This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

Sparse Diploid Spatial Biosignal Recovery for Genomic Variation Detection

Mario Banuelos*, Lasith Adhikari*, Rubi Almanza*, Andrew Fujikawa†, Jonathan Sahagún‡,
Katharine Sanderson§, Melissa Spence¶, Suzanne Sindi*, and Roummel F. Marcia*
*Applied Mathematics, University of California, Merced, 5200 North Lake Road, Merced, CA, USA
†California State University, Sacramento, 6000 J Street, Sacramento, CA, USA
‡California State University, Los Angeles, 5151 State University Drive, Los Angeles, CA, USA
§Montana State University, 211 Montana Hall, Bozeman, MT, USA
¶University of California, Davis, 1 Shields Ave, Davis, CA, USA

***Abstract*—Structural variants (SVs) - such as duplications, deletions and inversions - are rearrangements of an individual's genome relative to a given reference. The common method for detection of SVs is to sequence fragments from an individual's genome, map them to the appropriate reference and, by identifying discordant mappings, predict the locations and type of SV. However, errors in both the sequencing and mapping process will result in signals that look like SVs, resulting in inaccurate predictions. In addition, because of variation in sequencing coverage even when the evidence of an SV is present, determining if an individual has the SV present on one or both of their chromosomes is challenging. In our work, we seek to improve upon standard methods for SV detection in three ways. First, to reduce false-positive predictions, we simultaneously predict SVs in a parent and child using properties of inheritance to constrain the space of possible SVs. Second, we predict if a variant is homozygous (SV is on two chromosomes) or heterozygous (SV is on one chromosome). Third, we utilize a gradient-based optimization approach and constrain our solution with a sparsity-promoting ℓ_1 penalty (since SV instances should be rare). We demonstrate the improved performance of our computational approach on both simulated genomes as well as a parent-child trio from the 1000 Genomes Project.**

I. Introduction

The genome, or complete DNA sequence, of an individual consists of an ordered sequence of nucleotides (A,C,G,T). The total length of this sequence varies from millions of letters (for a bacteria) to billions of letters (for a mammalian genome) [1]. Some species, such as humans, are diploid – meaning they have two copies of their genome and receive one copy from each parent. While many parts of the genome are identical amongst all members of the same species, there are others that are flexible and may vary significantly. Structural variants (SVs) – such as duplications, deletions and inversions – represent a rearrangement of an individual's DNA sequence as compared to a reference. Once thought to be primarily associated with genetic diseases like cancer [2], there are many types of SVs, like inversions, deletions and duplications that have been identified in the genomes of healthy individuals [3], [4]. As such, SVs represent an important part of understanding our recent population history [5], [6].

The common approach for SV prediction from sequencing data has been to map the resulting sequences to a reference genome and computationally identify statistically significant deviations from the expected mapping signals consistent with each class of SV [7]. However, errors in both the sequencing and mapping process itself may cause inconsistencies in the data that falsely suggest the presence of an SV. As such, many computational approaches for SV detection suffer from high false-positive rates [7], [8]. In addition, despite the fact that for each SV an individual carries either 1 copy (heterozygous) or 2 copies (homozygous), most computational approaches do not distinguish between these two cases (see Fig. 1). Finally, despite the fact that the rate of novel germline SV is negligible [5], and therefore any SV present in a child must have been inherited from one of their parents, most computational SV pipelines consider only one individual at a time [9], [10], [11].

In this work, we develop a computational framework for predicting the presence of SVs by simultaneously analyzing related individuals, specifically a parent and a child. We utilize a likelihood-based approach for predicting the most likely SVs present in each individual's genome and constrain the space of possible predictions by those that are consistent with inheritance. We further enforce sparsity in our predictions through an ℓ_1 penalty term. This work builds upon our recent studies [12], [13], [14], but offers several improvements. First, we explicitly predict the number of copies each individual (0, 1 or 2) an individual carries of each SV. Second, our method of solving successive subproblems resulting from the gradient-based optimization framework is easily generalized beyond two individuals and thus could be adapted to trios, or larger population pedigree data. Finally, we have improved the computational speed of our method which further facilitates its adaptation to larger populations. In our work below, we develop our SV detection method as a constrained optimization problem. We discuss how we solve this problem and demonstrate the effectiveness of our method on both simulated parent-child genomes and real parent-child genomes from the 1000 Genomes Project [3].

