Detecting Inherited and Novel Structural Variants in Low-Coverage Parent-Child Sequencing Data

Melissa Spence,[†] Mario Banuelos,[⋆] Roummel F. Marcia,[†] and Suzanne Sindi[†][1]

[†]*Department of Applied Mathematics, University of California, Merced, Merced, CA 95343 USA*
[⋆]*Department of Mathematics, California State University, Fresno, Fresno, CA 93740 USA*

Abstract

Structural variants (SVs) are a class of genomic variation shared by members of the same species. Though relatively rare, they represent an increasingly important class of variation, as SVs have been associated with diseases and susceptibility to some types of cancer. Common approaches to SV detection require the sequencing and mapping of fragments from a test genome to a high-quality reference genome. Candidate SVs correspond to fragments with discordant mapped configurations. However, because errors in the sequencing and mapping will also create discordant arrangements, many of these predictions will be spurious. When sequencing coverage is low, distinguishing true SVs from errors is even more challenging. In recent work, we have developed SV detection methods that exploit genome information of closely related individuals – parents and children. Our previous approaches were based on the assumption that any SV present in a child's genome must have come from one of their parents. However, using this strict restriction may have resulted in failing to predict any rare but novel variants present only in the child. In this work, we generalize our previous approaches to allow the child to carry novel variants. We consider a constrained optimization approach where variants in the child are of two types either inherited - and therefore must be present in a parent - or novel. For simplicity, we consider only a single parent and single child each of which have a haploid genome. However, even in this restricted case, our approach has the power to improve variant prediction. We present results on both simulated candidate variant regions, parent-child trios from the 1000 Genomes Project, and a subset of the 17 Platinum Genomes.

Keywords: Sparse signal recovery, convex optimization, next-generation sequencing data, structural variants, computational genomics

[1] This work was supported by NSF Grant IIS 1741490. M. Spence and M. Banuelos contributed equally to this work and are both first authors of this manuscript. R. F. Marcia and S. Sindi are both the principal investigators of this research.

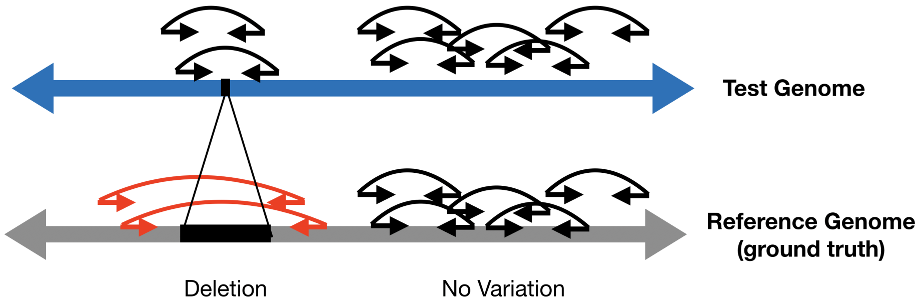


Figure 1: **Detecting Structural Variants.** To detect structural variants (SVs) in a test individual, fragments of DNA (black) are sampled from their (unknown) genome (top) and aligned to a reference genome (bottom). Fragments whose mappings are consistent with the underlying sampling process (right) 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.

1. Introduction

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) [1, 2]. 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 [3]. In humans, the accumulation of somatic mutations is commonly associated with the development of cancer [4] while the presence of certain germline variants has been shown to increase the susceptibility for certain types of cancer [5, 6]. Beyond cancer, genomic variants are associated with many significant biological outcomes for individuals including a variety of diseases in humans [7, 8], flowering behavior in plants [9] and have contributed to rates of adaptation and the emergence of new species [10].

