Detecting novel genomic structural variants through negative binomial optimization

Andrew Lazar
Dept. of Applied Mathematics
University of California, Merced
Merced, CA USA
alazar2@ucmerced.edu

Mario Banuelos
Dept. of Mathematics
California State University, Fresno
Fresno, CA USA
mbanuelos22@csufresno.edu

Suzanne Sindi
Dept. of Applied Mathematics
University of California, Merced
Merced, CA USA
ssindi@ucmerced.edu

Roummel F. Marcia
Dept. of Applied Mathematics
University of California, Merced
Merced, CA USA
rmarcia@ucmerced.edu

Abstract*—Structural variants (SVs) are short sequences of DNA, larger than one nucleotide, that can vary between members of the same species. Although SVs are relatively rare, compared to single nucleotide variants (SNVs) they are an important source of genetic variation and some SVs have been associated with diseases and susceptibility to certain types of cancer. SV detection is commonly performed by aligning sequenced fragments of an individual's genome to a high-quality reference genome. Candidate SVs correspond to discordant mapped configurations of fragments; however, errors in the sequencing also lead to potential discordant mappings. Because of this error, many candidate SVs are in fact false positives. When sequencing coverage is high, SV detection is more accurate, but this comes at higher sequencing cost. Sequencing at low coverage does reduce cost, but increases error and complexity of SV detection. The goal of our work is to use mathematical optimization to improve SV detection in low-coverage DNA sequencing data. Previous studies of SV detection have modeled coverage with a Poisson distribution, but this assumes the mean and variance are the same. In an effort more closely model the experimental data we use the negative binomial distribution, which allows for the mean and variance to differ, and contains the Poisson distribution as a special case. Our approach also control false positive predictions by simultaneously considering simultaneous SV prediction in a parent and child. We assume that most SVs carried by a child are inherited from a parent but a small fraction may be novel to the child. We balance the rarity of novel versus inherited SVs by enforcing sparsity through an l1-penalty and compare this negative binomial reconstruction algorithm to the Poisson reconstruction algorithm by testing both on the same simulated data sets.

***Index Terms*—Sparse signal recovery, structural variants, non-convex optimization, computational genomics, next-generation sequencing data**

I. Introduction

Structural variants (SVs) are regions of a genome (larger than a single nucleotide) that vary between individuals in the same species. SVs represent only one type of genomic variation and, though generally rare, they are an increasingly important class of variation in humans as they have been implicated in hereditary diseases and susceptibility to certain types of cancer [1]–[3]. Detecting SVs involves the sequencing and mapping of DNA fragments from a candidate genome to an established reference genome and analyzing the configuration of mapped fragments [4], [5]. Because SVs correspond to differences in individual genomes, relative to a reference, SVs are detected through identifying discordant mapping configurations [6]–[8]. However, SV prediction is far from perfect and true SVs can be challenging to separate from discordant configurations due to errors in sequencing and mapping. Furthermore, distinguishing true SVs from errors is made even more challenging in low-coverage sequencing settings [9]–[16]. In this work, we use a negative binomial framework to model the expected number of fragments covering any position in a genome [17]–[19].

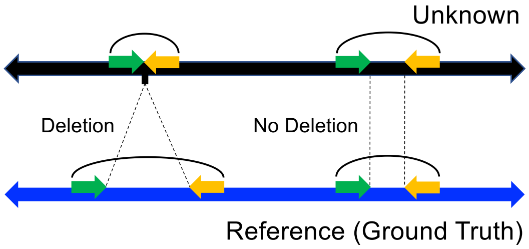

Fig. 1. Illustration of a structural variant in a genome sequence. When a fragment from an unknown genome does not map concordantly to the reference genome, this is considered a signal for a structural variant. In this illustration, a deletion (left) occurs when the fragment from the unknown genome maps to a larger region in the reference.

II. Problem Formulation

We now present a general framework for predicting structural variants (SVs) within sequencing data from one parent (p)

This work was supported by National Science Foundation Grant IIS-1741490.

and one child (c). For simplicity, we consider both individuals to be haploid (only one copy of each chromosome).