We note that hierarchial relationships have been used to infer the presence of somatic variants when comparing DNA taken from normal and tumor samples in the same individual [15], [16]. However, the framework is different for several reasons: (1) the goal is to detect cancer specific SVs rather than validate SV predictions in a normal genome, (2) cancer cells

978-1-5090-2984-6/17/$31.00 ©2017 IEEE

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on November 30,2021 at 19:43:13 UTC from IEEE Xplore. Restrictions apply.

often have chromosomal duplication events leading to more than diploid coverage, and (3) tumor samples are typically heterogeneous both with respect to clonal evolution of cancer and contamination by normal cells. Although sequencing costs have decreased, our method focuses on using low-coverage data to predict zygosity of structural variants, as this has shown to be a source of genetic diversity between populations [17].

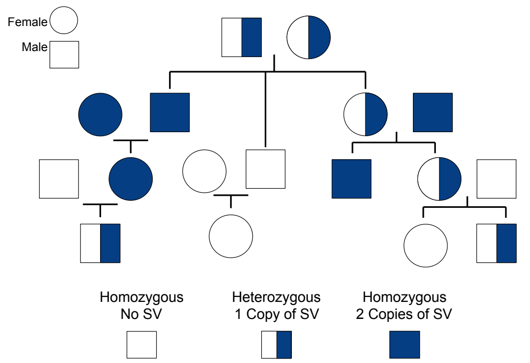

Fig. 1. **Familial Pedigree.** We illustrate a pedigree depicting the number of copies (0 = white, 1 = half shaded, 2 = full shaded) of a particular SV in each individual male (square) or female (circle). An individual is homozygous if both chromosomes have the same state. The number of parental copies determines the number of copies carried by the child. We also only depict the assumed constraints in this illustration (i.e., SVs can only be inherited from parents and new mutations are not considered).

II. Poisson Log-Likelihood Optimization

Here, we describe mathematically our computational framework for predicting the number of copies an individual carries of each SV.

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

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

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

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

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

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

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

$$\vec{s} \sim \text{Poisson}(A\vec{f}). \tag{1}$$

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

A. Continuous Relaxation and Constraints

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

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

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

B. Optimization Formulation

Under the model (1), the probability of observing $\vec{s}$ is

$$p(\vec{s} \,|A\vec{f}^*) = \prod_{i=1}^{4n} \frac{((A\vec{f}^*)_i)^{\vec{s}_i}}{\vec{s}_i!} \exp\left(-(A\vec{f}^*)_i \right). \tag{2}$$

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

$$\underset{\vec{f} \in \mathscr{F}}{\text{minimize}} \quad \phi(\vec{f}) \equiv F(\vec{f}) + \tau \|\vec{f}\|_1, \tag{3}$$

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

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

$\tau > 0$ is a regularization parameter, $\|\vec{f}\|_1 = \sum_{j=1}^{4n} |f_j|$ added to promote sparsity in the solution, and $0 < \epsilon \ll 1$ is a small parameter introduced to avoid the singularity at $\vec{f} = 0$. Using second-order Taylor series approximations around the current iterate $\vec{f}^k$ and approximating the Hessian matrix by a scalar multiple of the identity matrix, $\alpha_k I$ (see [18]), we solve a sequence of quadratic approximations to $F(\vec{f})$. Further, these quadratic subproblems can be simplified to the following form:

$$\vec{f}^{k+1} = \underset{\vec{f} \in \mathscr{F}}{\text{arg min}} \;\; \tfrac{1}{2} \|\vec{f} - \vec{r}^{\,k}\|_2^2 + \lambda \|\vec{f}\|_1, \tag{4}$$

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on November 30,2021 at 19:43:13 UTC from IEEE Xplore. Restrictions apply.

