Predicting Novel and Inherited Variants in 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*—Genomic variation shared by members of the same species that are longer than a single nucleotide are commonly called structural variants (SVs). Though relatively rare, they represent an increasingly important class of variation as SVs have been associated with diseases and susceptibility to some types of cancer. Common approaches to SV detection require the sequencing and mapping of fragments from a test genome to a high-quality reference genome. Candidate SVs correspond to fragments with discordant mapped configurations, but because errors in the sequencing and mapping will also create discordant arrangements, many of these predictions will be false. When sequencing coverage is low, distinguishing true SVs from errors is even more complicated. In recent work, we have developed SV detection methods that simultaneously consider the genomes of closely related individuals – parents and children. Our approaches control false positive SVs by requiring children inherit all SVs in their genome from a parent. However, in doing so, our models may have missed true novel variants acquired by the child. In this work, we generalize our previous approaches to allow the child to carry novel variants but enforce sparsity through an ℓ_1 penalty (since novel SVs in the child should be rare). We present results on both simulated genomes as well as two-sequenced parent-child trios from the 1000 Genomes Project.**

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

I. Introduction

The complete DNA sequence of an organism (the genome) is one or more ordered linear sequences of the letters A, C, G, or T. The total genome length is anywhere from millions (for bacteria) to billions (for mammals) of letters. Every cell in most multi-cellular organisms contains a complete and nearly identical copy of an organism's genome. When cells divide, the genome must be duplicated so each cell will have its own copy, but every time the genome is copied there is the opportunity for mutational processes to introduce variation. Genomic variation may consist of a modification to a single letter, termed single nucleotide variants (SNVs), or rearrangements of larger regions, termed structural variants (SVs) [1], [2]. For multi-cellular organisms, variants are often further classified into those which transmitted from parents to progeny, germline variants, or those which occur during cell division in the lifetime of an organism, somatic variants [3]. In humans, the accumulation of somatic mutations is commonly associated with the development of cancer [4] while the presence of certain germline variants have been shown to increase the susceptibility for certain types of cancer [5], [6]. Beyond cancer, genomic variants are associated with many significant biological outcomes for individuals including a variety of diseases in humans [7], [8], flowering behavior in plants [9] and have contributed to rates of adaptation and the emergence of new species [10].

The detection of genomic variants such as SVs remains a challenging scientific and computational problem. Even with modern DNA sequencing technologies, it is not possible to construct the complete genome of every cell. As such, the common practice has been to construct a high-quality reference genome for each species and then annotate this reference with sites of variation [11], [12]. The dominant method for identifying SNVs or SVs involves comparing fragments of DNA sequenced from a test (unknown) genome to a given reference [13], [14]. SVs are typically detected through indirect evidence – a fragment that maps to a larger than expected distance – and as such they are more difficult to identify than SNVs which may be directly observed through alignment of sequences from the test genome to the reference. However, predicting either type of variant is complicated by DNA sequencing and alignment errors (see Fig. 1). Because of these errors, algorithms for variant detection have suffered from high false-positive rates especially when the coverage – expected number fragments supporting each variant – is low [14], [13]. One hope for improving the ability to accurately predict SVs has come from methods that combine the information of many individuals [15]. This allows researchers to leverage

large-scale public efforts, such as the 1000 Genomes Project, that have made available sequencing data from thousands of individuals, including parent-child trios [16], [17]. Population level algorithms have the potential to improve variant detection because the signal of true SVs will be boosted, but only when variants are likely to be shared among multiple individuals. But, because of the massive population expansion, many variants in humans are rare and, as such, may only be shared by close relatives [18].

