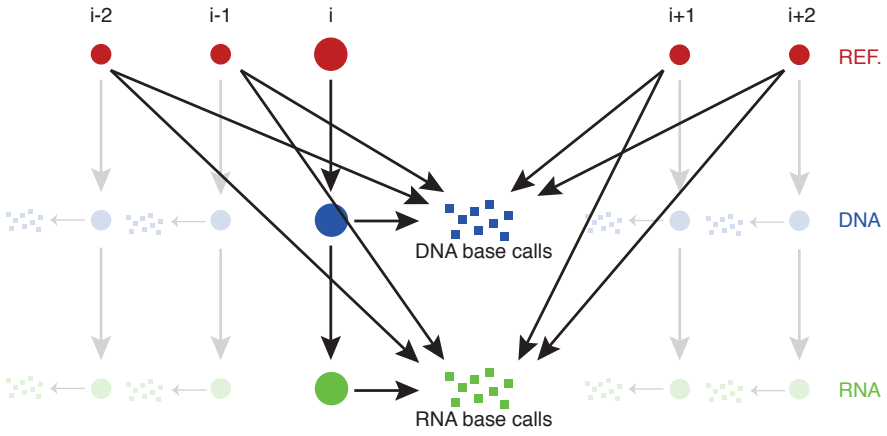

Figure 3.7: Conditional independence assumption for model simplification

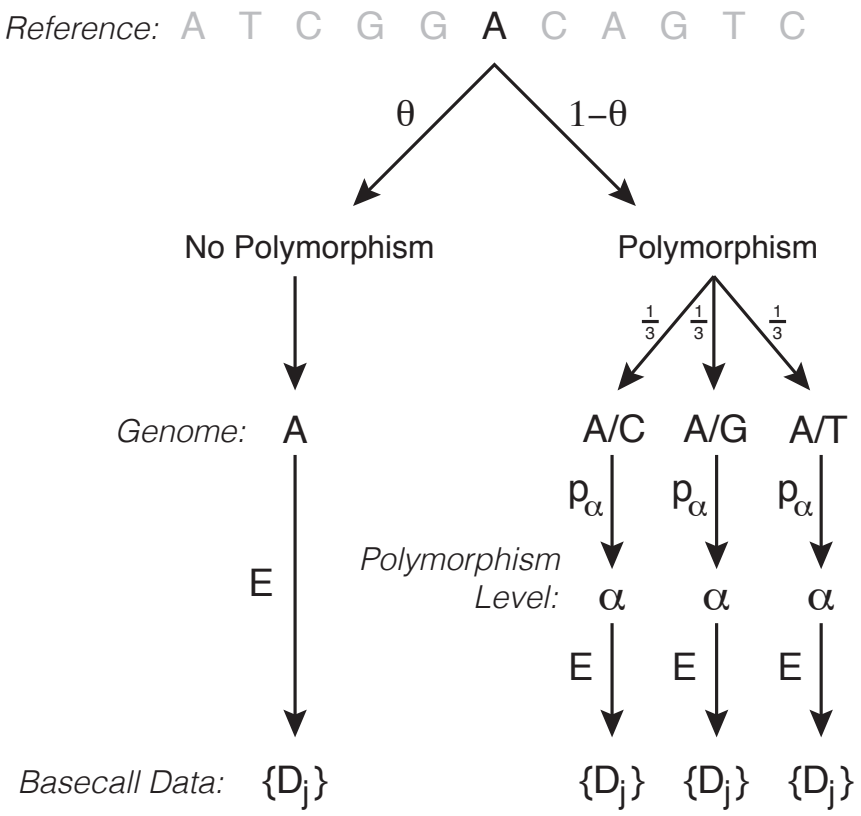

Figure 3.8: Structure of the polymorphism model

which is illustrated in Figure 3.8.

3.4.1 Model Parameters

Given an alignment of the reads to the reference genome, at each genomic position i, the reference nucleotide F_i emits a hidden state G_i, the genomic identity of the laboratory population at that position. There are two possibilities for G_i: most of the time the genome will match the reference, but occasionally the genomes of the laboratory flies will differ due to random mutation. These events are polymorphisms, which can either be fixed or unfixed in the population. Fixed polymorphisms are those that entered the population due to a mutation event in a single fly, then due ei-

ther to genetic drift or selective pressure, came to be present in the entire population at the current time. For a position with an unfixed polymorphism, the population as a whole carries both the reference nucleotide and a second, non-reference nucleotide in some proportion, sometimes referred to as the "allele frequency." Since flies are diploid, some may be homozygous for either the reference or alternative allele, and some may be heterozygous. We assume that positions are at most bi-allelic in the laboratory population and the reference combined, so that for an unfixed polymorphism in the lab population, one nucleotide must match the reference. This assumption is justified by the low probability of two independent mutation events in the exact same location.

We let θ represent the proportion of sites that match the genome, and we assume *a priori* that each of the non-reference nucleotides is equally likely, so that the prior probability of a polymorphism of one of the three types (A/C,A/G, or A/T in the example in the figure), whether fixed or unfixed, is $\frac{(1-\theta)}{3}$. At a polymorphic site we use α to represent the fraction of the population that differs from the reference. For example, $\alpha = 0.8$ corresponds to a site that is 20% reference and 80% alternative, while $\alpha = 1$ corresponds to a fixed polymorphism. We use p_α to represent the probabilities of each level of polymorphism for $\alpha = 0.1, 0.2, ..., 1$. We can think of the model up to this point as the prior model, which gives us the prior probability of any given genotype.

The hidden genomic identity G_i then emits all of the base calls $\{X_j^i\}_{j=1}^{M_i}$ that map to the corresponding reference position, where M_i is the number of base calls mapped to position i. In the example in Figure 3.5, the dataset at position i would be given by $\{X_{1:5}^i\} = \{G^W, G^W, A^W, C^C, T^C, T^C\}$, where the superscripts denote which strand the base call was mapped to. The probability of these base calls given their source nucleotide is based on a read error model that accounts for sequence context. We

use $E_{\eta_1,\eta_2,\eta_3}(N)$ to represent the probability of Illumina calling nucleotide N given that the nucleotide being read is η_3, and the two preceding nucleotides are $\eta_1\eta_2$. If base call X_j^i is mapped to the Watson strand, then

$$E_{\eta_1,\eta_2,\eta_3}(N) = P(X_j^i = N | g_i = \eta_3, F_{i-2} = \eta_1, F_{i-1} = \eta_2) \tag{3.2}$$

and if it is mapped to the Crick strand, then

$$E_{\eta_1,\eta_2,\eta_3}(N) = P(X_j^i = N | g_i^C = \eta_3, F_{i+2}^C = \eta_1, F_{i+1}^C = \eta_2) \tag{3.3}$$

where g_i is the actual nucleotide being sequenced, F_k is the reference nucleotide at position k, and the C superscript denotes complementarity (e.g. if $F_{i+1} = A$, then $F_{i+1}^C = T$). If there is no polymorphism, then $g_i = F_i$. We refer to $\eta_1\eta_2\eta_3$, the triplet ending in the nucleotide being sequenced, as the "sequence context" for position i. For base calls mapped to the Watson strand in Figure 3.5 (in the absence of a polymorphism) the sequence context given by the reference is GGA, while for those mapped to the Crick strand, the sequence context is TGT.

3.4.2 Likelihood Terms

We first describe how to infer polymorphisms from a dataset and a reference genome, for fixed population parameters θ, p_α, and E. In section 3.4.4, we will describe how to simultaneously infer these parameters alongside the polymorphisms using an expectation maximization algorithm.

If the DNA matches the reference at position i, a given call X_j at position i will have sequence context $F_{i-1}F_{i-1}F_i$ if it maps to the Watson strand, and con-

text $F^C_{i+2}F^C_{i+2}F^C_i$ if it maps to the Crick strand. Therefore, the probability of base call X_j is given by $E_{F_{i-2},F_{i-1},F_i}(X_j)$ if the call maps to the Watson strand, and by $E_{F^C_{i+2},F^C_{i+1},F^C_i}(X_j)$ if it maps to the Crick strand. Since we assume that base calls are conditionally independent given their source nucleotides, the probability of a set of calls $\{X_j\}$ from a genome matching the reference is given by the product over the independent probabilities:

$$P(\{X_j\}|G_i = F_i) = \prod_{j\in J^W} E_{F_{i-2},F_{i-1},F_i}(X_j) \prod_{j\in J^C} E_{F^C_{i+2},F^C_{i+1},F^C_i}(X_j) \tag{3.4}$$

where J^W indexes the base calls mapped to the Watson strand, and J^C indexes the calls mapped to the Crick strand.

If G_i is a polymorphism, the probability of the data depends on the level of the polymorphism, α. Given a particular α, the probability of a base call X_j mapped to the Watson strand is $\alpha E_{F_{i-2},F_{i-1},G_{\text{poly}}}(X_j) + (1-\alpha)E_{F_{i-2},F_{i-1},F_i}(X_j)$, where G_{poly} is the non-reference nucleotide of the polymorphism, which may be any of the three alternative nucleotides. Therefore, the probability of a set of calls emitted by a population of bi-allelic genomes containing nucleotides F_i and G_{poly} at specified level α is given by

$$P(\{X_j\}|G_i = F_i/G_{\text{poly}}, \text{level } \alpha) =$$

$$\prod_{j\in J^W} (\alpha E_{F_{i-2},F_{i-1},G_{\text{poly}}}(X_j) + (1-\alpha)E_{F_{i-2},F_{i-1},F_i}(X_j)) \tag{3.5}$$

$$\times \prod_{j\in J^C} (\alpha E_{F^C_{i+2},F^C_{i+1},G^C_{\text{poly}}}(X_j) + (1-\alpha)E_{F^C_{i+2},F^C_{i+1},F^C_i}(X_j))$$

where the shorthand $G_i = F_i/G_{\text{poly}}$ is used to indicate a polymorphism at position i where each allele in the population has either nucleotide F_i or G_{poly}.

We then find the probability of the data given polymorphism F_i/G_{poly} by marginalizing over α:

$$P(\{X_j\}|G_i = F_i/G_{\text{poly}}) = \sum_{\alpha} p_\alpha P(\{X_j\}|G_i = F_i/G_{\text{poly}}, \text{level } \alpha) \tag{3.6}$$

3.4.3 Inference of SNPs

In order to infer the locations of polymorphisms, we calculate the posterior probabilities of each of the four possible hidden genotypes (e.g. A, A/C, A/G, A/T for a position with $F_i = A$). The unnormalized posterior probability that $G_i = F_i$ (i.e. the DNA matches the reference) can be found using Bayes' rule, from which we derive the relationship $P(b|a) \propto P(a|b)P(b)$:

$$P(G_i = F_i|\{X_j\}) \propto P(\{X_j\}|G_i = F_i)P(G_i = F_i) = \theta P(\{X_j\}|G_i = F_i) \tag{3.7}$$

Likewise, we calculate the unnormalized posterior probability of each of the three possible polymorphisms:

$$\begin{aligned} P(G_i = F_i/G_{\text{poly}}|\{X_j\}) \propto & P(\{X_j\}|G_i = F_i/G_{\text{poly}})P(G_i = F_i/G_{\text{poly}}) \\ &= \frac{1-\theta}{3} P(\{X_j\}|G_i = F_i/G_{\text{poly}}) \end{aligned} \tag{3.8}$$

Since these are mutually exclusive and exhaustive possibilities for G_i under our model, we can normalize over the sum of all four to arrive at the posterior probabilities. We call a polymorphism if the maximum of these posterior probabilities corresponds to one of the bi-allelic alternatives.

We can also compute the posterior probability of a particular polymorphism level

α given a chosen polymorphism F_i/G_{poly}:

$$P(\text{level } \alpha | G_i = F_i/G_{\text{poly}}, \{X_j\}_i) = \frac{P(\{X_j\}|G_i = F_i/G_{\text{poly}}, \text{level } \alpha) p_\alpha}{P(\{X_j\}|G_i = F_i/G_{\text{poly}})} \tag{3.9}$$

Therefore, for called polymorphisms, we can also estimate the level α using the MAP (Maximum *a posteriori*) estimate of the level for an inferred polymorphism F_i/G_{poly}:

$$\hat{\alpha} = \text{argmax}_\alpha \, p_\alpha P(\{X_j\}|G_i = F_i/G_{\text{poly}}, \text{level } \alpha) \tag{3.10}$$

3.4.4 Parameter Learning with EM

Calling polymorphisms in the way described above requires knowledge of the parameters – the error probabilities for each context, the proportion of polymorphisms in the population, and the probability distribution across polymorphism levels. Since we do not know these *a priori*, we use an EM algorithm to simultaneously learn these parameters and identify polymorphisms. In each iteration of the algorithm, we use the current parameters values to compute the posterior distributions of the hidden states as described above, then we update the parameters with maximum likelihood estimates that use expected counts based on the posterior probabilities calculated for the hidden states.

The parameter updates are given in the following equations, where N is the total number of sites, $I^W_{\eta_1,\eta_2}$ is the set of genomic positions preceded on the Watson strand by $\eta_1\eta_2$, and $I^C_{\eta_1,\eta_2}$ is the set of positions preceded on the Crick strand by $\eta_1\eta_2$ (i.e. followed on the Watson strand by $\eta_2^C\eta_1^C$). $x^W_{i,\eta}$ and $x^C_{i,\eta}$ are the number of times nucleotide η is represented in the base calls mapped to position i on the Watson and

Crick strands respectively. M_i^W and M_i^C are the total number of base calls mapped to position i on the Watson and Crick strands respectively. All posterior distributions are calculated with current parameter values θ^t, p_α^t, and E^t (as explicitly expressed in the equation for $\theta^{(t+1)}$).

$$\theta^{(t+1)} \leftarrow \frac{\mathbb{E}\{\#\text{ Positions matching the reference}\}}{\#\text{ of positions}}$$

$$= \frac{\sum_{i=1}^{N} P(G_i = F_i|\{X_j\}_i\ \theta^t\ p_\alpha^t\ E^t)}{N} \tag{3.11}$$

This equation gives the expected number of positions matching the reference over the total number of positions. Since a given position either matches the reference or does not, the expected value at each position is simply given by the posterior probability that the position matches the reference.

$$E_{\eta_1,\eta_2,\eta_3}(\eta)^{(t+1)} \leftarrow \frac{\mathbb{E}\{\#\eta \text{ basecalls in context } \eta_1\eta_2\eta_3\}}{\mathbb{E}\{\#\text{ All base calls in context } \eta_1\eta_2\eta_3\}} \tag{3.12}$$

$$\frac{\sum_{i\in I^W_{\eta_1,\eta_2}} P(G_i = \eta_3|\{X_j\}_i)x^W_{i,\eta} + \sum_{i\in I^C_{\eta_1,\eta_2}} P(G^C_i = \eta_3|\{X_j\}_i)x^C_{i,\eta}}{\sum_{i\in I^W_{\eta_1,\eta_2}} P(G_i = \eta_3|\{X_j\}_i)M^W_i + \sum_{i\in I^C_{\eta_1,\eta_2}} P(G^C_i = \eta_3|\{X_j\}_i)M^C_i}$$

Here we have the expected number of base calls η with context $\eta_1\eta_2\eta_3$ over the expected number of total calls with context $\eta_1\eta_2\eta_3$. The first sum in the numerator gives the number of η base calls mapped to the Watson strand at positions preceded by $\eta_1\eta_2$ on the forward reference strand, weighted at each position by the probability that the genomic identity G_i is η_3. The second sum is the number of η base calls mapped to the Crick strand at positions preceded by $\eta_1\eta_2$ on the reverse strand, weighted at each position by the probability that the genomic identity G_i is η_3^C. The sums in the denominator are the same, substituting the total number of base calls

at each position for the number of η base calls.

$$p_\alpha{}^{(t+1)} \leftarrow \frac{\mathbb{E}\{\#\text{ Polymorphisms at level } \alpha\}}{\mathbb{E}\{\#\text{ Polymorphisms}\}} \tag{3.13}$$

$$\frac{\sum_{i=1}^{N}\sum_{G_{\text{poly}}} P(\text{level } \alpha | G_i = R_i/G_{poly}, \{D_j\}_i) P(G_i = R_i/G_{poly} | \{D_j\}_i)}{\sum_{i=1}^{N}\sum_{G_{poly}} P(G_i = R_i/G_{poly} | \{D_j\}_i)}$$

Here we have the expected number of polymorphisms at a specified level α over the expected total number of polymorphisms. Again, since a given position is a polymorphism at level α or it isn't, the expected value at each position is the probability that the position is a polymorphism at level α. To get this value, we marginalize over the three possible polymorphism types. The denominator can also be found as $\sum_{i=1}^{N}(1 - P(G_i = F_i|\{D_i\}_i))$, and is just the probability that the DNA does not match the reference, summed over all positions.

We can initialize the algorithm using generic starting guesses for p_α and θ, and using error rates calculated for the matching PhiX lane as a starting guess for E. For each iteration of the EM algorithm, we compute the posterior probabilities of the four possible genotypes, as well as the posterior distribution of the polymorphism level. We then compute new parameter estimates based on these posteriors, and iterate until the change in log likelihood falls below a threshold δ.

3.5 Results of SNP Inference

3.5.1 LoxP results and validation

We applied our algorithm to the LoxP dataset initialized with the corresponding PhiX error rates, and we found a surprising number of polymorphisms relative to the reference genome. The estimate for θ was 0.9938, corresponding to about one polymorphism every 161 nucleotides. Of the polymorphisms called, translations ($A \leftrightarrow G$ and $C \leftrightarrow T$) made up more about 53%, although they only represent one third of all possible changes.

The estimated distribution over polymorphism levels, p_α, is shown in Figure 3.9. What we see is that only about 38% of the polymorphisms are fixed in the population, while the rest are more or less evenly spread over polymorphism levels. A subset of the estimated error rates is shown in Figure 3.10. We see fairly low error rates, on average 0.0002, with the sort of variation we expect; the minimum error probability is 5.4×10^{-5} for $GGA \to GGC$ and the maximum is 0.0007 for $TTC \to TTT$.

In order to validate our polymorphism calls, we used Sanger sequencing to sequence randomly chosen 1.2 kB fragments of the genome. To choose these fragments, we drew a random sample of 20 predictions, then took the regions spanned by 600 bases up and downstream of the chosen prediction. In each of these regions, we filtered our predictions to those polymorphism calls which had posterior probability greater than 99.9% and had probability less than 1% of having polymorphism level $\alpha = 0.1$. The second filter was put in place because very low-level polymorphisms are difficult to detect using Sanger sequencing. These two filters together left us with about 75% of the original predictions.

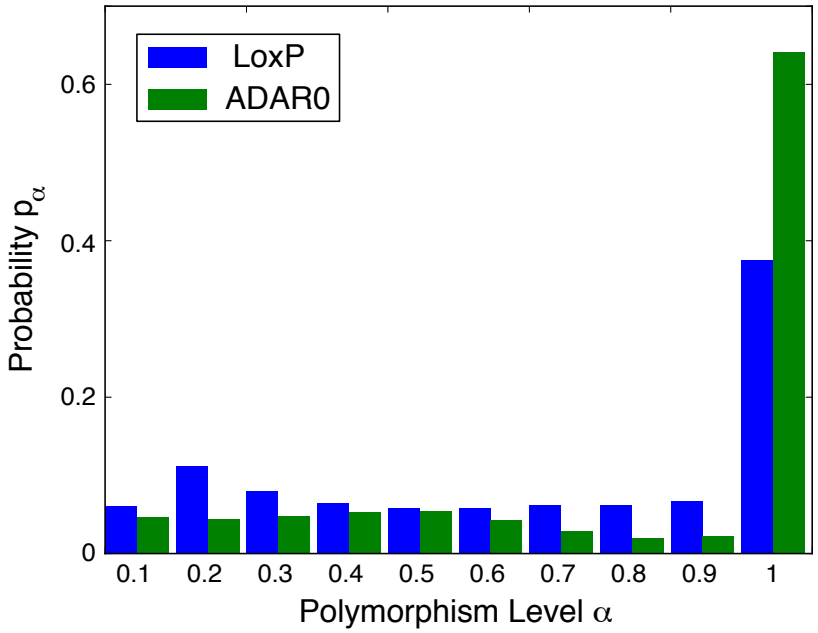

Figure 3.9: EM estimate of p_α for LoxP and ADAR0 datasets

This validation study actually caught an error in the preprocessing step of our algorithm. Despite using a masked genome to limit the scope of our search to unique sequence, some highly similar regions remained, leading to a fair number of false positives. The way we used the mapping software to map the DNA reads to the reference did not adequately account for these highly similar regions across chromosomes. We were able to fix the majority of these by adjusting our pre-processing step, after which we saw a validation rate of 83%, with a 95% confidence interval of $(0.76, 0.90)$. If we look at the individual regions in which we still call polymorphisms, most regions validate at 100%, while two regions validate at around 50%. The reason for these anomalies is unclear at this point, since they do not seem to be repetitive. One last region in which we continue to see false positives is a repetitive element called DNArep. We suspect that the reason for these particular false positives is that the masked genome contains only some DNArep elements and not others, forcing reads from the hidden elements to map to the visible elements.

