SPARSE NEGATIVE BINOMIAL SIGNAL RECOVERY FOR GENOMIC VARIANT PREDICTION IN DIPLOID SPECIES

Jocelyn Ornelas Munoz[⋆]*, Erica M. Rutter*[⋆]*, Mario Banuelos*[†]*, Suzanne S. Sindi*[⋆]*, Roummel F. Marcia*[⋆]

[⋆] Department of Applied Mathematics University of California, Merced
[†]Department of Mathematics California State University, Fresno

ABSTRACT

Structural variants (SVs) – such as insertions, deletions, and duplications of an individual's genome – are associated with genetic diseases and promotion of genetic diversity. Detecting SVs of an unknown genome is a mathematically challenging problem since SVs are rare and prone to low-coverage noise. We developed a computational method which seeks to improve existing SV detection methods in three ways: First, we implement an optimization approach consisting of a negative binomial log-likelihood objective function. Second, we use a block-coordinate descent approach to simultaneously predict if an SV is homozygous or heterozygous given genomic data of related individuals. Third, we model a biologically realistic scenario where variants in the child are either inherited or novel. We validate our framework with simulated data and demonstrate improvements in predicting SVs and detecting false positives.

***Index Terms*—** Structural variants, sparse signal recovery, nonconvex optimization, computational genomics

1. 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 in a human genome is approximately six billion letters [1]. Humans are diploid species, which means they have two copies of their genome and receive one copy from each parent. In a human organism, each cell contains a copy of the organism's genome, which is replicated through the process of cell division. As DNA molecules replicate, changes in the DNA sequence —genetic variants—may occur. Most of the time, genetic variations have no effect at all. However, sometimes the effect of these changes may be harmful and may be passed on from one generation to the next. Structural variants (SVs), a type of genetic variant characterized by insertions, deletions, inversions, etc., of > 50 letters, are rare occurrences of DNA rearrangements which can provide great insights into regulation of gene expression, ethnic diversity, large-scale chromosome evolution, and their role in disease susceptibility [1], [2], [3], [4].

The common approach for SV prediction has been to map the sequencing data to a reference genome and computationally identify statistically significant deviations from the expected mapping signals consistent with each class of SV [5], [6]. 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 [4], [5], [7], [8]. Despite the fact that the rate of *de novo* SVs is negligible [9], and therefore most SVs present in a child are inherited from one of their parents, most computational SV pipelines do not consider information from familial genomes [10], [11], [12], [13], [14].

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 Mendelian inheritance [15]. We further enforce sparsity in our predictions through an ℓ_1 penalty term. We describe the methodology in Section 2, provide results on simulated data in Section 3 and conclude in Section 4.

2. METHODS

Here, we describe our computational framework for predicting SVs for related individuals. We use diploid data from one parent (P) and one child (C) —which is separated to consider both inherited (H) and novel (N) SVs individually —for mathematical and computational simplicity. Each signal consists of n candidate locations in the genome where an SV may be present. For each individual signal $i \in \{P, H, N\}$ in our model, we consider two signals that take on binary values: a heterozygous indicator $\vec{y}_i \in \{0,1\}^n$ and a homozygous indicator $\vec{z}_i \in \{0,1\}^n$ such that the true signal $\vec{f}_i = 2\vec{z}_i + \vec{y}_i$ [16]. The heterozygous vector, $\vec{y}_i$, indicates that the signal of the individual has one copy of the SV in their genome while the homozygous vector, $\vec{z}_i$, indicates that the individual has two copies of the SV in their genome. Thus, if an individual is heterozygous for an SV at a 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 $(\vec{y}_i)_j = 0$ [16].

2.1. Observational Model

We will denote the observation vectors for the parent and the child by the vectors $\vec{s}_P \in \mathbb{R}^n$, $\vec{s}_C \in \mathbb{R}^n$, respectively. We assume the observed data follows a negative binomial distribution:

$$\begin{bmatrix}\vec{s}_P \\ \vec{s}_C\end{bmatrix} \sim \text{NegBin}\left(\begin{bmatrix}\vec{z}_P(2\lambda_P - \varepsilon) \\ \vec{z}_H(2\lambda_C - \varepsilon) + \vec{y}_H(\lambda_C - \varepsilon) \\ + \vec{z}_N(2\lambda_C - \varepsilon) + \vec{y}_N(\lambda_C - \varepsilon)\end{bmatrix}\right) \quad (1)$$

where λ_P, λ_C represent the sequencing coverage —the average number of reads that align to known reference bases—of the parent

This research is funded by NSF Grants DMS 1840265 and IIS 1741490. Approved for Public Release; Distribution Unlimited. Public Release Case Number 22-3373. J. Ornelas Munoz's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author.

and the child, respectively and $\varepsilon > 0$ is used to reflect the measurement errors incurred through the sequencing and mapping processes [16], [17].

