UNIVERSITY OF CALIFORNIA,
MERCED

DISSERTATION

Developing Statistical Models for the Analysis of Genomic Variants

by

Mario Banuelos

A technical report submitted
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Applied Mathematics

2018

Committee Members:
Professor Suzanne S. Sindi, Chair
Professor Roummel F. Marcia
Professor Arnold D. Kim

The Dissertation of Mario Banuelos is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Committee Member:

Professor Arnold D. Kim

Committee Member:

Professor Roummel F. Marcia

Committee Chair / Research Advisor:

Professor Suzanne S. Sindi

Date

Acknowledgments

I am indebted to my advisor Suzanne Sindi for her continual guidance and encouragement. Her mentorship has been one of the most influential parts of my life and a major reason why I fell in love with applied mathematics. I would also like to thank my committee members Roummel Marcia and Arnold Kim for their role in shaping my path as a scientist. Their often orthogonal approaches have transformed how I view both mathematics and mentorship. I also will never forget how graduate students in the applied mathematics group fostered a collegial and successful work environment, and I am fortunate to have spent my time learning and growing with them. Finally, I want to thank the two strongest women in my life. My mother always believed in the transformative power of education and she encouraged me to *sigue adelante*. My wife, Eliana Banuelos, has been my emotional and spiritual support throughout this academic journey, and I am blessed to have her in my life.

I gratefully acknowledge funding from the University of California, Merced's School of Natural Sciences, the UC Merced Fletcher Jones Fellowship, the Bill and Melinda Gates Foundation, and NSF Grant CMMI #1333326.

Contents

Developing Statistical Models for the Analysis of Genomic Variants

by

Mario Banuelos

Doctorate in Applied Mathematics

Suzanne Sindi, Chair

University of California, Merced

2018

Abstract

I develop a number of mathematical and statistical models for the study of genomic variation within and between species. These variants affect an organism's susceptibility to genetic diseases and are also responsible for speciation events. In particular, my work focuses on the following questions:

1. How does DNA causing genomic variation proliferate through the genome of a species?

2. For members of the same species, how can we leverage *a priori* information (i.e, relatedness and sparsity) to improve predictions of genomic variants?

I address these questions in the the context of using noisy and low-quality data. I begin with a review of models of DNA proliferation and detecting these genomic changes in Chapter 1. In Chapter 2, I answer the first question by developing a model which describes non-actively replicating repetitive elements in an organism's genome. Although they comprise a majority of many eukaryotic genomes, these elements are often ignored by models reviewed in Chapter 1. I answer the second question in Chapter 3 by developing a general optimization framework to detect genomic rearrangements in related individuals subject to different sequencing assumptions. In the context of limited and

noisy data, this work is one of the only methods (to my knowledge) that simultaneously predicts variants in a group of individuals instead of post-processing this information. Chapter 4 describes some of the convergence properties of the methods introduced in the previous chapter, and Chapter 5 summarizes this work and future projects.

List of Abbreviations and Symbols

TE	transposable element
SV	structural variant
SNV	single nucleotide variant
ODE	ordinary differential equation
PDE	partial differential equation
bp	basepair (i.e., A, G, C, or T)
α	replication rate of full-length transposable elements
D	deletion rate per unit length per unit time of a genome
I	insertion rate per unit length per unit time of a genome
L	length (in basepairs) of a transposable element
θ	The ratio of duplications to deletions α/DL
$u_L(t)$	the number of full-length transposable elements length L at time t under continuous assumptions
$u(x,t)$	the number of partial length transposable elements of length x at time t under continuous assumptions
$u_0(t)$	the zeroth moment of the distribution of $u(x,t)$
$u_1(t)$	the first moment of the distribution of $u(x,t)$
$U_L(t)$	the number of full-length transposable elements length L at time t under discrete assumptions
$U(i,t)$	the number of partial length transposable elements of length i at time t under discrete assumptions
$\eta(t)$	the zeroth moment of the distribution of $U(i,t)$
$\xi(t)$	the first moment of the distribution of $U(i,t)$
$\Pr[A]$	the probability of event A
$\Pr[A\|B]$	the probability of event A given event B

List of Figures

List of Tables

Curriculum Vitae

Education

2013–2018	PhD, Applied Mathematics. Advised by Prof. Suzanne Sindi. University of California, Merced. Merced, CA.
2012–2013	Single Subject Preliminary Credential, Mathematics. California State University, Bakersfield. Bakersfield, CA.
2008-2012	BA, Mathematics. California State University, Fresno. Fresno, CA.

Publications

- **M. Banuelos**, S. Sindi, and R. Marcia. "Structural variant prediction in extended pedigrees through sparse negative binomial genome signal recovery." Accepted to the *2018 International Conference of the IEEE Engineering in Medicine and Biology Society*.
- **M. Banuelos**, S. Sindi, and R. Marcia. "Negative binomial optimization for biomedical structural variant signal reconstruction." Accepted to the *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*.
- **M. Banuelos**, L. Adhikari, R. Almanza, A. Fujikawa, J. Sahagún, K. Sanderson, M. Spence, S. Sindi, and R. Marcia. "Sparse diploid spatial biosignal recovery for genomic variation detection." Proceedings of the *2017 IEEE International Symposium on Medical Measurements and Applications*.
- **M. Banuelos**, L. Adhikari, R. Almanza, A. Fujikawa, J. Sahagún, K. Sanderson, M. Spence, S. Sindi, and R. Marcia. "Nonconvex regularization for sparse genomic variant signal detection." Proc. of the *2017 IEEE International Symposium on Medical Measurements and Applications*.
- **M. Banuelos**, R. Almanza, L. Adhikari, S. Sindi, and R. Marcia. "Biomedical signal recovery: genomic variant detection in family lineages." Proc. of the *IEEE 5th Portuguese Meeting on Bioengineering*.
- **M. Banuelos**, R. Almanza, L. Adhikari, S. Sindi, and R. Marcia. "Constrained variant detection with sparc: sparsity, parental relatedness, and coverage." Proc. of the *2016 International Conference of the IEEE Engineering in Medicine and Biology Society*.
- **M. Banuelos**, R. Almanza, L. Adhikari, S. Sindi, and R. Marcia. "Sparse genomic structural variant detection: exploiting parent-child relatedness for signal recovery." Proc. of the *2016 IEEE Workshop on Statistical Signal Processing*.

- **M. Banuelos**, R. Almanza, L. Adhikari, S. Sindi, and R. Marcia. "Sparse signal recovery methods for variant detection in next-generation sequencing data." Proc. of the *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Teaching Experience

Spring 2015, 2016, Fall 2016

Graduate Teaching Assistant,
MATH122 Complex Analysis.
University of California, Merced.

Fall 2014, 2015 Graduate Teaching Assistant,
MATH141 Linear Analysis.
University of California, Merced.

Academic Year 2013-2014

Graduate Teaching Assistant,
MATH22 Calculus II for Physical Sciences and Engineering.
University of California, Merced.

Summer 2013 Graduate Teaching Assistant,
MATH11 Calculus I.
University of California, Merced.

Academic Year 2012-2013

High School Mathematics Teacher,
Algebra I and California High School Exit Exam Preparation.
Robert F. Kennedy High School, Delano, CA.

Chapter 1

Introduction

1.1 Genomic Variation Biology

The genome of an individual consists in sequences of nucleotides (A,C,G,T) that ranges in length from millions of letters (for a bacteria) or billions of letters (for a mammalian genome) [27]. However, the evolutionary processes of mutation, coupled with more complex heredity in sexually reproducing organisms, ensures variation between genomes of individuals within a species. Since genomic variation between species far exceeds variation among individuals from the same species, the common practice has been to develop a reference genome for each species along with an annotation of common sites of variation [2, 33]. Genomic variants are classified by their lengths and may either consist of a single letter (nucleotide), so called single nucleotide variants (SNVs), or rearrangements of larger regions of DNA, called structural variants (SVs). In both cases, variation is identified by comparing fragments of DNA sequenced from a test (unknown) genome to a given reference (see Fig. 1.1) [35, 30]. These genomic variants have often been associated with genetic diseases, such as cancer, but also have been attributed to promoting genetic diversity [18]. As such, SVs and variation in general, represents an important part of understanding the recent evolutionary history of a species [31, 40].

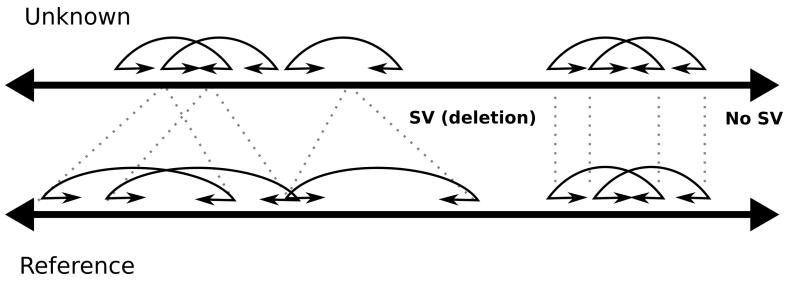

Figure 1.1: Illustration of a deletion in a test genome (unknown) relative to a reference genome (known). Deletions (and other SVs) are identified by sequencing both ends of fragments (of a particular length distribution) from the test genome and mapping them to the reference. Fragments whose mapped distance is significantly larger than expected (left) indicate a potential deletion while fragments which map to the reference (right) indicate no SV.

In humans, genomic diseases and disorders also accompany this diversity while heredity passes these traits from one generation to the next [39, 42]. The advent of next-generation sequencing (NGS) and the decreasing DNA sequencing costs – both producing more data – make detecting variation more amenable. As a result, many methods have emerged to detect these inherited changes. Fig. 1.2 illustrates this process in paired-end sequencing. Although sequencing technology advances, producing longer DNA reads, some emerging methods suffer from higher error rates [3]. The immediate solution is to sequence individuals multiple times to resolve ambiguities, but this comes at a higher financial and computational cost. Moreover, portable sequencing technology provides the opportunity to sequence many individuals at low coverage relatively quickly [24, 34]. In order to take advantage of these technologies, changes must be made in computational and mathematical methods.

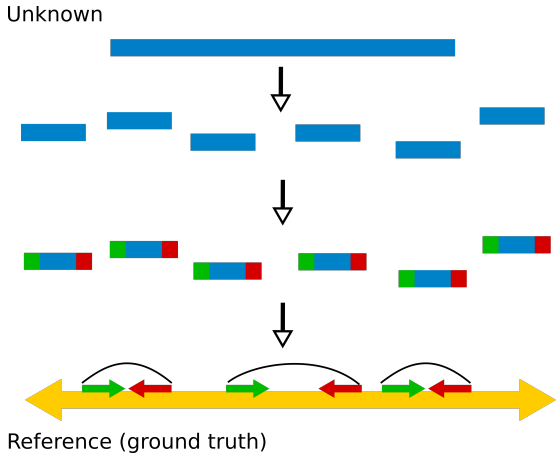

Figure 1.2: Illustration of DNA sequencing of an unknown genome. First, the unknown genome is fragmented. Second, the ends of the fragments are sequenced from both ends (red and green). Lastly, the sequenced ends are mapped to the reference genome. This process is repeated many times to resolve the errors in both sequencing and mapping to the reference genome.

In addition to their length, genomic variants are categorized as deletions, insertions, inversions, and translocations. A major driver of genomic insertions are transposable elements (TEs): mobile DNA sequences that encode their own self-replication [32, 19]. TEs vary in size, ranging from several hundred to several thousand bases. Originally discovered by Barbra McClintock in *Zea mays* [29], these elements comprise the majority of human and other primate genomes and are abundant in the genomes of many other organisms [41, 10, 1]. TEs are divided into two classes (I and II) based on the method they use to duplicate. Class I elements use a “copy-and-paste” mechanism via an RNA intermediate that results in two copies of the TE while Class II elements use “cut-and-paste” mechanism, in which the transposon excises itself and interrupts a target DNA sequence (see Fig. 2.1). TEs have been largely viewed as deleterious since both classes may cause considerable damage; for example, a TE insertion within a gene has been linked to Haemophilia A [17]. In addition, actively replicating TEs increase the size of the genome, and are shown to induce further structural variation such as

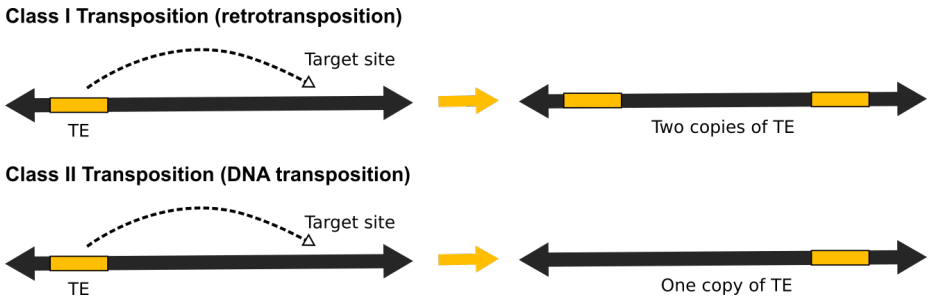

Figure 1.3: Transposable elements are self-replicating DNA sequences that are classified according to their replication process. Transposition occurs during DNA replication, where target sites are identified. (Top) The "copy-and-paste" mechanism yields two copies of the TE in the host genome and is the primary focus of our model. (Bottom) In contrast, the "cut-and-paste" mechanism is primarily found in plants, such as *Zea mays* and results in a new location of the DNA region in the host's genome.

inversions, duplications or deletions [38, 28, 7, 16]. However, the role of TEs is not entirely deleterious and there is increasing evidence of their benefit. For example, TEs may be responsible for genetic regulation as many human promoter regions contain TE-derived sequences [15]. Over time, both due to their self-duplicating nature and the gradual accumulation of mutations in the host genome, a copy of a TE may itself be mutated to the point that the copy loses the ability to replicate. Because of their unique self-replication nature, modeling TE dynamics has been an active area of research.

The vast complexities of molecular and hereditary factors in genomic variation provide many modeling and statistical inference opportunities. I outline existing mathematical literature for both prediction and modeling of genomic variants and then outline our own research in this context.

1.2 Mathematics of Modeling and Detecting Genomic Variants

1.2.1 Modeling Transposable Element Proliferation

The increasing availability of TE annotations for complete genomes allows for a more complete picture of the TE composition in a genome and offers an opportunity for quantitative assessment of different theories on the fitness impact of TEs. A number of mathematical models have been developed to study the dynamics of TEs and most focus on only on full-length replicating copies [5, 22, 21, 23]. Here, I summarize the first models proposed by [5] and more recent developments in modeling transposable element proliferation.

Most studies aimed at describing the evolutionary history of transposable elements have primarily focused on full-length elements, TEs that still encode the ability to replicate, and their copy number. Models have been of two varieties, depending on

their assumptions of how TEs impact the host genome. Neutral models assume TEs are neither detrimental or beneficial to the host, selection models incorporate the fitness of the population, and species-specific models take into account TE assumptions of the host species [22].

Neutral Models

The first neutral model proposed by Charlesworth and Charlesworth assumed TEs self-regulated their transposition according to their copy number, u_n, where u_n is a decreasing function of n (the number of elements in a diploid genome) [5]. For populations of infinite size, the average change in the mean copy number $\bar{n} = 2\sum_i x_i$, where x_i is the frequency of a TE at locus i and the summation is over all possible loci. The change per generation is then given

$$\triangle\bar{n} = \mathbb{E}[nu_n] - \bar{n}v, \tag{1.1}$$

where v is the probability that a TE is deleted and $\mathbb{E}$ is the expectation over all individuals. Expanding the right hand side about $\bar{n}$, the authors show that TE copy number will follow a binomial distribution with mean $\bar{n}$ and variance $\bar{n}(1 - \bar{n}/T)$, where T is the number of sites TEs can occupy. This yields

$$\triangle\bar{n} \approx \bar{n}(u_{\bar{n}} - v), \tag{1.2}$$

where the only nontrivial equilibrium is reached when $u_{\bar{n}^*} = v$ – when self-regulation equals the deletion rate. In finite populations, they show that the distribution of TE counts follow a beta or beta-like distribution [5].

Selection Models

Selection models consider the fitness of the population under the assumption that TEs result in deleterious effects on the genome, and they incorporate a number of selective forces (e.g., ectopic recombination) [4]. Charlesworth and Charlesworth also presented models including host fitness under the assumption that TEs are deleterious to the host [5]. As such, in finite populations, the mean copy number changes by

$$\triangle\bar{n} = \bar{n}\left(1 - \frac{\bar{n}}{T}\right)\frac{\partial \ln w_{\bar{n}}}{\partial\bar{n}} + \bar{n}(u - v), \tag{1.3}$$

where $w_{\bar{n}}$ represents the average population fitness and holds when T is much larger than the mean copy number. In their work, the authors incorporate fitness functions of the form $w_n = (1 - s)^n$ and also include strongly convex functions. An equilibrium exists local if the fitness function satisfies the constraint

$$-\frac{1}{n^*} < \frac{\partial^2 \ln w_{n^*}}{\partial n^{*2}} < 0.$$

and $\left|\frac{\partial \ln w_{\bar{n}}}{\partial\bar{n}}\Big|_{n=0}\right| + v < u$. For finite populations, the authors approximate the mean copy number as an approximate beta distribution, but note that both early neutral and

selection models focus on modeling full-length elements and their respective copy number. Since selective forces affecting TE proliferation vary, assumptions including deleterious insertion (allowing for varying selection coefficients), ectopic recombination, and deleterious transposition models have been explored [22].

Species-Specific Models

As more data became available through advances in sequencing technologies, TE models incorporated *a priori* information about the species in which the transposable element had invaded. These models use known features of TEs in specific species to inform assumptions. One recent model, for example, incorporates ecological frameworks to describe how certain TEs – primarily in primates – utilize existing replication machinery (i.e., RNA polymerase) [43]. The stochastic model addresses the competition between two TEs in a predator-prey setting, but does not include the deterioration of full-length elements into partial length copies (i.e., TEs are completely destroyed at a constant rate). The authors conclude that such models can describe oscillations of at least two populations of transposable elements when one TE hijacks the replicative mechanisms from another TE.

In bacteria, for example, horizontal transfer events are explicitly modeled, but may not be representative of TE dynamics in eukaryotes [13]. Indeed, the use of ecological models and birth-death processes can capture the stochasticity of how full-length elements evolve in time; however, when considering the non-replicating elements, this system grows exponentially and these models will be computationally expensive.

1.2.2 Structural Variation Detection

Structural variants (SVs) are typically defined as genomic rearrangements larger than a certain length (in bp). Although this may be useful in comparing predicted SVs with experimentally validated variants, their complexity may make this an ambiguous definition. Instead, in the most general sense, these rearrangements represent a set of *novel adjacencies* – two positions adjacent in the unknown genome but not in the reference. A deletion, for example, results in a single novel adjacency [36, 35]. As such, the signals associated with detecting SVs are an important component for the development of SV-detection methods. I summarize the three most common signals below in Fig. 1.4 with respect to deletions and describe methods utilizing primarily one signal (see [35] for a discussion on how these signals affect other types of SVs).

Signal-Specific SV Detection Methods

Early methods of structural variation detection focused on using one of these three signals as indicators for possible structural variants. Shortly after the completion of human sequencing efforts, many SV-detection tools emerged using paired-read sequencing signals [6, 37, 11, 33, 20]. Since my research program incorporates the output of the Geometric Analysis of Structural Variants (GASV) method, I focus on describing

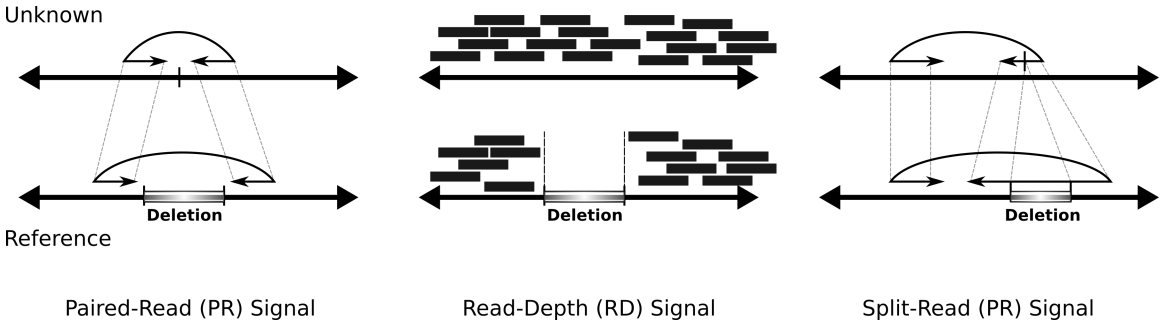

Figure 1.4: Three types of signals suggesting a possible deletion in the unknown genome. (*Left)* Paired-Read (PR) signals occur when DNA reads in the unknown genome map to regions in the reference and the resulting mapping distance is greater than expected distance. (*Center)* A drop in the number of reads mapping to the reference in comparison to unknown genome potentially signifies a deletion. (*Right)* Split-Read (SR) signals result when the expected mapping distance from at least one direction exceeds the expected distance.

this approach and its use of paired-read data [37]. The length of the DNA fragment is defined as L_C, where $L_c \in [L_{\min}, L_{\max}]$ and both $L_{\min}$ and $L_{\max}$ are known. Then, for a true breakpoint (a, b), we expect $L_C = (a - x_C) + (y_C - b)$ when $x_C \leq a,\ b \leq y_C$ (see Fig. 1.5). Since we know L is bounded above and below, we have

$$L_{\min} \leq (a - x_C) + (y_C - b) \leq L_{\max}.$$

Fig. 1.5 illustrates how this method begins to detect a deletions. In this case, one read corresponds to one trapezoid of feasible breakpoints. Since we expect multiple reads for a given region – resulting in many trapezoids – the authors introduced a geometric sweep-line algorithm to narrow the space of admissible breakpoints, and this yields increased sensitivity rates. Although I focus on deletions, GASV also addresses complex variants (e.g., inversions). The output of GASV is thus the number of fragments supporting a potential deletion for a set of genomic coordinates.

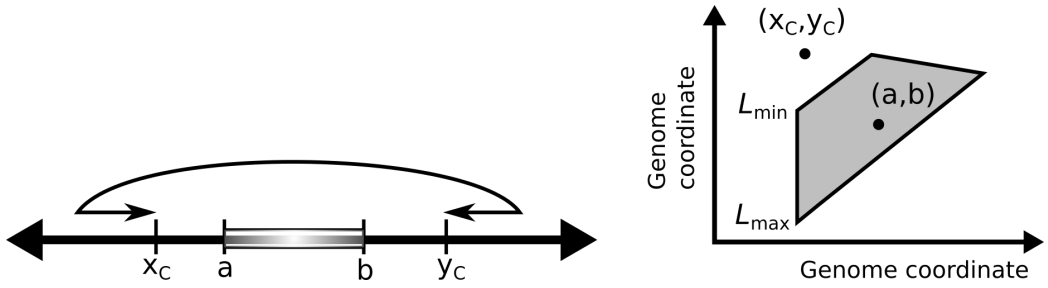

Figure 1.5: *(Left)* (a, b) represent the true breakpoints (deletion coordinates), while (x_C, y_C) indicate where unknown reads were mapped to the reference. In this case, $x_C \leq a,\ b \leq y_C$. *(Right)* Plot of genome coordinates (x_C, y_C) in comparison to (a, b). Since $L \in [L_{\min}, L_{\max}]$, the space of admissible breakpoints are outlined by the grey trapezoid. Note that the true breakpoint (a, b) is contained within this set.

One of the first methods to use read-depth (RD) as a signal for variant detection was developed by [45]. Methods using this type of signal may rely on non-overlapping or a sliding window approach. Not surprisingly, as the window length increases, the precision decreases. Yoon et al. incorporated a non-overlapping window approach on sequencing data that could be approximated by a normal distribution. Once the entire genome is searched, then an event-wise testing metric (via converting counts to a Z score) is defined and unusual events (those that meet criteria of statistical signficance) are categorized as deletions (see middle panel of Fig. 1.4).

Pindel, the first method to use split-read (SR) signals for SV detection, incorporated a pattern growth approach to detect large deletions as well as insertions using paired-end short reads [44]. This general procedure includes mapping all reads to the reference genome and assumes that one end (of the read) mapped to the reference. The user then defines the maximum deletion size and lengths for the minimum and maximum substrings before beginning the algorithm. If two or more reads support a deletion (i.e., thresholding), then the method reports the potential deletion. I summarize the initial pattern growth formulation in Algorithm 1. For a database S and pattern P in an alphabet A, C, G, T, Pindel outputs the minimum and maximum substrings that appear once in S. Once these substrings are reported, the method still relies on thresholding the number of fragments (> 2) to report potential deletions.

Hybrid and Population-level Methods

Current SV-detection methods now tend to rely on multiple signals and consider multiple samples at a time [9, 8, 25]. Given the three common types of SV signals, methods incorporating at least two for SV prediction (for a total of four possible improvements on previous methods) emerge. For example, HYDRA-Multi builds on a previous method may detect deletions in one individual by including information on 65 other individuals. More specifically, this method first identifies candidate SV clusters by independently clustering discordant alignments of individuals. These candidate SVs are then sorted by start position for each individual. Given these sorted candidates, sample-wide clusters are built (for individuals supporting the same variant) and subsequently reported [25]. These methods reduce the number of false positives without adding too much computational cost [9]. Other contemporary methods incorporate a more Bayesian approach, where SV signals are taken into account in a probabilistic framework [12]. In SV-Bay, after clustering discordant fragments, Bayes' rule is applied to detect the most likely model that explains the observed data of potential breakpoints (assuming discordant fragments follow a Poisson distribution). However, explicit heredity information is typically not modeled in these methods.

As sequencing of families has become more tractable, some recent methods have emerged to detect structural variation or single nucleotide variants (SNVs) in family lineages. For example, Canvas (Small Pedigree Workflow) SPW and TrioCNV primarily incorporate multiple SV signals in hidden Markov models to detect copy number variation in parent-offspring trios [26, 14]. As the cost of DNA sequencing decreases, methods amenable to handle genomic data of related individuals will be an important part of

Algorithm 1 Pindel ($P\{m\}$ returns the first m elements in the pattern P.)

```
1:  procedure PATTERN SEARCH
2:      S ← database
3:      P ← pattern
4:      i ← leftmost bp (letter) of P
5:      j ← length(i)
6:  top: return {min, max}
7:  loop:
8:      if j ≤ length(P) and |S_j| ≥ 1 then
9:          if i = 1 then
10:             Scan S for i.
11:             S_j ← projected S with locations of i.
12:             i ← P{i + 1}.
            else
13:         Scan S_j for i.
14:         j ← j + 1.
15:         S_j ← projected S with locations of i.
16:         if |S_j| = 1 then min = i.
17:         if |S_j| = 0 then max = P{i − 1}.
18:             goto top.
19:         i ← P{i + 1}.
20:         goto loop.
21:         close;
22:     goto top.
```

understanding the transmission of genomic variation throughout generations. When considering limited storage requirements in the case of hundreds or thousands of samples, SV methods incorporating noisy and low-quality data will be especially useful.

1.3 Motivation and Goals

Genomic rearrangements in all organisms lead to a wide range of observable phenomena. At times, these changes result in hereditary diseases that can be fatal to the organism. At other times, they lead to rich genetic diversity and add more complexity to the tree of life. With the use of statistical and mathematical models, we aim to model and detect this variation in organisms both between and within the same species.

As reviewed, modeling genomic variation through transposable element proliferation has been extensively explored. However, most models concerned with describing the proliferation of transposable elements tend to focus on actively replicating copies. Partial copies contribute a major source of repetitive DNA in many eukaryotes and some previous methods become less computationally tractable when considering all partial TEs in a genome.

Detecting these DNA rearrangements in organisms concerns itself with other half of trying to answer the disease and diversity question. As described, many of the original and some of the recent methods view variant prediction at the individual level. These methods offer high resolution of potential structural variants; however, we concern ourselves with concurrent prediction in related individuals where the sequencing data quality is obfuscated by noisy measurements.

In Chapter 2, I develop a neutral approach focused on describing the distribution of both transposable elements and non-transposing copies which are unable to proliferate in the host genome. Particularly, we introduce a discrete and continuous deterministic model to study the distribution of TEs for a specific TE length. We focus on Class I elements and their proliferation process. Using moments of the density distributions, we describe the full system. Solutions to the fragmentations equations are presented and our model is interpreted in the context of empirical distributions of genome data.

In Chapter 3, I develop an optimization framework for detecting structural variants in related individuals. This general formulation allows for different relatedness structures and DNA sequencing assumptions. I discuss my contributions to such methods and their applicability to genome sequencing data with different heredity constraints.

In Chapter 4, I discuss the convergence of convex methods introduced in Chapter 3 using the method of Lagrangian multipliers. Additionally, I also discuss the guarantee of local minima for nonconvex methods presented.

Finally, Chapter 5 outlines future work, with a focus on nonconvex optimization methods for structural variation detection. I also discuss future work on mathematical and statistical models for endemic diseases.

Bibliography

[1] Alan M Weiner. *SINEs and LINEs: the art of biting the hand that feeds you*. 2002. DOI: 10.1016/S0955-0674(98)80011-2.

[2] D. M. Altshuler et al. “A Map of Human Genome Variation from Population Scale Sequencing”. *Nature* 467.7319 (2010), pp. 1061–1073.

[3] D. Branton et al. “The potential and challenges of nanopore sequencing”. *Nature biotechnology* 26.10 (2008), pp. 1146–1153.

[4] P. Capy. *Dynamics and evolution of transposable elements*. North American distributor Chapman & Hall, 1998.

[5] B. Charlesworth and D. Charlesworth. “The population dynamics of transposable elements”. *Genetical Research* 42.01 (1983), pp. 1–27.

[6] K. Chen et al. “BreakDancer: an algorithm for high-resolution mapping of genomic structural variation”. *Nature methods* 6.9 (2009), pp. 677–681.

[7] T. A. Elliott and T. R. Gregory. “What’s in a genome? The C-value enigma and the evolution of eukaryotic genome content”. *Phil. Trans. R. Soc. B* 370.1678 (2015), p. 20140331.

[8] R. Elyanow, H.-T. Wu, and B. J. Raphael. “Identifying structural variants using linked-read sequencing data”. *Bioinformatics* 34.2 (2017), pp. 353–360.

[9] M. Hayes and J. S. Pearson. “Detecting large deletions at base pair level by combining split read and paired read data”. *BMC bioinformatics* 18.12 (2017), p. 413.

[10] D. J. Hedges and M. A. Batzer. “From the margins of the genome: mobile elements shape primate evolution”. *Bioessays* 27.8 (2005), pp. 785–794.

[11] F. Hormozdiari et al. “Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes”. *Genome research* 19.7 (2009), pp. 1270–1278.

[12] D. Iakovishina et al. “SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read map-pability”. *Bioinformatics* (2016), btv751.

[13] J. Iranzo et al. “Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes”. *PLoS computational biology* 10.6 (2014), e1003680.

[14] S. Ivakhno et al. “Canvas SPW: calling de novo copy number variants in pedigrees”. *Bioinformatics* (2017).

[15] I. Jordan et al. “Origin of a substantial fraction of human regulatory sequences from transposable elements”. *Trends in Genetics* 19.2 (2003), pp. 68–72. ISSN: 01689525. DOI: 10.1016/S0168-9525(02)00006-9.

[16] H. H. Kazazian. “Mobile elements: drivers of genome evolution”. *science* 303.5664 (2004), pp. 1626–1632.

[17] H. H. Kazazian et al. “Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man” (1988).

[18] J. M. Kidd et al. “Mapping and sequencing of structural variation from eight human genomes”. *Nature* 453.7191 (2008), pp. 56–64.

[19] M. G. Kidwell and D. R. Lisch. “Perspective: transposable elements, parasitic DNA, and genome evolution”. *Evolution* 55.1 (2001), pp. 1–24.

[20] J. O. Korbel et al. “PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data”. *Genome biology* 10.2 (2009), R23.

[21] M. Koroteev and J. Miller. “Scale-free duplication dynamics: A model for ultraduplication”. *Physical Review E* 84.6 (2011), p. 061919.

[22] A. Le Rouzic and G. Deceliere. “Models of the population genetics of transposable elements”. *Genetical research* 85.03 (2005), pp. 171–181.

[23] A. Le Rouzic, T. Payen, and A. Hua-Van. “Reconstructing the evolutionary history of transposable elements”. *Genome biology and evolution* 5.1 (2013), pp. 77–86.

[24] L. Liang et al. “Computational studies of DNA sequencing with solid-state nanopores: key issues and future prospects”. *Frontiers in chemistry* 2 (2014).

[25] M. R. Lindberg, I. M. Hall, and A. R. Quinlan. “Population-based structural variation discovery with Hydra-Multi”. *Bioinformatics* 31.8 (2014), pp. 1286–1289.

[26] Y. Liu et al. “Joint detection of copy number variations in parent-offspring trios”. *Bioinformatics* 32.8 (2015), pp. 1130–1137.

[27] M. Lynch and J. S. Conery. “The origins of genome complexity”. *science* 302.5649 (2003), pp. 1401–1404.

[28] X. Maside, S. Assimacopulos, and B. Charlesworth. “Rates of movement of transposable elements on the second chromosome of Drosophila melanogaster”. *Genetical research* 75.03 (2000), pp. 275–284.

[29] B. McClintock. “The origin and behavior of mutable loci in maize”. *Proceedings of the National Academy of Sciences* 36.6 (1950), pp. 344–355.

[30] P. Medvedev, M. Stanciu, and M. Brudno. “Computational methods for discovering structural variation with next-generation sequencing”. *Nature methods* 6 (2009), S13–S20.

[31] G. of the Netherlands Consortium et al. “Whole-genome sequence variation, population structure and demographic history of the Dutch population”. *Nature Genetics* 46.8 (2014), pp. 818–825.

[32] L. E. Orgel and F. H. Crick. “Selfish DNA: the ultimate parasite”. *Nature* 284 (1980), pp. 604–607.

[33] A. R. Quinlan et al. “Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome”. *Genome research* 20.5 (2010), pp. 623–635.

[34] J. Shendure et al. “DNA sequencing at 40: past, present and future”. *Nature* 550.7676 (2017), p. 345.

[35] S. S. Sindi and B. J. Raphael. “Identification of Structural Variation”. *Genome Analysis: Current Procedures and Applications* (2014), p. 1.

[36] S. S. Sindi et al. “An integrative probabilistic model for identification of structural variation in sequencing data”. *Genome biology* 13.3 (2012), R22.

[37] S. Sindi et al. “A geometric approach for classification and comparison of structural variants”. *Bioinformatics* 25.12 (2009), pp. i222–i230.

[38] R. K. Slotkin and R. Martienssen. “Transposable elements and the epigenetic regulation of the genome”. *Nature Reviews Genetics* 8.4 (2007), pp. 272–285.

[39] P. Stankiewicz and J. R. Lupski. “Structural variation in the human genome and its role in disease”. *Annual review of medicine* 61 (2010), pp. 437–455.

[40] H. Stefansson et al. “A common inversion under selection in Europeans”. *Nature genetics* 37.2 (2005), pp. 129–137.

[41] M. I. Tenaillon et al. “Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians”. *Genome biology and evolution* 3 (2011), pp. 219–229.

[42] J. Weischenfeldt et al. “Phenotypic impact of genomic structural variation: insights from and for human disease”. *Nature Reviews Genetics* 14.2 (2013), pp. 125–138.

[43] C. Xue and N. Goldenfeld. “Stochastic predator-prey dynamics of transposons in the human genome”. *Physical review letters* 117.20 (2016), p. 208101.

[44] K. Ye et al. “Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads”. *Bioinformatics* 25.21 (2009), pp. 2865–2871.

[45] S. Yoon et al. “Sensitive and accurate detection of copy number variants using read depth of coverage”. *Genome research* 19.9 (2009), pp. 1586–1592.

Chapter 2

Modeling Genomic Variation Across Species

This chapter is a submitted journal article (2018), "Sindi, S. S., & Banuelos, M. Modeling Transposable Element Dynamics with Fragmentation Equations. *Mathematical Biosciences*.", and is currently under review.

2.1 Abstract

Transposable elements (TEs), segments of DNA capable of self-replication, are abundant in the genomes of most organisms and thus serve as a record of past mutational events. While some work suggests TEs may serve a regulatory function for the host, most empirical and theoretical studies have shown that TEs often have deleterious effects on a host. Because they are not essential, the host genome consists of both full-length (actively replicating) and partial length (inactive remnant) copies of TEs. We developed a novel mathematical formulation of TE dynamics by modeling the density of full and partial length copies resulting from mutations (insertions and deletions) and TE replication within the host genome. More specifically, we model the time-evolution of the complete TE length distribution (full and partial elements) in a genome using fragmentation equations in both a discrete and continuous framework under two models of TE replication.

In the first case, we assume that full-length TEs replicate at a constant rate regardless of the number of full-length TEs present in the genome. While this assumption simplifies the underlying biological processes, it allows us to derive an explicit analytical form of the time-varying TE density, as well as the steady-state behavior, under both discrete and continuous formulations. Next, we take into account the potential deleterious effects of TEs by modeling TE replication with a logistic growth equation. Under this assumption, the number of actively replicating TEs in a genome is limited by a carrying capacity. While we are still able to derive to derive analytical forms for the time-varying TE density, for both discrete and continuous length formulations, these solutions are implicit. For all four proposed models, we prove existence and uniqueness of these

solutions describing TE length distributions. We compare both models and note that the logistic and exponential models initially agree. Since most TEs have not reached carrying capacity, we use the explicit exponential solution to quantify rates of replication to mutations. We apply our model to present day annotated collections of TEs from the genomes of species of fruit-flies, birds, and primates to uncover quantitative relationships of TE dynamics. With the increasing availability of complete genomes, our method is likely to uncover relationships of biological drivers of genomic variation in many species.