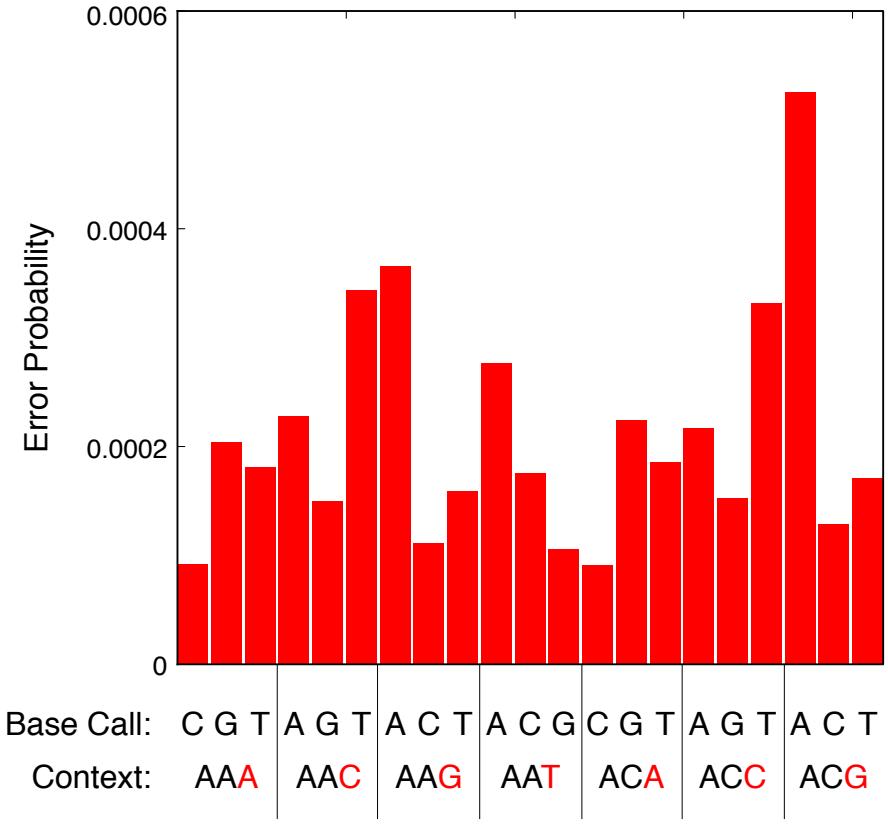

Figure 3.10: Subset of EM error rate estimates for LoxP dataset

3.5.2 ADAR0 results

We also applied our algorithm to the ADAR0 dataset, initialized with its matching PhiX error rates, and saw a similarly large polymorphism rate: $\theta = 0.9948$, or about one polymorphism every 191 nucleotides. In contrast to the LoxP line however, about 64% of the polymorphisms were fixed, as shown in Figure 3.9. Again, the other levels were approximately equally likely. The learned error parameters for the ADAR0 dataset have a mean of 9.7×10^{-5}, a minimum of 1.7×10^{-5} for $GGA \to GGC$, and a maximum of 0.0003 for $TTC \to TTT$. Notice that these are the same contexts for which we had a maximum and minimum error rate for the LoxP dataset as well. In fact, the error rates in the two populations have Spearman's rank correlation coefficient $\rho = 0.81$. The relationship between the ranked variables is shown in Figure 3.11, revealing a strong relationship between the qualitative effects of each sequence context between the two datasets. This is further confirmation that sequence context is playing a role in the error rates produced by the Illumina sequencing technology. It also illustrates the potential overall variation from run to run of the Illumina machine; the average error rate for LoxP is more than double that for ADAR0, but the relative contributions of each sequence context are roughly the same.

To compare ADAR0 and LoxP SNP calls, we first filtered the lists of calls using the same criteria as in Section 3.5.1 to collect calls with high posterior probability and moderate polymorphism levels. We also limited the lists to positions in which LoxP and ADAR0 had at least 5 overlapping reads each. We saw approximately 52% of the LoxP SNPs in the ADAR0, and approximately 55% of the ADAR0 SNPs in LoxP. Interestingly, of the fixed ADAR0 polymorphisms also found in LoxP, 38% were unfixed in LoxP. On the other hand, of the fixed LoxP polymorphisms also found in ADAR0, only 8% were unfixed in ADAR0. This observation, along with

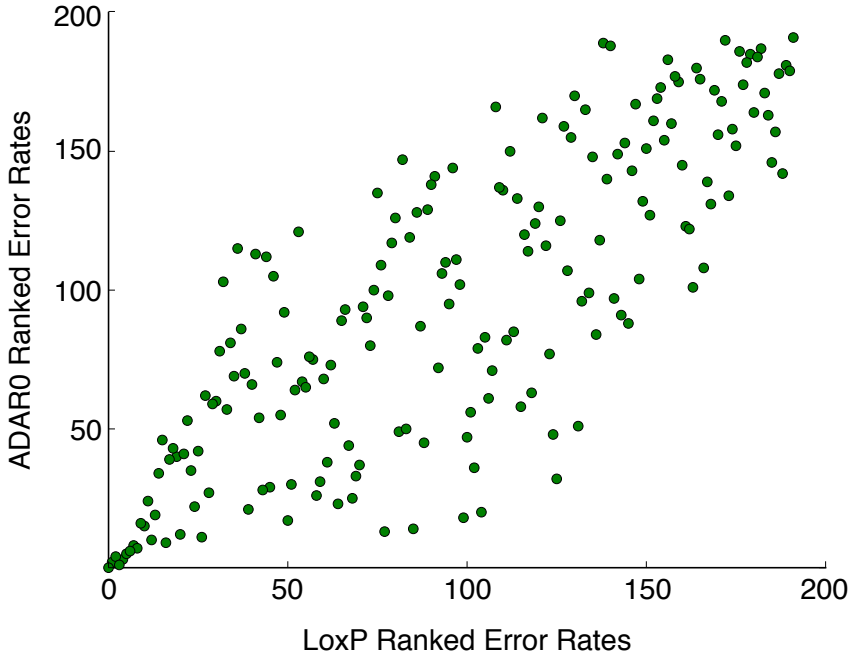

Figure 3.11: Relationship between ranked error rates by context for LoxP and ADAR0

the higher level of fixed polys in ADAR0 compared to LoxP, makes sense in light of the way ADAR0 was generated; a small number of flies displaying the ADAR null phenotype were isolated from the LoxP stock, resulting in an extreme population bottleneck, which would have the effect of fixing previously unfixed polymorphisms, or reverting them to the reference

3.5.3 Generic Initialization

Initialization with the error rates learned from matching PhiX runs leads to very fast convergence, but not all facilities run a PhiX control lane every time. We suspected, given the large amount of data and the relatively small number of parameters, that a generic initialization would converge fairly quickly to the same parameters. Indeed, initializing the ADAR0 dataset with error rates of 0.001 for all errors took about

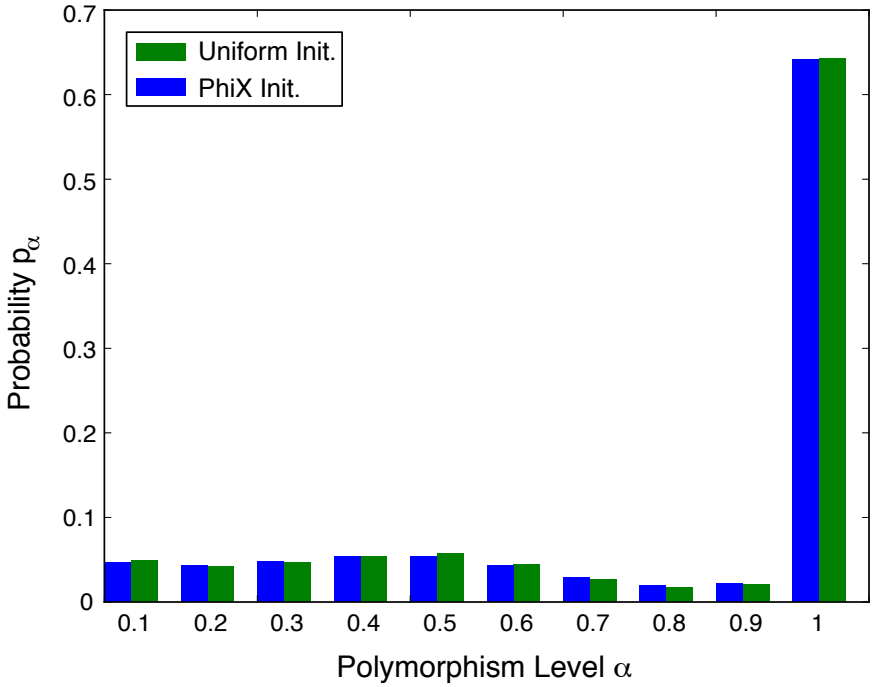

Figure 3.12: Comparison of p_α for different initializations of the EM algorithm

twice as long to converge, but gave nearly identical results. The average difference in the converged error rates between the two initializations is 1.06×10^{-15}, with variance 4.4×10^{-14}, only 5 thousandths of the variance in error rates, 2.1×10^{-9}. For parameter θ, the difference in values between initializations is 5.95×10^{-6}, representing an extra 6 polymorphisms every million positions, a very low number considering that we already expect over 6,000 polymorphisms in a million positions. The average difference in polymorphism level probabilities p_α is 5.0×10^{-14}, with variance 1.02×10^{-11}. Figure 3.12 shows the converged values for p_α for each initialization.

We see similar results for the LoxP dataset. With an initialization of all error probabilities at 0.001 compared to initialization with the matching PhiX control run, the difference in θ is 6.89×10^{-6}, corresponding to an extra 6-7 polymorphisms every million positions. The average difference in error rates is 1.6×10^{-7}, with a variance of 6.2×10^{-15}, in contrast with a variance in error rates of 1.6×10^{-8}. The average

difference in polymorphism level probabilities is 1.4×10^{-18} with variance 2.5×10^{-5}.

For each of these parameters, it seems quite likely that the differences would continue to shrink upon further iterations of the algorithm, with increasingly strict convergence criteria. This gives us increased confidence in two things: first, that our EM algorithm is converging to the maximum likelihood estimates of our parameters, and second, that initialization with a PhiX control is unnecessary, making our technique more broadly applicable for investigators who want to identify SNPs in their laboratory populations.

3.6 Full Model

We now move to the full model, incorporating both DNA and RNA data in order to infer the positions of both polymorphisms and editing events. Of course, the primary motivation is the discovery of editing sites, but there is a secondary motivation as well: the RNA provides us with more power to detect SNPs. This is especially useful when the DNA coverage is relatively low.

For the most part RNA reads can be treated just as DNA reads, but there are a couple of added features that we need to address. The first is one that we are interested in for its own sake: the editing of RNA molecules changing select As into Gs. Thus, in addition to SNPs and Illumina errors, we have a third process that can effect what we see in the data. Of course, editing can occur at different levels, so just as we modeled polymorphism levels in the DNA, we will need to model editing levels in the RNA. The second difficulty with RNA is that a single RNA molecule is transcribed from either the Watson or Crick strand. The consequence is that a

position can only be edited if its genome has an A and the RNA was transcribed from the Watson strand, *or* the genome has a T and the RNA was transcribed from the Crick strand. The trouble is that we do not directly observe which strand was transcribed, since cDNA is made from the RNA before amplification and sequencing, so we must infer it. We make the assumption here that at any given position, only one strand (either Watson or Crick) provides the template for all aligned base calls. We also assume that each strand is equally likely, and that the strand at a given position is independent from the strands at all other positions. The consequences of these assumptions are discussed in Section 3.7.7. We aim to use this model to infer the locations of polymorphisms as well as editing sites, while simultaneously learning the error rates for each run (for DNA and RNA separately), the frequency of polymorphisms, the frequency of editing events, and the distribution over the levels of polymorphisms and edits. Note: for notational simplicity, T will be used throughout the rest of this chapter to represent both a T in the DNA and a U in the RNA.

3.6.1 Prior Model

Suppose we have an A in the reference at position i. We assume 7 possible genotypes on the Watson strand, $G_i \in \{A, C, G, T, A/C, A/G, A/T\}$, where A/C, A/G, and A/T indicate unfixed polymorphisms, and C, G, and T indicated fixed polymorphisms. In the restricted model we grouped fixed and unfixed polymorphisms, but the assumptions we make for the full model require treating these two cases separately. We assign the following parameters: θ is the probability of no polymorphism (an A), as before. Given a polymorphism, ω is the probability that it is fixed in the population. Given a fixed polymorphism, each of the three possibilities (C,G,T)

is equally likely with probability $\frac{1}{3}$, and likewise for the unfixed polymorphisms (A/C,A/G,A/T). For the unfixed polymorphisms, we again let α be the proportion of non-reference nucleotides, and α follows distribution p_α.

We assume that editing events are only possible for two of these genotypes, $G_i = A$ and $G_i = T$, so that editing events at unfixed polymorphic sites are not allowed under our model. Furthermore, an editing event is only possible in the former case if the Watson strand is transcribed, and is only possible in the latter case if the Crick strand is transcribed. Therefore, for these two genotypes, we assume a $\frac{1}{2}$ probability that the RNA was transcribed from each of the strands. We let hidden variable S_i represent the strand transcribed at position i. If ($G_i = A$ and $S_i = W$) or ($G_i = T$ and $S_i = C$), then we let ϕ be the prior probability that position i is an edited site, leading to A/G or T/C in the RNA, respectively. We let δ be the level of editing, i.e. the proportion of $A \to G$ or $T \to C$ changes, and δ follows distribution q_δ. We use R_i for the hidden variable representing the identity of the RNA at position i. This generative prior model is illustrated in Figure 3.13 for the case in which the reference contains an A. Arrows in black have probability 1. For example, given that the DNA has an unfixed polymorphic site at level α, the RNA must have the same polymorphism at the same level α. The two entries in green show the possible editing events. Note that this model allows us to find potential editing events even when the reference does not have an A (i.e. when there is a fixed polymorphism to an A or a T).

This leads to the following prior probabilities on G_i and R_i (letting F_i be the

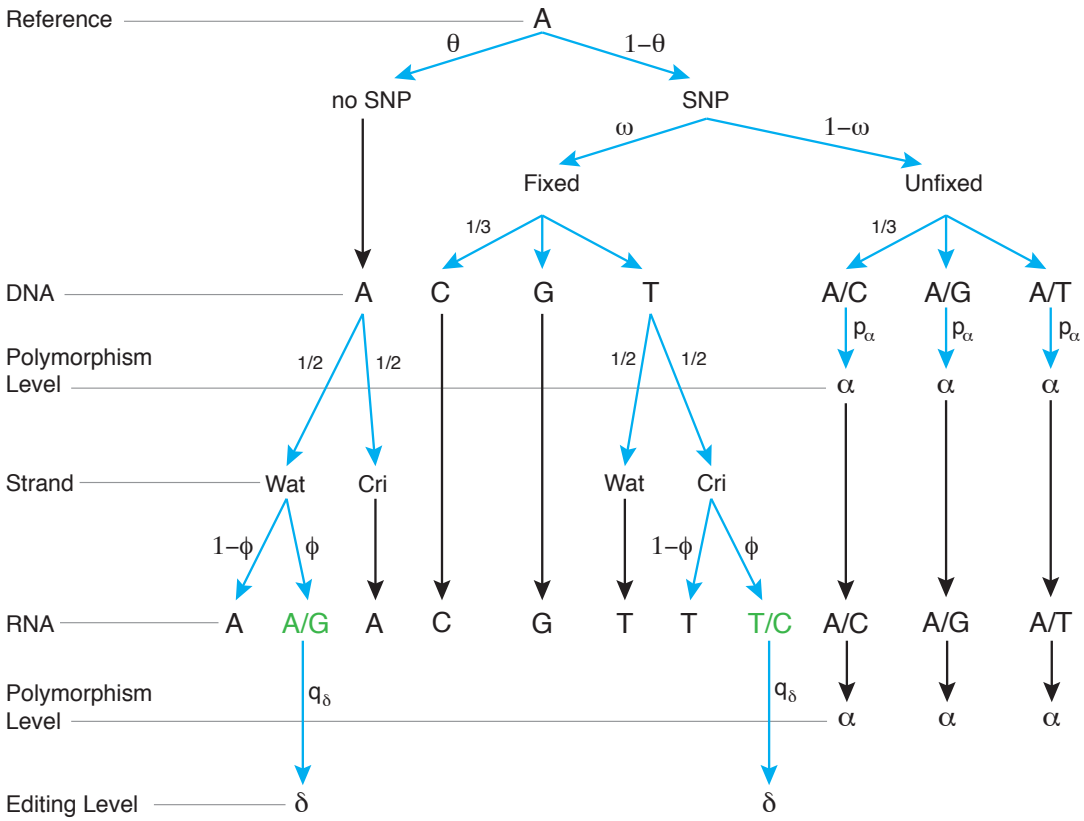

Figure 3.13: Diagram of the prior model for an A in the reference

reference base at position i) for the case where there is an A in the reference:

$$\begin{aligned}
P(G_i = A, R_i = A/G, \text{ level } \delta | F_i = A) &= \frac{1}{2}\theta\phi q_\delta \\
P(G_i = T, R_i = T/C, \text{ level } \delta | F_i = A) &= \frac{1}{6}(1-\theta)\omega\phi q_\delta \\
P(G_i = A, R_i = A | F_i = A) &= \frac{1}{2}\theta(2-\phi) \\
P(G_i = T, R_i = T | F_i = A) &= \frac{1}{6}(1-\theta)\omega(2-\phi) \\
P(G_i = C, R_i = C | F_i = A) &= \frac{1}{3}(1-\theta)\omega \\
P(G_i = G, R_i = G | F_i = A) &= \frac{1}{3}(1-\theta)\omega \\
P(G_i = A/C, \text{ level } \alpha, R_i = A/C, \text{ level } \alpha | F_i = A) &= \frac{1}{3}(1-\theta)(1-\omega)p_\alpha \\
P(G_i = A/G, \text{ level } \alpha, R_i = A/G, \text{ level } \alpha | F_i = A) &= \frac{1}{3}(1-\theta)(1-\omega)p_\alpha \\
P(G_i = A/T, \text{ level } \alpha, R_i = A/T, \text{ level } \alpha | F_i = A) &= \frac{1}{3}(1-\theta)(1-\omega)p_\alpha
\end{aligned} \tag{3.14}$$

All other possible combinations have probability zero.

