2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782090

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*
Email: {jornelasmunoz, erutter2, ssindi, rmarcia}@ucmerced.edu, mbanuelos22@csufresno.edu
* 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. Common approaches to detect SVs in an unknown genome require sequencing fragments of the genome, comparing them to a high-quality reference genome, and predicting SVs based on identified discordant fragments. We developed a computational method which seeks to improve existing SV detection methods in three ways: First, we implement an optimization approach using 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**

I. Introduction

The genome of an individual comprises a specific sequence of nucleotides (A, C, G, T) and is approximately six billion letters long in humans [1]. Humans are diploid organisms, inheriting two copies of their genome, one from each parent. Each cell in a human organism contains a replicated version of the genome, which is transmitted during cell division. As DNA molecules replicate, changes in the DNA sequence—genetic variants—may occur. Sometimes, these changes can have harmful effects that are inherited across generations. Structural variants (SVs) are a type of genetic variation characterized by insertions, deletions, inversions, and other alterations of more than 50 nucleotides. SVs are relatively rare occurrences but offer valuable insights into gene expression regulation, ethnic diversity, large-scale chromosome evolution, and their association with disease susceptibility [1], [2], [3], [4].

A common approach for SV prediction is to map individual sequencing data to a reference genome and computationally identify statistically significant deviations from the expected mapping signals [5], [6]. However, errors in both the sequencing and mapping process 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 leverage 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. Our approach extends previous Poisson-based methodologies [2], [15] by addressing a more realistic scenario, incorporating sequencing data that follows a negative binomial distribution from related diploid species with novel structural variants. Our contribution involves adapting the Sparse Poisson Intensity Reconstruction Algorithms (SPIRAL) [16] framework, initially developed for imaging applications, to meet the specific needs of our genomics application. We utilize instead likelihood-based approach to predict 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 [17]. We enforce sparsity in our predictions through an ℓ_1 penalty term. We describe the methodology in Section II, provide results on simulated data in Section III and conclude in Section IV.

II. 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. 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 $\vec{z}_i \in \{0,1\}^n$ indicating that the individual has one or two copies of the SV in their genome, respectively. The true signal is then $\vec{f}_i^* = 2\vec{z}_i + \vec{y}_i$ [18].

A. Observational Model

Observation vectors obtained from sequencing experiments conducted on both the parent and the child are given 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:

This material is based upon work supported by the National Science Foundation under Grants No. DMS-1840265 and IIS 1741490.

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

where λ_P, λ_C represent the sequencing coverage —the average number of reads that align to known reference bases—of the parent and the child, respectively and $\varepsilon > 0$ is used to reflect the measurement errors incurred through the sequencing and mapping processes [18], [19]. 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 $\vec{f} \in \{0,1\}^{6n}$. The general observation model is

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

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 as:

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

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

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

B. Optimization Formulation

We assume a Negative Binomial process to model the noise in the sequencing and mapping measurements. The negative binomial distribution is parameterized by 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 $r = 1$ to maximize the standard deviation. Thus, the probability of observing the observation vector $\vec{s}$ given the true signal $\vec{f}$, is given by

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

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. $\mathbf{0} \leq \vec{f} \leq \mathbf{1}$:

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

For simplicity, we assume inequalities read element-wise and denote $\mathbf{0}$ and $\mathbf{1}$ as the vector of 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

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

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, we require that

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

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.

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

$\mathcal{S}$ denotes the set of all vectors satisfying these constraints:

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

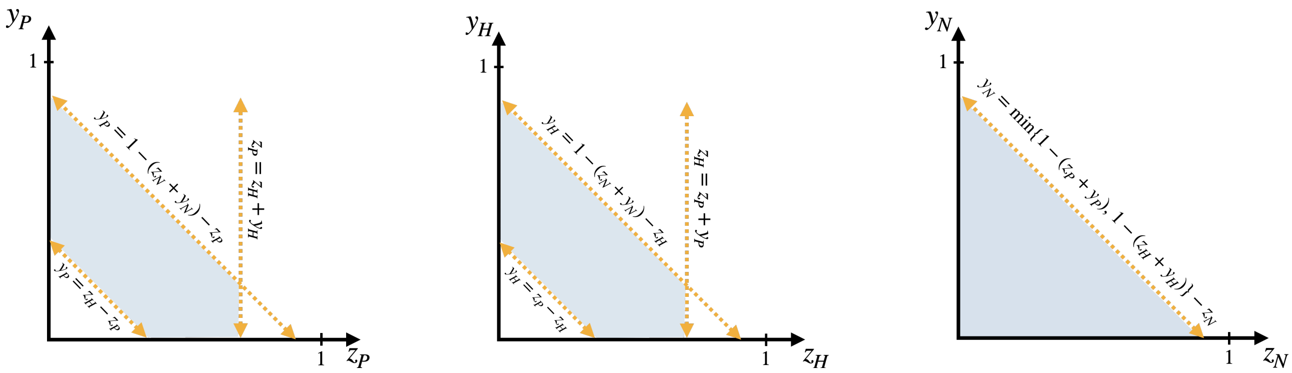
Sparsity-promoting ℓ_1 penalty. Since structural variants are rare in an individual's genome, a common challenge with SV recovery is predicting false positive SVs by mistaking fragments that are incorrectly mapped against the reference genome [18]. To model this, we incorporate an ℓ_1-norm penalty in 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. The penalty is:

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

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

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

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



(a) (b) (c)

Fig. 1. The feasible set is shown 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 (b) Step 2: We obtain the child's inherited indicator variables $\vec{z}_H$ and $\vec{y}_H$ by fixing $\vec{z}_P, \vec{y}_P, \vec{z}_N, \vec{y}_N$. (c) Step 3: We obtain the solution for the child's novel indicator variables $\vec{z}_N$ and $\vec{y}_N$ by fixing $\vec{z}_P, \vec{y}_P, \vec{z}_H, \vec{y}_H$.

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, $\alpha_k I$ [16]. 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

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

We reformulate this constrained quadratic subproblem into the following equivalent sequence of subproblems (see [16]):

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

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

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

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

$$\begin{aligned} \tilde{f}^{k+1} = &\min_{\substack{\tilde{f}=[z_P, z_H, z_N, \\ y_P, y_H, y_N]\in\mathbb{R}^6}} \frac{\tau}{\alpha_k}\Big\{|(\vec{z}_P)_j| + |(\vec{z}_H)_j| + \gamma|(\vec{z}_N)_j| \\ &+ |(\vec{y}_P)_j| + |(\vec{y}_H)_j| + \gamma|(\vec{y}_N)_j|\Big\} \\ &+ \frac{1}{2}\Big\{((\vec{z}_P - \vec{r}^k_{\vec{z}_P})_j)^2 + ((\vec{z}_H - \vec{r}^k_{\vec{z}_H})_j)^2 + ((\vec{z}_N - \vec{r}^k_{\vec{z}_N})_j)^2 \\ &+ ((\vec{y}_P - \vec{r}^k_{\vec{y}_P})_j)^2 + ((\vec{y}_H - \vec{r}^k_{\vec{y}_H})_j)^2 + ((\vec{y}_N - \vec{r}^k_{\vec{y}_N})_j)^2\Big\} \\ &\quad \text{subject to } \tilde{f} \in S \end{aligned} \qquad (8)$$

where z_i, y_i and r_{z_i}, r_{y_i} are scalar components of $\vec{z}_i$, $\vec{y}_i$ and $\vec{r}_{z_i}, \vec{r}_{y_i}$, respectively, at the same location; and S is the set of scalar constraints obtained from $\mathcal{S}$. Since Equation (8) has closed form solutions (obtained by completing the square and ignoring constant terms), the constrained minimizer is obtained by projecting the unconstrained solution to the feasible set.

C. Optimization Approach

We solve our problem using an alternating block-coordinate descent approach [18], [19], [20]. We fix all but one individual and solve Equation (7) over both indicator variables for that individual. We successively minimize 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 compute the unconstrained minimizer of Equation (7):

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

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

$$z_H = \text{mid}\{0, r^k_{z_H} - \frac{\tau}{\alpha_k}, 1\}, \quad z_N = \text{mid}\{0, r^k_{z_N} - \frac{\tau}{\alpha_k}\gamma, 1\},$$
$$y_H = \text{mid}\{0, r^k_{y_H} - \frac{\tau}{\alpha_k}, 1\}, \quad y_N = \text{mid}\{0, r^k_{y_N} - \frac{\tau}{\alpha_k}\gamma, 1\}$$

where mid$\{a, b, c\}$ chooses the middle value of the three arguments. Further, if $z_H + y_H > 1$, then we let $z_H = \hat{y}_H = 0.5$. We adjust z_N and y_N 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 $z_P = r^k_{z_P} - \frac{\tau}{\alpha_k}$ and $y_P = r^k_{y_P} - \frac{\tau}{\alpha_k}$.

Step 1: We project (z_P, y_P) onto the feasible set S with fixed inherited and novel variables to obtain the new parent indicator values $\hat{z}_P$ and $\hat{y}_P$.

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

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

We repeat Steps 1, 2, and 3 for every j to update $\vec{f}^{k+1}$ until the relative difference between consecutive iterates converges to $\|\vec{f}^{k+1} - \vec{f}^k\|/\|\vec{f}^k\| \leq 10^{-8}$.

III. Results

We introduce a novel algorithm named NEgative Binomial optimization Using ℓ_1 penalty Algorithm (NEBULA), developed through extensive modifications to the existing SPIRAL Poisson-based approach [16]. NEBULA incorporates the negative binomial statistical method for effectively solving the quadratic subproblems, marking a substantial advancement in our computational framework for predicting the presence of SVs in related individuals. Moreover, the flexibility and robustness of NEBULA render it generalizable to diverse applications beyond SV prediction [22]. 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 [18], [20]. In the parent signals, 5000 locations were chosen at random to be structural variants. While two parent signals were generated, only one parent signal was utilized for testing the method. The child signal was constructed by first applying a logical implementation of inheritance to $\lfloor 5000(1 - p) \rfloor$ randomly selected parent structural variants (where p is the percentage of novel SVs). Next, we randomly chose $\lfloor 5000p \rfloor$ 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 sequenced signals were generated by sampling from the Negative Binomial distribution with a given coverage and error. All code is available at `github.com/jornelasmunoz/structural_variants`.

