2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA) | 978-1-6654-9384-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/MeMeA57477.2023.10171853

GENETIC VARIANT DETECTION OVER GENERATIONS: SPARSITY-CONSTRAINED OPTIMIZATION USING BLOCK-COORDINATE DESCENT

Mohammed Aburidi[⋆] *Mario Banuelos*[†] *Suzanne Sindi*[⋆] *Roummel Marcia*[⋆]

[⋆] University of California Merced, Merced , CA, USA

[†] California State University Fresno, Fresno , CA, USA

ABSTRACT

Structural variants (SVs) are rearrangements of regions in an individual's genome signal. SVs are an important source of genetic diversity and disease in humans and other mammalian species. The SV detection process is susceptible to sequencing and mapping errors, especially when the average number of reads supporting each variant is low (i.e. low-coverage settings), which leads to high false-positive rates. Besides their rarity in the human genome, they are shared between related individuals. Thus, it's advantageous to devise algorithms that focus on close relatives. In this paper, we develop a constrained-optimization method to detect germline SVs in genetic signals by considering multiple related people. First, we exploit familial relationships by considering a biologically realistic scenario of three generations of related individuals (a grandparent, a parent, and a child). Second, we pose the problem as a constrained optimization problem regularized by a sparsity-promoting penalty. Our framework demonstrates improvements in predicting SVs in related individuals and uncovering true SVs from false positives on both simulated and real genetic signals from the 1000 Genomes Project with low coverage. Further, our block-coordinate descent approach produces results with equal accuracy to the 3D projections of the solution, demonstrating feasibility for more complex and higher-dimensional pedigrees.

Index Terms— Structural variants, nonconvex optimization, next- generation sequencing data, genetic signals.

1. INTRODUCTION

Structural variants (SVs) are a common type of genomics variation and it appears in a form of rearrangements of some of the DNA regions. SVs include inversions, translocations, insertions, and/or deletions [1]. Although SVs are rare, they have increasing importance; SV has a substantial impact on gene expression and it resulting in altered phenotypes and disease. In humans, they have been associated with the development of some hereditary diseases such as cancer [2]. Therefore, detecting SVs is beneficial for diagnostics and understanding cancer etiology.

The advent of next-generation sequencing (NGS) makes detecting variation amenable because of its high-throughput, low cost, and base-pair resolution [3], [4]. However, even with those modern DNA sequencing technologies, the SV detection process is still complicated, due to alignment errors - a step where DNA fragments from a candidate genome are mapped to high-quality reference genome. Due to this, most detection methods and algorithms have suffered from higher false-positive rates, especially when the sequencing coverage – the average number of reads supporting each variant – is low [5]. Therefore, more robust algorithms need to be developed. An important fact about SVs, besides their rarity, they are shared between related individuals as with other DNA markers of interest. Thus, it's advantageous to devise algorithms that focus on the close relatives [6]. For example, most of the SVs in a child's genome will be present in one of their parents.

In this study, we present an optimization framework for SV detection in the context of one grandparent, one parent, and one child. For simplicity, we consider simultaneous SV prediction in haploid genomes. We use a novel projection algorithm to enforce biological feasibility (allowable heritable patterns of SVs) while minimizing our objective over a 3-dimensional solution space. We promote rarity of SVs by enforcing sparsity in each signal using the l_1 norm. We validate our method using simulated and real data from 1000 Genomes Project. Our method improves the SV prediction problem for low-coverage individuals. Further, our block-coordinate descent approach produces results with equal accuracy to the 3D projections of the solution, demonstrating feasibility for more complex and high-dimensional pedigrees.

Related work: One of the earliest methods that used NGS data to detect SVs was [7], where they used statistical testing methodologies. Followed by the well-known readDepth algorithm that employed LOESS regression to predict SVs [8]. Whereas the CNVer algorithm used maximum-likelihood and graphic flow to detect variants in pair-end sequences [9]. Other methods based on Hidden Markov Models (HMMs) were developed for the same purpose [10], [11]. However,

This work is partly supported by National Science Foundation grants DMS 1840265 and IIS 1741490.

non of these approaches have considered familial relationships. Our group has developed optimization methods to improve SV prediction by considering pedigrees of related individual [12], [13]. However, all of the prior work was done on one generation, predicting the SVs on genome parent-child trois. In this work, we improve upon our previous methods by considering two generations of individuals.