If the base in the reference is not an A, the picture is similar. Take, for example, the case where the reference contains a C. Then θ is the probability that $G_i = C$, each fixed polymorphism A, G, T has probability $\frac{1}{3}(1-\theta)\omega$ and each unfixed polymorphism $C/A, C/G, C/T$ has probability $\frac{1}{3}(1-\theta)(1-\omega)$. Conditioned on G_i, the rest of the model is identical. editing events can only come from $G_i = A$ or $G_i = T$, and unfixed polymorphisms in the DNA must be present at the same level in the RNA. This case is illustrated in Figure 3.14. The prior probabilities on G_i and R_i for a C in the

reference are thus given by the following expressions:

$$\begin{aligned}
P(G_i = C, R_i = C | F_i = C) &= \theta \\
P(G_i = A, R_i = A/G, \text{ level } \delta | F_i = C) &= \frac{1}{6}(1-\theta)\omega\phi q_\delta \\
P(G_i = T, R_i = T/C, \text{ level } \delta | F_i = C) &= \frac{1}{6}(1-\theta)\omega\phi q_\delta \\
P(G_i = A, R_i = A | F_i = C) &= \frac{1}{6}(1-\theta)\omega(2-\phi) \\
P(G_i = T, R_i = T | F_i = C) &= \frac{1}{6}(1-\theta)\omega(2-\phi) \\
P(G_i = G, R_i = G | F_i = C) &= \frac{1}{3}(1-\theta)\omega \\
P(G_i = C/A, \text{ level } \alpha, R_i = C/A, \text{ level } \alpha | F_i = C) &= \frac{1}{3}(1-\theta)(1-\omega)p_\alpha \\
P(G_i = C/G, \text{ level } \alpha, R_i = C/G, \text{ level } \alpha | F_i = C) &= \frac{1}{3}(1-\theta)(1-\omega)p_\alpha \\
P(G_i = C/T, \text{ level } \alpha, R_i = C/T, \text{ level } \alpha | F_i = C) &= \frac{1}{3}(1-\theta)(1-\omega)p_\alpha
\end{aligned} \tag{3.15}$$

with all other possible combinations of G_i and R_i having probability zero.

3.6.2 Likelihood Terms

As before, we will first focus on the inference of polymorphisms and editing sites for a fixed set of parameters before discussing how to learn the parameters simultaneously via EM. Of course, the inference of polymorphisms and editing sites requires the computation of likelihood terms for each possible set of hidden variables.

Given G_i and R_i at a given position, we can calculate the likelihood of the data using our error model. Since the DNA and RNA data come from different runs of the Illumina machinery, we assume that the error rates may differ, but as before, both error models depend on the sequence context of the base call. We let $E^{\text{DNA}}_{\eta_1,\eta_2,\eta_3}(N)$

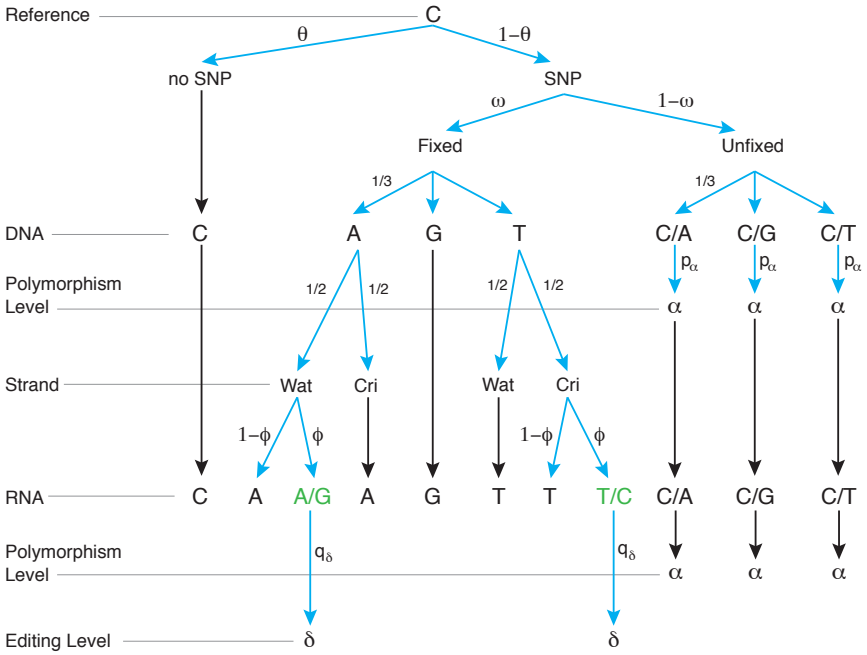

Figure 3.14: Diagram of the prior model for a C in the reference

and $E^{\text{RNA}}_{\eta_1,\eta_2,\eta_3}(N)$ be the probability of seeing base call N in sequence context $\eta_1\eta_2\eta_3$ in the DNA and RNA respectively.

Let $\{X_j\}_i$ be the DNA reads mapped to position i, and let $\{Y_j\}_i$ be the RNA reads mapped to position i. We also use J^W to index the DNA reads mapping to the Watson strand, and J^C to index the DNA reads mapping to the Crick strand. Likewise, we define K^W and K^C to index the RNA reads mapping to the Watson and Crick strands respectively. We also define an indicator variable ϵ to be 1 if the pair G_i, R_i represents an editing event and 0 otherwise. Since the DNA and RNA reads are conditionally independent given their source molecules G_i and R_i, we can define and compute the likelihood as follows:

$$P(\{X_j\}_i, \{Y_j\}_i | G_i, R_i) = P(\{X_j\}_i | G_i) P(\{Y_j\}_i | R_i, \epsilon) \tag{3.16}$$

We will consider these two terms separately. The first term can take two different forms, depending on whether G_i is fixed or unfixed. If G_i is fixed (N = A, C, G, or T):

$$P(\{X_j\}_i|G_i = N) = \prod_{j\in J^W} E^{\text{DNA}}_{\text{F}_{i-2},\text{F}_{i-1},N}(X_j) \prod_{j\in J^C} E^{\text{DNA}}_{\text{F}^C_{i+2},\text{F}^C_{i+1},N^C}(X_j) \tag{3.17}$$

Where superscript C on a nucleotide (i.e. F_{i+2}) denotes complementarity. If G_i is unfixed with reference nucleotide N_{ref} and alternative nucleotide N_{alt} at level α, then the probability of the DNA reads is given by the following (as in the restricted case):

$$\begin{aligned} P(\{X_j\}_i|G_i = N_{\text{ref}}/N_{\text{alt}},\ \text{level } \alpha) = \\ \prod_{j\in J^W} [(1-\alpha)E^{\text{DNA}}_{\text{F}_{i-2},\text{F}_{i-1},N_{\text{ref}}}(X_j) + \alpha E^{\text{DNA}}_{\text{F}_{i-2},\text{F}_{i-1},N_{\text{alt}}}(X_j)] \\ \times \prod_{j\in J^C} [(1-\alpha)E^{\text{DNA}}_{\text{F}^C_{i+2},\text{F}^C_{i+1},{N_{\text{ref}}}^C}(X_j) + \alpha E^{\text{DNA}}_{\text{F}^C_{i+2},\text{F}^C_{i+1},{N_{\text{alt}}}^C}(X_j)] \end{aligned} \tag{3.18}$$

Therefore, the probability of the data given polymorphism $N_{\text{ref}}/N_{\text{alt}}$ is:

$$P(\{X_j\}_i|G_i = N_{\text{ref}}/N_{\text{alt}}) = \sum_\alpha p_\alpha P(\{X_j\}_i|G_i = N_{\text{ref}}/N_{\text{alt}},\ \text{level } \alpha) \tag{3.19}$$

We can now move to the likelihood term involving the RNA reads. There are three possibilities: that the RNA is fixed, that there are two nucleotides present due to a polymorphism ($\epsilon = 0$), or that there are two nucleotides present due to editing ($\epsilon = 1$). Note that we do not allow for both bi-allelic possibilities simultaneously. The likelihoods for the first two possibilities are nearly identical to the DNA case: If R_i is fixed:

$$P(\{Y_j\}_i|R_i = N) = \prod_{j\in K^W} E^{\text{RNA}}_{\text{F}_{i-2},\text{F}_{i-1},N}(Y_j) \prod_{j\in K^C} E^{\text{RNA}}_{\text{F}^C_{i+2},\text{F}^C_{i+1},N^C}(Y_j) \tag{3.20}$$

If R_i is an unfixed polymorphism at level α ($\epsilon = 0$):

$$P(\{Y_j\}_i | R_i = N_{\text{ref}}/N_{\text{alt}}, \epsilon = 0, \text{level } \alpha) = \prod_{j \in K^W} [(1-\alpha) E^{\text{RNA}}_{\text{F}_{i-2},\text{F}_{i-1},N_{\text{ref}}}(Y_j) + \alpha E^{\text{RNA}}_{\text{F}_{i-2},\text{F}_{i-1},N_{\text{alt}}}(Y_j)] \times \prod_{j \in K^C} [(1-\alpha) E^{\text{RNA}}_{\text{F}^C_{i+2},\text{F}^C_{i+1},{N_{\text{ref}}}^C}(Y_j) + \alpha E^{\text{RNA}}_{\text{F}^C_{i+2},\text{F}^C_{i+1},{N_{\text{alt}}}^C}(Y_j)] \tag{3.21}$$

So for an unfixed polymorphism $N_{\text{ref}}/N_{\text{alt}}$, the probability of the RNA data is

$$P(\{Y_j\}_i | R_i = N_{\text{ref}}/N_{\text{alt}}, \epsilon = 0) = \sum_\alpha p_\alpha P(\{Y_j\}_i | R_i = N_{\text{ref}}/N_{\text{alt}}, \epsilon = 0, \text{ level } \alpha) \tag{3.22}$$

Finally, given an editing event [($G_i = A$ and $R_i = A/G$) or ($G_i = T$ and $R_i = T/C$)], we deal with errors the same way we deal with polymorphisms: each read has probability δ (instead of α) of coming from the edited nucleotide N_2, and probability $1-\delta$ of coming from the genomic nucleotide N_1, *a priori*. If the editing is at level δ, the probability of the data is:

$$P(\{Y_j\}_i | R_i = N_1/N_2, \epsilon = 1, \text{level } \delta) = \prod_{j \in K^W} [(1-\delta) E^{\text{RNA}}_{\text{F}_{i-2},\text{F}_{i-1},N_1}(Y_j) + \delta E^{\text{RNA}}_{\text{F}_{i-2},\text{F}_{i-1},N_2}(Y_j)] \times \prod_{j \in K^C} [(1-\delta) E^{\text{RNA}}_{\text{F}^C_{i+2},\text{F}^C_{i+1},{N_1}^C}(Y_j) + \delta E^{\text{RNA}}_{\text{F}^C_{i+2},\text{F}^C_{i+1},{N_2}^C}(Y_j)] \tag{3.23}$$

Therefore, the probability of the data over all possible editing levels is given by:

$$P(\{Y_j\}_i | R_i = N_1/N_2, \epsilon = 1) = \sum_\delta q_\delta P(\{Y_j\}_i | R_i = N_1/N_2, \epsilon = 1, \text{level } \delta) \tag{3.24}$$

3.6.3 Inference of Hidden States

Similar to the way in which we inferred polymorphisms in the restricted model, we infer the sites of polymorphisms and editing events by computing the posterior distributions of each possible pair of hidden variables (G_i, R_i). If the most probable pair represents an editing event or a polymorphism, then we make the corresponding prediction. Bayes' rule again lets us compute the posterior probability of each (G_i, R_i) pair:

$$P(G_i, R_i|\{X_j\}_i, \{Y_j\}_i, F_i) \propto P(\{X_j\}_i|G_i)P(\{Y_j\}_i|R_i, \epsilon)P(G_i, R_i|F_i) \tag{3.25}$$

Normalizing by the sum over all pairs gives us the proper posterior probabilities.

Given a particular polymorphism $N_{\text{ref}}/N_{\text{alt}}$, the un-normalized probability of a particular level α is given by:

$$P(\text{level } \alpha|\{X_j\}_i, \{Y_j\}_i, G_i = N_{\text{ref}}/N_{\text{alt}}, \epsilon = 0) \propto$$
$$P(\{X_j\}_i, \{Y_j\}_i|G_i = N_{\text{ref}}/N_{\text{alt}}, \text{ level } \alpha)p_\alpha \tag{3.26}$$

Likewise, the probability of an editing level δ for an appropriate N_1/N_2 (A/G or T/C) is given by:

$$P(\text{level } \delta|\{X_j\}_i, \{Y_j\}_i, G_i = N_1, R_i = N_1/N_2, \epsilon = 1) \propto$$
$$P(\{X_j\}_i, \{Y_j\}_i|G_i = N_1, R_i = N_1/N_2, \text{ level } \delta)q_\delta \tag{3.27}$$

In each case, normalization by all possible levels yields the proper posteriors.

We can also compute the posterior probability that each position was transcribed

from either the Watson or Crick strand in the case where $G_i = A$ or $G_i = T$, which will be necessary for the parameter updates:

$$P(S_i = \text{Wat}|G_i = A, \{Y_j\}) \propto \frac{1}{2}[(1-\phi)P(\{Y_j\}|R_i = A) + \phi P(\{Y_j\}_i|R_i = A/G)]$$
$$P(S_i = \text{Cri}|G_i = A, \{Y_j\}) \propto \frac{1}{2}P(\{Y_j\}|R_i = A) \quad (3.28)$$

$$P(S_i = \text{Cri}|G_i = T, \{Y_j\}) \propto \frac{1}{2}[(1-\phi)P(\{Y_j\}|R_i = T) + \phi P(\{Y_j\}|R_i = T/C)]$$
$$P(S_i = \text{Wat}|G_i = T, \{Y_j\}) \propto \frac{1}{2}P(\{Y_j\}|R_i = T) \quad (3.29)$$

3.6.4 Parameter Updates

Parameter Updates are also handled in a similar way as before, but with a few new parameters to learn: ω, the probability that a polymorphism is fixed, ϕ, the edit rate, E^{RNA}, the error model for the RNA reads, and q_δ, the distribution over editing levels. As before, we also learn θ, the polymorphism rate, E^{DNA}, the error model for the DNA reads, and p_α, the probability distribution over unfixed polymorphism levels. Of course, the expected values that we use to update these parameters are based on both the DNA and RNA data combined. These updates are given in the following equations, letting N be the total number of genomic positions, $x^W_{i,N}$ and $x^C_{i,N}$ being the number of DNA N basecalls mapped to position i on the Watson and Crick strands, respectively, and $y^W_{i,N}$ and $y^C_{i,N}$ as the number of RNA basecalls mapped to position i. We use D_i as a shorthand for all of the data at position i, $\{X_j\}_i, \{Y_j\}_i$.

$$\theta \leftarrow \frac{\mathbb{E}\{\#\text{ Polymorphisms}\}}{N} = \frac{\sum_i P(G_i \neq F_i|D_i)}{N} \tag{3.30}$$

$$\omega \leftarrow \frac{\mathbb{E}\{\#\text{ Fixed Polymorphisms}\}}{\mathbb{E}\{\#\text{ Polymorphisms}\}} = \frac{\sum_i \sum_{\eta \neq F_i} P(G_i = \eta|D_i)}{\sum_i P(G_i \neq F_i|D_i)} \tag{3.31}$$

$$\begin{aligned}\phi \leftarrow & \frac{\mathbb{E}\{\#\text{ Editing events}\}}{\mathbb{E}\{\#\text{ Transcribed As}\}} \\ & = \frac{\sum_i P(G_i = A, R_i = A/G|D_i) + P(G_i = T, R_i = T/C|D_i)}{\sum_i P(G_i = A, S_i = \text{Wat}|D_i) + P(G_i = T, S_i = \text{Cri}|D_i)} \\ & = \frac{\sum_i P(G_i = A, R_i = A/G|D_i) + P(G_i = T, R_i = T/C|D_i)}{\sum_i P(S_i = \text{Wat}|G_i = A, D_i)P(G_i = A|D_i) + P(S_i = \text{Cri}|G_i = T, D_i)P(G_i = T|D_i)}\end{aligned} \tag{3.32}$$

$$\begin{aligned}p_\alpha \leftarrow & \frac{\mathbb{E}\{\#\text{ Polymorphisms at Level } \alpha\}}{\mathbb{E}\{\#\text{ Unfixed Polymorphisms}\}} \\ & = \frac{\sum_i \sum_{\eta \neq F_i} P(\text{level } \alpha|D_i, G_i = F_i/\eta)P(G_i = F_i/\eta|D_i)}{\sum_i \sum_{\eta \neq F_i} P(G_i = F_i/\eta|D_i)}\end{aligned} \tag{3.33}$$

$$\begin{aligned}q_\delta \leftarrow & \frac{\mathbb{E}\{\#\text{ Editing Events at Level } \delta\}}{\mathbb{E}\{\#\text{ Editing Events}\}} \\ & = \frac{\begin{array}{r}\sum_i P(\text{level } \delta|D_i, G_i = A, R_i = A/G,)P(G_i = A, R_i = A/G|D_i) \\ + P(\text{level } \delta|D_i, G_i = T, R_i = T/C)P(G_i = T, R_i = T/C|D_i)\end{array}}{\sum_i P(G_i = A, R_i = A/G|D_i) + P(G_i = T, R_i = T/C|D_i)}\end{aligned} \tag{3.34}$$

$$E^{\text{DNA}}_{\eta_1\eta_2\eta_3}(N) \leftarrow \frac{\mathbb{E}\{\#\text{ DNA basecalls N in context } \eta_1\eta_2\eta_3\}}{\mathbb{E}\{\#\text{ DNA basecalls in context } \eta_1\eta_2\eta_3\}} \tag{3.35}$$

$$= \frac{\sum_{i\in I^W_{\eta_1\eta_2}} P(G_i=\eta_3|D_i)x^W_{i,N} + \sum_{i\in I^C_{\eta_1\eta_2}} P(G_i=\eta^C_3|D_i)x^C_{i,N}}{\sum_{i\in I^W_{\eta_1\eta_2}} P(G_i=\eta_3|D_i)(\sum_{\tilde{N}} x^W_{i,\tilde{N}}) + \sum_{i\in I^C_{\eta_1\eta_2}} P(G_i=\eta^C_3|D_i)(\sum_{\tilde{N}} x^C_{i,\tilde{N}})}$$

$$E^{\text{RNA}}_{\eta_1\eta_2\eta_3}(N) \leftarrow \frac{\mathbb{E}\{\#\text{ RNA basecalls N in context } \eta_1\eta_2\eta_3\}}{\mathbb{E}\{\#\text{ RNA basecalls in context } \eta_1\eta_2\eta_3\}} \tag{3.36}$$

$$= \frac{\sum_{i\in I^W_{\eta_1\eta_2}} P(R_i=\eta_3|D_i)y^W_{i,N} + \sum_{i\in I^C_{\eta_1\eta_2}} P(R_i=\eta^C_3|D_i)y^C_{i,N}}{\sum_{i\in I^W_{\eta_1\eta_2}} P(R_i=\eta_3|D_i)(\sum_{\tilde{N}} y^W_{i,\tilde{N}}) + \sum_{i\in I^C_{\eta_1\eta_2}} P(R_i=\eta^C_3|D_i)(\sum_{\tilde{N}} y^C_{i,\tilde{N}})}$$

We use EM to compute the posterior probabilities of each (G_i, R_i) pair, then update the population parameters, iterating until convergence. Given the set of parameters after convergence, we can infer the most probable (G_i, R_i) pair at each position i. We call a polymorphism if G_i does not match the reference F_i, and we call an editing event if $(G_i, R_i) = (A, A/G)$ or $(T, T/C)$.

3.7 Discussion

3.7.1 Duplicate Reads and Coverage

The Illumina sequencing technology depends on a polymerase chain reaction (PCR) step to amplify the DNA or cDNA to be sequenced. This can result in multiple reads representing the same source molecule, referred to as "PCR duplicates." These

duplicates provide independent information about sequencing errors, but do not provide independent information about polymorphisms or editing sites, and thus lead to inappropriately large sample sizes for inferring these phenomena. This is potentially further complicated by the use of a restriction enzyme in the initial processing of the DNA. Restriction enzymes typically have sequence preferences for cleavage, which lead to a bias in which fragments are made and eventually sequenced. As a result, it is difficult to distinguish between two reads that came from the same source molecule and two reads that came from different source molecules cut at the same place by the restriction enzyme. In the latter case, unlike the former, the reads do provide independent information about polymorphisms. The restriction enzyme used for the experiments discussed here, dsDNA Fragmentase (NEBNext®), is advertised to be unbiased, but our preliminary experiments suggest that this may not be the case. In future experiments, a different process called sonication will be used to shear the DNA with sound energy, removing the need for enzymes which can have sequence biases.