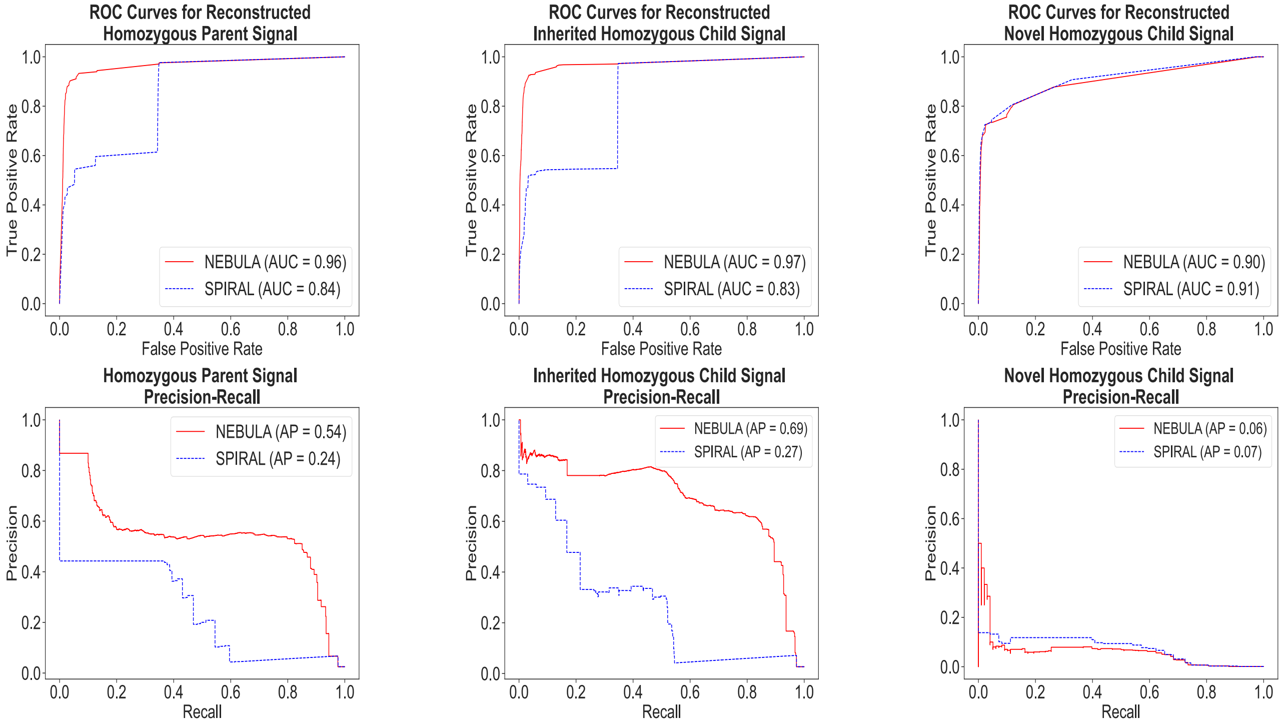


Fig. 2. ROC curves (top) and Precision-Recall curves (bottom) for the reconstructed homozygous parent signal (left), reconstructed inherited homozygous child signal (center), and reconstructed novel homozygous child signal (right) for our NEBULA algorithm (red) and the SPIRAL algorithm (blue). The regularization parameters used were $\tau = 1$, $\gamma = 2$, the percent of novel SVs is 4, and the coverage values for each individual are $(\lambda_P, \lambda_C) = (7, 3)$. The coverages were chosen based on low-coverage values explained in [21].

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. Our method demonstrates superior performance in reconstructing homozygous signals for each individual, even in the presence of significant sequencing and mapping errors ($\varepsilon = 0.5$). The Area Under the Curve (AUC) is employed to quantify the ability of both SPIRAL and NEBULA in distinguishing between classes. Since SVs are very rare, 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 [23]. We see improvements in AUC 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. However, neither method is able to accurately reconstruct the novel child signal. We hypothesize that including the information from both parents would enhance the ability to predict the child signal.

IV. 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. Our algorithm, called NEBULA, uses negative binomial statistical methods for optimization. This extends previous work (SPIRAL) 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. While the data we used in this proof-of-concept is synthetic, it is a natural extension to apply this method to real-world data such as the 1000 genomes project. In future studies, we intend to extend this work to a two parent and one child framework where we will be using real data.

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] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] Richard M. Durbin and et. al. Altshuler, David, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[22] Yu Lu and Roummel F. Marcia, "Negative binomial optimization for low-count overdispersed sparse signal reconstruction," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 1948–1952.

[23] 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.