Genomic Signal Processing for Variant Detection in Diploid Parent-Child Trios

Melissa Spence
Department of Applied Mathematics
University of California, Merced
Merced, USA
mspence5@ucmerced.edu

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

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

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

***Abstract*—Structural variants (SVs) are rearrangements in the DNA sequence of members within the same species. Detecting SVs is challenging because most approaches suffer from high-false positive rates. In this work, we improve the accuracy of SV detection by exploiting familial relationships and the rare occurence of these rearrangements. Mathematically, we pose SV detection as a constrained optimization problem regularized by a sparsity promoting term. Furthermore, we generalize our previous methods in two ways. First, we consider a biologically realistic scenario of a parent-child-trio, where each individual may carry zero, one, or two copies of any potential SV. Second, we employ a novel block-coordinate descent approach with orthogonal projection to efficiently minimize the objective and to enforce feasibility within the biological constraint space. Numerical results using both simulated and real trios demonstrate that our proposed approach improves our ability to separate true SVs from false positives.**

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

I. Introduction

The genome, the complete DNA sequence of an organism, consists of one (or more) lengthy sequences of nucleic acids represented by the letters A,C,G and T. Some species are diploid and inherit a complete genome from both parents whereas others (such as bacteria) are haploid with only a single copy of their genome. It was originally thought the genome was essentially identical, except for some single letter differences (single nucleotide variants, SNVs), between individuals. Today we understand that longer differences are common and the term structural variant (SV) has come to represent differences (typically > 50 letters) between genomes of the same species [1]. While the role of most SVs is unknown, many have been implicated in human disease [2].

SVs are typically identified through comparisons between an unknown genome and a given reference through a two phase process of sequencing and mapping. In the sequencing phase, fragments are sampled from the unknown genome. The types of samples drawn from the unknown genome depend on the specific DNA sequencing technology. The ends of the fragments are then sequenced, that is the DNA letters at either end are determined. This mapping process attempts to find the true position of the samples in the reference genome. SVs are identified through analysis of the mapped arrangements. For example, paired-ends which map too far apart in the reference indicate potential deletions. Many tools have been developed for SV identification, as reviewed in [3]. However, such methods are prone to erroneous predictions from errors in the sequencing and mapping process. When the sequencing coverage, expected number of samples containing each genome position, is high erroneous SV predictions are more easily separated from true SVs. However, increasing coverage greatly increases the cost of the sequencing experiment. In our work, we take an orthogonal approach to SV detection by simultaneously predicting SVs in multiple related individuals.

In large scale studies, such as the 1000 Genomes Project, the genomes of many related individuals are simultaneously sequenced including a number of parent-child trios. Because the *de novo* rate of formation of SVs is very low, it is expected that the vast majority of SVs in the genome of a child will be present in the genome of one of their parents. Moreover, if the child's genome does have an SV the number of copies will depend on the number of copies in the genomes of their parents because the child inherits a copy of their genome from each parent. In prior work, we have simultaneously considered the genome of parents and children [4], [5]. However, for mathematical and computational simplicity we considered approximations to the true biological system by assuming each individual has only one copy of their genome [6] or when allowing each individual to have two copies considering only one parent [5]. In this work, we consider simultaneous SV prediction in a diploid parent-child trio. We use a novel block-coordinate descent approach to enforce biological feasibility while minimizing our objective over a 6-dimensional solution

space and promote sparsity of SVs with an ℓ_1-norm.

II. Method

We consider a framework for refining structural variant (SV) recovery signals for multiple related individuals. This work considers diploid data from one father (F), one mother (M), and one child (C). We assume that each signal consists of n locations in the genome where an SV may occur. Humans have two copies of each chromosome, one inherited from each parent. If both parents have an SV at the same location, this impacts the probability that the child also has an SV at the same location. For each individual i in our model, we consider two signals that take on binary values: a heterozygous indicator $\vec{y}_i \in \{0,1\}^n$ and a homozygous indicator $\vec{z}_i \in \{0,1\}^n$. The heterozygous vector is an indicator that the individual has one copy of the SV while the homozygous vector indicates that the individual has two copies of the SV. If an individual is heterozygous for an SV at position j, 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 $(\vec{y}_i)_j = 0$.