The issue of PCR duplicates, however, will still remain. Tools exist to remove reads with identical coordinates, but this leaves us with fewer independent observations of the sequencing technology, and thus could lead to different parameter estimates and polymorphism calls. We tested this with our LoxP dataset, which has a moderate number of duplicates, as shown in the coverage plot in Figure 3.15 which displays a histogram of the base call depth at every unique position in the genome. We do indeed see slightly different parameters; the average original error rate is 0.00021, about 1.2 times the average error rate with duplicates removed, 0.00017, and θ, the prior probability of a genomic match is 0.9929 and 0.9931 with and without duplicates, respectively. This translates to a small change in the polymorphisms called by the algorithm. About 98% of polymorphisms called from the dataset with-

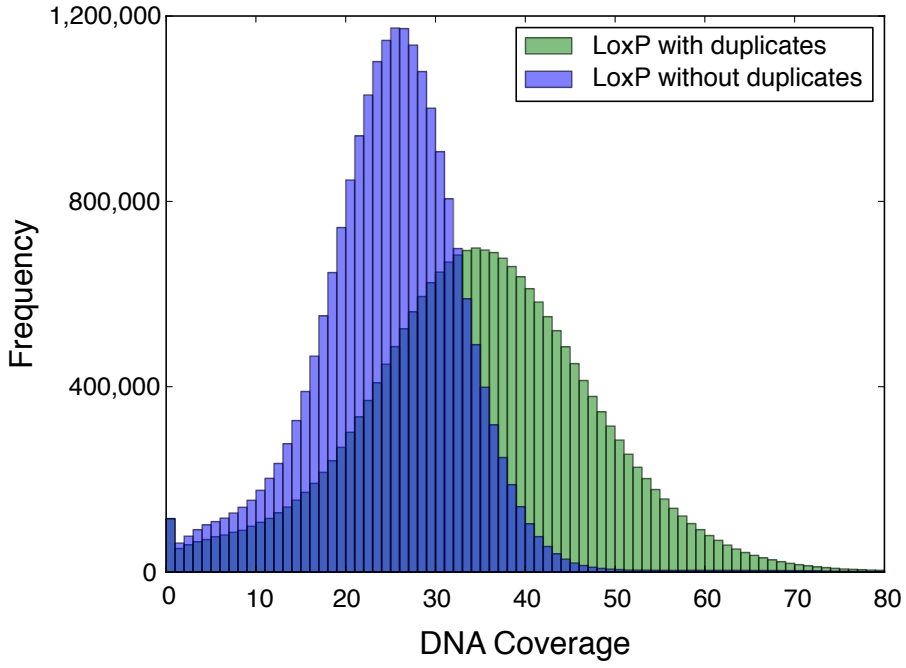

Figure 3.15: Coverage of LoxP DNA, chromosome 2L, with and without duplicates

out duplicates are also called from the dataset with duplicates, and about 95% of those called with duplicates are also called without duplicates.

The difference is also quite small for the ADAR0 dataset, which has fewer duplicates, as illustrated in the coverage plot in Figure 3.16. Here, the average original error rate is 9.7×10^{-5}, about 1.4 times the average error rate with duplicates removed, 6.8×10^{-5}. The difference in values for θ is 2.8×10^{-6}, and the polymorphism calls in each case are nearly identical: 98% of those called in one set are also called in the other.

3.7.2 Choice of Alignment Software

There are many different software packages designed for aligning short reads to a reference genome. Among the better known are Bowtie [38], BWA [140], Novoalign

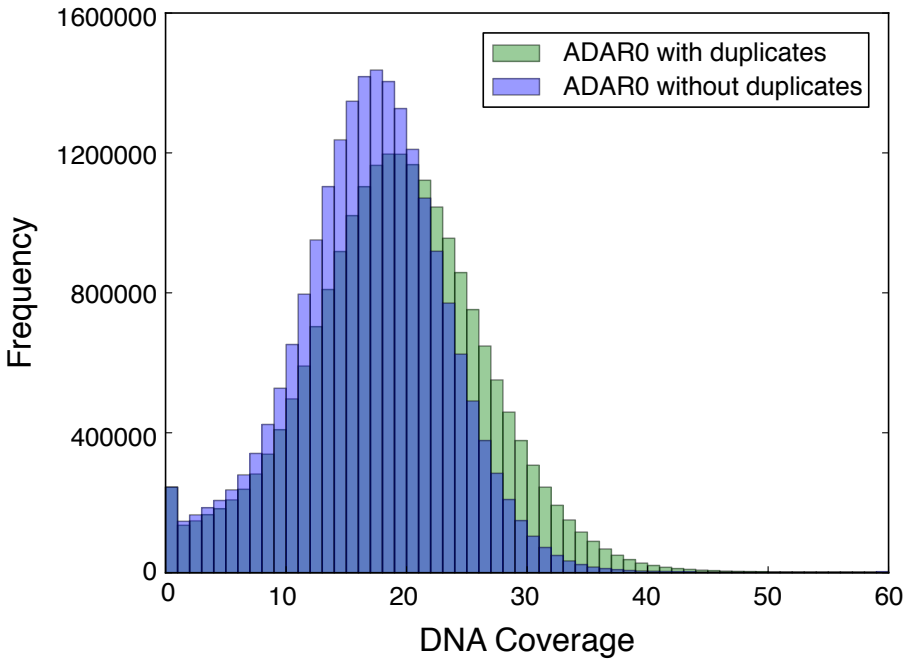

Figure 3.16: Coverage of ADAR0 DNA, chromosome 2L, with and without duplicates

(Novocraft), and Stampy [145]. For mapping DNA reads, we chose BWA, because it allows insertions and deletions in alignments. Because our algorithm takes the alignments of reads as given, changes in alignment software, or in the software settings, could potentially cause changes in the downstream parameter estimates and polymorphism calls.

For aligning RNA reads to the reference, we have to be a bit more careful than with DNA reads. Since many processed RNAs are spliced, and thus constitute discontinuous regions of the genome, RNA reads that span exon-exon boundaries will likely fail to be mapped. To address this issue, we will need to use splice-aware mapping software such as TopHat [172].

3.7.3 Repetitive Regions

Until this point, we have restricted ourselves to looking for polymorphisms and editing sites in unique regions of the genome. This has allowed us to take the alignments of reads to a reference genome as given, since a sequence of length 50 from a unique region is unlikely to map well to more than one place (there are approximately 10^{30} possible sequences of length 50, many orders of magnitude more than the number of nucleotides in the *Drosophila* genome, about 10^8). However, repetitive regions will be of interest in some applications, and are of particular interest to us in the study of ADAR and its interaction with the RNA interference pathway and transposon activity. As previously mentioned, long double-stranded RNAs made from repetitive sequences are used to silence other members of the same repetitive sequence family around the genome. ADAR interferes with this pathway, resulting in lower levels of silencing, and encouraging the activity of transposons [116].

To adapt our methods to repetitive sequence regions, we can no longer take alignments as given, and the source of a given read must become an additional hidden variable. Roughly, we would compute the probability that a read came from any one position using a "conditional aligner," an algorithm based on a hidden Markov model developed in our group with repetitive sequences in mind, which capitalizes on the fact that even highly repetitive sequences can be differentiated by the occasional SNP. Given each assignment for a particular read, the rest of the model would give the probability of the data, from which the hidden DNA and RNA sequences could be inferred. This version of the model would likely benefit from reads longer than 50 nucleotides, and from paired-end reads.

3.7.4 Parametric Model of p_α

We currently model the distribution over polymorphism levels α with a discrete distribution, learning a parameter for each polymorphism level
$\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. Alternatively, we could adopt a parametric model such as a beta distribution to describe the probability of any polymorphism level between 0 and 1. The beta distribution would be a natural choice, given that it is a flexible distribution over the interval [0,1]. The goal would then be to infer the parameters of this beta distribution. We might consider treating fixed polymorphisms separately, as we do in the full model, so that the beta distribution would give us the probability of an unfixed polymorphism at level $\alpha \in (0, 1)$.

3.7.5 Modeling Errors

We chose to model the sequencing errors based on sequence context, but there are a few other factors known to affect error rates: the factors typically considered are the quality score given by Illumina, and the "cycle" of the base call, how far along the read the base occurs, with errors tending to crop up in the first few positions and in later cycles [59, 69]. We can examine each of these effects using PhiX control runs.

The cycle effect is shown in Figure 3.17, which displays the average error rate for base calls in each of the 50 read cycles. The error rates in purple are for the original PhiX dataset, and those in blue are the errors after we apply our quality filter, disposing of any read that contains base calls with quality below 30. By imposing this restriction, the effect of cycle is greatly reduced, especially for late cycle base calls. The first few cycles still contain relatively high error rates, but in processing the data for EM, we throw out the first two base calls of each read since

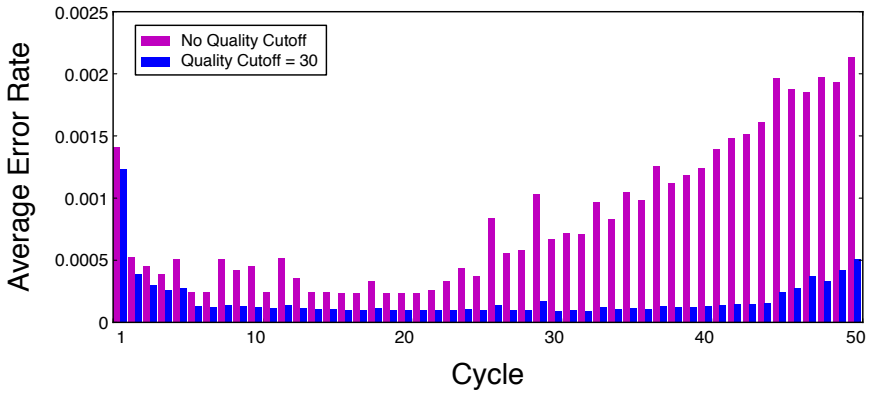

Figure 3.17: Average error rate by read cycle for PhiX control run, with and without quality cutoff

we do not have full 3-nucleotide sequence contexts for those positions. Nevertheless, a dependence on cycle could improve our algorithm, and would be relatively simple to implement by learning an error model like the one we learn now for each of the 50 cycle positions. This would take us from 256 error parameters to 12,800. Alternatively, we could model three difference cycle categories, beginning, middle, and end, for a total of 768 parameters.

Figure 3.18 displays the average mismatch rate by the reported quality score for a PhiX control run, and Figure 3.19 zooms in on those qualities above 30. We see a general, yet imperfect trend towards lower error rates with higher quality scores, and once we reach a quality score of 30, we see errors that are orders of magnitude less frequent than those at lower qualities. This supports our decision to use our strict quality cutoff. Even above quality score 30, however, errors continue to decline as quality score increases. This suggests that we may want to include the quality score as a variable in the future. As with the cycle variable, implementation would be relatively straightforward; we could learn a context-and-cycle specific error rate for each quality score, or for binned quality scores to reduce dimensionality.

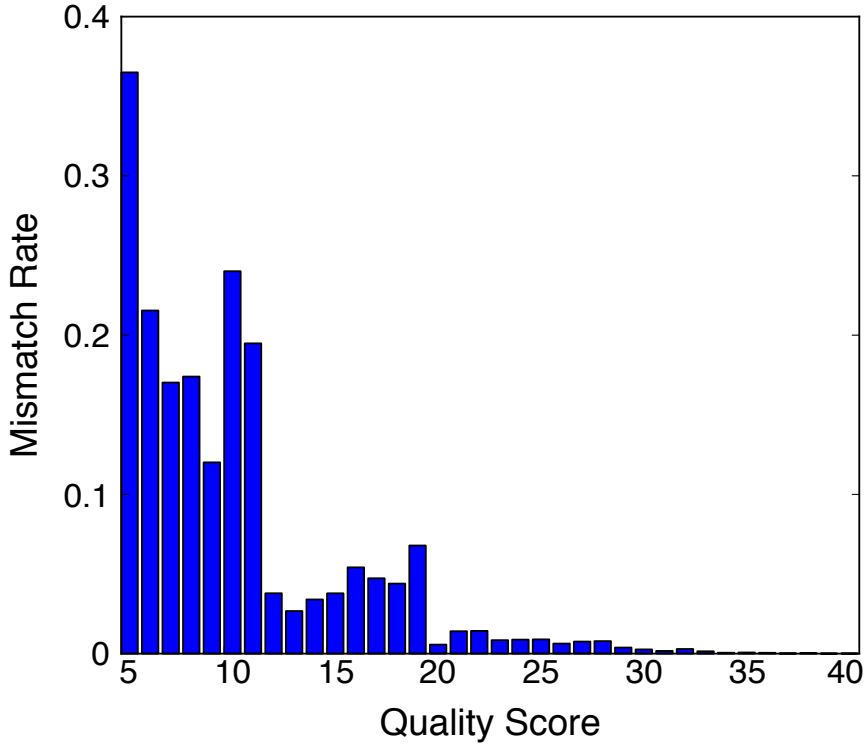

Figure 3.18: Average error rates by Illumina quality score

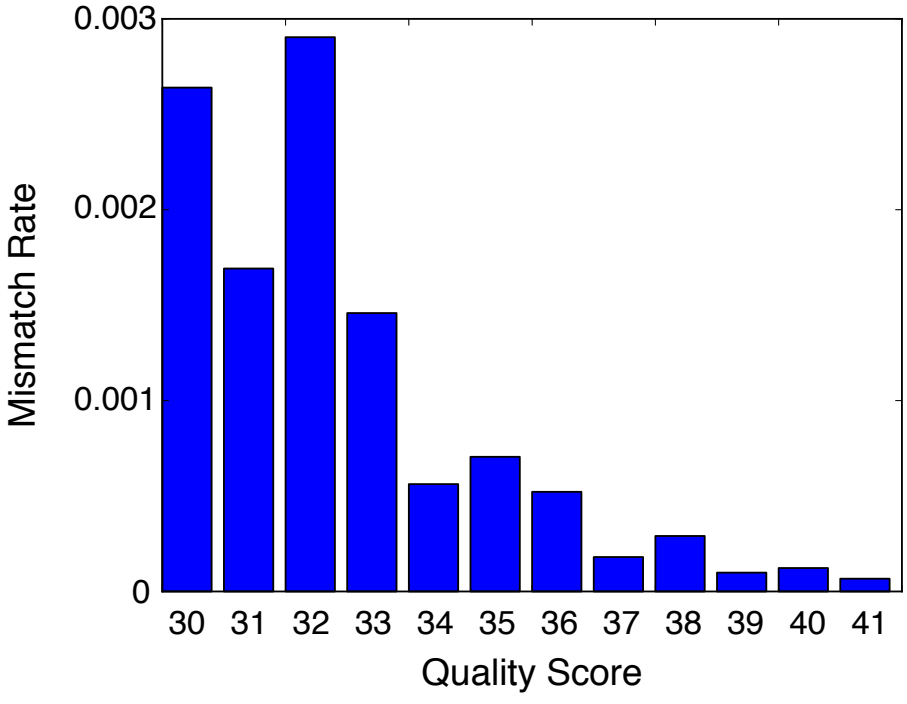

Figure 3.19: Average error rates by Illumina quality score, quality score ≥ 30

Given the large amount of data that Illumina provides, estimating this many parameters would likely be feasible, but it might not improve our model noticeably. One possible exception might be the inclusion of cycle-specific parameters in cases where a preliminary PhiX study shows spikes at particular cycles. This was something we noticed in a couple of PhiX runs, possibly indicating that something in the sequencing machinery went slightly awry for those particular cycles. If this effect is compounded by the generation by restriction enzymes of reads with the same coordinates, this could likely lead to false positives. Modeling cycle-specific error rates would adequately control for this.

3.7.6 Modeling SNPs

In the current model, we learn a single common probability for each possible type of polymorphism, $A \rightarrow C$, $C \rightarrow G$, $T \rightarrow A$, etc... There is no reason, however, that we cannot learn different mutation rates for each type, or for general categories of mutations. For example, the exchange of one purine (A or G) for another, and the exchange of one pyrimidine (C or T) for another, known as "transitions" are known to be more frequent than "transversions," substitutions from purine to pyrimidine and vice versa [133, 130] Indeed, we see this trend in our own polymorphism calls. This is partially due to the fact that transitions are more likely than transversions to result in "silent" substitutions, which do not result in an amino acid change, and are therefore less likely to be removed through selective pressure. A revised model of polymorphisms could learn three parameters in place of the one parameter θ that we currently learn: θ_1, the probability of no polymorphism, θ_2, the probability of a transition, and θ_3, the probability of a transversion.

3.7.7 Model Assumptions

We make a few assumptions in the full model in the interest of computational simplicity that are worth a closer look. We assume that at any given position, only one strand (either Watson or Crick) provides a template for the RNA. Given recent evidence of pervasive transcription across the genome [124], this assumption is likely incorrect, however we also assume that that the strand at a given position is independent from the strands transcribed at all other positions. While this is also not the case, since RNAs span many positions, this assumption has the effect of mitigating the first. The issue of which strand is transcribed only applies to positions in the genome with either an A or a T, only one of which can happen at a given time. If it happens that both strands are transcribed over a region containing edited As on both the forward and reverse strands, then at each position, we have the freedom to choose the strand containing the A without being biased by the fact that a nearby edit position came from the other strand, as we would without the strand independence assumption.

We make these assumptions because of the loss of strand-specific information through the generation of a cDNA library. However, some protocols generate strand-specific RNA reads, in which case neither of these assumptions is needed. In such a situation, the strand that the RNA read maps to represents the actual sequence that was in that RNA before it was turned into cDNA. Given strand-specific reads, we do not need to infer the strand transcribed at each location, so our model becomes a bit simpler, and we deal with the strand-specificity of ADAR at the read generation step of the model, instead of within the prior model. The changes to the model given strand-specific reads are described in Appendix E.

Our independence assumptions go past the transcribed strand. We also assume

that the random variables G_i and R_i at each position are independent of those at other positions conditioned on the reference sequence. This means that we take the probabilities of both a polymorphism and an editing event to be constant across the genome and transcriptome. However, polymorphisms are known to cluster in both fly and mammalian genomes in areas called "hotspots" [161, 30, 32], and ADAR binding sites are also known to cluster, with many targets in a single transcript, and many transcripts with no targets at all. A more complicated model could attempt learn the locations of these clusters via a hidden Markov model (HMM), or might model the probability of a polymorphism or editing event as a function of the probability of the corresponding event at previous positions. As it stands, our model will be conservative, possibly failing to call true events, but will be less likely to make false positive predictions.

CHAPTER **Four**

Inferring Secondary Structure from Multiple Related Sequences

4.1 Motivation and Background

It has long been appreciated that secondary structure plays an important role in the function of many RNAs. Furthermore, this functional importance means that structure is often conserved across multiple related RNAs, even when the sequence may not be. Understanding RNA secondary structure can help us understand cellular processes; predict RNA-DNA, RNA-RNA, and RNA-protein interactions; and identify members of RNA families that perform similar functions or are regulated in the same way. Accordingly, much work has been done to try to predict RNA secondary structure both for single sequences and for multiple related sequences. The algorithm described in this chapter, RNAG, finds a consensus structure from a set of

related, unaligned RNA sequences using a blocked Gibbs sampling algorithm, outperforming other structural prediction methods. The work described in this chapter was published in *Bioinformatics* in 2011 [176] and was presented at the American Society for Microbiology conference on regulating with RNA in bacteria in 2011.

4.2 Secondary structure prediction

At the most basic level, to infer the secondary structure of an RNA is to determine the set of base pairs that are present. Figure 4.1 shows two ways in which this base-pairing information is commonly displayed. A shows the structure in diagram format, which gives an idea of what the RNA looks like in two dimensions by bringing base-paired positions in proximity and connecting them. B shows the structure in circular format, focusing more directly on the set of base pairs that are present by representing the sequence on the outer circle and drawing arcs to connect paired bases.

Most folding algorithms impose two constraints; one is a physical restriction, and the other is artificial, but necessary to make computation feasible for sequences of even moderate length. For a sequence of length N, let $I_{i,j}$ be an indicator variable that is 1 if bases i and j are paired, and is 0 otherwise. Then the two constraints are as follows:

- $\forall i : \sum_{j=1}^{N} I_{i,j} \leq 1$
- if $i < j < k < l : \; I_{i,k} I_{j,l} = 0$

The first constraint says that any position may pair with at most one other position.

