University of California, Merced

Genomic Signal Processing for Structural Variant Detection in Related Individuals

A Thesis submitted in partial satisfaction of the requirements for the degree of Master of Science

in

Applied Mathematics

by

Melissa I. Spence

Committee in charge:
Professor Suzanne Sindi, Chair
Professor Roummel Marica
Professor Mario Banuelos
Professor Arnold Kim
Professor Erica Rutter

2020

This is to certify that I have examined a copy of a technical report by

Melissa I. Spence

and found it satisfactory in all respects, and that any and all revisions required by the examining committee have been made.

Applied Mathematics Graduate Studies Chair / Research Advisor:

Professor Roummel Marcia

Thesis Committee:

Professor Arnold Kim

Thesis Committee:

Professor Erica Rutter

Thesis Committee:

Professor Mario Banuelos

Committee Chair / Research Advisor:

Professor Suzanne Sindi

Date

Acknowledgments

I want to sincerely thank my wonderful advisors and mentors Suzanne Sindi and Roummel Marcia. They have both played a significant role in this chapter of my life and have helped me grow into a better scientist. I would also like to give significant thanks to Mario Banuelos, Arnold Kim and Erica Rutter for their collaboration, support and advice. I have been blessed to be surrounded by these amazing mentors during my time at UC Merced and many others such as Dean Marjorie Zatz. They have helped me define the leader I want to be as I move forward in my career and I appreciate all of the guidance they have given me. I also want to thank all of my fellow graduate students in applied mathematics. I am grateful that I got to go on this journey with each of you and experience your growth alongside mine, it is a relationship I continue to learn from and will never forget. Finally, I want to thank my husband, parents, and grandparents. They have been my biggest supporters and have helped me in any way possible through my entire academic journey. They continue to help me grow and evolve into a better version of myself and I am forever thankful. I would like to acknowledge funding from the University of California, Merced's School of Natural Sciences, and NSF Grant IIS-1741490.

Contents

List of Symbols

SV	structural variant
p	index of parent
p_1, p_2	index of parent 1 and parent 2 respectively
m	number of locations in the individuals genome
i	index of the inherited signal in the child
n	index of the novel signal in the child
j	index corresponding to the j^th location in the genome signal
λ	sequencing coverage
ϵ	measurement error corresponding to sequencing
$\vec{s}$	observed signal
$\vec{f}$	true signal
A	coverage matrix
F	negative Poisson log-likelihood function
$\mathcal{S}$	set of feasible points
γ	penalty weight for the novel variants in the child
τ	sparsity regularization parameter
$\mathcal{Q}(\vec{f})$	quadratic approximation to the objective function
α_k	learning rate
q	predicted new iterate along the steepest descent
ρ	percent overlap between the parent and child SV's in simulated data

List of Figures

List of Tables

Genomic Signal Processing for Structural Variant Detection in Related Individuals

by

Melissa I. Spence

Master of Science in Applied Mathematics

Suzanne Sindi, Committee Chair

University of California, Merced

2020

Abstract

In this work we develop a general optimization framework to more accurately recover structural variants (SVs) in low-coverage sequencing data from genomes of related individuals. In previous work the framework incorporated biological constraints that reflect relatedness between individuals and enforced sparsity to model the rarity of SVs. This framework operated under the assumption that the genomes were haploid, meaning that each individual had one copy of the genetic material. There are two main contributions of this thesis: First we propose an approach that allows the child signal to possess variants that are not present in either parent (i.e., novel SVs) under the assumption of haploid signals. Second, we propose an approach to reconstruct the signals of two parents and a child under the assumption of diploid genomes. We tested the effectiveness of these approaches on both simulated data and data from the 1000 Genomes Project.

Chapter 1

Introduction

1.1 Genetic Structural Variation

The complete DNA sequence of an organism (the genome) is one or more ordered linear sequences of the letters A,C,G, or T. The total genome length is anywhere from millions (for bacteria) to billions (for mammals) of letters. Every cell in most multi-cellular organisms contains a complete and nearly identical copy of an organism's genome. When cells divide, the genome must be duplicated so each cell will have its own copy, but every time the genome is copied there is the opportunity for mutational processes to introduce variation. Genomic variation may consist of a modification to a single letter, termed single nucleotide variants (SNVs), or rearrangements of larger regions, termed structural variants (SVs) [2,24]. For multi-cellular organisms, variants are often further classified into those which transmitted from parents to progeny, germline variants, or those which occur during cell division in the lifetime of an organism, somatic variants [29]. In humans, the accumulation of somatic mutations is commonly associated with the development of cancer [25] while the presence of certain germline variants has been shown to increase the susceptibility for certain types of cancer [27,28]. Beyond cancer, genomic variants are associated with many significant biological outcomes for individuals including a variety of diseases in humans [41, 45], flowering behavior in plants [46] and have contributed to rates of adaptation and the emergence of new species [18].

The detection of genomic variants such as SVs remains a challenging scientific and computational problem. Even with modern DNA sequencing technologies, it is not possible to construct the complete genome of every cell. As such, the common practice has been to construct a high-quality reference genome for each species and then annotate this reference with sites of variation [3, 14, 30]. The dominant method for identifying SNVs or SVs involves comparing fragments of DNA sequenced from a test (unknown) genome to a given reference (see Figure 1.1) [26, 33, 35, 43, 44]. SVs are typically detected through indirect evidence – a fragment that maps to a larger than expected distance – and as such they are more difficult to identify than SNVs which may be directly observed through alignment of sequences from the test genome

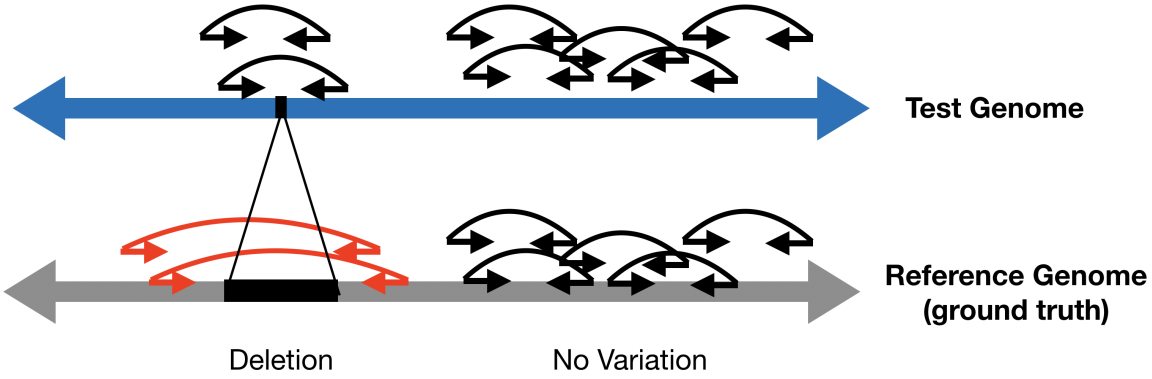


Figure 1.1: **Detecting Structural Variants.** To detect structural variants (SVs) in a test individual, fragments of DNA (black) are sampled from their (unknown) genome (top) and aligned to a reference genome (bottom). Fragments whose mappings are consistent with the underlying sampling process (right) suggest that test and reference genomes are the same. Fragments whose mappings are discordant indicate the presence of an SV. In our example (left) the test genome has a deletion relative to the reference. Two black fragments in the test genome that contain the deletion map to a much longer than expected length (red). Other variants, such as duplications and inversions, have their own unique discordant signal.

to the reference. However, predicting either type of variant is complicated by DNA sequencing and alignment errors. Because of these errors, algorithms for variant detection have suffered from high false-positive rates especially when the coverage – expected number fragments supporting each variant – is low [26, 35]. One hope for improving the ability to accurately predict SVs has come from methods that combine the information of many individuals [20]. This allows researchers to leverage large-scale public efforts, such as the 1000 Genomes Project, that have made available sequencing data from thousands of individuals, including parent-child trios [1, 23]. Population level algorithms have the potential to improve variant detection because the signal of true SVs will be boosted, but only when variants are likely to be shared among multiple individuals. Because of the massive population expansion, many variants in humans are rare and may only be shared by close relatives [19]. One approach for accurately detecting rare variants would be to simultaneously predict variants in a parent and a child. In particular, as shown in Figure 1.2, a parent and child will share many but not all SVs.

Our group has developed computational methods to improve SV prediction through considering pedigrees of related individuals [4, 6, 7, 9]. Our previous methods constrained the set of potential SVs through parent-child relationships by requiring that every variant present in the child was a germline variant transmitted from a parent. While these approaches have improved the ability to reduce false-positive predictions, they also increase the false-negative rate because they do not allow for novel variants (SVs that are not inherited from a parent) in the child genome.

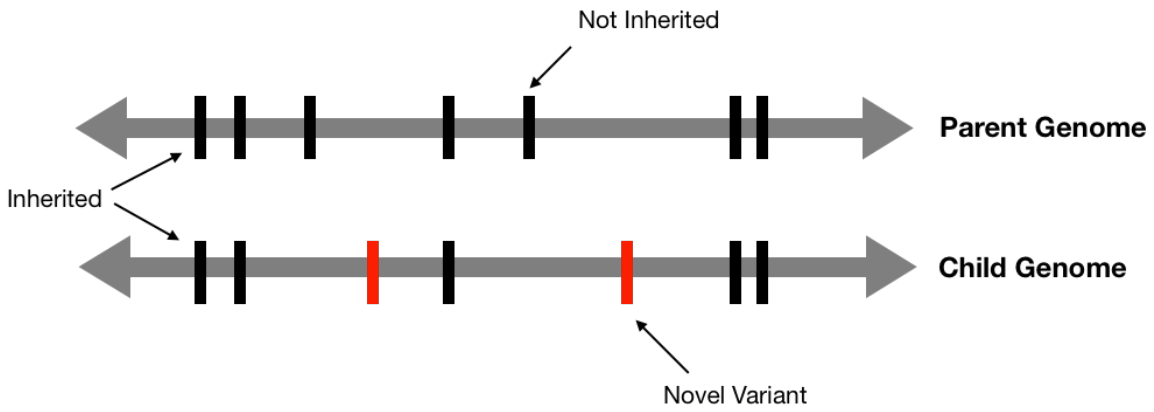


Figure 1.2: **Inheritance of Structural Variants.** Because germline structural variants (black and red bars) are transmitted from parents to their children, a child and parent will share many variants. However, because of recombination not all variants present in a parent's genome will be present in a child's genome and, although rare, a child may acquire novel variants not present in a parent (red).

The work presented in this thesis improves upon our previous work in two ways: First, we allow the child genome to possess novel variants. Secondly, we extend our assumptions to develop our model for diploid genomes. In Chapter 2 and 3 we will allow the child genome to possess novel variants. For simplicity in this work, we develop our model for haploid genomes so that at each potential SV site each individual either has the variant or does not. Chapter 2 focuses on building this framework for information from one parent and one child, while Chapter 3 extends this to include information from both parents. In Chapter 4 we make the step to consider diploid signals from each of the three individuals; however we do not allow the child to possess novel variants for simplicity. We consider a continuous relaxation of these discrete problems but favor sparse solutions through the use of the ℓ_1 norm. To demonstrate the effectiveness of the approach we present results from simulated data and data from the 1000 Genomes project. We also demonstrate that by a hierarchical approach it is possible to generalize our method to multiple generations. The results from the 1000 Genomes project were generated by Professor Mario Banuelos but are presented here for completeness.

Chapter 2

Haploid Genomes from Parent-Child Pairs with Novel Variants

Here we consider a general framework for detecting structural variants (SVs) given sequencing data from one parent (p) and one child (c). We assume that there are m locations in the genome that could be a potential SV for each individual. We assume that the variants in the child primarily come from the parent (inherited), but the child may have variants not present in the parent (novel). For simplicity, we consider each individual to be haploid (only one copy of each chromosome). As such, the true SV signal for the parent, $\vec{f}_p^* \in \{0,1\}^m$, has either a 0 at position j if the parent does not have an SV at location j or 1 otherwise. In contrast, the true SV signal for the child, $\vec{f}_c^* \in \{0,1\}^m$, comprises two vectors, i.e., $\vec{f}_c^* = \vec{f}_i^* + \vec{f}_n^*$, where $\vec{f}_i^* \in \{0,1\}^m$ is the vector of SVs that are inherited from the parent and $\vec{f}_n^* \in \{0,1\}^m$ is the vector of SVs that are novel. Specifically, the vector $\vec{f}_i^*$ has either a 1 at position j if an SV is inherited from the parent at position j or a 0 otherwise. Similarly, the vector $\vec{f}_n^*$ has a 1 if and only if there is a variant at position j that is not inherited from the parent and 0 otherwise. This approach is based on the paper [37, 39] coauthored with Professors Mario Banuelos, Roummel Marcia, and Suzanne Sindi published in *Methods*.