A. Observation Model

The observed data are the number of DNA fragments supporting each potential SV. In particular, we denote the observation vectors for the parents (father and mother) and child by the vectors $\vec{s}_F \in \mathbb{R}^n, \vec{s}_M \in \mathbb{R}^n$, and $\vec{s}_C \in \mathbb{R}^n$, respectively. We assume the data follow a Poisson distribution ([7], [8]):

$$\begin{bmatrix} \vec{s}_C \\ \vec{s}_F \\ \vec{s}_M \end{bmatrix} \sim \text{Poisson} \left(\begin{bmatrix} z_C(2\lambda_C - \epsilon) + y_C(\lambda_C - \epsilon) + \epsilon \\ z_F(2\lambda_F - \epsilon) + y_F(\lambda_F - \epsilon) + \epsilon \\ z_M(2\lambda_M - \epsilon) + y_M(\lambda_M - \epsilon) + \epsilon \end{bmatrix} \right), \quad (1)$$

where λ_C, λ_F, and λ_M are the sequencing coverage of the child, father, and mother, respectively, and $\epsilon > 0$ (see [9]). The parameter ϵ is reflective of measurement errors corresponding to the sequencing and mapping process. These errors are a large hindrance to accurate SV discovery methods and lead to a high false-positive discovery rate.

Letting

$$\vec{s} = \begin{bmatrix} \vec{s}_C \\ \vec{s}_F \\ \vec{s}_M \end{bmatrix}, \ \vec{z} = \begin{bmatrix} \vec{z}_C \\ \vec{z}_F \\ \vec{z}_M \end{bmatrix}, \ \vec{y} = \begin{bmatrix} \vec{y}_C \\ \vec{y}_F \\ \vec{y}_M \end{bmatrix}, \text{ and } \vec{f} = \begin{bmatrix} \vec{z} \\ \vec{y} \end{bmatrix},$$

we note that $\vec{f} \in \{0,1\}^{6n}$. Our general observation model (1) can be expressed as

$$\vec{s} \sim \text{Poisson}\left(Af + \epsilon \mathbf{1}\right),$$

where $\mathbf{1} \in \mathbb{R}^{3n}$ is the vector of ones and $A = [A_1 \ \ A_2] \in \mathbb{R}^{3n \times 6n}$ is the coverage matrix with

$$A_1 = \begin{bmatrix} (2\lambda_C - \epsilon) I_n & 0 & 0 \\ 0 & (2\lambda_F - \epsilon) I_n & 0 \\ 0 & 0 & (2\lambda_M - \epsilon) I_n \end{bmatrix}$$

and

$$A_2 = \begin{bmatrix} (\lambda_C - \epsilon) I_n & 0 & 0 \\ 0 & (\lambda_F - \epsilon) I_n & 0 \\ 0 & 0 & (\lambda_M - \epsilon) I_n \end{bmatrix}.$$

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

B. Problem Formulation

Assuming a Poisson process to model the noise in the measurements [10], the probability of observing the observation vector $\vec{s}$, given the true signal $\vec{f}$, is given by

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

We use the maximum likelihood principle to determine the unknown Poisson parameter $A\vec{f}$ such that the probability of observing the vector of Poisson data $\vec{s}$ in (2) is maximized. Specifically, we minimize the corresponding convex negative Poisson log-likelihood function

$$F(\vec{f}) = \sum_{j=1}^{3n} (A\vec{f})_j - \vec{s}_j \log((A\vec{f})_j + \epsilon). \quad (3)$$

To minimize $F(\vec{f})$, we apply a continuous relaxation of the variables and use gradient-based methods. Specifically, we let that the values of $\vec{f}$ to lie between 0 and 1, i.e., $\mathbf{0} \leq \vec{f} \leq \mathbf{1}$, or equivalently,

$$\mathbf{0} \leq \vec{z}_i, \vec{y}_i \leq \mathbf{1}, \quad (4)$$

where $\mathbf{0}$ is the vector of zeros, $i \in \{C, F, M\}$, and the inequalities are to be understood component-wise. We note that since a variant cannot be both heterozygous and homozygous simultaneously, we require further that

$$\mathbf{0} \leq \vec{z}_i + \vec{y}_i \leq \mathbf{1}. \quad (5)$$

C. Familial Constraints

We incorporate additional constraints that exploit information about the signal $\vec{f}$ to help improve the accuracy of our SV predictions. The constraints control for biological realities in each individual as well as constraints from the relatedness of individuals.

