Fine Tuning Sparsity Penalties to Improve Structural Variant Detection

Xiaolong Chen*, Hansell Perez, Melissa Spence, Roummel Marcia and Suzanne Sindi
Department of Applied Mathematics
University of California, Merced
Merced, United States
*Email: xchen46@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. Computational approaches for detecting SVs often involve parameters that describe certain relevant biological phenomena. In our work, such parameters relate the incidence of inherited and novel SVs to probabilistic models of observing these SVs. In the work presented here, we investigate the sensitivity of our computational framework to these parameters. In particular, we demonstrate the robustness of our method by identifying a wide range of parameter values that lead to high-accuracy SV predictions in simulated data.**

***Index Terms*—Computational genomics, next-generation sequencing data, structural variants, convex optimization, sparse signal recoery**

I. Introduction

Structural variants (SVs) are variations of genetic sequences that are larger than 1 kb. Recently, SVs have been associated with a variety of diseases, including various types of cancer [8]. The current method for detecting SVs is to sequence an individual's (unknown) genome and compare it with a known reference. These variations are commonly categorized into three main types: deletions, inversions, and novel insertions (see Fig. 1). When detecting SVs from alignments against a reference genome, we are interested in the number of fragments supporting any given position in the genome. This is commonly referred to as the coverage. Traditional sequencing methods offer high quality data that has been amplified to have high coverage, but it is costly and slow to generate. Next generation sequencing methods offer a faster, cheaper alternative to traditional sequencing methods, but at the cost of noisy data. Our goal is to develop methods that reduce the false positive discovery rate of structural variants due to low coverage.

One way for us to improve the ability to accurately predict SVs is to include information from related individuals whom we assume share similar features in their SV signals. While approaches like this have improved the ability to reduce false-positive predictions, they also increase the number false-negative predictions because they do not allow for novel variants (SVs that are not inherited from a parent) in the child genome. The method we present here builds a framework to reconstruct SV signals from one parent and one child under the assumption that the child possesses novel SVs. In this work, we examine the sensitivity of this model parameters.

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

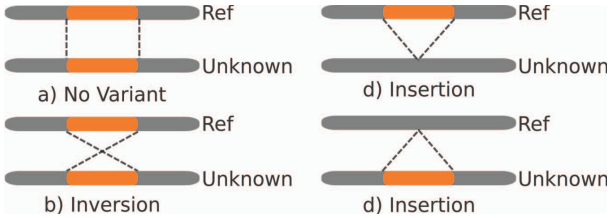

Fig. 1. Examples of different types of structural variations in an unknown being compared to a common reference genome.

A. Relation to Prior Work

Previously, we have developed frameworks to recover the signals of two parents and one child under the assumption that the child inherits all its SVs from one and/or both parents [4]. We hope to extend the following method to reconstruct the genetic SV signals of two parents and one child where the child can contain novel variants. When constructing these methods we want to have a thorough understanding of the parameters we use to help reflect the biological reality of the signals we are reconstructing.

II. Methods

For this work, we focus on tuning parameters for our framework to detect structural variants given sequencing data from one parent, $\vec{f}_p$, and one child, $\vec{f}_c$, where $\vec{f}_p$ and $\vec{f}_c$ are sequences both of length m. We allow the child to possess novel variants, which we assume are rarer than variants that are inherited. In this work we make the simplified assumption that each individual is haploid (only one copy of each chromosome). In our notation the true SV signal for each individual is a binary vector, $\vec{f}_I^{(j)} \in \{0,1\}^m$ where a 1 at position j indicates an SV and a 0 otherwise for individual $I \in \{p, c\}$. Fig. 2 shows an example of a true SV signal for one parent and one child. We represent the child signal as the sum of the novel variants signal, $\vec{f}_n$, namely, $\vec{f}_c = \vec{f}_i + \vec{f}_n$. The vectors of

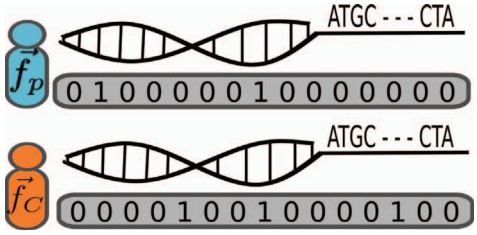

Fig. 2. Example of true SV signal in one parent $\vec{f}_p$ and child $\vec{f}_c$, $f_j = 1$ indicates a SV is present, 0 otherwise. Here we see an examples of the variants that are or are not inherited, as well as a variant in the child that is novel.

observed data $\vec{y}_p, \vec{y}_c \in \mathbb{R}^m$ are the number of DNA fragments supporting each potential SV. Let

$$\vec{y} = [\vec{y}_c; \vec{y}_p] \text{ and } \vec{f} = \left[\vec{f}_i; \vec{f}_n; \vec{f}_p\right].$$

We assume the sequencing coverage is low, and therefore, we assume the data follow a Poisson distribution. Thus the general observation model can be expressed as

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

where A is a linear operator that maps the true signal to the vector of measurements and $\mathbf{1}$ is the vector of ones.

