Novel structural variant genome detection in extended pedigrees 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 novel rearrangements in genomes of organisms and lead to a species' genomic heterogeneity. While rare, SVs represent an increasingly important class of genetic variation. To detect SVs, DNA fragments from a test genome are compared to a high-quality reference genome, where discordant mappings provide evidence of potential SVs. This process is susceptible to sequencing and mapping errors. In low-coverage settings, differentiating true SVs from errors is even more difficult. In this work, we consider SV detection within extended pedigrees by using a negative binomial framework to model the expected number of fragments covering any position in a genome and exploit familial relationships to improve detection accuracy.**

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

I. Introduction

Structural variants (SVs) are areas within a genome that are larger than a single nucleotide that can vary between individuals in the same species. SVs are a type of genomic variation, such as inversions, deletions and duplications; and, although generally rare, they form an increasingly important class of variation in human genomes as they have been associated with particular hereditary diseases and susceptibility to certain types of cancer [1]–[3]. SV detection is a process that involves (i) fragmenting candidate DNA, (ii) sequencing and mapping them to an established reference high-quality genome, and (iii) analyzing the positions of the mapped fragments [4]–[8]. However, due to errors in this process, false predictions may be made and true variants may be missed. Furthermore, identifying true SVs from sequencing errors is made even more difficult in low-coverage sequencing settings [9]–[16]. In this work, we build upon our previous work on using a negative binomial framework to model the expected number of fragments covering any position in a genome [17]–[19] and leverage parent-child-trio relationships to reduce the false-positive rate of SV predictions for both parents and child.

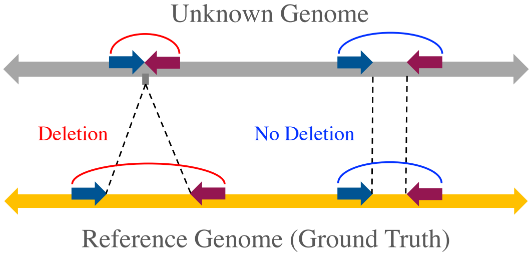


Fig. 1: Illustration of a structural variant in a genome sequence. When a fragment from an unknown genome does not map concordantly to the reference genome, this is considered a signal for a structural variant. In this illustration, a deletion (left) occurs when the fragment from the unknown genome maps 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 two parents (p_1 and p_2) and one child (c). For simplicity, we consider all individuals to be haploid (only one copy of each chromosome).

Statistical model. Consider two unrelated indviduals, p_1 and p_2, and their child, c. With $I \in \{p_1, p_2, c\}$, let the true signal $\vec{f}_I^* \in \{0,1\}^N$ be a binary-valued vector that indicates the presence of a structural variant in individual I's genome sequence, with $(\vec{f}_I^*)_j = 1$ if a variant is present at location

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

j and 0 otherwise. Furthermore, let the vector $\vec{y}_I \in \mathbb{R}^N$ correspond to the measurement vector with

$$\vec{y}_I \sim \text{NegBin}(\vec{\mu}_I, \vec{\sigma}_I^2),$$

where the mean μ_I and variance σ_I^2 depth of coverage are determined by the sequencing data of each respective individual. (Here, the notation $\vec{\sigma}^2$ is to be understood component-wise.) Consider the stacked measurement signal $\vec{y} = [\ \vec{y}_{p_1}\ ;\ \vec{y}_{p_2}\ ;\ \vec{y}_c\ ] \in \mathbb{R}^{3N}$ and corresponding mean and variance vectors, $\vec{\mu} \in \mathbb{R}^N$ and $\vec{\sigma}^2 \in \mathbb{R}^N$. Specifically, we have the following expressions for the components of $\vec{\mu}$ and $\vec{\sigma}^2$:

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

where A is a mapping that linearly projects the true signal $\vec{f}^* = [\ \vec{f}_{p_1}^*\ ;\ \vec{f}_{p_2}^*\ ;\ \vec{f}_c^*\ ] \in \{0,1\}^{3N}$ onto the $3N$-dimensional 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:

$$p(\vec{y}) = \prod_{j=1}^{3N} \binom{\vec{y}_j + \frac{\mu_j^2}{\sigma_j^2-\mu_j} - 1}{\vec{y}_j} \left(\frac{\mu_j}{\sigma_j^2}\right)^{\frac{\mu_j^2}{\sigma_j^2-\mu_j}} \left(1 - \frac{\mu_j}{\sigma_j^2}\right)^{\vec{y}_j}.$$