2.2 Introduction

The genome, the complete DNA sequence of an organism, may also contain numerous copies of transposable elements (TEs): mobile DNA sequences that encode their own self-replication [46, 31]. TEs vary in size, ranging from several hundred to several thousand bases. Originally discovered by Barbara McClintock in *Zea mays* [44], remnants of these elements comprise the majority of human and other primate genomes and are abundant in the genomes of many other organisms [58, 21, 2]. TEs are divided into two classes (I and II) based on the method they use to duplicate. Class I elements use a "copy-and-paste" mechanism via an RNA intermediate that results in two copies of the TE while Class II elements use "cut-and-paste" mechanism, in which the transposon excises itself and interrupts a target DNA sequence (see Figure 2.1). TEs have been largely viewed as deleterious since both classes may cause considerable damage; for example, a TE insertion within a gene has been linked to Haemophilia A [30]. Over time, both due to their self-duplicating nature and the gradual accumulation of mutations in the host genome, a copy of a TE may itself be mutated to the point that the copy loses the ability to replicate. Both empirical and theoretical studies suggest that TE often have deleterious effects on the host and negatively affect host fitness [46, 52, 12]. In small asexual populations, for example, deleterious TEs may lead to their extinction [12].

In addition, actively replicating TEs increase the size of the genome, and are shown to induce further structural variation such as inversions, duplications or deletions [51, 43, 13, 29]. However, the role of TEs is not entirely deleterious and there is increasing evidence of their benefit in the development of past and recent regulatory functions in humans [31, 24, 59]. For example, TEs may be responsible for genetic regulation as many human promoter regions contain TE-derived sequences [24]. The balance between the counteracting processes of mutations and TE replication rates is thus a growing area of interest [48]. Recent studies also attempt to reconstruct the phylogeny of TE copies [37, 52, 10] to gain a better insight on how these sequences persists through generations.

The increasing availability of TE annotations for complete genomes allows for a more complete picture of the TE composition in a genome and offers an opportunity for quantitative assessment of different theories on the fitness impact of TEs. Although a number of mathematical models have been developed to study the dynamics of TEs, most have focused only on full-length replicating copies [9, 36, 34, 37]. However, because most copies of TEs are partial (i.e., non-replicating) such models do not take full advantage of all annotated sequences. We take a different approach from prior

Figure 2.1: Transposable elements are self-replicating DNA sequences that are classified according to their replication process. Transposition occurs during DNA replication, where target sites are identified. The "copy-and-paste" mechanism yields two copies of the TE in the host genome and is the primary focus of our model.

methods and model both the full-length (active) TEs, capable of duplication, and the partial length (inactive) copies of TEs that are present in the genome. We focus on the Class I TEs that are capable of their own reproduction through the "copy-and-paste" mechanism and quantify the long-term length-distribution of all TE sequences. We study how the length distribution of full and partial TEs changes under both TE replications and mutations in the genome (insertions and deletions). To our knowledge, our work is the first mathematical framework to do so. We fit our model to present-day TE annotations and attempt to quantify TE proliferation and mutation.

More specifically, in this work we explore the dynamics of both partial length and full elements from different TE families using deterministic fragmentation equations. We consider both discrete and continuous formulations of the TE length distribution as well as two different models of TE replication dynamics. In Section 2, we review previous mathematical models for TE proliferation. In Section 3, we describe our assumptions on TE replication and the mutational processes of deletion and insertions and how all impact the length distribution of TEs. In Section 3.1, we derive discrete length models of TE dynamics under two TE replication conditions (exponential and logistic). In Section 3.2, we derive the analogous continuous length TE models under both models of TE dynamics. In both cases, we prove existence and uniqueness of solutions and derive analytical forms of the time-varying TE length distribution for our particular initial condition. In Section 4, we use the analytical solution of one of our models to estimate the ratio of replication to mutation rates in TE annotations from species of fruit flies, birds, and primates. By fitting the model to these present-day representative samples, we quantify relative transposition rates and provide insights into possible differences of TE mobility in different species. In Section 5, we conclude with the significance of our results and extendability of our method. With increasing availability of complete genomes, models such as the one we have developed seem likely to provide further quantitative insights by comparing to the genomic fossil-record TEs have left behind.

2.3 Previous Mathematical Models

Most previous studies modeling the evolution of transposable elements have focused on full-length elements, TEs that still encode the ability to replicate, and their copy number. These models fall into two categories based on their assumptions on how TEs impact the host: general or species-specific. General models may include neutral and

selection processes. Neutral models assume TEs are neither detrimental or beneficial to the host, selection models incorporate the fitness of the population. Species-specific models that take into account TE assumptions of the host species [36, 22, 62]. The first neutral model proposed by Charlesworth and Charlesworth proposed TEs self-regulated their transposition according to their copy number, u_n [9]. For populations of infinite size, the average change in the mean copy number n per generation was modeled by

$$\triangle \bar{n} \approx \bar{n}(u_{\bar{n}} - v) \tag{2.1}$$

where v is the constant rate of deletion. Equilibrium is reached when self-regulation equals the deletion rate. In finite populations, they show that elements follow a beta or beta-like distribution [36]. Selection models consider the fitness of the population under the assumption that TEs result in deleterious effects on the genome, and they incorporate a number of selective forces (e.g., ectopic recombination) [7].

Other models use known features of TEs in specific species to inform assumptions. One recent model, for example, incorporates ecological frameworks to describe how certain TEs – primarily in primates – utilize existing replication machinery (i.e., RNA polymerase) [62]. We note that this stochastic model considers the competition between two TE famlies, but does not include the deterioration of full-length elements into partial length copies (i.e., TEs are completely destroyed at a constant rate). Still, other models focus on using TE data and phylogenetic trees to account for TE loads in different species [56]. In this case, the authors describe how the effect of purifying selection and genetic drift in nematodes (and not any other processes) govern present-day TE loads. Although the authors incorporate a less stringent criterion for a transposable element, TEs of length $<$ 2000 base pairs (bp) are excluded in this study. Thus, we believe that our model serves to be complementary to this work as we consider shorter partial length copies. A recent model depicted TE evolution in bacteria and included the mechanism of horizontal gene transfer [22]. While intriguing, this model may not be representative of TE dynamics in eukaryotes. The use of ecological models and birth-death processes can capture the stochasticity of how full-length elements evolve in time; however, to include the dynamics of non-replicating elements results in an exponential growth in the size of the system and result in high computational cost to analyze.

A few methods have considered the dynamics of *partial length* or nonfunctioning TE copies, but have assumed that a nonfunctioning TE copy still utilize the TE transposition machinery to propagate throughout the host [36, 35]. In our work, we take an alternate approach by equating TE replication function with its length. In summary, the majority of published models focus on modeling TE growth in terms of copy number and rarely consider the distribution of non-transposing TE remnants. Our approach is novel because we focus on describing the distribution of non-transposing copies, which are unable to duplicate in the host genome, along with the actively replicating full-length TEs.

2.4 Mathematical Derivation and Model

2.4.1 Mathematical Model Assumptions

Present-day transposable element (partial and full-length) distributions in annotated genomes present a complex picture of organisms' evolutionary past. We assume that three evolutionary processes are responsible for the present-day length distributions: TE replication, insertions and deletions. In this work, we model the length distribution resulting from a single type of one TE in both a continuous and discrete setting and consider how the mutational processes (insertions and deletions) impact the number of full-length TEs and the distribution of partial TE lengths under two different models of TE replication. As shown in Figure 2.2, only full-length copies of TEs are capable of replication, creation of another full-length copy of the same TE elsewhere in the genome. But, other mutational processes, such as the insertion or deletion of a segment of DNA, may cause the complete loss or partial loss of a TE. We consider a host genome with one TE of full-length size L initially present. Our models consider the time evolution of the distribution of lengths of all instances of this TE in the genome. We first describe each of our evolutionary processes (deletions, insertions and replications) and then the resulting impact these processes have on the existing full and partial length TEs in the genome (see Figure 2.3). For simplicity, we present these details in the discrete length context. We then present our mathematical model of full and partial TE lengths in four different mathematical settings: discrete and continuous formulations of the full and partial TE length distribution as well as two biological assumptions on replication processes of full-length TEs.

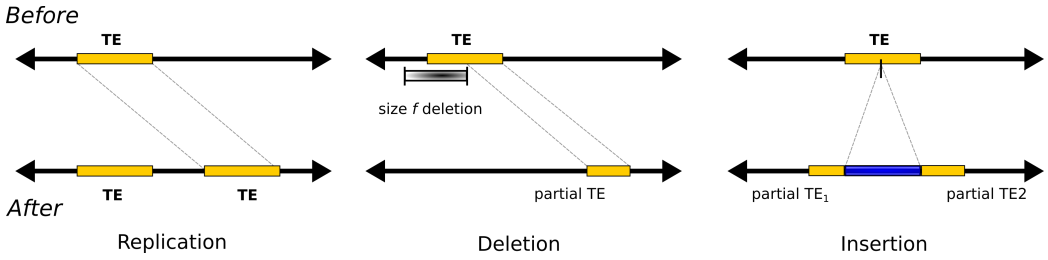

Figure 2.2: Three types of events impact the length distribution of TEs in host genomes: replications, deletions, and insertions. *(Left)* Full-length TEs are capable of self-replicating and inserting a full-length TE copy to a new location in the genome. *(Center)* Deletions of any length may delete all or part of an existing TE. *(Right)* Insertions of any length may occur, resulting in two (potentially different length) partial-length TEs.

Deletions. We assume deletions affect the host genome by selecting a starting position uniformly at random and subsequently deleting one or more positions after it in accordance with the deletion length. We assume that all deletion lengths are enumerated in the finite set $\mathbb{D}$ and that deletions of all lengths occur at a constant deletion rate per unit length per unit time D. (For example, if $\mathbb{D} = \{1, 3\}$, deletions of length 1bp and 3bp are equally likely.) Thus, the instantaneous rate of deletions of all possible sizes is given

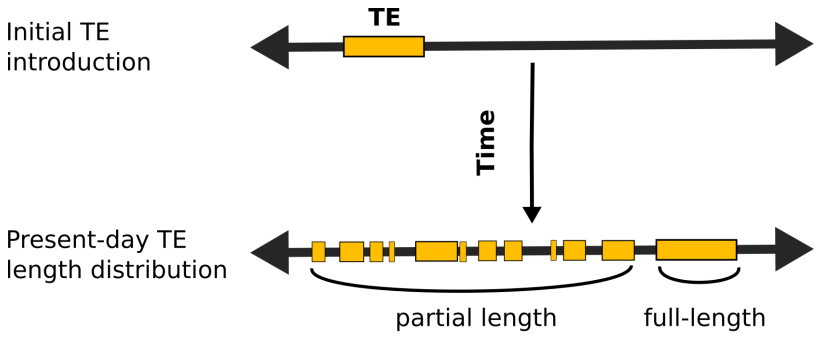

Figure 2.3: **TE Dynamics**. We assume a transposable element (TE) is introduced in a species' genome at an initial time. The three processes of TE replication, deletions, and insertions continue resulting in the complex TE distributions observed in present day genomes. Our model aims to enable quantitative analysis of present TE annotations (full and partial length TEs) to infer rates related to their past evolutionary history.

by: $cDG(t)$ where $c = |\mathbb{D}|$ is the cardinality of the set $\mathbb{D}$ and $G(t)$ is the host genome size. Our formulation allows for deletions of any length f where $1 \leq f$, thus $\mathbb{D} \subseteq \mathbb{Z}^+$. Since most genomic deletions tend to be less than 5 basepairs (bp) in length, a biologically relevant choice for $\mathbb{D}$ would favor deletions of length ≤ 5 [47] [1].

We next consider how deletions impact existing full length and partial length TEs. We distinguish between the rate at which the deletions cause the loss of existing full and partial length TEs and the rate at which the deletions increase the number of partial length TEs by reducing the length of a existing longer full or partial length TE. Consider an existing TE with length i and a deletion with length f. There are $(f + i - 1)$ starting positions in the genome for this deletion that would impact this TE, causing the loss of this length i TE. Thus, the probability that a deletion with length f causes the loss of an element of length i in a genome of size $G(t)$ is simply

$$\frac{f + i - 1}{G(t)}.$$

Since we allow deletions of any length $f \in \mathbb{D}$, the rate that a deletion of any allowable length causes the loss of an element of length i is given by

$$DG(t)\frac{\sum_{f\in\mathbb{D}}(f + i - 1)}{G(t)} = D\sum_{f\in\mathbb{D}}(f + i - 1)\,.$$

We next consider the rate at which deletions increase the number of TEs of length i by impacting a longer TE. For each deletion length f and TE with length $j > i$, there are two starting positions in the genome for this deletion that would result in the creation of a length i TE. For example, if $\mathbb{D} = \{1, 2\}$, a length 4 TE may be created from a length 5 TE through deletions of each possible length in two ways (see Figure 2.4). Thus, the

[1] As we discuss in Section 2.6, it is possible to consider different probabilities for different lengths.

probability of a deletion creating a length i TE through the partial deletion of a length j TE is

$$\frac{2c}{\sum_{f\in\mathbb{D}}(f+j-1)}.$$

Insertions. We assume genomic insertions in the host genome occur by selecting a random starting position (uniformly) and then inserting DNA at that position. Our formulation also allows for insertions of any length k where $k \geq 1$, where insertion lengths are enumerated in the finite set $\mathbb{I}$. We assume that insertions of all lengths (e.g., a single nucleotide or *other* TE replications) occur at the constant rate of insertions per unit length per unit time I [40, 41]. As such, the genomic insertion rate is $IG(t)$. In this way, insertions (of all lengths) impacting $G(t)$ at any starting position results in two regions – one on either side of the insertion (Figure 2.4). Because the length of the insertion is not relevant for the TE length problem, we need not specify our insertion length distribution at this point. In practice, however, insertions, like deletions, favor small sizes [47].

Next, we consider how insertions impact both full and partial length transposable elements. Similar to deletions, we differentiate between the rate of TE loss and the rate at which insertions increase the number of partial length TEs. For any insertion length, consider a full or partial length TE of length i. There are $(i-1)$ starting positions in the genome for an insertion to impact the loss of this length i element. Then, the probability of an insertion causing the loss of a length i element is given by

$$\frac{i-1}{G(t)},$$

and the rate that an insertion results in loss of an existing length i TEs is

$$IG(t)\frac{i-1}{G(t)} = I(i-1).$$

We next consider the rate insertions impact length $j > i$ elements, resulting in the increase in the number of length i elements. Consider a specific length j element. There are two starting positions in the genome where an insertion would create a length i element from this length j element. Regardless of the insertion length, the effect on TE mutation remains the same (see Figure 2.4). Given an insertion impacts a length j TE, the probability of an insertion creating a length i TE from this length j TE is given by:

$$\frac{2}{j-1}.$$

TE Replications. We assume that only TEs that retain their full-length (L) are capable of replicating within the host genome and that replication occurs through the selection of a random position in the genome and the insertion of that full-length TE to that location. We assume that replication occurs at a rate per full-length TE copy per unit time and consider two models of TE replication. In the first model, we assume the replication

rate per TE copy per unit time is a fixed constant α. This formulation will result in the exponential growth or decay of full-length TEs. In the second formulation, we follow prior approaches and assume that too many full-length TE copies become detrimental to host fitness and, as such, the replication rate itself is impacted by the number of full-length TEs already in the genome. In our model, we assume that TEs will not insert themselves into a copy from their same family, but this condition may be generalized [14].

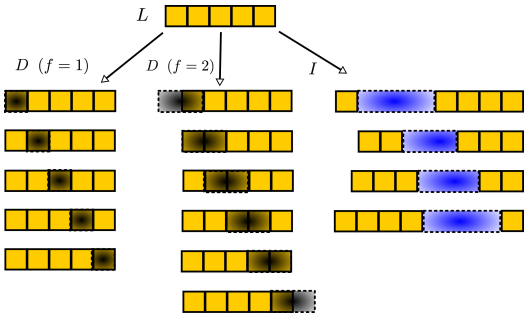

Figure 2.4: Illustration of the mutation mechanisms allowed in our model. In this case, the TE length $L = 5$. We consider the impact of deletions of length 1 and 2 and insertions of any length. Length 1 deletions may impact full-length elements in $L = 5$ ways, and length 2 deletions may impact them in $L+1 = 6$ ways, for a total of 11 potential deletion events. Thus, $\mathbb{D} = \{1, 2\}, c = |\mathbb{D}| = 2,$ and $F = \sum_{f\in\{1,2\}} f = 3$, and we note the quantity $F+c(L-1) = 3+2(4) = 11$ confirms all possible deletion events. Since insertions of any length create 2 partial-length elements *(right)*, I represents the rate of all insertions per unit length per unit time.

2.4.2 Discrete Dynamics

We next will develop our discrete length models by combining the mutational processes we developed above. We track the number of partial length TEs of length i as a function of time t; $U_i(t)$. Elements of the length i will be lost through deletions and insertions, but will be gained by deletions and insertions impacting elements of length $j > i$. To account for all possible lengths $j > i$, we will sum over all elements from $j = i + 1$ to L. Thus, the differential equation for each $U_i(t)$, when $i < L$, is given as follows:

$$\begin{aligned}\frac{dU_i}{dt} &= \underbrace{-D\left(\sum_{f\in\mathbb{D}}(f+i-1)\right)U_i}_{\text{Loss of } U_i \text{ from deletions}} - \underbrace{I(i-1)U_i}_{\text{Loss of } U_i \text{ from insertions}} \\ &+ \underbrace{\sum_{j=i+1}^{L} D\left(\sum_{f\in\mathbb{D}}(f+j-1)\right)\frac{2c}{\sum_{f\in\mathbb{D}}(f+j-1)}U_j}_{\text{Gain in } U_i \text{ from deletions of length } j>i} + \underbrace{\sum_{j=i+1}^{L} I(j-1)\frac{2}{j-1}U_j}_{\text{Gain in } U_i \text{ from insertions of length } j>i} \\ &= (-DF-(cD+I)(i-1))U_i + 2(cD+I)\sum_{j=i+1}^{L} U_j,\end{aligned}$$

where $c = |\mathbb{D}|$, $\sum_{f\in\mathbb{D}} f = F$, and the first term represents loss of elements of length i and subsequent terms account for the gain of elements of length i created by deletions of

$U_j(t)$; $i < j \leq L$. Letting $\gamma = 2cD + 2I$, we write the above more succinctly as

$$\frac{dU_i}{dt} = (-DF - \gamma(i-1))U_i + 2\gamma \sum_{j=i+1}^{L-1} U_j + 2\gamma U_L(t). \tag{2.2}$$

Since we are considering different models for $U_L(t)$, we separate $U_L(t)$ in Equation (2.2). We next describe the proliferation of full-length elements.

Full-length TEs. We now present two different models for the evolution of the number of full-length TEs, $U_L(t)$: exponential and logistic growth. We first consider the case where the number of full-length elements $U_L(t)$ replicate at a constant rate producing an exponentially growing number of full-length TEs. Recalling that $c = |\mathbb{D}|$ and $\sum_{f\in\mathbb{D}} f = F$, the number of full-length TEs, $U_L(t)$, changes as follows:

$$\begin{aligned}\frac{dU_L}{dt} &= \underbrace{\alpha U_L}_{\text{TE replication}} \underbrace{- D\left(\sum_{f\in\mathbb{D}}(f+L-1)\right)U_L}_{\text{Loss of } U_L \text{ from deletions}} - \underbrace{I(L-1)U_L}_{\text{Loss of } U_L \text{ from insertions}}, \\ &= (\alpha - DF - (cD+I)(L-1))\,U_L,\end{aligned}$$

where α is the rate per unit time that a transposable element replicates. We simplify this differential equation as

$$\frac{dU_L}{dt} = (\alpha - DF - \gamma(L-1))\,U_L. \tag{2.3}$$

Since previous studies indicate a potentially deleterious effect of TE proliferation on the host, we also consider a logistic growth model which limits the number of full-length TEs in the host at a given time. We accomplish this goal by considering a carrying capacity K and modeling TE proliferation through a logistic equation. Thus, we have

$$\begin{aligned}\frac{dU_L}{dt} &= \left(\underbrace{\alpha}_{\text{TE replication}} - \underbrace{\left(\sum_{f\in\mathbb{D}}(f+L-1)\right)}_{\text{Loss of } U_L \text{ from deletions}} - \underbrace{I(L-1)}_{\text{Loss of } U_L \text{ from insertions}}\right)\left(1 - \frac{U_L(t)}{K}\right)U_L(t) \\ &= (\alpha - DF - \gamma(L-1))\left(1 - \frac{U_L(t)}{K}\right)U_L(t)\end{aligned}$$

which simplifies to

$$\frac{dU_L}{dt} = r_d\left(1 - \frac{U_L(t)}{K}\right)U_L(t), \tag{2.4}$$

where $r_d = \alpha - DF - \gamma(L-1)$, K is the carrying capacity of full-length elements, α, D and I remain the same as before. We note as $K \to \infty$, the carrying capacity model approaches the exponential growth model. Partial-length elements are still modeled by Equation (2.2). We note that

$$U_L(t) = \frac{KU_L(0)e^{r_d t}}{K + U_L(0)[e^{r_d t} - 1]},$$

where $r_d = \alpha - DF - \gamma L(L-1)$, only depends on t.

In this general framework, Equation (2.2) and either Equation (2.3) or (2.4), we study TE dynamics by considering the zeroth and the first moment of the source distribution, which leads to the following lemma

Lemma 1. *For Equations* (2.2) *and* (2.3) *(exponential growth), the partial length distribution of transposable elements, $U(i,t)$, can be described completely by the zeroth and first moments defined as*

$$\eta(t) = \sum_{i=1}^{L-1} U_i(t), \qquad \xi(t) = \sum_{i=1}^{L-1} iU_i(t).$$

In particular, if full-length elements are governed solely by a time-varying solution $U_L(t)$, Equation (2.2) *has the following moment closure*

$$\frac{d\eta}{dt} = (-DF - \gamma)\eta(t) + \gamma\xi(t) + 2\gamma(L-1)U_L(t) \tag{2.5}$$

$$\frac{d\xi}{dt} = -DF\xi(t) + \gamma L(L-1)U_L(t). \tag{2.6}$$

When $U_L(t) = \exp\left[(\alpha - DF - \gamma(L-1))\,t\right]$ and for the initial condition $U_L(0) = 1$ and $U(i,0) = 0$, we present solutions for $\eta(t)$ and $\xi(t)$ as Equations (2.20) *and* (2.21)*. The solution $\eta(t)$ allows us to rewrite our discrete system as*

$$\vec{U}' = A\vec{U} + \vec{g}(t),$$

where $A \in \mathbb{R}^{(L-1)\times(L-1)}$ and $\vec{g}(t) \in \mathbb{R}^{(L-1)}$ are defined in Equation (2.24) *in 2.7.2. Thus, the explicit analytic solution $\vec{U}$ of Equation* (2.2) *is given by*

$$\vec{U}(t) = S\vec{x}_{\text{exp.}}(t), \tag{2.7}$$

where the matrix of left eigenvectors $S \in \mathbb{R}^{(L-1)\times(L-1)}$ of A is defined in Equation (2.26) *in 2.7.2 and for $i = 1, \ldots, L-1$,*

$$x_i(t) = \gamma(i^2+i)\left[K_2(U_L(t) - e^{\lambda_i t}) + \frac{C_1}{(1+i)\gamma}\left(e^{-DFt} - e^{\lambda_i t}\right) + \frac{C_2}{i\gamma}\left(e^{(-DF-\gamma)t} - e^{\lambda_i t}\right)\right],$$

where λ_i are the eigenvalues of A and constants K_2, C_1, and C_2 are defined in 2.7.1.

Moreover, when we consider the discrete logistic model, we derive an implicit analytic solution to Equation (2.2),

Lemma 2. *For Equations* (2.2) *and* (2.4) *(logistic growth), when $U_L(t) = K\left(1 + \exp\left[(\alpha - DF - \gamma(L-1))\,t\right](K-1)\right)^{-1}$ and for the initial condition $U_L(0) = 1$ and $U(i,0) = 0$, the partial length distribution of transposable elements, $U(i,t)$, can be described completely by the zeroth and first moments defined as*

$$\eta(t) = \sum_{i=1}^{L-1} U_i(t), \qquad \xi(t) = \sum_{i=1}^{L-1} iU_i(t).$$

Moreover, Equations (2.5) *and* (2.6) *describe how the zeroth and first moments change in time. The solution* $\eta(t)$ *allows us to rewrite our discrete system as*

$$\vec{U}' = A\vec{U} + \vec{g}(t),$$

where $A \in \mathbb{R}^{(L-1)\times(L-1)}$ *and* $\vec{g}(t) \in \mathbb{R}^{(L-1)}$ *are defined in Equation* (2.24) *in 2.7.2. Thus, the explicit analytic solution* $\vec{U}$ *of Equation* (2.2) *is given by*

$$\vec{U}(t) = S\vec{x}_{\log.}(t), \tag{2.8}$$

where $S \in \mathbb{R}^{(L-1)\times(L-1)}$ *is defined in Equation* (2.26) *in 2.7.2 and for* $i = 1, \ldots, L-1$,

$$x_i(t) = \tfrac{i(i+1)}{2} \int_0^t 2\gamma \left[e^{(-DF-\gamma)t} \left(\gamma \int \xi(s)ds + 2\gamma(L-1) \int U_L(s)ds \right) + U_L(\tau) \right] e^{\lambda_i(t-\tau)} d\tau.$$

We present solutions $\eta(t), \xi(t)$ and proofs for the solutions to the exponential and logistic models (Lemma 1 and Lemma 2) in 2.7.7.

2.4.3 Continuous Dynamics

Next, we consider the continuous analog of the discrete system presented in Section 2.4.2 for partial length elements $u(x,t)$ and full-length elements $u_L(t)$. The allowable set of deletion lengths $\mathbb{D}$ now may include any real-valued deletion lengths f. Hence, the rate of loss for elements of length x is given as $DF + \gamma x$ where F, c, and γ remain the same as in the discrete model. Next, we differentiate between how the length distributions of partial and full-length elements evolve in time.

Partial length TEs. For an element of fixed length $y > x$, the rate at which length x elements are created is 2γ. To account for all possible lengths $y > x$, we will integrate over all elements y to L. We now develop the continuous model of partial length TEs, $u(x,t)$. The continuous analog of Equation (2.2) is given by

$$x < L, \qquad \frac{\partial u}{\partial t} = (-DF - \gamma x)u(x,t) + 2\gamma \int_x^L u(y,t)\, dy + 2\gamma u_L(t), \tag{2.9}$$

where $u_L(t)$ represents the number of full-length TEs, $u(x,t)$ describes partial length TEs, $\gamma = cD + I$, and all other constants remain the same. We note that the allowable set of deletions $\mathbb{D}$ may now include non-integer deletion lengths. Because we are interested in the distribution of repetitive elements in time for a genome with initially no TEs present, we have initial conditions $u(x,0) = 0$ and $u_L(0) = 1$. Again, we consider two models for TE replication.

Full-length TEs. For the continuous model, we again consider two models of TE proliferation: exponential and logistic growth. In the exponential model, full-length elements $u_L(t)$ change as follows

$$\frac{du_L}{dt} = (\alpha - DF - \gamma L)u_L, \tag{2.10}$$

where α, F, and γ remain the same as in the discrete system. Next, we consider the model where full-length elements $u_L(t)$ are limited in their proliferation. As such, we have

$$\frac{du_L}{dt} = r_c \left(1 - \frac{u_L(t)}{K}\right) u_L(t), \tag{2.11}$$

where $r_c = \alpha - DF - \gamma L$, and α, K and D remain the same. The general solution for Equation (2.11) is thus

$$u_L(t) = \frac{K u_L(0) e^{r_c t}}{K + u_L(0)[e^{r_c t} - 1]}.$$

As in the discrete case, our model of TE dynamics is governed by the zeroth and first moment and we arrive at the following lemma

Lemma 3. *For Equations* (2.9) *and* (2.10) *or Equations* (2.9) *and* (2.11)*, the partial length distribution of transposable elements, $u(x,t)$, can be described completely by the zeroth and first moments defined as*

$$u_0(t) = \int_0^L u(x,t)\, dx, \qquad u_1(t) = \int_0^L xu(x,t)\, dx.$$

In particular, if full-length elements are governed solely by a time-varying solution $u_L(t)$, Equation (2.9) *has the following moment closure*

$$\begin{aligned} u_0'(t) &= -DFu_0(t) + \gamma u_1(t) + 2\gamma L u_L(t) \\ u_1'(t) &= -DFu_1(t) + \gamma L^2 u_L(t). \end{aligned} \tag{2.12}$$

Applying Lemma 3, we determine a solution $u(x,t)$ and prove its uniqueness as follows:

Theorem 1. *If $u(x,0) = 0$, $u_L(0) = u_{L_0} = 1$ and $\alpha > \gamma L$, there exists a solution $u(x,t)$ to the continuous TE length distribution model, Equations* (2.9) *and* (2.10)*, given by*

$$\begin{aligned} u(x,t) &= \mathscr{L}^{-1}\{V(x,s)\}, \\ u_L(t) &= \exp[(\alpha - DF - \gamma L)t], \end{aligned} \tag{2.13}$$

where $\mathscr{L}^{-1}$ is the inverse Laplace transform and $V(x,s) = J(s)(DF + s + \gamma x)^{-3}$ and $J(s)$ is defined in Equation (2.32)*. Moreover, when $F = L$, the solutions $u(x,t)$ and $u_L(t)$ are given by:*

$$\begin{aligned} u(x,t) = {} & \frac{-\gamma\left(e^{\nu}\left(2\alpha^2 + \omega^4 t^2 + 4\alpha^2\omega t - 2\omega^3 t(\alpha t - 1) + \alpha\omega^2 t(\alpha t - 6)\right)\right)}{(\alpha - \omega)^3} \\ & + \frac{-2\gamma\alpha^2 e^{t(\alpha - L(\gamma + D)}}{(\alpha - \omega)^3} \\ u_L(t) = {} & \exp[(\alpha - DL - \gamma L)t], \end{aligned} \tag{2.14}$$

where $\gamma = cD + I$, $\omega = \gamma(L - x)$ and $\nu = -t(DL + \gamma x)$. Moreover, this solution is unique.

We provide a proof of existence and uniqueness of the solution in 2.7.7. In the exponential model, we will use the explicit form for $u(x,t)$ to make inferences about the ratio of replication to mutation in real TE data. In the logistic case, we arrive at an implicit analytic solution defined by Laplace transforms and integrals,

Theorem 2. *If $u(x,0) = 0$, there exists a solution $u(x,t)$ to the continuous TE length distribution model, equations* (2.9) *and* (2.11)*, given by*

$$\begin{aligned} u(x,t) &= \mathscr{L}^{-1}\{V(x,s)\}, \\ u_L(t) &= K\left(1 + \exp[-(\alpha - DF - \gamma L)t](K-1)\right)^{-1}, \end{aligned} \tag{2.15}$$

where $\mathscr{L}^{-1}$ is the inverse Laplace transform and $V(x,s) = J(s)(DF + s + \gamma x)^{-3}$ and $J(s)$ is defined in Equation (2.43)*. Moreover, this solution is unique.*

2.5 Results

We now focus on comparing both discrete models and solutions we developed in Section 2.4. In particular, we present how each TE replication model leads to different TE length and probability distributions. We also discuss how the genome length is affected under these two different TE replication assumptions. Given these solutions and comparisons, we estimate the replication-to-mutation rate using real data of TEs in *Drosophila*, *Aves* and *Primates*.

2.5.1 Model Comparisons

We next compare the evolving distribution of TE lengths under our two models of TE replication. For ease in exposition, we focus on our discrete formulation, but note the same behavior holds in the continuous case. As might be expected, the initial dynamics of the TE length distribution under the logistic model is well approximated by a replication model with the same rate parameter:

$$r_d = \alpha - DF - \gamma(L-1).$$

When applying our initial condition $U_L(0) = 1$, we denote the difference of growth rates between the models at a fixed time t as

$$J(r_d,t) = r_d - r_d\left(1 - \frac{u_L(r_d,t)}{K}\right) = \frac{e^{r_d t}}{K + e^{r_d t} - 1}. \tag{2.16}$$

Further, we note that the second derivative of this increasing function,

$$\frac{\partial J}{\partial r_d} = \left(k + e^{r_d t} - 1\right)^{-3}(k-1)te^{r_d t}\left[(k-1)(r_d t + 2) + e^{r_d t}(2 - r_d t)\right],$$

has one zero when $K > 1$, namely

$$K = 1 + \frac{e^{r_d t}(r_d t - 2)}{r_d t + 2}. \tag{2.17}$$

We find that the exponential model remains a good approximation to the logistic model until the time at which the difference in their growth rates changes concavity. In practice, however, this equality will depend on the time t and not the set rates r_d and carrying capacity K. Thus, we approximate $\partial J / \partial r_d = 0$ as the first time the right hand side of Equation (2.17) exceeds K. For our subsequent comparisons, we assume $r_d = 0.0017$ ($\alpha = 0.0025, D = 5^{-8}, L = 5000, \mathbb{D} = \{1, 2, 3, 4, 5, L - 15\}$) and we vary K from 100 to 1000.

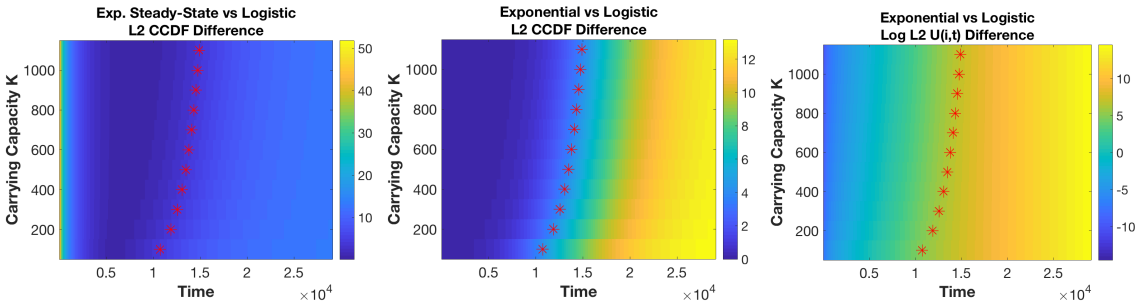

Figure 2.5: Heat maps of L2 difference in time from $t = 0$ to $t = 30000$ and where the carrying capacity K ranges from 100 to 1000 in the logistic model. Red stars indicate the first time the inflection point Equation (2.17) of Equation (2.16) is exceeded. Moreover, we see an increase of the rate of divergence in both models after this point time. *(Left)* L2 difference (CCDFs) of the exponential steady-state solution and the numerical logistic solution. *(Center)* L2 difference (CCDFs) of the numerical exponential solution and numerical logistic solution, Equation (2.19). *(Right)* L2 difference (Density) of the numerical exponential solution and numerical logistic solution, Equation (2.18).

Because the steady state distributions of both are power-law like, we compare the distributions $U_i(t)$ as well as the complementary cumulative distribution function. We consider the L2 difference in the number of full and partial length elements,

$$E_{\text{den.}}(t) = \left(\sum_{i=1}^{L} \left[U_{\text{exp.}}(i,t) - U_{\text{logistic}}(i,t) \right]^2 \right)^{1/2}, \tag{2.18}$$

as well as the differences in complementary cumulative distribution function (CCDF) between both models. First, we aggregate the counts for each partial and full-length TE for each model. We then convert this data to a complementary cumulative distribution, $C(i,t) = 1 - \left(\sum_{j=1}^{i} U_j(t) \right) / \sum_{j=1}^{L} U_j(t)$, (CCDF = 1 - CDF) and take the L2 difference,

$$E_{\text{dist.}}(t) = \left(\sum_{i=1}^{L} \left[C_{\text{exp.}}(i,t) - C_{\text{logistic}}(i,t) \right]^2 \right)^{1/2}. \tag{2.19}$$

Doing so accomplishes two goals: (1) We do not expect to see partial TEs of all possible lengths in real data and the CCDF smooths out this missing information, and (2) CCDFs provide a unified framework to compare TE families with one another. In our analysis, we compare the empirical TE CCDFs to the analytical CCDF derived in section 2.7.7.

We summarize the difference in the models in Figures 2.5, 2.6, and 2.7. We note that the numerical solutions to both the logistic and exponential model agree in both density and complementary distributions, but Equation (2.17) (denoted by red stars) marks the time at which the difference between the two models increases. In Figure 2.7, we observe similar TE length distributions in the models, but the exponential model results in continual addition of full and partial length TEs to the host genome. This results not only in an increasing difference in the number of total TEs, but also their respective length distributions (see Figure 2.6). Up until this time, both Figures 2.5 and 2.6 support relatively small $E_{\text{dist.}}(t)$ for a variety of carrying capacity K. In fact, we observe that as K increases, the time for divergence between the exponential and logistic model tends to grow logistically. We next consider the impact of exponential growth on the size of the host genome.