2. METHOD

Here, we describe mathematically our computational framework for detecting SVs given sequencing data from one grandparent (gp), one parent (p), and one child (c). For simplicity, we assume the haploid case (one-copy per chromosome). As such, the true SV signal $\vec{f}_I^* \in \{0,1\}^m$ is a binary vector indicating the presence of a genetic variant for each individual $I \in \{gp, p, c\}$. Thus, the corresponding grandparent $\vec{y}_{gp} \in \mathbb{R}^m$, parent $\vec{y}_p \in \mathbb{R}^m$ and child $\vec{y}_c \in \mathbb{R}^m$ observations. Each element in the observation vector represents the number of DNA fragments (i.e. sequencing coverage in real observations) supporting each potential SV. As in previous work [13], To consider the most realistic case in which the number of fragments covering any position (sequencing coverage) in the genome signal is low, this number should follow a Poisson distribution, thus, the general observation model can be expressed as follows:

$$\begin{bmatrix} (\vec{y}_c)_j \\ (\vec{y}_p)_j \\ (\vec{y}_{gp})_j \end{bmatrix} \sim \text{Poisson}\left(\begin{bmatrix} (\lambda_c - \epsilon)(\vec{f}_c^*)_j + \epsilon \\ (\lambda_p - \epsilon)(\vec{f}_p^*)_j + \epsilon \\ (\lambda_{gp} - \epsilon)(\vec{f}_{gp}^*)_j + \epsilon \end{bmatrix} \right) \quad (1)$$

where $j \in \{1, 2, ..., m\}$. The constants λ_{gp}, λ_p and λ_c represent the sequencing coverage of the grandparent, parent and child genome, respectively, and $\epsilon > 0$ is the error term in the measurement of the true signals corresponding to the sequencing processing, which is assumed to be the same for each observation. In matrix notation, we can express the general observation model as $\vec{y} \sim \text{Poisson}(A\vec{f}^* + \epsilon\mathbf{1})$, where $\vec{y} = \left[\vec{y}_c;\ \vec{y}_p^T;\ \vec{y}_{gp}\right]$, $\vec{f}^* = [\vec{f}_c^*;\ \vec{f}_p^*;\ \vec{f}_{gp}^*]$, $\mathbf{1} \in \mathbb{R}^{3m}$ is the vector of ones, and $A \in \mathbb{R}^{3m \times 3m}$ is the coverage matrix given by

$$A = \begin{bmatrix} (\lambda_c - \epsilon)I_m & 0 & 0 \\ 0 & (\lambda_p - \epsilon)I_m & 0 \\ 0 & 0 & (\lambda_{gp} - \epsilon)I_m \end{bmatrix}$$

where $I_m \in \mathbb{R}^{m \times m}$ is the identity matrix.

2.1. Problem formulation

Under the Poisson process model [13], the probability of observing the observation vector $\vec{y}$ given the true signal $\vec{f}$, is given by

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

To determine the unknown Poisson parameter $A\vec{f}^*$, we use the maximum likelihood principle such that the probability of observing the vector of Poisson data $\vec{y}$ in (2) is maximized. This is equivalent to minimizing the corresponding negative Poisson log-likelihood function

$$F(\vec{f}) = \boldsymbol{I}^T A\vec{f} - \sum_{j=1}^{3m} y_i \log((A\vec{f})_j + \epsilon).$$

In our approach for minimizing $F(f)$, we apply gradient-based optimization approaches and apply a continuous relaxation by allowing $\vec{f}$ to lie between 0 and 1, i.e., $0 \leq \vec{f} \leq 1$.

2.2. Familial constraints

To improve the accuracy of our SV predictions, we impose additional constraints that exploit information about the SV signal $\vec{f}$. This constraint corresponds to the biological assumptions we make, which is as follows:

$$0 \leq \vec{f}_c \leq \vec{f}_p \leq \vec{f}_{gp} \leq 1$$

This constraint exploits the biological assumption that variants in the child can be present only when the parent also has that SV and an SV in a parent can be present only when the grandparent also has that SV. Thus, an SV cannot be present in the child if neither grandparent nor parent has the SV.