2.1 Observational model

The observed data are the number of DNA fragments supporting each potential SV, and the vectors $\vec{s}_p \in \mathbb{R}^m$ and $\vec{s}_c \in \mathbb{R}^m$ are the observation vectors of the parent and child, respectively. As in previous work [8, 21, 32], we assume that the data follow a Poisson distribution,

$$\begin{bmatrix} (\vec{s}_c)_j \\ (\vec{s}_p)_j \end{bmatrix} \sim \text{Poisson}\left(\begin{bmatrix} (\lambda_c - \epsilon)\left\{ (\vec{f}_i)_j + (\vec{f}_n)_j \right\} + \epsilon \\ (\lambda_p - \epsilon)(\vec{f}_p)_j + \epsilon \end{bmatrix} \right) \tag{2.1}$$

where $j \in \{1, 2, \ldots, m\}$, λ_p and λ_c are the sequencing coverage of the parent and child, respectively, and $\epsilon > 0$ is the measurement error corresponding to the sequencing and mapping processes. Let

$$\vec{s} = \begin{bmatrix} \vec{s}_c \\ \vec{s}_p \end{bmatrix} \quad \text{and} \quad \vec{f}^* = \begin{bmatrix} \vec{f}_i^* \\ \vec{f}_n^* \\ \vec{f}_p^* \end{bmatrix}.$$

Then the general observation model can be expressed as

$$\vec{s} \sim \text{Poisson}(A\vec{f}^* + \epsilon\mathbb{1}), \tag{2.2}$$

where $\mathbb{1} \in \mathbb{R}^{2m}$ is the vector of ones and $A \in \mathbb{R}^{2m \times 3m}$ is the coverage matrix given by

$$A = \begin{bmatrix} (\lambda_c - \epsilon)I_m & (\lambda_c - \epsilon)I_m & 0 \\ 0 & 0 & (\lambda_p - \epsilon)I_m \end{bmatrix},$$

where $I_m \in \mathbb{R}^{m \times m}$ is the $m \times m$ identity matrix.

2.2 Problem formulation

Under the Poisson process model (3.1), the probability of observing $\vec{s}$ is given by

$$p(\vec{s}\,|A\vec{f}^*) = \prod_{j=1}^{2m} \frac{((A\vec{f}^*)_j + \epsilon)^{\vec{s}_j}}{\vec{s}_j!} \exp\left(-(A\vec{f}^*)_j + \epsilon \right). \tag{2.3}$$

We use the maximum likelihood principle to determine the unknown Poisson parameter $A\vec{f}^*$ such that the probability of observing the vector of Poisson data $\vec{s}$ in (2.3) is maximized. Specifically, we minimize the corresponding negative Poisson log-likelihood function

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

In our approach for minimizing $F(\vec{f})$, we will apply gradient-based methods and relax the domain of $\vec{f}$. In particular, rather than enforcing $\vec{f}$ to be binary in value, i.e., $\vec{f} \in \{0, 1\}^{3m}$, we only require the values of $\vec{f}$ to lie between 0 and 1, i.e., $\mathbf{0} \leq \vec{f} \leq \mathbb{1}$.

2.3 Familial constraints

To improve the accuracy of our SV predictions, we incorporate additional constraints that exploit information about the signal $\vec{f}$. First, if the child has a structural variant, then it must be from the parent or it must be novel, but not both, i.e.,

$$\mathbf{0} \leq \vec{f_i} + \vec{f_n} \leq \mathbb{1}.$$

Second, if the child has a structural variant from the parent, then that SV must be present in the parent, i.e.,

$$\mathbf{0} \leq \vec{f_i} \leq \vec{f_p} \leq \mathbb{1}.$$

Finally, we enforce that if there is a novel SV present in the child, it cannot be present in the parent, i.e.,

$$\mathbf{0} \leq \vec{f_n} \leq \mathbb{1} - \vec{f_p}.$$

We will denote the set of all vectors satisfying these constraints by $\mathcal{S}$, i.e.,

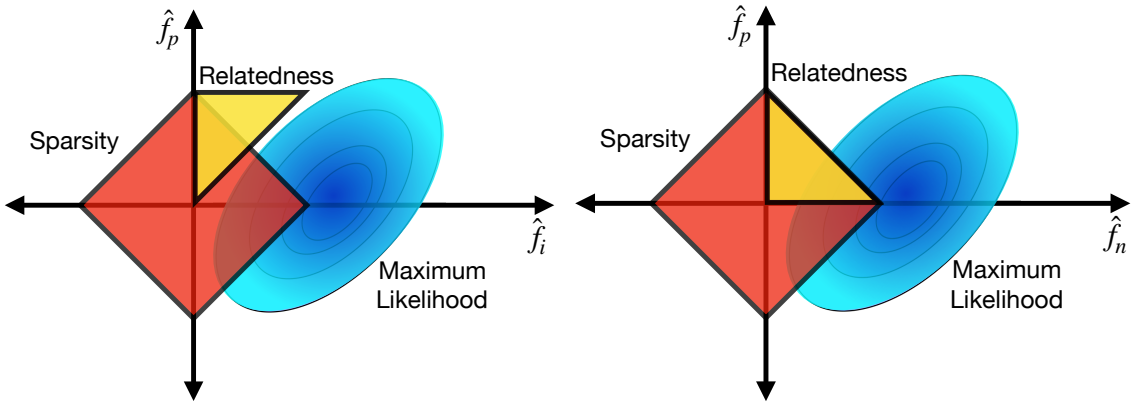
$$\mathcal{S} = \left\{ \begin{bmatrix} \vec{f_i} \\ \vec{f_n} \\ \vec{f_p} \end{bmatrix} \in \mathbb{R}^{3m} : \begin{array}{l} \mathbf{0} \leq \vec{f_i} + \vec{f_n} \leq \mathbb{1},\ \mathbf{0} \leq \vec{f_i} \leq \vec{f_p} \leq \mathbb{1}, \\ \mathbf{0} \leq \vec{f_n} \leq \mathbb{1} - \vec{f_p},\ \mathbf{0} \leq \vec{f_i}, \vec{f_n}, \vec{f_p} \leq \mathbb{1} \end{array} \right\}.$$

2.4 Sparsity

Structural variants are relatively rare in an individual's genome. Without incorporating how uncommon SVs are in a genome sequence, predictions result in false positives that mistake fragments that are incorrectly mapped to locations in the genome as SVs. In our work, we promote sparsity in our predictions by incorporating an ℓ_1-norm penalty term in our problem formulation, which is a common technique found in statistical literature [12, 13, 42]. What is particularly novel in our formulation is that while SVs are rare, SVs that are not inherited from a parent ($\vec{f_n}$ in our notation) are even rarer. To this end, we use two penalty terms: one for the parent SV ($\vec{f_p}$) and for the child SV inherited from the parent ($\vec{f_i}$), and another penalty term for the novel child SVs ($\vec{f_n}$). Mathematically, we express this penalty as

$$\text{pen}(\vec{f}) \ = \ \left(\|\vec{f_p}\|_1 + \|\vec{f_i}\|_1\right) + \gamma\|\vec{f_n}\|_1,$$

where $\gamma \gg 1$ is a penalty weight that places greater emphasis on $\vec{f_n}$ being much sparser than both $\vec{f_p}$ and $\vec{f_i}$, meaning the novel child SVs are much rarer than either the parent SVs or the inherited child SVs.



(a) No Novel Child Variants ($\hat{f}_n = 0$) (b) No Inherited Child Variants ($\hat{f}_i = 0$)

Figure 2.1: Illustration of feasible regions, sparsity penalty, and maximum likelihood surfaces for the two scenarios for child SVs: (a) When there is not a novel child variant ($\hat{f}_n = 0$), our approach reduces to our original model for germline structural variant prediction, where $0 \leq \hat{f}_i \leq \hat{f}_p \leq 1$, meaning an inherited child SV can only be present if the parent also has that SV. (b) When there is not an inherited child variant ($\hat{f}_i = 0$), a parent SV cannot be present where there is a novel child variant and vice versa, i.e., $0 \leq \hat{f}_p + \hat{f}_n \leq 1$.

2.5 Optimization problem

With these components defined, the genomic variants reconstruction problem has the following constrained optimization form:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{3m}}{\text{minimize}} \quad & F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} \quad & \vec{f} \in \mathcal{S} \end{aligned} \tag{2.4}$$

where $\tau > 0$ is a regularization parameter that balances the negative Poisson log-likelihood data fidelity term with the sparsity-promoting penalty term. This objective function in (2.4) will serve as the basis for each of the frameworks in the following chapters; however, F and $\mathcal{S}$ will change according to the assumptions we make in each chapter. Figure 2.1 provides a visualization of each of the components in our optimization framework: likelihood, sparsity and constraints.

We use the Sparse Poisson Intensity Reconstruction ALgorithm (SPIRAL) framework [16] to solve (2.4) by minimizing a sequence of quadratic models to the function $F(\vec{f})$. First we approximate $F(\vec{f})$ using a second-order Taylor series expansion at the current iterate $\vec{f}^k$:

$$F(\vec{f}) \approx F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^\top \nabla F(\vec{f}^k) + \tfrac{1}{2}(\vec{f} - \vec{f}^k)^\top \nabla^2 F(\vec{f}^k)(\vec{f} - \vec{f}^k). \tag{2.5}$$

The gradient of $F(\vec{f})$ is given by

$$\nabla F(\vec{f}) = \begin{bmatrix} \lambda_c\left(\mathbf{1} - D_c\vec{s}_c\right) \\ \lambda_c\left(\mathbf{1} - D_c\vec{s}_c\right) \\ \lambda_p\left(\mathbf{1} - D_p\vec{s}_p\right) \end{bmatrix},$$

where $\mathbf{1} \in \mathbb{R}^m$ is a column vector of ones and $D_c, D_p \in \mathbb{R}^{m\times m}$ are diagonal matrices with

$$\begin{aligned} (D_c)_{j,j} &= \frac{1}{\lambda_c(\vec{f}_i)_j + \lambda_c(\vec{f}_n)_j + \epsilon} \\ (D_p)_{j,j} &= \frac{1}{\lambda_p(\vec{f}_p)_j + \epsilon} \end{aligned}$$

for $1 \le j \le m$. We approximate the second-derivative Hessian matrix with a scalar multiple of the identity matrix $\alpha_k I$ where $\alpha_k > 0$ (see [10, 11] for details) and define the quadratic model

$$F^k(\vec{f}) \equiv F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^T\nabla F(\vec{f}^k) + \frac{\alpha_k}{2}||\vec{f} - \vec{f}^k||_2^2. \tag{2.6}$$

Now, each quadratic subproblem will be of the form

$$\begin{aligned} \vec{f}^{k+1} &= \underset{\vec{f}\in\mathbb{R}^{3m}}{\arg\min} \quad F^k(\vec{f}) + \tau\text{pen}(\vec{f}) \\ &\qquad \text{subject to} \quad \vec{f} \in \mathcal{S}. \end{aligned}$$

It can be shown that this constrained quadratic subproblem is equivalent to the following subproblem:

$$\begin{aligned} \vec{f}^{k+1} &= \underset{\vec{f}\in\mathbb{R}^{3m}}{\arg\min} \quad \mathcal{Q}(\vec{f}) = \tfrac{1}{2}\|\vec{f} - \vec{q}^{\,k}\|_2^2 + \tfrac{\tau}{\alpha_k}\text{pen}(\vec{f}) \\ &\qquad \text{subject to} \quad \vec{f} \in \mathcal{S}, \end{aligned} \tag{2.7}$$

where

$$\vec{q}^{\,k} = \begin{bmatrix} \vec{q}_i^{\,k} \\ \vec{q}_n^{\,k} \\ \vec{q}_p^{\,k} \end{bmatrix} = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k).$$

We note that the objective function $\mathcal{Q}(\vec{f})$ separates into the function

$$\mathcal{Q}(\vec{f}) = \sum_{j=1}^{m} \mathcal{Q}_j\big(\vec{f}_i, \vec{f}_n, \vec{f}_p\big),$$

where

$$\begin{aligned} \mathcal{Q}_j\big(\vec{f}_i, \vec{f}_n, \vec{f}_p\big) &= \tfrac{1}{2}\Big\{\big((\vec{f}_i - \vec{q}_i^{\,k})_j\big)^2 + \big((\vec{f}_n - \vec{q}_n^{\,k})_j\big)^2 + \big((\vec{f}_p - \vec{q}_p^{\,k})_j\big)^2\Big\} \\ &\quad + \tfrac{\tau}{\alpha_k}\Big\{\ |(\vec{f}_p)_j| + |(\vec{f}_i)_j| + \gamma|(\vec{f}_n)_j|\ \Big\}. \end{aligned}$$

