Sparse Negative Binomial Signal Recovery for Genomic Variant Prediction in Diploid Species

Jocelyn Ornelas-Munoz

Capstone submitted in partial satisfaction of the requirements for the degree of Master of Science in Applied Mathematics

Applied Mathematics
University of California, Merced

Committee in charge: Dr. Erica M. Rutter
Dr. Roummel F. Marcia

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

Jocelyn Ornelas-Munoz

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

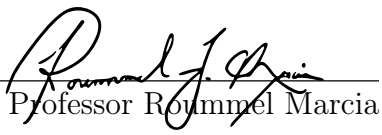
Applied Mathematics
Graduate Studies Chair:

Professor Maxime Theillard

Committee Co-Chair / Research Advisor:

Professor Erica Rutter

Committee Co-Chair / Research Advisor:



Professor Roummel Marcia

April 29, 2024

Date

Contents

1 Introduction

The genome, which is the complete sequence of DNA in an individual, is composed of a specific order of nucleotides (A, C, G, T). In the human genome, this sequence's total length amounts to approximately six billion letters [1]. Humans are diploid organisms, meaning they inherit two copies of the genome, with each parent contributing one copy. Inside a human organism, every cell carries a replica of the organism's genome, duplicated through the process of cell division. During DNA replication, alterations in the DNA sequence—referred to as genetic variants—can arise. While many of these variations have no significant impact, some changes can be detrimental and may be passed down through generations.

Structural variants (SVs) are a type of genetic variation characterized by insertions, deletions, inversions, and other rearrangements of DNA spanning over 50 letters (see Figure 1). These SVs are rare instances of DNA changes that offer valuable insights into gene expression regulation, ethnic diversity, extensive chromosome evolution, and their involvement in disease susceptibility [1, 2, 3, 4]. The primary approach for detecting structural variants (SVs) in an individual's genome involves sequencing their DNA, yielding numerous short DNA sequence reads. These reads are then aligned with a high-quality reference genome (see Figure 2). Any differences between the sequenced genome and the reference genome indicate SVs. The computational identification of SVs involves pinpointing clusters of reads exhibiting discordant arrangements [5, 6]. Despite advancements in DNA sequencing, methods for detecting SVs still suffer from high error rates during sequencing and mapping processes [5]. While one solution could be sequencing individuals at extremely high coverage, this approach incurs higher financial and computational costs. Therefore, our objective is to predict SV locations within a low-coverage framework. In the sequencing process, when genomic fragments are randomly chosen, the Poisson distribution describes the number of reads covering any genomic locus [7]. The Poisson assumption with a mean represented by the coverage also assumes the same variance. However, sequencing technologies introduce biases, leading to significant variation in coverage depth, especially in low-coverage scenarios [8, 9]. Studies suggest that in such settings, the two-parameter negative binomial distribution may more accurately describe the distribution of fragments [10, 11].

While the occurrence of newly arising (de novo) structural variants (SVs) is exceedingly rare [12], with the majority of SVs in a child being inherited from their parents, many computational SV analysis pipelines fail to leverage information from familial genomes [13, 14, 15, 16, 17].

In this study, we aim to bridge the gap in previous research by introducing a computational framework designed for predicting the presence of structural variants (SVs). We accomplish this by simultaneously analyzing diploid related individuals, with a specific focus on a parent and a child. Whereas previous work assumed mapped reads follow a Poisson distribution, we incorporate a negative binomial distribution to model the distribution of fragments. Instead of assuming equal mean and variance, we estimate both from the data and the negative binomial model captures large variability in the sequencing coverage. This allows us to forecast the most likely SVs within the genome of each individual. To refine our predictions, we constrain our predictions to only those SVs that conform to Mendelian inheritance patterns [18]. Additionally, we promote sparsity in our predictions by incorporating an ℓ_1 regularization penalty term, reflecting the biological reality that SVs are rare and reducing the risk of false positive predictions.