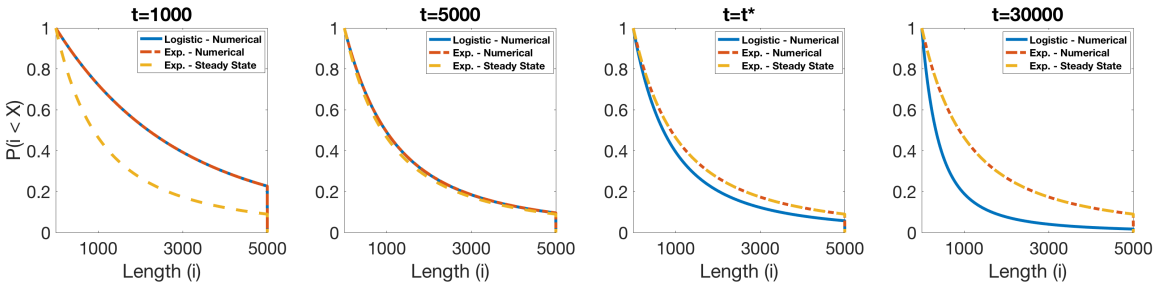

Figure 2.6: Comparison of 3 CCDFs (logistic, exponential, and exponential steady-state) at four distinct time points (from t= 1000 to t = 30000) for K =1000. We observe agreement between both the logistic and exponential model initially, but divergence increases rapidly after Equation (2.17) (t^* = 14903). Moreover, the numerical solution for the exponential model and the analytical steady state remain in agreement after this time.

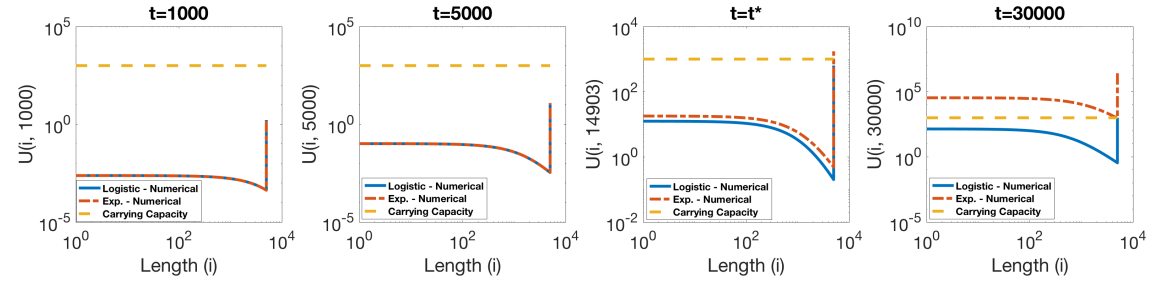

Figure 2.7: Comparison of two TE length distributions (logistic, exponential), along with the carrying capacity $K = 1000$, at four distinct time points (from $t = 1000$ to $t = 30000$). We observe agreement between both the logistic and exponential model initially, but divergence increases rapidly after Equation (2.17) ($t^* = 14903$).

Genome Size Dynamics. For both models, we compare how the length of the genome changes in time. As previously described, the genome size is affected by replications of

full-length TEs, insertions, and deletions. Thus, we have

$$\frac{dG}{dt} = LU_L(t)\ (\text{rate of full-length replication}) + IG(t)\sum_{k\in\mathbb{I}} k - DG(t)\sum_{f\in\mathbb{D}} f,$$

where $\mathbb{I}$ and $\mathbb{D}$ are the sets of allowable insertions and deletions, respectively. Since the rate of replication in the exponential model is constant, we may readily observe that $G(t) \to \infty$ as $t \to \infty$ unless the rate of growth and loss are equal. We acknowledge this is not biologically feasible, but rather shift our focus to the logistic model. In this case, however, we have

$$\frac{dG}{dt} = LU_L(t)r_d\left(1 - \frac{U_L(t)}{K}\right) + IG(t)\mathbb{E}[\text{insertion size}] - DG(t)\mathbb{E}[\text{deletion size}].$$

When $U_L(t)$ reaches the carrying capacity K, the host genome will only change based on genomic insertion and deletion rates. Since insertions are less prevalent than deletions, our general framework with the logistic model results in a bounded genome size in which TE impact on the genome length will disappear.

2.5.2 Parameter Estimation

Before applying our model to true TE annotations, we first investigated our ability to correctly recover parameters from simulations of TE evolution (see 2.8). We note that since the length L of a TE is a known property for each family of TE, we take this quantity as given. First, we attempted to independently estimate the parameters related to TE duplication and deletion (α, D, I) as well as the time t the TE entered the genome (t) using Equation (2.44), but found our least-squares formulation was ill-conditioned. In the logistic model, we had the added parameter of the carrying capacity K to consider and observed similar limitations. Hence, in this initial study, we focus on applying the exponential steady-state solution to TE data for three reasons: (1) since the logistic model is not explicitly solvable, we use the derived exponential solution presented in Theorem 1, (2) the logistic model, when not at carrying capacity, is consistent with the exponential steady-state, and (3) limited TE data does not support the hypothesis that TEs are close to reaching the carrying capacity in host genomes.

As such, we modified our parameter fitting approach in two ways: (1) we assume that the time since the first TE element was introduced is large enough that the logistic model may be described by the exponential steady-state solution, (2) we nondimensionalize (2.44) with

$$\theta = \frac{\alpha}{DF + \gamma L},$$

which is the ratio of replicative rate to mutation rate. Furthermore, we make the following simplifying assumption: $\mathbb{D} = \{1, 2, 3, 4, 5, L - 15\}$, which reflects the dominance of short deletion lengths in genomes but also allows for a TE to be almost deleted [47]. Thus, $\sum_{f\in\mathbb{D}} = F = L$ and $u(x,t)$ is of the form Equation (2.31). With these assumptions, $\gamma = 6D + I$ and we consider the case where $I \approx D$ (at most), as insertion-deletion

(indel) rates are often calculated together [55]. Moreover, there is support (particularly in *Drosophila*) that deletions outnumber insertion events [57, 20].

We note, as shown in Figure 2.8, our dimensionless parameter θ along with the TE length L controls the convergence of the time-varying solution to the steady-state solution. Because these tend to be longer TEs, such as the LINEs in the human genome, we therefore fit the data with our continuous model, Equation (2.44). We note this approach is complementary to traditional models of TE evolution which focus primarily on actively replicating TEs with many full-length copies in a genome.

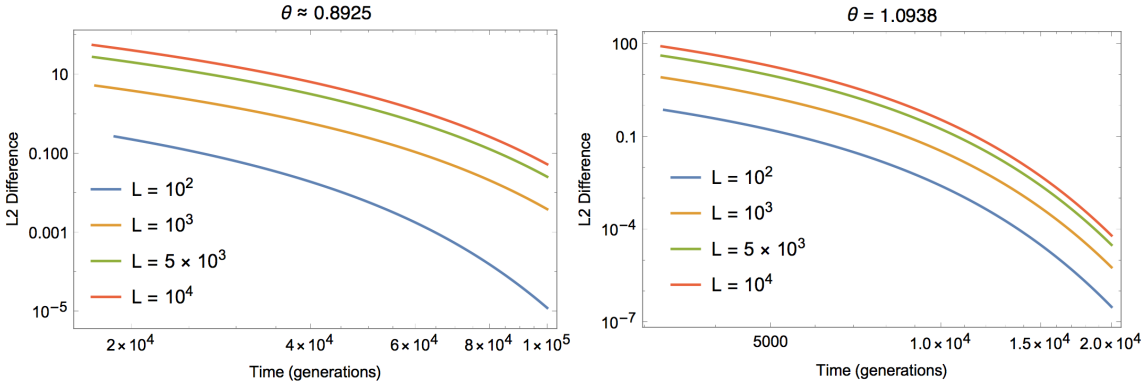

Figure 2.8: Log-Log plot of L2 difference between steady-state solution and TE length distribution Equation (2.31) in time with $\theta = \frac{\alpha}{DF+\gamma L} = 0.8925$ (left) and $\theta = 1.0938$ (right) for different TE lengths L. As described in the main text, we use the steady-state solution for all subsequent parameter estimation. Thus, we focus on TE distributions at or close to steady-state. We note that as θ increases, the number of generations until the difference between solutions approach zero decreases. Moreover, we observe a similar pattern as L decreases.

Before analyzing TE distributions from real genomes, we aggregate the counts for each partial and full-length TE for each TE family. We then convert this data to a complementary cumulative distribution (CCDF = 1 - CDF). For each species and each TE family considered, we determine the parameter θ by minimizing the sum of the residuals squared (RSS) between the data and the explicit solution using the Levenberg-Marquardt algorithm. In order to make comparisons between TE families in closely related species, we estimate θ for each species (for each TE) and compare the distribution of predicted replication/mutation estimates.

2.5.3 Drosophila roo elements

There are TE annotations available for the 12 genomes from the *Drosophila* (fruit-fly) clade. Previous studies have quantified the relationship of these twelve species, but have looked at only the distribution of total transposable elements across species [11]. Instead, we focus our analysis on the *roo* and *rooA* elements because of the abundance of partial-length copies of these elements in *Drosophila* genomes and since their evolutionary history has been previously studied [10, 32].

Of the *Drosophila* clade, *melanogaster* is reported to have at least 50 functioning copies of the transposable element *roo*. The remaining 11 species each have fewer than 5 active copies of the TE [10]. We use lengths of *roo* and *rooA* and calculate their empirical complementary cumulative distribution for each of the 12 *Drosophila* species. As described in the previous section, we process both the partial and full-length elements for both the *roo* and *rooA* elements for all 12 species. (Counts of full and partial TE copies are reported in Table 2.1.) We fit θ independently for each of the 24 cases (see Figures 2.14 and 2.15 for all fitted distributions) In addition, in Figure 2.9 we plot θ_{roo} for the average distributions of the 12 species.

Organism	*roo* **full/90%**	*roo* **partial >200bp**	*rooA* **full/90%**	*rooA* **partial >200bp**
D. simulans	-	28	-	171
D. sechellia	1/5	76	-/1	176
D. melanogaster	72	171/637	-/2	67
D. yakuba	-/1	138	-/5	271
D. erecta	1/1	119	-/18	324
D. ananassae	-	342	-	447
D. pseudoobscura	-	183	-	44
D. persimilis	-	486	-	303
D. willistoni	-	646	-	74
D. mojavensis	-	89	-	145
D. grimshawi	-	51	-	27
D. virilis	-	24	-	3

Table 2.1: For each genome in the twelve *Drosophila* species, we report the number of full and partial TE copies for *roo* and *rooA* elements. Column 2 and 4 report full-length elements (roo - 9092bp, rooA- 7621bp) along with elements that match 90% of the TE. Column 1 and 3 report all partial TE copies with lengths 200bp or longer, respectively.

Replication/Mutation Analysis. In order to interpret the relationship between *Drosophila* TE θ rates, we independently estimate θ, the replication/mutation parameter for each species. After doing so, we summarize these estimates in the violin plot shown in Figure 2.10. For the *roo* element, we observe an outlier from *D. melanogaster* affecting the distribution of θ estimates. In this case, the *D. melanogaster* roo element indicates TE replication is faster than mutation when compared to other members of the *Drosophila* clade. This is consistent with previous studies regarding it still being an relatively active transposon in this species [48, 32]. For *rooA*, we observe a longer tail as θ increases. This corresponds to higher estimates for *D. erecta* and *D. yakuba*. Since these two species are closely related, this phenomenon may suggest a change in replication and/or mutation dynamics in their recent evolutionary history. Together, this provides further support for recent transpositional activity of *rooA* and *roo* supported by Chaux et al. [10].

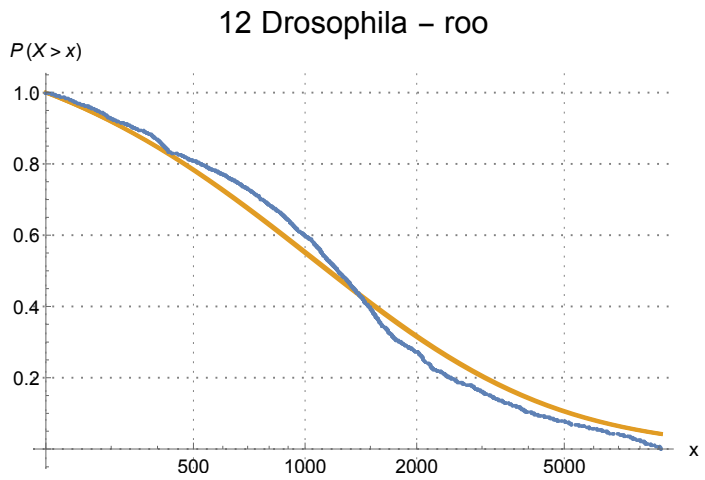

Figure 2.9: We aggregated counts for all 12 *Drosophila* species and we report the complementary cumulative distribution (1-CDF) as the data of interest (blue). We compare the empirical distribution to the exponential steady-state analytical fit (orange) with a p value of 2.815×10^{-14} from the Kolmogorov-Smirnov test, as described in 2.7.7. This Log plot reveals the low probability of observing full-length elements across all species considered, consistent with the transient assumption of our logistic model.

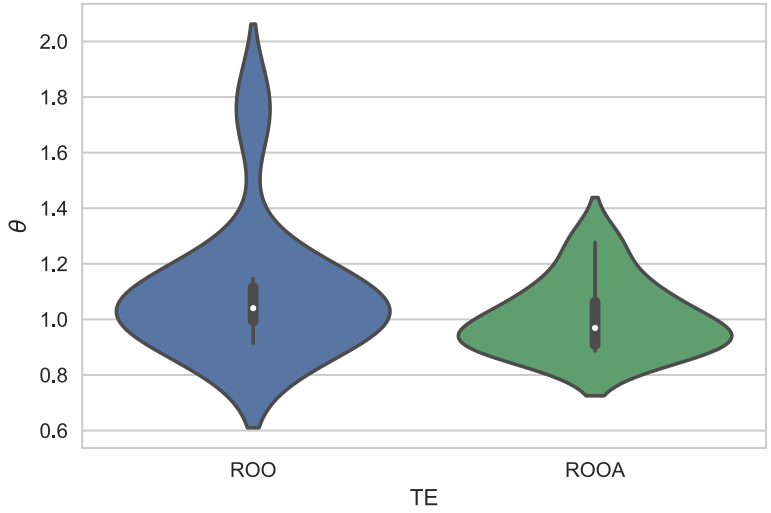

Figure 2.10: Violin plot comparing the distribution of θ (replication/mutation) estimates for the *roo* and *rooA* transposable elements for 10 drosophila species (*D. grimshawi* and *D. virilus* are excluded due to lack of data). We note that the outlier for *roo* corresponds to the high abundance in *D. melanogaster* and the tail in the distribution for *rooA* corresponds to *D. yakuba* and *D. erecta* θ estimates.

2.5.4 Avian retrotransposons

Retrotransposons, including long interspersed elements (LINEs) and long-terminal repeat (LTR) retrotransposons, in birds provide a unique perspective into the role of TEs in genomic evolution as bird genomes are less TE rich than other species. Recent studies suggest this may be due to the instability of avian genomes in TE-rich regions; as such, we explore retrotransposon transposition rates for the most common TEs present [18, 6]. We primarily focus on lineages of Chicken repeat 1 (CR1) retrotransposons, which are the most abundant superfamily in nearly all amniotes and have been used for previous phylogenic analyses [53]. Due to past transpositional activity and their highly conserved structure, we compare CR1 subfamily rates with other TEs from different superfamilies [53]. In particular, we explore the most common TE superfamily, chicken repeat 1 (CR1), in 23 bird species. Since CR1 TEs came from one common ancestor sequence, we compare parameter estimates with elements from the same superfamily [27, 28].

Replication/Mutation Analysis. Figure 2.11 illustrates the distribution of θ estimates for 13 different TEs (12 CR1 and 1 PSLINE) present in all 23 bird species. In particular, we note that estimates for CR1-D reveal higher replication to mutation rates than in other elements. CR1-D is often used as the reference for CR1 [60]. We also notice potentially two groups in CR1-C4, with birds such as the bald eagle, turkey vulture and Dalmatian pelican having higher estimates than others such as the bar-tailed trogon, barn owl, and Downy woodpecker. To further consider the evolutionary history of CR1-C4 in Aves, gathering more samples for these groups of birds may be worth exploring in future studies. With the exception of the PSLINE and subsequent TEs in Figure 2.11, the remaining θ estimates for the majority of the CR1 subfamilies result in similar values across all 23 bird species.

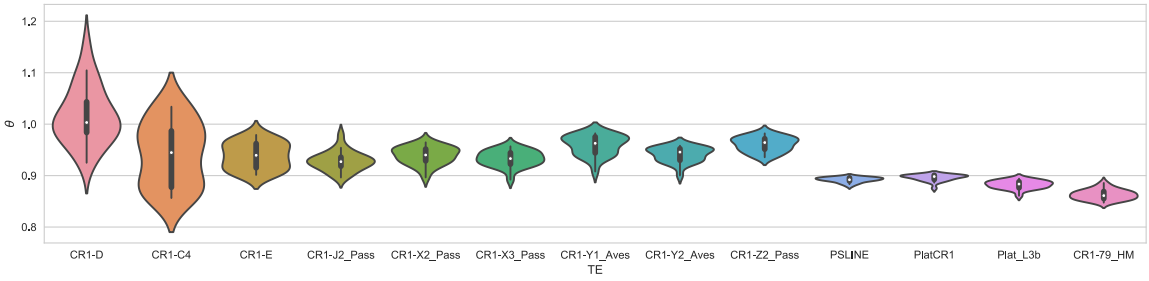

Figure 2.11: Violin plot comparing the distribution of θ (replication/mutation) estimates for a range of Chicken repeat (CR1) transposable elements as well as other LINEs (i.e., PSLINE). We observe that half of these estimates are approximately the same but note more variation for CR1-D, CR1-C4, CR1-E. Additionally, θ estimates for PSLINE and subsequent TEs are less than the majority of CR1 elements considered.

2.5.5 Primate LINE Elements

In contrast to birds, a large proportion of TEs comprise the majority of primate genomes [17, 16]. Since TEs play a role in shaping primate evolution, as evidenced by the primate-exclusive ALU element, we investigate the role of these mobile elements in a sample of humans and primates [21]. Most studies focus on actively transposing elements in these species. In humans, it is estimated that TEs that are currently transposing (L1 elements and ALUs) are doing so one per 100 generations [17]). These elements explain the large number of repetitive *k-mers* in genomes and studies using a transposon model, a master gene model, or a mixture provide insight into how TEs have shaped genomic statistical properties [49, 23]. However, since our work assumes a transposon model, and we focus our analysis on the transposition history of inactive primate TEs.

Specifically, we investigate nontransposing long interspersed nuclear elements, *LINEs* (e.g. L2 and L3) in these species, since L2 elements have stopped transposing for some time [50]. Although these elements are no longer replicating in the genomes of these organisms, studies have revealed that they are a driver of human genome length expansion. LINEs and other transposons also resulted in tandem repeats throughout genomes, and we aim to quantify past replicative rates in these replicating elements since they have evolutionarily impacted the evolution of primates [42, 1]. For non-reference samples, we run RepeatMasker independently with the `-q` option.

Replication/Mutation Analysis. Since we only consider 5 species (i.e., Chimp, Gibbon, Gorilla, Human, and Orangutan) of four LINEs, we plot the point estimates of θ in Figure 2.12. We note that these estimates are of the same order as previous estimates in birds and fruit flies. However, the scale of difference between all five primates' rates suggests similar transpositional dynamics and finer biological meaning is difficult to extract from such similar groups. Nevertheless, we see a separation of θ estimates for HAL1 elements in comparison to other LINEs. The HAL1, or "half-L1" elements may have independently originated from L1 elements several times [4]. Recent L1 transpositional activity thus may explain the higher θ estimate for HAL1 elements in comparison to older (non-replicating) LINEs.

2.6 Discussion and Conclusions

This paper presents a novel deterministic mathematical model and solutions for the evolution of TE dynamics. We consider a total of four models, discrete and continuous, with different assumptions about full-length TE replication dynamics. In the logistic case, we derive implicit analytical solutions for both discrete and continuous formulations. For exponential models, we derive explicit analytical forms of the time-varying TE density as well as the steady-state behavior. We compare both exponential and logistic model in the discrete case and observe agreement in number density and distribution until the difference in instantaneous growth increases. Moreover, we discuss the implications of each model on the host genome size. The logistic model guarantees a bounded genome size and provides a generalized model for future exploration of limited TE

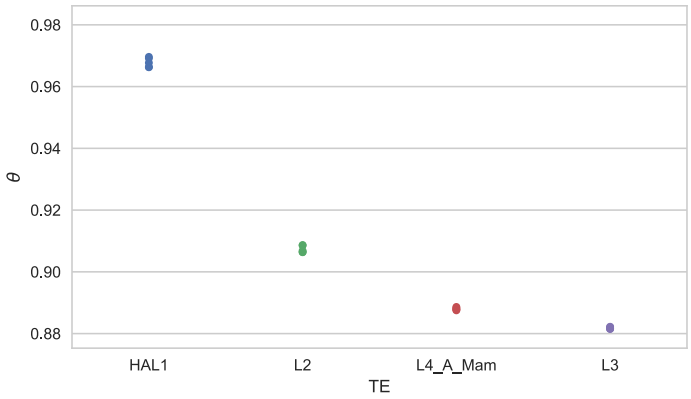

Figure 2.12: Plot of θ estimates for four LINEs in Chimp, Gibbon, Gorilla, Human, and Orangutan. We observe little fluctuation in estimates between species. However, we attribute differences in estimates in HAL1 to its origin from L1 elements as opposed to older LINEs (L2, L3, L4) in primates.

growth. Rather than focus on actively duplicating elements, our general framework allows for replication/mutation rate estimation using the full spectrum of available TE data. In particular, our solution $u(x,t)$ derived in 2.7.7 describes the change in number of non-replicating partial elements. Moreover, our model provides a means to quantify differences in replication to mutation rates of separate TEs in a variety of species. These differences provide an orthogonal view of the these evolutionary rates in these species and our non-species-specific model is extendable to different TEs for a variety of species.

Our fragmentation model has several limitations. We assume both constant replication and genome deletion rates in the proposed transposon model when transpositional bursts periods could be considered [19]. Moreover, our model assumes no spatial dependence, which may be an inaccurate representation if TEs are densely distributed. We intend to explore time-varying deletion rates as well as more accurate mutational processes in future work. While our model is capable of capturing time-varying dynamics, we found we lacked the power to independently estimate the time-varying parameters from data. In future studies, we will focus on incorporating the logistic model with real TE data as well as consider competing TEs in which each TE may have a different carrying capacity.

To compare to TE annotations, we require that the TE data must also support a quasi-steady-state distribution. We utilize RepeatMasker data for our analysis, but we also acknowledge that this procedure may generate wrongly-fragmented copies resulting in imperfect TE annotations. Finally, we note the genomic data used, annotated reference genomes, represent only one realization of TE dynamics for a species. As such, we are fitting a model to a single stochastic realization. We also note that genomes from different species are not independent realizations and future studies will incorporate a stochastic treatment instead of our deterministic framework. Although, our model represents the expected outcome of a stochastically varying process, because the logistic model is far from a steady-state distribution we should be able to effectively estimate parameters from a species reference genome. Nevertheless, organisms harbor a multitude of inactive

elements consistent with our model.

We use our model to estimate the replication to mutation rate θ in *Drosophila, Aves,* and *Primates*. In *Drosophila*, θ rates estimated for *roo* and *rooA* suggest a more nuanced dependency between these elements in closely related *D. yakuba* and D. erecta. Differences in rates across different groups of birds in the CR1 TE subfamily warrants further investigation of replication and mutation rates for the abundant CR1 transposable element [54]. Primate LINEs, although the most complex, do not vary substantially, but the origins of each of the LINE subfamilies may account for minute differences in TE parameter estimates. Heterogeneity of replication/mutation rates across multiple species and families may suggest a complex evolutionary history of TE proliferation.

Our deterministic model allows for quick replication/mutation rate comparison to explore the relationships of shared elements between species. Results indicate agreement between steady-state distributions and the proposed fragmentation equation framework. This works demonstrates that TE annotations, consisting of full-length and partial copies, allow us to quantify TE activity and promote exploratory analyses of less studied TEs in nonmodel organisms. We hope that further exploration into specific partial length repeat mutations and divergence from actively replicating elements will reveal a more complete evolutionary history of these abundant elements.

2.7 Appendix A: Mathematical Model for TE Dynamics

In Appendix A we provide mathematical derivations for lemmas and theorems in Section 2.4 the main text. In particular, we derive the moment closure for both discrete and continuous models of TE dynamics, prove existence and uniqueness of the continuous model of TE dynamics, and explicitly derive the time-varying TE length distribution for our continuous formulations. Lastly, we determine the form of the complementary cumulative distribution function (CCDF), which we use in all comparisons to data (Section 3.3.1).

2.7.1 Solving the Moment Closure System for Exponential Growth Discrete and Continuous Systems

Discrete System

As described in the text, the moment closure for the discrete model is given by Equations (2.5) and (2.6):

$$\begin{aligned} \frac{d\eta}{dt} &= (-DF-\gamma)\eta(t) + \gamma\xi(t) + 2\gamma(L-1)U_L(t), \\ \frac{d\xi}{dt} &= -DF\xi(t) + \gamma L(L-1)U_L(t), \end{aligned}$$

where $U_L(t)$ is the number of full length TEs in the discrete model and governed by Equation (2.3). Thus,

$$U_L(t) = U_L(0)e^{(\alpha - DF - \gamma(L-1))t}.$$

The first order system for the moments η and ξ can be explicitly solved through the use of integrating factors:

$$\begin{aligned}\eta(t) &= K_1 U_L(t) + C_1 e^{-DFt} + C_2 e^{(-DF-\gamma)t} \\ \xi(t) &= \frac{\gamma(L^2-L)}{\alpha-\gamma(L-1)} U_L(t) + C_1 e^{-DFt}.\end{aligned}$$

The constants C_1, C_2 and K_1 are determined by the initial conditions. In the case we are most interested in, we have only full length TEs at time $t = 0$, $U_L(0) = U_0$ and $\eta(0) = \xi(0) = 0$, and as such,

$$\eta(t) = K_1 U_L(t) + C_1 e^{-DFt} + C_2 e^{(-DF-\gamma)t} \tag{2.20}$$

$$\xi(t) = C_1 \left(U_0 e^{-DFt} - U_L(t)\right), \tag{2.21}$$

where $K_1 = \frac{\gamma(2\alpha-(L-2)\gamma)(L-1)}{(\alpha-\gamma(L-2))(\alpha-\gamma(L-1))}$, $C_1 = -\frac{\gamma(L^2-L)}{\alpha-\gamma(L-1)}$, and $C_2 = (-K_1 U_0 - C_1)$.

Continuous System

For the continuous system, Equations (2.12) and (2.13), we obtain the following system of two differential equations for the zeroth and first moments

$$\begin{aligned}u_0'(t) &= -DFu_0(t) + \gamma u_1(t) + 2\gamma L u_L(t) \\ u_1'(t) &= -DFu_1(t) + \gamma L^2 u_L(t).\end{aligned}$$

By Lemma 3, and because in our formulation $u_L(t)$ is not dependent on u_0 or u_1, Equations (2.12) and (2.13) describe how the zeroth and first moment change in time. We obtain the equation for $u_0(t)$ by integrating Equation (2.9) from 0 to L; the equation for $u_1(t)$ is obtained by multiplying Equation (2.9) by x and integrating. Because $u_L(t) = u_L(0)e^{(\alpha-DF-\gamma L)t}$, using e^{DFt} as an integrating factor allows us to solve (2.12) and (2.13):

$$u_0(t) = \frac{\gamma L(2\alpha-\gamma L)}{(\alpha-\gamma L)^2} u_L(t) + \gamma C_1 t e^{-DFt} + C_2 e^{-DFt} \tag{2.22}$$

$$u_1(t) = \frac{\gamma L^2}{\alpha-\gamma L} u_L(t) + C_1 e^{-DFt}. \tag{2.23}$$

As in the discrete case, applying the initial condition of $u(x,0) = 0$, (i.e., $u_0(0) = u_1(0) = 0$) allows us to determine the constants C_1 and C_2. In particular, we have $C_1 = -\frac{\gamma L^2 u_L(0)}{\alpha-\gamma L}$ and $C_2 = -\frac{\gamma L(2\alpha-\gamma L)u_L(0)}{(\alpha-\gamma L)^2}$.

2.7.2 Solution to the Exponential Growth Discrete System

Here we prove the existence and uniqueness of the time-varying solution to the discrete exponential model for TE length distribution. Using the zeroth moment $\eta(t)$ in Equation (2.20), we rewrite Equation (2.2) as

$$\frac{dU_i}{dt} = \left(-DF - \gamma(i+1)\right) U_i(t) - 2\gamma \sum_{k=1}^{i-1} U_k(t) + 2\gamma\eta(t) + 2\gamma U_L(t)$$

Thus, we rewrite our discrete system as

$$\vec{U}' = A\vec{U} + \vec{g}(t),$$

for $i = 1, 2, \dots, L-1$ and where A and $\vec{g}(t)$ are defined as

$$A = \begin{bmatrix} -DF - 2\gamma & & & & & \\ -2\gamma & -DF - 3\gamma & & & & \\ \vdots & -2\gamma & \ddots & & \mathbf{0} & \\ \vdots & \vdots & \ddots & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & -DF - (L-1)\gamma & \\ -2\gamma & \vdots & \dots & \dots & -2D & -DF - \gamma L \end{bmatrix}, \tag{2.24}$$

$$\vec{g(t)} = [2\gamma\eta(t) + 2\gamma U_L(t)]\,\mathbf{1},$$

where $\mathbf{1} \in \mathbb{R}^{(L-1)}$ is a vector of ones. By *Duhamel's principle*, the solution $\vec{U}(t)$ can be expressed as

$$\vec{U}(t) = e^{At}\vec{U}(0) + \int_0^t e^{A(t-\tau)}\vec{g}(\tau)\, d\tau,$$

and is a unique solution to the inhomogeneous constant-coefficient linear system [39]. Since A is a lower triangular matrix, then the diagonal elements are the eigenvalues of A, namely

$$\lambda_i = -DF - (i+1)\gamma \qquad \text{for } i = 1, 2, \dots, L-1.$$

Furthermore, A is diagonalizable since we have distinct $\lambda_i \neq 0$ and $L-1$ linearly independent eigenvectors and further, we apply our initial condition $U(0) = \mathbf{0}$. Then, $A = S\Lambda S^{-1}$ and our solution $\vec{U}$ simplifies to

$$\vec{U} = \int_0^t Se^{\Lambda t}S^{-1}\vec{g}(\tau)\, d\tau \tag{2.25}$$

We explicitly solve for the matrix of eigenvectors $S \in \mathbb{R}^{(L-1)\times(L-1)}$ and its inverse S^{-1} as

$$S = \begin{bmatrix} 1 & & & & & \\ -2 & 1 & & & \mathbf{0} & \\ 1 & -2 & \ddots & & & \\ & 1 & \ddots & \ddots & & \\ \mathbf{0} & & \ddots & \ddots & 1 & \\ & & & 1 & -2 & 1 \end{bmatrix}, \quad S^{-1} = \begin{bmatrix} 1 & & & & & \\ 2 & 1 & & & \mathbf{0} & \\ 3 & 2 & \ddots & & & \\ \vdots & 3 & \ddots & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & 1 & \\ L-1 & L-2 & \dots & 3 & 2 & 1 \end{bmatrix} \tag{2.26}$$

Let $\vec{x}_{\text{exp.}}(t) = \int_0^t g(\tau)e^{\Lambda(t-\tau)}S^{-1}\mathbf{1}\, d\tau$. Then, let $x_i(t)$ represent the ith entry in $\vec{x}_{\text{exp.}}(t)$. Thus,

$$x_i(t) = \sum_{k=1}^{i} k \int_0^t g(\tau)e^{\lambda_i(t-\tau)}\, d\tau = \frac{i(i+1)}{2}\int_0^t 2\gamma[\eta(\tau) + U_L(\tau)]e^{\lambda_i(t-\tau)}\, d\tau.$$

Applying the solution to the zeroth moment, Equation (2.20), we have

$$x_i(t) = \gamma(i^2+i)\left[K_2(U_L(t) - e^{\lambda_i t}) + \frac{C_1}{(1+i)\gamma}\left(e^{-DFt} - e^{\lambda_i t}\right) + \frac{C_2}{i\gamma}\left(e^{(-DF-\gamma)t} - e^{\lambda_i t}\right)\right],$$

where $K_2 = \frac{K_1+1}{\alpha-\gamma(L-2-i)}$, and C_1, C_2 are defined in 2.7.1. Thus, we arrive at the unique solution to Equation (2.2)

$$\vec{U}(t) = S\vec{x}_{\text{exp.}}(t).$$

2.7.3 Solution to the Exponential Growth Continuous System

Here we prove the existence and uniqueness of the time-varying solution to continuous model for TE length distribution. More specifically, we prove the following lemma and then the establish existence and uniqueness in the subsequent theorem.

Lemma 4. *For the continuous TE length distribution model, Equations* (2.9) *and* (2.10)*, the non-unique solution to the Laplace transformed Equation* (2.9) *is given by*

$$V(x,s) = \frac{J(s)}{(DF+s+\gamma x)^3}, \tag{2.27}$$

where $\gamma = cD + I, F = \sum_{f\in\mathbb{D}}$, and $J(s)$ is an undetermined function of s.

Proof. We begin by taking the Laplace transform of Equation (2.9) to obtain

$$2D\int_x^L V(y,s)\,dy - D(L+x)V(x,s) + \frac{2\gamma u_L(0)}{s-\alpha+DF+\gamma L} = sV(x,s) - u(x,0),$$

where $\frac{2\gamma u_L(0))}{s-\alpha+DF+\gamma L}$ is the Laplace tranform of the solution to Equation (2.10). We then apply $\partial_x[\cdot]$ to the transformed equation, with $u(x,0) = 0$, and we have

$$-3\gamma V = (s+DF+\gamma x)\frac{\partial V}{\partial x}. \tag{2.28}$$

□

The solution to Equation (2.28) follows a power-law, namely,

$$V_{\text{gen}}(x,s) = \frac{J(s)}{(DF+s+\gamma x)^3}, \tag{2.29}$$

and is only unique up to the function $J(s)$. With Lemma 4 proven, we are now able to prove Theorem 1 from the text. (We also re-state the theorem below.)

Theorem 1. *If $u(x,0) = 0$ and $\alpha > \gamma L$, there exists a solution $u(x,t)$ to the continuous TE length distribution model, Equations* (2.9) *and* (2.10)*, given by*

$$\begin{aligned} u(x,t) &= \mathscr{L}^{-1}\{V(x,s)\}, \\ u_L(t) &= \exp[(\alpha - DF - \gamma L)t], \end{aligned} \tag{2.30}$$

where $\mathscr{L}^{-1}$ is the inverse Laplace transform and $V(x,s) = J(s)(DF + s + \gamma x)^{-3}$ and $J(s)$ is defined in Equation (2.32)*. Moreover, when $F = L$, the solutions $u(x,t)$ and $u_L(t)$ are given by:*

$$\begin{aligned} u(x,t) &= \frac{-\gamma\left(e^{\nu}\left(2\alpha^2 + \omega^4 t^2 + 4\alpha^2\omega t - 2\omega^3 t(\alpha t - 1) + \alpha\omega^2 t(\alpha t - 6)\right)\right)}{(\alpha-\omega)^3} \\ &\quad + \frac{-2\gamma\alpha^2 e^{t(\alpha - L(\gamma + D)}}{(\alpha-\omega)^3} \\ u_L(t) &= \exp[(\alpha - DL - \gamma L)t], \end{aligned} \tag{2.31}$$

where $\gamma = cD + I$, $\omega = \gamma(L - x)$ and $\nu = -t(DL + \gamma x)$. Moreover, this solution is unique.

Proof. By the previous lemma, the zeroth moment of the solution to Equation (2.28) is given by

$$\int_0^L V(x,s) = \int_0^L \frac{J(s)}{(DF + s + \gamma x)^3} = J(s)\left(\frac{1}{2\gamma(DF + s)^2} - \frac{1}{2\gamma(DF + s + \gamma L)^2}\right). \tag{2.32}$$

Next, we apply the Laplace transform to the zeroth moment solution of Equation (2.12). Equating both sides, we determine $J(s)$ and thus, the unique solution to Equation (2.28)

$$V(x,s) = \frac{J(s)}{(DF + s + \gamma x)^3}, \tag{2.33}$$

where

$$J(s) = J_1\left(\frac{-2\alpha + \gamma L}{DL + s} + \frac{\gamma L(\gamma L - \alpha)}{(DL + s)^2} + \frac{2\alpha - \gamma L}{DF + s - \alpha + \gamma L}\right),$$

where

$$J_1 = \frac{2\gamma u_L(0)(DF + s)^2(DF + s + \gamma L)^2}{(\alpha - \gamma L)^2(2DF + 2s + \gamma L)}.$$

We determine Equation (2.30) by taking the inverse Laplace transform. In particular, if $\sum_{f\in\mathbb{D}} f = F = L$, the inverse Laplace transform is given by Equation (2.31).