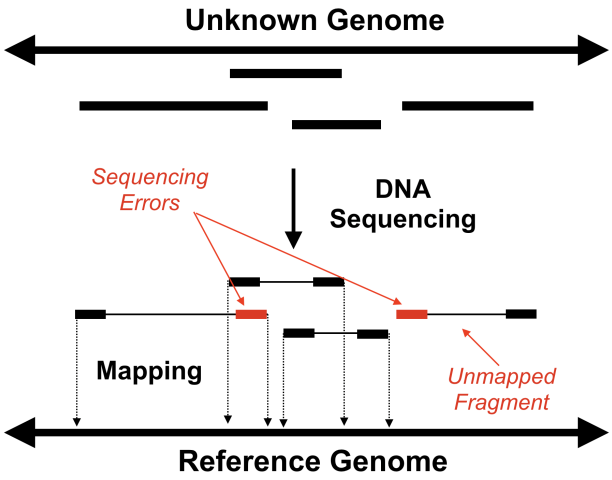

Fig. 1: Illustration of DNA sequencing and mapping process, beginning with an unknown genome. The unknown genome is fragmented and the ends of the fragments are sequenced. The sequenced reads are then compared and mapped to the reference genome. Measurement errors in both of these processes are depicted in red.

Our group has developed computational methods to improve SV prediction through considering pedigrees of related individuals [19], [20], [21], [22]. Our previous methods constrained the set of potential SVs through parent-child relationships by requiring that every variant present in the child was a germline variant transmitted from a parent. While these approaches have improved the ability to reduce false-positive predictions, they also increase the false-negative rate because they do not allow for novel variants (SVs that are not inherited from a parent) in the child genome.

This work improves upon our previous methods by allowing the child genome to possess novel variants. In Section 2, we develop our mathematical model and optimization framework for SV prediction in the context of one parent and one child. For simplicity in this work, we develop our model for haploid genomes so that at each potential SV site each individual either has the variant or does not. We consider a continuous relaxation of this discrete problem, but favor sparse solutions through the use of the ℓ_1 norm. In Section 3, we present the results of our method on both simulated and real sequencing data for parent-child trios from the 1000 Genomes Project. Even with our simplified haploid genome assumption, we find that our framework has the potential to improve SV prediction for low-coverage individuals.

II. Method

We now describe a structural variant (SV) detection framework given genomic data from both parents (father, F, and mother, M) and from one child (C). Let $\vec{f}^*_\iota \in \mathbb{R}^m$ be the vector of m locations of potential SVs for each individual $\iota \in \{F, M, C\}$. We make the following assumptions:

- **Inherited variants:** The variants in the child primarily come from the parents. In particular, if both parents have an SV at a particular location, the child must also have an SV at that location. Furthermore, we assume SVs are rare.
- **Novel variants:** On rarer occasions, the child may have variants not present in either parent.
- **Haploid genotype:** For simplicity, we consider each individual to be haploid (only one copy of each chromosome).
- **Low-coverage sequencing:** The expected number of fragments supporting each variant is low, and the observed measurements are governed by a Poisson process.

We denote the true SV signal for either parent, $P \in \{F, M\}$, by $\vec{f}^*_P \in \{0,1\}^m$, which has either a 1 at position j if the parent has an SV at location j or 0 otherwise. In contrast, the true SV signal for the child, $\vec{f}^*_C \in \{0,1\}^m$, is composed of two vectors:

$$\vec{f}^*_C = \vec{f}^*_I + \vec{f}^*_N,$$

where $\vec{f}^*_I \in \{0,1\}^m$ and $\vec{f}^*_N \in \{0,1\}^m$ denote the vector of SVs that are inherited from either parent and are novel, respectively. In particular, $(\vec{f}^*_I)_j$ has either a 1 if an SV is inherited from the parent at position j or a 0 otherwise. Similarly, $(\vec{f}^*_N)_j$ has a 1 if and only if there is a variant at position j that is not inherited from the parent and 0 otherwise. We note that at each location, $\vec{f}_I$ and $\vec{f}_N$ cannot simultaneously be both non-zero since a child variant can only either be inherited or be novel, but not both. In other words, the vectors $\vec{f}^*_I$ and $\vec{f}^*_N$ satisfy the complementary condition $(\vec{f}^*_I)_j(\vec{f}^*_N)_j = 0$ for $1 \leq j \leq m$.