Statistical model. Let the true signal $\vec{f}^* \in \{0,1\}^n$ for an individual be a binary-valued vector that indicates the presence of a genetic variant, with $\vec{f}^*_j = 1$ if a variant is present at location j and 0 otherwise [20]–[22]. Furthermore, let the vectors $\vec{y}_p$ and $\vec{y}_c$ correspond to the parent and child observed measurements, respectively, and be given by

$$\begin{aligned} \vec{y}_p &\sim \text{NegBin}(\vec{\mu}_p, \vec{\sigma}^2_p), \\ \vec{y}_c &\sim \text{NegBin}(\vec{\mu}_c, \vec{\sigma}^2_c), \end{aligned}$$

where the mean μ_i and variance σ_i^2, with $i \in \{p, c\}$, of depth of coverage are determined by the sequencing data of each respective individual. Consider the stacked parent-child signal $\vec{y} = [\ \vec{y}_p\ ;\ \vec{y}_c\ ]$ and corresponding mean and variance vectors, $\vec{\mu}$ and $\vec{\sigma}^2$, where the notation $\vec{\sigma}^2$ is to be understood component-wise. Specifically, we have the following expressions for the components of $\vec{\mu}$ and $\vec{\sigma}^2$:

$$(\mu)_j = \left(A\vec{f}^*\right)_j \quad \text{and} \quad (\sigma)^2_j = \left(A\vec{f}^*\right)_j + \frac{1}{r}\left(A\vec{f}^*\right)^2_j,$$

where A is a mapping that linearly projects the true signal $\vec{f}^*$ onto the set of observations, and r is the dispersion parameter of the negative binomial distribution. Under this model, the probability of observing $\vec{y}$ is given by the following expression:

$$p(\vec{y}) = \prod_{j=1}^{n} \begin{pmatrix} y_j + \frac{\mu_j^2}{\sigma_j^2 - \mu_j} - 1 \\ y_j \end{pmatrix} \left(\frac{\mu_j}{\sigma_j^2}\right)^{\frac{\mu_j^2}{\sigma_j^2 - \mu_j}} \left(1 - \frac{\mu_j}{\sigma_j^2}\right)^{y_j}.$$

To avoid using the gamma function, we assume that $r \in \mathbb{Z}^+$. In addition, we know $\sigma_j^2 = \mu_j + \frac{1}{r}\mu_j^2$, where σ_j^2 is maximized when $r = 1$. Thus, ignoring constant terms, the negative log-likelihood term, $F(\mu, \sigma^2)$, becomes

$$F(\mu) \equiv \sum_{j=1}^{n} (y_j + 1)\log\left(1 + \mu_j\right) - y_j \log\left(\mu_j\right).$$

However, knowing that the mean $\mu_j = e_i^T A f$ and adding the small parameter ε to represent sequencing or mapping error, we arrive at our negative log-likelihood objective function:

$$F\left(f\right) \equiv \sum_{j=1}^{n} (y_j{+}1)\log\left(1 + e_i^T A f + \varepsilon\right) {-} y_j \log\left(e_i^T A f + \varepsilon\right),$$

where e_i is the i^{th} column of the $n \times n$ identity matrix. In previous work, we assumed that a child will have an SV at a certain location only if the parent also has the SV at the same location. In this work, although we assume that the variants in the child primarily come from the parent (which we call inherited SVs), the child may also have variants not present in the parent (which we call novel SVs). To account for these two types of SVs, we decompose the SV signal for the child as $\vec{f}^*_c = \vec{f}^*_i + \vec{f}^*_n$, where $\vec{f}^*_i \in \{0,1\}^m$ is the vector of SVs that are *i*nherited from the parent and $\vec{f}^*_n \in \{0,1\}^m$ is the vector of SVs that are *n*ovel. In particular, the vector $\vec{f}^*_i$ has either a 1 at position j if an SV is inherited from the parent at position j or a 0 otherwise. Similarly, the vector $\vec{f}^*_n$ has a 1 if there is an SV at position j that is not inherited from the parent and 0 otherwise. (For an illustration, see Fig. 2.) Note that for every location, $\vec{f}_i$ and $\vec{f}_n$ cannot be both 1 simulatenously since an SV cannot be both inherited and novel.