We avoid using the gamma function by letting $r \in \mathbb{Z}^+$. Furthermore, since $\sigma_j^2 = \mu_j + \frac{1}{r}\mu_j^2$, we can maximize σ_j^2 by letting $r = 1$ to allow for the largest variance in the measurements. Thus, ignoring constant terms, the negative log-likelihood function, $F(\mu, \sigma^2)$, corresponding to the negative binomial probability above, is given by

$$F(\mu) \equiv \sum_{j=1}^{3N} (\vec{y}_j + 1)\log(1 + \mu_j) - \vec{y}_j \log(\mu_j).$$

However, knowing that the mean $\mu_j = (A\vec{f}^*)_j$ and incorporating a small parameter $\varepsilon > 0$ to represent sequencing or mapping error, we obtain our negative log-likelihood objective function:

$$F(\vec{f}) \equiv \sum_{j=1}^{3N} (\vec{y}_j + 1)\log\left(1 + (A\vec{f})_j + \varepsilon\right) - \vec{y}_j \log\left((A\vec{f})_j + \varepsilon\right).$$

To apply gradient-based optimization methods to minimize $F(\vec{f})$, we allow $\vec{f}$ to take on real values instead of being binary valued and require that

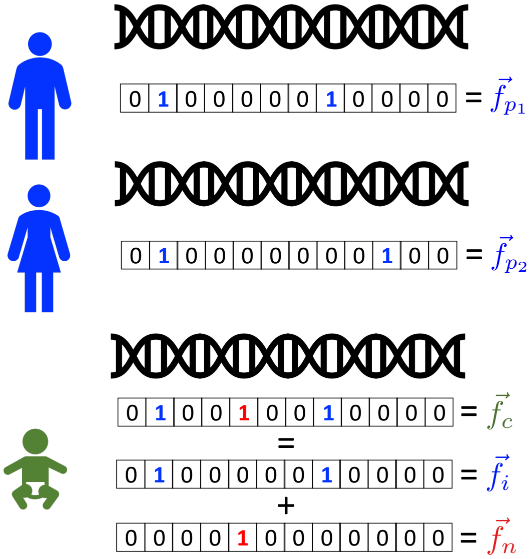
$$0 \leq \vec{f}_{p_1}, \vec{f}_{p_2}, \vec{f}_c \leq 1, \qquad (1)$$

where the inequalities are understood to be component-wise.

While child SVs can be inherited from a parent, that is not always the case. In previous work, we proposed expressing the child SV true signal as the following decomposition:

$$\vec{f}_c^* = \vec{f}_i^* + \vec{f}_n^*, \qquad (2)$$

where $\vec{f}_i^* \in \{0,1\}^N$ is the indicator vector of *inherited* SVs while $\vec{f}_n^* \in \{0,1\}^N$ is the indicator vector of *novel* SVs (or SVs that are not inherited). Note that because a child SV cannot be both inherited and novel simultaneously, $(\vec{f}_i^*)^\top \vec{f}_n^* = 0$.



0 1 0 0 0 0 0 1 0 0 0 0 = $\vec{f}_{p_1}$
0 1 0 0 0 0 0 0 0 1 0 0 = $\vec{f}_{p_2}$
0 1 0 0 1 0 0 1 0 0 0 0 = $\vec{f}_c$
=
0 1 0 0 0 0 0 1 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. Here, we describe the biological constraints that the SV signals must satisfy in addtion to those in (1) and formulate them mathematically. [20]. First, since $\vec{f}_i$ and $\vec{f}_n$ also take on 0 or 1 as values, we require that

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

Second, from (2), we have that

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

Third, if there is an SV in either 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 both parents, i.e.,

$$0 \leq \vec{f}_n \leq 1 - \vec{f}_{p_1} \quad \text{and} \quad 0 \leq \vec{f}_n \leq 1 - \vec{f}_{p_2}.$$