This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

where $\vec{r}^{\,k} = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$ and $\lambda = \frac{\tau}{\alpha_k}$. Then, the subproblem in (4) can be separated into scalar minimization problems (see [18] for details). Note that the objective function is separable on the variables $\vec{f}$. Thus (4) decouples into n four-dimensional problems of the form

$$\begin{aligned} f^{k+1} = \underset{f=[z_p;z_c;y_p;y_c]\in\mathbb{R}^4}{\arg\min} \quad & \tfrac{1}{2}\|f - r^k\|_2^2 + \lambda\|f\|_1 \\ \text{subject to} \quad & 0 \le y_p, y_c \le 1 \\ & 0 \le z_c \le z_p + y_p \le 1 \\ & 0 \le z_p \le z_c + y_c \le 1, \end{aligned} \tag{5}$$

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

C. Optimization Approach

In this paper, we propose solving (5) using two methods, both of which are based on block-coordinate descent approaches, which work as follows. Method I first fixes the homozygous indicator variables, z_σ, and minimizes over the heterozygous indicator variables, y_σ. In the next step, the heterozygous variables are fixed, and (5) is minimized over the homozygous variables. In contrast, Method II fixes all but one individual and minimizes (5) over the indicator variables for that particular individual. In subsequent steps, the variables corresponding to some other individual are minimized while fixing the variables for all other individuals. Both methods continue this block-coordinate descent approach until the iterates satisfy a pre-determined convergence criteria. We now describe each method in more detail.

METHOD I: This method solves (5) by alternating between homozygous and heterozygous indicator variables. In particular, it consists of the following steps.

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

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

$$\begin{aligned} (\hat{y}_p^{(i)}, \hat{y}_c^{(i)}) = \underset{y_p,y_c\in\mathbb{R}}{\arg\min} \quad & \tfrac{1}{2}(y_p - a_2)^2 + \tfrac{1}{2}(y_c - b_2)^2 \\ \text{subject to} \quad & 0 \le y_p, y_c \le 1 \\ & \hat{z}_c^{(i-1)} - \hat{z}_p^{(i-1)} \le y_p \le 1 - \hat{z}_p^{(i-1)} \\ & \hat{z}_p^{(i-1)} - \hat{z}_c^{(i-1)} \le y_c \le 1 - \hat{z}_c^{(i-1)}, \end{aligned} \tag{6}$$

where $a_2 = r^k_{y_p} - \lambda$ and $b_2 = r^k_{y_c} - \lambda$. Note that the bounds on y_p and y_c are simple bounds. Thus the feasible region is a simple rectangle (see Fig. 2).

Step 2: Suppose we have obtained $\hat{y}_p^{(i)}$ and $\hat{y}_c^{(i)}$ from Step 1. Then to obtain the solution for the current iteration $\hat{z}_p^{(i)}$ and $\hat{z}_c^{(i)}$, we complete the square and solve

$$\begin{aligned} (\hat{z}_p^{(i)}, \hat{z}_c^{(i)}) = \underset{z_p,z_c\in\mathbb{R}}{\arg\min} \quad & \tfrac{1}{2}(z_p - a_1)^2 + \tfrac{1}{2}(z_c - b_1)^2 \\ \text{subject to} \quad & z_p - \hat{y}_c^{(i)} \le z_c \le z_p + \hat{y}_p^{(i)}, \\ & 0 \le z_p \le 1 - \hat{y}_p^{(i)},\ 0 \le z_c \le 1 - \hat{y}_c^{(i)}. \end{aligned} \tag{7}$$

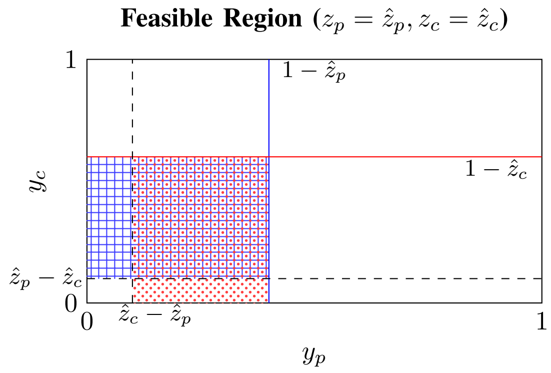

Fig. 2. Feasible region corresponding to the constraints in (6). The blue region represents the admissible set of solutions when $z_p - z_c \ge 0$ and the red region represents the feasible region when $z_p - z_c < 0$.