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 [11, 12, 13]. The dominant method for identifying SNVs or SVs involves comparing fragments of DNA sequenced from a test (unknown) genome to a given reference

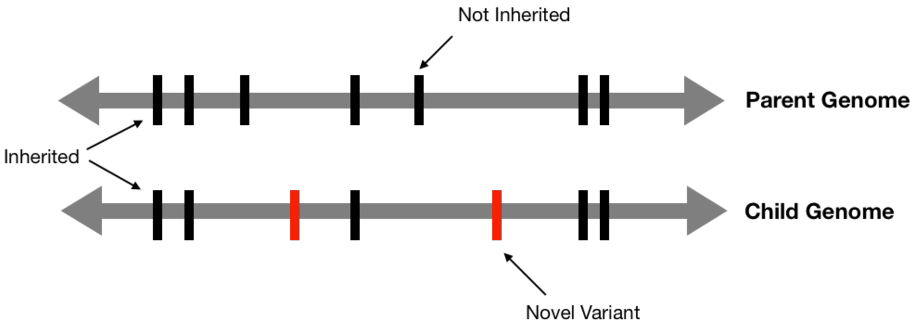


Figure 2: **Inheritance of Structural Variants.** Because germline structural variants (black and red bars) are transmitted from parents to their children, a child and parent will share many variants. However, because of recombination not all variants present in a parent's genome will be present in a child's genome and, although rare, a child may acquire novel variants not present in a parent (red).

(see Figure 1) [14, 15, 16, 17, 18]. 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 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 [14, 15]. One hope for improving the ability to accurately predict SVs has come from methods that combine the information of many individuals [19]. 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 [20, 21]. 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. But, because of the massive population expansion, many variants in humans are rare and, as such, may only be shared by close relatives [22]. 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 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 [23, 24, 25, 26]. 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.

This work improves upon our previous methods by allowing the child genome to possess novel variants. In Section 2, we develop our mathematical model and optimization framework for SV prediction in the context of one parent and one child. 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. We consider a continuous relaxation of this discrete problem, but favor sparse solutions through the use of the ℓ_1 norm. We also demonstrate that by a hierarchical approach it is possible to generalize our method to multiple generations. In Section 3, we show that, even with our simplified haploid genome assumption, our method improves SV detection on both simulated and real sequencing data for parent-child trios from the 1000 Genomes Project. Finally, we demonstrate that our hierarchical approach has the potential to improve SV prediction in extended pedigrees through analysis of a subset of the 17 Platinum Genomes.

2. Method

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 $\underline{\text{i}}$nherited from the parent and $\vec{f}_n^* \in \{0,1\}^m$ is the vector of SVs that are $\underline{\text{n}}$ovel. 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.

2.1. *Observational model*

The observed data are the number of DNA fragments supporting each potential SV, and the vectors $\vec{y}_p \in \mathbb{R}^m$ and $\vec{y}_c \in \mathbb{R}^m$ are the observation vectors of the parent and child, respectively. As in previous work [27, 28, 29], we assume that the data follow a Poisson distribution,

$$\begin{bmatrix}(\vec{y}_c)_j\\(\vec{y}_p)_j\end{bmatrix} \sim \text{Poisson}\left(\begin{bmatrix}(\lambda_c-\epsilon)\left\{(\vec{f}_i)_j+(\vec{f}_n)_j\right\}+\epsilon\\(\lambda_p-\epsilon)(\vec{f}_p)_j+\epsilon\end{bmatrix}\right) \tag{1}$$

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

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

Then the general observation model can be expressed as

$$\vec{y} \sim \text{Poisson}(A\vec{f}^* + \epsilon\mathbf{1}), \tag{2}$$

where $\mathbf{1} \in \mathbb{R}^{2m}$ 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 (2), the probability of observing $\vec{y}$ is given by

$$p(\vec{y}\,|A\vec{f}^*) = \prod_{j=1}^{2m} \frac{((A\vec{f}^*)_j + \epsilon)^{\vec{y}_j}}{\vec{y}_j!} \exp\Big(-(A\vec{f}^*)_j + \epsilon\Big). \tag{3}$$

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 $\vec{y}$ in (3) is maximized. Specifically, we minimize the corresponding *negative Poisson log-likelihood* function

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

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., $\mathbf{0} \le \vec{f} \le \mathbf{1}$.