Let

$$\vec{s} = \begin{bmatrix} \vec{s}_P \\ \vec{s}_C \end{bmatrix}, \quad \vec{z} = \begin{bmatrix} \vec{z}_P \\ \vec{z}_H \\ \vec{z}_N \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} \vec{y}_P \\ \vec{y}_H \\ \vec{y}_N \end{bmatrix}, \quad \vec{f} = \begin{bmatrix} \vec{z} \\ \vec{y} \end{bmatrix},$$

where $\vec{f} \in \{0,1\}^{6n}$. We express the general observation model as

$$\vec{s} \sim \text{NegBin}(A\vec{f} + \varepsilon \mathbf{1})$$

where $\mathbf{1} \in \mathbb{R}^{2n}$ is the vector of ones and $A = [A_1 \ A_2] \in \mathbb{R}^{2n \times 6n}$ is the sequence coverage matrix with A_1, A_2 given by:

$$A_1 = \left[\begin{array}{c|c|c} (2\lambda_P - \varepsilon)I_n & 0 & 0 \\ \hline 0 & (2\lambda_C - \varepsilon)I_n & (2\lambda_C - \varepsilon)I_n \end{array}\right]$$

$$A_2 = \left[\begin{array}{c|c|c} (\lambda_P - \varepsilon)I_n & 0 & 0 \\ \hline 0 & (\lambda_C - \varepsilon)I_n & (\lambda_C - \varepsilon)I_n \end{array}\right]$$

where $I_n \in \mathbb{R}^{n \times n}$ is the $n \times n$ identity matrix.

2.2. Optimization Formulation

We will assume a Negative Binomial process to model the noise in the sequencing and mapping measurements. The distribution can be parameterized in terms of its mean $\vec{\mu}_l = \vec{e_l}^T A\vec{f}$ and standard deviation $\vec{\sigma}_l^2 = (\vec{e_l}^T A\vec{f})_l + \frac{1}{r}(\vec{e_l}^T A\vec{f})_l^2$, $l = 1, \ldots, 2n$, where $\vec{e_l}$ represents the canonical standard basis vectors. We set the dispersion parameter $r = 1$ since this choice will maximize the standard deviation. With these considerations, the probability of observing the observation vector $\vec{s}$ given the true signal $\vec{f}$, is given by

$$p(\vec{s}\,|A\vec{f}) \ = \prod_{l=1}^{2n} \left(\frac{1}{1 + (A\vec{f})_l + \varepsilon}\right) \left(\frac{((A\vec{f})_l + \varepsilon)}{1 + (A\vec{f})_l + \varepsilon}\right)^{\vec{s}_l} \quad (2)$$

where $\varepsilon > 0$ represents the sequencing and mapping errors.

The solution space for inferring $\vec{f}$ from $\vec{s}$ is exponentially large for large n. Thus, we apply a continuous relaxation of $\vec{f}$ such that its elements lie between 0 and 1, i.e. $\mathbf{0} \leq \vec{f} \leq \mathbf{1}$, or equivalently,

$$\mathbf{0} \leq \ \vec{z}_i, \vec{y}_i \ \leq \mathbf{1}, \quad i \in \{P, H, N\}. \quad (3)$$

To simplify notation, we assume the inequalities read element-wise and denote $\mathbf{0}$ and $\mathbf{1}$ as the vector of all zeros and ones, respectively.