The feasible region is shown in Fig. 3.

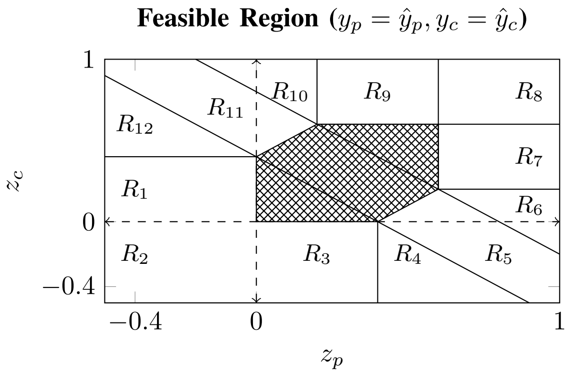

Fig. 3. Feasible region obtained from applying the constraints to (7), where the shaded grid region represents the admissible set of solutions when $\hat{y}_p \le 1 - \hat{y}_c$. If this condition is not satisfied, then we project onto the rectangular region obtained when $\hat{y}_p = 1 - \hat{y}_c$.

Solution. Note that both problems (6) and (7) have closed form solutions since the level sets of the objective functions in both problems are isotropic, and thus the minimizer can be easily obtained by projecting the unconstrained solution to the feasible set. In particular, the solution to Step 1 is given by

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

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

METHOD II: This method solves (5) by alternating between individuals. In particular, it consists of the following steps.

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

978-1-5090-2984-6/17/$31.00 ©2017 IEEE

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on November 30,2021 at 19:43:13 UTC from IEEE Xplore. Restrictions apply.

This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

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

TABLE I
SOLUTIONS TO (7) GIVEN a_1 AND b_1 FOR THE NONTRIVIAL REGION PROJECTIONS IN FIG. 3 WHEN $\hat{y}_p \leq 1 - \hat{y}_c$. HERE $r = a_1 + b_1,\ s = -a_1 + 2 - \hat{y}_c - 2\hat{y}_p,\ t = -a_1 + 2 - 2\hat{y}_c - \hat{y}_p$.

the following rule:

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

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

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

$$\begin{aligned} (\hat{z}_c^{(i)}, \hat{y}_c^{(i)}) \quad = \quad & \underset{z_c, y_c \in \mathbb{R}}{\text{arg min}} \quad \tfrac{1}{2}(z_c - b_1)^2 + \tfrac{1}{2}(y_c - b_2)^2 \qquad (8) \\ & \text{subject to} \quad 0 \leq z_c, y_c \leq 1 \\ & \qquad\qquad\quad z_c \leq \hat{z}_p^{(i-1)} + \hat{y}_p^{(i-1)} \leq 1 \\ & \qquad\qquad\quad \hat{z}_p^{(i-1)} \leq z_c + y_c \leq 1, \end{aligned}$$

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

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

$$\begin{aligned} (\hat{z}_p^{(i)}, \hat{y}_p^{(i)}) \quad = \quad & \underset{z_p, y_p \in \mathbb{R}}{\text{arg min}} \quad \tfrac{1}{2}(z_p - a_1)^2 + \tfrac{1}{2}(y_p - a_2)^2 \qquad (9) \\ & \text{subject to} \quad 0 \leq z_p, y_p \leq 1 \\ & \qquad\qquad\quad z_p \leq \hat{z}_c^{(i-1)} + \hat{y}_c^{(i-1)} \leq 1 \\ & \qquad\qquad\quad \hat{z}_c^{(i-1)} \leq z_p + y_p \leq 1, \end{aligned}$$

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

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

We note that the constraints Steps 1 and 2 are equivalent. Thus, the feasible region corresponding to the constraints are the same. For example, the feasible region for Step 1 is shown in Fig. 4.