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.,

$$\mathbf{0} \le \vec{f}_i + \vec{f}_n \le \mathbf{1}.$$

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

$$\mathbf{0} \le \vec{f}_i \le \vec{f}_p \le \mathbf{1}.$$

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

$$\mathbf{0} \le \vec{f}_n \le \mathbf{1} - \vec{f}_p.$$

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

$$\mathcal{S} = \left\{ \begin{bmatrix} \vec{f}_i \\ \vec{f}_n \\ \vec{f}_p \end{bmatrix} \in \mathbb{R}^{3m} : \begin{array}{l} \mathbf{0} \le \vec{f}_i + \vec{f}_n \le \mathbf{1},\ \mathbf{0} \le \vec{f}_i \le \vec{f}_p \le \mathbf{1}, \\ \mathbf{0} \le \vec{f}_n \le \mathbf{1} - \vec{f}_p,\ \mathbf{0} \le \vec{f}_i, \vec{f}_n, \vec{f}_p \le \mathbf{1} \end{array} \right\}.$$

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 [30, 31, 32]. 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.

2.5. Optimization problem

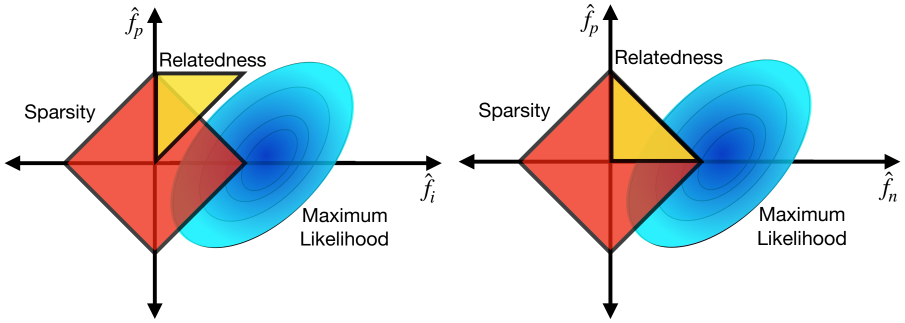
With these components defined, the genomic variants reconstruction problem has the following constrained optimization form:

$$\begin{array}{ll} \underset{\vec{f} \in \mathbb{R}^{3m}}{\text{minimize}} & F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \tag{4}$$

where $\tau > 0$ is a regularization parameter that balances the negative Poisson log-likelihood data fidelity term with the sparsity-promoting penalty term. Figure 3b 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 [33] to solve (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 $\vec{f}^k$:

$$F(\vec{f}) \;\approx\; F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^\top \nabla F(\vec{f}^k) + \tfrac{1}{2}(\vec{f} - \vec{f}^k)^\top \nabla^2 F(\vec{f}^k)(\vec{f} - \vec{f}^k). \tag{5}$$



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

Figure 3: 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., $0 \leq \hat{f}_p + \hat{f}_n \leq 1$.

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

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

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 $1 \leq j \leq m$. We approximate the second-derivative Hessian matrix with a scalar multiple of the identity matrix $\alpha_k I$ where $\alpha_k > 0$ (see [34, 35] for details) and define the quadratic model

$$F^k(\vec{f}) \equiv F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^T \nabla F(\vec{f}^k) + \frac{\alpha_k}{2} ||\vec{f} - \vec{f}^k||_2^2. \tag{6}$$

Now, each quadratic subproblem will be of the form

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

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

$$\begin{aligned} \vec{f}^{k+1} &= \underset{\vec{f}\in\mathbb{R}^{3m}}{\arg\min} \quad \mathcal{Q}(\vec{f}) = \tfrac{1}{2}\|\vec{f} - \vec{s}^{\,k}\|_2^2 + \tfrac{\tau}{\alpha_k}\mathrm{pen}(\vec{f}) \\ & \qquad \text{subject to} \;\; \vec{f} \in \mathcal{S}, \end{aligned} \tag{7}$$

where

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

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

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

Since the bounds that define the feasible set $\mathcal{S}$ are component-wise, then (7) separates into subproblems of the form

$$\begin{aligned} \underset{f_i, f_n, f_p \in \mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_i - s_i)^2 + \tfrac{1}{2}(f_n - s_n)^2 + \tfrac{1}{2}(f_p - s_p)^2 \\ & + \tfrac{\tau}{\alpha_k}|f_p| \; + \; \tfrac{\tau}{\alpha_k}|f_i| \; + \; \tfrac{\gamma\tau}{\alpha_k}|f_n| \\ \text{subject to} \quad & 0 \le f_i + f_n \le 1, \quad 0 \le f_i \le f_p \le 1, \\ & 0 \le f_n \le 1 - f_p, \quad 0 \le f_i, f_n, f_p \le 1, \end{aligned} \tag{8}$$