Observation model. We denote the vector of observations and the vector of true SV signals by $\vec{y} = [\vec{y}_C; \vec{y}_F; \vec{y}_M]$ and $\vec{f}^* = [\vec{f}^*_I; \vec{f}^*_N; \vec{f}^*_F; \vec{f}^*_M]$, where the entries in the measurement vector $\vec{y}_\iota$ correspond to the number of DNA fragments supporting each potential SV, and the vector $\vec{y}_\iota \in \mathbb{R}^m$, where $\iota \in \{C, F, M\}$, is the observation vectors for each individual. Because we assume that the sequence coverage is low, we expect that the number of fragments covering any position in the genome to follow a Poisson distribution (see e.g., [13], [23]). In particular, we can express the general observation model as

$$\vec{y} \sim \text{Poisson}(\mathbf{A}\vec{f}^* + \epsilon\mathbf{1}), \tag{1}$$

where $\mathbf{1} \in \mathbb{R}^{3m}$ is the vector of ones and $\mathbf{A} \in \mathbb{R}^{3m \times 4m}$ is the coverage matrix given by

$$\mathbf{A} = \begin{bmatrix} (\lambda_C - \epsilon) I_m & (\lambda_C - \epsilon) I_m & 0 & 0 \\ 0 & 0 & (\lambda_F - \epsilon) I_m & 0 \\ 0 & 0 & 0 & (\lambda_M - \epsilon) I_m \end{bmatrix},$$

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

Problem formulation. We use the *maximum likelihood principle* to determine $\vec{f}^*$ such that the probability of observing the vector of Poisson data $\vec{y}$ in (1) is maximized. More precisely, we minimize the corresponding *negative Poisson log-likelihood* function

$$\Phi(\vec{f}) = \sum_{j=1}^{3m} \left\{ \left(\mathbf{A}\vec{f} \right)_j - \vec{y}_j \log \left((\mathbf{A}\vec{f})_j + \epsilon \right) \right\}.$$

To apply gradient-based optimization approaches for minimizing $\Phi(\vec{f})$, we allow $\vec{f}$ to take on more than the binary values of 0 and 1 and instead be continuous in the interval $[0, 1]$.

Feasibility constraints. We impose the following constraints on the SV signal estimate $\vec{f}$, which correspond to the biological assumptions we make:

- Since each entry in an individual's SV signal is binary, i.e., $\vec{f}_C^*, \vec{f}_F^*, \vec{f}_M^* \in \{0,1\}^m$, and since $\vec{f}_C^* = \vec{f}_I^* + \vec{f}_N^*$ with $\vec{f}_I^*, \vec{f}_N^* \in \{0,1\}^m$, then we have $\mathbf{0} \leq \vec{f}_I, \vec{f}_N, \vec{f}_F, \vec{f}_M \leq \mathbf{1}$ and $\mathbf{0} \leq \vec{f}_I + \vec{f}_N \leq \mathbf{1}$.
- Because a novel variant in the child cannot be inherited from either parent, we have $\mathbf{0} \leq \vec{f}_N \leq \mathbf{1} - \vec{f}_F$ and $\mathbf{0} \leq \vec{f}_N \leq \mathbf{1} - \vec{f}_M$.
- If both parents have an SV, then the child must inherit the same SV: $\vec{f}_F + \vec{f}_M - 1 \leq \vec{f}_I$. Similarly, if neither parent has an SV, then the child cannot have an inherited SV: $\vec{f}_I \leq \vec{f}_F + \vec{f}_M$.

We will denote the set of $\vec{f}$ satisfying these constraints by $\mathcal{F}$.