The continuous relaxation allows us to apply a gradient-based maximum likelihood approach to recover the indicator values $\vec{z}_i$ and $\vec{y}_i$ by estimating $A\vec{f}$ such that the probability of observing the vector of negative binomial data $\vec{s}$ is maximized under our statistical model. We seek to minimize the corresponding Negative Binomial negative log-likelihood function

$$F(\vec{f}) \equiv \sum_{l=1}^{2n} (1 + \vec{s}_l) \log\left(1 + \vec{e_l}^T A\vec{f} + \varepsilon\right) - \vec{s}_l \log\left(\vec{e_l}^T A\vec{f} + \varepsilon\right) \quad (4)$$

Familial Constraints. We incorporate additional constraints to leverage biological information about $\vec{f}$ to improve accuracy of the model. Since a structural variant cannot be both homozygous and heterozygous at the same time, we require that

$$\mathbf{0} \leq \ \vec{z}_i + \vec{y}_i \ \leq \mathbf{1}, \quad i \in \{P, H, N\}.$$

Recall the signal of the child is comprised of both inherited and novel structural variants, $\vec{f}_C = \vec{z}_H + \vec{y}_H + \vec{z}_N + \vec{y}_N$, where a structural variant cannot be both inherited and novel simultaneously.

$$\mathbf{0} \leq \ \vec{z}_H + \vec{y}_H + \vec{z}_N + \vec{y}_N \ \leq \mathbf{1}.$$

To account for relatedness, we assume the child can have an inherited homogeneous SV only if the parent has at least a heterogeneous SV. Similarly, the child can only have an inherited heterogeneous SV if the parent has at least a heterogeneous SV. On the other hand, if the parent has a homogeneous SV at a particular location, then the child must have at least a heterozygous SV at that location:

$$\mathbf{0} \leq \ \vec{z}_H \ \leq \ \vec{z}_P + \vec{y}_P \ \leq \mathbf{1}$$

$$\mathbf{0} \leq \ \vec{z}_P \ \leq \ \vec{z}_H + \vec{y}_H \ \leq \mathbf{1}$$

Finally, we note that novel structural variants in the child cannot be inherited from the parent. Thus, for a location j, if $(\vec{z}_N)_j + (\vec{y}_N)_j = 1$, then $(\vec{z}_P)_j + (\vec{y}_P)_j = 0$. Similarly, if $(\vec{z}_P)_j + (\vec{y}_P)_j = 1$, then $(\vec{z}_N)_j + (\vec{y}_N)_j = 0$,

$$\mathbf{0} \leq \ \vec{z}_N + \vec{y}_N \ \leq \ 1 - (\vec{z}_P + \vec{y}_P) \ \leq \mathbf{1}$$

We denote the set of all vectors satisfying these constraints by $\mathcal{S}$:

$$\mathcal{S} = \left\{ \vec{f} = \begin{bmatrix} \vec{z}_P \\ \vec{z}_H \\ \vec{z}_N \\ \vec{y}_P \\ \vec{y}_H \\ \vec{y}_N \end{bmatrix} \in \mathbb{R}^{6n} : \begin{array}{l} \mathbf{0} \leq \ \vec{z}_i + \vec{y}_i \ \leq \mathbf{1} \\ \mathbf{0} \leq \ \vec{z}_H + \vec{y}_H + \vec{z}_N + \vec{y}_N \ \leq \mathbf{1} \\ \mathbf{0} \leq \ \vec{z}_H \ \leq \ \vec{z}_P + \vec{y}_P \ \leq \mathbf{1} \\ \mathbf{0} \leq \ \vec{z}_P \ \leq \ \vec{z}_H + \vec{y}_H \ \leq \mathbf{1} \\ \mathbf{0} \leq \ \vec{z}_N + \vec{y}_N \\ \qquad \leq \ 1 - (\vec{z}_P + \vec{y}_P) \ \leq \mathbf{1} \end{array} \right\}$$