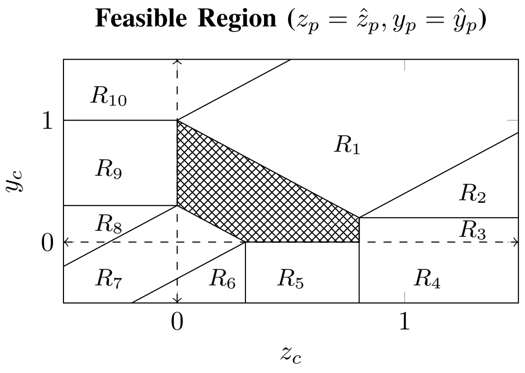

Fig. 4. Feasible region corresponding to the constraints, where the shaded grid region represents the admissible set of solutions.

TABLE II
SOLUTIONS TO (8) GIVEN b_1 AND b_2 FOR THE NONTRIVIAL PROJECTION REGIONS. HERE $\hat{q} = \hat{z}_p + \hat{y}_p$. THE VALUES (r_i, s_i) FOR $i = \{1, 2\}$ ARE GIVEN IN TABLE III.

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

Solution. The objective functions in both subproblems (8) and (9) are quadratic functions with the identity matrix as second-derivatives. Thus, the level curves of the objective functions are concentric circles centered around (b_1, b_2) and (a_1, a_2), respectively. This implies that the minimizer of the constrained subproblems are the orthogonal projections of (b_1, b_2) and (a_1, a_2) onto the feasible regional. Therefore, each subproblem has a closed-form solution and can be thus solved efficiently. In particular, for Step 1, the constrained solution is the projection of the unconstrained solution (b_1, b_2) onto the feasible set and can be found in Table II. The constrained solution to Step 2 is similarly defined.

III. RESULTS

In order to evaluate the effectiveness of our proposed method, we implemented it in MATLAB by modifying the existing SPIRAL approach [18] to solve subproblems and quantified our classification of SVs on both simulated and real genomes. We compare the two proposed methods in Sec. II-C. The regularization parameters (τ) for all experiments were chosen to obtain the maximum area under curve, AUC, for the receiver operating characteristic (ROC). The algorithm terminates if the relative difference between consecutive iterates converged to $\|\vec{f}^{k+1} - \vec{f}^k\|_2 / \|\vec{f}^k\|_2 \leq 10^{-8}$.

978-1-5090-2984-6/17/$31.00 ©2017 IEEE

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on November 30,2021 at 19:43:13 UTC from IEEE Xplore. Restrictions apply.

TABLE III
THE VALUES OF (r_1, s_1) AND (r_2, s_2) IN TABLE II.

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

A. *Simulated Data*

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

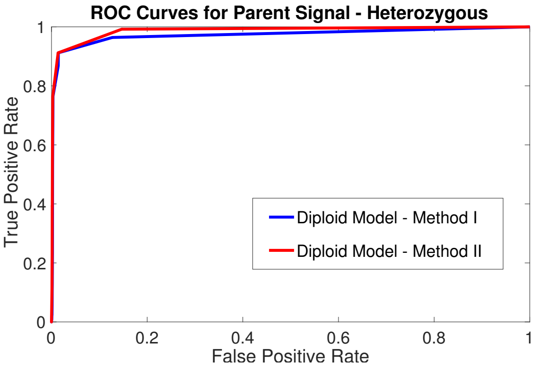

Fig. 5. ROC curves illustrating the false positive rate vs true positive rate for the reconstruction of the heterozygous parent signal in the simulated data with $k_p = 4$, $k_c = 4$, and 80% similarity of variants between parents using both methods with $\epsilon = 0.16$.

Analysis. We first examine the parent signal reconstruction. The false positive rate vs. true positive rate for the reconstruction of the heterozygous parent signal with coverages $k_p = 4$, $k_c = 4$, 80% similarity of variants between parents, and an error level of $\epsilon = 0.16$ is presented in Fig. 5. We note that Method II improves SV detection when compared to Method I (see Fig. 5). Based on AUC measurements, we observe that for the parent signal (both homozygous and heterozygous indicator variables) there was an improvement in reconstructions with the fix one individual method as we decrease the error level. Furthermore, we observed a higher accuracy for the homozygous parent signal reconstructions as we increase the percentage similarity between the parents. However, this pattern was the opposite for the heterozygous parent signal.