Optimization setup. With these components defined, the genomic variants reconstruction problem has the following constrained optimization form:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{4m}}{\text{minimize}} \quad & \Phi(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} \quad & \vec{f} \in \mathcal{F} \end{aligned} \tag{2}$$

where $\text{pen}(\vec{f})$ is a penalty that promotes sparsity in $\vec{f}$ and $\tau > 0$ is a regularization parameter that balances the negative Poisson log-likelihood term with the sparsity-promoting penalty term. We use the Sparse Poisson Intensity Reconstruction ALgorithm (SPIRAL) framework [24], [25] to solve (2), which involves solving a sequence of scalar quadratic subproblems of the form

$$\begin{aligned} \underset{\substack{f_I, f_N, f_F, f_M \\ \in \mathbb{R}}}{\text{minimize}} \quad & \tfrac{1}{2}(f_I - s_I)^2 + \tfrac{1}{2}(f_N - s_N)^2 + \tfrac{\tau}{\alpha_k}|f_I| + \tfrac{\tau\gamma}{\alpha_k}|f_N| + \\ & \tfrac{1}{2}(f_F - s_F)^2 + \tfrac{1}{2}(f_M - s_M)^2 + \tfrac{\tau}{\alpha_k}|f_F| + \tfrac{\tau}{\alpha_k}|f_M| \\ \text{subject to} \quad & 0 \leq f_I, f_N, f_F, f_M \leq 1, \quad 0 \leq f_I + f_N \leq 1 \\ & 0 \leq f_N \leq 1 - f_F, \qquad 0 \leq f_N \leq 1 - f_M, \\ & f_F + f_M - 1 \leq f_I \leq f_F + f_M. \end{aligned} \tag{3}$$

where at each iteration k,

- $\{f_I, f_N, f_F, f_M\}$ and $\{s_I, s_N, s_F, s_M\}$ are scalar components of the vectors $\vec{f}^k = \{\vec{f}_I^k, \vec{f}_N^k, \vec{f}_F^k, \vec{f}_M^k\}$ and $\vec{s}^k = \{\vec{s}_I^k, \vec{s}_N^k, \vec{s}_F^k, \vec{s}_M^k\}$, respectively, at the same location;
- α_k is the learning rate;
- $\vec{s}^k = \vec{f}^k - \frac{1}{\alpha_k} \nabla \Phi(\vec{f}^k)$ is the predicted new iterate along the steepest descent (negative gradient) from the current iterate with step length $1/\alpha_k$;
- $0 < \gamma < 1$ is a parameter that further amplifies sparsity on novel child SVs.

Note that because the constraints are more complex in (3) than in our previous work, we must use a different approach.

Optimization approach. We propose using an alternating block-coordinate descent approach to solve (3). Specifically, the proposed method solves (3) by alternating between child and parent indicator variables. First, we fix the parent structural variant signals, f_F and f_M, 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 indicator variables. The method continues until the difference between subsequent iterates falls below a specified threshold. We outline the steps below.

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

Step 1: Suppose we have obtained $\hat{f}_F^{(j-1)}$ and $\hat{f}_M^{(j-1)}$ from the previous iteration. The child indicator variables $\hat{f}_I^{(j)}$ and $\hat{f}_N^{(j)}$ are obtained from solving

$$\begin{aligned} \underset{f_I, f_N \in \mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_I - c_I)^2 + \tfrac{1}{2}(f_N - c_N)^2 \\ \text{subject to} \quad & 0 \leq f_I + f_N \leq 1 \\ & 0 \leq f_N \leq \min\left(1 - \hat{f}_F^{(j-1)}, 1 - \hat{f}_M^{(j-1)}\right) \\ & \max\left(0, \hat{f}_F^{(j-1)} + \hat{f}_M^{(j-1)} - 1\right) \leq f_I \\ & f_I \leq \min\left(1, \hat{f}_F^{(j-1)} + \hat{f}_M^{(j-1)}\right), \end{aligned} \tag{4}$$

where $c_I = s_I - \frac{\tau}{\alpha_j}$ and $c_N = s_N - \frac{\gamma\tau}{\alpha_j}$. The feasible region is shown in Fig. 2(a).