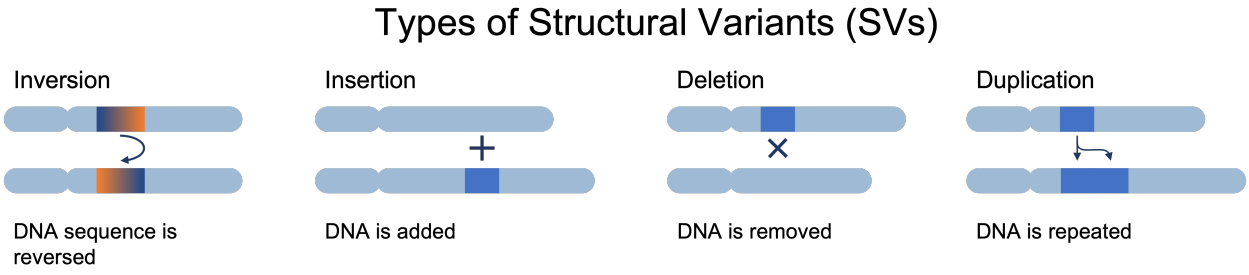


Figure 1: Types of structural variants.

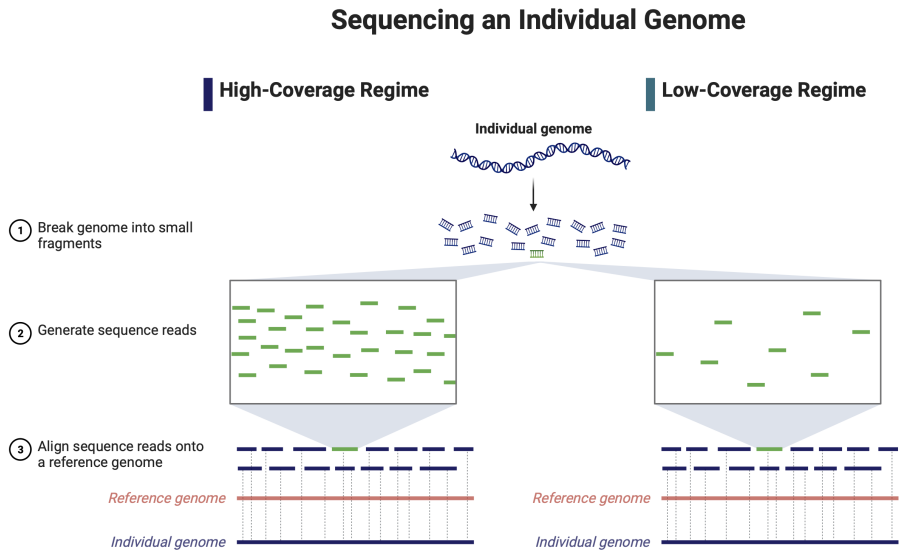


Figure 2: Sequencing an individual genome depicting a high-coverage and a low-coverage regime. In a high-coverage regime, millions of sequence reads are aligned onto reference genome while the low-coverage regime covers a mere fraction of the data per sample. Figure created with BioRender.

2 Proposed Method

Here, we describe our computational framework for predicting SVs for related individuals. We only use diploid data from one parent (P) and one child (C) for mathematical and computational simplicity. Each signal consists of n candidate locations in the genome where either 0, 1, or 2 copies of an SV may be present. We separate the signal from the child to consider both inherited (H) and novel (N) SVs individually. For this, we denote the true signal of the parent as $\vec{f}_P^* \in \{0,1,2\}^n$, and the true signal of the child as $\vec{f}_C^* = \vec{f}_H^* + \vec{f}_N^* \in \{0,1,2\}^n$, where $\vec{f}_H^* \in \{0,1,2\}^n$ and $\vec{f}_N^* \in \{0,1,2\}^n$ correspond to the vectors of inherited (H) and novel (N) structural variants in the child, respectively. For each $i \in \{P, H, N\}$ in our model, we consider two signals that take on binary values shown in Figure 3: a heterozygous indicator $\vec{y}_i \in \{0,1\}^n$ which signifies the presence of an SV in only one of the paired chromosomes at a specific location and a homozygous indicator $\vec{z}_i \in \{0,1\}^n$ the presence of an SV in both chromosomes at a particular location. Thus, if an individual is heterozygous for an SV at a position $j = 1, 2, \ldots, n$, 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 $(\vec{y}_i)_j = 0$ [19]. The true signal is then:

$$(\vec{f}_i^*)_j = 2(\vec{z}_i)_j + (\vec{y}_i)_j = \begin{cases} 2 \text{ presence of 2 copies of an SV at location } j \\ 1 \text{ presence of 0 copy of an SV at location } j \\ 0 \text{ otherwise} \end{cases}$$

2.1 Observational Model

Observation vectors for 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 [10]:

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

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 [19], [20]. 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 can then be written 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 $A = [A_z \; A_y] \in \mathbb{R}^{2n \times 6n}$ is the sequence coverage matrix with A_z, A_y as the block matrices:

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

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

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

2.2 Optimization Formulation

We assume a Negative Binomial process to model the noise in observed sequencing and mapping measurements. The negative binomial distribution is parameterized in terms of its mean $\vec{\mu}_l = \vec{e_l}^T A\vec{f}$ and variance $\vec{\sigma}_l^2 = (\vec{e_l}^T A\vec{f})_l + \frac{1}{r}(\vec{e_l}^T A\vec{f})_l^2$, $l = 1, \ldots, 2n$, where $\vec{e_l}$ represents the canonical standard basis vectors. When $r \to \infty$, we have $\sigma^2 = \mu$ and this reduces the negative binomial model to the Poisson case. Thus, we assume $r \in \mathbb{Z}^+$ and we set the dispersion parameter $r = 1$ to maximize the variance. With these considerations, the probability of observing the observation vector $\vec{s}$ given the true signal $\vec{f}$, is given by

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

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

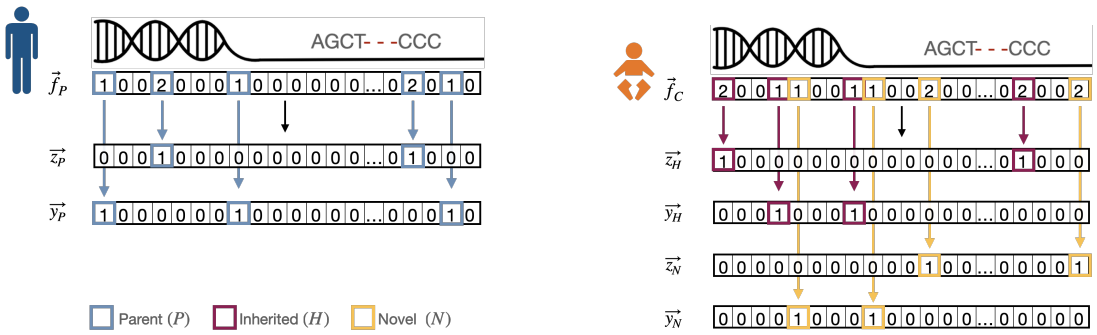


Figure 3: Visual representation of indicator vectors for the parent (left) and the child (right). Arrows illustrate structural variants (SVs), distinguishing between those inherited from the parent and those that are novel, as they are mapped onto the indicator vectors.

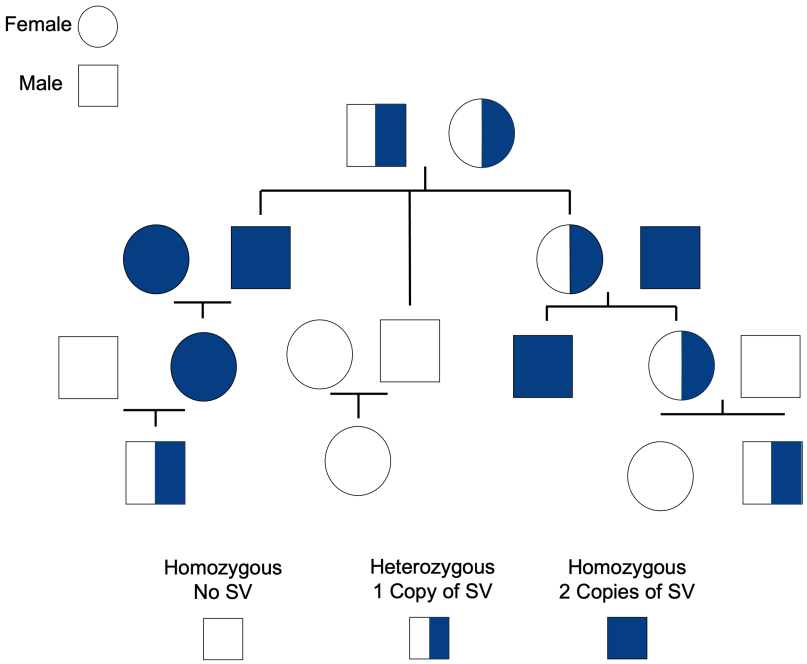


Figure 4: Family tree describing inheritance patterns of SVs. For example, if a heterozygous woman and a homozygous man have children, some of them will have 2 copies of SVs and some will only have 1 copy of an SV.