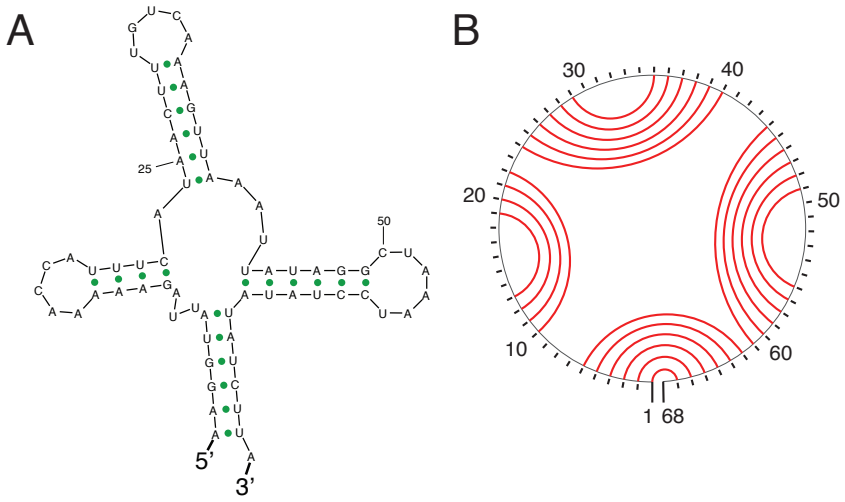

Figure 4.1: RNA secondary structure diagrams for Human tRNA

The second prevents so-called "pseudoknots," represented in the circular representation format as crossing arcs. Pseudoknots are not prevented from forming in actual RNAs, but most folding methods rely on dynamic programming algorithms, for which this constraint is necessary.

4.2.1 Single sequence

The first widely-accepted structural prediction algorithm, Mfold [184], uses thermodynamic parameters to evaluate the energies of small-scale structures like base pairs and stem loops, then employs a dynamic programming algorithm to compute the folding energy of whole sequences by recursively computing the energies of substructures, using a backtracing step to find the complete structure with the minimum free energy (MFE). Similarly, RNAfold [57] uses dynamic programming to predict the MFE structure. There is a significant downside, however, to relying on the MFE, which is a result of the high-dimensional space in which the prediction lives. In such

a space, even the most probable structure will likely have extremely low probability, and there is no principled reason why such a structure should be representative of the posterior-weighted space of structures [12]. To remedy this problem, other algorithms look for structures that are somehow representative of this posterior space. Sfold [18, 20] draws statistically representative structures from the posterior space, then uses a large set of sampled structures to compute a "centroid" estimate, the structure that minimizes the total base pair distance to all other structures. The centroid is also the closest valid structure to the posterior mean [12]. To further characterize the posterior space, Webb-Robertson *et al.* [174] introduce Bayesian credibility limits based on hamming distance that can characterize the uncertainty associated with an estimate such as the centroid. An alternative approach is taken for CONTRAfold [23], which relies solely on a discriminative probabilistic model in place of the thermodynamic model used in both Mfold and Sfold. They introduce the maximum expected accuracy (MEA) estimator, which allows a user-defined parameter γ to specify the relative weights of correctly paired and unpaired positions in the definition of "accuracy," then finds the structure that maximizes this accuracy under the distribution of structures conditioned on the sequence of interest. Hamada *et al.* [86] refine this idea and introduce the γ-centroid, the structure that maximizes the expected value of $\gamma TP + TN$, where TP and TN are the number of true positive and true negative base pairs, respectively.

4.2.2 Multiple sequences

When multiple sequences are thought to have a common, or "consensus" structure, the task of inferring this structure is in some ways easier and in other ways more difficult. It is easier because of the addition of comparative sequence information;

even early on, it was acknowledged that a single sequence alone was inadequate for predicting structure with much confidence [184]. The difficulty, however, comes in the form of an extra set of unknowns: the alignment of the multiple sequences.

Some algorithms avoid this issue entirely by starting from a fixed alignment. Most of these take advantage of sequence covariation, positions in the multiple sequences showing compensatory mutations, thereby indicating the presence of conserved base pairs. Gutell *et al.* [100] capitalize on this by computing the mutual information between two alignment columns, which can be interpreted as a log likelihood ratio for which statistical significance can be calculated. Cary and Stormo [13, 60] propose a graph theoretic algorithm which finds the set of base-pairing interactions maximizing the sum of mutual information, with the added benefit of allowing pseudoknots. Pfold [135, 136] uses a stochastic context free grammar together with a phylogenetic mutation model to compute a MAP estimate of the consensus structure.

Other algorithms incorporate the thermodynamic energy model to improve predictions. One such algorithm, KNetFold [7], uses a hierarchical network of k-nearest neighbor classifiers to predict whether two alignment columns form a pair in the consensus structure based on features including mutual information and the MFE structures predicted for each sequence by RNAfold [57]. An extension of Pfold, PETfold [165] uses thermodynamic parameters to find the structure that maximizes the expected combined accuracy of the consensus structure with respect to the Pfold model, and the individual sequence structures with respect to the energy model. RNAalifold [127, 107] modifies the recursions of the traditional thermodynamic energy model by replacing the pairing energy with a linear combination of the pairing energy and a base-pair conservation score. RNAalifold finds the centroid structure, and allows for stochastic backtracing for drawing representative sample structures.

Many algorithms also exist for multiple sequence alignment, including Clustal [96] and ProbCons [42], the latter of which takes advantage of the maximum expected accuracy alignment to avoid the pitfalls of maximum likelihood estimators. Neither of these algorithms, however, takes advantage of any structural knowledge, which can substantially improve alignment accuracy [151]. One method for inferring a multiple sequence alignment in a structure-aware manner involves folding each sequence individually, then comparing pairwise structures, building up to a multiple alignment using a library of weighted alignment edges [166, 119]. Another approach uses a covariance model (CM) [24], an extension of the probabilistic context-free grammar in which each node corresponds to an alignment column instead of a single nucleotide. Covariance models are implemented in the Infernal software package [152].

Of course, a reliable consensus structure depends on a reliable multiple alignment, and vice versa. Thus, the task of simultaneous alignment and folding of a set of unaligned sequences is a difficult one. Many strategies have been proposed; a popular class of algorithms is based on one by Sankoff [163] which uses a computationally intensive dynamic programming algorithm that is impractical in most situations. Modifications to the Sankoff algorithm include Foldalign [125], LocARNA [104], Murlet [55], and RAF (RNA Alignment and Folding) [22]. Other algorithms avoid having to simultaneously address structure and alignment by iterating between the two. Eddy and Durbin [24] use covariance models to iteratively update the alignment and consensus structure via expectation-maximization. CMfinder [180] extends this idea to RNA motif-finding. Others use Markov chain Monte Carlo methods, including SimulFold [149], which samples from the joint posterior distribution of RNA structures, alignments, and phylogenetic trees, and MASTR [143], which uses simulated annealing to minimize a cost function that considers sequence conservation,

covariation, and base-pairing probabilities. Another approach, RNA Sampler [178], iteratively samples aligned stems based on a stem conservation score, then updates the conservation scores based on the samples.

Gibbs sampling, introduced by Geman and Geman [118] is a Markov chain Monte Carlo (MCMC) procedure useful in situations where the conditional probability of each random variable given the rest is considerably easier to compute than the joint probability over all random variables. In the case where sets of random variables can be sampled together instead of individually, convergence is guaranteed in as few or fewer iterations [144]. This "blocked" Gibbs sampling strategy will translate to an improvement in computational expense as long as efficient methods exist for sampling each block of variables. What we introduce here is a blocked Gibbs sampling procedure that iteratively samples a consensus structure given a multiple alignment and vice versa. We then use representative samples to characterize the posterior space via hierarchical clustering and γ-centroid estimators.

4.3 Methods

4.3.1 RNAG

Let A represent a multiple alignment of sequences Q, and let S be their consensus structure. Our ultimate goal is to draw samples from the joint posterior distribution, $P(A, S|Q)$. RNAG achieves this goal by iteratively sampling from the conditional probabilities $P(S^{(t)}|A^{(t-1)}, Q, \Theta_S)$ and $P(A^{(t)}|S^{(t)}, Q, \Theta_A)$ at iteration t, where Θ_S and Θ_A are parameters involved in the structure and alignment sampling steps, respectively. We use RNAalifold [107] and the Infernal package [152] to carry out

each of these steps, and we find an initial multiple alignment from the unaligned sequences Q using ProbCons [42]. The algorithm proceeds as follows:

1. Initialize: create a multiple sequence alignment $A^{(0)}$ with ProbCons
2. Iterate:
    a) Sample $S^{(t)}$ from $P(S^{(t)}|A^{(t-1)}, Q, \Theta_S)$ with RNAalifold, with default covariation weight Θ_S
    b) Use cmbuild from the Infernal package to design a covariance model based on $A^{(t-1)}$ and $S^{(t)}$, learning parameters Θ_A via EM
    c) Use cmalign from the Infernal package to sample a multiple alignment according to $P(A^{(t)}|S^{(t)}, Q, \Theta_A)$

This algorithm uses RNAalifold and Infernal, but in theory, any algorithm that probabilistically samples from one of the conditional distributions could be substituted. We chose these two because there are few algorithms that permit statistical sampling from the posterior space. We chose ProbCons for the initial alignment step based on a preliminary study using a dataset from Kiryu *et al.* [55] in which ProbCons out-performed ClustalW [96]. Note that this is not actually a fully Bayesian procedure: instead of sampling the model parameters Θ_A from a conditional probability distribution given A and S, we use EM to optimize them. RNAG is available as a web service at http://ccmbweb.ccv.brown.edu/rnag.html.

4.3.2 Centroid Estimators

By nature, the RNAG algorithm allows for sampling of secondary structures, which can be used to characterize the posterior space. Since convergence is difficult to

assess, we allow for a burn-in period of 1,000 iterations, after which we draw samples. From 1,000 sampled structures, we compute empirical base pair frequencies from which centroid estimators [18] can be built. We also compute 95% credibility limits around the centroid estimators as in Newberg and Lawrence [153], interpreted as the radius of the smallest hypersphere centered at the centroid estimate and covering 95% of the posterior weighted space.

We can also use the samples to assess the bias-variance tradeoff inherent in predictions using finite data. Using hamming distance as our measure, which for binary variables is equivalent to squared-error distance and p-th power error distance [12], we measure bias (with respect to a reference structure) and variance using the centroid structure in place of the mean, since the mean is unlikely to be a valid secondary structure, and the centroid is the valid structure that is nearest to the mean.

We also calculate γ-centroid estimators [86] to explore the performance of RNAG over a range of sensitivities and positive predictive values, and compare this performance to other structural prediction methods. Given a reference structure, we let TP, TN, FP, and FN be true positive, true negative, false positive, and false negative base pairs, respectively. Then sensitivity (SEN) and positive predictive value (PPV) are defined as follows:

$$SEN = \frac{TP}{TP + FN} \tag{4.1}$$

$$PPV = \frac{TP}{TP + FP} \tag{4.2}$$

Note that as we include more base pairs in a prediction, we will improve sensitivity, but PPV will suffer. Following Do *et al.* [22], we average the PPV and sensitivity for each sequence, weighting all sequences equally. By varying $\gamma \in \{2^k : -5 \leq k \leq$

$10\} \cup \{6\}$ as in Hamada *et al.* [123], we can linearly interpolate a curve on the PPV-SEN plane and calculate the area under the curve as a qualitative measure of performance.

4.3.3 Clustering Analysis

We looked for clusters of structures in the posterior-weighted space, as often the ensemble of structures can be partitioned into distinct classes [20]. Following Hamada *et al.* [86], since comparison of consensus structures may be ambiguous, we mapped the consensus structures directly onto the individual sequences, then applied hierarchical clustering as in Ding *et al.* [19] using the Diana method [131] and the CH index [11] to determine the number of clusters. To assess how well separated clusters are, we use the following separation index:

$$S = \frac{D}{C_1 + C_2} \tag{4.3}$$

where D is the hamming distance between the centroids of two clusters, and C_1 and C_2 are the 95% credibility limits around the cluster centroids. We say that the clusters are well separated if $S \geq 1$, indicating that no more than 5% of the structures from one cluster are within the 95% credibility limit of the other.

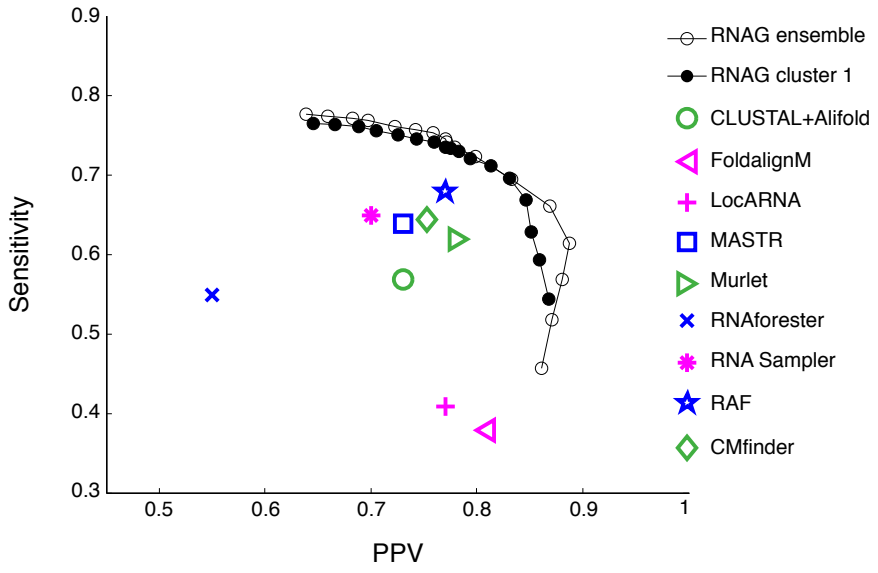

Figure 4.2: Comparison of secondary structure prediction methods on MASTR dataset

4.4 Results

4.4.1 Accuracy comparison

We compared RNAG with other structure prediction algorithms on two datasets, one from Lindgreen *et al.* (MASTR) [143] and one from Gardner *et al.* (BRAliBASE II) [94]. Structure predictions from other algorithms were given by Do *et al.* [22]. Both datasets consist of sets of sequences from RNA families in the Rfam database [103], so PPV and SEN results are averaged over the families. In both cases, reference structures were provided along with the datasets. Figure 4.2 shows the relative performance of RNAG on the MASTR dataset, and Figure 4.3 shows its relative performance on the BRAliBASE II dataset. In addition to the algorithms tested by Do *et al.* [22], we also tested CMfinder [180], run with default settings, even though its primary purpose is motif finding. CMfinder is also capable of predicting global structure, so we included it because of its algorithmic similarity to RNAG.

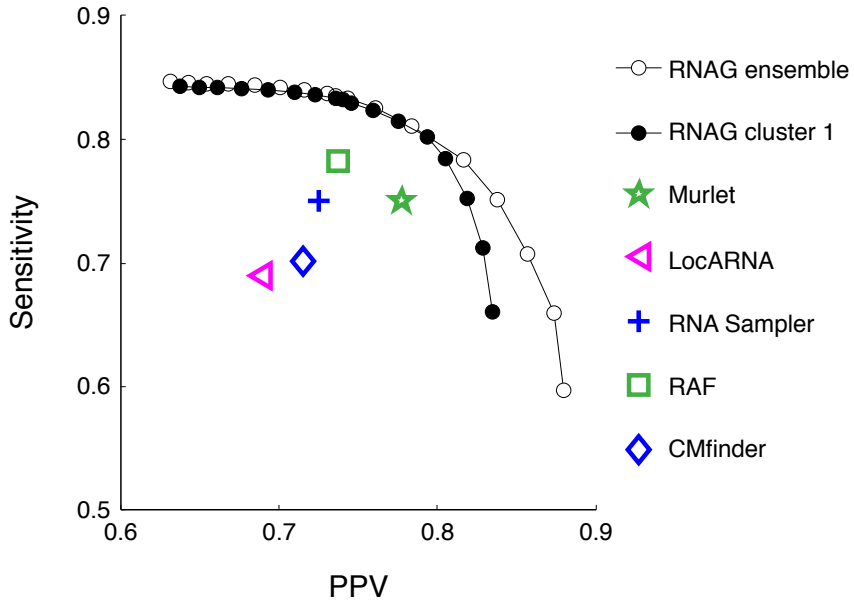

Figure 4.3: Comparison of prediction methods on BRAliBASE II dataset

Other prediction methods consistently fall beneath the RNAG frontier, indicating that RNAG is providing a better trade-off between sensitivity and positive predictive value, offering more accurate predictions overall. Unsurprisingly, this is not the case 100% of the time; if we separate the four Rfam families in the BRAliBASE II dataset, 13 out of 16 predictions are beneath the RNAG frontier, one lies on the frontier, and two are above, as shown in Figure 4.4.

4.4.2 RNAG performance characteristics

We used a third benchmark dataset curated by Kiryu *et al.* [55] to explore various aspects of RNAG. The dataset contains 85 reference alignments of 10 sequences each, representing 17 RNA families from the Rfam database [103]. Sequences ranged from 51 to 291 bases long, and sequence identity within families ranged from 40% to 94%.

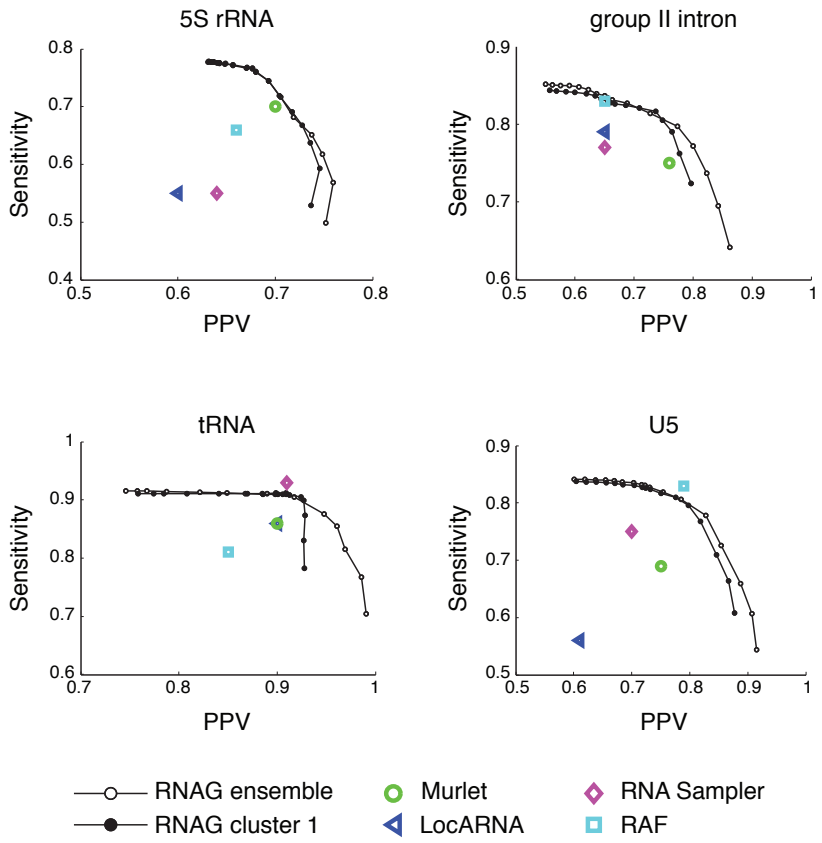

Figure 4.4: Comparison of prediction methods on specific families in BRAliBASE II dataset

We first looked at the effect of the number of input sequences on accuracy, as measured by the area under the interpolated PPV-SEN curve, as well as bias, variance, and credibility limits. For $2 \leq N < 10$, we drew 10 random sets of N sequences from each of the 85 reference alignments, and averaged the results from RNAG over the 10 runs. For $N = 10$, we ran RNAG once for the 10 sequences in each of the reference alignments. The results are in Table 4.1. Unsurprisingly, we see that as the number of input sequences increases, the structural predictions improve across the board, with diminishing returns as the number of sequences nears 10. This is further illustrated in Figure 4.5, which displays the average PPV-SEN curves for 2, 4, 6, 8, and 10 input sequences. From the table, we also see that the credibility limits shrink as we add additional sequences, as do bias and standard deviation. Finally, by looking at the number of the 1,000 sampled structures falling in the first and second clusters, we see that in general we are able to capture about 90% of the posterior-weighted space in just two clusters. When different sequence families are addressed independently, we find that the accuracy gains are greater for families with lower sequence identities, reflecting the fact that the additional sequences are providing valuable covariation information for the algorithm to capitalize on.