First, if one of the parents is homozygous for an SV at location j, i.e., $(\vec{z}_F)_j = 1$ or $(\vec{z}_M)_j = 1$, then the child must be at least heterozygous, i.e., $(\vec{z}_C)_j + (\vec{y}_C)_j = 1$. This means that

$$\begin{array}{ccccc} \mathbf{0} & \leq & \vec{z}_F & \leq & \vec{z}_C + \vec{y}_C \\ \mathbf{0} & \leq & \vec{z}_M & \leq & \vec{z}_C + \vec{y}_C. \end{array}$$

These constraints indicate that if the child does not have an SV in a particular location, then neither parent can have a homozygous SV at that location.

Second, the child can only be homozygous, i.e., $(\vec{z}_C)_j = 1$, if both of the parents are at least heterozygous, i.e., $(\vec{z}_F)_j + (\vec{y}_F)_j = 1$ and $(\vec{z}_M)_j + (\vec{y}_M)_j = 1$. Furthermore, the child must be homozygous if both parents are homozygous, i.e.,

$$\max\{\vec{z}_F + \vec{z}_M - \mathbf{1}, \mathbf{0}\} \leq \vec{z}_C \leq \min\{\vec{z}_F + \vec{y}_F, \vec{z}_M + \vec{y}_M\},$$

where $\max\{\cdot,\cdot\}$ and $\min\{\cdot,\cdot\}$ are to be understood componentwise.

Finally, the child can only be heterozygous if at lest one of the parents is at least heterozygous, and the child cannot have an SV if neither parent has an SV, i.e.,

$$\mathbf{0} \leq \vec{y}_C \leq \min\{\vec{z}_F + \vec{y}_F + \vec{z}_M + \vec{y}_M, \mathbf{1}\}.$$

We denote the set of all vectors satisfying these constraints by $\mathcal{S}$, i.e.,

$$\mathcal{S} = \left\{ \begin{bmatrix} \vec{z}_C \\ \vec{z}_F \\ \vec{z}_M \\ \vec{y}_C \\ \vec{y}_F \\ \vec{y}_M \end{bmatrix} \in \mathbb{R}^{6n} : \begin{aligned} \mathbf{0} \leq \vec{z}_i + \vec{y}_i \leq \mathbf{1},\ \mathbf{0} \leq \vec{z}_F \leq \vec{z}_C + \vec{y}_C, \\ \mathbf{0} \leq \vec{z}_M \leq \vec{z}_C + \vec{y}_C, \\ \max\{\vec{z}_F + \vec{z}_M - \mathbf{1}, \mathbf{0}\} \leq \vec{z}_C, \\ \vec{z}_C \leq \min\{\vec{z}_F + \vec{y}_F, \vec{z}_M + \vec{y}_M\}, \\ \mathbf{0} \leq \vec{y}_C \leq \min\{\vec{z}_F + \vec{y}_F + \vec{z}_M + \vec{y}_M, \mathbf{1}\} \end{aligned} \right\}.$$