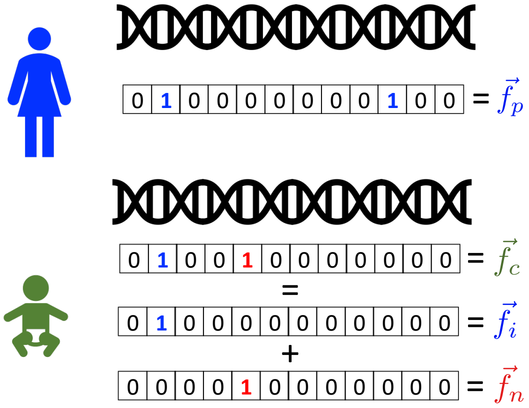
0 1 0 0 0 0 0 0 0 1 0 0 = $\vec{f}_p$
0 1 0 0 1 0 0 0 0 0 0 0 = $\vec{f}_c$
=
0 1 0 0 0 0 0 0 0 0 0 0 = $\vec{f}_i$
+
0 0 0 0 1 0 0 0 0 0 0 0 = $\vec{f}_n$

Fig. 2. The parent SV signal $\vec{f}_p$ and the child SV signal $\vec{f}_c$. The vector of child SVs inherited from the parent is denoted by $\vec{f}_i$, and the vector of novel SVs is denoted by $\vec{f}_n$. Note that $\vec{f}_c = \vec{f}_i + \vec{f}_n$.

Familial constraints. In this work, we use gradient-based optimization methods to minimize $F(f)$. As such, we allow f to take on real values instead of being binary valued. In addition, we formulate the biological constraints on the SV signals mathematically and incorporate them within the optimization problem [23].

Since $\vec{f}_i$ and $\vec{f}_n$ cannot be both 1 simulatenously at each location, the following must hold:

$$0 \leq \vec{f}_i + \vec{f}_n \leq 1,$$

where the inequalities are to be understood component-wise. Furthermore, an inherited SV must come from the parent. Therefore, if $(\vec{f}_p)_j = 0$, then $(\vec{f}_i)_j = 0$. Similarly, if $(\vec{f}_i)_j = 1$, then $(\vec{f}_p)_j = 1$. In other words, $\vec{f}_p$ and $\vec{f}_i$ must satisfy

$$0 \leq \vec{f}_i \leq \vec{f}_p \leq 1.$$

Moreover, if there is an SV in the parent at location j, then the child cannot have a novel SV at that location. Similarly, if there is a novel SV present in the child at location j, that SV cannot be present in the parent, i.e.,

$$0 \leq \vec{f}_n \leq 1 - \vec{f}_p.$$

Finally, since $\vec{f}$ should take on the values of either 0 or 1, we require that $0 \leq \vec{f} \leq 1$.

Combining all of these constraints, we define the set of all vectors satisfying these constraints by $\mathcal{S}$, given by

$$\mathcal{S} = \left\{ \begin{bmatrix} \vec{f}_p \\ \vec{f}_i \\ \vec{f}_n \end{bmatrix} \in \mathbb{R}^{3m} : \begin{array}{c} 0 \leq \vec{f}_i + \vec{f}_n \leq 1, \\ 0 \leq \vec{f}_i \leq \vec{f}_p \leq 1, \\ 0 \leq \vec{f}_n \leq 1 - \vec{f}_p, \\ 0 \leq \vec{f}_p, \vec{f}_i, \vec{f}_n \leq 1 \end{array} \right\}.$$

Parsimonious solutions. Genomes within the same species are highly similar. Therefore, structural variants are very rare. We incorporate this biological phenomenon in our mathematical model by imposing an ℓ_1-norm penalty term in our problem formulation, which is a common technique found in statistical literature to promote sparsity in the solution [24]–[26]. We further assume that novel SVs are even rarer. Thus, we associate a different (larger) regularization parameter with the novel SVs. Mathematically, we express this penalty term as

$$\text{pen}(\vec{f}) \;=\; \big(\|\vec{f}_p\|_1 + \|\vec{f}_i\|_1\big) + \gamma\|\vec{f}_n\|_1,$$

where $\gamma \gg 1$ is a penalty parameter that places greater weight on $\vec{f}_n$ to promote further sparsity.