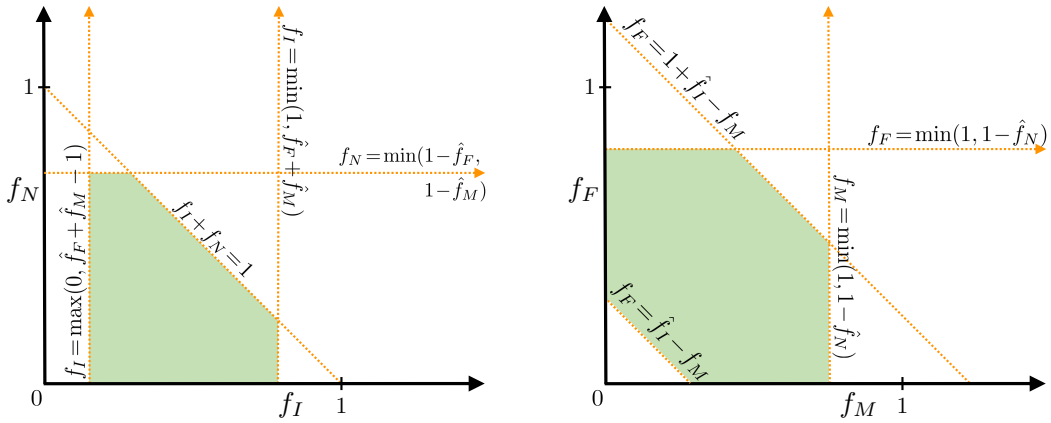

Fig. 2: The feasible set (indicated by the shaded region) for each step of the proposed block-coordinate minimization approach. (a) In Step 1, we minimize over the child indicator variables f_I and f_N given fixed parent indicator variables $\hat{f}_F$ and $\hat{f}_M$. (b) In Step 2, we minimize over the parent indicator variables f_F and f_M given fixed child indicator variables $\hat{f}_N$ and $\hat{f}_I$.

Step 2: Suppose we have obtained $\hat{f}_I^{(j)}$ and $\hat{f}_N^{(j)}$ from Step 1. To obtain the solution for the current iteration $\hat{f}_F^{(j)}$ and $\hat{f}_M^{(j)}$, we have

$$\begin{aligned} \underset{f_F, f_M \in \mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_F - c_F)^2 + \tfrac{1}{2}(f_M - c_M)^2 \\ \text{subject to} \quad & 0 \le f_F \le \min\left(1, 1 - \hat{f}_N^{(j)}\right), \\ & 0 \le f_M \le \min\left(1, 1 - \hat{f}_N^{(j)}\right), \\ & f_F + f_M - 1 \le \hat{f}_I^{(j)} \le f_F + f_M, \end{aligned} \tag{5}$$

where $c_F = s_F - \frac{\tau}{\alpha_j}$ and $c_M = s_M - \frac{\tau}{\alpha_j}$. The feasible region is shown in Fig. 2(b).

We note that both problems (4) and (5) have closed form solutions, where the minimizer is obtained by projecting the unconstrained solution to the feasible set (see e.g., [25]).

III. RESULTS

We implemented our method for variant detection in Matlab by extending our previous approach [25] based on the SPIRAL method [24]. We analyze the performance of our method on both simulated and real data by comparing our new method with two other variant prediction methods. We compare our previous method for variant prediction in the context of two-parents/one-child [20]. This method includes a sparsity promoting term τ, but did not specifically model novel variants in the child. Second, we include a comparison to the model that only enforces sparsity. The regularization parameter τ was chosen to be the same for all methods and γ was chosen when the area under the curve (AUC) was maximized. Each model was run with the same terminating criteria, checking if the relative difference between consecutive iterates converged to $\|\vec{f}_{k+1} - \vec{f}_k\|_2 / \|\vec{f}_k\|_2 \le 10^{-8}$.

A. Simulated Data

Because our model was developed in the simplified assumption of two-parent and one-child with haploid genomes, before applying it to real human data violating our assumptions, we studied its performance on data we simulated to match our assumptions. In these cases we simulated the true signal for both parents and the child and varied the fraction of similarity between parents and the number of novel variants in the child to study the performance of our model. We first created the parent signals and then derived the child with its novel variants. Each simulated true signal consisted of 10^5 potential SVs. For the parents, 500 locations were chosen at random to be true variants; the fraction of variants the parents had in common was varied according to their chosen percent similarity. For the child signal, if both parents had an SV at a particular location the child signal did as well. If only one parent had an SV at a location, the child had a 50% chance of inheriting that SV. Novel variants in the child were chosen randomly from locations where no parent had an SV. From these true signals, observed signals were created by sampling from the Poisson distribution with a given coverage and error.