Since the bounds that define the feasible set $\mathcal{S}$ are component-wise, then (2.7) separates into subproblems of the form

$$\begin{aligned} \underset{f_i,f_n,f_p\in\mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_i-q_i)^2+\tfrac{1}{2}(f_n-q_n)^2+\tfrac{1}{2}(f_p-q_p)^2 \\ & +\tfrac{\tau}{\alpha_k}|f_p| \;+\; \tfrac{\tau}{\alpha_k}|f_i| \;+\; \tfrac{\gamma\tau}{\alpha_k}|f_n| \\ \text{subject to} \quad & 0\le f_i+f_n\le 1, \quad 0\le f_i\le f_p\le 1, \\ & 0\le f_n\le 1-f_p, \quad 0\le f_i,f_n,f_p\le 1, \end{aligned} \tag{2.8}$$

where $\{f_i, f_n, f_p\}$ and $\{q_i, q_n, q_p\}$ are scalar components of the vectors $\{\vec{f_i}, \vec{f_n}, \vec{f_p}\}$ and $\{\vec{q_i}, \vec{q_n}, \vec{q_p}\}$, respectively, at the same location. Completing the squares and ignoring constant terms, the optimization problem (3.3) can be expressed as

$$\begin{aligned} \underset{f_i,f_n,f_p\in\mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_i-a)^2+\tfrac{1}{2}(f_n-b)^2+\tfrac{1}{2}(f_p-c)^2 \\ \text{subject to} \quad & 0\le f_i+f_n\le 1, \quad 0\le f_i\le f_p\le 1, \\ & 0\le f_n\le 1-f_p, \quad 0\le f_i,f_n,f_p\le 1, \end{aligned} \tag{2.9}$$

where $a = q_i - \frac{\tau}{\alpha_k}$, $b = q_n - \frac{\gamma\tau}{\alpha_k}$ and $c = q_p - \frac{\tau}{\alpha_k}$. The unconstrained minimizer of (2.9) is (a, b, c). If (a, b, c) is feasible with respect to the constraints, then it is also the constrained minimizer. If (a, b, c) is not feasible, then we obtain the feasible solution to (2.9) by orthogonally projecting (a, b, c) onto the three-dimensional feasible set, which is illustrated in Fig. 2.2. In particular, the f_i-f_n-f_p three-dimensional space partitions into **15** different regions that projects onto a vertex, edge, or surface of the feasible set for infeasible points. Tables 2.1 and 2.2 enumerate and define the regions of interest and the corresponding projections.

2.5.1 Forward and Backward Hierarchical Approaches

In application, observations for related individuals may span multiple generations. As such, we propose two approaches to address prediction of novel child variants. In the case we have observations $\vec{y}_c, \vec{y}_p$, and $\vec{y}_{gp}$, where $\vec{y}_{gp}$ is the observation vector of the grandparent signal, we describe both of these approaches below.

Forward Hierarchical (FH) Approach:

Step 1: Given $\vec{y}_p$ and $\vec{y}_{gp}$, reconstruct $\vec{f_p}$ and $\vec{f_{gp}}$.

Step 2: Use $\vec{f_p^0} \equiv \vec{f_p}$ from Step 1 as initialization to reconstruct $\vec{f_i}$, $\vec{f_n}$, and $\vec{f_p}$.

Backward Hierarchical (BH) Approach:

Step 1: Given $\vec{y}_c$ and $\vec{y}_p$, reconstruct $\vec{f_i}$, $\vec{f_n}$, and $\vec{f_p}$.

Step 2: Use $\vec{f_i^0} \equiv \vec{f_i}$ and $\vec{f_n^0} \equiv \vec{f_n}$ from Step 1 as initialization to reconstruct $\vec{f_i}$, $\vec{f_n}$, and $\vec{f_{gp}}$.

We note that for the backward hierarchical approach, the final novel variants are those not present in the grandparent signal.

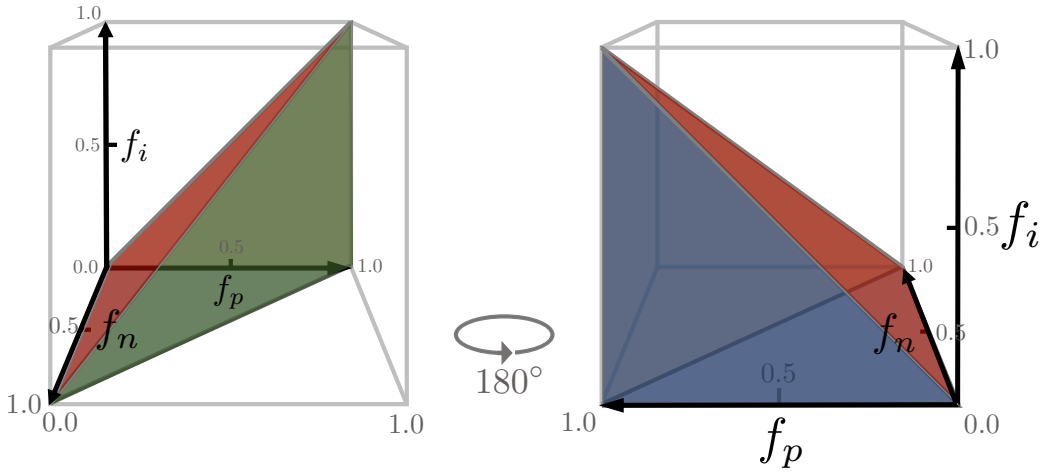


Figure 2.2: The three-dimensional feasible region of the minimization problem (2.9) on the f_i-f_n-f_p axis. Because novel child SVs are not present in the parent genome, i.e., $f_n \leq 1 - f_p$, $f_n \to 0$ as $f_p \to 1$. Similarly, because inherited SVs come from the parent genome, i.e., $f_i \leq f_p$, $f_i \to 0$ as $f_p \to 0$. Finally, because novel and inherited child SVs are mutually exclusive, i.e., $f_n + f_i \leq 1$, $f_n \to 0$ as $f_i \to 1$ and vice versa. These define the vertices of the feasible region, which is a polytope since the constraints are linear. Subproblem minimizers not satisfying the constraints are orthogonally projected onto this feasible region.

2.6 Results

2.6.1 Implementation Details

We implemented our method for variant detection in MATLAB by extending our previous approach [9] based on the SPIRAL method [16]. We next analyze the performance of our method on both simulated and real data. We compared the performance of our new method with two other variant prediction methods. First, we compare to our previously published method for variant prediction in the context of one-parent/one-child [8]. This method had a similar sparsity-promoting term τ, but required all predictions in the child to occur in the parent (i.e., did not allow for novel variants in the child). Second, we compare to the same method but with only sparsity constraints (i.e., no family constraints). When choosing τ we have found that there are a wide range of acceptable values for which the reconstruction will work well. If we are reconstructing the genome of a species for which the sparsity of the variants is well documented, then we can make an informed choice for τ. However if the sparsity is not well known, then we can choose a τ value in a reasonable range to enforce the sparsity. The regularization parameters τ were chosen to be the same for all methods and, when showing results for our new method, γ was chosen to maximize the area under the curve (AUC). In all cases, the SPIRAL algorithm was run with the same terminating criteria, if the relative difference between consecutive iterates

	Projection	a	b	c
Interior	(a, b, c)	$0 \leq a \leq c$	$0 \leq b$	$b + c - 1 \leq 0$
Vertex	$(0, 0, 0)$	$a \leq -c$	$b \leq 0$	$c \leq 0$
	$(0, 0, 1)$	$a \leq 0$	$b \leq c - 1$	$1 \leq c$
	$(0, 1, 0)$	$a \leq b - c - 1$	$1 \leq b$	$c \leq b - 1$
	$(1, 0, 1)$	$1 \leq a$	$b \leq a + c - 2$	$2 - a \leq c$
Edge	$(0, b, 0)$	$a < -c$	$0 < b < 1$	$c < 0$
	(u_1, v_1, w_1)	$2 - 2b - c < a$	$b < 1 + a + c$	$c < 2a - 1 + b$ $c < 2 - a + b$
	$(0, v_2, w_2)$	$a < 0$	$b < 1 + c$	$1 - b < c < b + 1$
	$(u_3, 0, w_3)$	$c < a < 2 - c$	$b < 0$	$-c < a$
Surface	(a, v_4, w_4)	$0 \leq a$	$b \leq 1 - 2a + c$	$b - 1 \leq c \leq b + 1$
	(u_5, b, w_5)	$\lvert c \rvert \leq a$	$0 \leq b \leq 1$	$c \leq -a - 2b + 2$

Table 2.1: The partitioning of the f_i-f_n-f_p space and the corresponding orthogonal projections onto the feasible set. The projection of the unconstrained minimizer (a, b, c) is the minimizer of (2.9). Projections onto edges and surfaces are represented as linear combinations of a, b, and c in Table 2.2.

converged to $\|\vec{f}_{k+1} - \vec{f}_k\|_2 / \|\vec{f}_k\|_2 \leq 10^{-8}$. For each trio, the numerical experiments took on average of 6 minutes to run in serial on a commodity machine. In contrast, in real experiments, the SV-caller GASV took an average of 180 minutes to process the .BAM files and 1.5 minutes to generate candidate SVs for each trio. In other words, the memory footprint of our method is extremely low and does not result in fatalistic warnings. In particular, the main computational overhead is in the generation of predictions. We are currently developing an open-source, parallel version of our method, but our MATLAB code is available upon request.

2.6.2 Simulated Experiments

Because our model was developed in the simplified assumption of one-parent and one-child with haploid genomes, before applying it to real human data violating our

	Projection	u	v	w
Edge	(u_1, v_1, w_1)	$\frac{1}{3}(1+a-b+c)$	$\frac{1}{3}(2-a+b-c)$	$\frac{1}{3}(1+a-b+c)$
	$(0, v_2, w_2)$	0	$\frac{1}{2}(1+b-c)$	$\frac{1}{2}(1-b+c)$
	$(u_3, 0, w_3)$	$\frac{1}{2}(a+c)$	0	$\frac{1}{2}(a+c)$
Surface	(a, v_4, w_4)	a	$\frac{1}{2}(1-c+b)$	$\frac{1}{2}(1+c-b)$
	(u_5, b, w_5)	$\frac{1}{2}(c+a)$	b	$\frac{1}{2}(c+a)$

Table 2.2: Orthogonal projections (u, v, w) of the unconstrained minimizer (a, b, c) onto the surfaces and edges of the feasible set.

assumptions, we studied its performance on data we simulated to match our assumptions. For simplicity we do not directly simulate the generation and mapping of reads, we only generated the sequencing depth (or coverage). In these cases we simulated the true signal for a parent and child by creating a vector of 10^5 potential SVs and selecting 500 locations to be true variants for the parent and child signal separately. We selected 500 locations uniformly at random to be the true SVs in the parent. The child signal was then generated by randomly selecting $\lfloor 500\rho \rfloor$ of the parent variants to be inherited (where ρ is the percent overlap between parent and child SVs) and then choosing $(500 - \lfloor 500\rho \rfloor)$ locations from the remaining $(10^5 - 500)$ locations that were not chosen as a parent variant to be novel variants in the child. In our experiments, we chose $0.5 \leq \rho \leq 1$.

Analysis

When the percentage of novel variants is ($< 10\%$) in the child, our method is better able to reconstruct the child signal. Hence, we are able to more accurately recover the SVs in the child reconstruction when we allow for novel variants. Figure 2.3 illustrates how our proposed method can adequately recover the parent signal under the assumptions that the novel variants are far more rare than the inherited variants in the child. As we were running test cases we noticed the trend that if we allow for a larger number of novel variants in the child ($\approx 50\%$), then our reconstruction is more reliant on the depth of the sequencing coverage. In these cases we need higher sequencing coverage ($\approx 10\times$) for both individuals to accurately recover their signals.

2.6.3 1000 Genomes Project Trio Data

To test our proposed method of novel variant detection, we apply our method to both father-mother-daughter trios sequencing data from the 1000 Genomes Project [3]. In the pilot study of the project, both the CEU (European ancestry) and YRI

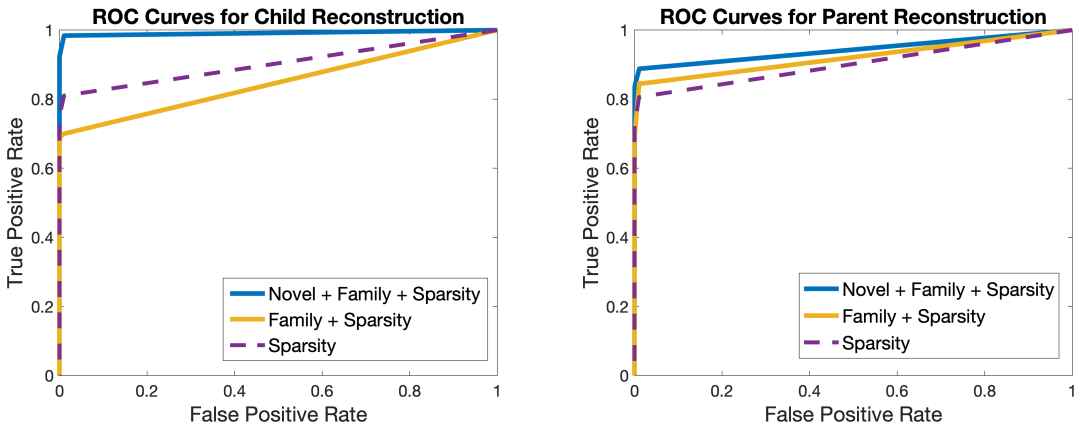


Figure 2.3: *Left.* ROC curves of three methods illustrating the false positive rate vs. the true positive rate in the simulated child reconstruction, where $\tau = 20$ and $\gamma = \frac{3}{2}$. *Right.* ROC curves of three methods illustrating the false positive rate vs. the true positive rate in the simulated parent reconstruction, where $\tau = 20$ and $\gamma = \frac{3}{2}$. These simulations were done with sequencing coverage of 4 in both individuals and $\rho = 0.9$ (so the child has 50 novel variants).