Sparsity-promoting ℓ_1 penalty. Structural variants are rare in an individual's genome. Thus, a common challenge with SV recovery is predicting false positive SVs by mistaking fragments that are incorrectly mapped against the reference genome [16]. To model this biological reality, we incorporate an ℓ_1-norm penalty in our objective function to enforce sparsity in our predictions. Further, we assume novel SVs are even more rare since they are not inherited from a parent. Therefore, we use two penalty terms: one for the parent SVs, $\vec{z}_P, \vec{y}_P$, and the child's inherited SVs, $\vec{z}_H, \vec{y}_H$, and another penalty term for the child's novel SVs, $\vec{z}_N, \vec{y}_N$. We define the penalty as:

$$\text{pen}(\vec{f}) = (\|\vec{z}_P\|_1 + \|\vec{z}_H\|_1 + \|\vec{y}_P\|_1 + \|\vec{y}_H\|_1) + \gamma(\|\vec{z}_N\|_1 + \|\vec{y}_N\|_1)$$

where $\gamma > 1$ is the penalty term that enforces greater sparsity in the child's novel SVs. Our objective function then takes the form:

$$\begin{array}{ll} \underset{\vec{f} \in \mathbb{R}^{6n}}{\text{minimize}} & F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \quad (5)$$

where $F(\vec{f})$ is the Negative Binomial negative log-likelihood function shown in Equation (4) and $\tau > 0$ is a regularization parameter. Our approach in solving the minimization problem in Equation (5) employs sequential quadratic approximations to the Negative Binomial negative log-likelihood $F(\vec{f})$. More specifically, at iteration k, we compute a separable quadratic approximation to $F(\vec{f})$ using its second-order Taylor series approximation at $\vec{f}^k$ and approximate the Hessian matrix by a scalar multiple of the identity matrix, $\alpha_k I$ [18]. This quadratic approximation is then defined as

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

which we use as a surrogate function for $F(\vec{f})$ in Equation (5). Using this approximation, the next iterate is given by

$$\vec{f}^{k+1} = \begin{array}{ll} \underset{\vec{f}\in\mathbb{R}^{6n}}{\arg\min} & F^k(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \tag{6}$$

We reformulate this constrained quadratic subproblem into the following equivalent sequence of subproblems (see [18]):

$$\vec{f}^{k+1} = \begin{array}{ll} \underset{\vec{f}\in\mathbb{R}^{6n}}{\arg\min} & \mathcal{Q}(\vec{f}) = \frac{1}{2}\|\vec{f} - \vec{r}^k\|_2^2 + \frac{\tau}{\alpha_k}\text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \tag{7}$$

where $\vec{r}^k = [\vec{r}^k_{z_P}, \vec{r}^k_{z_H}, \vec{r}^k_{z_N}, \vec{r}^k_{y_P}, \vec{r}^k_{y_H}, \vec{r}^k_{y_N}]^T = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$
Our objective function $\mathcal{Q}(\vec{f})$ is separable and decouples into the function $\mathcal{Q}(\vec{f}) = \sum_{j=1}^n \mathcal{Q}_j(\vec{z}_P, \vec{z}_H, \vec{z}_N, \vec{y}_P, \vec{y}_H, \vec{y}_N)$, where

$$\begin{aligned}
&\mathcal{Q}_j(\vec{z}_P, \vec{z}_H, \vec{z}_N, \vec{y}_P, \vec{y}_H, \vec{y}_N) = \\
&\frac{1}{2}\Big\{((\vec{z}_P - \vec{r}^k_{\vec{z}_P})_j)^2 + ((\vec{z}_H - \vec{r}^k_{\vec{z}_H})_j)^2 + ((\vec{z}_N - \vec{r}^k_{\vec{z}_N})_j)^2 \\
&\quad + ((\vec{y}_P - \vec{r}^k_{\vec{y}_P})_j)^2 + ((\vec{y}_H - \vec{r}^k_{\vec{y}_H})_j)^2 + ((\vec{y}_N - \vec{r}^k_{\vec{y}_N})_j)^2\Big\} \\
&+ \frac{\tau}{\alpha_k}\Big\{|(\vec{z}_P)_j| + |(\vec{z}_H)_j| + \gamma|(\vec{z}_N)_j| \\
&\quad + |(\vec{y}_P)_j| + |(\vec{y}_H)_j| + \gamma|(\vec{y}_N)_j|\Big\}
\end{aligned}$$