A. Structural Variant Signal Recovery

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 $\vec{y}$ is maximized. We do this by minimizing the corresponding *negative Poisson log-likelihood* given by

$$F(\vec{f}) = \sum_{j=1}^{2m} (A\vec{f})^{(j)} - \vec{y}^{\,(j)} \log\left((A\vec{f})^{(j)} + \epsilon\right).$$

The genomic variants reconstruction problem has the following constrained optimization form:

$$\begin{aligned} &\underset{\vec{f}\in\mathbb{R}^{3m}}{\text{minimize}} \quad F(\vec{f}) + \tau\text{pen}(\vec{f}) \\ &\text{subject to} \quad \vec{f} \in \mathcal{S} \end{aligned} \tag{2}$$

where $\text{pen}(\vec{f})$ is a sparsity enforcing penalty term and our biological constraints are given by

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

The biological constraint given by

$$0 \le \vec{f}_i + \vec{f}_n \le 1$$

controls for the childs SV's being either novel or inherited, but not both. The next constraint controls for the case when the child inherits an SV from the parent, then that structural variant must also be present in the parent:

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

Finally, if there is a novel SV in the child, we must enforce that there not be an SV at that same position in the parent:

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

Now we need to address the fact that structural variants are relatively rare in an individuals genome, especially structural variants that we are calling novel variants. We do this by incorporating an ℓ_1 penalty term into our objective function:

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

where $\gamma \gg 1$ is a penalty weight on $\vec{f}_n$ enforcing the more severe rarity of novel variants and $\tau > 0$ enforces sparsity on all types of SVs.

B. Fine Tuning Sparsity Penalties

We solve the optimization problem (2) using the Sparse Poisson Intensity Reconstruction ALgorithm (SPIRAL) [7], which generates a sequence of convex quadratic separable subproblems that have closed form solutions. In our pervious work, the τ and γ parameters were chosen based on a simple parameter search. We are motivated to explore the sensitivity of our method to these parameters to ensure that we are minimizing the false positive discovery rate.

We simulated data sets where $\vec{f}$ is a vector of length 10^5 with 500 SVs for both the parent and child. Then we varied the number of inherited and novel SVs in the child's signal. After solving for reconstrucing the SV signal for each data set, we compute the area under curve (AUC) of the receiver operating characteristic (ROC) for various values of τ and γ.

Fig 3. shows the AUC for each data set varying sparsity penalties. It indicates for a well-chosen τ values the accuracy of our framework remains consistent. The tuning of γ does not play a critical role, but it becomes more significant as the number of novel variants increases. In the figure, we observe a low accuracy region in $\tau = 1000$, which indicates the tuning of penalty parameter are critical to the accuracy of our framework.

III. Framework Sensitivity

After we analyzed the sensitivity of our method to τ and γ we were interested in seeing how sensitive our framework was to the composition of the child signal. For example if the child signal unrealistically only had novel SVs, nothing was inherited from the parent, then leveraging the parent's SV signal would not give us any information to reconstruct the child's signal. Alternatively if the overwhelming majority of the child's variants were inherited then leveraging the parent's signal would aid in our reconstruction and we would expect our performance to be much better.

To test these assumptions we generated data sets for which in the child's signal the number of inherited SVs ranged from 250-500 and the number of novel SVs ranged from 0-250. So in the most extreme cases the child has 0 novel SVs and 500 inherited SVs or the child has 250 novel SVs and 250 inherited SVs. We fixed $\gamma = 10$ and chose two different