where $\{f_i, f_n, f_p\}$ and $\{s_i, s_n, s_p\}$ are scalar components of the vectors $\{\vec{f}_i, \vec{f}_n, \vec{f}_p\}$ and $\{\vec{s}_i, \vec{s}_n, \vec{s}_p\}$, respectively, at the same location. Completing the squares and ignoring constant terms, the optimization problem (8) can be expressed as

$$\begin{aligned} \underset{f_i, f_n, f_p \in \mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_i - a)^2 + \tfrac{1}{2}(f_n - b)^2 + \tfrac{1}{2}(f_p - c)^2 \\ \text{subject to} \quad & 0 \le f_i + f_n \le 1, \quad 0 \le f_i \le f_p \le 1, \\ & 0 \le f_n \le 1 - f_p, \quad 0 \le f_i, f_n, f_p \le 1, \end{aligned} \tag{9}$$

where $a = s_i - \frac{\tau}{\alpha_k}$, $b = s_n - \frac{\gamma\tau}{\alpha_k}$ and $c = s_p - \frac{\tau}{\alpha_k}$. The unconstrained minimizer of (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 (9) by orthogonally projecting (a, b, c) onto the three-dimensional feasible set, which is illustrated in Fig. 4. In particular, the f_i-f_n-f_p

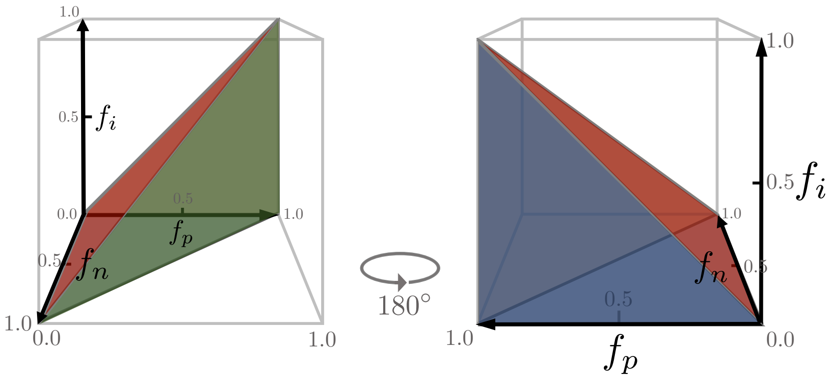


Figure 4: The three-dimensional feasible region of the minimization problem (9) on the f_i-f_n-f_p axis. Because novel child SVs are not present in the parent genome, i.e., $f_n \leq 1 - f_p$, $f_n \to 0$ as $f_p \to 1$. Similarly, because inherited SVs come from the parent genome, i.e., $f_i \leq f_p$, $f_i \to 0$ 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.

three-dimensional space partitions into **15** different regions that projects onto a vertex, edge, or surface of the feasible set for infeasible points. Tables 1 and 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.

	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 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 (9). Projections onto edges and surfaces are represented as linear combinations of a, b, and c in Table 2.

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

3. Results

3.1. Implementation Details

We implemented our method for variant detection in MATLAB by extending our previous approach [26] based on the SPIRAL method [33]. 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 [29]. 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). 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 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.

3.2. 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 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 $0.5 \leq \rho \leq 1$.