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

$$\begin{aligned}
\vec{f}^{k+1} = &\min_{\substack{z_P, z_H, z_N, \\ y_P, y_H, y_N \in \mathbb{R}}} \frac{\tau}{\alpha_k}\Big\{|(\vec{z}_P)_j| + |(\vec{z}_H)_j| + \gamma|(\vec{z}_N)_j| \\
&+ |(\vec{y}_P)_j| + |(\vec{y}_H)_j| + \gamma|(\vec{y}_N)_j|\Big\} \\
&+ \frac{1}{2}\Big\{((\vec{z}_P - \vec{r}^k_{\vec{z}_P})_j)^2 + ((\vec{z}_H - \vec{r}^k_{\vec{z}_H})_j)^2 + ((\vec{z}_N - \vec{r}^k_{\vec{z}_N})_j)^2 \\
&+ ((\vec{y}_P - \vec{r}^k_{\vec{y}_P})_j)^2 + ((\vec{y}_H - \vec{r}^k_{\vec{y}_H})_j)^2 + ((\vec{y}_N - \vec{r}^k_{\vec{y}_N})_j)^2\Big\} \\
&\text{subject to } \vec{f} \in S
\end{aligned} \tag{8}$$

where f_i and r_{z_i}, r_{y_i} are scalar components of $\vec{f}_i$ and $\vec{r}_{z_i}, \vec{r}_{y_i}$, respectively, at the same location.

Note that Equation (8) has closed form solutions (obtained by completing the square and ignoring constant terms), and thus the constrained minimizer can be easily obtained by projecting the unconstrained solution to the feasible set.

2.3. Optimization Approach

We solve our problem using an alternating block-coordinate descent approach, following the methods in [16], [17], [19]. We fix all but one individual and solve Equation (7) over both indicator variables for that individual. We continue successively minimizing both indicator variables for each individual while the other individuals are fixed. The feasible region for this step is illustrated in Figure 1.

Step 0: We begin by computing the unconstrained minimizer of Equation (7), which is given by

$$\begin{aligned}
\vec{f} = &[\vec{r}_{z_P} - \tfrac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{z_H} - \tfrac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{z_N} - \tfrac{\tau}{\alpha_k}\gamma\mathbf{1}_n, \\
&\vec{r}_{y_P} - \tfrac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{y_H} - \tfrac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{y_N} - \tfrac{\tau}{\alpha_k}\gamma\mathbf{1}_n]^T,
\end{aligned}$$

where $\mathbf{1}_n \in \mathbb{R}^n$. Next, we initialize the child's inherited and novel indicator variables by applying the following rule:

$$\begin{aligned}
\hat{z}_H^{(0)} &= \{0, r^k_{z_H} - \frac{\tau}{\alpha_k}, 1\}, \quad \hat{z}_N^{(0)} = \{0, r^k_{z_N} - \frac{\tau}{\alpha_k}\gamma, 1\}, \\
\hat{y}_H^{(0)} &= \{0, r^k_{y_H} - \frac{\tau}{\alpha_k}, 1\}, \quad \hat{y}_N^{(0)} = \{0, r^k_{y_N} - \frac{\tau}{\alpha_k}\gamma, 1\}
\end{aligned}$$

Further, if $\hat{z}_H^{(0)} + \hat{y}_H^{(0)} > 1$, then we let $\hat{z}_H^{(0)} = \hat{y}_H^{(0)} = 0.5$. We adjust $\hat{z}_N^{(0)}$ and $\hat{y}_N^{(0)}$ similarly. We apply these rules to ensure our initialization is consistent with the set of feasible solutions. To initialize the parent indicator variables, we let $\hat{z}_P^{(0)} = r^k_{z_P} - \frac{\tau}{\alpha_k}$ and $\hat{y}_P^{(0)} = r^k_{y_P} - \frac{\tau}{\alpha_k}$. We initialize the index with $k = 1$.

Step 1: Once we have obtained estimates for child's inherited and novel diploid indicator variables, $\hat{z}_H^{(k-1)}$, $\hat{y}_H^{(k-1)}$, $\hat{z}_N^{(k-1)}$, $\hat{y}_N^{(k-1)}$, from the previous iteration, we project $\hat{z}_P^{(k-1)}$ and $\hat{y}_P^{(k-1)}$ onto the feasible set S with fixed inherited and novel variables to obtain the new parent indicator values $\hat{z}_P^{(k)}$ and $\hat{y}_P^{(k)}$. This projection is similar to the projections done in [16, 19].