Number of sequences	Area under PPV-SEN curve			Bias	SD	95% credibility limit			Number of sampled structures		
	Ensemble	First cluster	Second cluster			Ensemble	First cluster	Second cluster	First cluster	Second cluster	First+second cluster
2	0.44	0.46	0.37	0.27	0.04	0.21	0.14	0.11	728.13	150.76	878.89
3	0.58	0.59	0.49	0.20	0.03	0.14	0.10	0.07	793.15	124.94	918.09
4	0.58	0.58	0.48	0.20	0.03	0.14	0.09	0.06	791.66	115.00	906.66
5	0.62	0.63	0.51	0.17	0.03	0.12	0.08	0.05	802.20	113.24	915.44
6	0.67	0.67	0.54	0.16	0.03	0.11	0.07	0.05	800.50	111.66	912.16
7	0.70	0.69	0.57	0.15	0.03	0.10	0.07	0.05	795.52	111.92	907.44
8	0.73	0.71	0.60	0.15	0.03	0.10	0.07	0.04	797.56	116.19	913.75
9	0.73	0.73	0.60	0.14	0.02	0.09	0.06	0.04	790.59	122.38	912.97
10	0.75	0.74	0.63	0.13	0.02	0.09	0.06	0.04	792.85	125.11	917.96

Table 4.1: RNAG results for increasing number of input sequences. We normalize bias, standard deviation (SD), and credibility limits by the average sequence length for the family. The last three columns give the number of the 1,000 sampled structures that fall within the specified cluster.

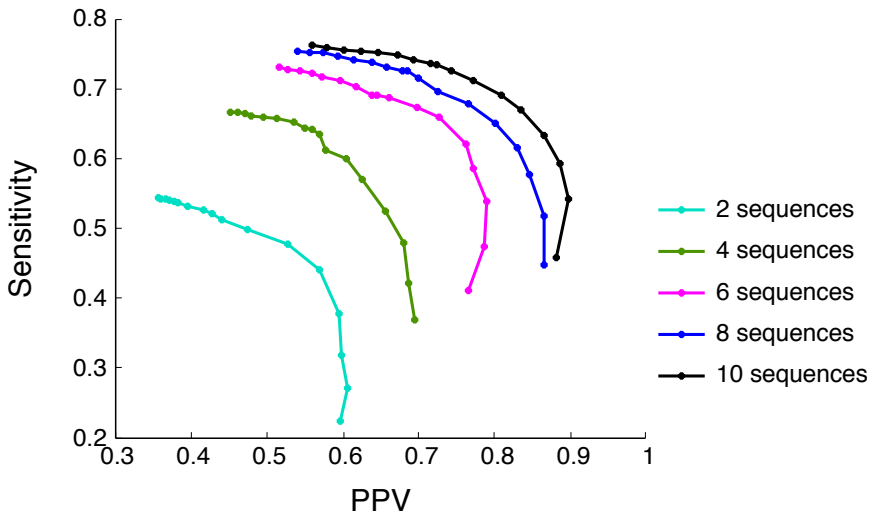

Figure 4.5: Improved PPV-SEN curves with increasing number of input sequences. Curves are shown for the ensemble centroids.

We learn more about the performance of RNAG and the characteristics of RNA families by evaluating RNAG on the families individually, as shown in Table 4.2. First, there is considerable variability in the under-curve areas and biases, which indicates that the accuracy with which we are able to predict the reference structure varies for different families. Figure 4.6 shows a strong correlation between bias and under-curve area, with larger bias associated with smaller area, both reflecting decreased accuracy. Second, the results of RNAG are more precise than they are accurate. The 95% credibility limits are small, indicating that clusters are tightly compact around their centroids, while the biases are larger than these limits. Finally, for 11 families, there are two well-separated clusters (with separation index ≥ 1), a further reminder that point-estimate prediction methods with no analysis of uncertainty will risk missing large portions of the posterior-weighted space of structures for these families.

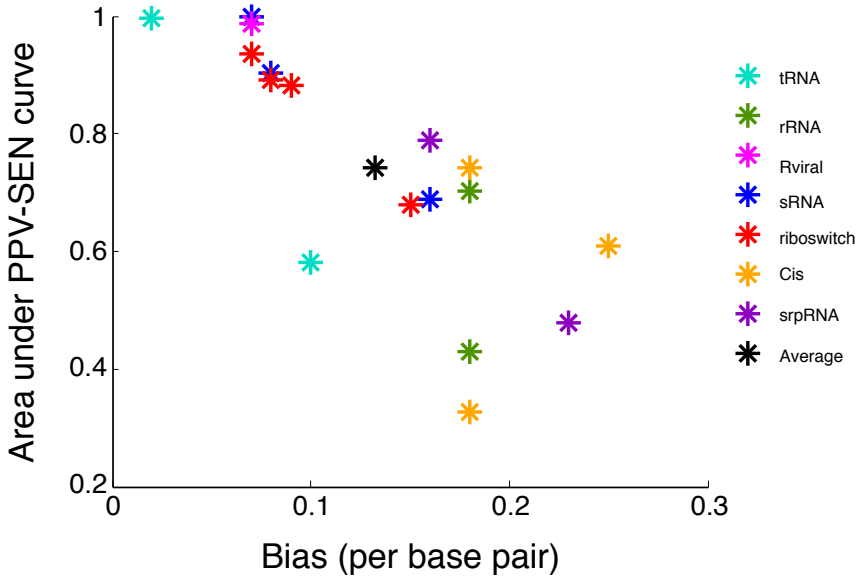

Figure 4.6: Bias vs area under PPV-SEN curve for ensemble centroids for 17 RNA families.

RNA family	RNA type	Average sequence length	Percent identity	Bias	SD	95% credibility limit			PPV-SEN area			Separation index
						Ensemble	First cluster	Second cluster	Ensemble	First cluster	Second cluster	
T-box	tRNA	244	45	0.10	0.01	0.06	0.04	0.02	0.58	0.55	0.47	1.00
t-RNA	tRNA	73	45	0.02	0.01	0.03	0.01	0.01	1.00	0.99	0.91	2.50
5S-rRNA	rRNA	116	57	0.17	0.02	0.07	0.05	0.03	0.70	0.70	0.67	0.88
5-8S-rRNA	rRNA	154	61	0.18	0.03	0.14	0.10	0.08	0.43	0.42	0.26	0.56
Retroviral-psi	Rviral	117	92	0.07	0.05	0.15	0.11	0.05	0.99	0.99	0.47	1.25
U1	sRNA	157	59	0.16	0.02	0.06	0.06	0.02	0.69	0.69	0.63	1.13
U2	sRNA	182	62	0.08	0.02	0.05	0.05	0.02	0.90	0.90	0.71	1.14
Sno-14q-I-II	sRNA	75	64	0.07	0.03	0.12	0.08	0.07	1.00	0.92	0.86	0.47
Lysine	riboswitch	181	49	0.07	0.02	0.06	0.05	0.03	0.94	0.93	0.84	0.88
RFN	riboswitch	140	66	0.15	0.03	0.11	0.06	0.06	0.68	0.64	0.60	0.58
THI	riboswitch	105	55	0.08	0.02	0.07	0.06	0.02	0.89	0.88	0.75	1.13
S-box	riboswitch	107	66	0.09	0.02	0.07	0.03	0.03	0.88	0.87	0.74	1.17
IRES-HCV	Cis	261	94	0.25	0.05	0.21	0.16	0.08	0.61	0.58	0.44	1.00
SECIS	Cis	64	41	0.17	0.02	0.08	0.02	0.02	0.74	0.71	0.72	1.50
UnaL2	Cis	54	73	0.18	0.03	0.06	0.02	0.02	0.33	0.62	0.61	1.00
SRP-bact	srpRNA	93	47	0.16	0.03	0.12	0.04	0.04	0.79	0.78	0.70	2.75
SRP-euk-arch	srpRNA	291	40	0.23	0.01	0.04	0.03	0.02	0.49	0.48	0.47	0.80
Average		142		0.13	0.02	0.09	0.06	0.04	0.76	0.74	0.63	0.90

Table 4.2: RNAG results for individual sequence families. Bias, SD, and credibility limits are normalized as in Table 4.1.

4.5 Extension to RNA motif finding

While RNAG looks for the global sequence and structural characteristics common to a set of sequences, often RNA structural signals are local. Many RNAs involved in regulation contain short structure and sequence motifs, existing as substrings of longer RNA transcripts. One example is the riboswitch, located in the 5' UTR of particular genes, which changes conformation based on the absence or presence of a ligand, thereby regulating the expression of its own downstream gene [179]. Another example is the CRISPR array, a set of near-identical stem loops nested between unique spacer regions, that play a crucial role in the bacterial immune system [147].

Finding multiple instances of a substring common to a set of sequences is known as the motif-finding problem, which has been very well-studied for sequence patterns in DNA. Motif-finding was introduced in 1989 for identifying recognition patterns for DNA-binding proteins like transcription factors given only a collection of DNA fragments, each known to contain exactly one binding site [169]. Since then, fully probabilistic Gibbs sampling algorithms [40] have been developed, as have expectation-maximization algorithms such as MEME [109]. For regulatory RNAs, the added complexity of structure makes motif-finding more difficult. Given a set of RNA sequences assumed to have a similar structure or function, either because they are orthologs across species [15] or are related sequences within a single species [76], we are now faced with the challenge of aligning substrings and inferring common structure and sequence patterns. Figure 4.7 illustrates this challenge with a toy example; we are tasked with identifying potentially subtle signals with primary base sequences that may differ substantially.

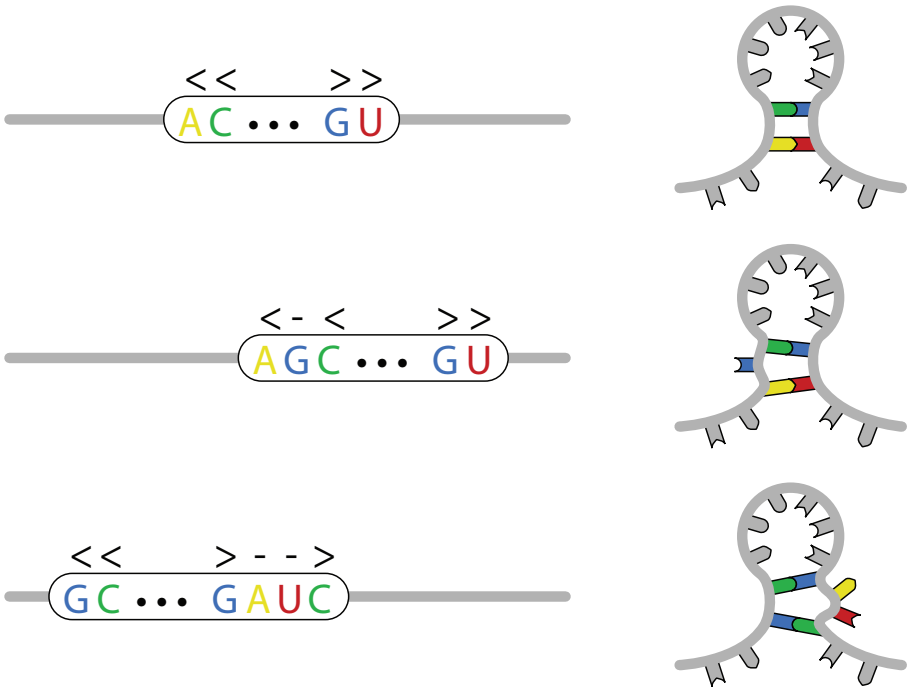

Figure 4.7: Illustration of the challenges of RNA motif-finding. Nested brackets indicate base pairs.

4.5.1 RGibbs

We implemented a preliminary RNA motif finder called RGibbs, which follows a similar framework to that of RNAG. We use a blocked Gibbs sampler to iteratively sample three parameters: the consensus structure (S), the multiple alignment of subsequences (A), and the location of the motifs in each sequence (L) for a set of input sequences Q. The algorithm is initialized with a structurally annotated local sequence alignment $(A^{(0)}, S^{(0)}, L^{(0)})$, and proceeds roughly as follows at iteration t:

1. Compute a background model $B^{(t)}$ of nucleotide composition outside of motif regions

2. Design a covariance model $CM^{(t)}$ for the motif regions based on $A^{(t-1)}$, $S^{(t-1)}$, and $L^{(t-1)}$, learning model parameters via EM with cmbuild (Infernal package)

3. Sample new motif locations $L^{(t)}$ from $P(L^{(t)}|CM^{(t)}, B^{(t)}, Q)$ using cmsearch (Infernal package)

4. Sample a multiple alignment $A^{(t)}$ from $P(A^{(t)}|L^{(t)}, CM^{(t)}, Q)$ with cmalign (Infernal package)

5. Sample a new consensus structure $S^{(t)}$ from $P(S^{(t)}|A^{(t)}, L^{(t)}, Q)$ with RNAalifold [107]

As with RNAG, RGibbs iterates for a burn-in period, then draws samples that can be used to make predictions. It extends RNAG to the problem of motif-finding by building into the covariance model a background state to and from which we can transition. This lets us predict multiple instances of a motif in a sequence, and allows for the possibility of some sequences having no motifs.

4.5.2 Analysis of Fly RNA Families

We conducted preliminary tests of RGibbs on sequence families from Rfam [103] to compare it to CMfinder [180], its major competitor. We first used CMfinder's own test set, 19 Rfam families, from which we selected ten sequences chosen to minimize average sequence identity (for one family, only 9 sequences were available). For each of these families, we embedded the sequences in 200 bp of flanking nucleotides, then predicted at most one site in each sequence. For preliminary purposes, we counted a true positive if the algorithm predicted a non-null structure— at least one base pair— and the motif site overlapped the true site by at least 50%. We counted a false positive if a non-null structure was predicted that did not overlap the true site by at least 50%. Both algorithms performed well at this task, as shown in Table 4.3.

	CMfinder		RGibbs	
	True Positives	False positives	True Positives	False positives
Site+Flanking	196 (93%)	1 (0.5%)	167 (88%)	2 (1%)
Flanking Only	N/A	134 (71%)	N/A	30 (16%)
Shuffled Control	N/A	142 (75%)	N/A	25 (13%)

Table 4.3: Performance of CMfinder and RGibbs on 19 Rfam families

	CMfinder		RGibbs	
	True Positives	False positives	True Positives	False positives
Site+Flanking	356 (92%)	18 (5%)	355 (93%)	10 (3%)
Flanking Only	N/A	293 (75%)	N/A	122 (31%)
Shuffled Control	N/A	273 (70%)	N/A	61 (16%)

Table 4.4: Performance of CMfinder and RGibbs on 36 *Drosophila* Rfam families

Next, we tested each algorithm's false positive rate in sequences with no motifs. We did this in two ways: first, we simply excised the true sites from the flanking sequence, and ran the motif finders on the remaining sequence. Second, we aligned the sequences with MUSCLE [25], a structure-unaware alignment algorithm, and shuffled the columns of the alignment before running the motif finders. The idea of the second control was to preserve sequence conservation, but break up any potential structural elements present in the flanking regions. The results from CMfinder and RGibbs are also in Table 4.3. For these tasks, CMfinder calls approximately 5 fold more false positives than RGibbs.

We carried out the same tests for a new dataset consisting of 36 Rfam families with sequences from *Drosophila* only, collecting sequences from up to twelve *Drosophila* species per family for a total of 389 sequences. We also used optimal sequence weights for the Infernal package, as suggested by Newberg *et al.* [154]. As shown in Table 4.4, both algorithms do well at finding real sites as before, and again, CMfinder has a harder time avoiding false positives in both controls.

Since we have only tested the accuracy of predicted motif-site locations with

RGibbs,, further work is needed to assess the relative accuracy of each algorithm's structural predictions. We are optimistic, however, since RGibbs follows the same framework as RNAG, which outperforms CMfinder and other global structural prediction methods.

4.5.3 Possible extensions

There are many ways in which RGibbs may be improved and extended. Currently, the model parameters for the covariance model are optimized through expectation maximization. Sampling them from a conditional probability distribution given the other variables would result in a fully Bayesian motif finder, allowing us to explore the full posterior space. There are also features of DNA motif finding that would be applicable here, including allowing for multiple distinct motifs included in sequences in a combinatorial fashion and considering the phylogenetic relationship between the input sequences.

4.6 Discussion

RNAG provides many advantages over other structural prediction methods working from unaligned sequences. First, it avoids maximum likelihood point estimates through the use of statistical sampling in place of optimization procedures. In addition, it has a theoretical convergence advantage over other Metropolis-Hastings algorithms due to the blocked Gibbs sampling approach. Unlike methods which use heuristics to restrict the space or attempt to align stems, RNAG is able to sample from the full space of multiple alignments and consensus structures. These advan-

tages are realized in prediction accuracy as illustrated in the PPV-sensitivity curves. In addition, RNAG allows for a fuller characterization of the posterior space than other methods, including cluster analysis and uncertainty analysis in the form of credibility limits and bias and variance calculations. Of course, since RNAG is an MCMC procedure, assessing convergence is a difficult task.

4.6.1 Dataset Limitations

The three datasets we selected were published independently, as were the results of other prediction methods— a choice we made to avoid self-serving biases. The only prediction method we applied was CMfinder because of its similarity to RNAG. It should be noted again that while CMfinder can be used to predict global consensus structure, it was designed for a different purpose: to find local secondary structure motifs within longer sequences. To ensure that we were using CMfinder properly, we first reproduced the results of Yao *et al.* [180] with default settings, then used the same settings for our tests.

It should also be noted that these three datasets are far from a random sample of all RNA families, and they are a relatively small sample at that. The most extensive dataset is that of Kiryu *et al.* [55] at only 17 families. Thus, we have only demonstrated the efficacy of RNAG on a small subset of all possible RNA families and generalizations should be drawn with caution.

4.6.2 Potential Improvements

Since we did not train the individual algorithms that make up RNAG, performance may be enhanced by tuning RNAalifold and Infernal on a training set to select particular parameter values or options. Of course, if other methods for drawing from the two relevant conditional probability distributions are proposed that outperform either RNAalifold or Infernal, the RNAG framework will easily accommodate an interchange, and will likely benefit in terms of predictive accuracy. Furthermore, a more accurate initial aligner would likely speed up convergence. RNAG would also benefit from methods for computational speed such as parallel implementation, as it currently runs about 3 times faster than RNA Sampler, but 10 times slower than RAF.

4.6.3 RNA family classification

The substantial biases that we see in the Kiryu *et al.* [55] dataset may reflect one of three things: that 1,000 iterations is not sufficient for convergence, that the models used in each of the sampling algorithms do not accurately model the biochemistry, or that the reference structures do not accurately represent the common structural characteristics of a family of RNAs. For this dataset, 13 of 17 reference structures were obtained via *in vitro* biochemical experiments that probe a specific structure, and thus may not adequately represent the common structural characteristics among multiple sequences. A more appropriate test might be how accurately a predicted consensus structure can classify sequences into the correct RNA families. The curation of a reference set for such a test is non-trivial; the Rfam database [103] would be a tempting place to start, but Rfam's sequence families themselves are obtained by

aligning the sequences to reference structures. Instead, independent methods such as immunoprecipitation studies would be needed to identify RNA families.

CHAPTER **Five**

Conclusion

In this thesis, we focus on three problems motivated by the study of RNA and ADAR. Chapters 2 and 3 are also motivated by the recent concern in the biological community about the issues of irreproducibility and false positives, especially in genome-scale studies.

In Chapter 2, we address the issue of irreproducibility directly by proposing a method by which investigators can assess the reproducibility of their own results, giving their audience a sense of what to expect if they want to build on those results. The hierarchical model is ideal for this purpose because it perfectly captures the relationship between experimental replicates, allowing them to vary due to biological or sample preparation variation, but recognizing the commonalities that exist due to the fact that each replicate follows the same experimental protocol, and is testing the same population of predictions. We use this hierarchical model to infer a predictive distribution of the results of a new replicate, quantifying reproducibility by

the mean, standard deviation, and 10^{th} percentile of this distribution. If we define better reproducibility by predictive distributions with high 10^{th} percentiles and low standard deviations, then our model allows us to choose the optimal number of replicates that will yield the best reproducibility. It seems likely that many opportunities for our software will arise increasingly often as biologists take advantage of modern technology to investigate every corner of the genome.

