Sparse Signal Recovery Methods for Variant Detection in Next-Generation Sequencing Data

Mario Banuelos, Rubi Almanza, Lasith Adhikari, Suzanne Sindi, and Roummel F. Marcia

Applied Mathematics, *University of California, Merced*, CA 95343 USA

Email: mbanuelos4@ucmerced.edu

Abstract

Recent advances in high-throughput sequencing technologies have led to the collection of vast quantities of genomic data. These sequencing data have the potential to answer questions about the evolutionary history of a species and the genomic basis of hereditary diseases. Structural variants (SVs) -- rearrangements of the genome larger than one letter such as inversions, insertions, deletions, and duplications -- are an important source of genetic variation and have been implicated in some genetic diseases. However, inferring SVs from sequencing data has proven to be challenging because true SVs are rare and are prone to low-coverage noise. We attempt to mitigate the deleterious effects of low-coverage sequences by following a maximum likelihood approach to SV prediction. Specifically, we model the noise using Poisson statistics and constrain the solution with a sparsity-promoting ℓ_1 penalty since SV instances should be rare. In addition, because offspring SVs inherit SVs from their parents, we incorporate familial relationships in the optimization problem formulation to increase the likelihood of detecting true SV occurrences. Numerical results are presented to validate our proposed approach.

DNA Sequencing and Genetic Variants

The 1000 Genomes Project, which catalogues human genomic variation in comprehensive detail is one example of large-scale sequencing studies. These massive repositories of data offer the potential to increase our understanding of the complex evolutionary history of different species, identify genetic basis of important phenotypes including disease and -- for humans -- usher in the era of personalized medicine.

The genome of organisms change throughout generations via deletions, mutations, or other replication processes. A promising class of genetic variant emerging from such studies are structural variants (SVs) -- rearrangements of the genome larger than one letter such as inversions, insertions, deletions, and duplications. We illustrate a few of these SVs below:

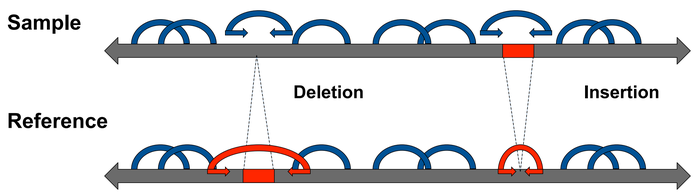


Figure 1: *Illustration of different structural variations in a sample genome in comparison to the reference genome. The sample genome is first fragmented. The ends of the fragments are then aligned to the reference genome. Fragments in the sample that do map to the reference are considered structural variants.*

Mathematical Model

We consider a general framework to detect SVs from sequencing data from a child and parent genome. We observe given fragments that support a potential SV. Then, our discrete stacked observations $\vec{y} = [\vec{y}_c; \vec{y}_p]$ for the parent and child signals can be described as

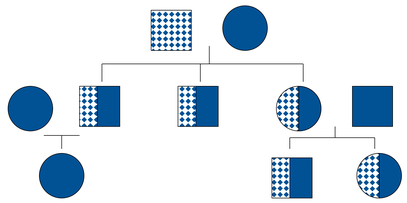
$$\vec{y} \sim \text{Poisson}(\hat{A}\vec{f}^*)$$

where $\hat{A}$ represents the expected genome coverage and $\vec{f}^*$ represents the true SV signal (0 if not present, 1 if variant present). We seek to maximize the probability of observing the data using the probability mass function of the Poisson distribution. Since each location *n* is independent, the probability of data is given as

$$p(\vec{y}|\hat{A}\vec{f}^*) = \prod_{i=1}^{2n} \frac{(\vec{e}_i^T \hat{A}\vec{f}^*)^{\vec{y}_i}}{\vec{y}_i!} \exp\left(-\vec{e}_i^T \hat{A}\vec{f}^*\right)$$

Sparsity and Familial Constraints

Large-scale sequencing studies often sequence populations of related individuals, including father-mother-child trios. Since spontaneous variants are rare, individuals inherit SVs from either a father or mother.



- *Homozygous Male (SV present)*
- *Heterozygous Male*
- *Homozygous Female (No SV)*

Figure 2: *Illustration of transmission of variants through generations where the offspring in the first generation are heterozygous for the structural variant.*

We use a maximum likelihood approach that incorporates the rarity of SVs with a penalty term and constrains parent and child signal reconstructions to reflect inheritance of variants. The resulting penalized constrained negative Poisson log-likelihood is given by

$$\underset{\vec{f}\in\mathbb{R}^{2n}}{\text{minimize}} \overbrace{\sum_{i=1}^{2n}(\hat{A}\vec{f})_i - \sum_{i=1}^{2n}\vec{y}_i \log\left((\hat{A}\vec{f})_i + \epsilon\right)}^{F(\vec{f})} + \tau\text{pen}(\vec{f}) \quad \text{subject to } 0 \le \vec{f}_c \le \vec{f}_p \le 1.$$

Based on [1], we solve this optimization problem by solving a sequence of quadratic sub-problems from the second-order Taylor series expansion at each iterate $\vec{f}^k$:

$$\vec{f}^{k+1} = \underset{\vec{f}\in\mathbb{R}^{2n}}{\arg\min} \ \frac{1}{2}\|\vec{f} - \vec{s}^k\|_2^2 + \frac{\tau}{\alpha_k}\text{pen}(\vec{f}) \quad \text{subject to } \ 0 \le \vec{f}_c \le \vec{f}_p \le 1,$$

where $\vec{s}^k = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$ and $\alpha_k > 0$. Because SVs are rare, we use $\text{pen}(\vec{f}) = \|\vec{f}\|_1$ to promote sparsity in our solution.