We prove uniqueness under a similar framework as in [38] by assuming we have two solutions to Equation (2.9): $u(x,t)$ and $v(x,t)$ where $u(x,0) = v(x,0)$. We define $g(x,t) = u(x,t) - v(x,t)$ and will show that $g(x,t) = 0$. First note that $g(x,t)$ must satisfy the following partial differential equation

$$\frac{\partial g}{\partial t} = (-DF - \gamma x)g(x,t) + 2\gamma\int_x^L g(y,t)\,dy. \tag{2.34}$$

We begin by forming a series of Volterra sequences satisfying the equation above. We will show that these sequences will ultimately be the solution to Equation (2.34). Let $g_0(x,t)$ be any arbitrary continuous function on $(0,L]\times[0,T]$ and let $M = \max_{(0,L]\times[0,T]} |g_0(x,t)|$. We then define $g_n(x,t)$ for $n \geq 1$ as follows

$$\frac{\partial g_n}{\partial t} = (-DF - \gamma x)g_n + 2\gamma\int_x^L g_{n-1}(y,t)\,dy, \tag{2.35}$$

where $g_n(x,0) = 0$. We can solve for $g_1(x,t)$ using an integrating factor and note that

$$\begin{aligned} g_1(x,t) &= e^{(-DF-\gamma x)t}\left[C_1(x) + 2\gamma \int_0^t e^{(-DF-\gamma x)\tau} \int_x^L g_0(y,\tau)\, dy\, d\tau\right] \\ &= e^{(-DF-\gamma x)t}\left[C_1(x) + 2\gamma \int_0^t \int_x^L e^{(-DF-\gamma x)\tau} g_0(y,\tau)\, dy\, d\tau\right]. \end{aligned}$$

Applying our initial condition, $g(x,0) = 0$, and because we know that $C_1(x) = 0$, we have

$$g_1(x,t) = 2\gamma e^{(-DF-\gamma x)t} \int_0^t \int_x^L e^{(-DF-\gamma x)\tau} g_0(y,\tau)\, dy\, d\tau.$$

In our bounded domain, $e^{(-DF-\gamma x)\tau} \le e^{(-DF-\gamma L)T}$ for all $x \in (0,L]$ and $t \in [0,T]$. Then

$$\begin{aligned} |g_1(x,t)| &\le 2\gamma \left|e^{(-DF-\gamma x)t}\right| \int_0^t \int_x^L e^{(-DF-\gamma L)T} \left| g_0(y,\tau)\right| dy\, d\tau \\ &\le 2\gamma \int_0^t \int_x^L e^{(-DF-\gamma L)T} |g_0(y,\tau)|\, dy\, d\tau \\ &\le M(2\gamma e^{(-DF-\gamma L)T})(L-x)t. \end{aligned} \tag{2.36}$$

We proceed to solve for $g_2(x,t)$ similarly,

$$\begin{aligned} \frac{\partial}{\partial t}\left(g_2(x,t)e^{(-DF-\gamma x)t}\right) &= 2\gamma e^{(-DF-\gamma x)t} \int_x^L g_1(y,\tau)\, dy \\ g_2(x,t) &= e^{(-DF-\gamma x)t}\left[C_2(x) + 2\gamma \int_0^t \int_x^L e^{(-DF-\gamma x)\tau} g_1(y,\tau)\, dy\, d\tau\right]. \end{aligned}$$

Applying the initial condition $g_2(x,0) = 0$, we have

$$g_2(x,t) = 2\gamma e^{(-DF-\gamma x)t} \int_0^t \int_x^L e^{(-DF-\gamma x)\tau} g_1(y,\tau)\, dy\, d\tau.$$

Then,

$$\begin{aligned} |g_2(x,t)| &\le 2\gamma \int_0^t \int_x^L e^{(-DF-\gamma L)T} |g_1(y,\tau)|\, dy\, d\tau \\ &\le 2\gamma e^{(-DF-\gamma L)T} \int_0^t \int_x^L 2\gamma e^{(-DF-\gamma L)T} M(L-y)\tau\, dy\, d\tau \\ &= M\left(2\gamma e^{(-DF-\gamma L)T}\right)^2 \left(\frac{((L-x)t)^2}{(2!)^2}\right). \end{aligned}$$

By induction, we note that

$$\begin{aligned} |g_n(x,t)| &\le M\left(2\gamma e^{(-DF-\gamma L)T}\right)^n \left(\frac{((L-x)t)^n}{(n!)^2}\right) \\ &\le M\frac{\left(2\gamma LT e^{(-DF-\gamma L)T}\right)^n}{(n!)^2}. \end{aligned}$$

Thus, $g_n(x,t)$ will converge to 0 as $n \to \infty$ as long as M and $2\gamma LT e^{(-DF-\gamma L)T}$ are finite.

We now wish to show $g(x,t) = 0$ is the unique solution to $\frac{\partial g}{\partial t} = (-DF - \gamma x)g + 2\gamma \int_x^L g(y,t)\, dy$, where $g(x,0) = 0$. It is straightforward to verify $g(x,t) = 0$ is a solution to the IVP. Let $\psi(x,t)$ also be a solution to the IVP and let $g_0(x,t) = \psi(x,t)$. Then,

$$\begin{aligned} \frac{\partial g_1}{\partial t} &= (-DF - \gamma x)g_1 + 2\gamma \int_x^L \psi(y,t)\, dy \\ \frac{\partial g_1}{\partial t} + (DF + \gamma x)g_1 &= \frac{\partial \psi}{\partial t} + (DF + \gamma x)\psi. \end{aligned}$$

Since $g_1(x,0) = \psi(x,0) = 0$, we can readily show that the only solution to the initial value problem $\frac{\partial g}{\partial t} + (DF + \gamma x)g = 0$, where $g(x,0) = 0$ is $g(x,t) = 0$. Thus, $g_1(x,t) = \psi(x,t)$. Furthermore, we have $g_n = \psi$ for all n. Since $\lim_{n\to\infty} g_n(x,t) = 0$ for all $x \in (0,L]$ and $t \in [0,T]$, we conclude $\psi = 0$. Thus, since $g = 0$ is the only solution to the homogeneous balance equation, the zeroth and first moments agree are zero for all time. Furthermore, we conclude

$$g(x,t) = 0 \Rightarrow u(x,t) = v(x,t),$$

proving uniqueness for (2.9). □

2.7.4 Solving the Moment Closure System for Logistic Growth Discrete and Continuous Systems

We next consider the moment closure for both discrete and continuous models under the assumption of logistic growth in the number of full length TEs.

Discrete System

As described in the text, the moment closure for the discrete model is still given by Equations (2.5) and (2.6), where $U_L(t)$ is the number of full length TEs in the discrete model and governed by Equation (2.4), with solution

$$U_L(t) = \frac{KU_L(0)e^{r_d t}}{K + U_L(0)[e^{r_d t} - 1]},$$

where $r_d = \alpha - DF - \gamma(L-1)$. We note that ξ may be explicitly solved through the use of integrating factors:

$$\xi(t) = A_1 U_L(t)\ {}_2F_1\left(1, 1 + \tfrac{DF}{r_d}, 2 + \tfrac{DF}{r_d}, -(K-1)^{-1}\exp[r_d t]\right) + C_1 e^{-DFt}, \tag{2.37}$$

where $A_1 = \frac{\gamma KL(L-1)}{(\alpha - \gamma[L-1])(K-1)}$ and ${}_2F_1(a,b;c;z)$ is the hypergeometric function defined as

$${}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k z^k}{(c)_k k!}\ ; |z| < 1.$$

The constant C_1 is determined by the initial conditions. In the case we are most interested in, we have only full length TEs at time $t = 0$, $U_L(0) = U_0$ and $\eta(0) = 0$, and as such,

$$C_1 = -A_1 U_L(0)\,{}_2F_1\left(1, 1 + \tfrac{DF}{r_d}, 2 + \tfrac{DF}{r_d}, -(K-1)^{-1}\right).$$

For the zeroth moment, $\eta(t)$, the solution may be represented as the following integral:

$$\eta(t) = e^{(-DF-\gamma)t}\left(\gamma \int \xi(t)\ dt + 2\gamma(L-1)\int U_L(t)\ dt\right), \tag{2.38}$$

where $\xi(t)$ and $U_L(t)$ are defined above.

Continuous System

For the continuous system, the moment-closed system $[u_0(t), u_1(t)]$, is described by Equations (2.12) and (2.13). We now consider the case where

$$u_L(t) = \frac{K u_L(0) e^{r_c} t}{K + u_L(0)[e^{r_c t} - 1]},$$

where $r_c = \alpha - DF - \gamma L$. Similar to the discrete case, we use e^{-DFt} as an integrating factor to obtain

$$u_1(t) = A_1 u_L(t)\ {}_2F_1\left(1, 1 + \tfrac{DF}{r_c}, 2 + \tfrac{DF}{r_c}, -(K-1)^{-1}\exp[r_d t]\right) + C_1 e^{-DFt}, \tag{2.39}$$

where $A_1 = \frac{\gamma K L^2}{(\alpha - \gamma L)(K-1)}$ and ${}_2F_1(a, b; c; z)$ remains the same as before. The constant C_1 is determined by the initial conditions. In the case we are most interested in, we have only full length TEs at time $t = 0$, $U_L(0) = U_0$ and $\eta(0) = 0$, and as such,

$$C_1 = -A_1 U_L(0)\,{}_2F_1\left(1, 1 + \tfrac{DF}{r_c}, 2 + \tfrac{DF}{r_c}, -(K-1)^{-1}\right).$$

For the zeroth moment, $u_0(t)$, we arrive at the following integral representation

$$u_0(t) = e^{-DFt}\left(\gamma \int u_1(t)\ dt + 2\gamma L \int u_L(t)\ dt\right). \tag{2.40}$$

2.7.5 Solution to the Logistic Growth Discrete System

We now determine the time-varying length distributions for the discrete model under the assumption of logistic growth in the number of full length TEs. In what follows, we use the explicit solution for these full-length dynamics, and the moment closures we obtained in the previous section to derive the partial length dynamics.

Here we prove the existence and uniqueness of the time-varying solution to the discrete logistic model for TE length distribution. Using the implicit solution to the zeroth moment $\eta(t)$ in Equation (2.38), we rewrite Equation (2.2) as

$$\frac{dU_i}{dt} = \left(-DF - \gamma(i+1)\right) U_i(t) - 2\gamma \sum_{k=1}^{i-1} U_k(t) + 2\gamma\,\eta(t) + 2\gamma U_L(t).$$

Thus, we rewrite our discrete system as

$$A\vec{U} + \vec{g}(t) = \vec{U}',$$

for $i = 1, 2, \ldots, L-1$ and where A and $\vec{g}(t)$ are defined as in Equation (2.24). Since $\vec{g}(t)$ is not identically zero, then by *Duhamel's principle*, the solution $\vec{U}(t)$ can be expressed as

$$\vec{U}(t) = e^{At}\vec{U}(0) + \int_0^t e^{A(t-\tau)}\vec{g}(\tau)\ d\tau,$$

and is a unique solution to the inhomogeneous constant-coefficient linear system [39]. Since A is a lower triangular matrix, then the diagonal elements are the eigenvalues of A, namely

$$\lambda_i = -DF - (i+1)\gamma \qquad \text{for } i = 1, 2, \ldots, L-1.$$

Furthermore, A is diagonalizable since we have distinct $\lambda_i \neq 0$ and $L-1$ linearly independent eigenvectors and further, we apply our initial condition $U(0) = \mathbf{0}$. Then, $A = S\Lambda S^{-1}$ and our solution $\vec{U}$ simplifies to

$$\vec{U} = \int_0^t S e^{\Lambda t} S^{-1} \vec{g}(\tau)\, d\tau$$

Let S, S^{-1} be defined as in Equation (2.26), and $\vec{x}_{\log.}(t) = \int_0^t g(\tau) e^{\Lambda(t-\tau)} S^{-1}\mathbf{1}\, d\tau$. Then, let $x_i(t)$ represent the ith entry in $\vec{x}_{\log.}(t)$. Thus,

$$x_i(t) = \sum_{k=1}^{i} k \int_0^t g(\tau) e^{\lambda_i(t-\tau)}\, d\tau = \frac{i(i+1)}{2} \int_0^t 2\gamma[\eta(\tau) + U_L(\tau)] e^{\lambda_i(t-\tau)}\, d\tau.$$

Applying the solution to the zeroth moment, Equation (2.38), we have

$$x_i(t) = \tfrac{i(i+1)}{2} \int_0^t 2\gamma \left[e^{(-DF-\gamma)t} \left(\gamma \int \xi(s) ds + 2\gamma(L-1) \int U_L(s) ds \right) + U_L(\tau) \right] e^{\lambda_i(t-\tau)} d\tau.$$

Thus, we arrive at the implicit solution to Equation (2.2)

$$\vec{U}(t) = S\vec{x}_{\log.}(t).$$

2.7.6 Solution to the Logistic Growth Continuous System

We now determine the time-varying length distributions for the continuous model under the assumption of logistic growth in the number of full length TEs. In what follows, we use the explicit solution for these full-length dynamics, and the moment closures we obtained in 2.7.4 to derive the partial length dynamics. We describe the form of the solution under the logistic growth model in the following theorem.

Theorem 2. *If $u(x,0) = 0$, there exists a solution $u(x,t)$ to the continuous TE length distribution model, Equations* (2.9) *and* (2.11)*, given by*

$$\begin{aligned} u(x,t) &= \mathscr{L}^{-1}\{V(x,s)\}, \\ u_L(t) &= K\left(1 + \exp[-(\alpha - DF - \gamma L)t](K-1)\right), \end{aligned} \tag{2.41}$$

where $\mathscr{L}^{-1}$ is the inverse Laplace transform, $V(x,s) = J(s)(DF + s + \gamma x)^{-3}$ and $J(s)$ is defined in Equation (2.43)*. Moreover, this solution is unique.*

Proof. From Lemma 4, the zeroth moment of the solution to Equation (2.28) is given by

$$\int_0^L V(x,s) = \int_0^L \frac{J(s)}{(DF + s + \gamma x)^3} = J(s)\left(\frac{1}{2\gamma(DF+s)^2} - \frac{1}{2\gamma(DF+s+\gamma L)^2} \right).$$

Next, we apply the Laplace transform to the zeroth moment solution of Equation (2.12) under the logistic model. Equating both sides, we determine $J(s)$ and thus, the unique solution to Equation (2.28)

$$V(x,s) = \frac{J(s)}{(DF + s + \gamma x)^3}, \tag{2.42}$$

where

$$J(s) = \mathscr{L}\{u_1(t)\}\left(\frac{L(2DF + 2s + \gamma L)}{2(DF + s)^2(DF + s + \gamma L)^2}\right), \tag{2.43}$$

where $\mathscr{L}$ is the Laplace transform. We may determine Equation (2.41) by taking the inverse Laplace transform. We prove uniqueness under a similar framework as in [38] and as the previous theorem by assuming we have two solutions to Equation (2.9): $u(x,t)$ and $v(x,t)$ where $u(x,0) = v(x,0)$. We define $g(x,t) = u(x,t) - v(x,t)$. By the same arguments as in Theorem 1, we know that $g(x,t) = 0$, thus establishing uniqueness. □

2.7.7 Complementary Cumulative Distribution Function

In the main text, when comparing models with one another or with data we use the complementary cumulative distribution function (CCDF). Give a distribution $u(x,t)$ and solution derived, we define the $C(x,t)$, the CCDF, as follows:

$$C(x,t) = \begin{cases} 1 - \frac{\int_{x_{\min}}^{x} u(y,t)\ dy}{\int_{x_{\min}}^{L} u(x,t)\ dx + u_L(t)} & \text{if } x < L \\ 0 & \text{if } x = L \end{cases}.$$

In the case for no insertions ($I = 0$), and $\mathbb{D} = \{D\}$ in Equations (2.9) and (2.10) , for $x < L$, $C(x,t)$ is given by

$$C(x,t) = \frac{Q(x,t)}{R(x,t)},$$

where

$$\begin{aligned} Q(x,t) &= (\alpha + D(x_{\min} - L))^2 e^{Dt(x_{\min} - x)} \\ &\quad \cdot \left(\alpha^2 e^{t(\alpha + D(x-L))} + D(L-x)(D(L-x)(Dt(L-x) - \alpha t + 1) - 2\alpha)\right) \\ R(x,t) &= (\alpha + D(x-L))^2 \\ &\quad \cdot \left(\alpha^2 e^{t(\alpha + D(x_{\min} - L))} + D(L - x_{\min})(D(L-x_{\min})(Dt(L-x_{\min}) - \alpha t + 1) - 2\alpha)\right). \end{aligned}$$

If $x_{\min} = 0$, then the ccdf is given by

$$C(x,t) = 1 - \frac{\int_0^x u(y,t)\ dy}{\int_0^L u(x,t)\ dx + u_L(t)} = 1 - \frac{\int_0^x u(y,t)\ dy}{u_0(t) + u_L(t)} \text{ if } x < L. \tag{2.44}$$

2.8 Appendix B: Simulations

In this section, we describe the stochastic simulations in the exponential model used for comparison to the solution of the fragmentation equations (2.2) and (2.3) presented in Section 2.4.2. Moreover, we detail three specific parameter choices for simulation runs and present model fits to these empirical distributions. Lastly, we show that the L2

difference between the analytical steady-state solution (2.44) and the simulations go to zero as $t \to \infty$.

To generate stochastic simulations of TE evolution, we implemented the Gillespie stochastic simulation algorithm [15] and considered each possible discrete TE length as a distinct biological species. In our implementation, we focus on the copy number of TEs in a given genome. That is, for a TE with length L we have L distinct species: 1 represents full-length TEs and the remaining $(L-1)$ for partial length TEs with length 1 to $(L-1)$. Moreover, we consider a simplified version of the general exponential model, with $I = 0$ and $\mathbb{D} = L$. As such, the number of copies of a partial length element i may increase when a deletion affects a TE with length $> i$ or decrease when any length i element is impacted by a deletion, which may happen in i different ways. In contrast, full-length elements only increase by replication rate α and, like length i elements, are impacted by deletions based on their length. In total, we have $\frac{L(L-1)}{2} + 1$ first-order reaction equations. We note that these first order reaction rates conform exactly to the rates associated with our discrete deterministic model formulation (Section 2.4). Finally, we note that neither our stochastic simulation, nor deterministic model, considers the spatial representation of TEs (i.e., a deletion might destroy more than 1 TE). As such, we may fail to faithfully represent TE dynamics in highly TE dense regions of genomes.

Simulations started with zero initial conditions for each repetitive element less than size L. As such, with probability 1, full-length elements are the only TEs we observe at the initial time. This is consistent with the assumption that initially one TE entered the genome at a given time (since we begin with the same complementary cumulative distribution). We terminated the simulation when the relative difference between the cumulative distributions fell below a set threshold. Letting

$$r_\beta = \alpha - 2DL,$$

we explore three different values of α and D with 500 realizations for each parameter combination. We first compare our method with the stochastic simulations for constant $r_\beta = 0.001$ values. This rate corresponds to a net growth for full length TEs in expectation. With fixed r_β and α values, we vary L and compare simulations with derived continuous CCDF (see Table 2.2 below). We use $\alpha = .0025$ in our simulations since experimentally verified TE rates range from 10^{-3} to 10^{-5} and other studies support these specific values [32, 45].

L	α	D
300	0.0025	2.5×10^{-6}
1000	0.0025	7.5×10^{-7}
7500	0.0025	1×10^{-7}

Table 2.2: Model parameters TE Length, L, transposition rate, α, and deletion rate D for stochastic simulations. For each parameter set above, we run 500 simulations until the relative difference between the cumulative distributions met a threshold. Each simulation follow constant $r_\beta = 0.001$ growth for full-length transposable elements.

Using the 500 trials, we plot the L2 difference in time for a single simulation and the mean distribution of all results from Gillespie SSA for all parameter values in Figure 2.13. Since our model represents the expected value of this stochastic process, we expect and confirm that the difference between the average distribution of simulations and our analytical solution is initially small and quickly decays to zero [61]. Additionally, we note our solution captures the behavior of a single stochastic simulation, which is representative of real genomic data.

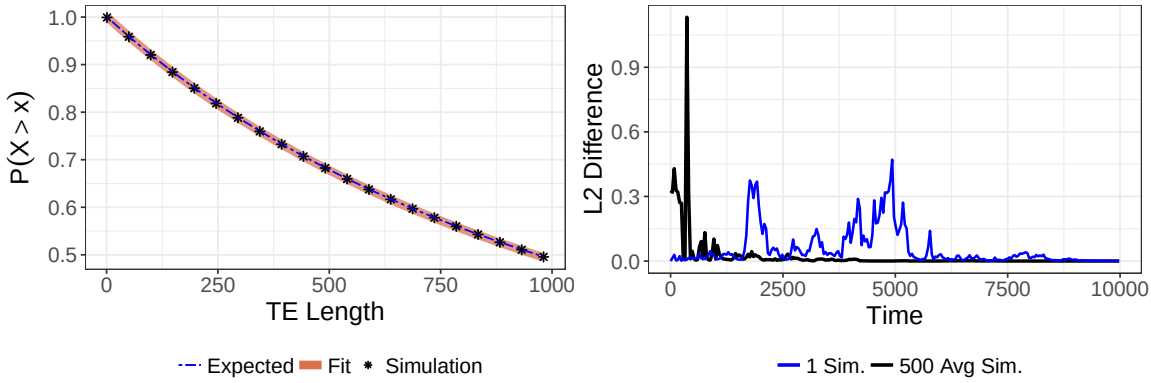

Figure 2.13: a) Complementary cumulative distribution plots of simulation (blue), steady-state analytical fit (orange), and true value (black) for one realization using a non-dimensionalized CCDF with the assumption that the replicative process has reached steady-state. b) Plot of the L2 difference between the derived steady-state CCDF and the empirical CCDF for $L = 1000$ for one (blue) and 500 simulations (black).

2.9 Appendix C: Data and Parameter Inference

In this Appendix, we list our data sources and processing steps for *Drosophila*, *Aves*, and *Primates* data as discussed in Section 3.3.1. We present specific parameter estimation with the steady-state analytical distribution described in 2.7.7 and corresponding fits for *Drosophila* roo and rooA elements.

2.9.1 Repeat data and processing

RepeatMasker Data We apply our deterministic model to a variety of RepeatMasker Data from different species. We note that other methods exists to detect and report repeats in genomes (see [5] for a review of TE-detection tools) and those may be incorporated into this framework; however, since RepeatMasker utilizes RepBase for consensus sequences (i.e. does not focus on *de novo* TE detection). We acknowledge that these RepeatMasker annotations may be influenced by improper genome assembly, but since data are readily available, we incorporate these data into our analytical framework [25, 33]. Of the three case studies we consider, most research focused on TEs in Drosophila and humans. Researchers posit the nontrivial role of population size in TE distribution

and transposition rate, and we consider three species with varying effective population sizes in our analysis [17, 8].

Drosophila. We obtained the current releases of RepeatMasker *.out* files obtained from the UCSC Genome Browser (`http://genome.ucsc.edu`), representing total repeats from the 12 completely sequenced genomes of *Drosophila* [**ucscGB**]. Each genome was run with *-s* option in RepeatMasker. Specifically, we focus on the distribution of all *roo* and *rooA* elements in the *Drosophila* clade. The canonical lengths for *roo* and *rooA* are 9092 basepairs (bp) and 7621 bp, respectively [26]. Previous studies suggested that the *rooA* element diverged early in the evolution of the *Drosophila* clade [10]. We report the number of partial and full copies of both *roo* and *rooA* in Table 2.1. We identify full-length elements as having 99% match to the canonical element. Our reported counts differ from that of [10] since we incorporate a more stringent criteria to classify an element as a full copy. We determine lengths of contiguous repeats using Perl scripts from [3] without the *strict* option since RepeatMasker classifies the internal part, ROO_I, of the TE separately from the flanking regions, ROO_LTR. For partial TE copies for all species, we only consider elements equal to 200bp or longer.

Aves. We gathered Avian RepeatMasker output from the Avian Phylogenomics Project (http://avian.genomics.cn/en/jsp/database.shtml) as well as RepeatMasker directly, resulting in repeat information for 23 species of birds. In comparison to other eukaryotes, TEs do not comprise the majority of bird genomes [27, 28]. We follow a similar procedure to *Drosophila* data, but instead we focus on a specific lineage of Chicken repeat 1 (CR1) retrotransposons, which are the most abundant superfamily in nearly all amniotes and have been used for previous phylogenic analyses [53].

Primates. We obtained Primate RepeatMasker output from RepeatMasker directly, resulting in repeat information for 5 species of primates. These include gorilla, human, gibbon, chimp, and orangutan species. We apply our method to non-active elements, namely, L2 and L3 long interspersed elements (LINEs) with canonical lengths 3082bp and 4099bp, respectively. These elements influenced the evolution of these species but do not currently replicate [17, 50]. As in the previous data sets, we consider partial length TEs those longer than 200bp.

2.9.2 Drosophila roo-rooA parameter inference

The following table and Figures represent the complementary cumulative distribution fit with the empirical distribution of the two TEs in *Drosophila* discussed in the text.

Table 2.3: Parameter estimation θ for *rooA* element in Drosophila.
*12 Dro reflects the averaged estimate of θ over all 12 fruit flies considered.

Species	θ Estimate	Standard Error	Confidence Interval
12 Dro*	1.1277	0.0013	[1.1251, 1.1303]
D.mel	1.0766	0.0139	[1.0488, 1.1044]
D.moj	0.9933	0.0031	[0.9871, 0.9995]
D.sim	0.9457	0.0022	[0.9413, 0.9501]
D.ere	1.2775	0.0093	[1.2589, 1.2961]
D.per	0.9135	0.0024	[0.9087, 0.9183]
D.yak	1.1445	0.0033	[1.1379, 1.1511]

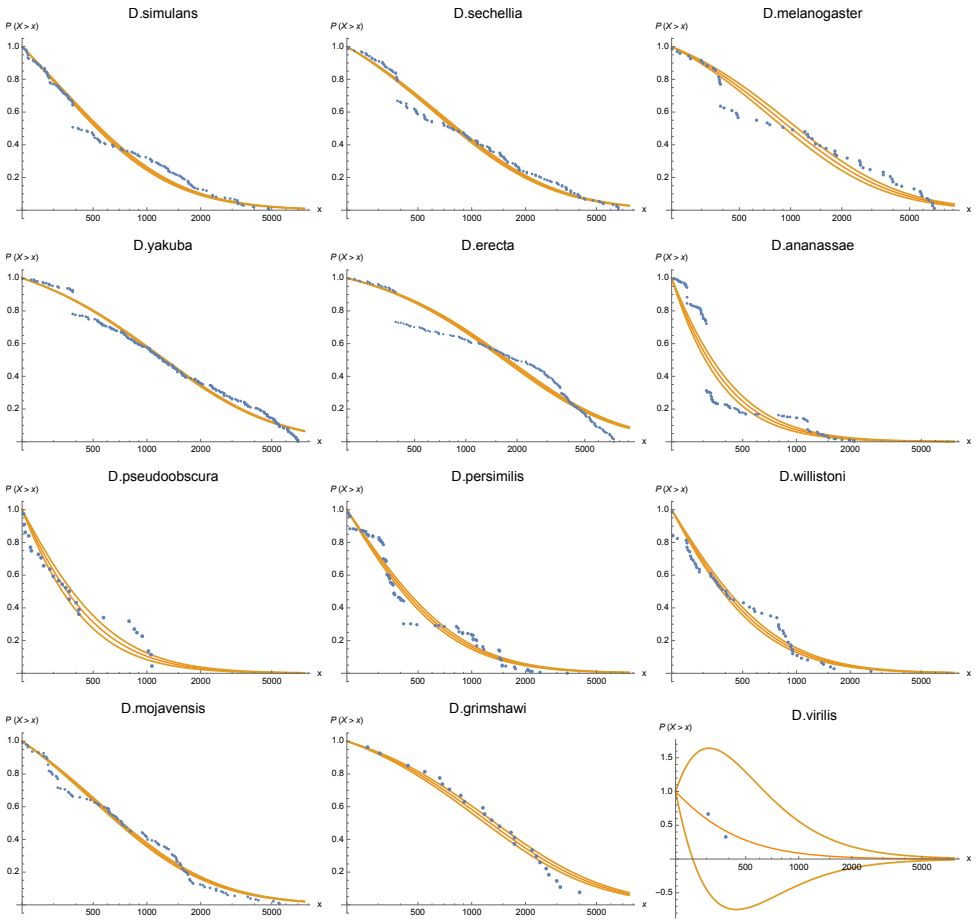

Figure 2.14: Complementary cumulative distribution (CCDF) plots of empirical *rooA* data (blue) and analytical fits (orange) for the 12 *Drosophila* species using a non-dimensionalized CCDF (2.44) with the assumption that the replicative process has reached steady-state. We consider only partial length TEs longer than 200bp in our parameter estimation. Since most species have relatively few full-length elements, our steady-state solution captures the qualitative behavior of these TE distributions. We exclude conclusions for *D. virilis* in Section 2.5.3 due to lack of TE data.

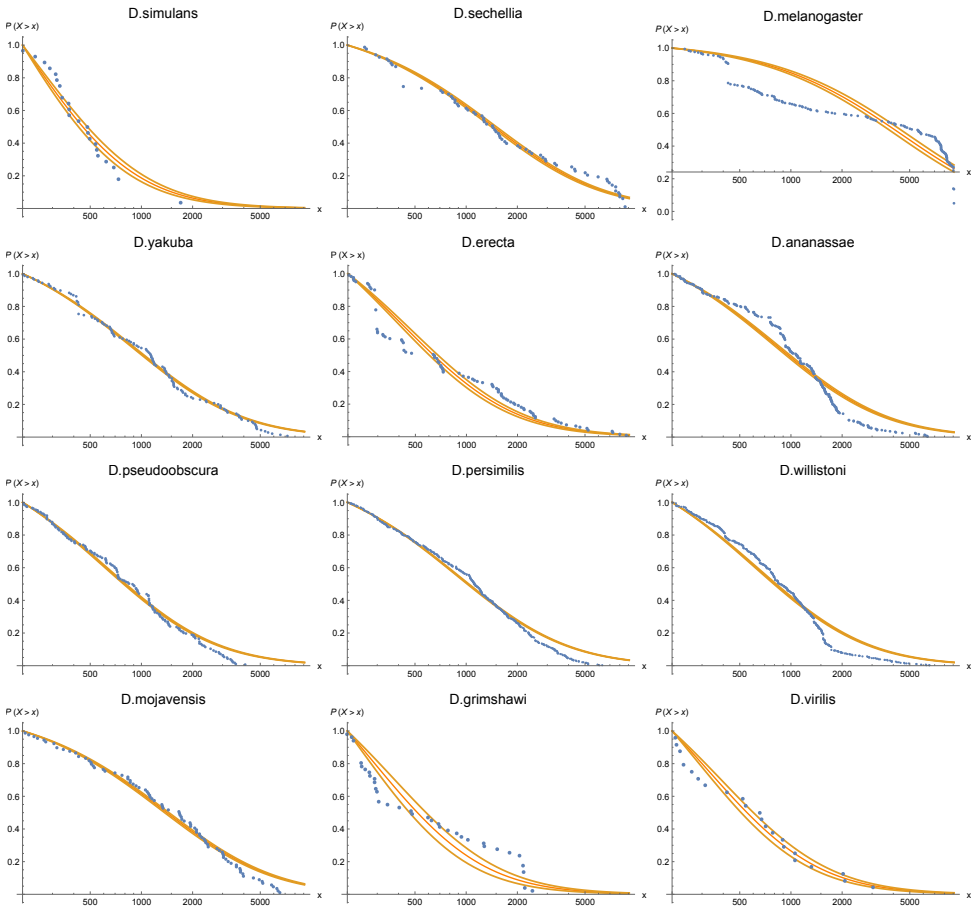

Figure 2.15: Complementary cumulative distribution (CCDF) plots of empirical *roo* data (blue) and analytical fits (orange) for the 12 *Drosophila* species using a non-dimensionalized CCDF (2.44) with the assumption that the replicative process has reached steady-state. Since most species have relatively few full-length roo elements, our steady-state solution captures the qualitative behavior of these TE distributions. For species with remaining actively replicating full-length copies (e.g., D. melanogaster), this assumption may not hold and results in a fit not reflecting the true distribution. We consider only partial length TEs longer than 200bp in our parameter estimation.

Bibliography

[1] M. Ahmed and P. Liang. “Transposable elements are a significant contributor to tandem repeats in the human genome”. *Comparative and functional genomics* 2012 (2012).

[2] Alan M Weiner. *SINEs and LINEs: the art of biting the hand that feeds you*. 2002. DOI: 10.1016/S0955-0674(98)80011-2.

[3] M. Bailly-Bechet, A. Haudry, and E. Lerat. ““One code to find them all”: a perl tool to conveniently parse RepeatMasker output files”. *Mobile DNA* 5:13 (2014).

[4] W. Bao and J. Jurka. “Origin and evolution of LINE-1 derived ?half-L1? retrotransposons (HAL1)”. *Gene* 465.1 (2010), pp. 9–16.

[5] C. M. Bergman and H. Quesneville. “Discovering and detecting transposable elements in genome sequences”. *Briefings in bioinformatics* 8.6 (2007), pp. 382–392.

[6] D. W. Burt. “Origin and evolution of avian microchromosomes.” *Cytogenetic and genome research* 96.1-4 (2002), pp. 97–112. ISSN: 1424-8581. DOI: 63018. URL: http://www.ncbi.nlm.nih.gov/pubmed/12438785.

[7] P. Capy. *Dynamics and evolution of transposable elements*. North American distributor Chapman & Hall, 1998.

[8] B. Charlesworth. “Effective population size and patterns of molecular evolution and variation”. *Nature Reviews Genetics* 10.3 (2009), pp. 195–205.

[9] B. Charlesworth and D. Charlesworth. “The population dynamics of transposable elements”. *Genetical Research* 42.01 (1983), pp. 1–27.

[10] N. de la Chaux and A. Wagner. “Evolutionary dynamics of the LTR retrotransposons roo and rooA inferred from twelve complete Drosophila genomes”. *BMC evolutionary biology* 9:205 (2009).

[11] A. G. Clark et al. “Evolution of genes and genomes on the Drosophila phylogeny”. *Nature* 450.7167 (2007), p. 203.

[12] E. S. Dolgin and B. Charlesworth. “The fate of transposable elements in asexual populations”. *Genetics* 174.2 (2006), pp. 817–827.

[13] T. A. Elliott and T. R. Gregory. “What’s in a genome? The C-value enigma and the evolution of eukaryotic genome content”. *Phil. Trans. R. Soc. B* 370.1678 (2015), p. 20140331.

[14] N. V. Fedoroff. “Transposable elements, epigenetics, and genome evolution”. *Science* 338.6108 (2012), pp. 758–767.

[15] D. T. Gillespie. “Exact stochastic simulation of coupled chemical reactions”. *The journal of physical chemistry* 81.25 (1977), pp. 2340–2361.

[16] J. Giordano et al. “Evolutionary History of Mammalian Transposons Determined by Genome-Wide Defragmentation”. *PLoS Computational Biology* 3.7 (2007), e137. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.0030137. URL: http://dx.plos.org/10.1371/journal.pcbi.0030137.

[17] J. González and D. A. Petrov. “Evolution of genome content: Population dynamics of transposable elements in flies and humans”. *Methods in Molecular Biology*. Ed. by M. Anisimova. Vol. 855. Humana Press, 2012, pp. 361–383. ISBN: 9781617795817. DOI: 10.1007/978-1-61779-582-4_13.

[18] T. R. Gregory et al. “The smallest avian genomes are found in hummingbirds”. *Proceedings of the Royal Society of London B: Biological Sciences* 276.1674 (2009). DOI: 10.1098/rspb.2009.1004.

[19] M. G. Guerreiro. “What makes transposable elements move in the Drosophila genome?” *Heredity* 108.5 (2012), p. 461.

[20] C. Haag-Liautard et al. “Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila”. *Nature* 445.7123 (2007), p. 82.

[21] D. J. Hedges and M. A. Batzer. “From the margins of the genome: mobile elements shape primate evolution”. *Bioessays* 27.8 (2005), pp. 785–794.

[22] J. Iranzo et al. “Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes”. *PLoS computational biology* 10.6 (2014), e1003680.

[23] L. J. Johnson and J. F. Brookfield. “A test of the master gene hypothesis for interspersed repetitive DNA sequences”. *Molecular biology and evolution* 23.2 (2005), pp. 235–239.

[24] I. Jordan et al. “Origin of a substantial fraction of human regulatory sequences from transposable elements”. *Trends in Genetics* 19.2 (2003), pp. 68–72. ISSN: 01689525. DOI: 10.1016/S0168-9525(02)00006-9.

[25] J. Jurka et al. “Repbase Update, a database of eukaryotic repetitive elements”. *Cytogenetic and genome research* 110.1-4 (2005), pp. 462–467.

[26] J. S. Kaminker et al. “The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective”. *Genome Biol* 3.12 (2002), RESEARCH0084.

[27] A. Kapusta and A. Suh. “Evolution of bird genomes?a transposon’s-eye view”. *Annals of the New York Academy of Sciences* (2016).

[28] A. Kapusta, A. Suh, and C. Feschotte. “The hidden elasticity of avian and mammalian genomes”. *bioRxiv* (2016). DOI: http://dx.doi.org/10.1101/081307. URL: http://dx.doi.org/10.1101/081307.

[29] H. H. Kazazian. “Mobile elements: drivers of genome evolution”. *science* 303.5664 (2004), pp. 1626–1632.

[30] H. H. Kazazian et al. “Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man” (1988).

[31] M. G. Kidwell and D. R. Lisch. “Perspective: transposable elements, parasitic DNA, and genome evolution”. *Evolution* 55.1 (2001), pp. 1–24.

[32] R. Kofler, V. Nolte, and C. Schlötterer. “Tempo and mode of transposable element activity in Drosophila”. *PLoS Genet* 11.7 (2015), e1005406.

[33] O. Kohany et al. “Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor”. *BMC bioinformatics* 7.1 (2006), p. 474.

[34] M. Koroteev and J. Miller. “Scale-free duplication dynamics: A model for ultraduplication”. *Physical Review E* 84.6 (2011), p. 061919.

[35] A. Le Rouzic and P. Capy. “Population genetics models of competition between transposable element subfamilies”. *Genetics* 174.2 (2006), pp. 785–793.

[36] A. Le Rouzic and G. Deceliere. “Models of the population genetics of transposable elements”. *Genetical research* 85.03 (2005), pp. 171–181.