2.3. Sparsity

Due to the rareness of the SVs in an individual's genome, predictions result in false positives that mistake fragments that are incorrectly mapped to locations in the genome as SVs. To avoid that, we incorporate how uncommon these SVs are in a genome sequence. To promote sparsity in our predictions we use a common technique found in literature [14], by which we incorporate an l_1-norm penalty term in our problem formulation. We use three penalty terms, one for the child SV ($\vec{f}_c$), one for the parent SV ($\vec{f}_p$) and one for the grandparent SV ($\vec{f}_{gp}$). What is particularly novel in our formulation is that while SVs are rare in grandparents, they should be rarer in parents and even more rare in children. Mathematically, we express this penalty as

$$\text{pen}(\vec{f}) = \|\vec{f}_{gp}\|_1 + \beta(\|\vec{f}_p\|_1 + \beta\|\vec{f}_c\|_1)$$

where $\beta \geq 1$ is a penalty weight that places greater emphasis on $\vec{f}_c$ being much sparser than both $\vec{f}_p$ and $\vec{f}_{gp}$.

2.4. Optimization Setup

With these components defined, our objective function takes the following constrained optimization form:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{3m}}{\text{minimize}} \quad & F(\vec{f}) + \tau\text{pen}(\vec{f}) \\ \text{subject to} \quad & 0 \leq \vec{f}_c \leq \vec{f}_p \leq \vec{f}_{gp} \leq 1 \end{aligned} \quad (3)$$

where $\tau > 0$ is a regularization parameter that balances the negative Poisson log-likelihood data fidelity term with the sparsity-promoting penalty term. We use the Sparse Poisson Intensity Reconstruction ALgorithm (SPIRAL) framework [12],[14], which is an iterative method that uses a second-order Taylor series approximation of $F(\vec{f})$ around the current iterate, to formulate a sequence of quadratic subproblems. In this approach, the Hessian matrix is approximated by a scalar multiple of the identity matrix, $\alpha_k I$, where $\alpha_k > 0$, which yields the following quadratic function:

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

This approximation leads to a sequence of quadratic subproblems of the following form:

$$\begin{aligned} \vec{f}^{k+1} \;=\; & \underset{\vec{f}\in\mathbb{R}^{3m}}{\arg\min} \quad \mathcal{Q}(\vec{f}) = \frac{1}{2}\|\vec{f} - \vec{s}^k\|_2^2 + \frac{\tau}{\alpha_k}\text{pen}(\vec{f}) \\ & \text{subject to} \quad 0 \le \vec{f}_c \le \vec{f}_p \le \vec{f}_{gp} \le 1 \end{aligned} \tag{4}$$

where $\vec{s}^k = [\vec{s}_c^k;\ \vec{s}_p^k;\ \vec{s}_{gp}^k] = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$.
Now, our objective function $\mathcal{Q}(\vec{f})$ in (4) is separable and decouples into the function

$$\begin{aligned} & \mathcal{Q}(\vec{f}) \\ & = \tfrac{1}{2}\sum_{j=1}^{m}\left\{\left((\vec{f}_c - \vec{s}_c^k)_j\right)^2 + \left((\vec{f}_p - \vec{s}_p^k)_j\right)^2 + \left((\vec{f}_{gp} - \vec{s}_{gp}^k)_j\right)^2\right\} \\ & \quad + \tfrac{\tau}{\alpha_k}\left\{\beta^2|(\vec{f}_c)_j| + \beta|(\vec{f}_p)_j| + |(\vec{f}_{gp})_j|\right\}. \end{aligned}$$

See [12] for more details. Since the bounds that define the feasible region are component wise, then (4) separates into subproblems of the form

$$\begin{aligned} \underset{f_c,f_p,f_{gp}\in\mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_c - s_c)^2 + \tfrac{1}{2}(f_p - s_p)^2 + \tfrac{1}{2}(f_{gp} - s_{gp})^2 \\ & + \tfrac{\beta^2\tau}{\alpha_k}|f_c| \;+\; \tfrac{\beta\tau}{\alpha_k}|f_p| \;+\; \tfrac{\tau}{\alpha_k}|f_{gp}| \\ \text{subject to} \quad & 0 \le f_c \le f_p \le f_{gp} \le 1 \end{aligned} \tag{5}$$

