SPARSE GENOMIC STRUCTURAL VARIANT DETECTION: EXPLOITING PARENT-CHILD RELATEDNESS FOR SIGNAL RECOVERY

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

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

ABSTRACT

Structural variants (SVs) – rearrangements of an individuals' genome – are an important source of heterogeneity in human and other mammalian species. Typically, SVs are identified by comparing fragments of DNA from a test genome to a known reference genome, but errors in both the sequencing and the noisy mapping process contribute to high false positive rates. When multiple related individuals are studied, their relatedness offers a constraint to improve the signal of true SVs. We develop a computational method to predict SVs given genomic DNA from a child and both parents. We demonstrate that enforcing relatedness between individuals and constraining our solution with a sparsity-promoting ℓ_1 penalty (since SV instances should be rare) results in improved performance. 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

1. INTRODUCTION

The genome (or complete DNA sequence) of an individual is vertically inherited from parent to child. However, the evolutionary processes of mutation, coupled with more complex heredity in sexually reproducing organisms, ensures variation between genomes of individuals within a species. Since genomes vary, the common practice has been to develop a reference genome for each species along with an annotation of common sites of variation [1, 2]. Genomic variation can either consist of a single letter (nucleotide), so called single nucleotide variants (SNVs), or rearrangements of larger regions of DNA, called structural variants (SVs). In both cases, variation is identified by comparing fragments of DNA sequenced from a test (unknown) genome to a given reference (see Figure 1) [4, 3] . Detecting SVs is typically more challenging than SNVs, and subject to errors from both DNA sequencing process as well as alignment to the reference genome. Because of this, the signal of a true SV may be compromised by noise and standard detection methods suffer from high false-positive rates [3, 4].

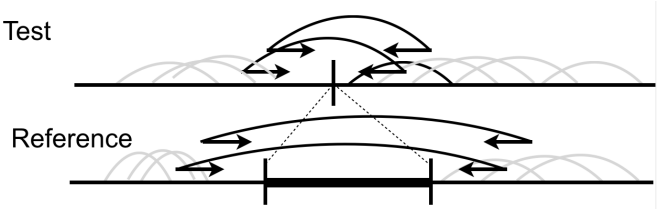

Fig. 1: Illustration of deletion in a test genome (unknown) relative to a reference genome (known). Deletions (and other SVs) are identified by sequencing fragments (of a particular length distribution) from the test genome and mapping them to the reference. Fragments whose mapped distance is significantly larger than expected (black fragments) indicate a potential deletion.

Recent advances in high-throughput DNA sequencing technologies have meant that researchers are now capable of obtaining DNA fragments from hundreds or thousands of individuals; typically, many of the individuals sampled are related [5, 6]. Since the rate of SV creation through genetic mutation is thought to be low [7], we can constrain the SV prediction problem by enforcing that when children and parents are sequenced, SVs present in the child must have been inherited from one of the parents.

In this paper, we extend the method from [8] by employing a similar maximum likelihood approach to variant detection but now include constraints from both parents as well as the child. In our model, we not only require that any SVs in the child are also present in one of the parents, but also exclude the possibility of the child having a variant if neither parent possesses the variant. Numerical results of both simulated and real sequencing data improves SV detection for low-coverage individuals using the familially constrained one-parent method.

2. METHOD

IHere we consider a general framework for detecting structural variants (SVs) given sequencing data from two parents (p_1 and p_2) and one child (c). We assume that there are n

This work was supported by NSF Grant CMMI 1333326.

locations in the genome that could be a potential SV. For simplicity, we consider each individual to be haploid (only one copy of each chromosome). As such, the true SV signal for each individual at each location is either a 0, if they do not have an SV at that location, or 1 if they do. The observed data are the number of DNA fragments supporting each potential SV, $\vec{y}_{p_1}, \vec{y}_{p_2}, \vec{y}_c \in \mathbb{R}^n$ for both parents and the child, and the data are assumed to follow a Poisson distribution [4], $\vec{y}_i \sim \text{Poisson}(A_i \vec{f}_i^*)$, where $i \in \{p_1, p_2, c\}$ and $A_i = (k_i - \epsilon)\mathbb{I} \in \mathbb{R}^{n \times n}$ is a linear projection of the true genomic variants $\vec{f}_i^* \in \mathbb{R}^n$ to the observation $\vec{y}_i$. The constants k_{p_1}, k_{p_2}, and k_c are the sequencing coverage for each individual, the mean of the Poisson distribution. Finally, we assume that the error in measurement $\epsilon > 0$ is the same for all observations. We stack the true variant signals and observations in the form $\vec{f}^* = [\vec{f}_c^*; \vec{f}_{p_1}^*; \vec{f}_{p_2}^*]$ and $\vec{y} = [\vec{y}_c; \vec{y}_{p_1}; \vec{y}_{p_2}]$, the general observation model is be expressed as

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

where $\hat{A} \in \mathbb{R}^{3n \times 3n}$ is a block-diagonal matrix with $\hat{A} = \text{diag}(A_c, A_{p_1}, A_{p_2})$.