One contributing factor to irreproducibility of results is false positives, and a potential source of false positives in genomic research is the failure to account for genetic variation. This problem arises predominantly due to the use of a reference genome, the DNA sequence of an organism from a particular species that is commonly used to map short sequence reads from high-throughput sequencing. The existence of a reference genome makes most modern genomic studies possible, but there is often a dangerous implicit assumption that the reference genome is the genomic sequence of the organisms under study. This is a particularly bad assumption for *Drosophila*, which must be maintained as live stocks, reproducing every 2-4 weeks. Our collaborators estimate that decades may have passed since the most recent common ancestor of their laboratory flies and the reference fly. The risk of failing to identify genetic polymorphisms, especially unfixed polymorphisms, is highest for investigations of proteins like ADAR, for which the signals are highly localized. An unaccounted-for A to G polymorphism will appear as an editing event and will lead to a false positive prediction. It is likely that this has been a major problem for edit-site discovery papers, which typically fail to account for how common polymorphisms are, and commonly find little overlap with other published studies. In Chapter 3, we build a model of the biological processes involved in genetic polymorphisms, ADAR-mediated editing, and sequencing errors, which may also look like editing sites if not properly dealt with, and we use an EM algorithm to learn the model parameters and

posterior probabilities so that we can make rigorous statements about the probability of a polymorphism or editing event at a given site.

A typical way to validate editing predictions from high-throughput studies is to use Sanger sequencing to isolate the region of the RNA and look for overlapping peaks, as was done in our companion paper. While this will indeed validate the presence of mixed A and G in the RNA, it does nothing to control for the possibility of a polymorphism beyond any precautions taken in the original study. In other words, this validation scheme will weed out false positives due to high-throughput sequencing errors, but not those due to genetic variation, which we show to be far more common than the former. The result is a misleading level of confidence in the set of predictions. Therefore, while validation experiments are absolutely crucial to the scientific method, they are not a silver bullet. Care needs to be taken to ensure that these experiments control for all potential sources of false positives.

While Chapters 2 and 3 are concerned with RNA primarily as a passive target of regulation, in Chapter 4 we begin to appreciate RNA for its versatility and for its own active role in regulation. In particular, we capitalize on the close relationship between structure and function in RNA. Instead of trying to predict the structure of a single sequence, we look to RNA families that have similar function or are regulated in the same way. This allows us to look for common structural characteristics across the multiple sequences, taking advantage of both single-sequence folding tendencies and information about sequence conservation to predict structure. The main obstacle that remains is the uncertainty in the alignment of the related sequences. Our approach uses a blocked Gibbs sampler that alternately samples a consensus structure given an alignment, and an alignment given a consensus structure. We show that this algorithm, RNAG, outperforms other methods. We also describe an extension of RNAG to an RNA motif-finder, RGibbs, that can find local structures

common to the transcripts.

One interesting use of RGibbs would be to try to improve the understanding of ADAR's binding preferences. An intriguing aspect of editing sites is that they tend to cluster; most transcripts are completely untouched by ADAR, while some are edited in multiple places. This suggests a possible mechanism for ADAR binding: one motif is used to recognize the transcript targets while another guides ADAR to particular adenosines. RGibbs could easily be extended to look for two motifs in each targeted transcript, with the potential for multiple instances of the second motif in a given transcript.

While the knowledge of more editing sites would improve our motif prediction, we could also use knowledge of ADAR's secondary structure requirements to improve our prediction of editing sites. Our current model assumes that any A in the RNA is an equally likely target of ADAR, but a future algorithm could incorporate structural predictions as part of the prior model.

APPENDIX **A**

Discarding an RNA pool

The original ADAR dataset contained 6 RNA pools. A skilled investigator well-versed in the protocol created the five that we included in our analysis, and the sixth was created by an investigator with much less experience. We suspected that the sixth pool was of much lower quality than the others, so we set out to test this hypothesis. Using the first pool as a benchmark, we collected the predictions submitted for validation in pools 2-6 and re-validated them in pool 1. We kept track of false negatives (predictions validated in pool 1, but not in the original pool), and true positives (predictions that validated in both). The results are in the table below. A Fisher exact test of the hypothesis that pool 6 differed from the rest gave a p-value of 0.026. Accordingly, we discarded this pool and worked with pools 1-5 for the remainder of the analysis.

	Validated in original pool (True positive)	Did not validate in original pool (False Negative)
Pools 2-5	33	4
Pool 6	2	3

It is worth noting here that while discarding a pool created by a different investigator may seem counterintuitive when the object of our analysis is to assess the reproducibility of the validation experiments in the context of sources of variation including individual experimenter technique, it is also important to define exactly the population to which the reproducibility results apply. It is reasonable to assume that the population of interest is that of scientists and technicians who are technically proficient and also well-versed in the particular protocol.

APPENDIX **B**

Drawing from the posterior distribution on α and β

The un-normalized posterior distribution of α and β given data k_i, $i = 1, ..., m$ is given by the following expression:

$$f(\alpha, \beta | k_1^m) \propto f(\alpha, \beta) \prod_{i=1}^{m} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k_i)\Gamma(\beta+N_i-k_i)}{\Gamma(\alpha+\beta+N_i)}$$

where $f(\alpha, \beta)$ represents the prior. We choose the prior recommended by Gelman *et al.*, a uniform distribution over $(\frac{\alpha}{\alpha+\beta}, \frac{1}{\sqrt{\alpha+\beta}})$, which translates to $f(\alpha, \beta) = (\alpha + \beta)^{-\frac{5}{2}}$. To compute the posterior distribution, we reparameterize to the $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ plane, so that the distribution we are concerned with is the following:

$$f(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta) | k_1^m) \propto \alpha\beta(\alpha+\beta)^{-\frac{5}{2}} \prod_{i=1}^{m} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k_i)\Gamma(\beta+N_i-k_i)}{\Gamma(\alpha+\beta+N_i)}$$

We compute these un-normalized values on a 200x200 grid in the $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ plane, begining with method of moments estimates for α and

β as a starting point for centering the grid, then shifting and rescaling the grid as needed until the contour plot occupies 80% of the width and height of the grid window. Figure B.1 illustrates the possible ways in which the contour plot may require adjusting. If the grid spacing is too large, the resolution will be too low to capture the important elements of the distribution. If the grid spacing is too small, or the grid is off-center, the contour plot will be truncated, and we will be missing parts of the posterior distribution. We shift and resize until we get the last contour plot, perfectly centered and scaled, with a buffer that constitutes 10% of the grid window on all sides. We then interpolate between grid points using step functions and normalize to create a valid probability distribution.

To draw samples, a row of the 200x200 grid is chosen in proportion to its marginal probability, then a column of the grid is chosen in proportion to its conditional probability given the chosen row. From that gridpoint, independent uniform horizontal and vertical jitters are applied with center at the gridpoint and width equal to the grid spacing. This coordinate (x, y) is converted back into the (α, β) plane by the following transformations:

$$\begin{aligned} \alpha &= \frac{e^{x+y}}{e^x + 1} \\ \beta &= \frac{e^y}{e^x + 1} \end{aligned}$$

For more detail, see Gelman *et al.*, 2003 [35].

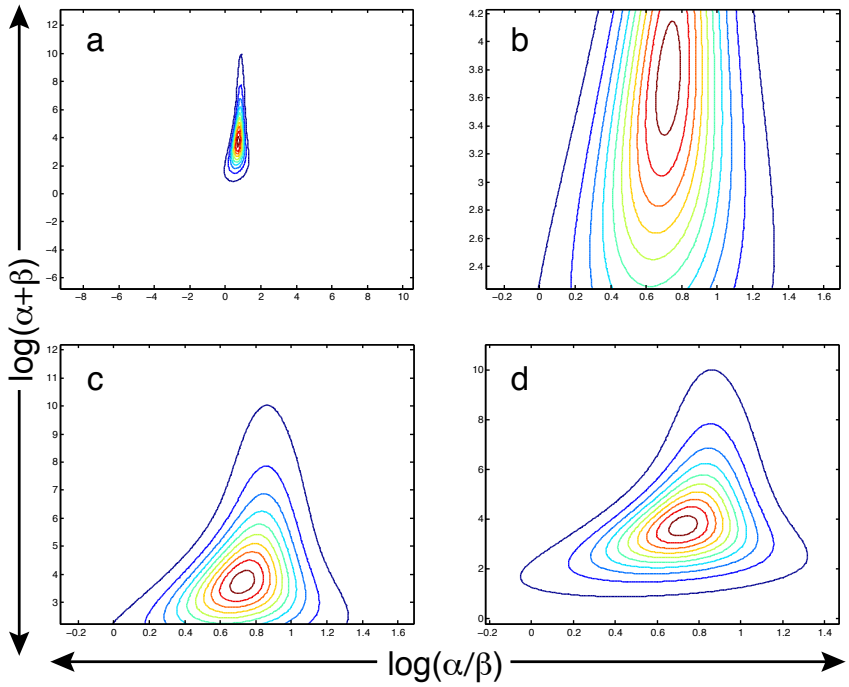

Figure B.1: Contour plot adjustments

APPENDIX **C**

Bayesian cross-validation with pooled samples

The main text describes the results of cross validation of the five replicate pools in the ADAR study, first using our model, then using standard frequentist confidence limits. To illustrate that our results were not solely due to Bayesian paradigm, we also performed the same leave-one-out cross validation using Bayesian credibility limits for the pooled replicates. This consisted of calculating the posterior distribution of the proportion p, given the data and a prior distribution on p. We used the non-informative Jeffreys prior for Bernoulli trials, a beta distribution with $\alpha = \beta = \frac{1}{2}$. The posterior probability of p then follows a beta distribution with parameters $\alpha =$

RNA Pool	Valid/Total	CV 80% CI	CV 90% CI	CV 95% CI
1	151/205 (74%)	(0.59,0.70)	(0.58,0.72)	(0.56,0.73)
2	17/32 (53%)	(0.69,0.76)	(0.68,0.76)	(0.67,0.77)
3	21/30 (70%)	(0.67,0.74)	(0.66,0.75)	(0.65,0.75)
4	24/32 (75%)	(0.66,0.73)	(0.65,0.74)	(0.65,0.75)
5	18/29 (62%)	(0.68,0.74)	(0.67,0.75)	(0.66,0.76)

$n + \frac{1}{2}$, $\beta = N - n + \frac{1}{2}$ where N is the total number of experiments in the pool and n is the number of those that successfully validate. Given this distribution, we can find credibility intervals by computing the appropriate percentiles (e.g. 80% CI is given by the 10^{th} and 90^{th} percentiles). These credibility limits are given in the table, where the left-out pool is the row pool. These intervals are very similar to those given by the frequentist confidence limits, and like those limits, do not contain the left-out value most of the time. Here, 4 out of 5 times, the left-out proportion lies outside all three intervals, and only 1 out of 5 lies within the 80% credibility interval.

APPENDIX **D**

Splitting Experiments

To find the optimal number of replicates, we run our predictive inference procedure many times for different assignments of replicate number and experiments per replicate. This table shows the division of 96 total experiments into different numbers of replicates. The numbers shown in the table are chosen to be a reasonably-spaced representation of all possibilities. For example, 17 replicates is not included, since it results in 5-6 experiments per replicate and will thus likely have very similar results to 19 replicates. Specifically, if there is a range of numbers $n_1, ..., n_j, ..., n_J$ such that floor($\frac{96}{n_j}$) is the same for all $j-1, ..., J$, we choose the maximum over all of the n_j to include in our calculations.

Number of Replicates (N_{rep})	3	4	5	6	7	8	9
Experiments per Replicate (N_{exp})	32	24	19-20	16	13-14	12	10-11

10	12	13	16	19	24	32	48
9-10	8	7-8	6	5-6	4	3	2

APPENDIX **E**

Dealing with Strand-specific RNA Reads

The full model described in Chapter 3 assumes that we do not get to see which strand each RNA read was generated from. Some protocols, however, yield strand-specific reads that provide this information. If this is the case, our model changes slightly. The new prior model is shown in Figure E.1, the only difference being the absence of the step where we draw the Watson or Crick strand with equal probabilities at A or T sites.

Given an A in the reference, we have the following prior probabilities on the DNA

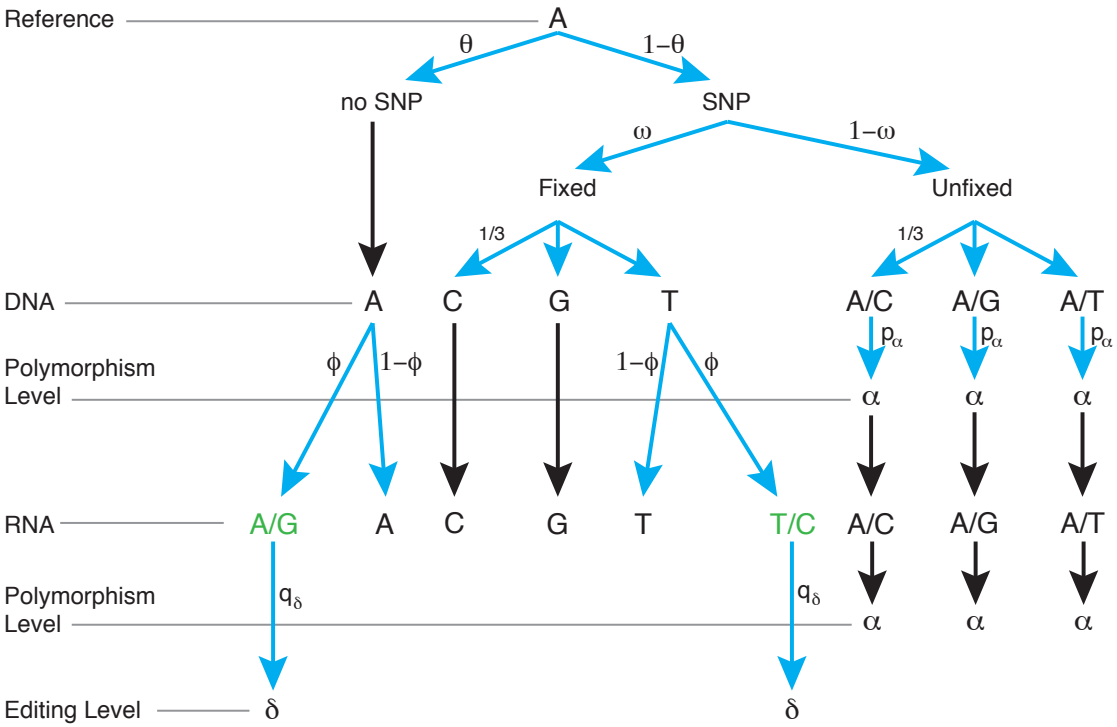

Figure E.1: Prior model for strand-specific RNA reads

and RNA:

$$
\begin{aligned}
P(G_i = A, R_i = A/G, \text{ level } \delta | F_i = A) &= \theta\phi q_\delta \\
P(G_i = T, R_i = T/C, \text{ level } \delta | F_i = A) &= \frac{1}{3}(1-\theta)\omega\phi q_\delta \\
P(G_i = A, R_i = A | F_i = A) &= \theta(1-\phi) \\
P(G_i = T, R_i = T | F_i = A) &= \frac{1}{3}(1-\theta)\omega(1-\phi) \\
P(G_i = C, R_i = C | F_i = A) &= \frac{1}{3}(1-\theta)\omega \\
P(G_i = G, R_i = G | F_i = A) &= \frac{1}{3}(1-\theta)\omega \\
P(G_i = A/C, \text{ level } \alpha, R_i = A/C, \text{ level } \alpha | F_i = A) &= \frac{1}{3}(1-\theta)(1-\omega)p_\alpha \\
P(G_i = A/G, \text{ level } \alpha, R_i = A/G, \text{ level } \alpha | F_i = A) &= \frac{1}{3}(1-\theta)(1-\omega)p_\alpha \\
P(G_i = A/T, \text{ level } \alpha, R_i = A/T, \text{ level } \alpha | F_i = A) &= \frac{1}{3}(1-\theta)(1-\omega)p_\alpha
\end{aligned}
\tag{E.1}
$$

All other possible combinations of G_i and R_i have probability zero.

The likelihood terms also change slightly, since we now need to account for the strand-specificity of ADAR itself. Briefly, $A \to G$ evidence for editing at a position with $G_i = A$ can only come from base calls mapped to the Watson strand. Likewise, evidence for editing at a position with $G_i = T$ can only come from base calls mapped to the Crick strand. Therefore, Equation 3.23 splits into two cases:

$$
\begin{aligned}
P(\{Y_j\}_i | R_i = A/G, \epsilon = 1, \text{ level } \delta) = & \\
& \prod_{j \in K^W} \left[(1-\delta) E^{\text{RNA}}_{F_{i-2}, F_{i-1}, A}(Y_j) + \delta E^{\text{RNA}}_{F_{i-2}, F_{i-1}, G}(Y_j) \right] \\
& \times \prod_{j \in K^C} E^{\text{RNA}}_{F^C_{i+2}, F^C_{i+1}, T}(Y_j)
\end{aligned}
\tag{E.2}
$$

$$P(\{Y_j\}_i|R_i = T/C, \epsilon = 1, \text{ level } \delta) = \prod_{j \in K^W} E^{\text{RNA}}_{F_{i-2},F_{i-1},A}(Y_j) \prod_{j \in K^C} [(1-\delta)E^{\text{RNA}}_{F^C_{i+2},F^C_{i+1},T}(Y_j) + \delta E^{\text{RNA}}_{F^C_{i+2},F^C_{i+1},C}(Y_j)] \tag{E.3}$$

Whereas before, an editing event could be supported by As and Gs on the Watson strand and Ts and Cs on the Crick strand simultaneously, this likelihood model forces the bases generated from the un-editable strand to be generated from the unedited nucleotide (A or T).

For EM, the ϕ update in Equation 3.32 takes on a slightly simpler form:

$$\begin{aligned} \phi \leftarrow & \frac{\mathbb{E}\{\#\text{ Editing events}\}}{\mathbb{E}\{\#\text{ Transcribed As}\}} \\ = & \frac{\sum_i P(G_i = A, R_i = A/G|D_i) + P(G_i = T, R_i = T/C|D_i)}{\sum_i P(G_i = A|D_i)} \end{aligned} \tag{E.4}$$

All other equations in Chapter 3 are unchanged for strand-specific RNA data.

BIBLIOGRAPHY

[1] Alison Abbott. Disputed results a fresh blow for social psychology. *Nature*, 497(7447):16, 2013.

[2] Nature Announcement. Reducing our irreproducibility. *Nature*, 496:398, 2013.

[3] Paul L Auer and RW Doerge. Statistical design and analysis of rna sequencing data. *Genetics*, 185(2):405–416, 2010.

[4] Monya Baker. Independent labs to verify high-profile papers. *Nature News*, 2012.

[5] Brenda L Bass and Harold Weintraub. An unwinding activity that covalently modifies its double-stranded rna substrate. *Cell*, 55:1089–1098, 1988.

[6] CG Begley and LM Ellis. Drug development: raise standards for preclinical cancer research. *Nature*, 483:531–533, 2012.

[7] Eckart Bindewald and Bruce A Shapiro. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, 12(3):342–352, 2006.

[8] Douglas L Black. Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry*, 72(1):291–336, 2003.

[9] Michael A Black and RW Doerge. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics*, 18(12):1609–1616, 2002.

[10] Anne-Laure Boulesteix and Martin Slawski. Stability and aggregation of ranked gene lists. *Briefings in bioinformatics*, 10(5):556–568, 2009.

[11] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[12] Luis E Carvalho and Charles E Lawrence. Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proceedings of the National Academy of Sciences*, 105(9):3209–3214, 2008.

[13] Robert B Cary and Gary D Stormo. Graph-theoretic approach to RNA modeling using comparative data. *ISMB*, 95:75–80, 1995.

[14] Sonia Casillas, Natalia Petit, and Antonio Barbadilla. DPDB: a database for the storage, representation and analysis of polymorphism in the *drosophila* genus. *Bioinformatics*, 21:ii26–ii30, 2005.

[15] Sean Conlan, Charles E Lawrence, and Lee Ann McCue. Rhodopseudomonas palustris regulons detected by cross-species analysis of alphaproteobacterial genomes. *Applied and environmental microbiology*, 71(11):7442–7452, 2005.