D. Optimization Formulation

A common difficulty with SV recovery is predicting false positive SVs by mistaking fragments that are incorrectly mapped against the reference genome. Since SVs are rare in an individual's genome, we enforce sparsity in our predictions by incorporating an ℓ_1-norm penalty term in our objective function (see [11]). Our objective function takes the following form:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{6m}}{\text{minimize}} \quad & F(\vec{f}) + \tau ||\vec{f}||_1 \\ \text{subject to} \quad & \vec{f} \in \mathcal{S} \end{aligned} \tag{6}$$

where $F(\vec{f})$ is the negative Poisson log-likelihood function shown in (3) and $\tau > 0$ is a regularization parameter. We then use a second-order Taylor series approximation around the current iterate $\vec{f}^k$ to formulate a sequence of quadratic subproblems. In this approach, we approximate the Hessian matrix by a scalar multiple of the identity matrix, $\alpha_k I$, where $\alpha_k > 0$ (see [12] for details) for how to compute α_k), and define the function

$$F^k(\vec{f}) = F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^T \nabla F(\vec{f}^k) + \frac{\alpha_k}{2} ||\vec{f} - \vec{f}^k||_2^2, \tag{7}$$

which we use as a surrogate function for $F(\vec{f})$ in (6). This approximation leads to the following equivalent subproblem formulation:

$$\begin{aligned} \vec{f}^{k+1} = \underset{\vec{f} \in \mathbb{R}^{6n}}{\arg\min} \quad & \frac{1}{2} ||\vec{f} - \vec{r}^k||_2^2 + \gamma ||\vec{f}||_1 \\ \text{subject to} \quad & \vec{f} \in \mathcal{S} \end{aligned} \tag{8}$$

where $\vec{r}^k = \vec{f}^k - \frac{1}{\alpha_k} \nabla F(\vec{f}^k)$ and $\gamma = \frac{\tau}{\alpha_k}$. This approach is based on [13], [14]. Note that the objective function in (8) is separable in f. Thus, (8) can be solved in batches. In particular, at each candidate SV position, we solve

$$\begin{aligned} f^{k+1} = \underset{f \in \mathbb{R}^6}{\arg\min} \quad & \frac{1}{2} ||f - r^k||_2^2 + \gamma ||f||_1 \\ \text{subject to} \quad & f \in S \end{aligned} \tag{9}$$

where the vectors $r^k = [r^k_{z_C}; r^k_{z_F}; r^k_{z_M}; r^k_{y_C}; r^k_{y_F}; r^k_{y_M}]$ and $f = [z_C; z_F; z_M; y_C; y_F; y_M]$ correspond to the components of $\vec{r}^k$ and $\vec{f}$, respectively, and the set S is similar to the feasible set $\mathcal{S}$ but restricted to the particular candidate SV position.

E. Optimization Approach

Here we propose solving our problem using a block-coordinate descent approach. Following methods used in previous work (see [5]), we fix all but one individual and solve (9) over both indicator variables for that individual. In subsequent steps, the variables corresponding to some other individual are minimized while the other individuals signals are fixed. This block-coordinate descent approach continues until the iterates satisfy a pre-determined convergence criteria.

Step 0: First, we compute the unconstrained minimizer of (9), which is given by $\hat{f}^{(0)} = r^k - \gamma \mathbf{1}$. Then we initialize the parent indicator variables by $\hat{z}_I^{(0)} =| \{0, r^k_{z_I} - \gamma, 1\}$ and $\hat{y}_I^{(0)} =| \{0, r^k_{y_I} - \gamma, 1\}$, where $I \in \{F, M\}$ and $\text{mid}\{\cdot,\cdot,\cdot\}$ takes on the value that is in the middle to ensure that the constraint in (4) is satisfied. To ensure that the constraint in (5) is satisfied, if $\hat{z}_F^{(0)} + \hat{y}_F^{(0)} > 1$, then we let $\hat{z}_F^{(0)} = \hat{y}_F^{(0)} = 0.5$. We adjust $\hat{z}_M^{(0)}$ and $\hat{y}_M^{(0)}$ similarly. To initialize the child indicator variables we let $\hat{z}_C^{(0)} = r^k_{z_C} - \gamma$ and $\hat{y}_C^{(0)} = r^k_{y_C} - \gamma$. We initialize the index with $i = 1$.