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

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

For the ease of notation, 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. In particular, we seek to minimize the corresponding Negative Binomial negative log-likelihood function

$$F(\vec{f}) \equiv \sum_{l=1}^{2n} (1 + \vec{s}_l) \log\left(1 + \vec{e}_l^T A\vec{f} + \varepsilon\right) - \vec{s}_l \log\left(\vec{e}_l^T A\vec{f} + \varepsilon\right) \tag{3}$$

Familial Constraints. We incorporate additional constraints based on Mendelian inheritance patterns to leverage biological information about $\vec{f}$ and improve accuracy of the model. These inheritance patterns are illustrated in Figure 4. 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. 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:

$$\begin{aligned}\mathbf{0} \leq \vec{z}_H &\leq \vec{z}_P + \vec{y}_P \leq \mathbf{1} \\ \mathbf{0} \leq \vec{z}_P &\leq \vec{z}_H + \vec{y}_H \leq \mathbf{1}\end{aligned}$$

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 $(\vec{z}_N)_j + (\vec{y}_N)_j = 0$,

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

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

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

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 [19]. To model this, we incorporate multiple ℓ_1-norm penalties 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 follows:

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

where $\gamma > 1$ is the penalty term that enforces greater sparsity in the child's novel SVs.

Our objective function then takes the form:

$$\begin{array}{ll} \underset{\vec{f} \in \mathbb{R}^{6n}}{\text{minimize}} & F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \tag{4}$$

where $F(\vec{f})$ is the Negative Binomial negative log-likelihood function shown in Equation (3) and $\tau > 0$ is a regularization parameter. Following the SPIRAL framework for sparse Poisson reconstruction [21], we solve Equation (4) by minimizing 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, $\alpha_k I$ [21]. 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 (4). Using this approximation, the next iterate is given by

$$\vec{f}^{k+1} = \begin{array}{ll} \underset{\vec{f} \in \mathbb{R}^{6n}}{\arg\min} & F^k(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \tag{5}$$

We reformulate this constrained quadratic subproblem into the following equivalent sequence of subproblems (see [21]):

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

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

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

Since the bounds defining the region $\mathcal{S}$ are component-wise, then Equation (6) separates into subproblems of the form:

$$\begin{aligned}\vec{f}^{k+1} = \underset{z_P,z_H,z_N,y_P,y_H,y_N\in\mathbb{R}}{\arg\min} &\frac{1}{2}\Big\{((\vec{z}_P - \vec{r}^k_{\vec{z}_P})_j)^2 + ((\vec{z}_H - \vec{r}^k_{\vec{z}_H})_j)^2 + ((\vec{z}_N - \vec{r}^k_{\vec{z}_H})_j)^2 \\ &+ ((\vec{y}_P - \vec{r}^k_{\vec{y}_P})_j)^2 + ((\vec{y}_H - \vec{r}^k_{\vec{y}_H})_j)^2 + ((\vec{y}_N - \vec{r}^k_{\vec{y}_H})_j)^2\Big\} \\ &+ \frac{\tau}{\alpha_k}\Big\{|(\vec{z}_P)_j|+|(\vec{z}_H)_j|+\gamma|(\vec{z}_N)_j|+|(\vec{y}_P)_j|+|(\vec{y}_H)_j|+\gamma|(\vec{y}_N)_j|\Big\} \\ &\text{subject to} \vec{f} \in S\end{aligned} \tag{7}$$

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 (7) 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.