Separable Subproblems

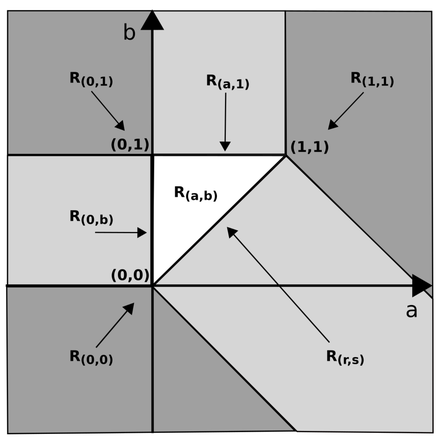
Let $\lambda = \tau/\alpha$. The objective function decouples in each variable and can be optimized separately, which results in the following *scalar* optimization:

$$\underset{f_p,\, f_c\in\mathbb{R}}{\text{minimize}} \ \frac{1}{2}(f_p - s_p)^2 + \lambda|f_p| + \frac{1}{2}(f_c - s_c)^2 + \lambda|f_c| \quad \text{subject to } \ 0 \le f_c \le f_p \le 1.$$

Let $a = s_c - \lambda, b = s_p - \lambda$. Completing the squares and ignoring constant terms yields

$$(f_c^*, f_p^*) = \underset{f_p, f_c\in\mathbb{R}}{\arg\min} \ \frac{1}{2}(f_p - a)^2 + \frac{1}{2}(f_c - b)^2 \qquad \text{subject to } 0 \le f_c \le f_p \le 1.$$

The minimizer (f_c^*, f_p^*) depends on (a, b), which is given by Figure 3.



(a)

Region	Condition a	Condition b	(f_c^*, f_p^*)
$R_{(a,b)}$	$0 < a < b$	$0 < b < 1$	(a, b)
$R_{(0,b)}$	$a < 0$	$0 \le b \le 1$	$(0, b)$
$R_{(a,1)}$	$0 \le a \le 1$	$b > 1$	$(a, 1)$
$R_{(0,1)}$	$a < 0$	$b > 1$	$(0, 1)$
$R_{(0,0)}$	$a \le -b$	$b < 0$	$(0, 0)$
$R_{(1,1)}$	$a > 1$	$b \ge -a + 2$	$(1, 1)$
$R_{(r,s)}$	$a > \lvert b\rvert$	$b < -a + 2$	(r, s)

(b)

Figure 3: ***a)*** *Plot of the a-b plane, feasible region, and graphical representation of projected minimizers of optimization subproblems.* ***b)*** *Table representing all regions and conditions in the a-b plane, along with corresponding minimizers and* $r = s = (a + b)/2$.

Simulated Data Results

We validated our method using simulated data. Given 10^5 possible SV locations and 500 true variants, we compare the results with and without familial constraints.

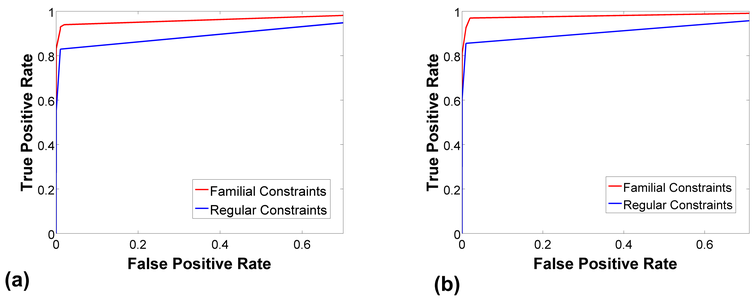


Figure 4: *ROC curves depicting the False Positive Rate vs. True Positive Rate for the reconstruction of the parent signals with low coverage and* ***a)*** *70%,* ***b)*** *90% similarity of variants using both methods with* $\tau = 1.553$ *for regular constraints and* $\tau = 1.474$ *for family constraints.*

1000 Genomes Data Results

We apply our method to low-coverage (~ 4X) sequencing data for the CEU trio from the 1000 Genomes Project [2] using regular and familial constraints. Using the GASV [3] method on this dataset, we obtain a set of possible SVs. For the parent signals, we report higher specificity and sensitivity rates with our method incorporating familial constraints.

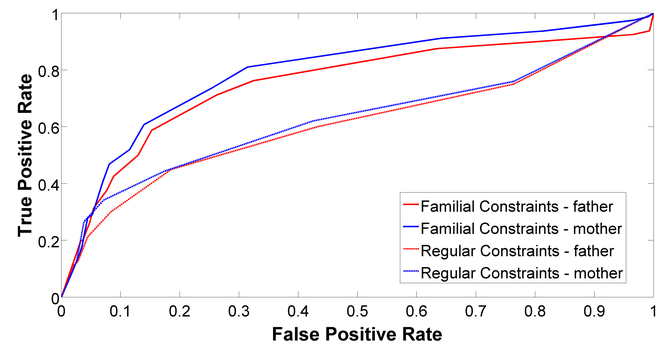


Figure 5: *Plot of ROC curves depicting False Positive Rate vs. True Positives for Chromosome 1 of CEU Parents comparing familial constraints with regular constraints and* $\tau = 2.65$. *True deletions were experimentally validated by the 1000 Genomes Project.*

Concluding Remarks

We introduce a novel SV discovery method using a maximum likelihood approach that incorporates sparsity and familial constraints. We demonstrated the effectiveness of our approach on both simulated data and data from the 1000 Genomes Project. We intend to consider a general relatedness parameter to predict structural variants in a population.

References

[1] 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 Im- age Processsing*, vol. 21, pp. 1084 – 1096, 2011.
[2] 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.
[3] 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.

Acknowledgements

This research was supported by NSF Grant CMMI 1333326.