Step 1: Once we have obtained estimates for both parents' diploid indicator variables, $\hat{z}_F^{(i-1)}, \hat{y}_F^{(i-1)}, \hat{z}_M^{(i-1)}$ and $\hat{y}_M^{(i-1)}$, from the previous iteration, we project $\hat{z}_C^{(i-1)}$ and $\hat{y}_C^{(i-1)}$ onto the feasible set S with fixed parent variables to obtain the new child indicator variables $\hat{z}_C^{(i)}$ and $\hat{y}_C^{(i)}$. This projection is similar to the projections done in [5]. The feasible region for this step is illustrated in Fig. 1(a).

Step 2: After obtaining the new estimates for the child's diploid indicator variables, $\hat{z}_C^{(i)}$ and $\hat{y}_C^{(i)}$, from Step 1, we project $\hat{z}_F^{(i-1)}$ and $\hat{y}_F^{(i-1)}$ onto our feasible set S with fixed child and mother indicator variables to obtain the new father indicator variables $\hat{z}_F^{(i)}$ and $\hat{y}_F^{(i)}$. This projection is also similar to the projections done in [5]. The feasible region for this step is similar to that illustrated in Fig. 1(b).

Step 3: After obtaining the new estimates for the father's diploid indicator variables, $\hat{z}_F^{(i)}$ and $\hat{y}_F^{(i)}$, from Step 2, we project $\hat{z}_M^{(i-1)}$ and $\hat{y}_M^{(i-1)}$ onto our feasible set S with fixed child and father indicator variables to obtain the new mother indicator variables $\hat{z}_M^{(i)}$ and $\hat{y}_M^{(i)}$. This projection is also similar to the projections done in [5]. The feasible region for this step is illustrated in Fig. 1(b).

Steps 1, 2 and 3 are repeated in an alternating cycle until some convergence criteria are satisfied. In our numerical experiments, we saw that iterates did not change after three cycles. Thus, we terminated each cycle after three iterations. Note that Steps 2 and 3 are equivalent and result in identical feasible regions.

III. RESULTS

A. Simulated Data

Before applying our method to real human data, we first tested the performance on simulated data to match our assumptions. To do this we simulated two parent signals with a

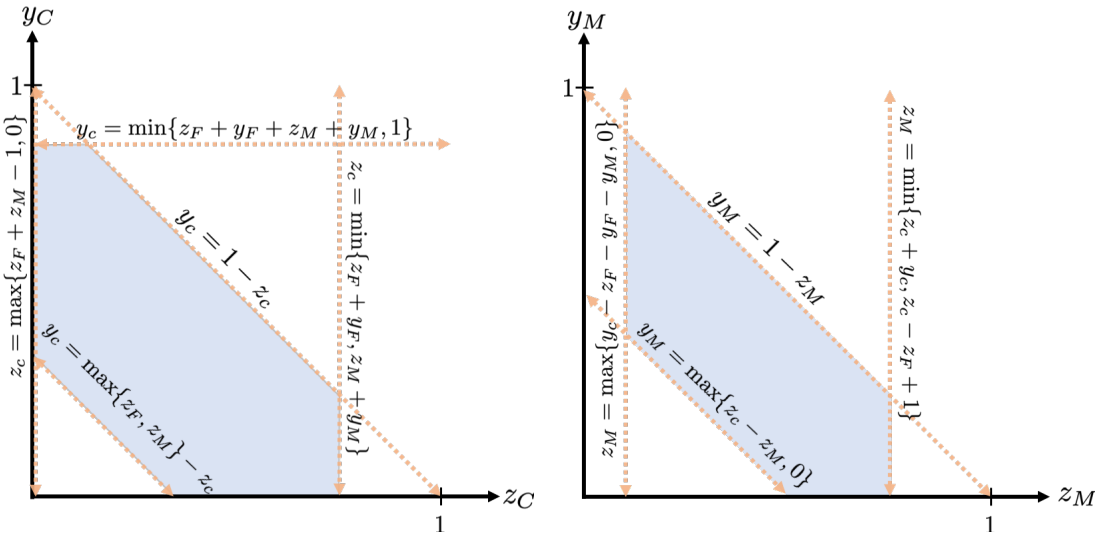

Fig. 1: The feasible set (shown above by the shaded region) for each step of the proposed block-coordinate minimization approach. (a) In Step 1, we obtain the solution for the child's variables z_C and y_C given fixed parent indicator variables z_F, y_F, z_M and y_M. (b) In Step 3, we obtain the solution for the mother's variables z_M and y_M given fixed indicator variables z_C, y_C, z_F and y_F. The feasible set represented in Step 2 is similar to that in Step 3.