Optimization approach. Assuming that these SVs are rare, we express the SV prediction problem as the following sparse signal constrained optimization problem:

$$\begin{aligned} &\underset{\vec{f}\in\mathbb{R}^{3n}}{\text{minimize}} && \psi(\vec{f}) \equiv F(\vec{f}) + \tau\text{pen}(\vec{f}) \\ &\text{subject to} && \vec{f}\in\mathcal{S}, \end{aligned} \tag{1}$$

where $\vec{f} = [\vec{f}_p; \vec{f}_i; \vec{f}_n]$ and $\tau > 0$ is a regularization parameter that balances the data-fidelity $F(f)$ term with the sparsity-promoting penalty term. We solve (1) using the Sparse Poisson Intensity Reconstruction ALgorithm (SPIRAL) framework [27] by minimizing a sequence of quadratic models to the function $F(\vec{f})$. First we first define the second-order Taylor series approximation $F^k(f)$ to $F(f)$ at the current iterate $\vec{f}^k$:

$$\begin{aligned} F^k(\vec{f}) \;=\; & F(\vec{f}^k) + (\vec{f}-\vec{f}^k)^\top\nabla F(\vec{f}^k) \\ & + \tfrac{1}{2}(\vec{f}-\vec{f}^k)^\top\nabla^2F(\vec{f}^k)(\vec{f}-\vec{f}^k). \end{aligned} \tag{2}$$

The gradient of $F(\vec{f})$ is given by

$$\nabla F(f) = \sum_{j=1}^{n}\frac{y_j+1}{1+e_j^TAf+\varepsilon}A^Te_j - \frac{y_j}{e_j^TAf+\varepsilon}A^Te_j, \tag{3}$$

where $A \in \mathbb{R}^{2m\times 3m}$ is the coverage matrix given by

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

where $I_m \in \mathbb{R}^{m\times m}$ is the $m\times m$ identity matrix, λ_p and λ_c are the sequencing coverage of the parent and child, respectively, and $\epsilon > 0$ is the measurement error corresponding to the sequencing processing. To simplify our quadratic model, we approximate the second-derivative Hessian matrix with a scalar multiple of the identity matrix $\alpha_k I$, where $\alpha_k > 0$ (see [28], [29] for details). We define the quadratic model

$$\widetilde{F}^k(\vec{f}) \equiv F(\vec{f}^k) + (\vec{f}-\vec{f}^k)^T\nabla F(\vec{f}^k) + \frac{\alpha_k}{2}||\vec{f}-\vec{f}^k||_2^2. \tag{4}$$

Now, each quadratic subproblem will be of the form

$$\begin{aligned} \vec{f}^{k+1} \;=\; & \underset{\vec{f}\in\mathbb{R}^{3m}}{\arg\min}\;\; F^k(\vec{f}) + \tau\text{pen}(\vec{f}) \\ & \text{subject to } \vec{f}\in\mathcal{S}. \end{aligned}$$

This constrained quadratic subproblem is equivalent to the following subproblem:

$$\begin{aligned} \vec{f}^{k+1} \;=\; & \underset{\vec{f}\in\mathbb{R}^{3m}}{\arg\min}\quad \mathcal{Q}(\vec{f}) = \frac{1}{2}\|\vec{f}-\vec{s}^{\,k}\|_2^2 + \frac{\tau}{\alpha_k}\text{pen}(\vec{f}) \\ & \text{subject to }\; \vec{f}\in\mathcal{S}, \end{aligned} \tag{5}$$

where

$$\vec{s}^{\,k} = \begin{bmatrix} \vec{s}_p^{\,k} \\ \vec{s}_i^{\,k} \\ \vec{s}_n^{\,k} \end{bmatrix} = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$$