(Yoruba population) genomes were sequenced at high coverage using three sequencing platforms with a mean mapped depth of 43.14, and 40.05, respectively. This data was subsequently subsampled to $\approx 4\times$ coverage and aligned to NCBI36. In particular, we use the .bam files corresponding to SLX (Illumina Genome Analyser ABI SOLiD system), with 36 - 50 bp reads. We incorporate the SV-caller GASV to obtain candidate variant positions for all six individuals [34]. The preprocessing of GASV, BamToGASV, was run with default settings and candidate variants were obtained with the –batch option in GASV for candidate deletions. In addition to comparing our method to other constrained models (i.e., sparsity and sparsity with family constraints), we benchmark our work against GASV output by thresholding at each observed number of fragments supporting a potential SV. As such, our model mitigates the high false positive rates of previous SV-calling tools.

For the true signals $\vec{f}^*$, the study reported deletions passing filters associated with a post-beagle 95% confident call rate and a Hardy-Weinberg equilibrium p-value $<$ 0.01 in each of the populations. Additionally, we filter out *LowQual* deletions near centromeres or telomeres longer than 250bp of the reported validated deletion set. Moreover, variants in the child signal not in one of the parents represent the the novel deletion signal we aim to reconstruct. In particular, the child has an average of 8.55% and 6.26% novel variants (of total variants) for the YRI and CEU trios, respectively.

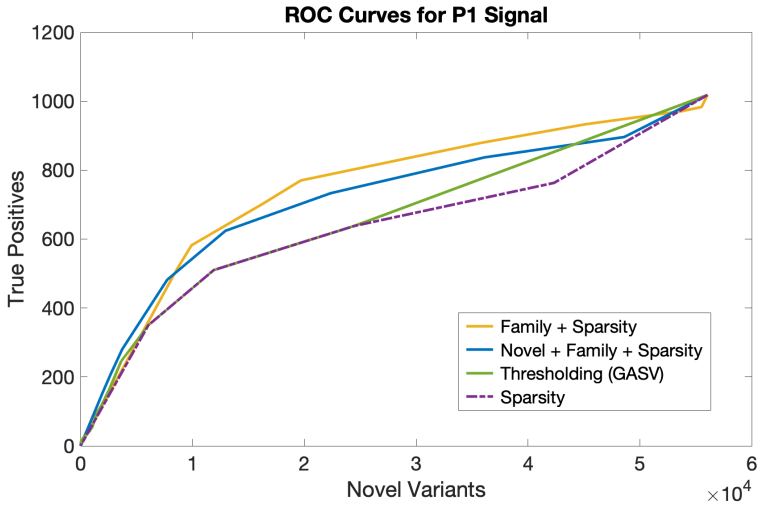


Figure 2.4: ROC curves of four methods illustrating the novel deletions (validated set of deletions may be incomplete) vs. true positives in the signal of the CEU parent NA12891, where $\tau = 0.0129$ and $\gamma = 10$. We observe an increase of true positive predictions when the number of novel predictions < 1000.

Analysis

For parent signal reconstructions, we note an initial improvement in true positive prediction of our proposed model when the number of novel predictions is low. Figure 2.4 illustrates our findings for the CEU parent NA12891. Although the area under the curve (AUC) for the ROC curve is less for our proposed method, we note an improvement from simple thresholding techniques (GASV). Moreover, constraints from our initial model favor parent signal recovery [8]. Next, we focus on the reconstruction of the entire child signal $\vec{f}_c$. Figure 2.5 illustrates the novel variant predictions against validated deletions in the YRI child NA19240 considering the same four methods. We observe comparable results with enforcing only sparsity (i.e., no inheritance constraints) and an improvement over previous methods. Since the rate of novel variants is less than 10% in this low coverage regime, this is consistent with our simulated experiments.

2.6.4 Platinum Genomes

We also apply our method to low-coverage ($\approx 5\times$) sequencing data for the three-generation, 17-member CEU pedigree (dbGaP accession phs001224.v1.p1) using the same four models as before [14]. All 17 family members' DNA was originally sequenced on an Illumina HiSeq2000 to an average depth of $50\times$ using 2×100 bp reads and PCR-free sample preparation. Although originally sequenced at high coverage, we use Samtools to subsample and achieve low coverage of approximately $5\times$ [22].

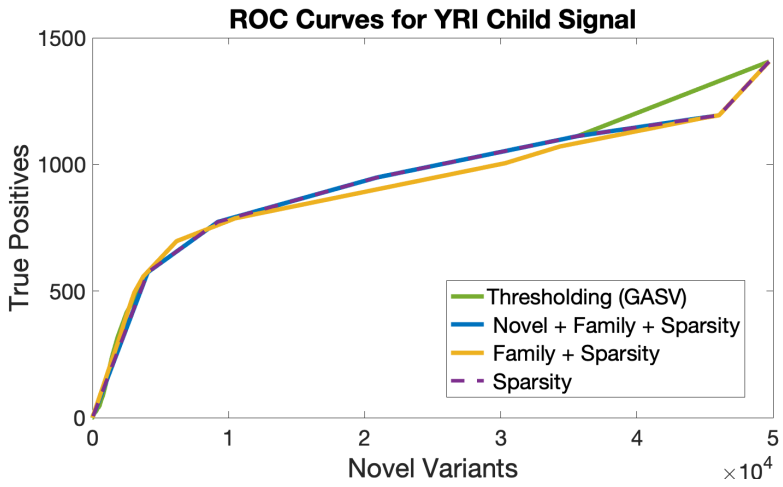


Figure 2.5: ROC curves of four methods illustrating the novel deletions vs. true positives in the combined child signal $\vec{f_c}$ of the child in the YRI trio (NA19240), where $\tau = 1$ and $\gamma = 10$. We note comparable performance of our proposed model with only enforcing sparsity.

We determine true SVs with the intersection of GASV and Delly SV calls [31, 34]. In particular, we look at the deletions from the grandparent-parent-child (NA12889, NA12877, and NA12882) and apply our method using the proposed hierarchical approaches. As before, we benchmark our method by comparing to the thresholding of GASV candidate structural variants.

Analysis

For both the forward and backward hierarchical approaches, we find similar patterns for parent and grandparent signal reconstruction, namely less predictive power of true positives. Fig 2.6 illustrates the novel child signal reconstructions for NA12882. We note that the forward hierarchical (FH) approach achieves competitive AUC values when initializing the parent signal from one application of our method (with parent with the grandparent signals). We highlight that the backward hierarchical (BH) method results in an increase of the true positive predicted for the novel child signal. The BH approach is also compared to applying the method once (with child and parent observations) and note that it outperforms all other methods. When we considered higher coverage in this data set ($\approx 10\times$), we observe similar performance for the backward hierarchical approach for novel deletions and less improvement in sensitivity when compared to thresholding GASV deletion call set (data not shown).

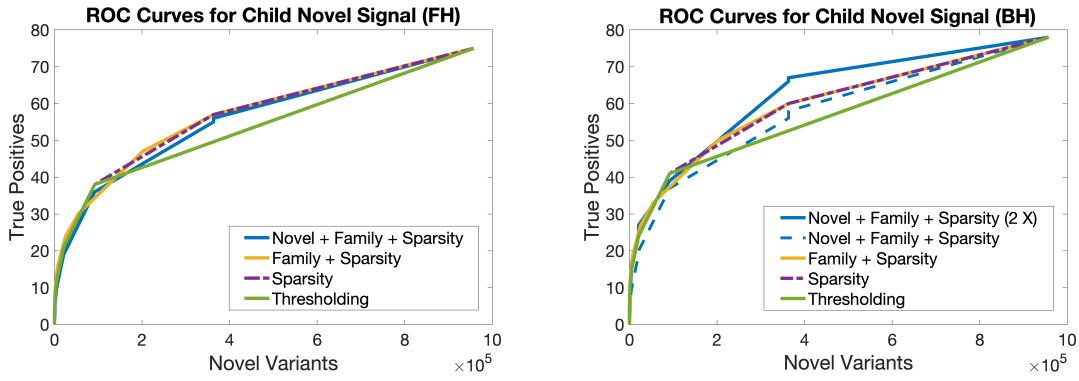


Figure 2.6: ROC curves of four methods illustrating the novel deletions vs. true positives in the signal of the CEU child NA12882, where $\tau = 0.0129$ and $\gamma = 10$. *Left.* Using the forward hierarchical (FH) approach, we observe comparable detection of novel variants. *Right.* With the backward hierarchical (BH) approach, we note an increase in true positive rate compared to applying our method only once (dashed blue line).

2.7 Conclusions

We propose a new method to detect novel structural variants – SVs present in a child not inherited from a parent – from sequencing data in parent-child pairs. Our method incorporates both relatedness and sparsity constraints, allowing for varying penalty parameters in the reconstruction of the child signal. By doing so, our new model is less sensitive to our regularization parameters. Although parent signal recovery resulted in reduced predictive capacity, our proposed method improved true positive predictions in the child. We present our results for both simulated, real data from the 1000 Genomes Project and a subset of the Platinum Genomes, and suggest further exploration in varying sequencing coverage for future parent-offspring data. In future studies, we intend to incorporate other SV-calling tools, larger family structures, and a general relatedness parameter in our methods.

Chapter 3

Haploid Genomes from Parent-Child Trios with Novel Variants

We now describe a structural variant (SV) detection framework given genomic data from both parents (p_1, and p_2) and from one child (c). Let $\vec{f}_I^* \in \mathbb{R}^m$ be the vector of m locations of potential SVs for each individual $I \in \{p_1, p_2, c\}$. We make the following assumptions:

- **Inherited variants:** The variants in the child primarily come from the parents. In particular, if both parents have an SV at a particular location, the child must also have an SV at that location. Furthermore, we assume SVs are rare.

- **Novel variants:** On rarer occasions, the child may have variants not present in either parent.

- **Haploid genotype:** For simplicity, we consider each individual to be haploid (only one copy of each chromosome).

- **Low-coverage sequencing:** The expected number of fragments supporting each variant is low, and the observed measurements are governed by a Poisson process.

We denote the true SV signal for either parent, $P \in \{p_1, p_2\}$, by $\vec{f}_P^* \in \{0,1\}^m$, which has either a 1 at position j if the parent has an SV at location j or 0 otherwise for $j = 0, \ldots, m$. In contrast, the true SV signal for the child, $\vec{f}_c^* \in \{0,1\}^m$, is composed of two vectors:

$$\vec{f}_c^* = \vec{f}_i^* + \vec{f}_n^*,$$

where $\vec{f}_i^* \in \{0,1\}^m$ and $\vec{f}_n^* \in \{0,1\}^m$ denote the vector of SVs that are inherited from either parent and are novel, respectively. In particular, $(\vec{f}_i^*)_j$ has either a 1 if an SV

is inherited from the parent at position j or a 0 otherwise. Similarly, $(\vec{f_n^*})_j$ has a 1 if and only if there is a variant at position j that is not inherited from the parent and 0 otherwise. We note that at each location, $\vec{f_i}$ and $\vec{f_n}$ cannot simultaneously be both non-zero since a child variant can only either be inherited or be novel, but not both. In other words, the vectors $\vec{f_i^*}$ and $\vec{f_n^*}$ satisfy the complementary condition $(\vec{f_i^*})_j(\vec{f_n^*})_j = 0$ for $1 \leq j \leq m$.

This approach is based on the paper [38] coauthored with Professors Mario Banuelos, Roummel Marcia, and Suzanne Sindi published in the conference proceedings for 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA).

3.1 Observation model

We denote the vector of observations and the vector of true SV signals by $\vec{s} = [\vec{s}_c; \vec{s}_{p_1}; \vec{s}_{p_2}]$ and $\vec{f^*} = [\vec{f_i^*}; \vec{f_n^*}; \vec{f_{p_1}^*}; \vec{f_{p_2}^*}]$, where the entries in the measurement vector $\vec{s}_I$ correspond to the number of DNA fragments supporting each potential SV, and the vector $\vec{s}_I \in \mathbb{R}^m$, where $I \in \{c, p_1, p_2\}$, is the observation vectors for each individual. Because we assume that the sequence coverage is low, we expect that the number of fragments covering any position in the genome to follow a Poisson distribution (see e.g., [8, 35]).