Fourth, we assume that if both parents have a variant in the same location, the child will inherit this variant, meaning

$$\vec{f}_{p_1} + \vec{f}_{p_2} - 1 \leq \vec{f}_i.$$

Similarly, if the neither parent has a variant present, we assume the child will not have an inherited variant, meaning

$$\vec{f}_i \leq \vec{f}_{p_1} + \vec{f}_{p_2}.$$

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

$$\mathcal{S} = \left\{ \begin{bmatrix} \vec{f}_{p_1} \\ \vec{f}_{p_2} \\ \vec{f}_i \\ \vec{f}_n \end{bmatrix} \in \mathbb{R}^{4N} \ : \begin{array}{l} 0 \leq \vec{f}_i + \vec{f}_n \leq 1, \\ 0 \leq \vec{f}_n \leq 1 - \vec{f}_{p_1}, \\ 0 \leq \vec{f}_n \leq 1 - \vec{f}_{p_2}, \\ \vec{f}_{p_1} + \vec{f}_{p_2} - 1 \leq \vec{f}_i \leq \vec{f}_{p_1} + \vec{f}_{p_2}, \\ 0 \leq \vec{f}_{p_1}, \vec{f}_{p_2}, \vec{f}_i, \vec{f}_n \leq 1 \end{array} \right\}.$$

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 [21]–[23]. 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

$$\text{pen}(\vec{f}) \;=\; \left(\|\vec{f}_{p_1}\|_1 + \|\vec{f}_{p_2}\|_1 + \|\vec{f}_i\|_1\right) + \gamma\|\vec{f}_n\|_1,$$

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:

$$\begin{array}{ll} \underset{\vec{f}\in\mathbb{R}^{4N}}{\text{minimize}} & \psi(\vec{f}) \equiv F(\vec{f}) + \tau\text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S}, \end{array} \tag{3}$$

where $\vec{f} = [\vec{f}_{p_1}; \vec{f}_{p_2}; \vec{f}_i; \vec{f}_n]$ and $\tau > 0$ is a regularization parameter that balances the data-fidelity $F(\vec{f})$ term with the sparsity-promoting penalty term. We solve (3) using the Sparse Poisson Intensity Reconstruction ALgorithm (SPIRAL) framework [24] by minimizing a sequence of quadratic models to the function $F(\vec{f})$. First we first define the second-order Taylor series approximation $F^k(\vec{f})$ to $F(\vec{f})$ at the current iterate $\vec{f}^k$:

$$\begin{aligned} F^k(\vec{f}) \;&=\; F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^\top \nabla F(\vec{f}^k) \\ &\quad + \tfrac{1}{2}(\vec{f} - \vec{f}^k)^\top \nabla^2 F(\vec{f}^k)(\vec{f} - \vec{f}^k). \end{aligned} \tag{4}$$

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

$$\nabla F(\vec{f}) = \sum_{j=1}^{4N} \frac{\vec{y}_j + 1}{1 + e_j^T A\vec{f} + \varepsilon} A^T e_j - \frac{\vec{y}_j}{e_j^T A\vec{f} + \varepsilon} A^T e_j, \tag{5}$$

where e_j is the j^{th} column of the $3N \times 3N$ identity matrix, $A \in \mathbb{R}^{3N\times 4N}$ is the coverage matrix given by

$$A = \begin{bmatrix} (\lambda_{p_1}-\epsilon)I_N & 0 & 0 & 0 \\ 0 & (\lambda_{p_2}-\epsilon)I_N & 0 & 0 \\ 0 & 0 & (\lambda_c-\epsilon)I_N & (\lambda_c-\epsilon)I_N \end{bmatrix},$$

where $I_N \in \mathbb{R}^{N\times N}$ is the $N \times N$ identity matrix, and λ_{p_1}, λ_{p_2}, and λ_c are the sequencing coverage of the parents and child, respectively. To further 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 [25], [26] for details). We define the quadratic model

$$\widetilde{F}^k(\vec{f}) \equiv F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^T \nabla F(\vec{f}^k) + \frac{\alpha_k}{2}\|\vec{f} - \vec{f}^k\|_2^2. \tag{6}$$

Now, each quadratic subproblem will be of the form

$$\begin{array}{rcl} \vec{f}^{k+1} & = & \underset{\vec{f}\in\mathbb{R}^{4N}}{\arg\min} \quad F^k(\vec{f}) + \tau\text{pen}(\vec{f}) \\ & & \text{subject to } \vec{f} \in \mathcal{S}. \end{array}$$