(see [27] for details). Note that $\mathcal{Q}(\vec{f})$ separates into the sum

$$\mathcal{Q}(\vec{f}) = \sum_{j=1}^{m}\mathcal{Q}_j\big((\vec{f}_p)_j, (\vec{f}_i)_j, (\vec{f}_n)_j\big),$$

where $\mathcal{Q}_j\colon \mathbb{R}^3 \to \mathbb{R}$ and

$$\begin{aligned} & \mathcal{Q}_j\big((\vec{f}_p)_j, (\vec{f}_i)_j, (\vec{f}_n)_j\big) \\ & = \; \frac{1}{2}\Big\{\big((\vec{f}_i-\vec{s}_i^{\,k})_j\big)^2 + \big((\vec{f}_n-\vec{s}_n^{\,k})_j\big)^2 + \big((\vec{f}_p-\vec{s}_p^{\,k})_j\big)^2\Big\} \\ & \quad + \frac{\tau}{\alpha_k}\Big\{\; |(\vec{f}_p)_j| + |(\vec{f}_i)_j| + \gamma|(\vec{f}_n)_j|\Big\}. \end{aligned}$$

Note that the bounds for $\mathcal{S}$ are component-wise. Therefore, (5) separates into subproblems of the form

$$\begin{aligned} \underset{f_p,f_i,f_n\in\mathbb{R}}{\text{minimize}}\;\; & \frac{1}{2}(f_p-s_p)^2 + \frac{1}{2}(f_i-s_i)^2 + \frac{1}{2}(f_n-s_n)^2 \\ & + \frac{\tau}{\alpha_k}|f_p| + \frac{\tau}{\alpha_k}|f_i| + \frac{\gamma\tau}{\alpha_k}|f_n| \\ \text{subject to}\;\; & 0 \le f_i + f_n \le 1, \quad 0 \le f_i \le f_p \le 1, \\ & 0 \le f_n \le 1 - f_p, \quad 0 \le f_i, f_n, f_p \le 1, \end{aligned} \tag{6}$$

where $\{f_p, f_i, f_n\}$ and $\{s_p, s_i, s_n\}$ are scalar components of the vectors $\{\vec{f}_p, \vec{f}_i, \vec{f}_n\}$ and $\{\vec{s}_p, \vec{s}_i, \vec{s}_n\}$, respectively, at the same location. The constrained optimization problem (5) can be solved analytically by compling the square in the ojbective function and orthogonally projecting onto the feasible set (see [30] for details).

III. RESULTS

We implemented our method for variant detection using the Negative Binomial-based SPIRAL method (SPNB), similar to previous approaches [23]. We analyzed the results on simulated data and compared the results to the Poisson based SPIRAL (SPP) method. Similar to previously published methods, we observed the variant predictions in a one-parent/one-child model [23], [31]. Our method contained a sparsity promoting term τ. This method has a second regularization parameter, γ, which is chosen to promote more sparsity within the novel variants, f_n. In every case, the SPIRAL algorithm was run with the terminating criteria, if the relative difference between consecutive iterates converged to $||\vec{f}^{k+1} - \vec{f}^k||_2/||\vec{f}^k||_2 \le 10^{-8}$.

τ/γ	**2**	**10**	**20**	**50**	**100**	**200**	**500**
0.01	0.905	**0.906**	**0.906**	**0.906**	**0.906**	**0.906**	**0.906**
0.1	**0.906**	**0.906**	**0.906**	**0.906**	**0.906**	**0.906**	**0.906**
1	**0.906**	**0.906**	**0.906**	**0.906**	**0.906**	0.575	0.575
10	0.895	0.797	0.758	**0.906**	**0.906**	**0.906**	**0.906**
100	0.880	0.880	0.880	0.880	0.880	0.880	0.880
1000	0.513	0.517	0.517	0.517	0.517	0.517	0.517