In particular, we can express the general observation model as

$$\vec{s} \sim \text{Poisson}(\mathbf{A}\vec{f^*} + \epsilon\mathbb{1}), \tag{3.1}$$

where $\mathbb{1} \in \mathbb{R}^{3m}$ is the vector of ones and $\mathbf{A} \in \mathbb{R}^{3m \times 4m}$ is the coverage matrix given by

$$\mathbf{A} = \begin{bmatrix} (\lambda_C - \epsilon)I_m & (\lambda_C - \epsilon)I_m & 0 & 0 \\ 0 & 0 & (\lambda_F - \epsilon)I_m & 0 \\ 0 & 0 & 0 & (\lambda_M - \epsilon)I_m \end{bmatrix},$$

where $I_m \in \mathbb{R}^{m \times m}$ is the $m \times m$ identity matrix.

3.2 Problem formulation

We use the maximum likelihood principle to determine $\vec{f^*}$ such that the probability of observing the vector of Poisson data $\vec{s}$ in (3.1) is maximized. More precisely, we minimize the corresponding negative Poisson log-likelihood function

$$F(\vec{f}) = \sum_{j=1}^{3m} \left\{ \left(\mathbf{A}\vec{f}\right)_j - \vec{s}_j \log\left((\mathbf{A}\vec{f})_j + \epsilon \right) \right\}.$$

To apply gradient-based optimization approaches for minimizing $F(\vec{f})$, we allow $\vec{f}$ to take on more than the binary values of 0 and 1 and instead be continuous in the interval $[0, 1]$.

3.3 Feasibility constraints

We impose the following constraints on the SV signal estimate $\vec{f}$, which correspond to the biological assumptions we make:

- Since each entry in an individual's SV signal is binary, i.e., $\vec{f}_c^*, \vec{f}_{p_1}^*, \vec{f}_{p_2}^* \in \{0,1\}^m$, and since $\vec{f}_c^* = \vec{f}_i^* + \vec{f}_n^*$ with $\vec{f}_i^*, \vec{f}_n^* \in \{0,1\}^m$, then we have $\mathbf{0} \leq \vec{f}_i, \vec{f}_n, \vec{f}_{p_1}, \vec{f}_{p_2} \leq \mathbf{1}$ and $\mathbf{0} \leq \vec{f}_i + \vec{f}_n \leq \mathbf{1}$.

- Because a novel variant in the child cannot be inherited from either parent, we have $\mathbf{0} \leq \vec{f}_n \leq \mathbf{1} - \vec{f}_{p_1}$ and $\mathbf{0} \leq \vec{f}_n \leq \mathbf{1} - \vec{f}_{p_2}$.

- If both parents have an SV, then the child must inherit the same SV: $\vec{f}_{p_1} + \vec{f}_{p_2} - 1 \leq \vec{f}_i$. Similarly, if neither parent has an SV, then the child cannot have an inherited SV: $\vec{f}_i \leq \vec{f}_{p_1} + \vec{f}_{p_2}$.

We will denote the set of $\vec{f}$ satisfying these constraints by $\mathcal{S}$.

3.4 Optimization Setup

With these components defined, the genomic variants reconstruction problem has the following constrained optimization form:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{4m}}{\text{minimize}} \quad & F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} \quad & \vec{f} \in \mathcal{S} \end{aligned} \tag{3.2}$$

where $\text{pen}(\vec{f})$ is a penalty that promotes sparsity in $\vec{f}$ and $\tau > 0$ is a regularization parameter that balances the negative Poisson log-likelihood term with the sparsity-promoting penalty term. We use the Sparse Poisson Intensity Reconstruction ALgorithm (SPIRAL) framework [16, 37] to solve (2.4), which involves solving a sequence

of scalar quadratic subproblems of the form

$$\begin{aligned} \underset{\substack{f_i,f_n,f_{p_1},f_{p_2}\\ \in\mathbb{R}}}{\text{minimize}} \quad & \tfrac{1}{2}(f_i-q_i)^2+\tfrac{1}{2}(f_n-q_n)^2+\tfrac{\tau}{\alpha_k}|f_i|+\tfrac{\tau\gamma}{\alpha_k}|f_n| + \\ & \tfrac{1}{2}(f_{p_1}-q_{p_1})^2 + \tfrac{1}{2}(f_{p_2}-q_{p_2})^2+\tfrac{\tau}{\alpha_k}|f_{p_1}|+\tfrac{\tau}{\alpha_k}|f_{p_2}| \\ \text{subject to} \quad & 0 \le f_i, f_n, f_{p_1}, f_{p_2} \le 1, \quad 0 \le f_i + f_n \le 1 \\ & 0 \le f_n \le 1 - f_{p_1}, \qquad 0 \le f_n \le 1 - f_{p_2}, \\ & f_{p_1} + f_{p_2} - 1 \le f_i \le f_{p_1} + f_{p_2}. \end{aligned} \tag{3.3}$$

where at each iteration k,

- $\{f_i, f_n, f_{p_1}, f_{p_2}\}$ and $\{q_i, q_n, q_{p_1}, q_{p_2}\}$ are scalar components of the vectors $\vec{f}^{\,k} = \{\vec{f}_i^{\,k}, \vec{f}_n^{\,k}, \vec{f}_{p_1}^{\,k}, \vec{f}_{p_2}^{\,k}\}$ and $\vec{q}^{\,k} = \{\vec{q}_i^{\,k}, \vec{q}_n^{\,k}, \vec{q}_{p_1}^{\,k}, \vec{q}_{p_2}^{\,k}\}$, respectively, at the same location;

- α_k is the learning rate;

- $\vec{q}^{\,k} = \vec{f}^{\,k} - \frac{1}{\alpha_k}\nabla F(\vec{f}^{\,k})$ is the predicted new iterate along the steepest descent (negative gradient) from the current iterate with step length $1/\alpha_k$;

- $0 < \gamma < 1$ is a parameter that further amplifies sparsity on novel child SVs.

Note that because the constraints are more complex in (3.3) than in our previous work, we must use a different approach.

3.5 Optimization approach

We propose using an alternating block-coordinate descent approach to solve (3.3). Specifically, the proposed method solves (3.3) by alternating between child and parent indicator variables. First, we fix the parent structural variant signals, f_{p_1} and f_{p_2}, and solve the resulting minimization problem for the child signal, f_i and f_n. Next, we fix the child signal and minimize over the parent indicator variables. The method continues until the difference between subsequent iterates falls below a specified threshold. We outline the steps below.

Step 0: Initially, we fix the values for the parent indicator variables by setting $f_{p_1}^{(0)} = f_{p_2}^{(0)} = 0.5$ for each candidate SV location.

Step 1: Suppose we have obtained $\hat{f}_{p_1}^{(j-1)}$ and $\hat{f}_{p_2}^{(j-1)}$ from the previous iteration. The

child indicator variables $\hat{f}_i^{(j)}$ and $\hat{f}_n^{(j)}$ are obtained from solving

$$\begin{aligned} \underset{f_i, f_n \in \mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_i - c_i)^2 + \tfrac{1}{2}(f_n - c_n)^2 \\ \text{subject to} \quad & 0 \le f_i + f_n \le 1 \\ & 0 \le f_n \le \min\left(1 - \hat{f}_{p_1}^{(j-1)}, 1 - \hat{f}_{p_2}^{(j-1)}\right) \\ & \max\left(0, \hat{f}_{p_1}^{(j-1)} + \hat{f}_{p_2}^{(j-1)} - 1\right) \le f_i \\ & f_i \le \min\left(1, \hat{f}_{p_1}^{(j-1)} + \hat{f}_{p_2}^{(j-1)}\right), \end{aligned} \tag{3.4}$$

where $c_i = q_i - \frac{\tau}{\alpha_j}$ and $c_n = q_n - \frac{\gamma\tau}{\alpha_j}$. The feasible region is shown in Fig. 3.1(a).

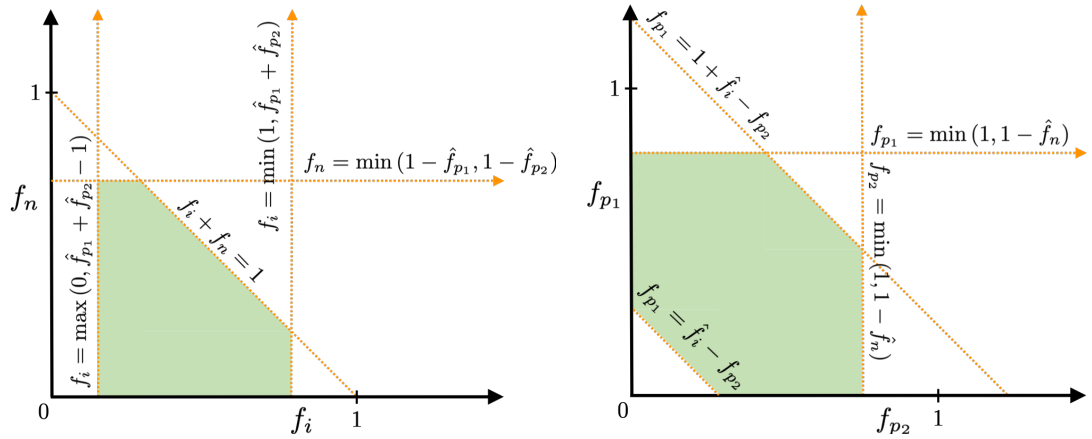


Figure 3.1: The feasible set (indicated by the shaded region) for each step of the proposed block-coordinate minimization approach. (a) In Step 1, we minimize over the child indicator variables f_i and f_n given fixed parent indicator variables $\hat{f}_{p_1}$ and $\hat{f}_{p_2}$. (b) In Step 2, we minimize over the parent indicator variables f_{p_1} and f_{p_2} given fixed child indicator variables $\hat{f}_n$ and $\hat{f}_i$.

Step 2: Suppose we have obtained $\hat{f}_i^{(j)}$ and $\hat{f}_n^{(j)}$ from Step 1. To obtain the solution for the current iteration $\hat{f}_{p_1}^{(j)}$ and $\hat{f}_{p_2}^{(j)}$, we have

$$\begin{aligned} \underset{f_{p_1}, f_{p_2} \in \mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_{p_1} - c_{p_1})^2 + \tfrac{1}{2}(f_{p_2} - c_{p_2})^2 \\ \text{subject to} \quad & 0 \le f_{p_1} \le \min\left(1, 1 - \hat{f}_n^{(j)}\right), \\ & 0 \le f_{p_2} \le \min\left(1, 1 - \hat{f}_n^{(j)}\right), \\ & f_{p_1} + f_{p_2} - 1 \le \hat{f}_i^{(j)} \le f_{p_1} + f_{p_2}, \end{aligned} \tag{3.5}$$

where $c_{p_1} = q_{p_1} - \frac{\tau}{\alpha_j}$ and $c_{p_2} = q_{p_2} - \frac{\tau}{\alpha_j}$. The feasible region is shown in Fig. 3.1(b).

We note that both problems (3.4) and (3.5) have closed form solutions, where the minimizer is obtained by projecting the unconstrained solution to the feasible set (see e.g., [37]).

3.6 Results

We implemented our method for variant detection in Matlab by extending our previous approach [37] based on the SPIRAL method [16]. We analyze the performance of our method on both simulated and real data by comparing our new method with two other variant prediction methods. We compare our previous method for variant prediction in the context of two-parents/one-child [7]. This method includes a sparsity promoting term τ, but did not specifically model novel variants in the child. Second, we include a comparison to the model that only enforces sparsity. The regularization parameter τ was chosen to be the same for all methods and γ was chosen when the area under the curve (AUC) was maximized. Each model was run with the same terminating criteria, checking if the relative difference between consecutive iterates converged to $\|\vec{f}_{k+1} - \vec{f}_k\|_2 / \|\vec{f}_k\|_2 \leq 10^{-8}$.

3.6.1 Simulated Data

Because our model was developed in the simplified assumption of two-parent and one-child with haploid genomes, before applying it to real human data violating our assumptions, we studied its performance on data we simulated to match our assumptions. In these cases we simulated the true signal for both parents and the child and varied the fraction of similarity between parents and the number of novel variants in the child to study the performance of our model. We first created the parent signals and then derived the child with its novel variants. Each simulated true signal consisted of 10^5 potential SVs. For the parents, 500 locations were chosen at random to be true variants; the fraction of variants the parents had in common was varied according to their chosen percent similarity. For the child signal, if both parents had an SV at a particular location the child signal did as well. If only one parent had an SV at a location, the child had a 50% chance of inheriting that SV. Novel variants in the child were chosen randomly from locations where no parent had an SV. From these true signals, observed signals were created by sampling from the Poisson distribution with a given coverage and error.

Analysis. When the percentage of novel variants is small in the child ($< 10\%$), we observe better performance of our new method. In Figure 3.2 we show an ROC curve for a simulated data set where the parents were chosen to have 50% similarity and the child had 50 novel variants. We note that the area under the curve for our proposed method is higher than our other methods for the child reconstruction. Hence, we are able to more accurately recover the SVs in the child reconstruction when we allow for novel variants. We also note that while our performance is reduced

for reconstructing the parent genome as compared to our previous method, our new method still outperforms sparsity constraints alone as can be seen in Figure 3.3. We also observed that for the child reconstruction, our proposed method is more stable under varying τ values.