Analysis. When the percentage of novel variants is small in the child ($< 10\%$), we observe better performance of our new method. In Figure 3 we show an ROC curve for a simulated data set where the parents were chosen to have 50% similarity and the child had 50 novel variants. We note that the area under the curve for our proposed method is higher than our other methods for the child reconstruction. Hence, we are able to more accurately recover the SVs in the child reconstruction when we allow for novel variants. We also note that while our performance is reduced for reconstructing the parent genome as compared to our previous method, our new method still outperforms sparsity constraints alone (data not shown). We also observed that for the child reconstruction, our proposed method is more stable under varying τ values.

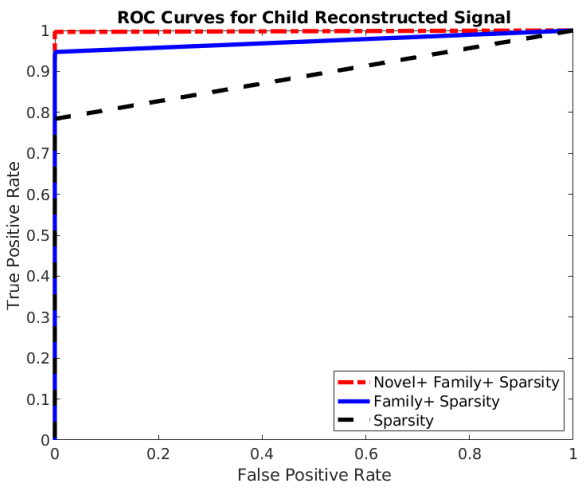

Fig. 3: ROC curves of three methods illustrating the false positive rate vs. the false true positive rate in the simulated child reconstruction, where $\lambda_M = \lambda_F = 8$, $\lambda_C = 10$, $\epsilon = .01$, $\tau = 100$ and $\gamma = 500$.

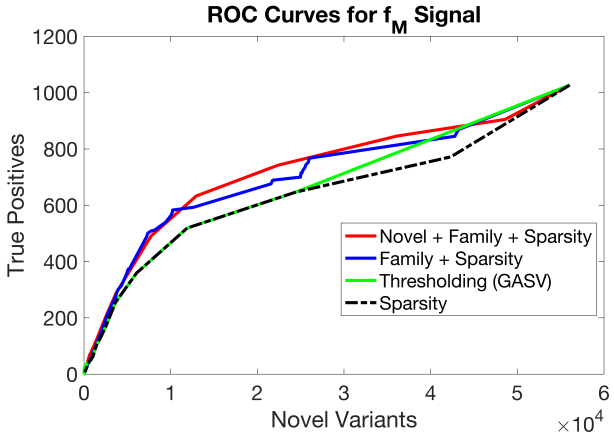

Fig. 4: ROC curves of four methods illustrating novel variants vs. true positives (experimentally validated) in the signal of the CEU mother NA12891, where $\tau = 10$ and $\gamma = \frac{1}{10}$. We observe an overall improvement in correctly classifying SVs compared to previous methods.

B. 1000 Genomes Project Trio Data

To validate our method, we consider trio data of two separate populations from the 1000 Genomes Project [11]. Both the European (CEU) and Yoruba (YRI) father-mother-daughter trio genomes were sequenced at $\approx 4\times$ coverage and aligned to NCBI36. We obtain our candidate set of SVs from the GASV pipeline, but our method is also applicable to other SV callers [26]. We filter the validated set of variants by eliminating experimentally validated SVs shorter than 250bp and which are classified as low quality. We note that only 15 of the validated variants in the child signal not present in either of the parents constitute our novel child signal $\vec{f}_N^*$.