TABLE I

THE TABLE ABOVE SHOWS THE AUCS FOR THE CHILD WITH 5% NOVEL VARIANTS USING THE SPNB ALGORITHM. THE VALUES ALONG EACH COLUMN ARE γ, WHILE THE VALUES ALONG EACH ROW ARE τ. THE HIGHEST AUC IS IN BOLDFACE. WE NOTICE A ROBUSTNESS IN THE VALUES OF τ AND γ WHICH ACHIEVE THE HIGHEST AUC

τ/γ	**2**	**10**	**20**	**50**	**100**	**200**	**500**
0.01	0.905	0.905	0.874	0.876	**0.906**	**0.906**	0.895
0.1	**0.906**	**0.906**	**0.906**	0.895	**0.906**	**0.906**	**0.906**
1	**0.906**	**0.906**	**0.906**	**0.906**	**0.906**	0.522	0.522
10	**0.906**	**0.906**	0.797	**0.906**	**0.906**	**0.906**	**0.906**
100	0.880	0.880	0.880	0.880	0.880	0.880	0.880
1000	0.517	0.517	0.517	0.517	0.517	0.517	0.517

TABLE II

THE TABLE ABOVE SHOWS THE AUCS FOR THE CHILD WITH 5% NOVEL VARIANTS USING THE SPP ALGORITHM. WE NOTICE A LESS ROBUSTNESS, WHEN COMPARED TO THE SPNB TABLE, OF THE HIGHEST AUC

A. Simulated Data

Similar to our previous approach, the model was developed in the form of a one-parent and one-child with a haploid genome assumption. Before applying it to real human data, with diploid genomes that violate our assumptions, we studied the performance on data we simulated that matches our assumptions. We simulated the true signal for the parent and child by creating the vector, $\vec{f}$ of size 10^6 and selecting 500 locations to be true variants for the parent and child. We control the number of novel SVs in the child by by first selecting 500 locations at random to be the true SVs in the parent. We construct the child signal by randomly selecting $\lfloor 500p \rfloor$ (where p is the percentage of novel variants), of the parent variants to be inherited and then choosing $(500 - \lfloor 500p \rfloor)$ locations of the remaining $(10^6 - 500)$ locations to be novel [27].

B. Analysis

We compared the performance of both SPNB and SPP when reconstructing Negative Binomial distributed data and Poisson distributed data. In most cases SPNB produced an area under the curve that was equal to SPP. Figure 2 and Figure 3 illustrate these findings. For the parent signal, we were able to find higher accuracy in the AUC compared to the child. In few cases we found that the AUC was different between both methods. In regards to the value of the AUC, we observed the highest of the highest AUCs for 2% novel data and the lowest of the highest AUCs for 20% novel data in the reconstruction of the child and parent. We highlight results for the parent and child together as well as each of them individually. We tested a variety of different values for τ and γ to determine the regularization parameters' effect on our results. We found that for SPNB there was a more robust interval of τ and γ for which the highest AUC was achieved when compared to SPP. Notice in Table I, the block of boldface AUCs which represent the highest AUCs for that percentage and individual. When compared to Table II, we see slightly more variance of AUCs and less robust intervals. We observed this mostly in cases where the percentage of novel variants was small ($< 10\%$).

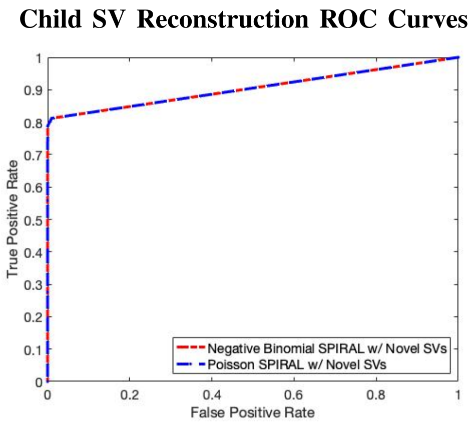

Fig. 3. ROC curves for the child for 2% novel variants which illustrate the true postive rate vs. the false positive rate. We observe for SPNB, the AUC is 0.9049 and for SPP we have the same AUC. We have $\tau = 1$ and $\gamma = 50$