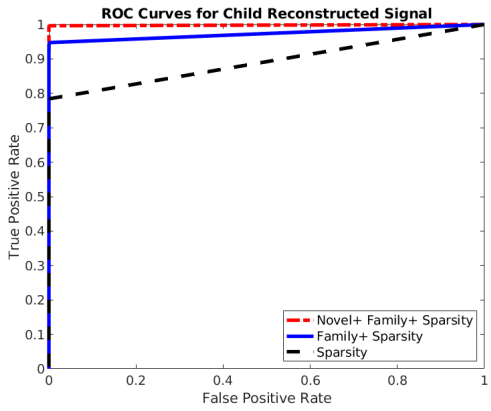
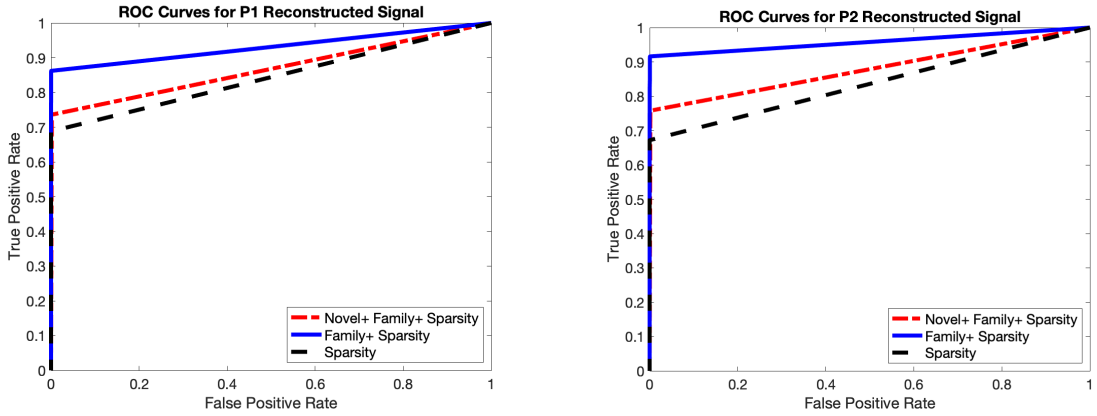


Figure 3.2: ROC curves of three methods illustrating the false positive rate vs. the true positive rate in the simulated child reconstruction, where $\lambda_{p_1} = \lambda_{p_2} = 8$, $\lambda_c = 10$, $\epsilon = .01$, $\tau = 100$ and $\gamma = 500$.



(a) (b)

Figure 3.3: ROC curves of three methods illustrating the false positive rate vs. the true positive rate in the simulated parent reconstructions, where $\lambda_{p_1} = \lambda_{p_2} = 8$, $\lambda_c = 10$, $\epsilon = .01$, $\tau = 100$ and $\gamma = 500$.

3.6.2 1000 Genomes Project Trio Data

To validate our method, we consider trio data of two separate populations from the 1000 Genomes Project [3]. Both the European (CEU) and Yoruba (YRI) father-mother-daughter trio genomes were sequenced at $\approx 4\times$ coverage and aligned to

NCBI36. We obtain our candidate set of SVs from the GASV pipeline, but our method is also applicable to other SV callers [34]. We filter the validated set of variants by eliminating experimentally validated SVs shorter than 250bp and which are classified as low quality. We note that only 15 of the validated variants in the child signal not present in either of the parents constitute our novel child signal $\vec{f}_n^*$.

Analysis. For child (inherited and novel) signal reconstructions, we achieve competitive sensitivity with our previous methods. When reconstructing the parent signals, we improve on our previous 2 parent - 1 child model, whose iterates are updated by non-alternating closed-form projections [7]. Figure 3.4 illustrates this improvement with predicted novel deletions against experimentally validated variants in the CEU mother NA12891 when comparing against previous models. We also find that our new method is stable under changes of τ and γ values.

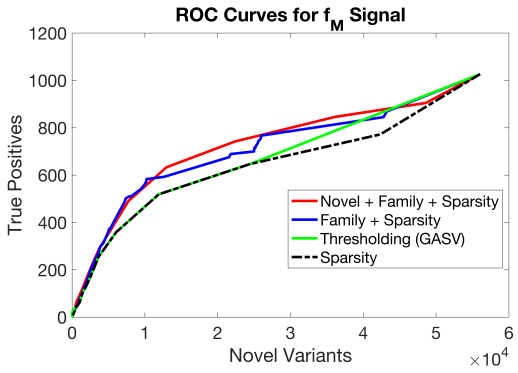


Figure 3.4: ROC curves of four methods illustrating novel variants vs. true positives (experimentally validated) in the signal of the CEU mother NA12891, where $\tau = 10$ and $\gamma = \frac{1}{10}$. We observe an overall improvement in correctly classifying SVs compared to previous methods.

3.7 Conclusions

We propose a new method to detect novel structural variants – SVs present in a child not inherited from a parent – from sequencing data in parent-child trios. Our method incorporates both relatedness and sparsity constraints, allowing for varying penalty parameters in the reconstruction of the child signal. By doing so, our new model is less sensitive to our regularization parameters. With real data, our method achieves competitive true positive predictions in the child and improves parent signal recovery, and we intend on exploring this with further simulated data studies. We present our results for both simulated and real data from the 1000 Genomes Project and suggest further exploration in varying sequencing coverage for future parent-offspring data. In future studies, we intend to incorporate other SV-calling tools, larger family structures, and a general relatedness parameter in our methods.

Chapter 4

Diploid Genomes from Parent-Child Trios

We consider a framework for refining structural variant (SV) recovery signals for multiple related individuals. This work considers diploid data from one father (p_1), one mother (p_2), and one child (c). We assume that each signal consists of m locations in the genome where an SV may occur. Humans have two copies of each chromosome, one inherited from each parent. If both parents have an SV at the same location, this impacts the probability that the child also has an SV at the same location. For each individual I in our model, we consider two signals that take on binary values: a heterozygous indicator $\vec{y}_I \in \{0,1\}^m$ and a homozygous indicator $\vec{z}_I \in \{0,1\}^m$. The heterozygous vector is an indicator that the individual has one copy of the SV while the homozygous vector indicates that the individual has two copies of the SV. If an individual is heterozygous for an SV at position j, then $(\vec{y}_I)_j = 1$ and $(\vec{z}_I)_j = 0$. Similarly, if an individual is homozygous for an SV at position j, then $(\vec{z}_I)_j = 1$ and $(\vec{y}_I)_j = 0$.

This approach is based on the paper [40] coauthored with Professors Mario Banuelos, Roummel Marcia, and Suzanne Sindi under review for the 2020 European Signal Processing Conference (EUSIPCO) proceedings.

4.1 Observation Model

The observed data are the number of DNA fragments supporting each potential SV. In particular, we denote the observation vectors for the parents (father and mother) and child by the vectors $\vec{s}_{p_1} \in \mathbb{R}^m, \vec{s}_{p_2} \in \mathbb{R}^m$, and $\vec{s}_c \in \mathbb{R}^m$, respectively. We assume the data follow a Poisson distribution ([21, 36]):

$$\begin{bmatrix} \vec{s}_c \\ \vec{s}_{p_1} \\ \vec{s}_{p_2} \end{bmatrix} \sim \text{Poisson}\left(\begin{bmatrix} z_c(2\lambda_c - \epsilon) + y_c(\lambda_c - \epsilon) + \epsilon \\ z_{p_1}(2\lambda_{p_1} - \epsilon) + y_{p_1}(\lambda_{p_1} - \epsilon) + \epsilon \\ z_{p_2}(2\lambda_{p_2} - \epsilon) + y_{p_2}(\lambda_{p_2} - s\epsilon) + \epsilon \end{bmatrix} \right), \tag{4.1}$$

where λ_c, λ_{p_1}, and λ_{p_2} are the sequencing coverage of the child, father, and mother, respectively, and $\epsilon > 0$ (see [8]). The parameter ϵ is reflective of measurement errors corresponding to the sequencing and mapping process. These errors are a large hindrance to accurate SV discovery methods and lead to a high false-positive discovery rate.

Letting

$$\vec{s} = \begin{bmatrix} \vec{s}_c \\ \vec{s}_{p_1} \\ \vec{s}_{p_2} \end{bmatrix}, \ \vec{z} = \begin{bmatrix} \vec{z}_c \\ \vec{z}_{p_1} \\ \vec{z}_{p_2} \end{bmatrix}, \ \vec{y} = \begin{bmatrix} \vec{y}_c \\ \vec{y}_{p_1} \\ \vec{y}_{p_2} \end{bmatrix}, \text{ and } \vec{f} = \begin{bmatrix} \vec{z} \\ \vec{y} \end{bmatrix},$$

we note that $\vec{f} \in \{0,1\}^{6m}$. Our general observation model (4.1) can be expressed as

$$\vec{s} \sim \text{Poisson}\left(Af + \epsilon \mathbf{1}\right),$$

where $\mathbf{1} \in \mathbb{R}^{3m}$ is the vector of ones and $A = [A_1 \ \ A_2] \in \mathbb{R}^{3m \times 6m}$ is the coverage matrix with

$$A_1 = \begin{bmatrix} (2\lambda_c - \epsilon)I_m & 0 & 0 \\ 0 & (2\lambda_{p_1} - \epsilon)I_m & 0 \\ 0 & 0 & (2\lambda_{p_2} - \epsilon)I_m \end{bmatrix}$$

and

$$A_2 = \begin{bmatrix} (\lambda_c - \epsilon)I_m & 0 & 0 \\ 0 & (\lambda_{p_1} - \epsilon)I_m & 0 \\ 0 & 0 & (\lambda_{p_2} - \epsilon)I_m \end{bmatrix}.$$

Here, $I_m \in \mathbb{R}^{m \times m}$ is the $m \times m$ identity matrix.

4.2 Problem Formulation

Assuming a Poisson process to model the noise in the measurements [24], the probability of observing the observation vector $\vec{s}$, given the true signal $\vec{f}$, is given by

$$p(\vec{s}|A\vec{f}) = \prod_{j=1}^{3m} \frac{((A\vec{f})_j + \epsilon)^{\vec{s}_j}}{\vec{s}_j!} \exp\left(-(A\vec{f})_j + \epsilon\right). \tag{4.2}$$

We use the maximum likelihood principle to determine the unknown Poisson parameter $A\vec{f}$ such that the probability of observing the vector of Poisson data $\vec{s}$ in (4.2) is maximized. Specifically, we minimize the corresponding negative Poisson log-likelihood function

$$F(\vec{f}) = \sum_{j=1}^{3m} (A\vec{f})_j - \vec{s}_j \log((A\vec{f})_j + \epsilon). \tag{4.3}$$

To minimize $F(\vec{f})$, we apply a continuous relaxation of the variables and use gradient-based methods. Specifically, we let that the values of $\vec{f}$ to lie between 0 and 1, i.e., $\mathbf{0} \leq \vec{f} \leq \mathbf{1}$, or equivalently,

$$\mathbf{0} \leq \vec{z}_I, \vec{y}_I \leq \mathbf{1}, \tag{4.4}$$

where $\mathbf{0}$ is the vector of zeros, $I \in \{c, p_1, p_2\}$, and the inequalities are to be understood component-wise. We note that since a variant cannot be both heterozygous and homozygous simultaneously, we require further that

$$\mathbf{0} \leq \vec{z}_I + \vec{y}_I \leq \mathbf{1}. \tag{4.5}$$

4.3 Familial Constraints

We incorporate additional constraints that exploit information about the signal $\vec{f}$ to help improve the accuracy of our SV predictions. The constraints control for biological realities in each individual as well as constraints from the relatedness of individuals.

First, if one of the parents is homozygous for an SV at location j, i.e., $(\vec{z}_{p_1})_j = 1$ or $(\vec{z}_{p_2})_j = 1$, then the child must be at least heterozygous, i.e., $(\vec{z}_c)_j + (\vec{y}_c)_j = 1$. This means that

$$\begin{array}{rcccl} \mathbf{0} & \leq & \vec{z}_{p_1} & \leq & \vec{z}_c + \vec{y}_c \\ \mathbf{0} & \leq & \vec{z}_{p_2} & \leq & \vec{z}_c + \vec{y}_c. \end{array}$$

These constraints indicate that if the child does not have an SV in a particular location, then neither parent can have a homozygous SV at that location.

Second, the child can only be homozygous, i.e., $(\vec{z}_c)_j = 1$, if both of the parents are at least heterozygous, i.e., $(\vec{z}_{p_1})_j + (\vec{y}_{p_1})_j = 1$ and $(\vec{z}_{p_2})_j + (\vec{y}_{p_2})_j = 1$. Furthermore, the child must be homozygous if both parents are homozygous, i.e.,

$$\max\{\vec{z}_{p_1} + \vec{z}_{p_2} - \mathbf{1}, \mathbf{0}\} \leq \vec{z}_c \leq \min\{\vec{z}_{p_1} + \vec{y}_{p_1}, \vec{z}_{p_2} + \vec{y}_{p_2}\},$$

where $\max\{\cdot,\cdot\}$ and $\min\{\cdot,\cdot\}$ are to be understood componentwise.