This constrained quadratic subproblem is equivalent to the following subproblem:

$$\begin{array}{rcl} \vec{f}^{k+1} & = & \underset{\vec{f}\in\mathbb{R}^{4N}}{\arg\min} \quad \mathcal{Q}(\vec{f}) = \dfrac{1}{2}\|\vec{f} - \vec{s}^{\,k}\|_2^2 + \dfrac{\tau}{\alpha_k}\text{pen}(\vec{f}) \\ & & \text{subject to } \;\vec{f} \in \mathcal{S}, \end{array} \tag{7}$$

where $\vec{s}^{\,k} = [\vec{s}^{\,k}_{p_1}; \vec{s}^{\,k}_{p_2}; \vec{s}^{\,k}_i; \vec{s}^{\,k}_n] = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$ (see [24] for details). Note that $\mathcal{Q}(\vec{f})$ separates into the sum

$$\mathcal{Q}(\vec{f}) = \sum_{j=1}^{N} \mathcal{Q}_j\big((\vec{f}_{p_1})_j, \vec{f}_{p_2})_j, (\vec{f}_i)_j, (\vec{f}_n)_j\big),$$

where $\mathcal{Q}_j\colon \mathbb{R}^4 \to \mathbb{R}$ and

$$\begin{aligned} \mathcal{Q}_j\big((\vec{f}_{p_1})_j, (\vec{f}_{p_2})_j, (\vec{f}_i)_j, (\vec{f}_n)_j\big) &= \frac{1}{2}\Big\{\big((\vec{f}_{p_1} - \vec{s}^{\,k}_{p_1})_j\big)^2 \\ &+ \big((\vec{f}_{p_2} - \vec{s}^{\,k}_{p_2})_j\big)^2 + \big((\vec{f}_i - \vec{s}^{\,k}_i)_j\big)^2 + \big((\vec{f}_n - \vec{s}^{\,k}_n)_j\big)^2\Big\} \\ &+ \frac{\tau}{\alpha_k}\Big\{\;|(\vec{f}_{p_1})_j| + |(\vec{f}_{p_2})_j| + |(\vec{f}_i)_j| + \gamma|(\vec{f}_n)_j|\Big\}. \end{aligned}$$

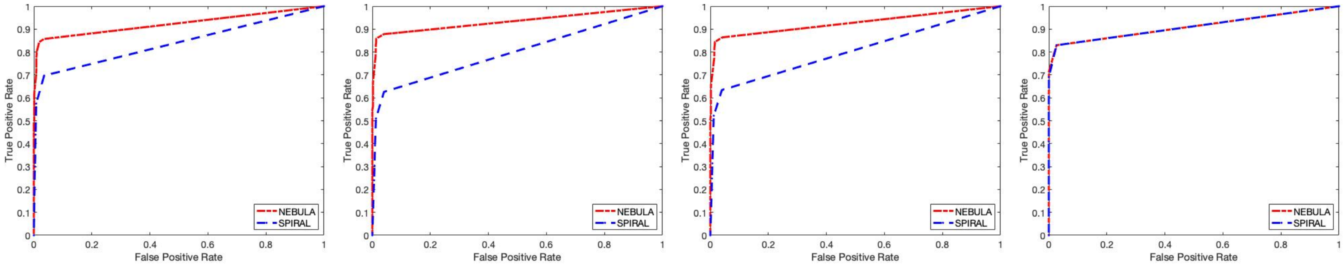
Note that the bounds for $\mathcal{S}$ are component-wise. Therefore, (7) separates into subproblems of the form

$$\begin{array}{ll} \underset{f_{p_1}, f_{p_2}, f_i, f_n \in \mathbb{R}}{\text{minimize}} & \dfrac{1}{2}\Big\{(f_{p_1} - s_{p_1})^2 + (f_{p_2} - s_{p_2})^2 \\ & \quad + (f_i - s_i)^2 + (f_n - s_n)^2\Big\} \\ & \quad + \dfrac{\tau}{\alpha_k}\Big\{|f_{p_1}| + |f_{p_1}| + |f_i| + \gamma|f_n|\Big\} \\ \text{subject to} & 0 \le f_i + f_n \le 1 \\ & 0 \le f_n \le 1 - f_{p_1}, \\ & 0 \le f_n \le 1 - f_{p_2}, \\ & f_{p_1} + f_{p_2} - 1 \le f_i \le f_{p_1} + f_{p_2} \\ & 0 \le f_{p_1}, f_{p_2}, f_i, f_n \le 1. \end{array} \tag{8}$$

where $\{f_{p_1}, f_{p_2}, f_i, f_n\}$ and $\{s_{p_1}, s_{p_2}, s_i, s_n\}$ are scalar components of the vectors $\{\vec{f}_{p_1}, \vec{f}_{p_2}, \vec{f}_i, \vec{f}_n\}$ and $\{\vec{s}_{p_1}, \vec{s}_{p_2}, \vec{s}_i, \vec{s}_n\}$, respectively, at the same location.