3.2.1. 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

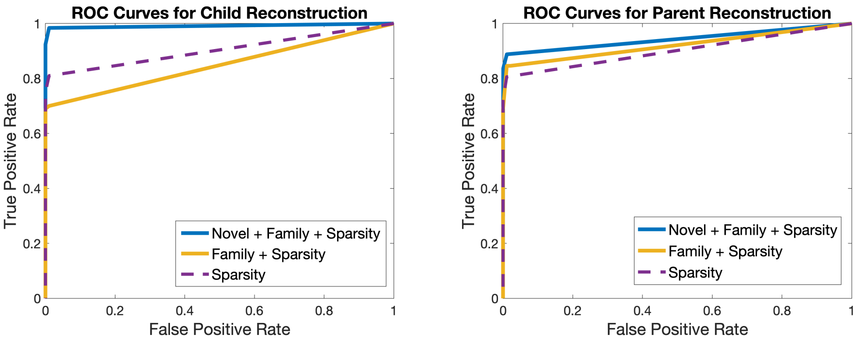


Figure 5: *Left.* ROC curves of three methods illustrating the false positive rate vs. the false true positive rate in the simulated child reconstruction, where $\tau = 20$ and $\gamma = \frac{3}{2}$. *Right.* ROC curves of three methods illustrating the false positive rate vs. the false true positive 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).

recover the SVs in the child reconstruction when we allow for novel variants. Figure 5 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. 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 (data not shown).

3.3. 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 [11]. In the pilot study of the project, both the CEU (European ancestry) and YRI (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 [36]. 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*

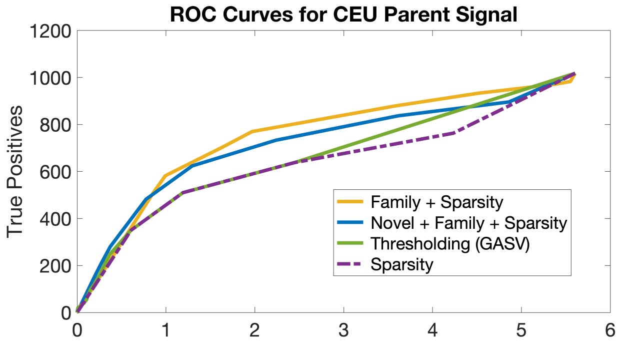


Figure 6: ROC curves of four methods illustrating the novel deletions (validated set of deletions may be incomplete) vs. true positives in the signal of the CEU parent NA12891, where $\tau = 0.0129$ and $\gamma = 10$. We observe an increase of true positive predictions when the number of novel predictions < 1000.

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.

3.3.1. 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 6 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 [29]. Next, we focus on the reconstruction of the entire child signal $\vec{f_c}$. Figure 7 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.

3.4. 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 [13]. All 17 family members' DNA was originally sequenced on an Illumina HiSeq2000 to an average depth of 50× using 2×100 bp reads and PCR-free sample preparation. Although originally

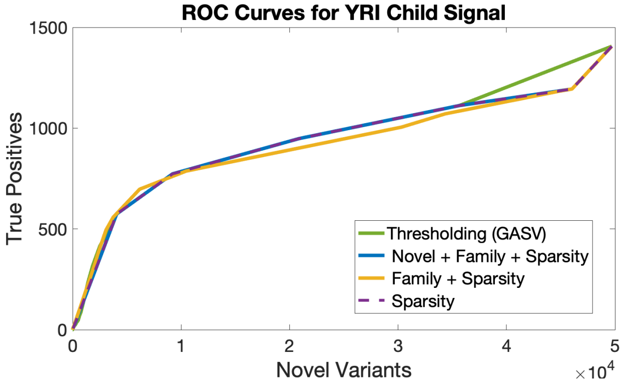


Figure 7: ROC curves of four methods illustrating the novel deletions vs. true positives in the combined child signal $\vec{f_c}$ of the child in the YRI trio (NA19240), where $\tau = 1$ and $\gamma = 10$. We note comparable performance of our proposed model with only enforcing sparsity.

sequenced at high coverage, we use Samtools to subsample and achieve low coverage of approximately 5× [37]. We determine true SVs with the intersection of GASV and Delly SV calls [36, 38]. 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.

3.4.1. 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 8 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 (≈ 10×), 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).

4. 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.