Finally, the child can only be heterozygous if at lest one of the parents is at least heterozygous, and the child cannot have an SV if neither parent has an SV, i.e.,

$$\mathbf{0} \leq \vec{y}_c \leq \min\{\vec{z}_{p_1} + \vec{y}_{p_1} + \vec{z}_{p_2} + \vec{y}_{p_2}, \mathbf{1}\}.$$

We denote the set of all vectors satisfying these constraints by $\mathcal{S}$, i.e.,

$$\mathcal{S} = \left\{ \begin{bmatrix} \vec{z}_c \\ \vec{z}_{p_1} \\ \vec{z}_{p_2} \\ \vec{y}_c \\ \vec{y}_{p_1} \\ \vec{y}_{p_2} \end{bmatrix} \in \mathbb{R}^{6m} : \begin{array}{r} \mathbf{0} \leq \vec{z}_I + \vec{y}_I \leq \mathbf{1},\ \mathbf{0} \leq \vec{z}_{p_1} \leq \vec{z}_c + \vec{y}_c, \\ \mathbf{0} \leq \vec{z}_{p_2} \leq \vec{z}_c + \vec{y}_c, \\ \max\{\vec{z}_{p_1} + \vec{z}_{p_2} - \mathbf{1}, \mathbf{0}\} \leq \vec{z}_c, \\ \vec{z}_c \leq \min\{\vec{z}_{p_1} + \vec{y}_{p_1}, \vec{z}_{p_2} + \vec{y}_{p_2}\}, \\ \mathbf{0} \leq \vec{y}_c \leq \min\{\vec{z}_{p_1} + \vec{y}_{p_1} + \vec{z}_{p_2} + \vec{y}_{p_2}, \mathbf{1}\} \end{array} \right\}.$$

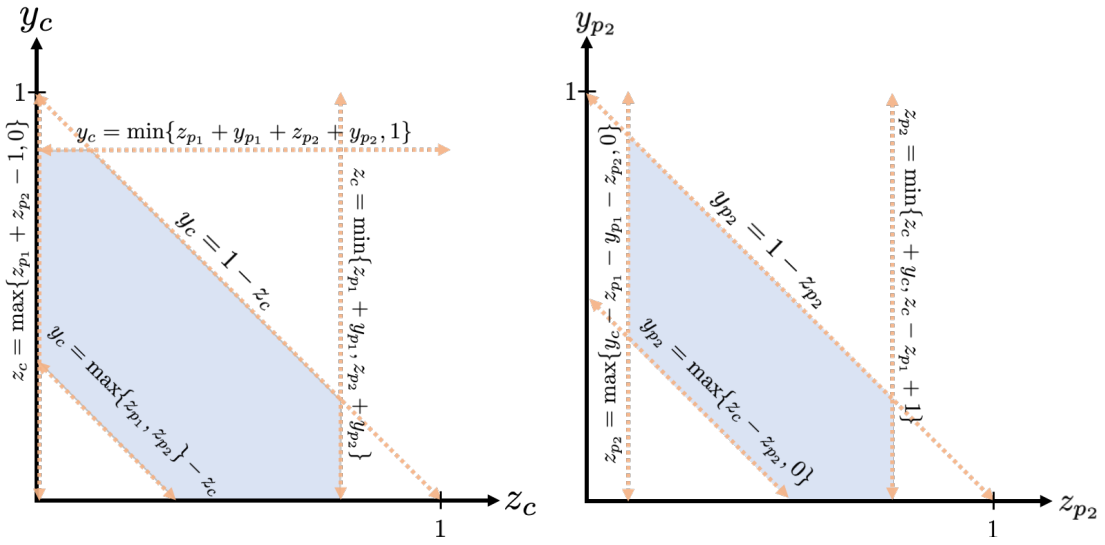


Figure 4.1: The feasible set (shown above by the shaded region) for each step of the proposed block-coordinate minimization approach. (a) In Step 1, we obtain the solution for the child's variables z_c and y_c given fixed parent indicator variables $z_{p_1}, y_{p_1}, z_{p_2}$ and y_{p_2}. (b) In Step 3, we obtain the solution for the mother's variables z_{p_2} and y_{p_2} given fixed indicator variables z_c, y_c, z_{p_1} and y_{p_1}. The feasible set represented in Step 2 is similar to that in Step 3.

4.4 Optimization Setup

A common difficulty with SV recovery is predicting false positive SVs by mistaking fragments that are incorrectly mapped against the reference genome. Since SVs are rare in an individual's genome, we enforce sparsity in our predictions by incorporating an ℓ_1-norm penalty term in our objective function (see [42]). Our objective function takes the following form:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{6m}}{\text{minimize}} \quad & F(\vec{f}) + \tau ||\vec{f}||_1 \\ \text{subject to} \quad & \vec{f} \in \mathcal{S} \end{aligned} \tag{4.6}$$

where $F(\vec{f})$ is the negative Poisson log-likelihood function shown in (4.3) and $\tau > 0$ is a regularization parameter. We then use a second-order Taylor series approximation around the current iterate $\vec{f}^k$ to formulate a sequence of quadratic subproblems. In this approach, we approximate the Hessian matrix by a scalar multiple of the identity matrix, $\alpha_k I$, where $\alpha_k > 0$ (see [10] for details) for how to compute α_k), and define the function

$$F^k(\vec{f}) = F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^T \nabla F(\vec{f}^k) + \frac{\alpha_k}{2} ||\vec{f} - \vec{f}^k||_2^2, \tag{4.7}$$

which we use as a surrogate function for $F(\vec{f})$ in (2.4). This approximation leads to the following equivalent subproblem formulation:

$$\begin{aligned} \vec{f}^{k+1} = \underset{\vec{f}\in\mathbb{R}^{6m}}{\arg\min} \quad & \frac{1}{2}||\vec{f} - \vec{r}^{k}||_2^2 + \gamma||\vec{f}||_1 \\ \text{subject to} \quad & \vec{f} \in \mathcal{S} \end{aligned} \tag{4.8}$$

where $\vec{r}^{k} = \vec{f}^{k} - \frac{1}{\alpha_k}\nabla F(\vec{f}^{k})$ and $\gamma = \frac{\tau}{\alpha_k}$ (see [15, 17] for details). Note that the objective function in (4.8) is separable in f. Thus, (4.8) can be solved in batches. In particular, at each candidate SV position, we solve

$$\begin{aligned} f^{k+1} = \underset{f\in\mathbb{R}^{6}}{\arg\min} \quad & \frac{1}{2}||f - r^{k}||_2^2 + \gamma||f||_1 \\ \text{subject to} \quad & f \in S \end{aligned} \tag{4.9}$$

where the vectors $r^k = [r^k_{z_c}; r^k_{z_{p_1}}; r^k_{z_{p_2}}; r^k_{y_c}; r^k_{y_{p_1}}; r^k_{y_{p_2}}]$ and $f = [z_c; z_{p_1}; z_{p_2}; y_c; y_{p_1}; y_{p_2}]$ correspond to the components of $\vec{r}^{k}$ and $\vec{f}$, respectively, and the set S is similar to the feasible set $\mathcal{S}$ but restricted to the particular candidate SV position.

4.5 Optimization Approach

Here we propose solving our problem using a block-coordinate descent approach. Following methods used in previous work (see [5]), we fix all but one individual and solve (4.9) over both indicator variables for that individual. In subsequent steps, the variables corresponding to some other individual are minimized while the other individuals signals are fixed. This block-coordinate descent approach continues until the iterates satisfy a pre-determined convergence criteria.

Step 0: First, we compute the unconstrained minimizer of (4.9), which is given by

$$\hat{f}^{(0)} = r^k - \gamma\mathbf{1}.$$

Then we initialize the parent indicator variables by

$$\hat{z}_I^{(0)} =| \{0, r^k_{z_I} - \gamma, 1\} \text{ and } \hat{y}_I^{(0)} =| \{0, r^k_{y_I} - \gamma, 1\},$$

where $I \in \{p_1, p_2\}$ and $\text{mid}\{\cdot,\cdot,\cdot\}$ takes on the value that is in the middle to ensure that the constraint in (4.4) is satisfied. To ensure that the constraint in (4.5) is satisfied, if

$$\hat{z}^{(0)}_{p_1} + \hat{y}^{(0)}_{p_1} > 1,$$

then we let $\hat{z}^{(0)}_{p_1} = \hat{y}^{(0)}_{p_1} = 0.5$. We adjust $\hat{z}^{(0)}_{p_2}$ and $\hat{y}^{(0)}_{p_2}$ similarly. To initialize the child indicator variables we let

$$\hat{z}^{(0)}_c = r^k_{z_c} - \gamma \text{ and } \hat{y}^{(0)}_c = r^k_{y_c} - \gamma.$$

We initialize the index with $i = 1$.

Step 1: Once we have obtained estimates for both parents' diploid indicator variables, $\hat{z}_{p_1}^{(i-1)}, \hat{y}_{p_1}^{(i-1)}, \hat{z}_{p_2}^{(i-1)}$ and $\hat{y}_{p_2}^{(i-1)}$, from the previous iteration, we project $\hat{z}_c^{(i-1)}$ and $\hat{y}_c^{(i-1)}$ onto the feasible set S with fixed parent variables to obtain the new child indicator variables $\hat{z}_c^{(i)}$ and $\hat{y}_c^{(i)}$. This projection is similar to the projections done in [5]. The feasible region for this step is illustrated in Fig. 4.1(a).

Step 2: After obtaining the new estimates for the child's diploid indicator variables, $\hat{z}_c^{(i)}$ and $\hat{y}_c^{(i)}$, from Step 1, we project $\hat{z}_{p_1}^{(i-1)}$ and $\hat{y}_{p_1}^{(i-1)}$ onto our feasible set S with fixed child and mother indicator variables to obtain the new father indicator variables $\hat{z}_{p_1}^{(i)}$ and $\hat{y}_{p_1}^{(i)}$. This projection is also similar to the projections done in [5]. The feasible region for this step is similar to that illustrated in Fig. 4.1(b).

Step 3: After obtaining the new estimates for the father's diploid indicator variables, $\hat{z}_{p_1}^{(i)}$ and $\hat{y}_{p_1}^{(i)}$, from Step 2, we project $\hat{z}_{p_2}^{(i-1)}$ and $\hat{y}_{p_2}^{(i-1)}$ onto our feasible set S with fixed child and father indicator variables to obtain the new mother indicator variables $\hat{z}_{p_2}^{(i)}$ and $\hat{y}_{p_2}^{(i)}$. This projection is also similar to the projections done in [5]. The feasible region for this step is illustrated in Fig. 4.1(b).

Steps 1, 2 and 3 are repeated in an alternating cycle until some convergence criteria are satisfied. In our numerical experiments, we saw that iterates did not change after three cycles. Thus, we terminated each cycle after three iterations. Note that Steps 2 and 3 are equivalent and result in identical feasible regions.

4.6 Results

4.6.1 Simulated Data

Before applying our method to real human data, we first tested the performance on simulated data to match our assumptions. To do this we simulated two parent signals with a set number of structural variants and a set similarity between the parent signals. The simulated true signals all consisted of 10^5 potential SVs. In the parent signals 500 locations were chosen at random to be variants; the percentage of variant sites the parents had in common was varied for testing. We then formed the child signal using a logical implementation of inheritance. If both parents were homozygous for an SV at position j then the child is homozygous for an SV at position j. If one parent was homozygous for an SV at position j and the other parent was heterozygous for an SV at position j then the child was at least heterozygous for an SV at that position, and had a 50% chance of being homozygous for an SV at position j. After forming the true signals for each individual, the observed signals were created by sampling from the Poisson distribution with a given coverage and error.

Analysis. Given an optimal τ value, our method is better able to reconstruct the homozygous signals for each individual. In Figure 4.2 we show an ROC curve gener-

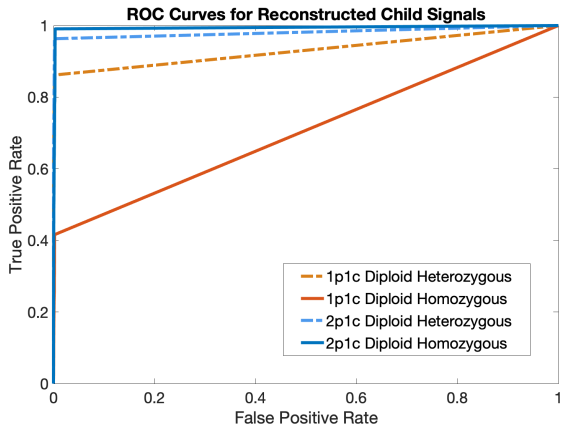


Figure 4.2: ROC curves of two methods illustrating the false positive rate vs. the false true positive rate in the child reconstruction broken into the heterozygous signal and the homozygous signal, where $\tau = 150$, the parents share 90% of their SVs and 30% of each parents SVs are homozygous.The coverage values for each individual are as follows $(\lambda_c, \lambda_{p_1}, \lambda_{p_2}) = (5, 10, 10)$.

ated for a simulated data set where the parents share 90% of their SVs and 30% of their SVs are homozygous. The area under the curve for each signal recovered from our method is greater than that of our previous diploid model which only includes information from one parent and one child [5]. We found that given optimal τ we were able to better recover not only the child signal, but also each of the parents as compared to our previous method.