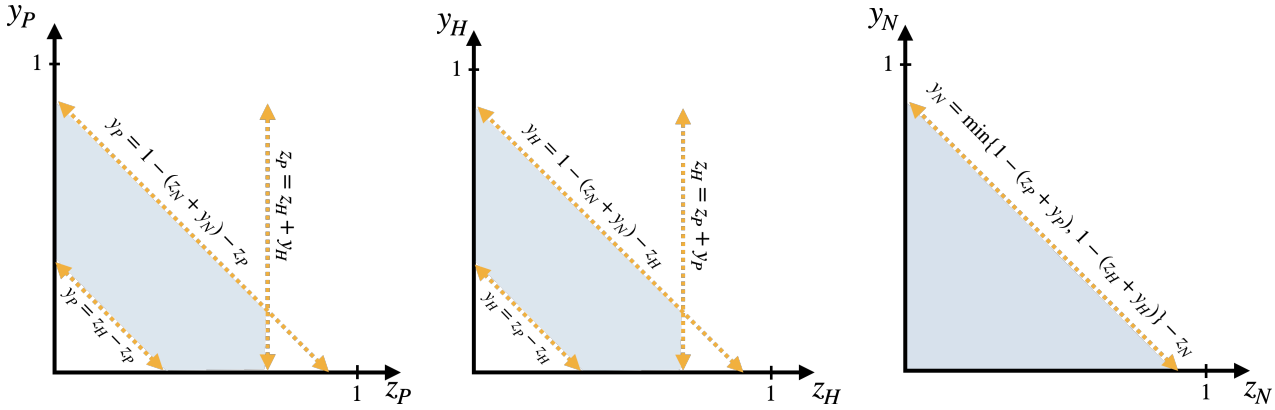


Figure 5: The feasible set is shown by the shaded region for each step of the proposed block-coordinate descent approach. (a) Step 1: We obtain the solution for the parent's variables $\vec{z}_P$ and $\vec{y}_P$ given fixed child inherited and novel indicator variables (b) Step 2: We obtain the child's 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 $\vec{z}_P, \vec{y}_P, \vec{z}_H, \vec{y}_H$.

2.3 Optimization Approach

We solve our problem using an alternating block-coordinate descent approach inspired by the methods in [19], [20], [22]. We fix all but one pair of indicator variables and solve Equation (6). We successively minimize both indicator variables for each P, H, N while the other variables are fixed. The feasible region for this step

is illustrated in Figure 5 and the optimization approach is shown in Figure 6.

Step 0: We compute the unconstrained minimizer of Equation (6):

$$\vec{f} = \left[\vec{r}_{z_P} - \frac{\tau}{\alpha_k}\mathbf{1}_n;\ \vec{r}_{z_H} - \frac{\tau}{\alpha_k}\mathbf{1}_n;\ \vec{r}_{z_N} - \frac{\tau}{\alpha_k}\gamma\mathbf{1}_n;\ \vec{r}_{y_P} - \frac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{y_H} - \frac{\tau}{\alpha_k}\mathbf{1}_n,\ \vec{r}_{y_N} - \frac{\tau}{\alpha_k}\gamma\mathbf{1}_n\right]^T$$

where $\mathbf{1}_n \in \mathbb{R}^n$.

Next, steps 1-3 are done for every j in $\vec{z}_P, \vec{z}_H, \vec{z}_N, \vec{y}_P, \vec{y}_H, \vec{y}_N$. Here, again, z_i, y_i correspond to scalar components of $\vec{z}_i$ and $\vec{y}_i$ at the same location. First, 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 $\text{mid}\{a, b, c\}$ chooses the middle value of the three arguments to stay consistent with the constraints in $\mathcal{S}$. To initialize the parent indicator variables, we let $z_P = r^k_{z_P} - \frac{\tau}{\alpha_k}$ and $y_P = r^k_{y_P} - \frac{\tau}{\alpha_k}$, the values obtained from the unconstrained minimizer.

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 $\hat{z}_P$ and $\hat{y}_P$, 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 $\hat{z}_H$ and $\hat{y}_H$ 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 $\|\vec{f}^{k+1} - \vec{f}^k\|/\|\vec{f}^k\| \leq 10^{-8}$.

3 Results

We modified the existing SPIRAL approach [21] to include the negative binomial statistical method for solving the quadratic sub-problems. The implementation is done in MATLAB. We refer to the new algorithm as NEgative Binomial optimization Using ℓ_1-penalty Algorithm (NEBULA). 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 [19, 22]. 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 $\lfloor 5000(1-p) \rfloor$ randomly selected parent structural variants (where p is the percentage of novel SVs). Next, we 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 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. The data simulation code was implemented in Python. All code is available on GitHub[1].