[16] AP Dempster, NM Laird, and DB Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society, series B*, 39:1–38, 1977.

[17] Brian DeVeale, Derek van der Kooy, and Tomas Babak. Critical evaluation of imprinted gene expression by RNA–Seq: a new perspective. *PLoS genetics*, 8(3):e1002600, 2012.

[18] Ye Ding, Chi Yu Chan, and Charles E Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *Rna*, 11(8):1157–1166, 2005.

[19] Ye Ding, Chi Yu Chan, and Charles E Lawrence. Clustering of RNA secondary structures with application to messenger RNAs. *J. Mol. Biol.*, 359:554–571, 2006.

[20] Ye Ding and Charles E Lawrence. A statistical sampling algorithm for rna secondary structure prediction. *Nucleic acids research*, 31(24):7280–7301, 2003.

[21] Chuong B. Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature Biotechnology*, 26:897–899, 2008.

[22] Chuong B Do, Chuan-Sheng Foo, and Serafim Batzoglou. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, 24(13):i68–i76, 2008.

[23] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

[24] Sean R Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic acids research*, 22(11):2079–2088, 1994.

[25] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[26] Nature Editorial. Error prone. *Nature*, 487:406, 2012.

[27] Nature Editorial. Further confirmation needed. *Nature Biotechnology*, 30:806, 2012.

[28] Nature Editorial. Must try harder. *Nature*, 483:509, 2012.

[29] Julie M Eggington, Tom Greene, and Brenda L Bass. Predicting sites of ADAR editing in double-stranded RNA. *Nature Communications*, 2:doi:10.1038/ncomms1324, 2011.

[30] Hans Ellegren, Nick GC Smith, and Matthew T Webster. Mutation rate variation in the mammalian genome. *Current opinion in genetics & development*, 13(6):562–568, 2003.

[31] Aaron McKenna *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20:1297–1303, 2010.

[32] Adrian E Platts *et al.* Massively parallel resequencing of the isogenic Drosophila melanogaster strain w1118; iso-2; iso-3 identifies hotspots for mutations in sensory perception genes. *Fly*, 3(3):192, 2009.

[33] Andreas Massouras *et al.* Genomic variation and its impact on gene expression in *drosophila melanogaster*. *PLoS Genetics*, 8:e1003055, 2012.

[34] Andrew Fire *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *nature*, 391(6669):806–811, 1998.

[35] Andrew Gelman *et al. Bayesian data analysis*. CRC press, 2013.

[36] AW Bell *et al.* A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nature Methods*, 6:423–430, 2009.

[37] Barry Hoopengardner *et al.* Nervous system targets of RNA editing identified by comparative genomics. *Science*, 301(5634):832–836, 2003.

[38] Ben Langmead *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[39] Brenton R Graveley *et al.* The developmental transcriptome of *drosophila melanogaster*. *Nature*, 471:473–479, 2011.

[40] Charles E Lawrence *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[41] Christopher Gregg *et al.* High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, 329(5992):643–648, 2010.

[42] Chuong B Do *et al.* ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research*, 15(2):330–340, 2005.

[43] Dale Hedges *et al.* Exome sequencing of a multigenerational human pedigree. *PLoS one*, 4:e8232, 2009.

[44] David R Shanks *et al.* Priming intelligent behavior: An elusive phenomenon. *PloS one*, 8(4):e56515, 2013.

[45] Dirk Feldmeyer *et al.* Neurological dysfunctions in mice expressing different levels of the Q/R site-unedited AMPAR subunit GluR-B. *Nature neuroscience*, 2(1):57–64, 1999.

[46] Doris Chen *et al.* FLYSNPdb: a high-density SNP database of *drosophila melanogaster*. *Nucleic Acids Research*, 37:D567–D570, 2009.

[47] E R Martin *et al.* SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, 26:2803–2810, 2010.

[48] Eleftheria Zeggini *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40(5):638–645, 2008.

[49] Emily Bernstein *et al.* Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366, 2001.

[50] Erin D Pleasance *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463:191–197, 2010.

[51] Georges St Laurent *et al.* Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila. *Nature structural & molecular biology*, 20(11):1333–1339, 2013.

[52] Gillian I Rice *et al.* Mutations in ADAR1 cause Aicardi-Goutieres syndrome associated with a type I interferon signature. *Nature genetics*, 44(11):1243–1248, 2012.

[53] Gokul Ramaswami *et al.* Accurate identification of human *alu* and non-*alu* RNA editing sites. *Nature Methods*, 9:579–581, 2012.

[54] Gokul Ramaswami *et al.* Identifying RNA editing sites using RNA sequencing data alone. *Nature methods*, 10(2):128–132, 2013.

[55] Hisanori Kiryu *et al.* Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, 23(13):1588–1598, 2007.

[56] I Surolia *et al.* Functionally defective germline variants of sialic acid acetylesterase in autoimmunity. *Nature*, 466:243–247, 2010.

[57] IL Hofacker *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, 125:167–188, 1994.

[58] Ilona Gurevich *et al.* Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron*, 34(3):349–356, 2002.

[59] Irina Abnizova *et al.* Analysis of context-dependent errors for Illumina sequencing. *Journal of Bioinformatics and Computational Biology*, 10:1241005, 2012.

[60] Jack E Tabaska *et al.* An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–699, 1998.

[61] Jae Hoon Bahn *et al.* Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Research*, 22:142–150, 2012.

[62] James EC Jepson *et al.* Visualizing adenosine-to-inosine RNA editing in the Drosophila nervous system. *Nature methods*, 9(2):189–194, 2012.

[63] Jeffrey C Barrett *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nature genetics*, 40(8):955–962, 2008.

[64] Jeffrey T Leek *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[65] Jin Billy Li *et al.* Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Research*, 19:1606–1615, 2009.

[66] John H Livingston *et al.* A type I interferon signature identifies bilateral striatal necrosis due to mutations in ADAR1. *Journal of medical genetics*, 51(2):76–82, 2014.

[67] JP Ioannidis *et al.* Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41:149–155, 2009.

[68] Julia Zeitlinger *et al.* RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nature genetics*, 39(12):1512–1516, 2007.

[69] Juliane C. Dohm *et al.* Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36:e105, 2008.

[70] Jun Wang *et al.* The diploid genome sequence of an asian individual. *Nature*, 456:60–66, 2008.

[71] K Nishikura *et al.* Substrate specificity of the dsRNA unwinding/modifying activity. *EMBO*, 10:3523–3532, 1991.

[72] KA Hunt *et al.* Rare and functional SIAE variants are not associated with auotimmune disease risk in up to 66,924 individuals of European ancestry. *Nature Genetics*, 44:3–5, 2012.

[73] Katherine S Button *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

[74] Keith A Baggerly *et al.* Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, 19:1477–1483, 2003.

[75] Lauren A Sugden *et al.* Assessing the validity and reproducibility of genome-scale predictions. *Bioinformatics*, 29(22):2844–2851, 2013.

[76] Lee Ann McCue *et al.* Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic acids research*, 29(3):774–782, 2001.

[77] Leming Shi *et al.* The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9):1151–1161, 2006.

[78] Li C Xia *et al.* Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC systems biology*, 5(Suppl 2):S15, 2011.

[79] Li Kang *et al.* Identification of miRNAs associated with sexual maturity in chicken ovary by Illumina small RNA deep sequencing. *BMC Genomics*, 14:352, 2013.

[80] Lisa M McShane *et al.* Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469, 2002.

[81] Magnus A Rosenblad *et al.* SRPDB: signal recognition particle database. *Nucleic Acids Research*, 31:363–365, 2003.

[82] Mary E Winn *et al.* Gene expression profiling of human whole blood samples with the Illumina WG-DASL assay. *BMC Genomics*, 12:412, 2011.

[83] Mathias Sprinzl *et al.* Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research*, 26(1):148–153, 1998.

[84] Michael J Palladino *et al.* A-to-I Pre-mRNA Editing in *Drosophila* Is Primarily Involved in Adult Nervous System Function and Integrity. *Cell*, 102(4):437–449, 2000.

[85] Michael Palladino *et al.* dADAR, a Drosophila double-stranded RNA-specific adenosine deaminase is highly developmentally regulated and is itself a target for RNA editing. *Rna*, 6(7):1004–1018, 2000.

[86] Michiaki Hamada *et al.* Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, 25(4):465–473, 2009.

[87] Min Zhang *et al.* Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, 25(13):1662–1668, 2009.

[88] MJ Clark *et al.* U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet*, 6:e1000832, 2010.

[89] N J Barrows *et al.* Factors affecting reproducibility between genome-scale siRNA-based screens. *J Biomol Screen*, 15:735–747, 2010.

[90] N Tian *et al.* A structural determinant required for rna editing. *Nucleic Acids Research*, 39:5669–5681, 2011.

[91] Nicolas Nègre *et al.* A comprehensive map of insulator elements for the Drosophila genome. *PLoS genetics*, 6(1):e1000814, 2010.

[92] NJ Barrows *et al.* Factors affecting reproducibility between genome-scale siRNA-based screens. *J Biomol Screen*, 15:735–747, 2010.

[93] Olivier Harismendy *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 10:R32, 2009.

[94] PP Gardner *et al.* A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, 22:2433–2439, 2005.

[95] Ramal Moonesinghe *et al.* Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proceedings of the National Academy of Sciences*, 105(2):617–622, 2008.

[96] Ramu Chenna *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic acids research*, 31(13):3497–3500, 2003.

[97] Ricardo ZN Vêncio *et al.* Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). *BMC bioinformatics*, 5(1):119, 2004.

[98] Roger A Hoskins *et al.* Single nucleotide polymorphism markers for genetic mapping in *drosophila melanogaster*. *Genome Research*, 11:1100–1113, 2001.

[99] Roger A Hoskins *et al.* Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome research*, 21(2):182–192, 2011.

[100] RR Gutell *et al.* Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic acids research*, 20(21):5785–5795, 1992.

[101] Ruiqiang Li *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 19:1124–1132, 2009.

[102] Sacha Van Hijum *et al.* A generally applicable validation scheme for the assessment of factors involved in reproducibility and quality of dna-microarray data. *BMC genomics*, 6(1):77, 2005.

[103] Sam Griffiths-Jones *et al.* Rfam: an RNA family database. *Nucleic acids research*, 31(1):439–441, 2003.

[104] Sebastian Will *et al.* Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS computational biology*, 3(4):e65, 2007.

[105] ST Sherry *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29:308–311, 2001.

[106] Stella Dracheva *et al.* RNA editing and alternative splicing of human serotonin 2C receptor in schizophrenia. *Journal of neurochemistry*, 87(6):1402–1412, 2003.

[107] Stephan H Bernhart *et al.* RNAalifold: improved consensus structure prediction for RNA alignments. *BMC bioinformatics*, 9(1):474, 2008.

[108] Takuto Hideyama *et al.* Profound downregulation of the RNA editing enzyme ADAR2 in ALS spinal motor neurons. *Neurobiology of disease*, 45(3):1121–1128, 2012.

[109] Timothy L Bailey *et al.* MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research*, 34(suppl 2):W369–W373, 2006.

[110] TM Keane *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477:289–294, 2011.

[111] Wei Lin *et al.* Comment on "Widespread RNA and DNA sequence differences in the human transcriptome. *Science*, 335:1302, 2012.

[112] Winston Patrick Kuo *et al.* A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nature biotechnology*, 24(7):832–840, 2006.

[113] Xinan Yang *et al.* Similarities of ordered gene lists. *Journal of Bioinformatics and Computational Biology*, 4(03):693–708, 2006.

[114] Xue-Jun Zhang *et al.* Seven novel mutations of the ADAR gene in Chinese families and sporadic patients with dyschromatosis symmetrica hereditaria (DSH). *Human mutation*, 23(6):629–630, 2004.

[115] Yiannis A Savva *et al.* Auto-regulatory RNA editing fine-tunes mRNA recoding and complex behaviour in *Drosophila*. *Nature Communications*, 3:790, 2012.

[116] Yiannis A Savva *et al.* RNA editing regulates transposon-mediated heterochromatic gene silencing. *Nature communications*, 4, 2013.

[117] Zhiyu Peng *et al.* Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology*, 30(3):253–260, 2012.

[118] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721–741, 1984.

[119] Robert Giegerich, Björn Voß, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic acids research*, 32(16):4843–4851, 2004.

[120] Walter Gilbert. The RNA world. *Nature*, 319:618, 1986.

[121] Peter Glaus, Antti Honkela, and Magnus Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012.

[122] C Guthrie and B Patterson. Spliceosomal snRNAs. *Annual review of genetics*, 22(1):387–419, 1988.

[123] Michiaki Hamada, Kengo Sato, and Kiyoshi Asai. Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucleic acids research*, 39(2):393–402, 2011.

[124] Matthew J Hangauer, Ian W Vaughn, and Michael T McManus. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS genetics*, 9(6):e1003569, 2013.

[125] Jakob H Havgaard, Rune B Lyngsø, and Jan Gorodkin. The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic acids research*, 33(suppl 2):W650–W653, 2005.

[126] Erika Check Hayen. RNA studies under fire. *Nature*, 484:428, 2012.

[127] Ivo L Hofacker, Martin Fekete, and Peter F Stadler. Secondary structure prediction for aligned RNA sequences. *Journal of molecular biology*, 319(5):1059–1066, 2002.

[128] Darrel C Ince, Leslie Hatton, and John Graham-Cumming. The case for open computer programs. *Nature*, 482:485–488, 2012.

[129] George Johnson. New truths that only one can see. *The New York Times*, 2014.

[130] Thomas H Jukes. Transitions, transversions, and the molecular evolutionary clock. *Journal of Molecular Evolution*, 26:87–98, 1987.

[131] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[132] Kathleen M Kerr and Gary A Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences*, 98(16):8961–8965, 2001.

[133] Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[134] Claudia L Kleinman and Jacek Majewski. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome. *Science*, 335:1302, 2012.

[135] B Knudsen and J Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15:446–454, 1999.

[136] Bjarne Knudsen and Jotun Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic acids research*, 31(13):3423–3428, 2003.

[137] Christian Laing and Tamar Schlick. Computational approaches to RNA structure prediction, analysis and design. *Current opinion in structural biology*, 21:306–318, 2011.

[138] KA Lehmann and BL Bass. The importance of internal loops within RNA substrates of ADAR1. *J. Mol. Biol.*, 291:1–13, 1999.

[139] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27:2987–2993, 2011.

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

[141] Mingyao Li, Isabel X Wang, Yun Li, Alan Bruzel, Allison L Richards, Jonathan M Toung, and Vivian G Cheung. Widespread rna and dna sequence differences in the human transcriptome. *Science*, 333(6038):53–58, 2011.

[142] Qunhua Li, James B Brown, Haiyan Huang, Peter J Bickel, et al. Measuring reproducibility of high-throughput experiments. *The annals of applied statistics*, 5(3):1752–1779, 2011.

[143] Stinus Lindgreen, Paul P Gardner, and Anders Krogh. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics*, 23(24):3304–3311, 2007.

[144] Jun S Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[145] Gerton Lunter and Martin Goodson. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research*, 21(6):936–939, 2011.

[146] Daniel MacArthur. Face up to false positives. *Nature*, 487:427–428, 2012.

[147] Luciano A Marraffini and Erik J Sontheimer. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature Reviews Genetics*, 11(3):181–190, 2010.

[148] ES Maxwell and MJ Fournier. The small nucleolar RNAs. *Ann. Rev. Biochem.*, 35:897–934, 1995.

[149] Irmtraud M Meyer and István Miklós. SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS computational biology*, 3(8):e149–e149, 2007.

[150] François Michel, Umesono Kazuhiko, and Ozeki Haruo. Comparative and functional anatomy of group II catalytic intronsa review. *Gene*, 82(1):5–30, 1989.

[151] Eric P Nawrocki and Sean R Eddy. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Computational Biology*, 3(3):e56, 2007.

[152] Eric P Nawrocki, Diana L Kolbe, and Sean R Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, 2009.

[153] Lee A Newberg and Charles E Lawrence. Exact calculation of distributions on integers, with application to sequence alignment. *Journal of Computational Biology*, 16(1):1–18, 2009.

[154] Lee A Newberg, Lee Ann McCue, and Charles E Lawrence. The relative inefficiency of sequence weights approaches in determining a nucleotide position weight matrix. *Statistical Applications in Genetics and Molecular Biology*, 4:doi:10.2202/1544–6115.1135, 2005.

[155] Evgeny Nudler and Alexander S Mironov. The riboswitch control of bacterial metabolism. *Trends in biochemical sciences*, 29(1):11–17, 2004.

[156] Roman Pahl, Helmut Schäfer, and Hans-Helge Müller. Optimal multistage designsa general framework for efficient genome-wide association studies. *Biostatistics*, 10(2):297–309, 2009.

[157] Wei Pan, Jizhen Lin, and Chap T Le. How many replicates of arrays are required to detect gene expression changes in microarray experiments? a mixture model approach. *Genome Biol*, 3(5):1–0022, 2002.

[158] Joseph K Pickrell, Yoav Gilad, and Jonathan K Pritchard. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome. *Science*, 335:1302, 2012.

[159] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712–712, 2011.

[160] Joseph Rodriguez, Jerome S Menet, and Michael Rosbash. Nascent-seq indicates widespread cotranscriptional rna editing in *drosophila*. *Molecular cell*, 47(1):27–37, 2012.

[161] Igor B Rogozin and Youri I Pavlov. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Reviews in Mutation Research*, 544(1):65–85, 2003.

[162] Jonathan F Russell. If a job is worth doing, it is worth doing twice. *Nature*, 496:7, 2013.

[163] David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.

[164] ADJ Scadden and Christopher WJ Smith. RNAi is antagonized by A→ I hyper-editing. *EMBO reports*, 2:1107–1111, 2001.

[165] Stefan E Seemann, Jan Gorodkin, and Rolf Backofen. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic acids research*, 36(20):6355–6362, 2008.

[166] Sven Siebert and Rolf Backofen. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, 21(16):3352–3359, 2005.

[167] Mark Stapleton, Joseph W Carlson, and Susan E Celniker. RNA editing in Drosophila melanogaster: New targets and functional consequences. *RNA*, 12(11):1922–1932, 2006.

[168] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[169] Gary D Stormo and George W Hartzell. Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences*, 86(4):1183–1187, 1989.

[170] Ramanjulu Sunkar, Thomas Girke, and Jian-Kang Zhu. Identification and characterization of endogenous small interfering RNAs from rice. *Nucleic Acids Research*, 33:4443–4454, 2005.

[171] Robert Tibshirani. A simple method for assessing sample sizes in microarray experiments. *Bmc Bioinformatics*, 7(1):106, 2006.

[172] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[173] David L Vaux. Know when your numbers are significant. *Nature*, 492:180–181, 2012.

[174] Bobbie-Jo M Webb-Robertson, Lee Ann McCue, and Charles E Lawrence. Measuring global credibility with application to local sequence alignment. *PLoS computational biology*, 4(5):e1000077, 2008.

[175] Caimiao Wei, Jiangning Li, and Roger E Bumgarner. Sample size for detecting differentially expressed genes in microarray experiments. *BMC genomics*, 5(1):87, 2004.

[176] Donglai Wei, Lauren V Alpert, and Charles E Lawrence. RNAG: a new Gibbs sampler for predicting RNA secondary structure for unaligned sequences. *Bioinformatics*, 27:2486–2493, 2011.

[177] Hermann Weidner, Raymond Yuan, and Donald M Crothers. Does 5S RNA function by a switch between two secondary structures? *Nature*, 266:193–194, 1977.

[178] Xing Xu, Yongmei Ji, and Gary D Stormo. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics*, 23(15):1883–1891, 2007.

[179] Charles Yanofsky. Attenuation in the control of expression of bacterial operons. *Nature*, 289:751–758, 1981.

[180] Zizhen Yao, Zasha Weinberg, and Walter L Ruzzo. CMfindera covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–452, 2006.

[181] Ed Yong. Bad copy. *Nature*, 485:298–300, 2012.

[182] Xiao yong Li *et al.* Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS biology*, 6(2):e27, 2008.

[183] Eleftheria Zeggini and John PA Ioannidis. Meta-analysis in genome-wide association studies. *Pharmacogenomics*, 10:191–201, 2009.

[184] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, 1981.