4.6.2 1000 Genomes Project Trio Data

We next apply our diploid method to the 1000 Genomes Project CEU trio data [1]. The father-mother-daughter (NA12891-12892-12878) trio was sequence at approximately 4 $\times$ coverage and structural variants were experimentally validaded for these individuals. To create $\vec{z}$ and $\vec{y}$, we filter *LowQual* predictions and incorporated the genotype to separate heterozygous from homozygous reported deletions. Moreover, we only consider deletions longer than 250bp in the experimentally validated set.

Analysis. For each CEU genome, there are $n = 57,078$ candidate deletion locations. Of these GASV predictions, 686, 637, and 724 are validated deletions (heterozygous and homozygous combined) in the father, mother, and child, respectively. Whereas our previous method fixes one individual at a time, our new method simultaneously predicts all three individuals while improving the heterozygous signal reconstruction for the child (see Fig. 4.3). Moreover, we see comparable performance for the reconstruction of both heterozygous and homozygous signals for both parents. Fig. 4.4 is representative of the slightly improved predictions for the parent signals for varying values of τ.

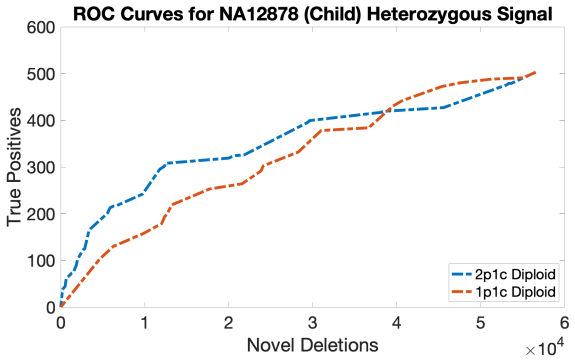


Figure 4.3: ROC curves for the reconstruction of the heterozygous child signal, $\vec{y}_c$, where $\lambda_c = \lambda_{p_1} = \lambda_{p_2} = 4$, $\tau = 1 \times 10^{-4}$, and $\epsilon = 0.01$. Since the validated set may not contain all true deletions, we plot novel deletions against validated true positives. We observe a considerable improvement in the detection of true positives with our proposed method.

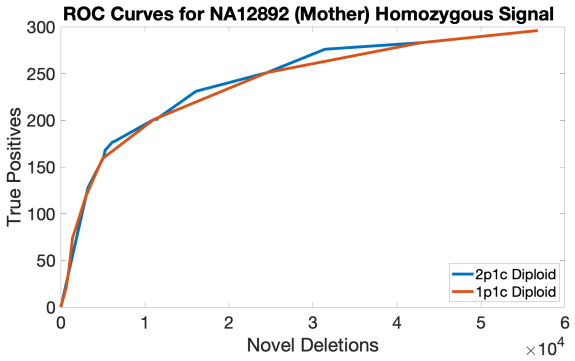


Figure 4.4: ROC curves for the reconstruction of the homozygous mother signal, $\vec{z}_{p_2}$, where the coverage is approximately 4 × for all individuals, $\tau = 1 \times 10^{-4}$, and $\epsilon = 0.01$. We note a marginal improvement over our previous method in this reconstruction.

4.7 Conclusions

We present an optimization method to detect SVs in sequencing data from parent-child trios. This method leverages relatedness between the individuals to improve signal reconstruction of noisy data. This extends previous work that focused on diploid signals from one parent and a child. We present results for both simulated and real data from the 1000 Genomes Project. We demonstrate that we are able to capture variants for which the individual possessed two copies. In future studies we intend to apply this work to a multi-generational framework with multiple offspring.

Chapter 5

Conclusions

The overall goal of this work was to use information about biological relatedness to improve the ability to predict structural variants. The two main contributions of this thesis are the following:

1. In Chapters 2 and 3, we formulated a framework for detecting novel variants in a child genome ,i.e., variants that are not present in the parent but are in the child.

2. In Chapter 4, we developed a framework to work with diploid genomic data.

For these frameworks we employed an approach that is traditionally used in image reconstruction. We modified these methods to handle constraints that ensure our results are biologically feasible. In the first framework we began with a one parent, one child model which results in a three-dimensional optimization formulation for every potential SV. Solving this problem required us to calculate orthogonal projections to return to our feasible region. We employed a different approach in our two parent, one child model. In this case our constraints led to a four-dimensional problem which required an alternating projection process to minimize each subproblem. Finally, in Chapter 4 we focused on developing a two parent, one child diploid model. This model resulted in a six-dimensional problem which we solved similarly to the previous model through an alternating projection process. Each of these Chapters has generalized a previous problem which increases the dimension of our subproblems but increases the accuracy of our SV predictions. We have shown the benefit of our methods on both simulated and real data from the 1000 Genomes project.

The optimization approach we developed for SV detection is composed of many subproblems. These subproblems are low dimensional, however with the addition of these biologically relevant constraints solving the subproblems becomes increasingly challenging. In particular we had to introduce an alternating block coordinate descent approach (in Chapter 3 and 4) for the frameworks where the feasibility sets were more than 3 dimensions. We have built these methods one at a time because each additional feature represents a dimension added to the problem and the projections. This means that as we add biologically relevant constraints or we add individuals we

increase the number of dimensions and constraints, hereby further complicating the process to ensure we have a feasible solution. A general method is needed to solve a quadratic subproblem subject to n-dimensional constraints for low values of n in order to effectively include all of the proposed features in one model. We believe that the block-coordinate descent approach that we use may be a valuable tool when developing a general method.

References

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

[2] C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. Nature Reviews Genetics, 12(5):363, 2011.

[3] D. M. Altshuler, E. S. Lander, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, et al. A map of human genome variation from population scale sequencing. Nature, 467(7319):1061–1073, 2010.

[4] M. Banuelos, L. Adhikari, R. Almanza, A. Fujikawa, J. Sahagún, K. Sanderson, M. Spence, S. Sindi, and R. F. Marcia. Nonconvex regularization for sparse genomic variant signal detection. In 2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pages 281–286, 2017.

[5] M. Banuelos, L. Adhikari, R. Almanza, A. Fujikawa, J. Sahagún, K. Sanderson, M. Spence, S. Sindi, and R. F. Marcia. Sparse diploid spatial biosignal recovery for genomic variation detection. In 2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pages 275–280, 2017.

[6] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, and S. Sindi. Constrained variant detection with sparc: Sparsity, parental relatedness, and coverage. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 3490–3493, 2016.

[7] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, and S. Sindi. Sparse genomic structural variant detection: Exploiting parent-child relatedness for signal recovery. In 2016 IEEE Statistical Signal Processing Workshop (SSP), pages 1–5, 2016.

[8] M. Banuelos, R. Almanza, L. Adhikari, S. Sindi, and R. F. Marcia. Sparse signal recovery methods for variant detection in next-generation sequencing data. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 864 – 868, 2016.

[9] M. Banuelos, R. Almanza, L. Adhikari, S. Sindi, and R. F. Marcia. Biomedical signal recovery: Genomic variant detection in family lineages. In 2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG), pages 1–4, 2017.

[10] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. IMA J. Numer. Anal., 8(1):141–148, 1988.

[11] E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. SIAM Journal on Optimization, 10(4):1196–1211, 2000.

[12] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 59(8):1207–1223, 2006.

[13] D. L. Donoho. Compressed sensing. IEEE Transactions on Information Theory, 52(4):1289–1306, 2006.

[14] M. A. Eberle, E. Fritzilas, P. Krusche, M. Källberg, B. L. Moore, M. A. Bekritsky, Z. Iqbal, H. Chuang, S. J. Humphray, and A. L. Halpern. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome research, 27(1):157–164, 2017.

[15] Z. T. Harmany, R. F. Marcia, and R. M. Willett. Sparse Poisson intensity reconstruction algorithms. In Proceedings of IEEE Statistical Signal Processing Workshop, Cardiff, Wales, UK, September 2009.

[16] 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, 21:1084 – 1096, 2011.

[17] 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, 21:1084 – 1096, 2011.

[18] A. A. Hoffmann and L. H. Rieseberg. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? Annual review of ecology, evolution, and systematics, 39:21–42, 2008.

[19] A. Keinan and A. G. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science, 336(6082):740–743, 2012.

[20] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding. Varscan: variant detection in

massively parallel sequencing of individual and pooled samples. Bioinformatics, 25(17):2283–2285, 2009.

[21] E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 2(3):231–239, 1988.

[22] B. Li, H.and Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. Bioinformatics, 25(16):2078–2079, 2009.

[23] 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):1–3, 2014.

[24] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer. The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic acids research, 42(D1):D986–D992, 2013.

[25] I. Martincorena and P. J. Campbell. Somatic mutation in cancer and normal cells. Science, 349(6255):1483–1489, 2015.

[26] P. Medvedev, M. Stanciu, and M. Brudno. Computational methods for discovering structural variation with next-generation sequencing. Nature methods, 6:S13–S20, 2009.

[27] A. Meindl, H. Hellebrand, C. Wiek, V. Erven, B. Wappenschmidt, D. Niederacher, M. Freund, P. Lichtner, L. Hartmann, H. Schaal, et al. Germline mutations in breast and ovarian cancer pedigrees establish rad51c as a human cancer susceptibility gene. Nature Genetics, 42(5):410, 2010.

[28] Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, W. Ding, et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. Science, 266(5182):66–71, 1994.

[29] B. Milholland, X. Dong, L. Zhang, X. Hao, Y. Suh, and J. Vijg. Differences between germline and somatic mutation rates in humans and mice. Nature Communications, 8:15183, 2017.

[30] A. R. Quinlan, R. A. Clark, S. Sokolova, M. L. Leibowitz, Y. Zhang, M. E. Hurles, J. .C. Mell, and I. M. Hall. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome research, 20(5):623–635, 2010.

[31] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics, 28(18):i333–i339, 2012.

[32] Jarupon Fah Sathirapongsasuti, Hane Lee, Basil AJ Horst, Georg Brunner, Alistair J Cochran, Scott Binder, John Quackenbush, and Stanley F Nelson. Exome sequencing-based copy-number variation and loss of heterozygosity detection: Exomecnv. Bioinformatics, 27(19):2648–2654, 2011.

[33] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods, 15(6):461–468, 2018.

[34] S. Sindi, E. Helman, A. Bashir, and B. J. Raphael. A geometric approach for classification and comparison of structural variants. Bioinformatics, 25(12):i222–i230, 2009.

[35] S. S. Sindi and B. J. Raphael. Identification of structural variation. Genome Analysis: Current Procedures and Applications, page 1, 2014.

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

[37] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi. Detecting novel structural variants in genomes by leveraging parent-child relatedness. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 943–950, Dec 2018.

[38] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi. Predicting novel and inherited variants in parent-child trios. In 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2019.

[39] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi. Detecting inherited and novel structural variants in low-coverage parent-child sequencing data. Methods, 173:61–68, 2020.

[40] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi. Genomic signal processing for variant detection in diploid parent-child trios. Under review for the 2020 European Signal Processing Conference (EUSIPCO) proceedings, 2020.

[41] P. Stankiewicz and J. R. Lupski. Structural variation in the human genome and its role in disease. Annual review of medicine, 61:437–455, 2010.

[42] R. Tibshirani. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B, 58(1):267–288, 1996.

[43] Jeremiah A Wala, Pratiti Bandopadhayay, Noah F Greenwald, Ryan O'Rourke, Ted Sharpe, Chip Stewart, Steve Schumacher, Yilong Li, Joachim Weischenfeldt, Xiaotong Yao, et al. Svaba: genome-wide detection of structural variants and indels by local assembly. Genome research, 28(4):581–591, 2018.

[44] Xin Wang, Huan Zhang, and Xiaojing Liu. Defind: Detecting genomic deletions by integrating read depth, gc content, mapping quality and paired-end mapping signatures of next generation sequencing data. Current Bioinformatics, 14(2):130–138, 2019.

[45] J. Weischenfeldt, O. Symmons, F. Spitz, and J. O. Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. Nature Reviews Genetics, 14(2):125, 2013.

[46] Z. Zhang, L. Mao, H. Chen, F. Bu, G. Li, J. Sun, S. Li, H. Sun, C. Jiao, R. Blakely, et al. Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. The Plant Cell, pages tpc–114, 2015.