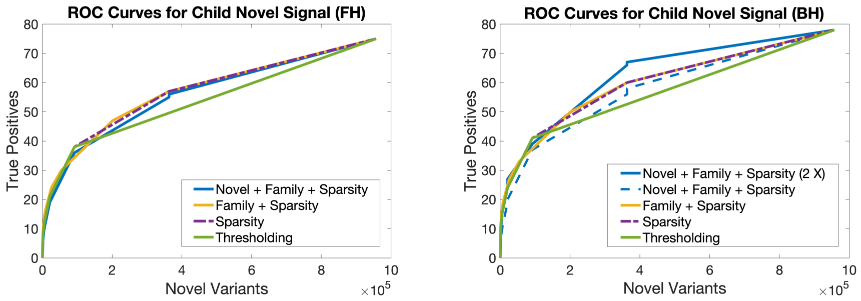


Figure 8: ROC curves of four methods illustrating the novel deletions vs. true positives in the signal of the CEU child NA12882, where $\tau = 0.0129$ and $\gamma = 10$. *Left.* Using the forward hierarchical (FH) approach, we observe comparable detection of novel variants. *Right.* With the backward hierarchical (BH) approach, we note an increase in true positive rate compared to applying our method only once (dashed blue line).

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.

References

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

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

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

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

[5] 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) (2010) 410.

[6] 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) (1994) 66–71.

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

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

[9] 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 (2015) tpc–114.

[10] A. A. Hoffmann, 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 (2008) 21–42.

[11] 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) (2010) 1061–1073.

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

[13] M. A. Eberle, E. Fritzilas, P. Krusche, M. Källberg, B. L. Moore, M. A. Bekritsky, Z. Iqbal, H. Chuang, S. J. Humphray, 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) (2017) 157–164.

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

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

[16] J. A. Wala, P. Bandopadhayay, N. F. Greenwald, R. O'Rourke, T. Sharpe, C. Stewart, S. Schumacher, Y. Li, J. Weischenfeldt, X. Yao, et al., Svaba: genome-wide detection of structural variants and indels by local assembly, Genome research 28 (4) (2018) 581–591.

[17] F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, M. C. Schatz, Accurate detection of complex structural variations using single-molecule sequencing, Nat Methods 15 (6) (2018) 461–468.

[18] X. Wang, H. Zhang, X. 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) (2019) 130–138.

[19] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, L. Ding, Varscan: variant detection in massively parallel sequencing of individual and pooled samples, Bioinformatics 25 (17) (2009) 2283–2285.

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

[21] 1000 Genomes Project Consortium and others, An integrated map of genetic variation from 1,092 human genomes, Nature 491 (7422) (2012) 56–65.

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

[23] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, 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), 2016, pp. 3490–3493. `doi:10.1109/EMBC.2016.7591480`.

[24] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, S. Sindi, Sparse genomic structural variant detection: Exploiting parent-child relatedness for signal recovery, in: 2016 IEEE Statistical Signal Processing Workshop (SSP), 2016, pp. 1–5. `doi:10.1109/SSP.2016.7551828`.

[25] M. Banuelos, L. Adhikari, R. Almanza, A. Fujikawa, J. Sahagún, K. Sanderson, M. Spence, S. Sindi, R. F. Marcia, Nonconvex regularization for sparse genomic variant signal detection, in: 2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2017, pp. 281–286. `doi:10.1109/MeMeA.2017.7985889`.

[26] M. Banuelos, R. Almanza, L. Adhikari, S. Sindi, R. F. Marcia, Biomedical signal recovery: Genomic variant detection in family lineages, in: 2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG), 2017, pp. 1–4. `doi:10.1109/ENBENG.2017.7889453`.

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

[28] J. F. Sathirapongsasuti, H. Lee, B. A. Horst, G. Brunner, A. J. Cochran, S. Binder, J. Quackenbush, S. F. Nelson, Exome sequencing-based copy-number variation and loss of heterozygosity detection: Exomecnv, Bioinformatics 27 (19) (2011) 2648–2654.

[29] M. Banuelos, R. Almanza, L. Adhikari, S. Sindi, 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), 2016, pp. 864 – 868. doi:10.1109/ICASSP.2016.7471798.

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

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

[32] E. Candes, J. Romberg, 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) (2006) 1207–1223.

[33] Z. T. Harmany, R. F. Marcia, R. M. Willett, This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms—theory and practice, IEEE Trans. on Image Processsing 21 (2011) 1084 – 1096.

[34] J. Barzilai, J. M. Borwein, Two-point step size gradient methods, IMA J. Numer. Anal. 8 (1) (1988) 141–148. doi:10.1093/imanum/8.1.141.

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

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

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

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