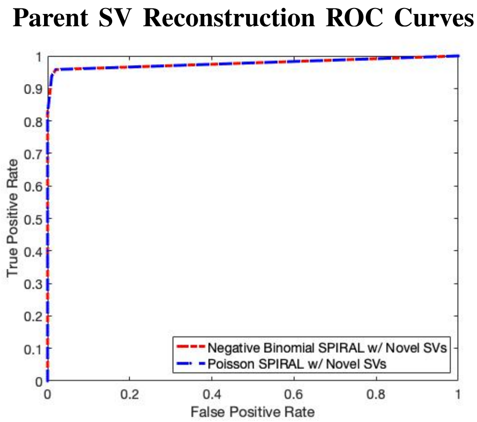

Fig. 4. ROC curves for the parent for 2% novel variants which illustrate the true postive rate vs. the false positive rate. We observe for SPNB, the AUC is 0.9777 and for SPP we have the same AUC. We have $\tau = 1$ and $\gamma = 50$

IV. CONCLUSIONS

We propose a method which builds on our previously developed SPIRAL method, which reconstructs signals arising from the Negative Binomial distribution rather than the Poisson distribution. This method detects both inherited and novel variants within the child. Both relatedness and sparsity are incorporated into our method. We found a robustness of best results (highest AUC) by considering various factors, including the percent of novel structural variants, penalty parameters τ and γ, and the comparison of SPNB versus SPP.

ACKNOWLEDGMENT

The authors would like to thank Melissa Anisko for helpful conversations regarding this work.

References

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

[2] J. Weischenfeldt, F. Symmons, O.and Spitz, and J.O. Korbel, "Phenotypic impact of genomic structural variation: insights from and for human disease," *Nature Reviews Genetics*, vol. 14, no. 2, pp. 125–138, 2013.

[3] L. R. Pal and J. Moult, "Genetic basis of common human disease: Insight into the role of missense SNPs from genome-wide association studies," *Journal of Molecular Biology*, vol. 427, no. 13, pp. 2271–2289, 2015.

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

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

[6] D. M. Altshuler, E. S. Lander, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, et al., "A map of human genome variation from population scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[7] 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.

[8] 1000 Genomes Project Consortium et al., "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[9] 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*, p. btv751, 2016.

[10] S. Yoon, V. Xuan, Z.and 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.

[11] V. Boeva, A. Zinovyev, K. Bleakley, J.-P. Vert, I. Janoueix-Lerosey, O. Delattre, and E. Barillot, "Control-free calling of copy number alterations in deep-sequencing data using gc-content normalization," *Bioinformatics*, vol. 27, no. 2, pp. 268–269, 2011.

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

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

[14] F. Hormozdiari, C. Alkan, E. E. Eichler, and S. C. Sahinalp, "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes," *Genome research*, vol. 19, no. 7, pp. 1270–1278, 2009.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] M. Banuelos, S. Sindi, and R. F Marcia, "Negative binomial optimization for biomedical structural variant signal reconstruction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 906–910.

[20] M. Banuelos, R. Almanza, L. Adhikari, S. Sindi, and R. F. Marcia, "Sparse signal recovery methods for variant detection in next-generation sequencing data," 2016, Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[21] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, and S. Sindi, "Constrained variant detection with sparc: Sparsity, parental relatedness, and coverage.," in *EMBC*, 2016, pp. 3490–3493.

[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 *Medical Measurements and Applications (MeMeA), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 275–280.

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

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[25] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[26] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006.

[27] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms—theory and practice," *IEEE Trans. on Image Processsing*, vol. 21, pp. 1084 – 1096, 2011.

[28] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA J. Numer. Anal.*, vol. 8, no. 1, pp. 141–148, 1988.

[29] E. G. Birgin, J. M. Martínez, and M. Raydan, "Nonmonotone spectral projected gradient methods on convex sets," *SIAM Journal on Optimization*, vol. 10, no. 4, pp. 1196–1211, 2000.

[30] 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.

[31] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, and S. Sindi, "Sparse genomic structural variant detection: Exploiting parent-child relatedness for signal recovery," in *Statistical Signal Processing Workshop (SSP), 2016 IEEE*. IEEE, 2016, pp. 1–5.