We solve (8) using an alternating block-coordinate descent approach, which utilizes alternating steps between child and parent variables [27]. We start by fixing the parent signals f_{p_1} and f_{p_2}, and solve the resulting minimization problem for the child signal, f_i and f_n. Next, we fix the child signal and minimize over the parent variables. We continue this method until the subsequent iterates falls below a specified threshold. The steps are as follows.

Step 0: Initially, we fix the values for the parent variables by setting $f_{p_1}^{(0)} = f_{p_2}^{(0)} = 0.5$ for each candidate SV location.



(a) All Individuals ROC (b) Parent 1 ROC (c) Parent 2 ROC (d) Child ROC

Fig. 3: ROC curves for reconstructions for data drawn from a negative binomial distribution, illustrating the true positive rate versus the false positive rate with 5% novel variants where $\tau = 0.1$ and $\gamma = 15$. (a) The area under the curve (AUC) for all the individuals combined for NEBULA is 0.9236 while the AUC for SPIRAL is 0.8382. (b) The AUC for Parent 1 for NEBULA is 0.9341 while the AUC for SPIRAL is 0.7993. (c) The AUC for Parent 2 for NEBULA is 0.9263 while the AUC for SPIRAL is 0.8037. (d) The AUC for the child for both NEBULA and SPIRAL are 0.9107.

Step 1: Suppose we have obtained $f_{p_1}^{(j-1)}$ and $f_{p_2}^{(j-1)}$ from the previous iteration. The child variables $f_i^{(j)}$ and $f_n^{(j)}$ are obtained by solving the following:

$$\begin{array}{ll} \underset{f_i, f_n \in \mathbb{R}}{\text{minimize}} & \frac{1}{2}(f_i - c_i)^2 + \frac{1}{2}(f_n - c_n)^2 \\ \text{subject to} & 0 \leq f_n, f_i \leq 1, \qquad 0 \leq f_i + f_n \leq 1, \\ & 0 \leq f_n \leq 1 - f_{p_1}^{(j-1)}, \quad 0 \leq f_n \leq 1 - f_{p_2}^{(j-1)}, \\ & f_{p_1}^{(j-1)} + f_{p_2}^{(j-1)} - 1 \leq f_i \leq f_{p_1}^{(j-1)} + f_{p_2}^{(j-1)}, \end{array}$$

where $c_i = s_i - \frac{\tau}{\alpha_j}$ and $c_n = s_n - \frac{\gamma\tau}{\alpha_j}$.

Step 2: Suppose we have obtained $f_i^{(j)}$ and $f_n^{(j)}$ from the previous step. We obtain the solution for the current iteration $f_{p_1}^{(j)}$ and $f_{p_2}^{(j)}$ are obtained by solving the following:

$$\begin{array}{ll} \underset{f_{p_1}, f_{p_2} \in \mathbb{R}}{\text{minimize}} & \frac{1}{2}(f_{p_1} - c_{p_1})^2 + \frac{1}{2}(f_{p_2} - c_{p_2})^2 \\ \text{subject to} & 0 \leq f_{p_1} \leq 1 - f_n^{(j)}, \quad 0 \leq f_{p_2} \leq 1 - f_n^{(j)}, \\ & f_{p_1} + f_{p_2} - 1 \leq f_i^{(j)} \leq f_{p_1} + f_{p_2}, \\ & 0 \leq f_{p_1}, f_{p_2} \leq 1, \end{array}$$

where $c_{p_1} = s_{p_1} - \frac{\tau}{\alpha_j}$ and $c_{p_2} = s_{p_2} - \frac{\tau}{\alpha_j}$.

We note that both steps have closed form solutions, which can be obtained by projecting the unconstrained solution to the corresponding feasible set (see [27] for details).