[1] `https://github.com/jornelasmunoz/structural_variants`

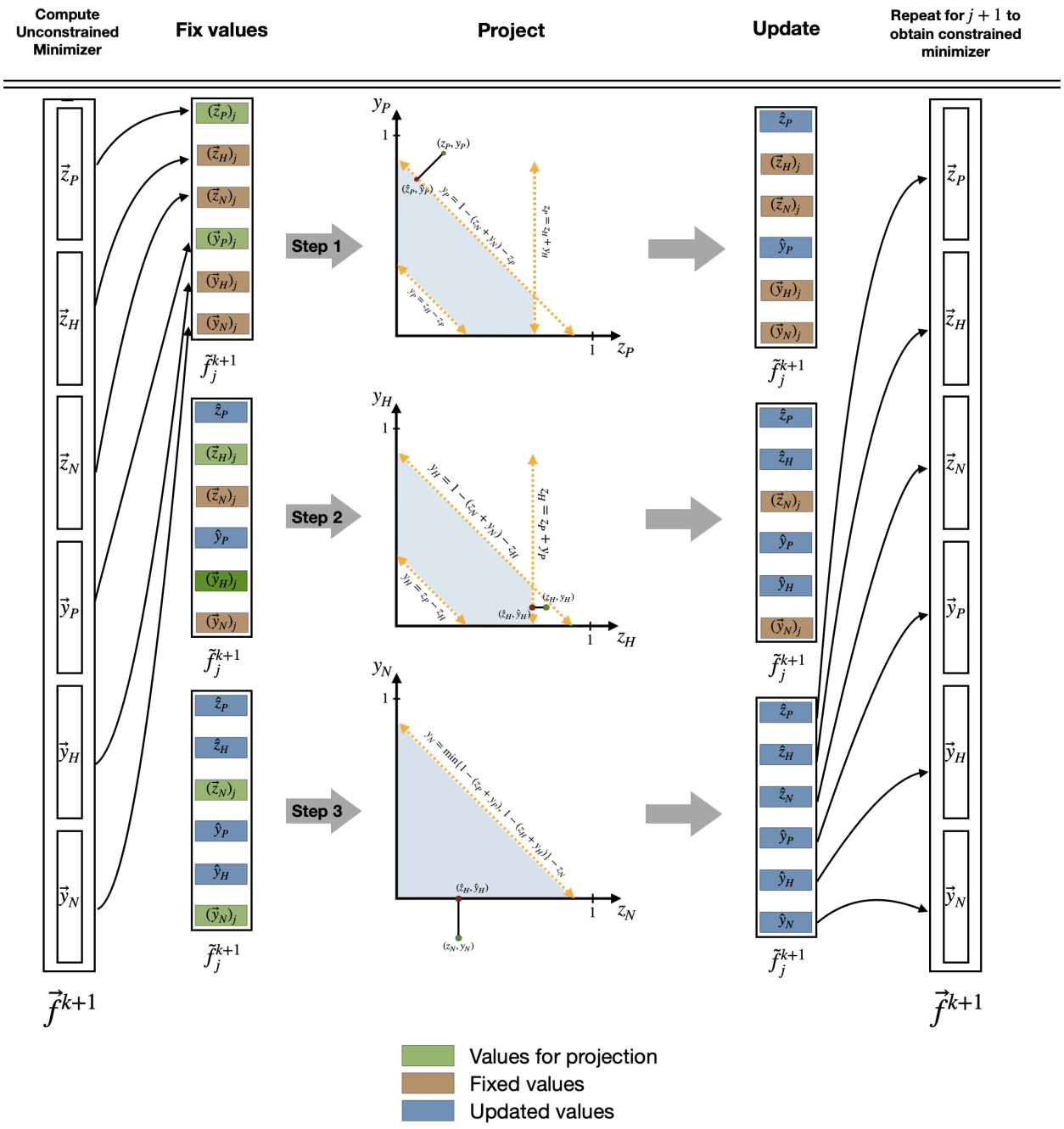


Figure 6: Optimization approach explained in Section 2.3. The green values represent the indicator variables used for projection while the brown represent fixed values and the blue represent updated values. Step 1 is shown in the top panel, Step 2 in the middle, and Step 3 in the bottom.