where $\{f_c, f_p, f_{gp}\}$ and $\{s_c, s_p, s_{gp}\}$ are scalar components of the vectors $\{\vec{f}_c, \vec{f}_p, \vec{f}_{gp}\}$ and $\{\vec{s}_c, \vec{s}_p, \vec{s}_{gp}\}$, respectively, at the same location. Since the variables are non-negative, we ignore the absolute values. By completing the squares and ignoring the constant terms, the optimization problem (5) yields to

$$\begin{aligned} \underset{f_i,f_p,f_{gp}\in\mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_c - a)^2 + \tfrac{1}{2}(f_p - b)^2 + \tfrac{1}{2}(f_{gp} - c)^2 \\ \text{subject to} \quad & 0 \le f_c \le f_p \le f_{gp} \le 1 \end{aligned} \tag{6}$$

where $a = s_c - \frac{\beta^2\tau}{\alpha_k}$, $b = s_p - \frac{\beta\tau}{\alpha_k}$ and $c = s_{gp} - \frac{\tau}{\alpha_k}$. The unconstrained minimizer of (6) is (a, b, c). If (a, b, c) satisfies the the constraints, then it is also the constrained minimizer. If not, we obtain the feasible solution to (6) by orthogonally projecting (a, b, c) onto the three-dimensional feasible set, which is shown in Fig. 1.

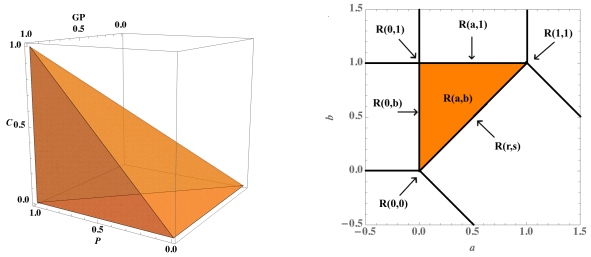


Fig. 1. Left: The three-dimensional feasible region of the minimization problem (6) on the f_c-f_p-f_{gp} axis. Subproblem minimizers not satisfying the constraints are projected onto the region Δ. Right: Plot of the a-b plane, where regions are defined in Table 1. $R_{(a,b)}$ represents the feasible region when fixing c.

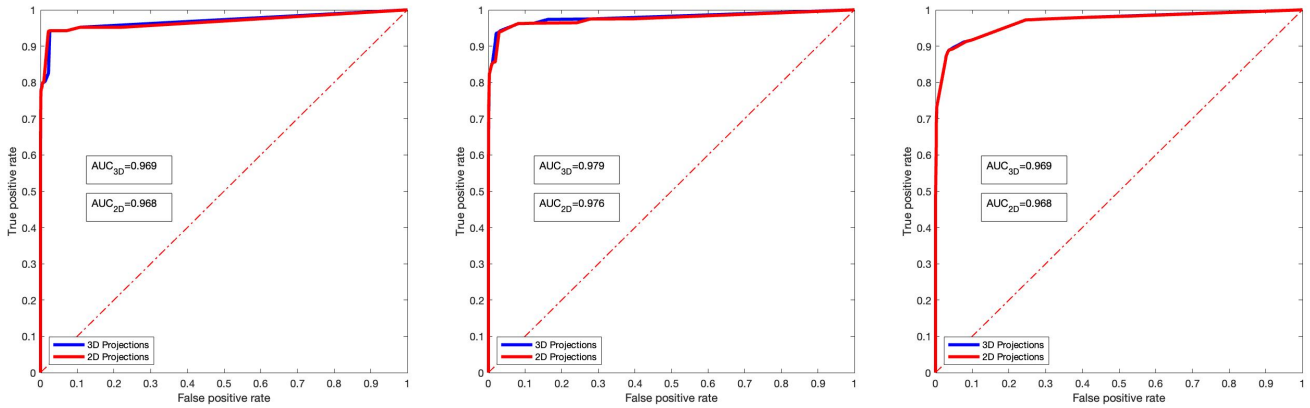
Table 1. Table representing the optimization problem as a function of a and b. Here, $r = s = (c + p)/2$

Region	Condition on a	Condition on b	(f_c^*, f_b^*)
$R(a,b)$	$0 < a < b$	$0 < a < 1$	(a,b)
$R(0,b)$	$a < 0$	$0 < b < 1$	$(0,b)$
$R(a,1)$	$0 \le a \le 1$	$b > 1$	$(b,1)$
$R(0,1)$	$a < 0$	$b > 1$	$(0,1)$
$R(0,0)$	$a \le -b$	$b < 0$	$(0,0)$
$R(1,1)$	$a > 1$	$b \ge -a + 2$	$(1,1)$
$R(r,s)$	$a > \lvert b\rvert$	$b < -a + 2$	(r,s)