III. Results

We implemented our proposed method for variant detection called NEgative Binomial Optimization Using ℓ_1 Penalty ALgorithms (NEBULA) and compared its results to the SPIRAL method. Similar to previously published methods, we observed the variant predictions in a two-parent/one-child model [20]. Our method used a sparsity promoting parameter τ. This method has a second regularization parameter, γ, which was chosen to promote further sparsity within the novel variants, f_n. The methods were terminated if the relative difference between consecutive iterates $||\vec{f}^{\,k+1} - \vec{f}^{\,k}||_2/||\vec{f}^{\,k}||_2 \leq 10^{-8}$.

We studied the performance on data we simulated that match our assumptions. We simulated the true signal for the parents and child by creating the vector, $\vec{f}$ of size 10^6 and selecting 500 locations to be true variants for the parents 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. For the child signal, we made the assumption that if both parents have a SV at a particular location, the child does as well. However, if only one parent has a SV at a particular location, the child has a 50% chance of inherting that SV [27]. The novel variants in the child are chosen randomly from locations where the parents do not have a SV. We created our observed signals by sampling from a negative binomial distribution (Fig. 3) and a Poisson distribution (Fig. 4) based upon a given coverage and error.

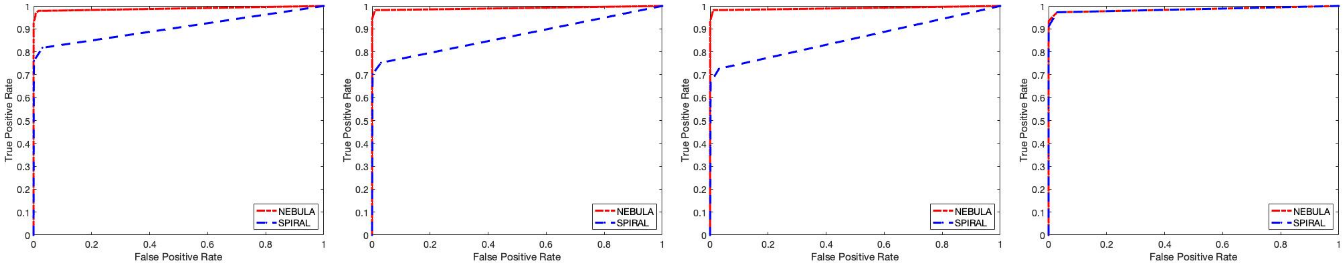
Analysis. Compared to our work in the one-parent/one-child model [28], we noticed a significant improvement with the AUCs for NEBULA over those for SPIRAL. Also, both NEBULA and SPIRAL produced higher AUCs when the data are drawn from a Poisson distribution than those drawn from a negative binomial distribution. Moreover, when considering the algorithms and the individuals we found both parents will have relatively the same level of reconstruction accuracy (i.e., Parents 1 and 2 have similar AUC values). Finally, we found that for the Parent 1 and Parent 2, the AUCs for NEBULA where higher than those from the existing SPIRAL method.

IV. Conclusions

We propose a method which builds on our previously developed negative binomial optimization method, which reconstructs signals arising from the negative binomial distribution rather than the Poisson distribution. This method aims to detect both inherited and novel variants within a child's genome when genomic information from both parents is available. Overall, our proposed method improves upon our previous work for predicting structural variants.

Acknowledgment

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



(a) All Individuals ROC (b) Parent 1 ROC (c) Parent 2 ROC (d) Child ROC

Fig. 4: ROC curves for reconstructions for data drawn from a negative binomial distribution, illustrating the true positive rate versus the false positive rate with 5% novel variants where $\tau = 0.1$ and $\gamma = 15$. (a) The area under the curve (AUC) for all the individuals combined for NEBULA is 0.9886 while the AUC for SPIRAL is 0.9046. (b) The AUC for Parent 1 for NEBULA is 0.9904 while the AUC for SPIRAL is 0.8708. (c) The AUC for Parent 2 for NEBULA is 0.9903 while the AUC for SPIRAL is 0.8571. (d) The AUC for the child for for NEBULA is 0.9852 while the AUC for SPIRAL 0.9848.

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. 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.

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

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

[23] 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.

[24] 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.

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

[26] 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.

[27] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi, "Predicting novel and inherited variants in parent-child trios," in *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2019, pp. 1–6.

[28] A. Lazar, M. Banuelos, S. Sindi, and R. F. Marcia, "Detecting novel genomic structural variants through negative binomial optimization," in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, 2020, pp. 511–515.