[37] A. Le Rouzic, T. Payen, and A. Hua-Van. “Reconstructing the evolutionary history of transposable elements”. *Genome biology and evolution* 5.1 (2013), pp. 77–86.

[38] G. Ledder. “A Simple Introduction to Integral Equations”. *Mathematics magazine* 69.3 (1996), pp. 172–181.

[39] R. J. LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. Vol. 98. Siam, 2007.

[40] A. Levy, S. Schwartz, and G. Ast. “Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements”. *Nucleic acids research* 38.5 (2009), pp. 1515–1530.

[41] R. S. Linheiro and C. M. Bergman. “Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster”. *PloS one* 7.2 (2012), e30008.

[42] G. Liu et al. “Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome”. *Genome research* 13.3 (2003), pp. 358–368.

[43] X. Maside, S. Assimacopulos, and B. Charlesworth. “Rates of movement of transposable elements on the second chromosome of Drosophila melanogaster”. *Genetical research* 75.03 (2000), pp. 275–284.

[44] B. McClintock. “The origin and behavior of mutable loci in maize”. *Proceedings of the National Academy of Sciences* 36.6 (1950), pp. 344–355.

[45] S. V. Nuzhdin. “Sure facts, speculations, and open questions about the evolution of transposable element copy number”. *Genetica* 107.1-3 (1999), p. 129.

[46] L. E. Orgel and F. H. Crick. “Selfish DNA: the ultimate parasite”. *Nature* 284 (1980), pp. 604–607.

[47] D. A. Petrov. “Mutational equilibrium model of genome size evolution”. *Theoretical population biology* 61.4 (2002), pp. 531–544.

[48] D. A. Petrov et al. “Population genomics of transposable elements in Drosophila melanogaster”. *Molecular biology and evolution* 28.5 (2011), pp. 1633–1644.

[49] M. Sheinman et al. “Evolutionary dynamics of selfish DNA explains the abundance distribution of genomic subsequences”. *Scientific reports* 6 (2016).

[50] J. Silva et al. “Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR-and L2-derived sequences within the mouse and human genomes”. *Genetical research* 82.01 (2003), pp. 1–18.

[51] R. K. Slotkin and R. Martienssen. “Transposable elements and the epigenetic regulation of the genome”. *Nature Reviews Genetics* 8.4 (2007), pp. 272–285.

[52] C. J. Struchiner et al. “The tempo and mode of evolution of transposable elements as revealed by molecular phylogenies reconstructed from mosquito genomes”. *Evolution* 63.12 (2009), pp. 3136–3146.

[53] A. Suh. “The specific requirements for CR1 retrotransposition explain the scarcity of retrogenes in birds”. *Journal of molecular evolution* 81.1-2 (2015), pp. 18–20.

[54] A. Suh et al. “Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes”. *Nature communications* 7 (2016).

[55] W. Sung et al. “Evolution of the insertion-deletion mutation rate across the tree of life”. *G3: Genes, Genomes, Genetics* 6.8 (2016), pp. 2583–2591.

[56] A. Szitenberg et al. “Genetic drift, not life history or RNAi, determine long-term evolution of transposable elements”. *Genome biology and evolution* 8.9 (2016), pp. 2964–2978.

[57] K. Tamura, S. Subramanian, and S. Kumar. “Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks”. *Molecular biology and evolution* 21.1 (2004), pp. 36–44.

[58] M. I. Tenaillon et al. “Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians”. *Genome biology and evolution* 3 (2011), pp. 219–229.

[59] B. G. Thornburg, V. Gotea, and W. Makałowski. “Transposable elements as a significant source of transcription regulating signals”. *Gene* 365 (2006), pp. 104–110. ISSN: 03781119. DOI: 10.1016/j.gene.2005.09.036.

[60] T. L. Vandergon and M. Reitman. “Evolution of chicken repeat 1 (CR1) elements: evidence for ancient subfamilies and multiple progenitors.” *Molecular Biology and Evolution* 11.6 (1994), pp. 886–898.

[61] D. J. Wilkinson. *Stochastic modelling for systems biology*. CRC press, 2011.

[62] C. Xue and N. Goldenfeld. “Stochastic predator-prey dynamics of transposons in the human genome”. *Physical review letters* 117.20 (2016), p. 208101.

Chapter 3

Detecting Genomic Variation Within Species

In this chapter, I develop a statistical and optimization framework for detecting structural variations (SVs) in related individuals of the same species (i.e., humans). Furthermore, these methods are tested and validated on simulated and real genomic datasets, respectively. The work described in in this chapter is based on the papers by Banuelos et. al [10, 9, 6, 8, 5, 7, 3, 4].

3.1 Introduction

Recent advances in high-throughput sequencing technologies have led to the collection of vast quantities of genomic data. The 1000 Genomes Project [2], which catalogues human genomic variation in comprehensive detail, and the 3000 Rice Genomes Project [24, 28], which reports an international resequencing effort of 3,000 rice genomes, are two successful examples of such large-scale sequencing studies. These massive repositories of data offer the potential to increase our understanding of the complex evolutionary history of different species, identify genetic basis of important phenotypes including disease and – for humans – usher in the era of personalized medicine [33, 47]. A promising class of genetic variant emerging from such studies are structural variants (SVs) – rearrangements of the genome larger than one letter such as inversions, insertions, deletions, and duplications (see Fig. 3.1).

SVs are typically predicted by sequencing fragments from an unknown individual genome and mapping those fragments to a previously identified reference genome [29, 41]. If the starting points of the genomic fragments are chosen uniformly and randomly from the genome, then the expected number of fragments covering any position in the genome may be modeled by a Poisson distribution [27]. I consider multiple assumptions on this sequencing process in Section 3.2. The mean of the sequencing distribution is referred to as the *coverage* of the genome. Since, in most large sequencing studies, many individuals will be sequenced at low coverage, even if an individual carries a genetic variant, we may not sample a fragment from that particular region of the genome.

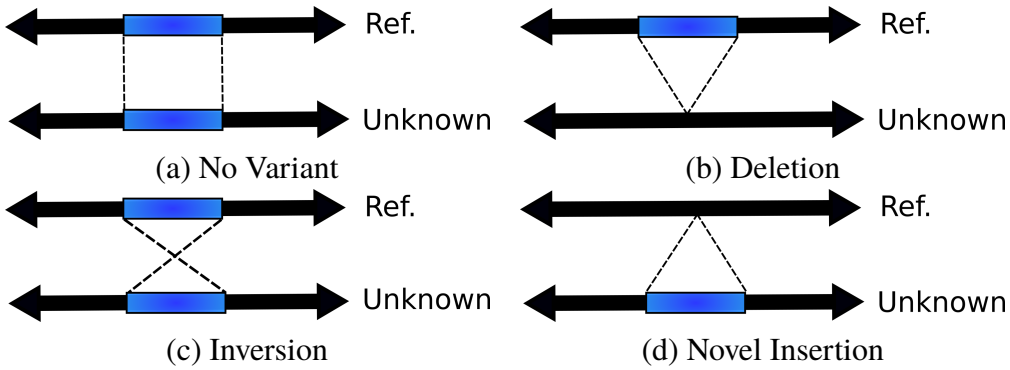

Figure 3.1: Example of different structural variations (b) - (d) in an unknown genome in comparison to the reference genome. When there is no difference between the reference genome and the unknown genome, then there is no variant present (a).

Similarly, if we observe a single fragment supporting a variant, it may represent an erroneous mapping rather than a true observation.

There have been many published methods to identify SVs from sequencing data (see, e.g., [43, 23, 15, 35, 34]). However, these approaches almost universally rely on high-coverage of a single individual genome and not on the scenario emerging from many large-scale sequencing efforts where there is low-coverage of many individuals. In addition, prior approaches when applied to populations typically consider each individual in isolation when – in fact – common variants would be shared by many individuals. Finally, most methods utilize a threshold – minimum number of supporting fragments – to prioritize predicted variants rather than a likelihood based statistic. Indeed, inferring SV information from sequencing data has proven to be challenging because true SVs are rare and are prone to low-coverage noise.

In this work, I attempt to mitigate the challenges of low-coverage sequencing by following a maximum likelihood approach to SV prediction. Specifically, I model the noise using Poisson and Negative Binomial statistics and constrain the solution to promote sparsity, i.e., SV instances should be rare. Further, I consider multiple individuals and use relatedness among individuals as a constraint on the solution space – to our knowledge, this is the first SV detection algorithm to do so. Specifically, in our work below I use the assumption that a parent and child are sequenced and require that any SVs predicted in the child be present in the parent. Numerical analysis of both simulated and real sequencing data suggest that our approach has the promise to improve SV detection in studies of many low-coverage individuals.

3.2 General Optimization Framework

To detect genomic variants, we first obtain sequencing data from related individuals. This data represents the counts for candidate variant positions and does not represent all loci of an individual's genome. I proceed with a maximum likelihood approach to

maximize the probability of observing these counts with the following framework:

$$\begin{aligned} &\text{maximize} && P(\text{data} \mid \mathbb{S}) \\ &\text{subject to} && \mathscr{H} \end{aligned}$$

where $\mathbb{S}$ represents the assumed sequencing model for the distribution of the count number of DNA fragments supporting a potential SV and $\mathscr{H}$ represents the heredity constraints reflected by different family structures (e.g., one parent and one child). Thus, $\mathbb{S}$ alters the objective function and $\mathscr{H}$ changes the feasible region in the constrained optimization problem. In the subsequent sections, I describe methods based on two different sequencing frameworks. First, the Poisson distribution models the process of DNA sequencing. Under this framework, I consider changing the constraint assumption to include different family structures, haploid, diploid, and nonconvex formulations. Second, I consider changing the assumption to model sequencing as a negative binomial process. This, in turn, results in a nonconvex formulation, and I present my contributions to address the modified optimization problem for a variety of assumptions on heredity assumptions $\mathscr{H}$.

3.3 DNA Sequencing as a Poisson Process

3.3.1 Haploid One Parent-One Child Method

The work described in this section is based on the paper by Banuelos et al. [10]. Let $\vec{f}_i^* \in \{0,1\}^n$ be the vector of genomic variants for an individual i, i.e., $\vec{f}_{i,j}^* = 1$ if individual i has genomic variant j and is 0 otherwise. This classification framework thus considers the haploid – one copy of a genomic variant per individual – inheritance of variants from parent to child. Let $\vec{y}_i \in \mathbb{Z}_+^n$ be the vector of observations for individual i. The variables $\vec{y}_{i,j}$ obey a Poisson distribution [44] whose mean, c_i, is equal to the sequencing coverage of individual i. In this work, we specifically consider the structural variants for two individuals who are related, namely a parent and child. Let $\vec{f}_p^*$ and $\vec{f}_c^*$ be the true genomic variants for a parent and child, respectively. Then the corresponding observations, denoted by $\vec{y}_p$ and $\vec{y}_c$, are given by

$$\begin{aligned} \vec{y}_p &\sim \text{Poisson}\left(A_p \vec{f}_p^*\right) \\ \vec{y}_c &\sim \text{Poisson}\left(A_c \vec{f}_c^*\right), \end{aligned} \tag{3.1}$$

where $A_p = \left(k_p - \varepsilon\right)\mathbb{I}, A_c = \left(k_c - \varepsilon\right)\mathbb{I} \in \mathbb{R}^{n\times n}$ linearly transforms $\vec{f}_p^*,\ \vec{f}_c^*$ onto an n-dimensional set of observations $\vec{y}_p,\ \vec{y}_c \in \mathbb{Z}_+^n$. The constants k_p and k_c represent the sequencing coverage of the parent and child genome, respectively. It is assumed that ε, the error term in the measurement of the true signals, is the same for both observations.

We consider a general framework for the recovery of variant detection given sequencing data from one haploid parent and one haploid child. Our observation $\vec{y}$ will be considered a stacked signal in the form $\begin{bmatrix}\vec{y}_p^T & \vec{y}_c^T\end{bmatrix}^T$, where $\vec{y}_p$ and $\vec{y}_c$ represent observations

of parent and child, respectively. Since the true signal $\vec{f}^*$ is also stacked, our observation model is given by

$$\vec{y} \sim \text{ Poisson}\left(\hat{A}\,\vec{f}^*\right), \tag{3.2}$$

where $\hat{A} \in \mathbb{R}^{2n\times 2n}$ is a block-diagonal matrix with upper-left block A_p and lower-left block A_c.

Problem formulation

Under this Poisson model (3.2), the probability of observing $\vec{y}$ is given by

$$p(\vec{y}|\hat{A}\vec{f}^*) = \prod_{i=1}^{2n} \frac{(\vec{e}_i^T\hat{A}\vec{f}^*)^{\vec{y}_i}}{\vec{y}_i!} \exp\left(-\vec{e}_i^T\hat{A}\vec{f}^*\right), \tag{3.3}$$

where $\vec{e}_i$ is the ith canonical basis vector. Under a similar framework in [22] and ignoring constant terms $\log(\vec{y}_i!)$, we minimize the negative Poisson log-likelihood given by

$$F(\vec{f}) = \mathbf{1}^T\hat{A}\vec{f} - \sum_{i=1}^{2n} \vec{y}_i \log\left(\vec{e}_i^T\hat{A}\vec{f} + \varepsilon\right), \tag{3.4}$$

with gradient

$$\nabla F(\vec{f}) = \hat{A}^T\mathbf{1} - \sum_{i=1}^{2n} \frac{y_i}{\vec{e}_i^T\hat{A}\vec{f} + \varepsilon}\hat{A}^T\vec{e}_i, \tag{3.5}$$

where $\mathbf{1}$ is a vector of ones. Hence, we focus on solving the following constrained optimization problem:

$$\begin{aligned} &\underset{\vec{f}\in\mathbb{R}^{2n}}{\text{minimize}} \quad \phi(\vec{f}) \equiv F(\vec{f}) + \tau\ \text{pen}(\vec{f}) \\ &\text{subject to} \quad 0 \le \vec{f}_c \le \vec{f}_p \le 1, \end{aligned} \tag{3.6}$$

where $\vec{f} = \begin{bmatrix}\vec{f}_p \\ \vec{f}_c\end{bmatrix}$, $\tau > 0$ is a regularization parameter, and pen is usually a non-differentiable penalty functional. Here, we impose the constraint $0 \le \vec{f}_c \le \vec{f}_p \le 1$ element-wise to enforce the continuous variables $\vec{f}_c$ and $\vec{f}_p$ to lie between 0 and 1 (i.e., SVs are either present or not), but in addition, to require that a variant in the child genome can be present only when the parent genome also has that variant.

Sparsity penalty

Our approach to solving (3.6) is based on SPIRAL [22, 20, 21], which is an iterative method whose iterates are defined from minimizing a sequence of quadratic subproblems. This approach utilizes the second-order Taylor expansion of the Poisson log-likelihood, $F(\vec{f})$, around the current iterate $\vec{f}^k$ and approximates the second derivative matrix by a

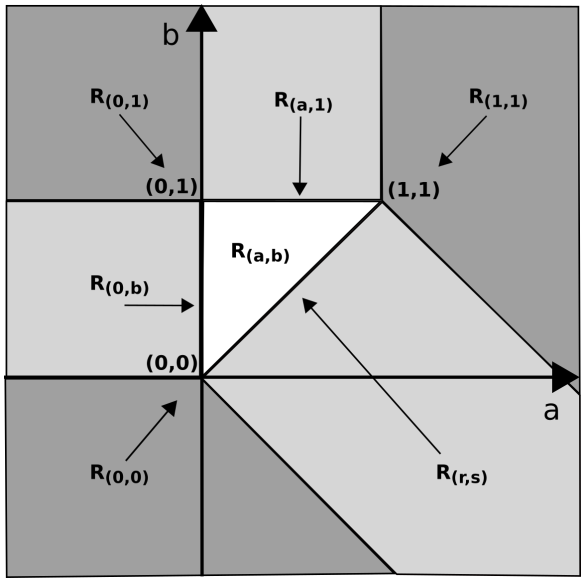

Figure 3.2: Plot of a-b plane, where regions are defined in Table 3.1 and $R_{(a,b)}$ represents the feasible region for the solution of the separable subproblem (3.25).

scalar multiple of the identity matrix $\alpha_k I$, $\alpha_k > 0$ [22, 11, 48]. Thus, the next iterate is given by

$$\vec{f}^{k+1} = \begin{bmatrix} \vec{f}_p^{k+1} \\ \vec{f}_c^{k+1} \end{bmatrix} = \underset{\vec{f} \in \mathbb{R}^{2n}}{\arg\min} \quad F^k(\vec{f}) + \tau \ \text{pen}(\vec{f}) \qquad \text{subject to } \ 0 \le \vec{f}_c \le \vec{f}_p \le 1, \tag{3.7}$$

where

$$F^k(\vec{f}) = \nabla F(\vec{f}^k)^T(\vec{f} - \vec{f}^k) + \frac{\alpha_k}{2}\|\vec{f} - \vec{f}^k\|_2^2.$$

Manipulating $F^k(\vec{f})$ leads to the following equivalent optimization formulation:

$$\vec{f}^{k+1} = \underset{\vec{f} \in \mathbb{R}^{2n}}{\arg\min} \quad \frac{1}{2}\|\vec{f} - \vec{s}^k\|_2^2 + \frac{\tau}{\alpha_k}\text{pen}(\vec{f}) \qquad \text{subject to } \ 0 \le \vec{f}_c \le \vec{f}_p \le 1, \tag{3.8}$$

where

$$\vec{s}^k = \begin{bmatrix} \vec{s}_p^k \\ \vec{s}_c^k \end{bmatrix} = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k). \tag{3.9}$$

When $\text{pen}(\vec{f}) = \|\vec{f}\|_1 = \sum_{i=1}^n |\vec{f}_i|$, the objective function in (3.8) decouples in each variable and can be optimized separately, which results in the following *scalar* optimization:

$$\underset{f_p, f_c \in \mathbb{R}}{\text{minimize}} \ \frac{1}{2}(f_p - s_p)^2 + \lambda|f_p| + \frac{1}{2}(f_c - s_c)^2 + \lambda|f_c| \qquad \text{subject to } \ 0 \le f_c \le f_p \le 1, \tag{3.10}$$

where f_p and f_c correspond to each scalar element of $\vec{f}_p$ and $\vec{f}_c$, respectively. Since both f_p and f_c are non-negative, the absolute values in (3.10) can be dropped. Completing the squares in (3.10) and ignoring constant terms yield

$$\begin{aligned} &\underset{f_p, f_c \in \mathbb{R}}{\text{minimize}} \ \psi(f_p, f_c) = \frac{1}{2}(f_p - b)^2 + \frac{1}{2}(f_c - a)^2 \\ &\text{subject to } \ 0 \le f_c \le f_p \le 1, \end{aligned} \tag{3.11}$$

where $a = s_c - \lambda, \ b = s_p - \lambda$. The solution to (3.25) can be obtained by partitioning the a-b plane into different regions (see Fig. 3.2). Then the minimizer of (3.25) depends on the region in which the point (a, b) lies. For example, if $(a, b) \in R_{(a,b)}$, i.e., $0 \le a \le b \le 1$, then the minimizer, (f_c^*, f_p^*), of (3.25) is the point (a, b). The complete set of minimizers is listed in Table 3.1.

Region	Condition a	Condition b	$\left(f_c^*, f_p^*\right)$
$R_{(a,b)}$	$0 < a < b$	$0 < b < 1$	(a, b)
$R_{(0,b)}$	$a < 0$	$0 \le b \le 1$	$(0, b)$
$R_{(a,1)}$	$0 \le a \le 1$	$b > 1$	$(a, 1)$
$R_{(0,1)}$	$a < 0$	$b > 1$	$(0, 1)$
$R_{(0,0)}$	$a \le -b$	$b < 0$	$(0, 0)$
$R_{(1,1)}$	$a > 1$	$b \ge -a + 2$	$(1, 1)$
$R_{(r,s)}$	$a > \lvert b \rvert$	$b < -a + 2$	(r, s)

Table 3.1: Table representing the solution to (3.25) as a function of a and b. Here, $r = s = (a + b)/2$.

Numerical Results

The solution to the problem proposed in the previous section was implemented using the SPIRAL-ℓ_1 algorithm in [22] with the appropriate modifications to accommodate for the different constraints (see (3.6)). The results obtained are compared to those of the original SPIRAL-ℓ_1 approach in order to evaluate the validity of the proposed approach on both simulated and real genomic data.

Two simulated test signals, $\vec{f}_p$ and $\vec{f}_c$, of length $n = 10^5$ were used to examine the effectiveness of the proposed approach. We varied the coverage of both between 2 and 10, and the child is chosen to have between 70% to 90% of the variants in the parent. The true signal for the parent $\vec{f}_p$, is set to be 0.5% sparse, so that only 500 variants are present. Furthermore, consistent with the assumption of similar error term in the measurement of the true signals, a single value of $\varepsilon = 0.01$ was selected. On the simulated data, we are able to select the optimum value for τ and found on this data the optimal τ occurred between 0.5 and 3. Further, we observed limited sensitivity to τ as the model with and without family constraints had a similar τ range.

We first examined the parent signal reconstruction. Fig. 3.3 illustrates a small segment ($n = 2.5 \times 10^4$) of the parent signal with $k_p = 2$, $k_c = 2$, and 90% similarity of

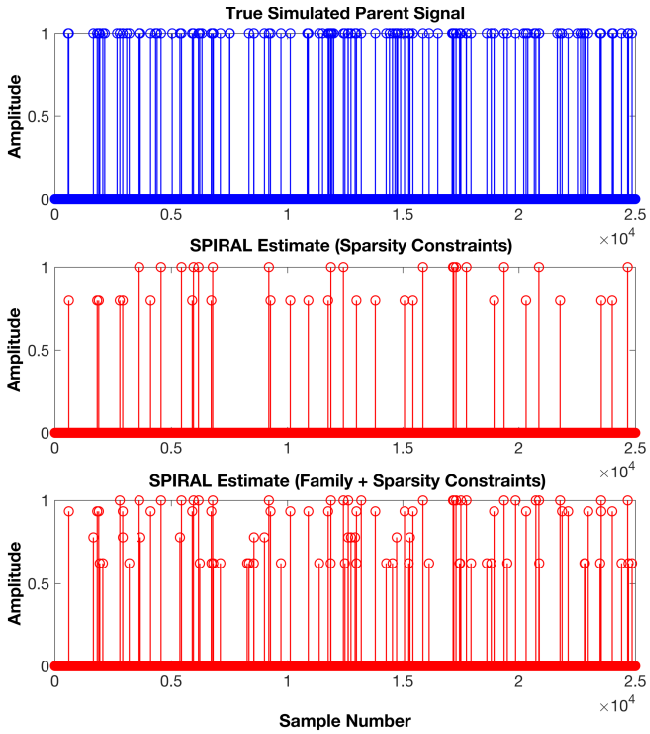

Figure 3.3: From top to bottom: A small segment of the parent signal with $k_p = 2$, $k_c = 2$, and 90% similarity of variants; reconstruction using the sparsity SPIRAL constraints with $\tau = 1.779$ yielded 152 correctly identified out of 500; and reconstruction using the family and sparsity constraints with $\tau = 1.221$ yielded 349 correctly identified out of 500.

variants, the reconstructed signal obtained by the sparsity-only SPIRAL constraints, and the reconstructed signal obtained by the family+sparsity SPIRAL constraints both at a threshold value of 0.5308. The improvement in variant predictions is visually clear from this figure.

We observed that an increase in the coverage of either child or parent helps improve the quality of the predictions. Moreover, the greater the similarity between parent and child, the more helpful adding the familial constraints results. Fig. 3.4 further illustrates how the familiarly constrained model ranks all true predictions above all false predictions.

We apply our method to the previously sequenced genomes of the father-mother-daughter CEU trio (NA12891, NA12892, NA12878) from the 1000 Genomes Project [2]. These genomes were sequenced to low coverage ($\approx$ 4×) in Pilot 1 of the study and high coverage ($\approx$ 40×) in Pilot 2. Both were aligned to NCBI36. We compared our reconstructions against the reported validated set of low coverage Chromosome 1 deletions longer than 250bp. In addition, we filtered the set of validated deletions by removing cases that overlapped the centromeres or telomeres and removed cases where a reported deletion was marked *LowQual* for all three individuals.

We used the GASV [43] method on this dataset as observations to predict the set of

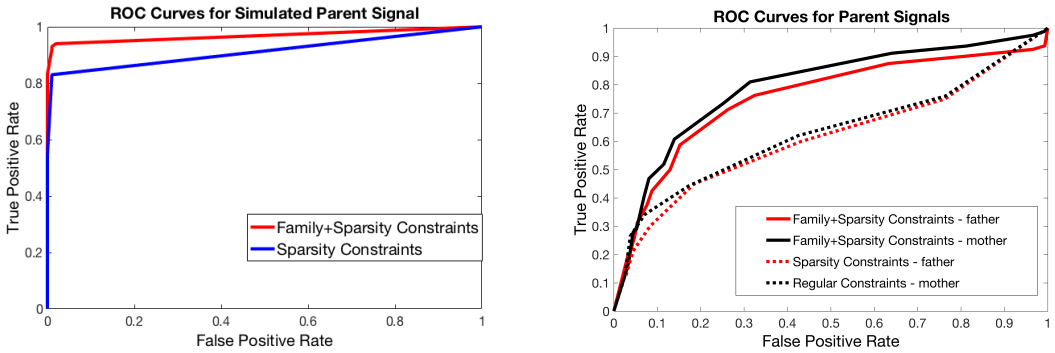

Figure 3.4: (Left). ROC curves depicting the False Positive Rate vs True Positive Rate for the reconstruction of the parent signal with $k_p = 2$, $k_c = 5$, and 70% similarity of variants using both methods with $\tau = 1.553$ for sparsity constraints and $\tau = 1.474$ for family constraints. (Right). ROC curves depicting the False Positive Rate vs True Positive Rate for the reconstruction of both CEU parent Chromosome 1 signals using both methods with $\tau = 2.65$.

possible SVs. We filtered out SVs predicted to lie in the centromere or telomeres. We took the filtered set of predictions as the observed signals, and the true signals for each individual were constructed by determining if the validated deletions lie in the region predicted by GASV.

In the reconstruction of the parent signals, we separately use the child (NA12878) observed signal to constrain the parent signals. As shown in Figure 3.4, the reconstructions of both parent signals improve with the added familial constraints proposed by our method. Since NA12878 shares 90% and 92.5% of deletions with NA12891 (father) and NA12892 (mother), respectively, we observe higher true positive rates for false positive rates > 0.1 in the reconstructions with added child data than the other method.

Conclusions

This work presents a novel approach for inferring structural variants (SVs) from noise-corrupted data sets. We exploit the rare occurrence of SVs by incorporating a sparsity-promoting ℓ_1 penalty regularization term. Furthermore, we mitigate the deleterious effects of low-coverage sequences by following a maximum likelihood approach to SV prediction, and, in particular, using Poisson statistics to model the noise. Finally, we incorporate the relatedness of individuals as a constraint on the solution space. Specifically, we use the assumption that a parent and child are sequenced and require that any SVs predicted in the child be present in the parent. To our knowledge, our proposed approach is the first SV detection algorithm to do so. We demonstrated the effectiveness of our approach on both synthetic data and data from the 1000 Genomes Project.

3.3.2 Haploid Two Parents-One Child Method

The work described in this section is based on the paper by Banuelos et al. [9]. Here we consider a general framework for detecting structural variants (SVs) given sequencing data from two parents (p_1 and p_2) and one child (c). We assume that there are n locations in the genome that could be a potential SV. For simplicity, we consider each individual to be haploid (only one copy of each chromosome). As such, the true SV signal for each individual at each location is either a 0, if they do not have an SV at that location, or 1 if they do. The observed data are the number of DNA fragments supporting each potential SV, $\vec{y}_{p_1}, \vec{y}_{p_2}, \vec{y}_c \in \mathbb{R}^n$ for both parents and the child, and the data are assumed to follow a Poisson distribution [41], $\vec{y}_i \sim \text{Poisson}\left(A_i \vec{f}_i^*\right)$, where $i \in \{p_1, p_2, c\}$ and $A_i = (k_i - \epsilon)\mathbb{I} \in \mathbb{R}^{n\times n}$ is a linear projection of the true genomic variants $\vec{f}_i^* \in \mathbb{R}^n$ to the observation $\vec{y}_i$. The constants k_{p_1}, k_{p_2}, and k_c are the sequencing coverage for each individual, the mean of the Poisson distribution. Finally, we assume that the error in measurement $\epsilon > 0$ is the same for all observations. We stack the true variant signals and observations in the form $\vec{f}^* = [\vec{f}_c^*; \vec{f}_{p_1}^*; \vec{f}_{p_2}^*]$ and $\vec{y} = [\vec{y}_c; \vec{y}_{p_1}; \vec{y}_{p_2}]$, the general observation model is be expressed as

$$\vec{y} \sim \text{Poisson}(\hat{A}\vec{f}^*), \tag{3.12}$$

where $\hat{A} \in \mathbb{R}^{3n\times 3n}$ is a block-diagonal matrix with $\hat{A} = \text{diag}(A_c, A_{p_1}, A_{p_2})$.

Familial constraints. We require the (continuous) elements for each individual lie within 0 and 1, i.e., $0 \le \vec{f} \le 1$. The continuous relaxation of the reconstruction $\vec{f}$ thus allows us to to apply gradient-based techniques. Further, we impose the element-wise constraints that if both parents have the SV, then the child must also, i.e., $\vec{f}_{p_1} + \vec{f}_{p_2} - 1 \le \vec{f}_c$ for each location. Finally, if neither parent has the SV, then the SV cannot be present in the child, i.e., $\vec{f}_c \le \vec{f}_{p_1} + \vec{f}_{p_2}$.

Problem Formulation

Under the Poisson process model (3.12), the probability of observing $\vec{y}$ is given by

$$p(\vec{y}\,|\hat{A}\vec{f}^*) = \prod_{i=1}^{3n} \frac{(\vec{e}_i^T \hat{A}\vec{f}^*)^{\vec{y}_i}}{\vec{y}_i!} \exp\left(-\vec{e}_i^T \hat{A}\vec{f}^*\right), \tag{3.13}$$

where $\vec{e}_i$ is the i-th column of the $3n \times 3n$ identity matrix. The *maximum likelihood principle* is used to determine the unknown Poisson parameter $\hat{A}\vec{f}^*$ such that the probability of observing the vector of Poisson data $\vec{y}$ in (3.13) is maximized. Thus, the genomic variants reconstruction problem has the following constrained optimization form:

$$\begin{aligned} \underset{\vec{f}\in\mathbb{R}^{3n}}{\text{minimize}} \quad & \phi(\vec{f}) \equiv F(\vec{f}) + \tau\,\text{pen}(\vec{f}) \\ \text{subject to} \quad & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \le \vec{f}_c \le \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & 0 \le \vec{f} \le 1 \end{aligned} \tag{3.14}$$

where $F(\vec{f})$ is the negative Poisson log-likelihood function

$$F(\vec{f}) = \mathbf{1}^T\hat{A}\vec{f} - \sum_{i=1}^{3n} \vec{y}_i \log\left(\vec{e}_i^T\hat{A}\vec{f} + \epsilon\right),$$

where $\vec{f} = [\vec{f}_c;\vec{f}_{p_1};\vec{f}_{p_2}]$, $\tau > 0$ is a regularization parameter, $\mathbf{1}$ is a vector of ones, and pen is usually a sparsity enforcing penalty functional.

Since true variants are rare, the penalty functional pen($\vec{f}$) in (3.14) can thus be replaced by sparsity-promoting ℓ_1-norm of $\vec{f}$, i.e., $\|\vec{f}\|_1$. In the SPIRAL framework [22], the solution of (3.14) is obtained by minimizing a sequence of quadratic models to the function $F(\vec{f})$. In these models, the Hessian in the second-order Taylor series expansion of $F(\vec{f})$ at the current iterate $\vec{f}^k$ is replaced by a scaled identity matrix $\alpha_k I$ with $\alpha_k > 0$ (see [11, 12] for details). This quadratic approximation can be simplified to a subproblem of the form:

$$\begin{aligned} \vec{f}^{k+1} = \underset{\vec{f}\in\mathbb{R}^{3n}}{\arg\min} \quad & \tfrac{1}{2}\|\vec{f} - \vec{s}^k\|_2^2 + \tfrac{\tau}{\alpha_k}\|\vec{f}\|_1 \\ \text{subject to} \quad & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \le \vec{f}_c \le \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & 0 \le \vec{f}_c, \vec{f}_{p_1}, \vec{f}_{p_2} \le 1, \end{aligned} \tag{3.15}$$

where $\vec{s}^k = [\vec{s}_c^k; \vec{s}_{p_1}^k; \vec{s}_{p_2}^k] = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$. Then, the subproblem in (3.15) can be separated into scalar minimization problems (see [22] for details).

Completing the squares and ignoring constant terms, we have

$$\begin{aligned} \underset{f_c, f_{p_1}, f_{p_2}\in\mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_c - c)^2 + \tfrac{1}{2}(f_{p_1} - p_1)^2 + \tfrac{1}{2}(f_{p_2} - p_2)^2 \\ \text{subject to} \quad & f_{p_1} + f_{p_2} - 1 \le f_c \le f_{p_1} + f_{p_2}, \\ & 0 \le f_c, f_{p_1}, f_{p_2} \le 1 \end{aligned} \tag{3.16}$$

where $c = s_c - \lambda$, $p_1 = s_{p_1} - \lambda$, and $p_2 = s_{p_2} - \lambda$. The feasible solution to (3.16) is obtained by orthogonally projecting the solution (c, p_1, p_2) to a three-dimensional feasible region (see Fig. 3.5).

In particular, there are 27 regions to be considered in the f_c-f_{p_1}-f_{p_2} tri-dimensional space according to the constraints of (3.16). If (c, p_1, p_2) satisfy the constraints of (3.16), then the minimizer for the subproblem (3.16) is (c, p_1, p_2). Otherwise, the subproblem solution corresponds to the minimizers given in Table 3.2 in Appendix 3.5.

3.3.3 Generalized Haploid Formulation

We extend the framework presented in [9] to detect SVs from one trio to a family lineage involving both parents (p_1 and p_2) and d children ($c_1, c_2, \ldots, c_d$). In particular, we let $\vec{f}_{\mathcal{I}}^* \in [0,1]^n$ be the vector of true genomic variants for individual $\mathcal{I} \in \{p_1, p_2, c_1, c_2, \ldots, c_d\}$. The data for each individual $\mathcal{I}$ are assumed to follow a Poisson distribution (cf. [41, 27]), with $\vec{y}_{\mathcal{I}} \sim \text{Poisson}(A_{\mathcal{I}}\vec{f}_{\mathcal{I}}^*)$, where $A_{\mathcal{I}} = (k_{\mathcal{I}} - \epsilon)\mathbb{I} \in \mathbb{R}^{n\times n}$ is a linear projection matrix that maps the true genomic variants $\vec{f}_{\mathcal{I}}^*$ to the observation $\vec{y}_{\mathcal{I}}$

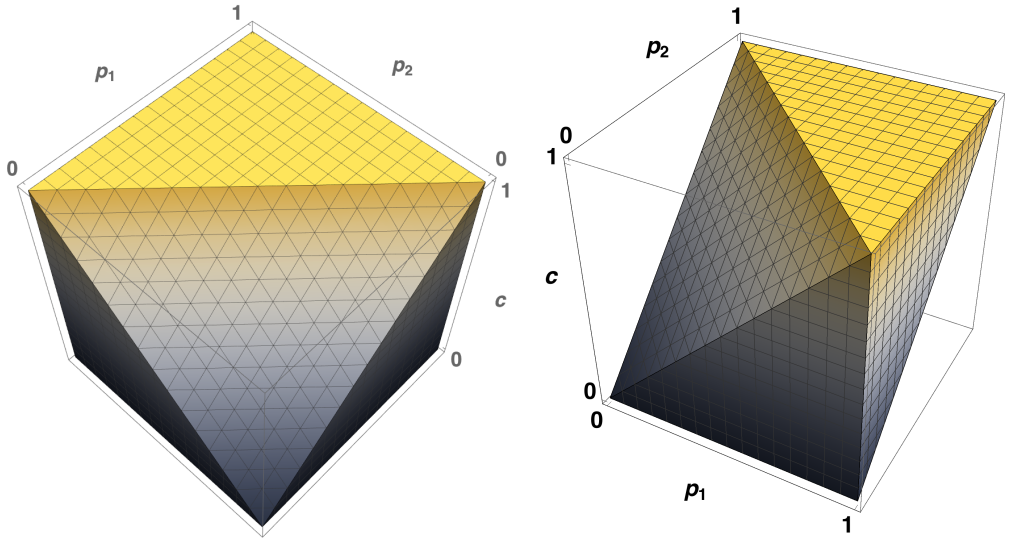

Figure 3.5: The three-dimensional feasible region of the minimization problem (3.16) on the f_c–f_{p_1}–f_{p_2} axis. Subproblem minimizers not satisfying the constraints are projected onto the region. Left : Front view. Right : Side view.

and the constant $k_{\mathcal{T}}$, the mean of the Poisson distribution, is the sequencing coverage for individual $\mathcal{T}$.