Step 2: Using our estimates for the parent diploid indicator variables from Step 1, we project $\hat{z}_H^{(k-1)}, \hat{y}_H^{(k-1)}$ onto our feasible set S with

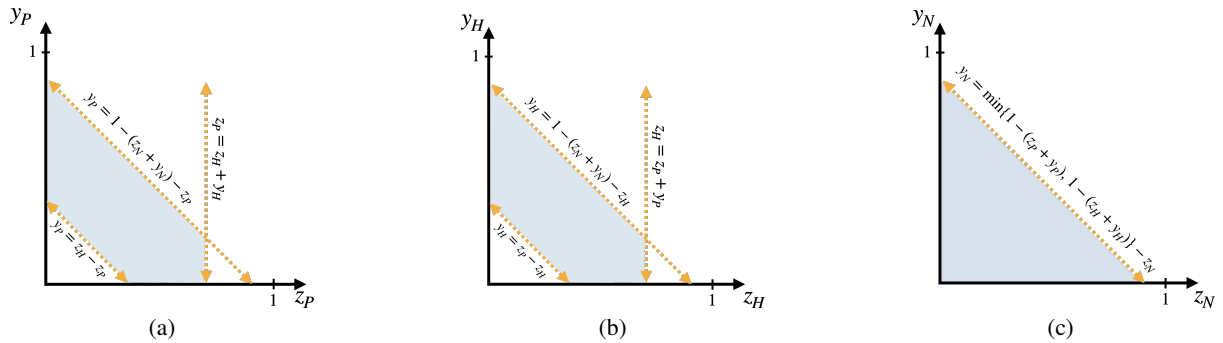

Fig. 1. The feasible set is shown above by the shaded region for each step of the proposed block-coordinate descent approach. (a) Step 1: We obtain the solution for the parent's variables $\vec{z}_P$ and $\vec{y}_P$ given fixed child inherited and novel indicator variables $\vec{z}_H, \vec{z}_N, \vec{y}_H$, and $\vec{y}_N$. (b) Step 2: We obtain the child's inherited indicator variables $\vec{z}_H$ and $\vec{y}_H$ by fixing $\vec{z}_P, \vec{y}_P, \vec{z}_N$, and $\vec{y}_N$. (c) Step 3: We obtain the solution for the child's novel indicator variables $\vec{z}_N$ and $\vec{y}_N$ by fixing $\vec{z}_P, \vec{y}_P, \vec{z}_H$, and $\vec{y}_H$.

fixed parent and child's novel indicator variables to obtain the new child's inherited indicator variables $\hat{z}_H^{(k)}$ and $\hat{y}_H^{(k)}$.

Step 3: Using estimates for the child's inherited diploid indicator variables from Step 2, we project $\hat{z}_N^{(k-1)}, \hat{y}_N^{(k-1)}$ onto our feasible set S with fixed parent and child's inherited indicator variables to obtain the new child's novel indicator variables $\hat{z}_N^{(k)}$ and $\hat{y}_N^{(k)}$.

We repeat Steps 1, 2, and 3 until some convergence criteria are satisfied. Steps 1 and 2 result in identical feasible regions. In this work, our convergence criteria is that relative difference between consecutive iterates converged to $\|\vec{f}^{k+1} - \vec{f}^{k}\|/\|\vec{f}^{k}\| \leq 10^{-8}$.

3. RESULTS

To evaluate the effectiveness of our proposed method, we implemented it in MATLAB by modifying the existing SPIRAL approach [18] to include the negative binomial statistical method [18] to solve the quadratic subproblems. We refer to the new algorithm as NEgative Binomial optimization Using ℓ_1 penalty Algorithm (NEBULA). We compared the Poisson-based predictions with the Negative Binomial-based predictions.The regularization parameters (τ, γ) were hyperparameters selected to maximize the area under the curve (AUC) for the receiver operating characteristic (ROC).

Simulated Data. Similar to previous approaches, we simulated two parent signals of size 10^5 with a set number of structural variants and a set similarity of 80% between the parent signals [16, 19]. In the parent signals, 5000 locations were chosen at random to be structural variants. We then constructed the child signal by first applying a logical implementation of inheritance to $\lfloor 5000p \rfloor$ randomly selected parent structural variants (where p is the percent overlap between parent and child SVs). Next, we chose $(5000 - \lfloor 5000p \rfloor)$ locations from the remaining $(10^5 - 5000)$ that were not chosen as a parent variant to be novel variants in the child. After forming the true signals for each individual, the observed signals were generated by sampling from the Negative Binomial distribution with a given coverage and error. For the purpose of testing the proposed approach, only one parent signal was used.