set number of structural variants and a set similarity between the parent signals. The simulated true signals all consisted of 10^5 potential SVs. In the parent signals 500 locations were chosen at random to be variants; the percentage of variant sites the parents had in common was varied for testing. We then formed the child signal using a logical implementation of inheritance. If both parents were homozygous for an SV at position j then the child is homozygous for an SV at position j. If one parent was homozygous for an SV at position j and the other parent was heterozygous for an SV at position j then the child was at least heterozygous for an SV at that position, and had a 50% chance of being homozygous for an SV at position j. After forming the true signals for each individual, the observed signals were created by sampling from the Poisson distribution with a given coverage and error.

Analysis. Given an optimal τ value, our method is better able to reconstruct the homozygous signals for each individual. In Figure 2 we show an ROC curve generated for a simulated data set where the parents share 90% of their SVs and 30% of their SVs are homozygous. The area under the curve for each signal recovered from our method is greater than that of our previous diploid model which only includes information from one parent and one child [5]. We found that given optimal τ we were able to better recover not only the child signal, but also each of the parents as compared to our previous method.

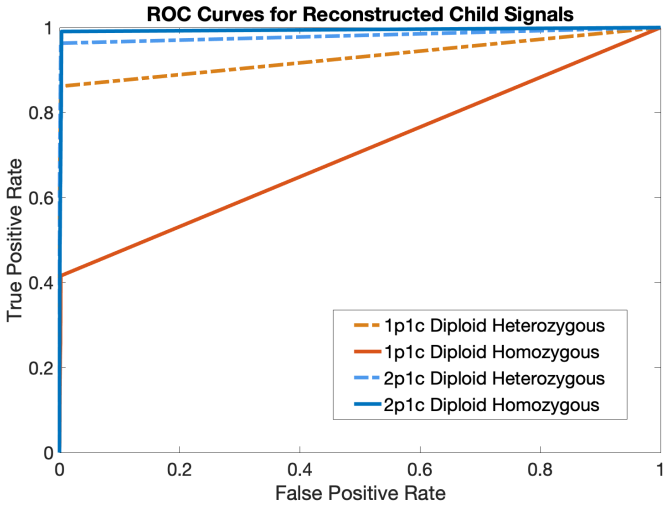

Fig. 2: ROC curves of two methods illustrating the false positive rate vs. the true positive rate in the child reconstruction broken into the heterozygous signal and the homozygous signal, where $\tau = 150$, the parents share 90% of their SVs and 30% of each parents SVs are homozygous. The coverage values for each individual are as follows $(\lambda_C, \lambda_F, \lambda_M) = (5, 10, 10)$.

B. 1000 Genomes Project Trio Data

We next apply our diploid method to the 1000 Genomes Project CEU trio data [15]. The father-mother-daughter (NA12891-12892-12878) trio was sequence at approximately $4\times$ coverage and structural variants were experimentally validaded for these individuals. To create $\vec{z}$ and $\vec{y}$, we filter *LowQual* predictions and incorporated the genotype to separate heterozygous from homozygous reported deletions. Moreover, we only consider deletions longer than 250bp in the experimentally validated set.

Analysis. For each CEU genome, there are $n = 57,078$ candidate deletion locations. Of these GASV predictions, 686, 637, and 724 are validated deletions (heterozygous and homozygous combined) in the father, mother, and child, respectively. Whereas our previous method fixes one individual at a time, our new method simultaneously predicts all three individuals while improving the heterozygous signal reconstruction for the child (see Fig. 3). Moreover, we see comparable performance for the reconstruction of both heterozygous and homozygous signals for both parents. Fig. 4 is representative of the slightly improved predictions for the parent signals for varying values of τ.