More compactly, we let $\vec{f}^* = [\vec{f}^*_{c_1}; \cdots; \vec{f}^*_{c_d}; \vec{f}^*_{p_1}; \vec{f}^*_{p_2}]$, and $\vec{y} = [\vec{y}_{c_1}; \cdots; \vec{y}_{c_d}; \vec{y}_{p_1}; \vec{y}_{p_2}]$. Furthermore, we let $\hat{A} \in \mathbb{R}^{(d+2)n \times (d+2)n}$ be a block-diagonal matrix with $\hat{A} = \text{diag}(A_{c_1}, \cdots, A_{c_d}, A_{p_1}, A_{p_2})$. Thus, the probability of observing $\vec{y}$ is given by

$$p(\vec{y}|\hat{A}\vec{f}^*) = \prod_{i=1}^{(d+2)n} \frac{(\vec{e}_i^T \hat{A}\vec{f}^*)^{\vec{y}_i}}{\vec{y}_i!} \exp\left(-\vec{e}_i^T \hat{A}\vec{f}^*\right), \tag{3.17}$$

where $\vec{e}_i$ is the i-th column of the $(d+2)n \times (d+2)n$ identity matrix. We seek to maximize the probability of observing $\vec{y}$ in (3.17) using a *maximum likelihood* approach in reconstructing $\vec{f}^*$.

Continuous relaxation. The true signal $\vec{f}^*$ is a binary-valued vector, with discrete values 0 and 1. Since maximizing the probability of observing $\vec{y}$ is a combinatorial problem that is generally difficult to solve, we allow for solutions to include vectors with continuous values, i.e., $\vec{f} \in \mathbb{R}^{(d+2)n}$. As a result, we can apply gradient-based optimization techniques in SV detection.

Familial constraints. Since each component of $\vec{f}^*$ is a discrete quantity in $\{0, 1\}$, our continuous approximation $\vec{f}$ must satisfy the element-wise constraint $0 \leq \vec{f} \leq 1$. Moreover, the constraints must also reflect inheritance of a variant in a population. Specifically, If both parents possess a variant, then each child $\vec{f}_{c_i}$, $i \in \{1, 2, \ldots, d\}$, must inherit it as well, i.e., $\vec{f}_{p_1} + \vec{f}_{p_2} - 1 \leq \vec{f}_{c_i}$. Likewise, if neither parent has the SV, it cannot be present in either child, i.e., $\vec{f}_{c_i} \leq \vec{f}_{p_1} + \vec{f}_{p_2}$.

Gradient-based optimization. Under the maximum likelihood model, we can formulate

the SV detection problem as the following constrained optimization problem:

$$\begin{aligned} \underset{\vec{f}\in\mathbb{R}^{(d+2)n}}{\text{minimize}} \quad & \phi(\vec{f}) \equiv F(\vec{f}) + \tau\,\text{pen}(\vec{f}) \\ \text{subject to} \quad & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \le \vec{f}_{c_1} \le \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \le \vec{f}_{c_2} \le \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \le \vec{f}_{c_d} \le \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & 0 \le \vec{f} \le 1 \end{aligned} \tag{3.18}$$

where $F(\vec{f})$ is the negative Poisson log-likelihood function

$$F(\vec{f}) = \mathbf{1}^T\hat{A}\vec{f} - \sum_{i=1}^{(d+2)n} \vec{y}_i \log\left(\vec{e}_i^T\hat{A}\vec{f} + \epsilon\right),$$

and $\tau > 0$ is a regularization parameter, $\mathbf{1}$ is a vector of ones, and pen is usually a sparsity enforcing penalty functional. In our case, we use the widely-used sparsity promoting ℓ_1 penalty term.

Our proposed approach for solving the optimization problem (3.18) builds upon the SPIRAL framework [22] by incorporating a more complex feasible domain. As in previously defined methods, we use a sequence of quadratic subproblems from the second-order Taylor series approximations to $F(\vec{f})$ at the current iterate $\vec{f}^k$ and approximate the Hessian matrix by a scalar multiple of the identity matrix, $\alpha_k I$. The resulting quadratic subproblem is separable in each SV location, meaning its minimizer can be obtained by solving the following optimization problem:

$$\begin{aligned} \underset{f_{c_1},\ldots,f_{c_d},f_{p_1},f_{p_2}\in\mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}\sum_{i=1}^{2}(f_{c_i} - \mu_{c_i})^2 + \tfrac{1}{2}\sum_{i=1}^{d}(f_{p_i} - \mu_{p_i})^2 \\ \text{subject to} \quad & f_{p_1} + f_{p_2} - 1 \le f_{c_1} \le f_{p_1} + f_{p_2}, \\ & \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ & f_{p_1} + f_{p_2} - 1 \le f_{c_d} \le f_{p_1} + f_{p_2}, \\ & 0 \le f_{c_1},\ldots,f_{c_d},f_{p_1},f_{p_2} \le 1 \end{aligned} \tag{3.19}$$

where $(\mu_{c_1},\ldots,\mu_{c_d},\mu_{p_1},\mu_{p_2}) = (s_{c_1} - \frac{\tau}{\alpha_k},\ldots,s_{c_d} - \frac{\tau}{\alpha_k},s_{p_1} - \frac{\tau}{\alpha_k},s_{p_2} - \frac{\tau}{\alpha_k})$. If $(f_{c_1},\ldots,f_{c_d},f_{p_1},f_{p_2}) = (\mu_{c_1},\ldots,\mu_{c_d},\mu_{p_1},\mu_{p_2})$ is feasible, then it is the minimizer of (3.19). Otherwise, we use an alternating iterative method based on block-coordinate descent to calculate $\vec{f}^{k+1}$. We now describe the method in more detail for a family quartet.

Alternating minimization. At each iteration k, we alternate between fixing all but one child signal $f_{c_i}, i \in \{1,2,\ldots,d\}$ and solve the resulting minimization subproblem:
Step 1: Initially, we fix $\hat{f}_{c_2},\ldots,\hat{f}_{c_d}$. Then, we solve

$$\begin{aligned} \underset{f_{c_1},f_{p_1},f_{p_2}\in\mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_{c_1} - \mu_{c_1})^2 + \tfrac{1}{2}\sum_{i=1}^{2}(f_{p_i} - \mu_{p_i})^2 \\ \text{subject to} \quad & f_{p_1} + f_{p_2} - 1 \le f_{c_1} \le f_{p_1} + f_{p_2}, \\ & 0 \le f_{c_1},f_{p_1},f_{p_2} \le 1 \end{aligned} \tag{3.20}$$

which can be solved using orthogonal projections described in Section 3.3.2.

Step 2: Suppose we have obtained $\hat{f}_{c_1}, \hat{f}_{p_1}, \hat{f}_{p_2}$ from Step 1. For $f_{c_i},\ i \in \{2, \dots, d\}$, on the i^{th} child, we fix all child signals except $\hat{f}_i$ and solve

$$\begin{aligned} \underset{f_{c_i} f_{p_1} f_{p_2} \in \mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_{c_i} - \mu_{c_i})^2 + \tfrac{1}{2}\sum_{i=1}^{2}(f_{p_i} - \hat{f}_{p_i})^2 \\ \text{subject to} \quad & f_{p_1} + f_{p_2} - 1 \le f_{c_i} \le f_{p_1} + f_{p_2}, \\ & 0 \le f_{c_i} \le 1 \end{aligned} \tag{3.21}$$

which can be similarly solved in Step 1 using orthogonal projections. We separate Step 2 from Step 1 since f_{p_1} and f_{p_2} satisfy the boundedness constraints. This results in a reduction of the 27 regions considered in the initial step. Steps 1 and 2 are repeated at each iterate k until the relative difference between subsequent iterates are below a set threshold. To avoid biases, Step 1 should rotate fixing the initial child signal from f_{c_1} to f_{c_d}.

3.3.4 Nonconvex Regularization of Haploid Models

The work described in this section is based on the paper by Banuelos et al. [7]. For the previous models presented, we have focused on incorporating the ℓ_1 sparsity-promoting penalty functional. We now consider nonconvex formulations of the results presented in [10, 9, 6] for detecting structural variants (SVs) from sequencing data from one two-parent one-child trio (p_1, p_2, c, respectively). Although we focus on the trio optimization problem in this section, the results may be extended to the generalized haploid formulation presented in Section 3.3.3. To reflect the rarity of SVs in the genome, we incorporate a sparsity-promoting penalty in our problem formulation. Previous methods [10, 9, 6] use the widely used ℓ_1 regularization term [46], which is a convex relaxation of the ℓ_0 counting semi-norm. To promote further sparsity and thus decrease the false positives classified as true variants, we use a nonconvex q-norm penalty functional, $\|\vec{f}\|_q^q = \sum_{i=1}^{3n} |f_i|^q$, where $0 < q < 1$ (see e.g., [1]). As $q \to 0$, we expect to identify the true support of $\vec{f}^*$ more accurately and, thus, capture reconstructions closer to the true signal $\vec{f}^*$.

Optimization Approach

Our proposed approach for solving the optimization problem (3.14) is based on the SPIRAL framework [22]. As before, these quadratic subproblems can be simplified to the following form:

$$\begin{aligned} \vec{f}^{k+1} = \ \underset{\vec{f} \in \mathbb{R}^{3n}}{\arg\min} \quad & \tfrac{1}{2}\|\vec{f} - \vec{s}^{\,k}\|_2^2 + \lambda \|\vec{f}\|_q^q \\ \text{subject to} \quad & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \le \vec{f}_c \le \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & 0 \le \vec{f}_c, \vec{f}_{p_1}, \vec{f}_{p_2} \le 1, \end{aligned} \tag{3.22}$$

where $\vec{s}^k = [\vec{s}_c^k; \vec{s}_{p_1}^k; \vec{s}_{p_2}^k] = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$ and $\lambda = \frac{\tau}{\alpha_k}$. Note that the objective function is separable in $\vec{f}$. Since the constraints are component-wise bounds on the variables, (3.22) decouples into n quadratic subproblems of the form:

$$\begin{aligned} f^{k+1} = \underset{f\in\mathbb{R}^3}{\arg\min} \quad & \mathcal{Q}(f) = \tfrac{1}{2}\|f - s^k\|_2^2 + \lambda \sum_{i\in\{c,p_1,p_2\}} |f_i|^q \\ \text{subject to} \quad & f_{p_1} + f_{p_2} - 1 \le f_c \le f_{p_1} + f_{p_2}, \\ & 0 \le f_c, f_{p_1}, f_{p_2} \le 1, \end{aligned} \tag{3.23}$$

where $f = [f_c; f_{p_1}; f_{p_2}]$ and s^k correspond to components of $\vec{f}$ and $\vec{s}^k$, respectively. Each of these decoupled subproblems now depends only on the three scalar variables, f_c, f_{p_1}, and f_{p_2}. However, even without the constraints, the subproblem (3.23) does not have a closed form solution. Thus, we make an additional approximation, where we use the first-order Taylor series expansion of the penalty term:

$$|f_i|^q \approx T_1(f_i) = |f_i^k|^q + q|f_i^k|^{q-1}(f_i - f_i^k), \tag{3.24}$$

where $i \in \{c, p_1, p_2\}$ (see Fig. 3.6). We investigated using a second-order Taylor expansion

$$T_2(f_i) = |f_i^k|^q + q|f_i^k|^{q-1}(f_i - f_i^k) + \tfrac{1}{2}q(q-1)|f_i^k|^{q-2}(f_i - f_i^k)^2.$$

However, this approximation complicated later calculations and did not improve the performance of our proposed algorithm, and so we used the simpler $T_1(f)$ approximation. At the first iteration, we expand around $\frac{1}{2}$, and for subsequent iterates, this approximation is centered at the previous iterate f^k. Substituting (3.24) in (3.22), completing the squares, and ignoring constant terms, we have

$$\begin{aligned} \underset{f_c, f_{p_1}, f_{p_2}\in\mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_c - \sigma_c)^2 + \tfrac{1}{2}(f_{p_1} - \sigma_{p_1})^2 + \tfrac{1}{2}(f_{p_2} - \sigma_{p_2})^2 \\ \text{subject to} \quad & f_{p_1} + f_{p_2} - 1 \le f_c \le f_{p_1} + f_{p_2}, \\ & 0 \le f_c, f_{p_1}, f_{p_2} \le 1 \end{aligned} \tag{3.25}$$

where

$$\sigma_c = s_c^k - \lambda q r_1^{q-1}; \quad \sigma_{p_1} = s_{p_1}^k - \lambda q r_2^{q-1}; \quad \text{and} \quad \sigma_{p_2} = s_{p_2}^k - \lambda q r_3^{q-1};$$

and $r_1 = r_2 = r_3 = \frac{1}{2}$ initially and $[r_1; r_2; r_3] = [f_c^k, f_{p_1}^k, f_{p_2}^k]$ at subsequent iterates k.

The solution to (3.25) can be obtained in the following way. The level sets of the objective function are isotropic. Thus, to compute the solution, we simply have to project the unconstrained minimizer (c, p_1, p_2) onto the truncated cube defined by the constraints in (3.25). The projection can be done by partitioning the three-dimensional c-p_1-p_2 space into 27 subregions where if (c, p_1, p_2) satisfies the constraints, it is the constrained minimizer, and otherwise, it is projected onto a vertex, an edge, or a face of the three-dimensional feasible region. For further details on how this can be accomplished, see Section 3.3.2.

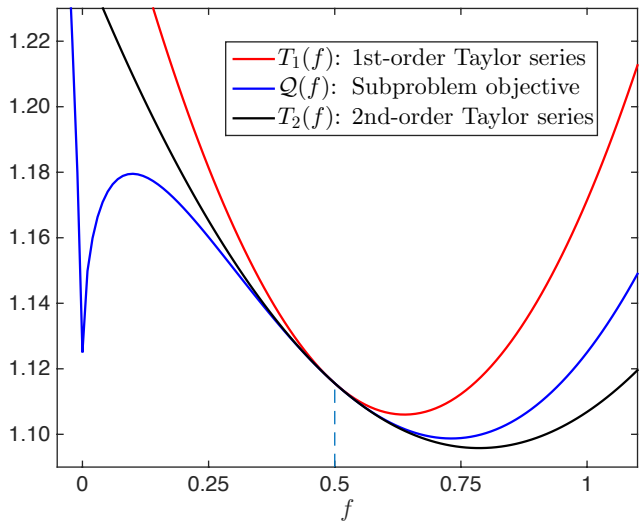

Figure 3.6: Plots of the subproblem objective function $\mathcal{Q}(f)$ from (3.23) with q-norm approximations using first- and second-order Taylor series expansions, $T_1(f)$ and $T_2(f)$, respectively, centered around $f = 0.5$.

Related methods for Poisson reconstruction

We note that other methods exist for recovering signals from data corrupted by Poisson noise, most notably from the medical and astronomy imaging communities. For example, in [31, 30], Nowak and Kolaczyk describe a multiscale Bayesian approach in an Expectation-Maximization (EM) framework for Poisson reconstruction. Proximal functions can also be applied to solve these functions [32, 14] as well as split Bregman approaches [18, 17, 38]. However, these methods use convex penalty terms such as the ℓ_1 norm or the total variations norm [36]. Furthermore, it is not clear how non-trivial constraints can be incorporated into these approaches. In contrast, our current framework allows us to impose constraints on the computed solution without significant computational overhead costs.

3.3.5 Diploid One Parent-One Child Method

For the proceeding models, we focus on haploid individuals (detecting one copy of a variant). However, humans are diploid (two copies inherited from parents) organisms and as such, we shift our focus to detecting the number of structural variants in a group of related individuals. This work presented in this section is based on the paper by Banuelos et al. [8]. Specifically, we describe mathematically our computational framework for predicting the number of copies an individual carries of each SV for one diploid parent and child.

Problem formulation. First, let n be the length of the vector of genetic variants for every individual. At each location i ($1 \leq i \leq n$), we let $z_\sigma^{(i)}$ and $y_\sigma^{(i)}$ be indicator variables for

individual σ such that

$$\begin{aligned} z_\sigma^{(i)} &= 1 \quad \text{if individual } \sigma \text{ has two copies of an SV} \\ y_\sigma^{(i)} &= 1 \quad \text{if individual } \sigma \text{ has one copy of an SV.} \end{aligned}$$

If the individual has no copies of the SV at location i, then $z_\sigma^{(i)} = y_\sigma^{(i)} = 0$. Thus, the observation, $s_\sigma^{(i)}$, at the i^{th} location for SV copies, is modeled as

$$s_\sigma^{(i)} \sim \text{Poisson}\left(z_\sigma^{(i)}(2k_\sigma - \epsilon) + y_\sigma^{(i)}(k_\sigma - \epsilon) + \epsilon\right),$$

where k_σ is the sequencing coverage for individual σ and ϵ is the measurement error. We can write this compactly as follows: Let

$$\vec{z}_p = \begin{bmatrix} z_p^{(1)} \\ \vdots \\ z_p^{(n)} \end{bmatrix}, \ \vec{z}_c = \begin{bmatrix} z_c^{(1)} \\ \vdots \\ z_c^{(n)} \end{bmatrix}, \ \vec{y}_p = \begin{bmatrix} y_p^{(1)} \\ \vdots \\ y_p^{(n)} \end{bmatrix}, \text{ and } \vec{y}_c = \begin{bmatrix} y_c^{(1)} \\ \vdots \\ y_c^{(n)} \end{bmatrix}.$$

Let $\vec{z} = [\vec{z}_p; \vec{z}_c] \in \Re^{2n}$ and $\vec{y} = [\vec{y}_p; \vec{y}_c] \in \Re^{2n}$. Define $\vec{f} = [\vec{z}; \vec{y}] \in \Re^{4n}$. Now let $A_2 = (2k_j - \epsilon)I_{2n}$ and $A_1 = (k_j - \epsilon)I_{2n}$, where I_{2n} is the $2n \times 2n$ identity matrix. Define $A = [\, A_2 \ \ A_1\,] \in \Re^{2n \times 4n}$. Thus, the vector of observations is modeled by

$$\vec{s} \sim \text{Poisson}(A\vec{f}). \tag{3.26}$$

where is the linear projection of true heterozygous and homozygous SVs onto our observed signal $\vec{s}$.

Continuous Relaxation and Constraints For large n, the solution space for inferring $\vec{f}$ from $\vec{s}$ is exponentially large since $f \in \{0, 1\}^{4n}$. Thus, rather than solving this problem combinatorially, we relax our problem formulation to *continuous* variables so that we can apply calculus of variations approaches. In particular, we apply a gradient-based maximum likelihood approach to recover the true indicator variables z_σ and y_σ where $\sigma = p$ if individual σ is the parent and $\sigma = c$ if σ is the child. Since z_σ and y_σ are either 0 or 1, we allow for solutions in the interval $[0, 1]$, i.e., $0 \leq z_\sigma, y_\sigma \leq 1$. Moreover, since an individual can only have 0, 1, or 2 copies, we enforce the following constraint $0 \leq z_\sigma + y_\sigma \leq 1$. To incorporate relatedness of individuals, we assume a child cannot have two copies of the SV if one parent does not have at least one copy of the SV (since *de-novo* mutations are rare), i.e., $0 \leq z_c \leq z_p + y_p$. Additionally, if the parent has two copies of the SV (i.e., $z_p = 1$), then the child must have at least one copy of the SV, i.e., $z_p \leq z_c + y_c$. Thus, we define our feasible set as

$$\mathscr{F} = \left\{ \vec{f} = [\vec{z}; \vec{y}] \in \Re^{4n} : \begin{array}{c} 0 \leq \vec{y}_p, \vec{y}_c \leq \mathbf{1}, \\ 0 \leq \vec{z}_c \leq \vec{z}_p + \vec{y}_p \leq \mathbf{1}, \\ 0 \leq \vec{z}_p \leq \vec{z}_c + \vec{y}_c \leq \mathbf{1} \end{array} \right\},$$

where $\mathbf{1}$ is a vector of ones.

Optimization Formulation and Approach

Under the model (3.26), the probability of observing $\vec{s}$ is

$$p(\vec{s}\,|A\vec{f}^*) = \prod_{i=1}^{4n} \frac{((A\vec{f}^*)_i)^{\vec{s}_i}}{\vec{s}_i!} \exp\left(-(A\vec{f}^*)_i\right). \tag{3.27}$$

Following a maximum likelihood approach, reconstructing genomic variants has the following constrained optimization form:

$$\underset{\vec{f}\in\mathcal{F}}{\text{minimize}} \quad \phi(\vec{f}) \equiv F(\vec{f}) + \tau\|\vec{f}\|_1, \tag{3.28}$$

where $F(\vec{f})$ is the negative Poisson log-likelihood function

$$F(\vec{f}) = \sum_{j=1}^{2n} (A\vec{f})_j - \sum_{j=1}^{4n} \vec{s}_j \log\left((A\vec{f})_j + \epsilon\right),$$

$\tau > 0$ is a regularization parameter, $\|\vec{f}\|_1 = \sum_{j=1}^{4n} |f_j|$ added to promote sparsity in the solution, and $0 < \epsilon \ll 1$ is a small parameter introduced to avoid the singularity at $\vec{f} = 0$. As before, we solve a sequence of quadratic approximations to $F(\vec{f})$. Further, these quadratic subproblems can be simplified to the following form:

$$\vec{f}^{k+1} = \underset{\vec{f}\in\mathcal{F}}{\arg\min} \ \tfrac{1}{2}\|\vec{f} - \vec{r}^{\,k}\|_2^2 + \lambda\|\vec{f}\|_1, \tag{3.29}$$

where $\vec{r}^{\,k} = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$ and $\lambda = \frac{\tau}{\alpha_k}$. Then, the subproblem in (3.29) can be separated into scalar minimization problems (see [22] for details). Note that the objective function is separable on the variables $\vec{f}$. Thus (3.29) decouples into n four-dimensional problems of the form

$$\begin{aligned} f^{k+1} = \underset{f=[z_p;z_c;y_p;y_c]\in\mathbb{R}^4}{\arg\min} \ & \tfrac{1}{2}\|f - r^k\|_2^2 + \lambda\|f\|_1 \\ \text{subject to} \quad & 0 \le y_p, y_c \le 1 \\ & 0 \le z_c \le z_p + y_p \le 1 \\ & 0 \le z_p \le z_c + y_c \le 1, \end{aligned} \tag{3.30}$$

where $r^k = [r^k_{z_p}; r^k_{z_c}; r^k_{y_p}; r^k_{y_c}]$ and $f = [z_p; z_c; y_p; y_c]$ correspond to components of $\vec{r}^{\,k}$ and $\vec{f}$, respectively.

In this work, we propose solving (3.30) using two methods, both of which are based on block-coordinate descent approaches, which work as follows. Method I first fixes the homozygous indicator variables, z_σ, and minimizes over the heterozygous indicator variables, y_σ. In the next step, the heterozygous variables are fixed, and (3.30) is minimized over the homozygous variables. In contrast, Method II fixes all but one individual and minimizes (3.30) over the indicator variables for that particular individual. In subsequent steps, the variables corresponding to some other individual are minimized while fixing the variables for all other individuals. Both methods continue this

block-coordinate descent approach until the iterates satisfy a pre-determined convergence criteria. We now describe each method in more detail.

METHOD I: This method solves (3.30) by alternating between homozygous and heterozygous indicator variables. In particular, it consists of the following steps.

Step 0: Initially, we fix the values for the homozygous indicator variables by setting $z_p^{(0)} = z_c^{(0)} = 0.5$ for each candidate SV location. We then proceed to Step 1.

Step 1: Suppose we have obtained $\hat{z}_p^{(i-1)}$ and $\hat{z}_c^{(i-1)}$ from the previous iteration. Then to obtain the solution for the current iteration $\hat{y}_p^{(i)}$ and $\hat{y}_c^{(i)}$, we solve

$$\begin{aligned}(\hat{y}_p^{(i)}, \hat{y}_c^{(i)}) \;=\; \underset{y_p, y_c \in \mathbb{R}}{\arg\min} \;\; & \tfrac{1}{2}(y_p - a_2)^2 + \tfrac{1}{2}(y_c - b_2)^2 \\ \text{subject to} \quad & 0 \le y_p, y_c \le 1 \\ & \hat{z}_c^{(i-1)} - \hat{z}_p^{(i-1)} \le y_p \le 1 - \hat{z}_p^{(i-1)} \\ & \hat{z}_p^{(i-1)} - \hat{z}_c^{(i-1)} \le y_c \le 1 - \hat{z}_c^{(i-1)},\end{aligned} \tag{3.31}$$

where $a_2 = r_{y_p}^k - \lambda$ and $b_2 = r_{y_c}^k - \lambda$. Note that the bounds on y_p and y_c are simple bounds. Thus the feasible region is a simple rectangle (see Fig. 3.7).

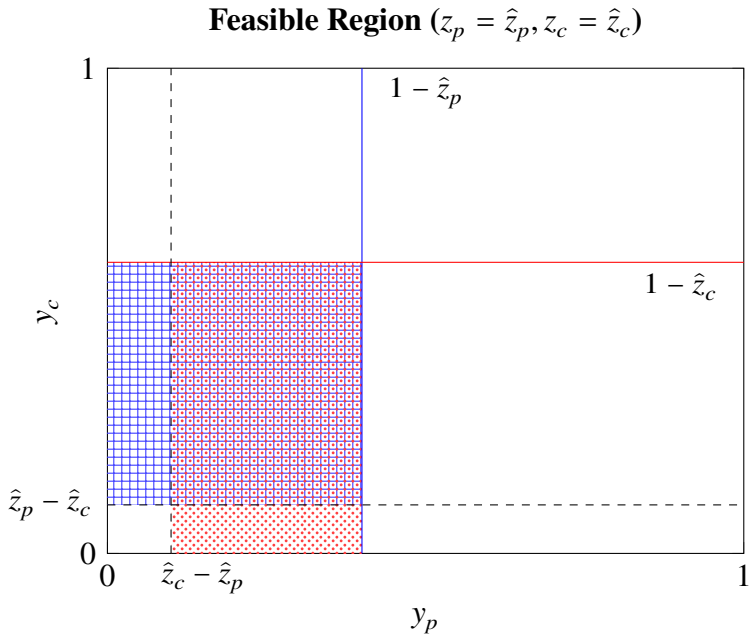

Figure 3.7: Feasible region corresponding to the constraints in (3.31). The blue region represents the admissible set of solutions when z_p - z_c $\geq$ 0 and the red region represents the feasible region when $z_p - z_c < 0$.

Step 2: Suppose we have obtained $\hat{y}_p^{(i)}$ and $\hat{y}_c^{(i)}$ from Step 1. Then to obtain the solution

for the current iteration $\hat{z}_p^{(i)}$ and $\hat{z}_c^{(i)}$, we complete the square and solve

$$\begin{aligned}(\hat{z}_p^{(i)}, \hat{z}_c^{(i)}) = \underset{z_p, z_c \in \mathbb{R}}{\arg\min} \quad & \tfrac{1}{2}(z_p - a_1)^2 + \tfrac{1}{2}(z_c - b_1)^2 \\ \text{subject to } & z_p - \hat{y}_c^{(i)} \le z_c \le z_p + \hat{y}_p^{(i)}, \\ & 0 \le z_p \le 1 - \hat{y}_p^{(i)},\ 0 \le z_c \le 1 - \hat{y}_c^{(i)}.\end{aligned} \tag{3.32}$$

The feasible region is shown in Fig. 3.8.

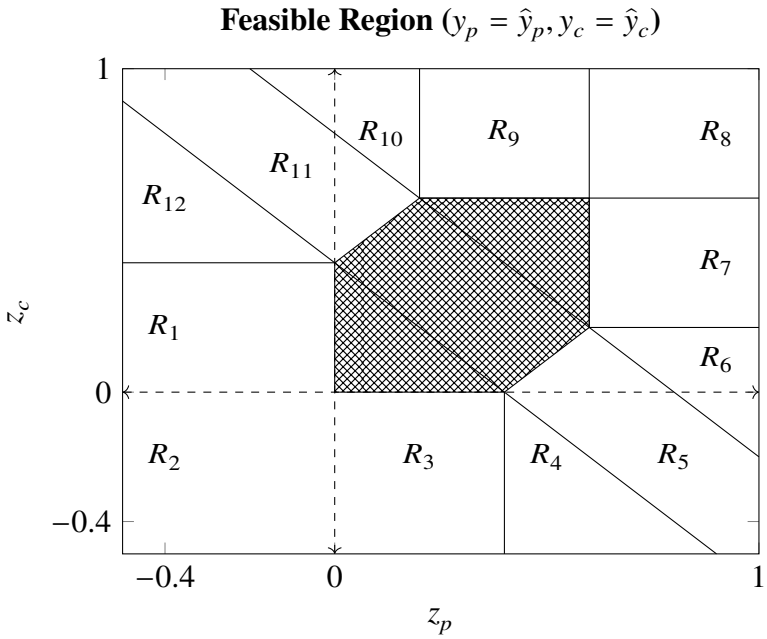

Figure 3.8: Feasible region obtained from applying the constraints to (3.32), where the shaded grid region represents the admissible set of solutions when $\hat{y}_p \le 1 - \hat{y}_c$. If this condition is not satisfied, then we project onto the rectangular region obtained when $\hat{y}_p = 1 - \hat{y}_c$.

Solution. Note that both problems (3.31) and (3.32) have closed form solutions since the level sets of the objective functions in both problems are isotropic, and thus the minimizer can be easily obtained by projecting the unconstrained solution to the feasible set. In particular, the solution to Step 1 is given by

$$\begin{aligned}\hat{y}_p^{(i)} &= \text{mid}\,\{1 - \hat{z}_p^{(i-1)}, a_2, \max\{0, \hat{z}_c^{(i-1)} - \hat{z}_p^{(i-1)}\}\} \\ \hat{y}_c^{(i)} &= \text{mid}\,\{1 - \hat{z}_c^{(i-1)}, b_2, \max\{0, \hat{z}_p^{(i-1)} - \hat{z}_c^{(i-1)}\}\},\end{aligned}$$

where the operator $\text{mid}\{a, b, c\}$ chooses the middle value of the three arguments. The solution to Step 2 can be found in Table 3.4, where the projection, $(z_p^{(i)}, z_c^{(i)})$, of the unconstrained solution (a_1, b_1) is explicitly computed.

METHOD II: This method solves (3.30) by alternating between individuals. In particular, it consists of the following steps.

Step 0: Initially, we fix the values for the parent. Let $f^k = [z_p^k; z_c^k; y_p^k; y_c^k]$ be the elements of $\vec{f}^k$ corresponding to the variables $f = [z_p; z_c; y_p; y_c]$. To initialize z_p and y_p, we apply the following rule:

$$\begin{aligned} \hat{z}_p^{(0)} &= \text{mid}\,\{0, r_{z_p}^k - \lambda, 1\}, \\ \hat{y}_p^{(0)} &= \text{mid}\,\{0, r_{y_p}^k - \lambda, 1\}. \end{aligned}$$

Thus, for each candidate SV location, our initialization is consistent with the set of feasible solutions with the intent of reducing false-positives in our model.

Step 1: Once we have obtained estimates for the parent's diploid indicator variables, $\hat{z}_p^{(i-1)}$ and $\hat{y}_p^{(i-1)}$, from the previous iteration, we obtain the solution for the the child's diploid indicator variable, $\hat{z}_c^{(i)}$ and $\hat{y}_c^{(i)}$, by solving

$$\begin{aligned} (\hat{z}_c^{(i)}, \hat{y}_c^{(i)}) = \underset{z_c, y_c \in \mathbb{R}}{\arg\min} \quad & \tfrac{1}{2}(z_c - b_1)^2 + \tfrac{1}{2}(y_c - b_2)^2 \\ \text{subject to} \quad & 0 \le z_c, y_c \le 1 \\ & z_c \le \hat{z}_p^{(i-1)} + \hat{y}_p^{(i-1)} \le 1 \\ & \hat{z}_p^{(i-1)} \le z_c + y_c \le 1, \end{aligned} \tag{3.33}$$

where $b_1 = r_{z_c}^k - \lambda$ and $b_2 = r_{y_c}^k - \lambda$.

Step 2: Once we have obtained estimates for the child's diploid indicator variables, $\hat{z}_c^{(i)}$ and $\hat{y}_c^{(i)}$ from Step 1, we solve

$$\begin{aligned} (\hat{z}_p^{(i)}, \hat{y}_p^{(i)}) = \underset{z_p, y_p \in \mathbb{R}}{\arg\min} \quad & \tfrac{1}{2}(z_p - a_1)^2 + \tfrac{1}{2}(y_p - a_2)^2 \\ \text{subject to} \quad & 0 \le z_p, y_p \le 1 \\ & z_p \le \hat{z}_c^{(i-1)} + \hat{y}_c^{(i-1)} \le 1 \\ & \hat{z}_c^{(i-1)} \le z_p + y_p \le 1, \end{aligned} \tag{3.34}$$

to obtain the solution for the parent's diploid indicator variable, $\hat{z}_p^{(i)}$ and $\hat{y}_p^{(i)}$, where $a_1 = r_{z_p}^k - \lambda$ and $a_2 = r_{y_p}^k - \lambda$.

Steps 1 and 2 are repeated alternatingly until some convergence criteria are satisfied.

We note that the constraints Steps 1 and 2 are equivalent. Thus, the feasible region corresponding to the constraints are the same. For example, the feasible region for Step 1 is shown in Fig. 3.9.

Solution. The objective functions in both subproblems (3.33) and (3.34) are quadratic functions with the identity matrix as second-derivatives. Thus, the level curves of the objective functions are concentric circles centered around (b_1, b_2) and (a_1, a_2), respectively. This implies that the minimizer of the constrained subproblems are the orthogonal projections of (b_1, b_2) and (a_1, a_2) onto the feasible regional. Therefore, each subproblem has a closed-form solution and can be thus solved efficiently. In particular, for Step 1, the constrained solution is the projection of the unconstrained solution (b_1, b_2) onto the feasible set and can be found in Table 3.5. The constrained solution to Step 2 is similarly defined.

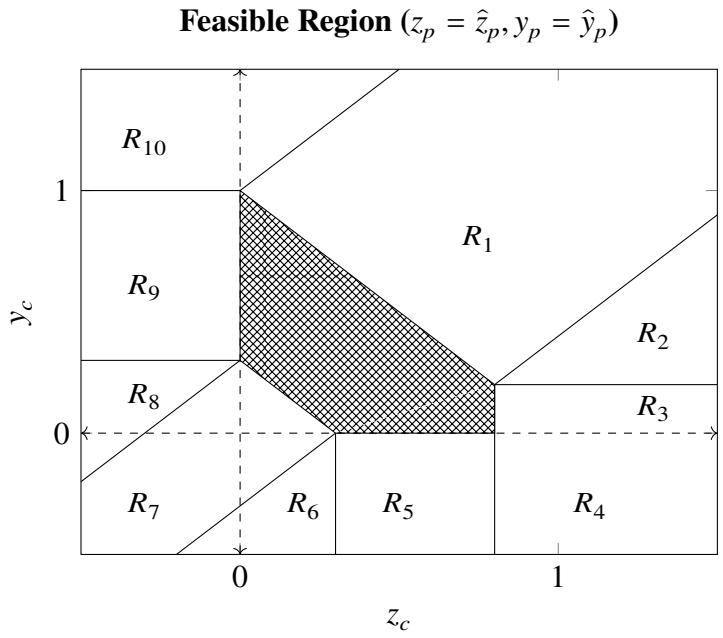

Figure 3.9: Feasible region corresponding to the constraints, where the shaded grid region represents the admissible set of solutions.

Numerical Results

For the experiments with the simulated data, a child signal was generated from two parent signals, which shared a percent similarity (e.g. if both parents were homozygous for a variant, then the child would also be homozygous and so on) ranging from 60% to 100% in 20% increments. However, for the purpose of testing the proposed approach, only one parent signal was used. Furthermore, we considered 0.02, 0.08, and 0.16 error term ϵ in obtaining measurements from the forward model.

We first examine the parent signal reconstruction. The false positive rate vs. true positive rate for the reconstruction of the heterozygous parent signal with coverages $k_p = 4, k_c = 4$, 80% similarity of variants between parents, and an error level of $\epsilon = 0.16$ is presented in Fig. 3.10. We note that Method II improves SV detection when compared to Method I (see Fig. 3.10). Based on AUC measurements, we observe that for the parent signal (both homozygous and heterozygous indicator variables) there was an improvement in reconstructions with the fix one individual method as we decrease the error level. Furthermore, we observed a higher accuracy for the homozygous parent signal reconstructions as we increase the percentage similarity between the parents. However, this pattern was the opposite for the heterozygous parent signal.

Next, we examined the child signal reconstruction. Fig. 3.10 illustrates the false positive rate vs. true positive rate for the reconstruction of the homozygous child signal with coverages $k_p = 4, k_c = 4$, 80% similarity of variants between parents, and an error level of $\epsilon = 0.08$. For false positive rate values > 0.10 and true positive rate values < 0.90, no significant difference could be discerned. Accordingly, the axes were reduced

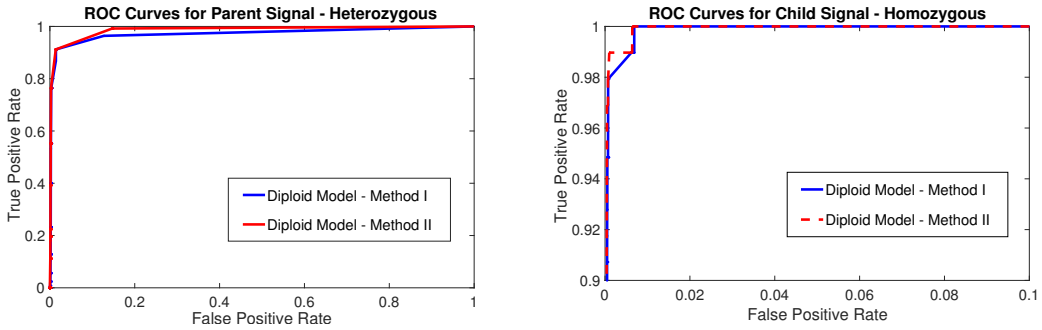

Figure 3.10: (Left) ROC curves illustrating the false positive rate vs true positive rate for the reconstruction of the heterozygous parent signal in the simulated data with k_p=4, k_c=4, and 80% similarity of variants between parents using both methods with $\epsilon = 0.16$. (Right) ROC curves illustrating the false positive rate vs true positive rate for the reconstruction of the homozygous child signal in the simulated data with $k_p = 4$, $k_c = 4$, and 80% similarity of variants between parents using both methods with $\epsilon = 0.08$.

in order to provide a more detailed view of the comparison between the two methods. In this case, we notice both methods perform similarly (see Fig. 3.10). Based on AUC measurements, we observed that the homozygous child signal reconstructions improved as we increase the percentage similarity between the parents. We did not observe this type of improvement in signal reconstruction for the child signal when varying the error level.