Analysis. Figure 2 displays the Receiving Operating Characteristic (ROC) (top) and Precision-Recall (PR) (bottom) curves obtained for a simulated data set where the parents share 80% of their SVs. Given an optimal τ, γ values, our method is better able to reconstruct the homozygous signals for each individual despite large sequencing and mapping error, $\varepsilon = 0.5$. Similar to previous work, we use the area under the curve (AUC) for the ROC curves to measure the ability of SPIRAL and NEBULA to distinguish between classes. Since SVs are very rare, our dataset has few true positives. A more informative metric is to examine Precision-Recall curves to gain a deeper understanding of the performance of our algorithm as it relates to false positives [20]. We see improvements in area under the curve and average precision for the parent and child's inherited signals. We also see comparable performance for the reconstruction of the child's novel signal. We observe that neither method is able to accurately reconstruct the novel child signal. We note that as this work only considers the relationship between one parent and one child. We hypothesize that including the information from both parents would enhance the ability to predict the child signal.

4. CONCLUSION

We present an optimization method for detecting both structural variants and their genotype (homozygous or heterozygous) from low-coverage DNA sequencing data in related individuals. This method leverages Mendelian inheritance to improve signal reconstruction of noisy data. This extends previous work that focused on a Poisson-based optimization algorithm. We compare our method to SPIRAL and applied them to simulated data to reconstruct heterozygous and homozygous signals. Overall, we achieve improved precision rates for total SV detection with our method. In future studies, we intend to apply this work to a two parent and one child framework.

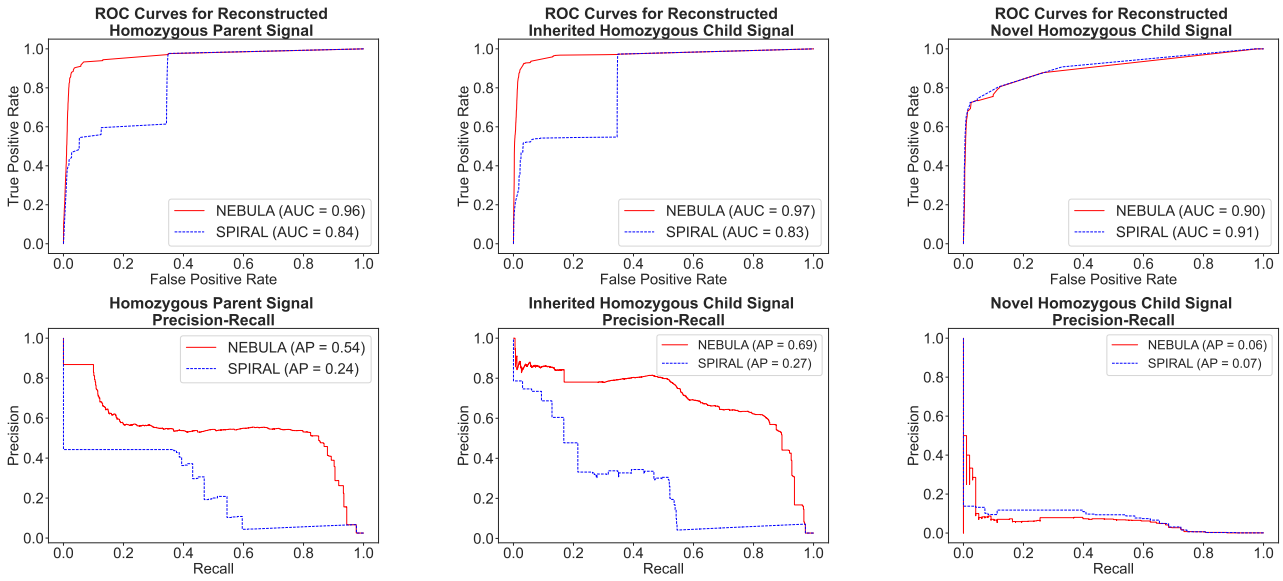

Fig. 2. ROC curves (top) and Precision-Recall curves (bottom) for the reconstructed homozygous parent signal (left), reconstructed inherited homozygous child signal (center), and reconstructed novel homozygous child signal (right) for our NEBULA algorithm (red) compared with the SPIRAL algorithm (blue). The regularization parameters used were $\tau = 1$, $\gamma = 2$, the percent of novel SVs is 4, and the coverage values for each individual are as follows $(\lambda_P, \lambda_C) = (7, 3)$.