Analysis. For child (inherited and novel) signal reconstructions, we achieve competitive sensitivity with our previous methods. When reconstructing the parent signals, we improve on our previous 2 parent - 1 child model, whose iterates are updated by non-alternating closed-form projections [20]. Figure 4 illustrates this improvement with predicted novel deletions against experimentally validated variants in the CEU mother NA12891 when comparing against previous models. We also find that our new method is stable under changes of τ and γ values.

IV. Conclusions

We propose a new method to detect novel structural variants – SVs present in a child not inherited from a parent – from sequencing data in parent-child trios. Our method incorporates both relatedness and sparsity constraints, allowing for varying penalty parameters in the reconstruction of the child signal. By doing so, our new model is less sensitive to our regularization parameters. With real data, our method achieves competitive true positive predictions in the child and improves parent signal recovery, and we intend on exploring this with further simulated data studies. We present our results for both simulated and real data from the 1000 Genomes Project and suggest further exploration in varying sequencing coverage for future parent-offspring data. In future studies, we intend to incorporate other SV-calling tools, larger family structures, and a general relatedness parameter in our methods.

Acknowledgment

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

References

[1] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping," *Nature Reviews Genetics*, vol. 12, no. 5, pp. 363, 2011.

[2] 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.

[3] B. Milholland, X. Dong, L. Zhang, X. Hao, Y. Suh, and J. Vijg, "Differences between germline and somatic mutation rates in humans and mice," *Nature Communications*, vol. 8, pp. 15183, 2017.

[4] I. Martincorena and P. J. Campbell, "Somatic mutation in cancer and normal cells," *Science*, vol. 349, no. 6255, pp. 1483–1489, 2015.

[5] A. Meindl, H. Hellebrand, C. Wiek, V. Erven, B. Wappenschmidt, D. Niederacher, M. Freund, P. Lichtner, L. Hartmann, H. Schaal, et al., "Germline mutations in breast and ovarian cancer pedigrees establish rad51c as a human cancer susceptibility gene," *Nature Genetics*, vol. 42, no. 5, pp. 410, 2010.

[6] Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, W. Ding, et al., "A strong candidate for the breast and ovarian cancer susceptibility gene brca1," *Science*, vol. 266, no. 5182, pp. 66–71, 1994.

[7] 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.

[8] J. Weischenfeldt, O. Symmons, F. 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, 2013.

[9] Z. Zhang, L. Mao, H. Chen, F. Bu, G. Li, J. Sun, S. Li, H. Sun, C. Jiao, R. Blakely, et al., "Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber," *The Plant Cell*, pp. tpc–114, 2015.

[10] A. A. Hoffmann and L. H. Rieseberg, "Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation?," *Annual review of ecology, evolution, and systematics*, vol. 39, pp. 21–42, 2008.

[11] 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.

[12] 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.

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

[14] 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.

[15] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding, "Varscan: variant detection in massively parallel sequencing of individual and pooled samples," *Bioinformatics*, vol. 25, no. 17, pp. 2283–2285, 2009.

[16] J.-Y. Li, J. Wang, and R. S. Zeigler, "The 3,000 rice genomes project: new opportunities and challenges for future rice research," *GigaScience*, vol. 3, no. 1, pp. 1–3, 2014.

[17] 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.

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

[19] 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.

[20] 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.

[21] M. Banuelos, L. Adhikari, R. Almanza, A. Fujikawa, J. Sahagún, K. Sanderson, M. Spence, S. Sindi, and R. F. Marcia, "Nonconvex regularization for sparse genomic variant signal detection," in *Medical Measurements and Applications (MeMeA), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 281–286.

[22] M. Banuelos, R. Almanza, L. Adhikari, S. Sindi, and R. F. Marcia, "Biomedical signal recovery: Genomic variant detection in family lineages," in *Bioengineering (ENBENG), 2017 IEEE 5th Portuguese Meeting on*. IEEE, 2017, pp. 1–4.

[23] 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*.

[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] 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.

[26] S. Sindi, E. Helman, A. Bashir, and B. J. Raphael, "A geometric approach for classification and comparison of structural variants," *Bioinformatics*, vol. 25, no. 12, pp. i222–i230, 2009.