Next, we apply our proposed methods to 1000 Genomes Project [2] father-mother-daughter CEU trio data (NA12891, NA12892, NA12878). The genomes in Pilot 1 were aligned to NCBI36 and sequenced at approximately 4× coverage. Experimentally validated (reported) deletions longer than 250bp are taken as the true signal. We use the reported genotype, unless marked with *LowQual*, to determine whether the reported deletion was either heterozygous or homozygous. After applying this filtering, we create the vectors $\vec{z}, \vec{y}$, representing the indicator variables for the genotype at each location.

The number of candidate deletion locations is $n = 57,078$ for each CEU genome. The total number of deletions, both heterozygous and homozygous, were 686, 637, and 724 for the father, mother, and child, respectively. In the 1000 Genomes data and the simulated data, we observe similar improvement trends in our proposed method for both parent heterozygous signals and child homozygous signals. Since the experimentally validated set of deletions may not be complete, we compare the number of predicted heterozygous novel deletions to validated SVs in Fig. 3.11. Further, if we consider the broader question of correctly identifying a deletion, regardless of genotype, we achieve similar results to the original proposed method (see Fig. 3.11). These similar results are achieved at a lower computational cost in a more general framework. We report the improved computational cost of our method in comparison to the original method in Fig. 3.12.

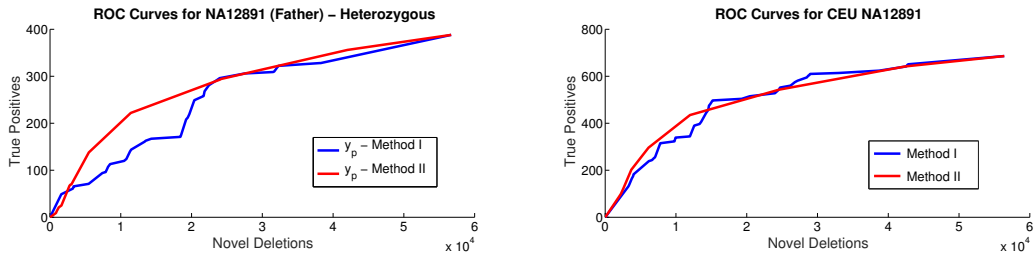

Figure 3.11: (Left) ROC curves depicting novel deletions vs true positives for the reconstruction of heterozygous CEU NA12891 (father) signal. (Right) ROC curves depicting novel deletions vs true positives for the reconstruction the combined heterozygous and homozygous CEU NA12891 (father) signal. In both, $k_{p_1} = 4$, $k_c = 4$, with $\tau = 2.34 \times 10^{-10}$ and $\epsilon = 0.01$.

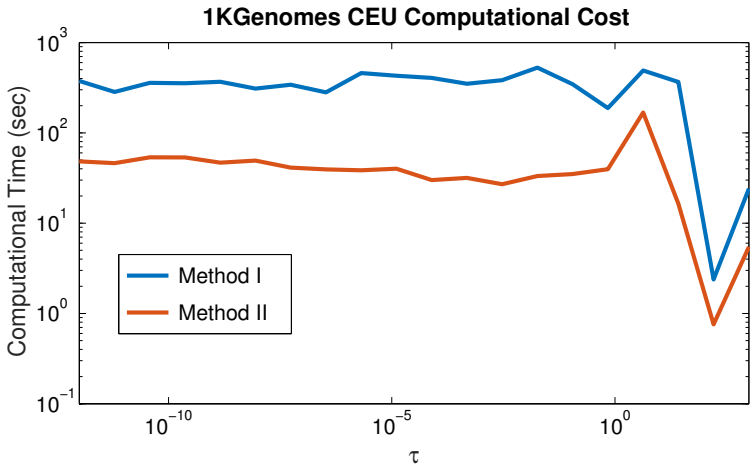

Figure 3.12: Given the heterozygous and homozygous observations $\vec{s}$, we plot the computational time (in seconds) in reconstructing the true signal for the CEU dataset (NA12891 and NA12878). We observe a general reduction of computational cost for Method II for a range of penalty values τ.

Conclusions

We present a generalized approach for detecting both structural variants (SVs) and their genotype (heterozygous or homozygous) from low coverage DNA sequencing data. While it is possible that our methods could be adopted to cancer studies, this is outside the scope of the current study. We enforce sparsity of variants – since *de novo* mutations are rare – as well as include relatedness constraints between individuals. Moreover, this framework can consider lineages of individuals while keeping computational costs low. We present and compare two methods and applied them to both real and simulated data to reconstruct heterozygous and homozygous signals and conclude that we achieve comparable recall rates for total SV detection with Method II for less CPU time.

3.4 DNA Sequencing as a Negative Binomial Process

3.4.1 Integer-Valued Dispersion

The following work is based on the paper by Banuelos et al. [3]. During the sequencing process, if genomic fragments are randomly chosen from the genome, then the Poisson distribution describes the number of reads covering any genomic locus [27]. The Poisson assumption with a mean represented by the *coverage* also assumes the same variance. However, sequencing technologies are known to be biased, resulting in large variation of coverage depth. This is particularly true in low-coverage settings [26, 49, 13]. In this regime, studies suggest that the two parameter negative binomial distribution may be more accurate in describing the distribution of fragments [37, 40].

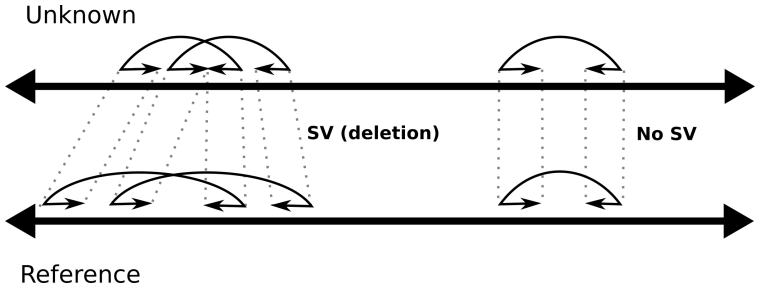

Figure 3.13: Illustration of regions in sequenced genome where there is a deletion (*left*) and no deletion (*right*) relative to a reference genome (ground truth). When sequenced fragments of the unknown genome do not map concordantly to the reference genome, we consider this a signal for a potential deletion or other structural variants (SVs). Note that for a deletion, the fragment from the individual maps to a larger than expected region in the reference. Fragments aligning to the reference in a concordant fashion indicate there is no genomic variation.

We note that many computational methods exist for processing mapped fragments and predicting SVs [34, 43, 35, 15, 23]; however most are based on only the mapped fragments and do not utilize other information about SVs if available. For example, SVs are relatively rare in an individual's genome, but most methods do not attempt to rank or prioritize predictions by how likely they are. This results in many false positive predictions because fragments that have been mapped to incorrect locations in the genome are likely to be mistaken as an SV [16, 19, 25]. In addition, when analyzing related individuals, who should share SVs, variant detection methods only use relatedness to filter calls as a post-processing step [34, 35, 15]. While some computational methods utilize the probability of arrangements of fragments, allowing them to estimate the probability a prediction is false or to rank their predictions by likelihood, most methods rely on the assumption of Poisson coverage [42]. Overall, most computational methods suffer from high false positive rates, but high-coverage and high quality data tend to resolve many false calls [41, 39].

In this work, we aim to improve upon past SV prediction methods in primarily three ways. Whereas previous work assumed mapped reads follow a Poisson distribution, we incorporate a negative binomial distribution to model the distribution of fragments [37,

10, 9, 6]. Instead of assuming equal mean and variance, we estimate both from the data and the negative binomial model captures the large variability in the sequencing coverage. Fig. 3.14, for example, provides empirical examples of this phenomena from the 1000 Genomes Project [16]. Secondly, we incorporate low-coverage data instead of relying on high-quality genomic data. Finally, we concurrently consider sequencing data of related individuals and enforce inheritance of variants through inequality constraints.

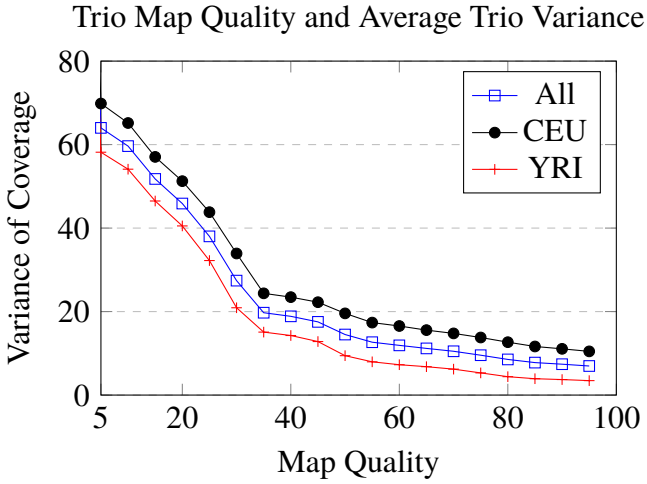

Figure 3.14: Plot of the map quality vs depth of coverage variance (mean per trio reported) for European (CEU) trio, Yoruba (YRI) trio, and both trios (father-mother-child) genomes from the 1000 Genomes Project. Varying the minimum map quality of reads, we calculate the depth of coverage for each genomic locus. The data show a much higher variance than the expected coverage of ≈ 4X.

Negative Binomial Log-Likelihood Optimization.

We consider the true signal $\vec{f}^* \in \{0,1\}^n$ to be a binary vector indicating the presence of a genetic variant, with $\vec{f}^*_j = 1$ if a variant is present at location j and 0 otherwise [10, 6, 8]. Thus, the corresponding parent $\vec{y}_p$ and child $\vec{y}_c$ observations are given by

$$\vec{y}_p \sim \text{NegBin}(\vec{\mu}_p, \vec{\sigma}^2_p) \quad \text{and} \quad \vec{y}_c \sim \text{NegBin}(\vec{\mu}_c, \vec{\sigma}^2_c), \tag{3.35}$$

where mean μ_i and variance $\sigma^2_i, (i \in \{p,c\})$ of depth of coverage will be determined by the sequencing data of each respective individual. We consider the stacked child-parent signal $\vec{y} = [\ \vec{y}_p^T \quad \vec{y}_c^T\ ]^T$ and corresponding mean and variance vectors, $\vec{\mu}$ and $\vec{\sigma}^2$. (Here, the notation $\vec{\sigma}^2$ is to be understood component-wise.) In particular, we have the following expressions for the components of $\vec{\mu}$ and $\vec{\sigma}^2$:

$$(\mu)_j = (A\vec{f}^*)_j \quad \text{and} \quad (\sigma)^2_j = (A\vec{f}^*)_j + \frac{1}{r}(A\vec{f}^*)^2_j,$$

where A, representing expected sequencing coverage, linearly projects the true signal $\vec{f}^*$ onto the n-dimensional set of observations, and r is the dispersion parameter of the negative binomial distribution.

Problem Formulation. When $r \to \infty$, we have $\sigma = \mu$ and this reduces to the Poisson case. If we choose to estimate these parameters from the sample, then we must observe a variance higher than the mean. Under this model, the probability of observing $\vec{y}$ is given by the following:

$$p(\vec{y}) = \prod_{j=1}^{n} \binom{y_j + \frac{\mu_j^2}{\sigma_j^2 - \mu_j} - 1}{y_j} \left(\frac{\mu_j}{\sigma_j^2} \right)^{\frac{\mu_j^2}{\sigma_j^2 - \mu_j}} \left(1 - \frac{\mu_j}{\sigma_j^2} \right)^{y_j}. \tag{3.36}$$

Ignoring constant terms, the negative log-likelihood term, $F(\mu, \sigma^2)$, becomes

$$\begin{aligned} F(\mu, \sigma^2) \quad \equiv \quad & \sum_{j=1}^{n} -\log \left(\frac{\left[y_j + \frac{\mu_j^2}{\sigma_j^2 + \mu_j} - 1 \right]!}{(y_j)! \left[\frac{\mu_j^2}{\sigma_j^2 + \mu_j} - 1 \right]!} \right) \\ & - \frac{\mu_j^2}{\sigma_j^2 - \mu_j} \log \left(\frac{1}{\sigma_j^2} \mu_j \right) - y_j \log \left(1 - \frac{1}{\sigma_j^2} \mu_j \right). \end{aligned} \tag{3.37}$$

Maximizing variance. Without reverting to the use of Gamma functions for $r \in \mathbb{R}$, we assume $r \in \mathbb{Z}^+$ and we know $\sigma_j^2 = \mu_j + \frac{1}{r} \mu_j^2$, where σ_j^2 is maximized when $r = 1$. Thus, we can rewrite the probability (3.43) of observing $\vec{y}$ as

$$p(\vec{y}) = \prod_{j=1}^{n} \left(\frac{1}{1 + \mu_j} \right) \left(\frac{\mu_j}{1 + \mu_j} \right)^{y_j}, \tag{3.38}$$

with associated negative log-likelihood,

$$F \equiv \sum_{j=1}^{n} (y_j + 1) \log \left(1 + \mu_j \right) - y_j \log \left(\mu_j \right).$$

However, we know the mean $\mu_j = e_i^T A f$. Then, adding the small parameter ε to represent sequencing or mapping error, we have

$$F(f) \equiv \sum_{j=1}^{n} (y_j + 1) \log \left(1 + e_i^T A f + \varepsilon \right) - y_j \log \left(e_i^T A f + \varepsilon \right), \tag{3.39}$$

with gradient

$$\nabla F(f) = \sum_{j=1}^{n} \frac{y_j + 1}{1 + e_i^T A f + \varepsilon} A^T e_i - \frac{y_j}{e_i^T A f + \varepsilon} A^T e_i. \tag{3.40}$$

Continuous Relaxation. To apply calculus of variations approaches in this classification problem, we allow for f to take on continuous values in $[0, 1]$. Otherwise, the combinatorial optimization problem may be intractable with a maximum-likelihood approach. As such, the negative binomial reconstruction algorithm takes the following form of the following constrained optimization problem for a one-parent and one-child (P, C) model:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{2n}}{\text{minimize}} \quad & \psi(\vec{f}) \equiv F(\vec{f}) + \tau \, \|\vec{f}\|_1 \\ \text{subject to} \quad & 0 \le \vec{f}_c \le \vec{f}_p \le \mathbf{1}, \end{aligned} \tag{3.41}$$

where $\vec{f} = [\,\vec{f}_p^{\,T}\ \vec{f}_c^{\,T}\,]^T$, $\mathbf{1}$ is a vector of ones, and τ is a regularization parameter. We assume that a child will have an SV at a certain location only if the parent also has the SV at the same location. We enforce this through the linear constraint $0 \leq \vec{f}_c \leq \vec{f}_p \leq 1$. Using a gradient-descent approach, the next iterate in our estimation is given by

$$\vec{f}^{k+1} = \left[\vec{f}^k - \alpha_k \nabla F(\vec{f}^k) + \tau \mathbf{1}\right]_{P,C}, \tag{3.42}$$

with step size (learning rate) α_k and the operation $[\cdot]_{P,C}$ is a projection onto the feasible set defined by the linear constraints in (3.47) (see [10] for further details).

Results. We evaluate the effectiveness of the proposed method on both simulated and real genomic data and compare our reconstructions with thresholding observations $\vec{y}$ and previous Poisson models. The proposed method is implemented in Python 3.6. We explored ten logarithmically-spaced regularization parameters τ from a 10^{-2} to 10^2 grid and chose the value yielding the largest average maximum area under curve for the receiver operating characteristic (ROC) using 5-fold cross-validation. To determine the number of true and false positives, we threshold the reconstructed signal – thereby un-relaxing our continuous assumption. For all experiments, we set $\alpha = 0.01$. The algorithm terminates if the relative difference between consecutive iterates $\|\vec{f}^{k+1} - \vec{f}^k\|/\|\vec{f}^k\|_2 \leq 10^{-6}$ or exceeds the maximum number of iterates.

We simulated two signals $\vec{f}_p$ and $\vec{f}_c$, representing the parent and child signals respectively. Each candidate set of SVs were drawn from a negative binomial distribution with dispersion parameter $r = 1$ and mean $\mu_p = \mu_c = 4$. We observe $n = 10^5$ potential SV candidates for each individual, with 500 true variants for $\vec{f}_p$, and 250 inherited variants for $\vec{f}_c$. This reflects a 50% similarity level and we set $\varepsilon = 0.01$ to represent the mapping and sequencing error in the forward model.

We first examine the parent signal reconstruction. Fig. 3.15 presents the number of false positive vs true positives for the reconstruction of the parent and child signals with mean coverage $\mu_p = \mu_c = 4$, $r = 1$, and $\varepsilon = 0.01$. Although $n = 10^5$, we focus on a more detailed view in the ROC curve to discern differences in prediction. Based on AUC measurements, we immediately observe an improvement in the number of true predictions over thresholding with our proposed model. For the 1 Parent-1 Child model, we expect parental reconstructions to be more informed by the child signal [10].

For the reconstructed child signal, we observe a marginal improvement when the number of false positives is relatively low. We note, however, that both the parent and child reconstructions incorporate a penalty of $\tau = 1$. This is an improvement on our previous methods, which typically resulted in tuning τ for each individual [5, 7].

We applied our method to both sequenced genomes of the father-mother-daughter trios from European (CEU) and Yoruba (YRI) populations. All six individuals from the 1000 Genomes Project were sequenced to $\approx$ 4X in Pilot 1 and aligned to NCBI36 [16]. We consider experimentally validated deletions meeting the following criteria as the true deletions: longer than 250bp, not *LowQual*, and non-overlapping with centromere and telomere regions.

We implemented GASV [43] on this data to obtain the candidate variant set. The intersection between the candidate SVs and true deletions results in the true signal $\vec{f}^*$. We

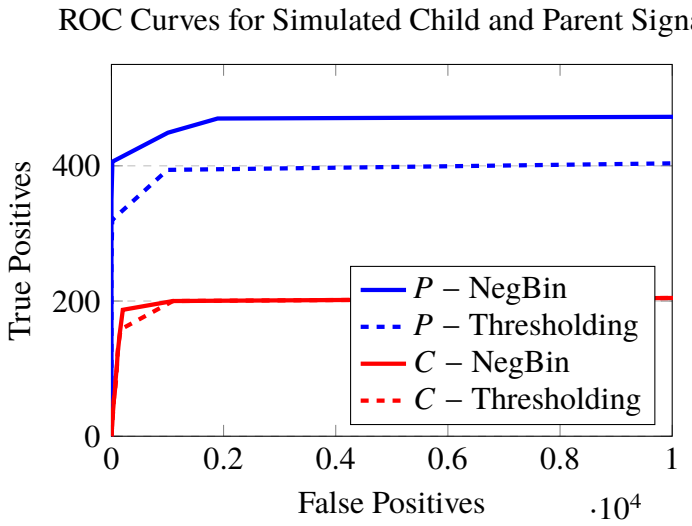

Figure 3.15: ROC curves illustrating the number of false positives vs the number of true positives for both the parent and child signal reconstruction with $\mu_p = 4$, $\mu_c = 4$, $\varepsilon = 0.01$, and 50% similarity. In both reconstructions, we set $\tau = 1.6681$ based on 5-fold cross validation. We observe more true positives using our proposed method when compared to thresholding the signal. This is particularly true in the first thousand predictions.

observe high variability in expected coverage in Fig. 3.14 and thus threshold the minimum map quality at 10 for all individuals.

We note a higher area under the curve in ROCs of the reconstructed signals for both CEU and YRI populations in Fig. 3.16 and 3.17 in comparison to the previous Poisson model. Additionally, this Fig. 3.16 depicts the number of novel deletions vs. true (experimentally validated) deletions since not the true set may be incomplete. Although not pictured, we observe similar trends for p_2 in CEU and YRI populations. Next, we consider the reconstruction for the child signals for both CEU and YRI populations. Fig. 3.17 illustrates a small but measurable difference in true predictions for both with the same τ across all individuals.

We propose a novel optimization method to detect structural variants from sequencing data of related individuals. Our method addresses mean and variance assumptions of previous methods and incorporates both relatedness and sparsity into the signal reconstruction. In the next section, we will relax the integer assumption on the dispersion parameter r to generalize our method and accommodate higher variances in sequencing data.

3.4.2 Real-Valued Dispersion

The work described below is based on the submitted work by Banuelos et al. [4]. With $r \in \mathbb{R}$, the non-integer number of failures reflects an increase in variance (i.e., much larger than the mean) and the distribution may still be represented by its probability mass

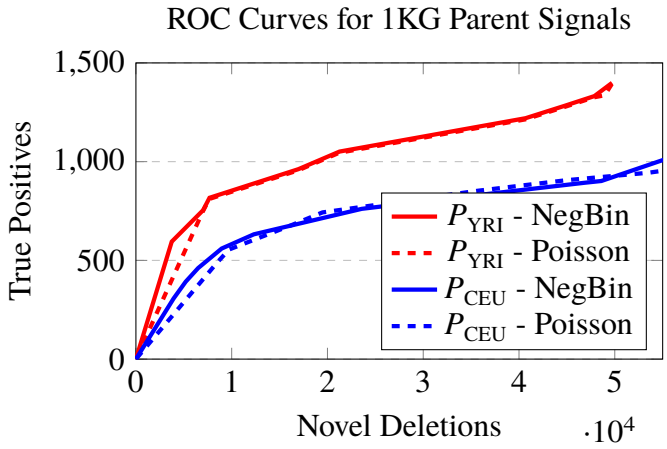

Figure 3.16: ROC curves illustrating the number of false positives vs the number of true positives for parent signal reconstruction with $\mu_p = \mu_c = 4$, $\varepsilon = 0.01$, $\tau = 0.01$ for both CEU and YRI populations. We observe an improvement in true predictions across both signals of interest.

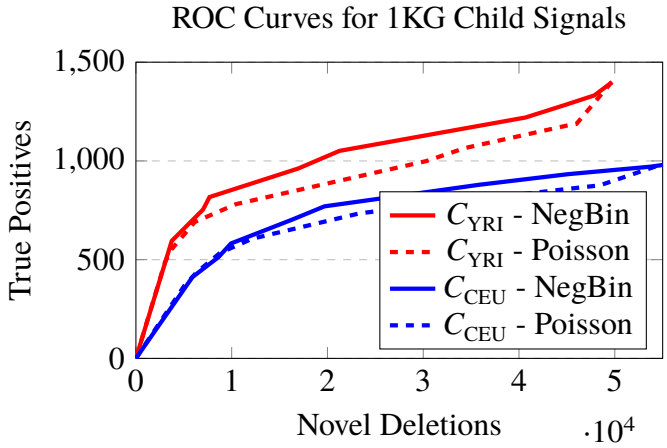

Figure 3.17: ROC curves illustrating the number of false positives vs the number of true positives for child signal reconstructions with $\mu_p = \mu_c = 4$, $\varepsilon = 0.01$, $\tau = 0.01$ for both CEU and YRI populations. For a fixed number of novel deletions, we report a higher number of true positives.

function. Thus, we seek to maximize the probability of observing $\vec{y}$, given by:

$$p(\vec{y}) = \prod_{i=1}^{n} \left(\frac{\Gamma(y_i + \frac{\mu_i^2}{\sigma_i^2 - \mu_i})}{y_i!\, \Gamma(\frac{\mu_i^2}{\sigma_i^2 + \mu_i})} \right) \left(\frac{\mu_i}{\sigma_i^2} \right)^{\frac{\mu_i^2}{\sigma_i^2 - \mu_i}} \left(1 - \frac{\mu_i}{\sigma_i^2} \right)^{y_i}, \tag{3.43}$$

where $\Gamma(z)$ is the gamma function given by $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x$. Subsequently, the negative log-likelihood term, $J(\mu, \sigma^2)$, becomes

$$\begin{aligned} J(\mu, \sigma^2) \equiv & \\ & - \sum_{i=1}^{n} \log \left(\frac{\Gamma(y_i + \frac{\mu_i^2}{\sigma_i^2 - \mu_i})}{(y_i)! \left[\Gamma(\frac{\mu_i^2}{\sigma_i^2 + \mu_i}) \right]!} \right) \\ & - \sum_{i=1}^{n} \left(\frac{\mu_i^2}{\sigma_i^2 - \mu_i} \log \left(\frac{1}{\sigma_i^2} \mu_i \right) - y_i \log \left(1 - \frac{1}{\sigma_i^2} \mu_i \right) \right). \end{aligned} \tag{3.44}$$

Estimating Dispersion. Since sequencing data support that $\sigma^2 > \mu$, we take an orthogonal approach to estimating the dispersion parameter r [45]. Instead of calculating r for windows along a genome for each individual, we estimate dispersion according to $\sigma_i = \mu_i + \frac{1}{r} u_i^2$. Thus, after estimating r and omitting constant terms, we can rewrite the negative log-likelihood as the following:

$$J(\mu) \equiv \sum_{i=1}^{n} (y_i + r) \log (r + \mu_i) - y_i \log (\mu_i). \tag{3.45}$$

Since the mean can be described by the linear transformation by the coverage matrix A on the true signal $\vec{f}^*$, we know $\mu_i = e_i^\top A f$. To model errors in both the sequencing and mapping process, we introduce the small parameter $\varepsilon > 0$ and the updated likelihood function becomes

$$J(\vec{f}) \equiv \sum_{i=1}^{n} (y_i + r) \log \left(r + e_i^\top A \vec{f} + \varepsilon \right) - y_i \log \left(e_i^\top A \vec{f} + \varepsilon \right). \tag{3.46}$$

Imposing Parent-Child Constraints. To minimize (3.46) using gradient-based approaches, we relax the set of feasible $\vec{f}$ from $\vec{f} \in \{0, 1\}^n$ to $0 \leq \vec{f} \leq 1$. This relaxation allows us to compute the first derivative of $J(\vec{f})$ and update each iterate accordingly. In addition, we constrain the space of feasible solutions by making two assumptions. First, we assume that a child must inherit an SV if both parents have the SV. Mathematically, we express this constraint as

$$\vec{f}_{p_1} + \vec{f}_{p_2} - 1 \leq \vec{f}_c.$$

In other words, if both parents have the SV at location i, then $(\vec{f}_{p_1})_i = 1$ and $(\vec{f}_{p_2})_i = 1$, and therefore $(\vec{f}_c)_i = 1$. Second, we assume that the child cannot have a variant if both parents do not have the SV at that location. We represent this as the following constraint:

$$\vec{f}_c \leq \vec{f}_{p_1} + \vec{f}_{p_2}.$$

In other words, if both parents do not have the SV at location i, then $(\vec{f}_{p_1})_i = 0$ and $(\vec{f}_{p_2})_i = 0$, and therefore $(\vec{f}_c)_i = 0$.

Optimization Formulation. The generalized negative binomial reconstruction algorithm takes the following form of the following constrained optimization problem in family lineages:

$$\begin{aligned} \underset{\vec{f}\in\mathbb{R}^{3n}}{\text{minimize}} \quad & \psi(\vec{f}) \equiv J(\vec{f}) + \tau \, \|\vec{f}\|_1 \\ \text{subject to} \quad & \vec{f}_{p_1} + \vec{f}_{p_2} - 1 \le \vec{f}_c \le \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & 0 \le \vec{f}_c, \vec{f}_{p_1}, \vec{f}_{p_2} \le 1, \end{aligned} \tag{3.47}$$

where $\vec{f} = [\, \vec{f}_{p_1}^{\top} \; \vec{f}_{p_2}^{\top} \; \vec{f}_c^{\top} \,]^{\top}$, and $\tau > 0$ is a regularization parameter that balances the data-fidelity term $J(\vec{f})$ with the sparsity-promoting penalty term $\|\vec{f}\|_1$.

We implement this method iteratively using a gradient descent approach and inheritance constraints – and corresponding projections – are enforced as in [9]. Specifically, at each iteration, we compute the gradient

$$\nabla J(\vec{f}) = \sum_{j=1}^{n} \frac{y_j + r}{r + e_i^{\top} A\vec{f} + \varepsilon} A^{\top} e_i - \frac{y_j}{e_i^{\top} A\vec{f} + \varepsilon} A^{\top} e_i, \tag{3.48}$$

and use it to compute the next iterate, defined by

$$\vec{f}^{k+1} = \text{Proj}\left[\vec{f}^{k} - \alpha_k \nabla J(\vec{f}^{k}) + \tau \mathbf{1}\right], \tag{3.49}$$

where α_k is the step size (learning rate), $\mathbf{1}$ is the vector of ones, and the operation Proj$[\,\cdot\,]$ is a projection onto the feasible set defined by the linear constraints in (3.47). Fig. 3.5 is a three-dimensional representation of this feasible set. For further details on projecting onto this feasible set, see [10].

3.5 Appendix: Tables of Minimizers and Projections

This appendix summarizes the minimizers for the aforementioned methods with the respective tables describing the analytical solutions to optimization subproblems presented in Sections 3.3.2 and 3.3.5.

Table 3.2: **Haploid Two Parents-One Child Minimizers.** Solutions to (3.16) given (c, p_1, p_2) corresponding to the feasible region in Fig. 3.5

REGION	c	p_1	p_2	Minimizer
Inside Region				
$R_{(c,p_1,p_2)}$	$0 \le c \le 1, c \le p_2 + p_1, c \ge p_2 + p_1 - 1$	$0 \le p_1 \le 1$	$0 \le p_2 \le 1$	(c, p_1, p_2)
Vertices				
$R_{(0,0,0)}$	$c \le -p_2, c \le -p_1$	$p_1 \le 0$	$p_2 \le 0$	$(0, 0, 0)$
$R_{(0,0,1)}$	$c \le 0, c \le -p_1$	$p_1 \le p_2 - 1$	$p_2 \ge 1$	$(0, 0, 1)$
$R_{(0,1,0)}$	$c \le 0, c \le -p_2$	$p_1 \ge 1$	$p_2 \le p_1 - 1$	$(0, 1, 0)$
$R_{(1,0,1)}$	$c \ge 1, c \ge 2 - p_2$	$p_1 \le 0$	$p_2 \ge p_1 + 1$	$(1, 0, 1)$
$R_{(1,1,0)}$	$c \ge 1, c \ge -p_1 + 2$	$p_1 \ge p_2 + 1$	$p_2 \le 0$	$(1, 1, 0)$
$R_{(1,1,1)}$	$c \ge -p_1 + 2, c \ge -p_2 + 2$	$p_1 \ge 1$	$p_2 \ge 1$	$(1, 1, 1)$
Edges				
$R_{(0,0,p_2)}$	$c < 0$	$p_1 < 0$	$0 < p_2 < 1$	$(0, 0, p_2)$
$R_{(0,p_1,0)}$	$c < 0$	$0 < p_1 < 1$	$p_2 < 0$	$(0, p_1, 0)$
$R_{(0,s_1,t_1)}$	$c < -\frac{1}{2}p_2 - \frac{1}{2}p_1 + \frac{1}{2}$	$p_1 < p_2 + 1, p_1 > p_2 - 1$	$p_2 > 1 - p_1$	$(0, s_1, t_1)$
$R_{(c,0,1)}$	$0 < c < 1$	$p_1 < 0$	$p_2 > 1$	$(c, 0, 1)$
$R_{(c,1,0)}$	$0 < c < 1$	$p_1 > 1$	$p_2 < 0$	$(c, 1, 0)$
$R_{(1,s_2,t_2)}$	$c > -\frac{1}{2}p_2 - \frac{1}{2}p_1 + \frac{3}{2}$	$p_1 < p_2 + 1, p_1 > p_2 - 1$	$p_2 < 1 - p_1$	$(1, s_2, t_2)$
$R_{(1,p_1,1)}$	$c > 1$	$0 < p_1 < 1$	$p_2 > 1$	$(1, p_1, 1)$
$R_{(1,1,p_2)}$	$c > 1$	$p_1 > 1$	$0 < p_2 < 1$	$(1, 1, p_2)$
$R_{(r_3,0,t_3)}$	$c > -p_2, c > p_2$	$p_1 < \frac{1}{2}(p_2 - c)$	$p_2 < 2 - c$	$(r_3, 0, t_3)$
$R_{(r_4,s_4,0)}$	$c > -p_1, c > p_1$	$p_1 < 2 - c$	$p_2 < \frac{1}{2}(p_1 - c)$	$(r_4, s_4, 0)$
$R_{(r_5,s_5,1)}$	$c > -p_1, c < p_1,$	$p_1 < 2 - c$	$p_2 > \frac{1}{2}(2 + p_1 - c)$	$(r_5, s_5, 1)$
$R_{(r_6,1,t_6)}$	$c > -p_2, c < p_2,$	$p_1 > \frac{1}{2}(2 + p_2 - c)$	$p_2 < 2 - c$	$(r_6, 1, t_6)$
Surfaces				
$R_{(c,0,p_2)}$	$c \ge 0, c \le p_2$	$p_1 \le 0$	$0 \le p_2 \le 1$	$(c, 0, p_2)$
$R_{(c,p_1,0)}$	$c \ge 0, c \le p_1$	$0 \le p_1 \le 1$	$p_2 \le 0$	$(c, p_1, 0)$
$R_{(1,p_1,p_2)}$	$c \ge 1$	$p_1 \le 1, p_1 \ge 1 - p_2$	$0 \le p_2 \le 1$	$(1, p_1, p_2)$
$R_{(0,p_1,p_2)}$	$c \le 0$	$p_1 \ge 0, p_1 \le 1 - p_2$	$0 \le p_2 \le 1$	$(0, p_1, p_2)$
$R_{(c,1,p_2)}$	$c \le 1, c \ge p_2$	$p_1 \ge 1$	$0 \le p_2 \le 1$	$(c, 1, p_2)$
$R_{(c,p_1,1)}$	$c \le 1, c \ge p_1$	$0 \le p_1 \le 1$	$p_2 \ge 1$	$(c, p_1, 1)$
$R_{(r_7,s_7,t_7)}$	$c \ge p_2 + p_1, c \le -\frac{1}{2}p_2 - \frac{1}{2}p_1 + \frac{3}{2}$	$p_1 \ge \frac{1}{2}(p_2 - c)$	$p_2 \ge \frac{1}{2}(p_1 - c)$	(r_7, s_7, t_7)
$R_{(r_8,s_8,t_8)}$	$c \le p_2 + p_1 - 1, c \ge -\frac{1}{2}p_2 - \frac{1}{2}p_1 + \frac{1}{2}$	$p_1 \le \frac{1}{2}(p_2 - c + 2)$	$p_2 \le \frac{1}{2}(p_1 - c + 2)$	(r_8, s_8, t_8)

Table 3.3: **Haploid Two Parents-One Child Projections** Edge and Surface projections of (c, p_1, p_2) corresponding to (3.16) and Table 3.2.

REGION	r_i	s_i	t_i	Minimizer
Edges				
$R_{(0,s_1,t_1)}$	0	$s_1 = \frac{1}{2}(p_1 - p_2 + 1)$	$t_1 = \frac{1}{2}(p_2 - p_1 + 1)$	$(0, s_1, t_1)$
$R_{(1,s_2,t_2)}$	1	$s_2 = \frac{1}{2}(p_1 - p2 + 1)$	$t_2 = \frac{1}{2}(p_2 - p_1 + 1)$	$(1, s_2, t_2)$
$R_{(r_3,0,t_3)}$	$r_3 = \frac{1}{2}(c + p_2)$	0	$t_3 = \frac{1}{2}(c + p_2)$	$(r_3, 0, t_3)$
$R_{(r_4,s_4,0)}$	$r_4 = \frac{1}{2}(c + p_1)$	$s_4 = \frac{1}{2}(c + p_1)$	0	$(r_4, s_4, 0)$
$R_{(r_5,s_5,1)}$	$r_5 = \frac{1}{2}(c + p_1)$	$s_5 = \frac{1}{2}(c + p_1)$	1	$(r_5, s_5, 1)$
$R_{(r_6,1,t_6)}$	$r_6 = \frac{1}{2}(c + p_2)$	1	$t_6 = \frac{1}{2}(c + p_2)$	$(r_6, 1, t_6)$
Surfaces				
$R_{(r_7,s_7,t_7)}$	$r_7 = \frac{2}{3}c + \frac{1}{3}p_1 + \frac{1}{3}p_2$	$s_7 = \frac{1}{3}c + \frac{2}{3}p_1 - \frac{1}{3}p_2$	$t_7 = \frac{1}{3}c - \frac{1}{3}p_1 + \frac{2}{3}p_2$	(r_7, s_7, t_7)
$R_{(r_8,s_8,t_8)}$	$r_8 = \frac{2}{3}c + \frac{1}{3}p_1 + \frac{1}{3}p_2 - \frac{1}{3}$	$s_8 = \frac{1}{3}c + \frac{2}{3}p_1 - \frac{1}{3}p_2 + \frac{1}{3}$	$t_8 = \frac{1}{3}c - \frac{1}{3}p_1 + \frac{2}{3}p_2 + \frac{1}{3}$	(r_8, s_8, t_8)

Table 3.4: **Diploid One Parent-One Child (Method 1) Minimizers**. Solutions to (3.32) given a_1 and b_1 for the region projections in Fig. 3.8 when $\hat{y}_p \leq 1 - \hat{y}_c$. Here $r = a_1 + b_1$, $s = -a_1 + 2 - \hat{y}_c - 2\hat{y}_p$, $t = -a_1 + 2 - 2\hat{y}_c - \hat{y}_p$.