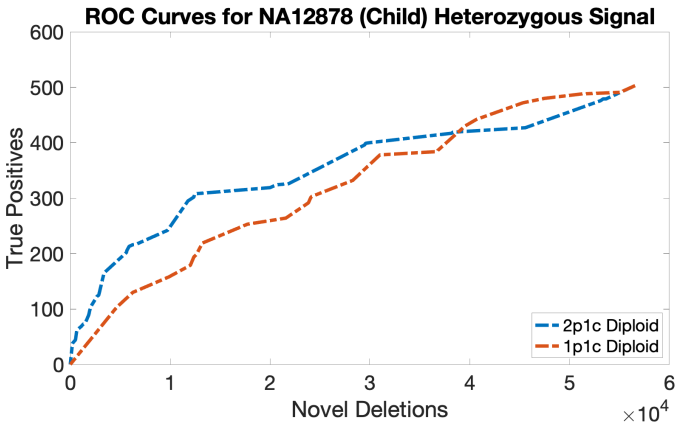

Fig. 3: ROC curves for the reconstruction of the heterozygous child signal, $\vec{y}_C$, where $\lambda_C = \lambda_F = \lambda_M = 4$, $\tau = 1 \times 10^{-4}$, and $\epsilon = 0.01$. Since the validated set may not contain all true deletions, we plot novel deletions against validated true positives. We observe a considerable improvement in the detection of true positives with our proposed method.

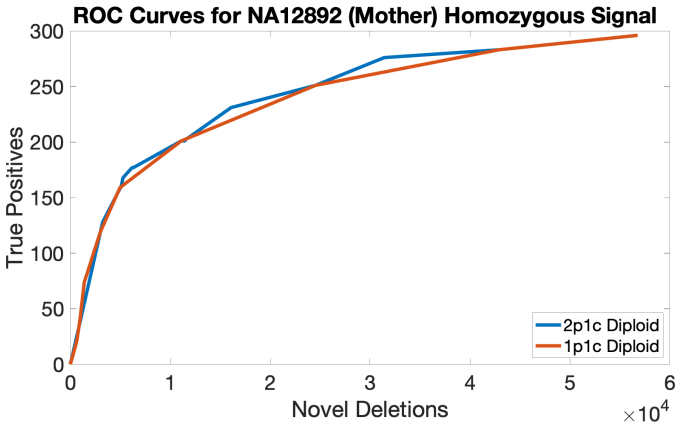

Fig. 4: ROC curves for the reconstruction of the homozygous mother signal, $\vec{z}_M$, where the coverage is approximately 4 $\times$ for all individuals, $\tau = 1 \times 10^{-4}$, and $\epsilon = 0.01$. We note a marginal improvement over our previous method in this reconstruction.

IV. Conclusions

We present an optimization method to detect SVs in sequencing data from parent-child trios. This method leverages relatedness between the individuals to improve signal reconstruction of noisy data. This extends previous work that focused on diploid signals from one parent and a child. We present results for both simulated and real data from the 1000 Genomes Project. We demonstrate that we are able to capture variants for which the individual possessed two copies. In future studies we intend to apply this work to a multi-generational framework with multiple offspring.

Acknowledgment

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

References

[1] S. S. Ho, A. E. Urban, and R. E. Mills, "Structural variation in the sequencing era," *Nature Reviews Genetics*, pp. 1–19, 2019.

[2] P. Stankiewicz and J. R. Lupski, "Structural variation in the human genome and its role in disease," *Annual review of medicine*, vol. 61, pp. 437–455, 2010.

[3] 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. 117, 2019.

[4] 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)*, Dec 2018, pp. 943–950.

[5] 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.

[6] 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.

[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. Snyder, *Random Point Processes*, Wiley-Interscience, New York, NY, 1975.

[9] 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*.

[10] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, "The database of genomic variants: a curated collection of structural variation in the human genome," *Nucleic acids research*, vol. 42, no. D1, pp. D986–D992, 2013.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

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

[13] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "Sparse Poisson intensity reconstruction algorithms," in *Proceedings of IEEE Statistical Signal Processing Workshop*, Cardiff, Wales, UK, September 2009.

[14] 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.

[15] 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.