Next, we examined the child signal reconstruction. Fig. 6 illustrates the false positive rate vs. true positive rate for the reconstruction of the homozygous child signal with coverages $k_p = 4$, $k_c = 4$, 80% similarity of variants between parents, and an error level of $\epsilon = 0.08$. For false positive rate values > 0.10 and true positive rate values < 0.90, no significant difference could be discerned. Accordingly, the axes were reduced in order to provide a more detailed view of the comparison between the two methods. In this case, we notice both methods perform similarly (see Fig. 6). Based on AUC measurements, we observed that the homozygous child signal reconstructions improved as we increase the percentage similarity between the parents. We did not observe this type of improvement in signal reconstruction for the child signal when varying the error level.

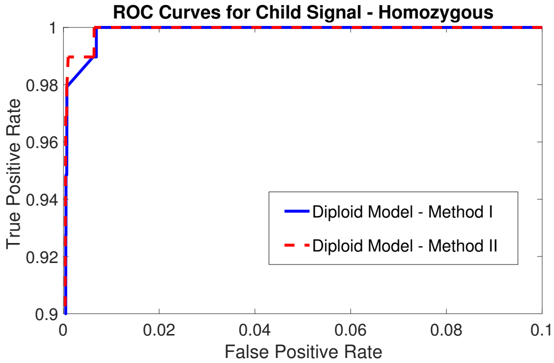

Fig. 6. ROC curves illustrating the false positive rate vs true positive rate for the reconstruction of the homozygous child signal in the simulated data with $k_p = 4$, $k_c = 4$, and 80% similarity of variants between parents using both methods with $\epsilon = 0.08$.

B. *1000 Genomes Project Trio Data*

Next, we apply our proposed methods to 1000 Genomes Project [19] father-mother-daughter CEU trio data (NA12891, NA12892, NA12878). The genomes in Pilot 1 were aligned to NCBI36 and sequenced at approximately 4× coverage. Experimentally validated (reported) deletions longer than 250bp are taken as the true signal. We use the reported genotype, unless marked with *LowQual*, to determine whether the reported deletion was either heterozygous or homozygous. After applying this filtering, we create the vectors $\vec{z}, \vec{y}$, representing the indicator variables for the genotype at each location.

Analysis. The number of candidate deletion locations is $n = 57,078$ for each CEU genome. The total number of deletions, both heterozygous and homozygous, were 686, 637, and 724 for the father, mother, and child, respectively. In the 1000 Genomes data and the simulated data, we observe similar improvement trends in our proposed method for both parent heterozygous signals and child homozygous signals. Since the experimentally validated set of deletions may not be complete, we compare the number of predicted heterozygous novel deletions to validated SVs in Fig. 7. Further, if we consider the broader question of correctly identifying a deletion, regardless of genotype, we achieve similar results to the original proposed method (see Fig. 8). These similar results are achieved at a lower computational cost in a more general framework. We report the improved computational cost of our method in comparison to the original method in Fig. 9.

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on November 30,2021 at 19:43:13 UTC from IEEE Xplore. Restrictions apply.

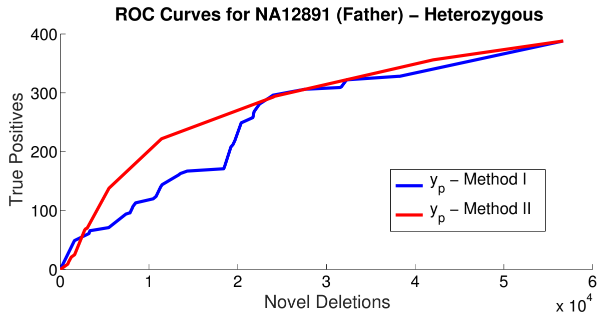

Fig. 7. ROC curves depicting novel deletions vs true positives for the reconstruction of heterozygous CEU NA12891 (father) signal. Here, $k_{p_1} = 4, k_c = 4$, with $\tau = 2.34 \times 10^{-10}$ and $\epsilon = 0.01$.