2.5. Computational Projection Approach

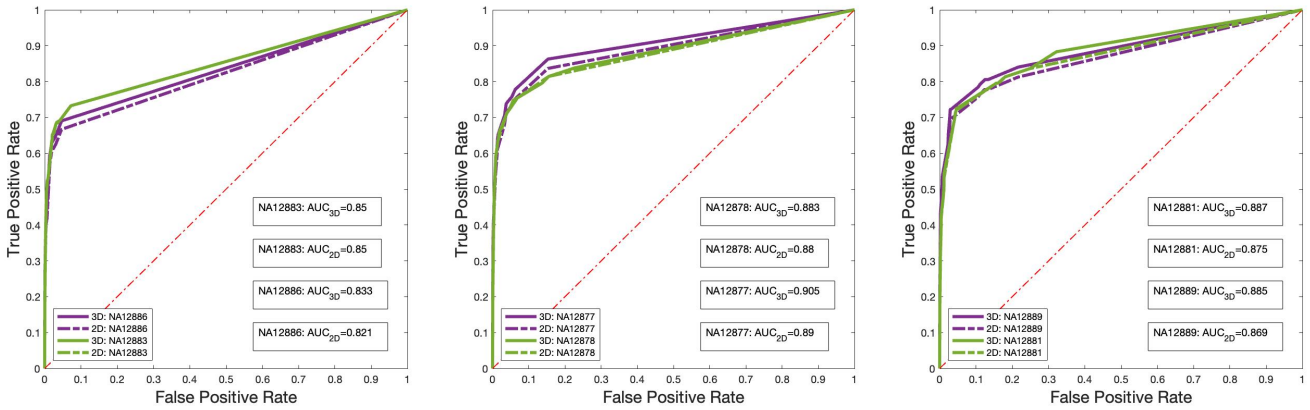
The feasible solution to (6) is obtained by orthogonally projecting the solution (a, b, c) to a three-dimensional feasible region Δ (see Fig. 1). In particular, the three-dimensional space partitions into 15 different regions that projects onto a vertex, edge, or surface of the feasible set for infeasible points. Alternatively, we propose posing the constraints using our projection algorithm that is inspired by methods used in previous work (see [13]). Here, we describe our algorithm of solving the optimization problem. At each iteration k, we alternate between fixing one individual (i.e. one dimension) and projecting onto the remaining two individuals. In particular, the process consists of the following three steps:
For each 3D point in SPIRAL solution (a, b, c):

Step 1: Fix c, and orthogonally project the solution to the two-dimensional feasible region formed by (a and b). This two-dimensional space partitions into 6 different regions that projects onto a point or an edge.
Step 2: Repeat the process by fixing the second individual b and projecting the solution orthogonally to the two-dimensional feasible region formed by a and c.
Step 3: Fix the third individual a and project the solution



(a) Child signal (b) Parent signal (c) Grandparent signal

Fig. 2. ROC curves of individuals' reconstructed signals using simulated data with coverage $\lambda_c = \lambda_p = \lambda_{gp} = 3$.



(a) Child signal (b) Parent signal (c) Grandparent signal

Fig. 3. ROC curves of individuals' reconstructed signals using 1000 Genomes Project data.

orthogonally to the two-dimensional feasible region formed by b and c. Repeat the process until it converges, i.e., the point is inside the 3D feasible region.

3. RESULTS

Experiment I: Simulated Data. We first tested the performance on simulated data to match our assumptions. We generated binary signals for each individual with the sequencing depth (or coverage). We simulated the true signal for a grandparent, a parent, and a child by creating a vector of 10^5 potential SVs. We then selected uniformly at random 5000 locations to be true variants for the grandparent. The parent signal was then generated by randomly selecting 50% of the grandparent variants to be inherited, that is $\sim$2500 SVs in the parent. Finally, 50% of the parent variants were randomly selected to be inherited in the child signal, ~ 1750 SVs in the child signal. Doing it that way confirms that these SVs follow the Mendelian rules.