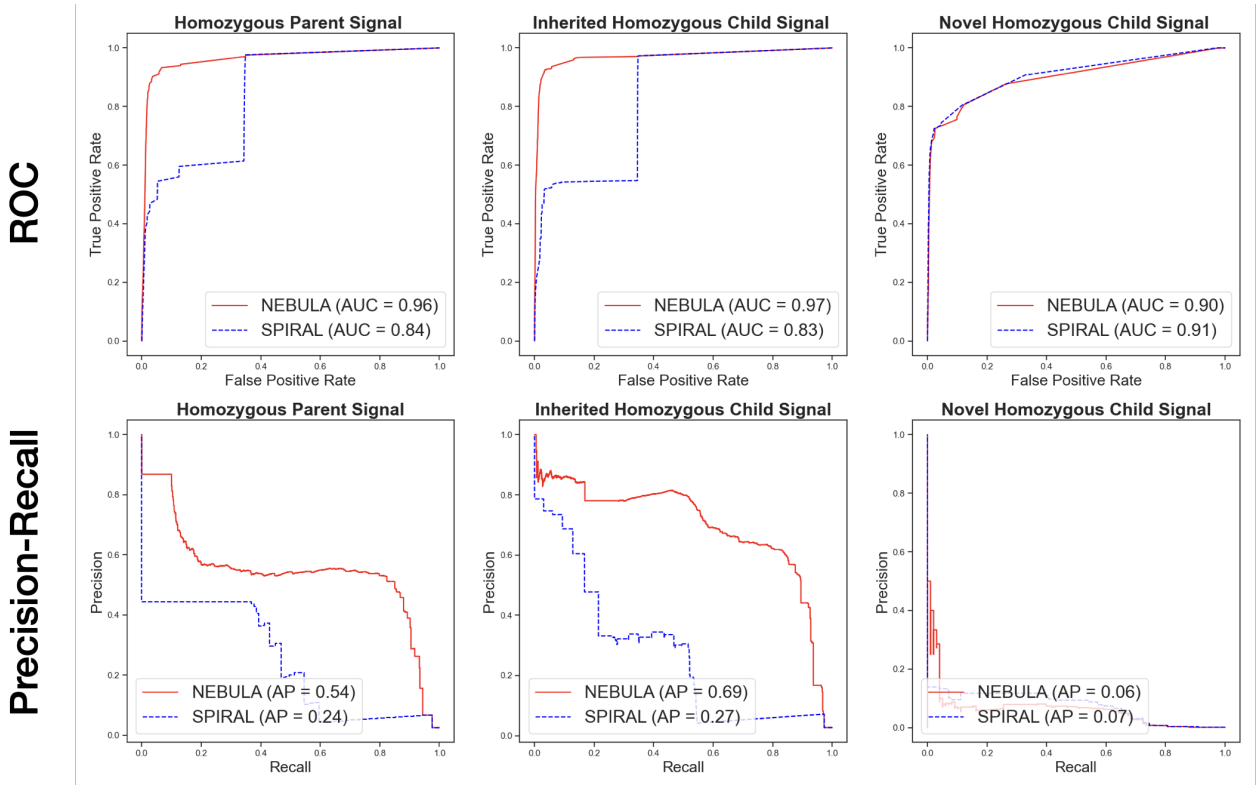


Figure 7: ROC curves (top) and Precision-Recall curves (bottom) for the reconstructed homozygous parent signal (left), reconstructed inherited homozygous child signal (center), and reconstructed novel homozygous child signal (right) for our NEBULA algorithm (red) and the SPIRAL algorithm (blue). The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

Analysis. Figure 7 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 is better able to reconstruct the homozygous signals for each individual despite large sequencing and mapping error, $\varepsilon = 0.5$. We use the AUC to measure the ability of SPIRAL and NEBULA to distinguish 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 note that as this work only considers the relationship between one parent and one child. This result is similar to the heterozygous results. 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 extend this work to a two parent and one child framework.