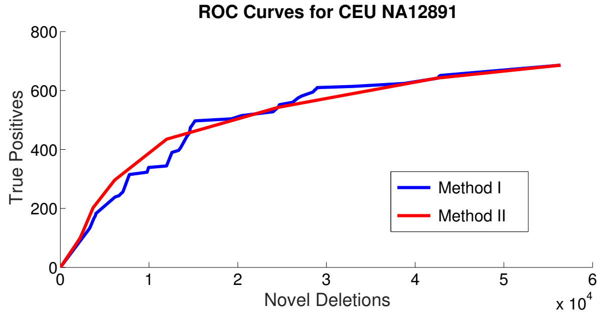

Fig. 8. ROC curves depicting novel deletions vs true positives for the reconstruction the combined heterozygous and homozygous CEU NA12891 (father) signal. Here, we illustrate comparable recall rates between our generalized method to our previous approach with $k_{p_1} = 4, k_c = 4, \tau = 2.34 \times 10^{-10}$, and $\epsilon = 0.01$.

IV. Conclusions

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

Acknowledgments: This work was supported by NSF Grant DMS 1359484.

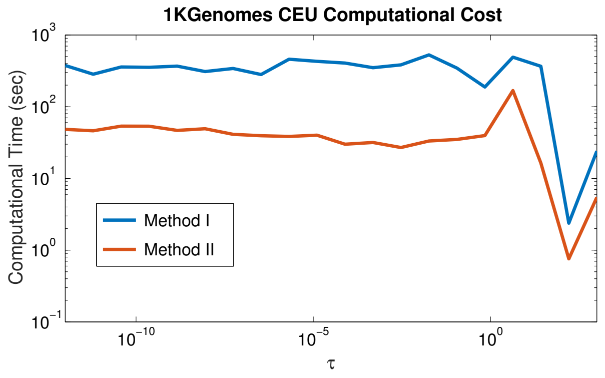

Fig. 9. Given the heterozygous and homozygous observations $\vec{s}$, we plot the computational time (in seconds) in reconstructing the true signal for the CEU dataset (NA12891 and NA12878). We observe a general reduction of computational cost for Method II for a range of penalty values τ.

References

[1] J. Pevsner, *Bioinformatics and functional genomics*. John Wiley & Sons, 2015.

[2] 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*, vol. 14, no. 2, pp. 125–138, 2013.

[3] . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[4] J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci *et al.*, "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, no. 7191, pp. 56–64, 2008.

[5] G. of the Netherlands Consortium *et al.*, "Whole-genome sequence variation, population structure and demographic history of the dutch population," *Nature Genetics*, vol. 46, no. 8, pp. 818–825, 2014.

[6] H. Stefansson, A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. G. Gudnadottir *et al.*, "A common inversion under selection in europeans," *Nature genetics*, vol. 37, no. 2, pp. 129–137, 2005.

[7] S. S. Sindi and B. J. Raphael, "Identification of structural variation," *Genome Analysis: Current Procedures and Applications*, p. 1, 2014.

[8] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing," *Nature methods*, vol. 6, pp. S13–S20, 2009.

[9] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke *et al.*, "Breakdancer: an algorithm for high-resolution mapping of genomic structural variation," *Nature methods*, vol. 6, no. 9, pp. 677–681, 2009.

[10] 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*, vol. 28, no. 18, pp. i333–i339, 2012.

[11] 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*, vol. 20, no. 5, pp. 623–635, 2010.

[12] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[13] 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 *Proceedings of IEEE Workshop on Statistical Signal Processing*, 2016.

[14] ——, "Constrained variant detection with sparc: Sparsity, parental relatedness, and coverage," in *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2016.

[15] M. A. Doyle, J. Li, K. Doig, A. Fellowes, and S. Q. Wong, "Studying cancer genomics through next-generation dna sequencing and bioinformatics," *Clinical Bioinformatics*, pp. 83–98, 2014.

[16] B. J. Raphael, "Structural variation and medical genomics," *PLoS Comput Biol*, vol. 8, no. 12, p. e1002821, 2012.

[17] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, pp. 75–81, 2015.

[18] 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*, vol. 21, pp. 1084 – 1096, 2011.

[19] 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*, vol. 467, no. 7319, pp. 1061–1073, 2010.

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on November 30,2021 at 19:43:13 UTC from IEEE Xplore. Restrictions apply.