Given an optimal or near-optimal τ and β values ($\tau \approx 1$ and $\beta \approx 1$) and high noise $\epsilon = 0.3$, our method is able to reconstruct the signals for each individual well. In Fig. 2 we show ROc curves generated for a simulated data set. Moreover, our block-coordinate descent approach was able to give results as good as the actual 3D projections, blue lines match the red line in ROCs.

Experiment II: 1000 Genomes Project Data. To validate our method, we consider a three-generation, 17- member family pedigree of European ancestry (CEU) from a recent study to test our mathematical model with relatedness and rarity of SVs [15]. We obtain our candidate set of SVs from the GASV pipeline. We run two experiments on two groups of individuals of three generations: (Child ID, Parent ID, Grandparent ID) = (NA12883, NA12878, NA12881) and (NA12886, NA12877, NA12889). Our method is able to well reconstruct the signals for each individual with relatively high AUC as shown in the left column of Fig. 3. For both experiments, the AUC ranged from $\sim 85\%$ to $\sim 90\%$, which are high compared to previous studies in literature [13, 7]. Huge part of

these errors are false positives (an SV detected to be there, but it should not).

4. CONCLUSIONS

We present a generalized approach to predict germline SVs in related individuals over two generations. Our proposed model leverages both sparsity and relatedness, by such it reduces the number of false positives predicted in simulated and real genomic data of multiple generations. Moreover, our proposed block-coordinate approach is able to produce results as accurate as the 3D projections of the solution, demonstrating feasibility for more complex and high dimensional pedigrees when the direct high dimensional projection is not possible. In future work, we intend to apply this work on the diploid case of a multi-generational framework with multiple offspring.

5. REFERENCES

[1] Jeffrey R. MacDonald, Robert Ziman, Ryan K. C. Yuen, Lars Feuk, and Stephen W. Scherer, "The database of genomic variants: a curated collection of structural variation in the human genome," *Nucleic acids research*, vol. 42, pp. D986–D992, 2013.

[2] Iñigo Martincorena and Peter J Campbell, "Somatic mutation in cancer and normal cells," *Science*, vol. 349, no. 6255, pp. 1483–1489, Sept. 2015.

[3] 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," *Nature*, vol. 20, pp. 623–635, 2010.

[4] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston, "DNA sequencing at 40: past, present and future," *Nature*, vol. 550, no. 7676, pp. 345–353, Oct. 2017.

[5] Richard M Durbin, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, Oct. 2010.

[6] Alon Keinan and Andrew G Clark, "Recent explosive human population growth has resulted in an excess of rare genetic variants," *Science*, vol. 336, no. 6082, pp. 740–743, May 2012.

[7] Chao Xie and Martti T Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BMC Bioinformatics*, vol. 10, no. 1, pp. 80, Mar. 2009.

[8] Christopher A Miller, Oliver Hampton, Cristian Coarfa, and Aleksandar Milosavljevic, "ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads," *PLOS ONE*, vol. 6, no. 1, pp. e16327, Jan. 2011.

[9] Paul Medvedev, Marc Fiume, Misko Dzamba, Tim Smith, and Michael Brudno, "Detecting copy number variation with mated short reads," *Genome Res*, vol. 20, no. 11, pp. 1613–1622, Aug. 2010.

[10] Alberto Magi, Matteo Benelli, Seungtai Yoon, Franco Roviello, and Francesca Torricelli, "Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm," *Nucleic Acids Res*, vol. 39, no. 10, pp. e65, Feb. 2011.

[11] Jared T Simpson, Rebecca E McIntyre, David J Adams, and Richard Durbin, "Copy number variant detection in inbred strains from short read sequence data," *Bioinformatics*, vol. 26, no. 4, pp. 565–567, 2009.

[12] Melissa Spence, Mario Banuelos, Roummel F. Marcia, and Suzanne Sindi, "Detecting novel structural variants in genomes by leveraging parent-child relatedness," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 943–950.

[13] Mario Banuelos, Lasith Adhikari, Rubi Almanza, Andrew Fujikawa, Jonathan Sahagún, Katharine Sanderson, Melissa Spence, Suzanne Sindi, and Roummel F. Marcia, "Nonconvex regularization for sparse genomic variant signal detection," in *2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2017, pp. 281–286.

[14] 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.

[15] Michael Eberle et al, "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree," *Genome Res*, vol. 27, no. 1, pp. 157–164, Nov. 2016.