4.1 Acknowledgements

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

5 Appendix

5.1 Structural Variant Detection Notation

Notation	Description
P	Subscript denoting parent
C	Subscript denoting child
H	Subscript denoting inherited
N	Subscript denoting novel
$\vec{f}_P^*$	$\in \{0,1,2\}^n$ Vector representing true parent signal
$\vec{f}_C^*$	$\in \{0,1,2\}^n$ Vector representing true child signal
$\vec{f}_H^*$	$\in \{0,1,2\}^n$ Vector representing true child *inherited* signal
$\vec{f}_N^*$	$\in \{0,1,2\}^n$ Vector representing true child *novel* signal
$\vec{z}_i$	$\in \{0,1\}^n$ Homozygous indicator vector (i.e. 2 copies)
$\vec{y}_i$	$\in \{0,1\}^n$ Heterozygous indicator vector (i.e. 1 copy)
$\vec{s}_P$	$\in \mathbb{R}^n$ Observation vector for parent
$\vec{s}_C$	$\in \mathbb{R}^n$ Observation vector for child
$\vec{s}$	$= [\vec{s}_P; \vec{s}_C]^T$ Vector of all observation vectors
$\vec{z}$	$= [\vec{z}_P; \vec{z}_H; \vec{z}_N]^T$ Vector of all homozygous indicator vectors
$\vec{y}$	$= [\vec{y}_P; \vec{y}_H; \vec{y}_N]^T$ Vector of all heterozygous indicator vectors
$\vec{f}$	$= [\vec{z}_P; \vec{z}_H; \vec{z}_N; \vec{y}_P; \vec{y}_H; \vec{y}_N]^T$ Vector of all indicator vectors
A	$= [A_1 A_2] \in \mathbb{R}^{2n \times 6n}$ Sequence coverage matrix
λ_P	sequencing coverage for parent
λ_C	sequencing coverage for child
ε	Measurement errors incurred in sequencing and mapping
$\mathbf{1}$	$\in \mathbb{R}^{2n}$ Vector of ones
$\mathbf{0}$	$\in \mathbb{R}^{2n}$ Vector of zeros
I_n	$\in \mathbb{R}^{n \times n}$ Identity matrix
$\vec{\mu}$	$\in \mathbb{R}^{2n}$ Mean vector of Neg. Binom. distribution
$\vec{\sigma}^2$	$\in \mathbb{R}^{2n}$ Standard deviation vector of Neg. Binom. distribution
$\vec{e}_l$	$\in \mathbb{R}^{2n}$ Canonical basis vector with 1 in l-th position
$F(\vec{f})$	Negative binomial negative log-likelihood function
$F^k(\vec{f})$	Second-order Taylor series approximation of F at $\vec{f}^k$
$\mathcal{Q}(\vec{f})$	Reformulation of objective function. Separable function
$\mathcal{S}$	Set of all vectors satisfying familial constraints
S	Set of scalar constraints obtained from $\mathcal{S}$
$\text{pen}(\vec{f})$	Sparsity-promoting ℓ_1 penalty
τ	> 0 Regularization parameter
γ	> 1 Regularization parameter for novel SVs
α_k	Scalar used to approximate Hessian
$\vec{r}$	$\vec{f}^k - \frac{1}{\alpha_k} \nabla F(\vec{f}^k)$
z_i	Scalar component of $\vec{z}_i$ at a particular location
y_i	Scalar component of $\vec{y}_i$ at a particular location
r_{zi}	Scalar component of $\vec{r}_{zi}$ at a particular location
r_{yi}	Scalar component of $\vec{r}_{yi}$ at a particular location
p	percentage of novel SVs
r	Dispersion parameter of Neg. Binom.
n	Number of SV candidate locations.
i	$\in \{P, H, N\}$ index for parent, inherited, novel
j	$= 1, 2, \ldots, n$ index for vector position
k	Iteration number
l	$= 1, 2, \ldots, 2n$ index for observed vector position

References

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

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

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

[4] S. Kosugi, Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani, "Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing," *Genome Biology*, vol. 20, no. 1, pp. 1–18, 2019.

[5] S. S. Sindi, S. Önal, L. C. Peng, H.-T. Wu, and B. J. Raphael, "An integrative probabilistic model for identification of structural variation in sequencing data," *Genome biology*, vol. 13, pp. 1–25, 2012.

[6] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing," *Nature methods*, vol. 6, no. Suppl 11, pp. S13–S20, 2009.

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

[8] D. Iakovishina, I. Janoueix-Lerosey, E. Barillot, M. Regnier, and V. Boeva, "Sv-bay: structural variant detection in cancer genomes using a bayesian approach with correction for gc-content and read mappability," *Bioinformatics*, vol. 32, no. 7, pp. 984–992, 2016.

[9] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome research*, vol. 19, no. 9, pp. 1586–1592, 2009.

[10] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, "Sequencing depth and coverage: key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 121–132, 2014.

[11] J. Sampson, K. Jacobs, M. Yeager, S. Chanock, and N. Chatterjee, "Efficient study design for next generation sequencing," *Genetic epidemiology*, vol. 35, no. 4, pp. 269–277, 2011.

[12] T. G. 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.

[13] 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.

[14] 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.

[15] 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.

[16] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall, "Lumpy: a probabilistic framework for structural variant discovery," *Genome Biology*, vol. 15, no. 6, pp. 1–19, 2014.

[17] K. Uguen, C. Jubin, Y. Duffourd, C. Bardel, V. Malan, J.-m. Dupont, L. El Khattabi, N. Chatron, A. Vitobello, P.-A. 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, p. e1114, 2020.

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

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

[20] A. Lazar, M. Banuelos, S. Sindi, and R. F. Marcia, "Novel structural variant genome detection in extended pedigrees through negative binomial optimization," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 563–567, 2021.

[21] Z. T. Harmany, R. F. Marcia, and R. 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.

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

[23] T. Saito and M. 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, p. e0118432, 2015.