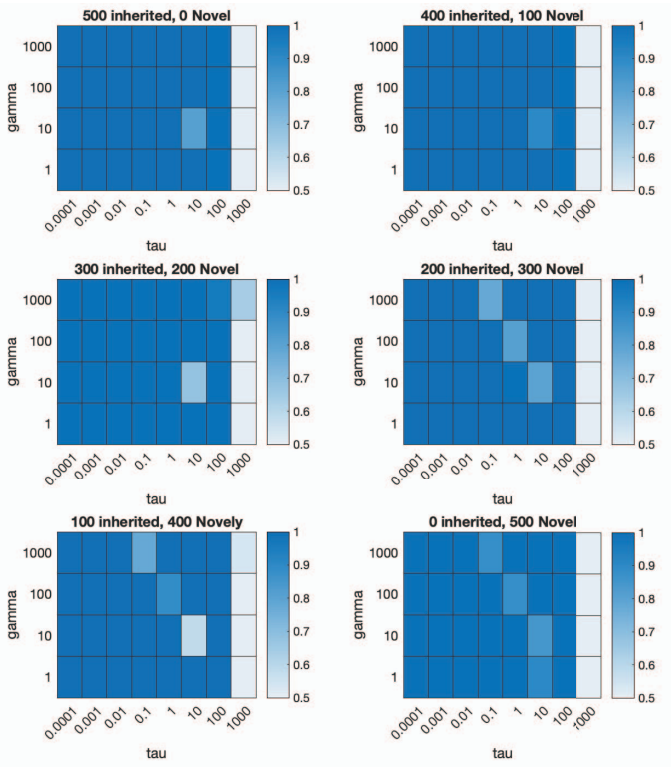

Fig. 3. AUC shown for different values of the penalty terms under varying assumptions of child signal structure.

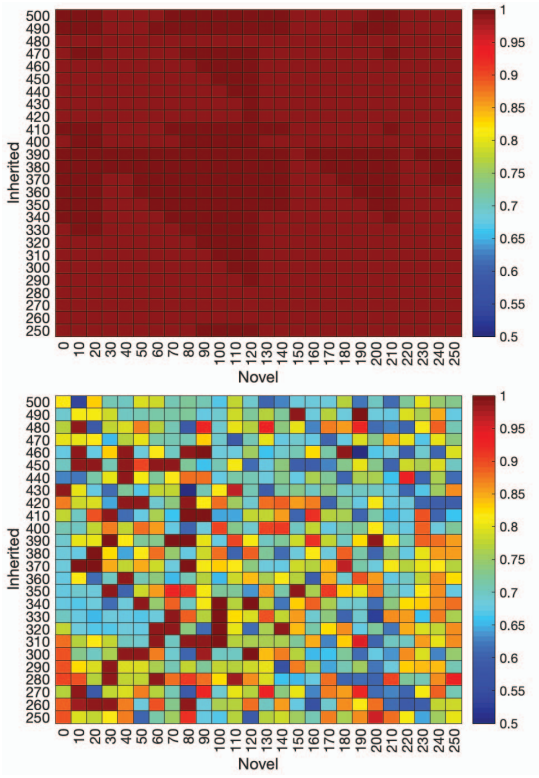

Fig. 4. **Top** $\tau = 0.01$. **Bottom** $\tau = 10$. Area under the curve for ROC heatmap. The inherited SVs range from $250 - 500$ while the novel SVs range from $0 - 250$.

tau values so we could create a heatmap of performance ranging over the different compositions of the child signal. The results of this can be seen in Fig. 4, where for each data set we plotted the AUC. We find that our method is robust to the number of inherited and novel Svs for carefully chosen values of τ. However, if τ is chosen to be too large then we are over enforcing sparsity on our solution which leads to less uniform performance.

IV. Conclusion

In the beginning of this analysis, we suspected that the performance of our method heavily depends on the ratio of inherited to novel SVs in the child. However, the above analysis suggests that well chosen parameter values has a larger influence on performance that the composition of the signal being reconstructed. To increase the accuracy of SV predictions, choosing appropriate parameters is crucial. In the future we would like to extend this sensitivity analysis to real data from the 1000 Genomes Project [1] since we suspect the simplicity of our simulated data may hinder us in uncovering all of the patterns in performance given changes in each parameter.

References

[1] 1000 Genomes Project Consortium and others. "An integrated map of genetic variation from 1,092 human genomes." Nature, vol. 491, no. 7422, 2012.

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

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

[4] Banuelos, M., Adhikari, L., Fujikawa, A., Sahagun, J., Sanderson, K., Spence, M., Almanza, R., Sindi, S., Marcia, RF. "Nonconvex Regularization for Sparse Genomic Variant Signal Detection" *IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2017.

[5] Banuelos, M., Adhikari, L., Fujikawa, A., Sahagun, J., Sanderson, K., Spence, M., Almanza, R., Sindi, S., Marcia, RF. "Sparse Diploid Spatial Biosignal Recovery for Genomic Variant Detection" *IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2017.

[6] Spence, M., Banuelos, M., Marcia, RF., and Sindi, S. "Detecting Novel Structural Variants in Genomes by Leveraging Parent-Child Relatedness" *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018.

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

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