Region	Condition a_1	Condition b_1	$\left(z_p^{(i)}, z_c^{(i)}\right)$
R_{a_1,b_1}	$0 \leq a_1 \leq 1 - \hat{y}_p$	$a_1 - \hat{y}_c \leq b_1$, $b_1 \leq a_1 + \hat{y}_c$, $b_1 \leq 1 - \hat{y}_c$ $0 \leq b_1$	(a_1, b_1)
R_1	$a_1 \leq 0$	$0 \leq b_1 \leq \hat{y}_p$	$(0, b_1)$
R_2	$a_1 < 0$	$b_1 < 0$	$(0, 0)$
R_3	$0 < a_1 \leq \hat{y}_c$	$b_1 \leq 0$	$(a_1, 0)$
R_4	$a_1 > \hat{y}_c$	$b_1 < s$	$(\hat{y}_c, 0)$
R_5	$a_1 \geq b_1 + \hat{y}_c$	$-a_1 + \hat{y}_c \leq b_1$ $b_1 \leq s$	$(\frac{1}{2}(r + \hat{y}_c), \frac{1}{2}(r - \hat{y}_c))$
R_6	$a_1 > 1 - \hat{y}_p$	$s < b_1$ $b_1 < 1 - \hat{y}_c - \hat{y}_p$	$(1 - \hat{y}_p, 1 - \hat{y}_c - \hat{y}_p)$
R_7	$a_1 \geq 1 - \hat{y}_p$	$1 - \hat{y}_p - \hat{y}_c < b_1$ $b_1 < 1 - \hat{y}_c$	$(1 - \hat{y}_p, b_1)$
R_8	$a_1 > 1 - \hat{y}_p$	$b_1 > 1 - \hat{y}_c$	$(1 - \hat{y}_p, 1 - \hat{y}_c)$
R_9	$1 - \hat{y}_c - \hat{y}_p < a_1$ $a_1 < 1 - \hat{y}_p$	$b_1 > 1 - \hat{y}_c$	$(a_1, 1 - \hat{y}_c)$
R_{10}	$a_1 < 1 - \hat{y}_c - \hat{y}_p$	$b_1 > t$	$(1 - \hat{y}_c - \hat{y}_p, 1 - \hat{y}_c)$
R_{11}	$a_1 \leq b_1 - \hat{y}_p$	$-a_1 + \hat{y}_p \leq b_1$ $b_1 \leq t$ $b_1 \leq -a_1$	$(\frac{1}{2}(r - \hat{y}_p), \frac{1}{2}(r + \hat{y}_p))$
R_{12}	$a_1 < 0$	$\hat{y}_p < b_1$ $b_1 < -a_1 + \hat{y}_p$	$(0, \hat{y}_p)$

Table 3.5: **Diploid One Parent-One Child (Method 2) Minimizers**. Solutions to (3.33) given b_1 and b_2 for the nontrivial projection regions. Here $\hat{q} = \hat{z}_p + \hat{y}_p$. The values (r_i, s_i) for $i = \{1, 2\}$ are given in Table 3.6.

Region	Condition b_1	Condition b_2	$(z_c^{(i)}, y_c^{(i)})$
$R_{(b_1,b_2)}$	$0 \leq b_1 \leq \hat{q}$	$b_2 \leq 1 - b_1$ $b_2 \geq \hat{z}_p - b_1$ $b_2 \geq 0$	(b_1, b_2)
R_1	$b_1 > 1 - b_2$	$b_2 > b_1 + 1 - 2\hat{q}$ $b_2 < b_1 + 1$	(r_1, s_1)
R_2	$b_1 > \hat{q}$	$b_2 < b_1$+1–$2\hat{q}$ $b_2 >$ $1 - \hat{q}$	$(\hat{q}, 1 - \hat{q})$
R_3	$b_1 > \hat{q}$	$0 < b_2 < 1 - \hat{q}$	$(\hat{q}, b_2)$
R_4	$b_1 > \hat{q}$	$b_2 < 0$	$(\hat{q}, 0)$
R_5	$\hat{z}_p < b_1 < \hat{q}$	$b_2 < 0$	$(b_1, 0)$
R_6	$b_1 < \hat{z}_p$	$b_2 < 0$ $b_2 < b_1 - \hat{z}_p$	$(\hat{z}_p, 0)$
R_7	$b_1 < \hat{z}_p - b_2$	$b_2 > b_1 - \hat{z}_p$ $b_2 < b_1 + \hat{z}_p$	(r_2, s_2)
R_8	$b_1 < 0$	$b_2 < \hat{z}_p$ $b_2 > b_1 + \hat{z}_p$	$(0, \hat{z}_p)$
R_9	$b_1 < 0$	$\hat{z}_p < b_2 < 1$	$(0, b_2)$
R_{10}	$b_1 < b_2 - 1$	$b_2 > 1$	$(0, 1)$

Table 3.6: **Diploid One Parent-One Child (Method 2) Projections**.The values of (r_1, s_1) and (r_2, s_2) in Table 3.5.

Region	Variable r_i	Variable s_i	$(z_c^{(i)}, y_c^{(i)})$
R_1	$\frac{1}{2}(b_1 - b_2 + 1)$	$\frac{1}{2}(b_2 - b_1 + 1)$	(r_1, s_1)
R_7	$\frac{1}{2}(b_1 - b_2 + \hat{z}_p)$	$\frac{1}{2}(b_2 - b_1 + \hat{z}_p)$	(r_2, s_2)

Bibliography

[1] L. Adhikari and R. F. Marcia. "Nonconvex relaxation for poisson intensity reconstruction". *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, pp. 1483–1487.

[2] D. M. Altshuler et al. "A Map of Human Genome Variation from Population Scale Sequencing". *Nature* 467.7319 (2010), pp. 1061–1073.

[3] M. Banuelos, S. Sindi, and R. F. Marcia. "Negative binomial optimization for biomedical structural variant signal reconstruction". Accepted to the *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2018.

[4] M. Banuelos, S. Sindi, and R. F. Marcia. *Structural variant prediction in extended pedigrees through sparse negative binomial genome signal recovery*. Accepted to the *International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2018.

[5] M. Banuelos et al. "Biomedical signal recovery: Genomic variant detection in family lineages". *Bioengineering (ENBENG), 2017 IEEE 5th Portuguese Meeting on*. IEEE. 2017, pp. 1–4.

[6] M. Banuelos et al. "Constrained Variant Detection with SPaRC: Sparsity, Parental Relatedness, and Coverage". Proceedings of *International Conference of the IEEE Engineering in Medicine and Biology Society*. 2016.

[7] M. Banuelos et al. "Nonconvex regularization for sparse genomic variant signal detection". *Medical Measurements and Applications (MeMeA), 2017 IEEE International Symposium on*. IEEE. 2017, pp. 281–286.

[8] M. Banuelos et al. "Sparse diploid spatial biosignal recovery for genomic variation detection". *Medical Measurements and Applications (MeMeA), 2017 IEEE International Symposium on*. IEEE. 2017, pp. 275–280.

[9] M. Banuelos et al. "Sparse Genomic Structural Variant Detection: Exploiting Parent-Child Relatedness For Signal Recovery". Proceedings of *IEEE Workshop on Statistical Signal Processing*. 2016.

[10] M. Banuelos et al. "Sparse Signal Recovery Methods for Variant Detection in Next-Generation Sequencing Data". Proceedings of *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2016.

[11] J. Barzilai and J. M. Borwein. “Two-Point Step Size Gradient Methods”. *IMA J. Numer. Anal.* 8.1 (1988), pp. 141–148. DOI: 10.1093/imanum/8.1.141.

[12] E. G. Birgin, J. M. Martínez, and M. Raydan. “Nonmonotone spectral projected gradient methods on convex sets”. *SIAM Journal on Optimization* 10.4 (2000), pp. 1196–1211.

[13] V. Boeva et al. “Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization”. *Bioinformatics* 27.2 (2011), pp. 268–269.

[14] C. Chaux, J.-C. Pesquet, and N. Pustelnik. “Nested Iterative Algorithms for Convex Constrained Image Recovery Problems”. *SIAM J. Imag. Sci.* 2.2 (2009), pp. 730–762. DOI: 10.1137/080727749. URL: http://link.aip.org/link/?SII/2/730/1.

[15] K. Chen et al. “BreakDancer: an algorithm for high-resolution mapping of genomic structural variation”. *Nature methods* 6.9 (2009), pp. 677–681.

[16] 1. G. P. Consortium et al. “An integrated map of genetic variation from 1,092 human genomes”. *Nature* 491.7422 (2012), pp. 56–65.

[17] M. Figueiredo and J. Bioucas-Dias. “Deconvolution of Poissonian images using variable splitting and augmented Lagrangian optimization”. *Proc. IEEE Workshop on Statistical Signal Processing*. Available at http://arxiv.org/abs/0904.4868. 2009.

[18] T. Goldstein and S. Osher. “The split Bregman method for $L1$-regularized problems”. *SIAM J. Imaging Sci.* 2.2 (2009), pp. 323–343. ISSN: 1936-4954.

[19] D. F. Gudbjartsson et al. “Large-scale whole-genome sequencing of the Icelandic population”. *Nature Genetics* 47.5 (2015), pp. 435–444.

[20] Z. T. Harmany, R. F. Marcia, and R. M. Willett. “Sparse Poisson intensity reconstruction algorithms”. *Proceedings of IEEE Statistical Signal Processing Workshop*. Cardiff, Wales, UK, Sept. 2009.

[21] Z. T. Harmany, R. F. Marcia, and R. M. Willett. “Sparsity-regularized photon-limited imaging”. *Proceedings of IEEE International Symposium on Biomedical Imaging*. Rotterdam, The Netherlands, Apr. 2010.

[22] Z. T. Harmany, R. F. Marcia, and R. M. Willett. “This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms—Theory and Practice”. *IEEE Trans. on Image Processsing* 21 (2011), pp. 1084–1096.

[23] F. Hormozdiari et al. “Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes”. *Genome research* 19.7 (2009), pp. 1270–1278.

[24] X. Huang, T. Lu, and B. Han. “Resequencing rice genomes: an emerging new era of rice genomics”. *Trends in Genetics* 29.4 (2013), pp. 225–232.

[25] J. Huddleston and E. E. Eichler. “An Incomplete Understanding of Human Genetic Variation”. *Genetics* 202.4 (2016), pp. 1251–1254.

[26] D. Iakovishina et al. “SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read map-pability”. *Bioinformatics* (2016), btv751.

[27] E. S. Lander and M. S. Waterman. “Genomic mapping by fingerprinting random clones: a mathematical analysis”. *Genomics* 2.3 (1988), pp. 231–239.

[28] J.-Y. Li, J. Wang, and R. S. Zeigler. “The 3,000 rice genomes project: new opportunities and challenges for future rice research”. *GigaScience* 3.1 (2014), pp. 1–3.

[29] P. Medvedev, M. Stanciu, and M. Brudno. “Computational methods for discovering structural variation with next-generation sequencing”. *Nature methods* 6 (2009), S13–S20.

[30] R. Nowak and E. Kolaczyk. “A multiscale MAP estimation method for Poisson inverse problems”. *32nd Asilomar Conf. Signals, Systems, and Comp.* Vol. 2. Pacific Grove, CA, 1998, pp. 1682–1686.

[31] R. Nowak and E. Kolaczyk. “A Multiscale Statistical Framework for Poisson Inverse Problems”. *IEEE Trans. Info. Theory* 46 (2000), pp. 1811–1825.

[32] P. L. Combettes and J.-C. Pesquet. “A proximal decomposition method for solving convex variational inverse problems”. *Inverse Prob.* 24.6 (2008), p. 065014.

[33] L. R. Pal and J. Moult. “Genetic Basis of Common Human Disease: Insight into the Role of Missense SNPs from Genome-Wide Association Studies”. *Journal of molecular biology* (2015).

[34] A. R. Quinlan et al. “Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome”. *Genome research* 20.5 (2010), pp. 623–635.

[35] T. Rausch et al. “DELLY: structural variant discovery by integrated paired-end and split-read analysis”. *Bioinformatics* 28.18 (2012), pp. i333–i339.

[36] L. I. Rudin, S. Osher, and E. Fatemi. “Nonlinear total variation based noise removal algorithms”. *Physica D: Nonlinear Phenomena* 60.1 (1992), pp. 259–268.

[37] J. Sampson et al. “Efficient study design for next generation sequencing”. *Genetic epidemiology* 35.4 (2011), pp. 269–277.

[38] S. Setzer, G. Steidl, and T. Teuber. “Deblurring Poissonian images by split Bregman techniques”. *J. Visual Commun. Image Represent.* (2009). In Press. ISSN: 1047-3203. DOI: DOI:10.1016/j.jvcir.2009.10.006.

[39] S. Shetty. “Structural Variant Detection”. PhD thesis. Arizona State University, 2014.

[40] D. Sims et al. “Sequencing depth and coverage: key considerations in genomic analyses”. *Nature Reviews Genetics* 15.2 (2014), pp. 121–132.

[41] S. S. Sindi and B. J. Raphael. “Identification of Structural Variation”. *Genome Analysis: Current Procedures and Applications* (2014), p. 1.

[42] S. S. Sindi et al. “An integrative probabilistic model for identification of structural variation in sequencing data”. *Genome biology* 13.3 (2012), R22.

[43] S. Sindi et al. “A geometric approach for classification and comparison of structural variants”. *Bioinformatics* 25.12 (2009), pp. i222–i230.

[44] D. Snyder. *Random Point Processes*. New York, NY: Wiley-Interscience, 1975.

[45] J. P. Szatkiewicz et al. “Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation”. *Nucleic acids research* 41.3 (2012), pp. 1519–1532.

[46] R. Tibshirani. “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[47] J. Weischenfeldt et al. “Phenotypic impact of genomic structural variation: insights from and for human disease”. *Nature Reviews Genetics* 14.2 (2013), pp. 125–138.

[48] S. J. Wright, R. D. Nowak, and M. Figueiredo. “Sparse reconstruction by separable approximation”. *IEEE Trans. on Signal Processing* 57.7 (2009), pp. 2479–2493.

[49] S. Yoon et al. “Sensitive and accurate detection of copy number variants using read depth of coverage”. *Genome research* 19.9 (2009), pp. 1586–1592.

Chapter 4

Convergence Analysis of Poisson Process Methods

In this chapter, I prove the convergence of methods discussed in Sections 3.3.1 and 3.3.2 by incorporating the method of Lagrange multipliers.

4.1 Haploid One Parent-One Child Method

Since we are concerned with this binary classification problem, we know both the minimum and maximum of the signal intensity and our reconstruction takes the form:

$$\begin{aligned} \hat{\vec{f}} = &\underset{\vec{f}\in\mathscr{F}}{\arg\min} \quad \phi(\vec{f}) \equiv F(\vec{f}) + \tau\|\vec{f}\|_1, \\ &\text{subject to } \ \mathbf{0} \le \vec{f} \le \mathbf{1}, \end{aligned} \tag{4.1}$$

where $F(\vec{f})$ is the negative Poisson log-likelihood function and $\mathscr{F}$ represents the feasible set of solutions. Following the SPIRAL framework [5] , this decomposes into the following subproblem:

$$\begin{aligned} \left[\hat{f}_c, \hat{f}_p\right] = &\underset{f_p, f_c\in\mathbb{R}}{\arg\min} \quad \frac{1}{2}(f_p - b)^2 + \frac{1}{2}(f_c - a)^2 \\ &\text{subject to } \ 0 \le f_c \\ &\qquad\qquad\ \ 0 \le f_p - f_c \\ &\qquad\qquad\ \ 0 \le 1 - f_p. \end{aligned} \tag{4.2}$$

Theorem 3. *The separable subproblem* (4.2) *has a unique minimizer* $[f_c^*, f_p^*]$ *for* $a, b \in \mathbb{R}$.

Proof. The associated Lagrangian for (4.2) is given by

$$\begin{aligned} \mathscr{L}(f_c, f_p, \lambda_1, \lambda_2, \lambda_3) = &\frac{1}{2}(f_p - b)^2 + \frac{1}{2}(f_c - a)^2 \\ &- \lambda_1 f_c - \lambda_2(f_p - f_c) \\ &- \lambda_3(1 - f_p), \end{aligned}$$

where $\lambda_1, \lambda_2, \lambda_3 \geq 0$ are the Lagrange multipliers corresponding to the previous constraints. Since the primal problem is convex and Slater's condition holds for the primal problem, the duality gap is zero. Thus, the primal and dual values are equal and since the level sets of Equation (4.2) are concentric circles, then the orthogonal projection method proposed in Section 3.3.1 solves both the primal and dual problem [3, 4]. For completeness, we consider solving the dual problem. To satisfy the complementary slackness conditions, we consider which constraints are not active, which directly depends on the value of the unconstrained minimum (a, b). We proceed by cases, with a specific example for the interior, edges, and vertices of the feasible region:

Case 1 - Interior Region - (a, b) **satisfy all inequality constraints.** In this case, all the constraints are not active and $\lambda_1 = \lambda_2 = \lambda_3 = 0$. Thus,

$$\left[f_c^*, f_p^*\right] = [a, b].$$

Case 2 - Vertices. We consider the unconstrained minimum, where $a \geq 1, b \geq -a + 2$. Since $a \geq 1$, we know $\lambda_1 = 0$. Furthermore, we must consider the following subcases *Subcase* $a > b$. The modified Lagrangian becomes

$$\begin{aligned}\mathscr{L}_{V_1}(f_c, f_p, 0, \lambda_2, \lambda_3) = \frac{1}{2}(f_p - b)^2 + \frac{1}{2}(f_c - a)^2 \\ - \lambda_2(f_p - f_c) - \lambda_3(1 - f_p).\end{aligned}$$

Differentiating $\mathscr{L}_{V_1}$ with respect to f_c and f_p and setting the derivatives to zero yields

$$\begin{aligned} f_c &= a - \lambda_2 \\ f_p &= b + \lambda_2 - \lambda_3. \end{aligned} \tag{4.3}$$

Then, substituting (4.3) in $\mathscr{L}_{V_1}$, the Lagrangian dual problem becomes

$$\begin{aligned}\mathscr{J}_{V_1}(\lambda_2, \lambda_3) = a\lambda_2 + b(\lambda_3 - \lambda_2) - \lambda_2^2 + \lambda_2\lambda_3 - \frac{\lambda_3^2}{2} - \lambda_3 \\ \text{subject to } \lambda_2, \lambda_3 \geq 0.\end{aligned} \tag{4.4}$$

Computing the gradient of $\mathscr{J}_{V_1}$ with respect to λ_2 and λ_3 yields

$$\begin{aligned}\nabla_{\lambda_2}\mathscr{J}_{V_1} &= \lambda_3 + a - b - 2\lambda_2 \\ \nabla_{\lambda_3}\mathscr{J}_{V_1} &= \lambda_3 + a - b - 2\lambda_2.\end{aligned}$$

Solving for λ_2 and λ_3, we have $\lambda_2 = -1 + a$ and $\lambda_3 = -2 + a + b$. Note that we have

$$\begin{aligned} f_c = a - \lambda_2 &= a - (-1 + a) \\ &= 1, \end{aligned}$$

and

$$\begin{aligned} f_p = b + \lambda_2 - \lambda_3 &= b + (-1 + a) - (-2 + a + b) \\ &= 1. \end{aligned}$$

Thus, since (4.2) is strictly convex, the global minimum for $\mathscr{L}_{V_1}(f_p, f_c, 0, \lambda_2, \lambda_3)$ is $(1, 1, 0, -1+a, -2+a+b)$, which satisfies the KKT conditions and agrees with projections outlined in [2]. We obtain consistent results for remaining vertices. *Subcase* $a \leq b$. We have $\lambda_2 = 0$ and the modified Lagrangian becomes

$$\begin{aligned}\mathscr{L}_{V_2}(f_c, f_p, 0, 0, \lambda_3) = \frac{1}{2}(f_p - b)^2 + \frac{1}{2}(f_c - a)^2 \\ - \lambda_3(1 - f_p).\end{aligned}$$

Differentiating $\mathscr{L}_{V_2}$ with respect to f_c and f_p and setting the derivatives to zero yields

$$\begin{aligned} f_c &= a \\ f_p &= b - \lambda_3. \end{aligned} \tag{4.5}$$

Then, for the dual problem, we have

$$\begin{aligned}\mathscr{J}_{V_2}(\lambda_3) = \frac{1}{2}\lambda_3(-\lambda_3 - 2 + 2b) \\ \text{subject to } \ \lambda_3 \geq 0.\end{aligned}$$

Differentiating $\mathscr{J}_{V_2}$ with respect to λ_3, setting it to zero, and solving for λ_3, we obtain $\lambda_3 = b - 1$. Note that we have

$$f_p = b - \lambda_3 = b - (b - 1) = 1$$

and since $a \geq 1$ and $a \leq b$, we have $f_c = a = 1$. Thus, the global minimum for $\mathscr{L}_{V_1}(f_p, f_c, 0, 0, \lambda_3)$ is $(1, 1, 0, 0, b - 1)$.

Case 3 - Edges. We consider the unconstrained minimum, where $a \geq |b|, b < -a+2$. Since $a \geq 0$, we know $\lambda_1 = 0$. Moreover, we have $b \leq 1$, which results in the third constraint as not active and we have $\lambda_3 = 0$. Then, the modified Lagrangian becomes

$$\begin{aligned}\mathscr{L}_{E_1}(f_c, f_p, 0, \lambda_2, 0) = \frac{1}{2}(f_p - b)^2 + \frac{1}{2}(f_c - a)^2 \\ - \lambda_2(f_p - f_c)\end{aligned}$$

Differentiating $\mathscr{L}_{E_1}$ with respect to f_c and f_p and setting the derivatives to zero yields

$$\begin{aligned} f_c &= a - \lambda_2 \\ f_p &= b + \lambda_2. \end{aligned} \tag{4.6}$$

Then, substituting (4.6) in $\mathscr{L}_{E_1}$, the Lagrangian dual problem becomes

$$\begin{aligned}\mathscr{J}_{E_1}(\lambda_2) = a\lambda_2 - b\lambda_2 - \lambda_2^2 \\ \text{subject to } \ \lambda_2 \geq 0.\end{aligned} \tag{4.7}$$

Computing the gradient of $\mathscr{J}_{E_1}$ with respect to λ_2, setting to zero and solving yields

$$\lambda_2 = \frac{a - b}{2} \Rightarrow f_c = f_p = \frac{a + b}{2}.$$

Thus, the optimal point for $\mathscr{L}_{E_1}(f_p, f_c, 0, \lambda_2, 0)$ is $(\frac{a+b}{2}, \frac{a+b}{2}, 0, \frac{a-b}{2}, 0)$. Using a similar approach, we uniquely identify the optimal solution for the remaining two edges. □

As a result of the unique minimizer for the subproblem above, we obtain the main convergence result for the Haploid One Parent-One Child (P_1, C_1) method.

Theorem 4. *Let* $\left\{f_c^{(j)}, f_p^{(j)}\right\}_{j\geq 0}$ *be the sequence generated by iteratively solving the separable subproblem* (4.2)*. Then all accumulation points are critical points and hence the* P_1, C_1 *method converges to a minimizer of* (4.1)*.*

Proof. This proof follows identically as the proof of Theorem 1 in [5] where this problem satisfies all the conditions outlined (e.g. F is proper convex) and no modifications are required for the constraints within the proof. □

4.2 Haploid Two Parent-One Child Method

Following the SPIRAL framework [5], Equation (3.16) decomposes into the following subproblem:

$$\begin{aligned}
\left[\hat{f}_c, \hat{f}_{p_1}, \hat{f}_{p_2}\right] = \underset{f_p, f_c \in \mathbb{R}}{\arg\min} \quad & \frac{1}{2}(f_c - c)^2 + \frac{1}{2}(f_{p_1} - p_1)^2 + \frac{1}{2}(f_{p_2} - p_2)^2 \\
\text{subject to} \quad & 0 \leq f_c \\
& 0 \leq f_{p_1} \\
& 0 \leq f_{p_2} \\
& 0 \leq 1 - f_c \\
& 0 \leq 1 - f_{p_1} \\
& 0 \leq 1 - f_{p_2} \\
& 0 \leq f_c - f_{p_1} - f_{p_2} + 1 \\
& 0 \leq f_{p_1} + f_{p_2} - f_c.
\end{aligned} \tag{4.8}$$

Theorem 5. *The separable subproblem* (4.8) *has a unique minimizer* $[f_c^*, f_{p_1}^*, f_{p_2}^*]$ *for* $c, p_1, p_2 \in \mathbb{R}$*.*

Proof. The associated Lagrangian for (4.8) is given by

$$\begin{aligned}
\mathscr{L}(f_c, f_{p_1}, f_{p_2}, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8) = & \frac{1}{2}(f_c - c)^2 + \frac{1}{2}(f_{p_1} - p_1)^2 + \frac{1}{2}(f_{p_2} - p_2)^2 \\
& - \lambda_1 f_c - \lambda_2 f_{p_1} - \lambda_3 f_{p_2} \\
& - \lambda_4(1 - f_c) - \lambda_5(1 - f_{p_1}) - \lambda_6(1 - f_{p_2}) \\
& - \lambda_7(f_c - f_{p_1} - f_{p_2} + 1) - \lambda_8(f_{p_1} + f_{p_2} - f_c),
\end{aligned}$$

As in the previous section, the primal problem is convex and Slater's condition holds for the primal problem. Hence, the duality gap is zero and the primal and dual values are equal. Since the level sets of Equation (4.8) are concentric circles, then the orthogonal projection method proposed in Section 3.3.2 solves both the primal and dual problem [3, 4]. As before, we consider the cases for solving the dual problem for completeness.

Case 1 - Interior Region - (c, p_1, p_2) **satisfy all inequality constraints.** In this case, all the constraints are not active and $\lambda_i = 0$ for $i = 1, \dots, 8$. Thus,

$$\left[f_c^*, f_{p_1}^*, f_{p_2}^*\right] = [c, p_1, p_2].$$

Case 2 - Vertices. We consider the unconstrained minimum, where $c \geq 1$, $c \geq -p_1 + 2$, $p_1 \geq p_2 + 1$, and $p_2 \leq 0$. At minimum, this results in $\lambda_1 = \lambda_6 = 0$ for the inactive constraints. Moreover, we consider the following three subcases:
subcase V1: $0 < p_1 < 1$**.** Since p_1 satisfies the second and fifth constraints, we also have $\lambda_2 = \lambda_5 = 0$. Consequently, we know that $c - p_1 - p_2 + 1 \geq 0$ and thus $\lambda_7 = 0$ as an inactive constraint. Then, the modified Lagrangian becomes

$$\begin{aligned}\mathscr{L}_{V_1}(f_c, f_{p_1}, f_{p_2}, 0, 0, \lambda_3, \lambda_4, 0, 0, 0, \lambda_8) = \frac{1}{2}(f_c - c)^2 + \frac{1}{2}(f_{p_1} - p_1)^2 + \frac{1}{2}(f_{p_2} - p_2)^2 \\ - \lambda_3 f_{p_2} - \lambda_4(1 - f_c) - \lambda_8(f_{p_1} + f_{p_2} - f_c).\end{aligned}$$

Differentiating $\mathscr{L}_{V_1}$ with respect to f_c, f_{p_1} and f_{p_2}, setting the derivatives to zero yields

$$\begin{aligned} f_c &= c - \lambda_4 - \lambda_8 \\ f_{p_1} &= \lambda_8 + p_1 \\ f_{p_2} &= \lambda_3 + \lambda_8 + p_2. \end{aligned} \tag{4.9}$$

Substituting (4.9) into $\mathscr{L}_{V_1}$, the Lagrangian dual problem becomes

$$\begin{aligned}\mathscr{J}_{V_1}(\lambda_3, \lambda_4, \lambda_8) = &\frac{1}{2}\left(2\lambda_4(c - \lambda_8 - 1) - \lambda_8(-2c + 3\lambda_8 + 2(p_1 + p_2))\right. \\ &\left. - \lambda_3^2 - 2\lambda_3(\lambda_8 + p_2) - \lambda_4^2\right) \\ &\text{subject to } \ \lambda_3, \lambda_4, \lambda_8 \geq 0.\end{aligned} \tag{4.10}$$

Computing the gradient of $\mathscr{J}_{V_1}$ with respect to λ_3, λ_4, and λ_8, setting the derivatives equal to zero, we obtain

$$\begin{aligned} \lambda_3 &= p_1 - p_2 - 1 \\ \lambda_4 &= c + p_1 - 2 \\ \lambda_8 &= 1 - p_1, \end{aligned} \tag{4.11}$$

which results in $[f_c, f_{p_1}, f_{p_2}] = [1, 1, 0]$. Thus, the optimal point for $\mathscr{L}_{V_1}(f_c, f_{p_1}, f_{p_2}, 0, 0, \lambda_3, \lambda_4, 0, 0, 0, \lambda_8$ is $(1, 1, 0, 0, 0, p_1 - p_2 - 1, c + p_1 - 2, 0, 0, 0, 1 - p_1)$, as described in [1].
subcase V2: $p_1 \geq 1$**.** In this case, we have $\lambda_2 = 0$ as the inactive constraint. However, this only occurs if $p_2 = 0$. Thus, we conclude $\lambda_3 = 0$. With $p_2 = 0$ and having λ_4 and λ_5 as active constraints ensures that the seventh constraint is met (i.e., $\lambda_7 = 0$). Thus, the modified Lagrangian becomes

$$\begin{aligned}\mathscr{L}_{V_2}(f_c, f_{p_1}, f_{p_2}, 0, 0, 0, \lambda_4, \lambda_5, 0, 0, \lambda_8) = \frac{1}{2}(f_c - c)^2 + \frac{1}{2}(f_{p_1} - p_1)^2 + \frac{1}{2}(f_{p_2} - p_2)^2 \\ - \lambda_4(1 - f_c) - \lambda_5(1 - f_{p_1}) - \lambda_8(f_{p_1} + f_{p_2} - f_c).\end{aligned}$$

Differentiating $\mathscr{L}_{V_2}$ with respect to f_c,f_{p_1} and f_{p_2}, setting the derivatives to zero and recalling $p_2 = 0$, yields

$$\begin{aligned} f_c &= c - \lambda_4 - \lambda_8 \\ f_{p_1} &= -\lambda_5 + \lambda_8 + p_1 \\ f_{p_2} &= \lambda_8 + p_2 = \lambda_8. \end{aligned} \tag{4.12}$$

Substituting (4.12) into $\mathscr{L}_{V_2}$, the Lagrangian dual problem becomes

$$\begin{aligned} \mathscr{J}_{V_2}(\lambda_4, \lambda_5, \lambda_8) = &\frac{1}{2}\,(2c\lambda_8 - 3\lambda_8^2 - 2\lambda_8 p_1 + 2\lambda_5(p_1 - 1 + \lambda_8) - 2\lambda_8 p_2 + p_2^2 \\ &- \lambda_4^2 + 2\lambda_4(c - 1 - \lambda_8) - \lambda_5^2\,) \\ &\text{subject to } \ \lambda_4, \lambda_5, \lambda_8 \geq 0. \end{aligned} \tag{4.13}$$

Computing the gradient of $\mathscr{J}_{V_2}$ with respect to λ_4, λ_5, and λ_8, setting the derivatives equal to zero, we obtain

$$\begin{aligned} \lambda_4 &= c - 1 - \lambda_8 = c - 1 \\ \lambda_5 &= \lambda_8 - 1 + p_1 = p_1 - 1 \\ \lambda_8 &= 0, \end{aligned} \tag{4.14}$$

which results in $[f_c,f_{p_1},f_{p_2}] \ = \ [1, 1, 0]$. Thus, the optimal point for $\mathscr{L}_{V_1}(f_c,f_{p_1},f_{p_2}, 0, 0, 0, \lambda_4, \lambda_5, 0, 0, \lambda_8)$ is $(1, 1, 0, 0, 0, 0, c - 1, p_1 - 1, 0, 0, 0)$, as described in [1].

subcase V3: $p_1 \leq 0$**.** In this case, we have $\lambda_5 = 0$ as an inactive constraint. Consequently, we have $\lambda_7 = 0$. Moreover, we know that $p_1 \geq p_2 + 1$ in this regime and so having the active constraint of λ_3 ensures $p_1 \geq 0$ (i.e., $\lambda_2 = 0$). Thus, the modified Lagrangian is given by

$$\begin{aligned} \mathscr{L}_{V_3}(f_c,f_{p_1},f_{p_2}, 0, 0, \lambda_3, \lambda_4, 0, 0, 0, \lambda_8) = &\frac{1}{2}(f_c - c)^2 + \frac{1}{2}(f_{p_1} - p_1)^2 + \frac{1}{2}(f_{p_2} - p_2)^2 \\ &- \lambda_3 f_{p_2} - \lambda_4(1 - f_c) - \lambda_8(f_{p_1} + f_{p_2} - f_c). \end{aligned}$$

However, $\mathscr{L}_{V_3} = \mathscr{L}_{V_1}$ with the same optimal solution as above. Edges and surfaces follow similarly as the case for vertices by considering the modified Lagrangian for three subcases for each region of the truncated cube. □

Establishing the unique minimizer for the subproblem above leads to the main convergence result for the Haploid Two Parent-One Child (P_2, C_1) method:

Theorem 6. *Let* $\left\{f_c^{(j)},f_{p_1}^{(j)},f_{p_2}^{(j)}\right\}_{j\geq 0}$ *be the sequence generated by iteratively solving the separable subproblem* (4.8)*. Then all accumulation points are critical points and hence the* P_2, C_1 *method converges to a minimizer of* (4.1)*.*

Proof. This proof follows identically as the proof of Theorem 1 in [5] where this problem satisfies all the conditions outlined (e.g. F is proper convex) and no modifications are required for the constraints within the proof. □

Bibliography

[1] M. Banuelos et al. “Sparse Genomic Structural Variant Detection: Exploiting Parent-Child Relatedness For Signal Recovery”. Proceedings of *IEEE Workshop on Statistical Signal Processing*. 2016.

[2] M. Banuelos et al. “Sparse Signal Recovery Methods for Variant Detection in Next-Generation Sequencing Data”. Proceedings of *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2016.

[3] J. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[4] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[5] Z. T. Harmany, R. F. Marcia, and R. M. Willett. “This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms—Theory and Practice”. *IEEE Trans. on Image Processsing* 21 (2011), pp. 1084–1096.

Chapter 5

Conclusions and Future Work

5.1 Conclusion

Genomic variation, responsible for a variety of evolutionary phenomena, remains a substantial field for mathematical and statistical models. Once regarded as junk DNA, understanding how transposable elements proliferate (and proliferated in the past) may uncover their abundance across many species. Moreover, detecting such genomic rearrangements has only become feasible with advances in DNA sequencing capabilities. Detection methods, as a result, will likely uncover potential genomic changes responsible for inherited traits and diseases.

This dissertation outlined my contributions to the field of modeling and detecting genomic variation within and between species. In particular, this work was guided and organized by two fundamental questions: 1) how does DNA causing genomic variation proliferate through the genome of a species? and 2) how can we leverage *a priori* information to improve predictions of genomic variants?

Chapter 2 addressed the first question through the development of a mathematical model of transposable element length distributions. Both discrete and continuous models accurately depicts the distribution of TEs in a variety of species. By explicitly modeling partial length elements, the proposed model incorporated data often ignored by previous models.

In Chapter 3, I developed a general optimization framework to detect genomic variants subject to heredity constraints. In this general context, I presented a total of five different models reflecting a change in family structure and DNA sequencing assumptions. Both simulated and real data results were presented for a subset of these models. Lastly, Chapter 4 addressed the convergence of Haploid models with Poisson sequencing assumptions using the method of Lagrange multipliers.

These statistical models and methods, however, raise even more questions. Next, I outline some of these emerging questions and how I will create and build upon mathematical models to begin answering them.

5.2 Future Work

5.2.1 Nonconvex Methods for Variant Detection

The end of Chapter 3 introduced using the negative binomial distribution to model DNA sequencing. As a result, this assumption resulted in a nonconvex formulation of the objective function. Although a local minimum will be obtained using gradient-descent methods, I am interested in applying methods from nonconvex optimization to structural variant detection. In particular, machine learning approaches present a tractable way to address this classification problem.

The methods discussed in Chapter 3 also only focused on traditional family structures, but these relatedness constraints may be relaxed to include non-parental and non-sibling constraints. If data were collected from a geographic region, for example, I look forward to exploring how this changes the proposed models in the context of both noisy and low-quality genomic data.

5.2.2 Malarial Resistance and Signs of Selection

To characterize the role of genomic variants in humans, I am also interested in applying statistical and analytical models to endemic diseases, like malaria. I am currently using sequencing data to obtain an unbiased sample for evidence of pathogenic resistance (e.g., malaria). Most studies that have been conducted have focused on sequencing the DNA of individuals with the intent of analyzing specific genes associated with malaria resistance. However, most large-scale sequencing data repositories have the broader of goal of categorizing human genetic variation. I intend to analyze both genes associated with such resistance but also include any regions related to blood regulation in humans. To accomplish this goal and view high-dimensional data, I plan to use dimension-reduction techniques, such as principal component analysis (PCA), in a local genomic setting.

5.2.3 A Mathematical Model of Central Valley Fever

My introduction to fungal genomics at the Joint Genome Institute has led to questions in modeling Coccidioides, the cause of Valley fever, and applying classification techniques to Valley fever patient-data to better inform our limited understanding of this disease.