5. REFERENCES

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

[2] Melissa Spence, Mario Banuelos, Roummel F. Marcia, and Suzanne Sindi, "Detecting inherited and novel structural variants in low-coverage parent-child sequencing data," *Methods*, vol. 173, pp. 61–68, 2020.

[3] Alhafidz Hamdan and Ailith Ewing, "Unravelling the tumour genome: the evolutionary and clinical impacts of structural variants in tumourigenesis," *The Journal of Pathology*, 2022.

[4] Shunichi Kosugi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani, "Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing," *Genome Biology*, vol. 20, no. 1, pp. 1–18, 2019.

[5] Eric S Lander and Michael S Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, 1988.

[6] Lorenzo Tattini, Romina D'Aurizio, and Alberto Magi, "Detection of genomic structural variants from next-generation sequencing data," *Frontiers in bioengineering and biotechnology*, vol. 3, pp. 92, 2015.

[7] Paul Medvedev, Monica Stanciu, and Michael Brudno, "Computational methods for discovering structural variation with next-generation sequencing," *Nature Methods*, vol. 6, no. 11, pp. S13–S20, 2009.

[8] Michael M Khayat, Sayed Mohammad Ebrahim Sahraeian, Samantha Zarate, Andrew Carroll, Huixiao Hong, Bohu Pan, Leming Shi, Richard A Gibbs, Marghoob Mohiyuddin, Yuanting Zheng, et al., "Hidden biases in germline structural variant detection," *Genome Biology*, vol. 22, no. 1, pp. 1–15, 2021.

[9] The Genome of the Netherlands Consortium, "Whole-genome sequence variation, population structure and demographic history of the Dutch population," *Nature Genetics*, vol. 46, no. 8, pp. 818–825, 2014.

[10] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin 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.

[11] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel, "Delly: structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics*, vol. 28, no. 18, pp. i333–i339, 2012.

[12] Aaron R Quinlan, Royden A Clark, Svetlana Sokolova, Mitchell L Leibowitz, Yujun Zhang, Matthew E Hurles, Joshua C Mell, and Ira 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.

[13] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall, "Lumpy: a probabilistic framework for structural variant discovery," *Genome Biology*, vol. 15, no. 6, pp. 1–19, 2014.

[14] Kévin Uguen, Claire Jubin, Yannis Duffourd, Claire Bardel, Valérie Malan, Jean-michel Dupont, Laila El Khattabi, Nicolas Chatron, Antonio Vitobello, Pierre-Antoine Rollat-Farnier, et al., "Genome sequencing in cytogenetics: Comparison of short-read and linked-read approaches for germline structural variant detection and characterization," *Molecular Genetics & Genomic Medicine*, vol. 8, no. 3, pp. e1114, 2020.

[15] Genetic Alliance, "Understanding genetics: a district of columbia guide for patients and health professionals," 2010.

[16] Melissa Spence, Mario Banuelos, Roummel F. Marcia, and Suzanne Sindi, "Genomic signal processing for variant detection in diploid parent-child trios," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1318–1322.

[17] Andrew Lazar, Mario Banuelos, Suzanne Sindi, and Roummel F. Marcia, "Novel structural variant genome detection in extended pedigrees through negative binomial optimization," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*, 2021, pp. 563–567.

[18] Zachary T. Harmany, Roummel F. Marcia, and Rebecca M. Willett, "This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms—Theory and Practice," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1084–1096, 2012.

[19] Mario Banuelos, Lasith Adhikari, Rubi Almanza, Andrew Fujikawa, Jonathan Sahagún, Katharine Sanderson, Melissa Spence, Suzanne Sindi, and Roummel F. Marcia, "Sparse diploid spatial biosignal recovery for genomic variation detection," in *2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2017, pp. 275–280.

[20] Takaya Saito and Marc Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, pp. e0118432, 2015.