Familial constraints. We require the (continuous) elements for each individual lie within 0 and 1, i.e., $0 \leq \vec{f} \leq 1$. The continuous relaxation of the reconstruction $\vec{f}$ thus allows us to to apply gradient-based techniques. Further, we impose the element-wise constraints that if both parents have the SV, then the child must also, i.e., $\vec{f}_{p_1} + \vec{f}_{p_2} - 1 \leq \vec{f}_c$ for each location. Finally, if neither parent has the SV, then the SV cannot be present in the child, i.e., $\vec{f}_c \leq \vec{f}_{p_1} + \vec{f}_{p_2}$.

2.1. Problem Formulation

Under the Poisson process model (1), the probability of observing $\vec{y}$ is given by

$$p(\vec{y} \,|\hat{A}\vec{f}^*) = \prod_{i=1}^{3n} \frac{(\vec{e}_i^T \hat{A}\vec{f}^*)^{\vec{y}_i}}{\vec{y}_i!} \exp\left(-\vec{e}_i^T \hat{A}\vec{f}^*\right), \tag{2}$$

where $\vec{e}_i$ is the i-th column of the $3n \times 3n$ identity matrix. The *maximum likelihood principle* is used to determine the unknown Poisson parameter $\hat{A}\vec{f}^*$ such that the probability of observing the vector of Poisson data $\vec{y}$ in (2) is maximized. Thus, the genomic variants reconstruction problem has the following constrained optimization form:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{3n}}{\text{minimize}} \quad & \phi(\vec{f}) \equiv F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} \quad & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \leq \vec{f}_c \leq \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & 0 \leq \vec{f} \leq 1 \end{aligned} \tag{3}$$

where $F(\vec{f})$ is the negative Poisson log-likelihood function

$$F(\vec{f}) = \mathbf{1}^T \hat{A}\vec{f} - \sum_{i=1}^{3n} \vec{y}_i \log\left(\vec{e}_i^T \hat{A}\vec{f} + \epsilon\right),$$

where $\vec{f} = [\vec{f}_c; \vec{f}_{p_1}; \vec{f}_{p_2}]$, $\tau > 0$ is a regularization parameter, $\mathbf{1}$ is a vector of ones, and pen is usually a sparsity enforcing penalty functional.

Since true variants are rare, the penalty functional pen($\vec{f}$) in (3) can thus be replaced by sparsity-promoting ℓ_1-norm of $\vec{f}$, i.e., $\|\vec{f}\|_1$. In the SPIRAL framework [9], the solution of (3) is obtained by minimizing a sequence of quadratic models to the function $F(\vec{f})$. In these models, the Hessian in the second-order Taylor series expansion of $F(\vec{f})$ at the current iterate $\vec{f}^k$ is replaced by a scaled identity matrix $\alpha_k I$ with $\alpha_k > 0$ (see [10, 11] for details). This quadratic approximation can be simplified to a subproblem of the form:

$$\begin{aligned} \vec{f}^{k+1} = \underset{\vec{f} \in \mathbb{R}^{3n}}{\arg\min} \quad & \tfrac{1}{2}\|\vec{f} - \vec{s}^k\|_2^2 + \tfrac{\tau}{\alpha_k}\|\vec{f}\|_1 \\ \text{subject to} \quad & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \leq \vec{f}_c \leq \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & 0 \leq \vec{f}_c, \vec{f}_{p_1}, \vec{f}_{p_2} \leq 1, \end{aligned} \tag{4}$$

where $\vec{s}^k = [\vec{s}_c^k; \vec{s}_{p_1}^k; \vec{s}_{p_2}^k] = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$. Then, the subproblem in (4) can be separated into scalar minimization problems (see [9] for details). Completing the squares and ignoring constant terms, we have

$$\begin{aligned} \underset{f_c, f_{p_1}, f_{p_2} \in \mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_c - c)^2 + \tfrac{1}{2}(f_{p_1} - p_1)^2 + \tfrac{1}{2}(f_{p_2} - p_2)^2 \\ \text{subject to} \quad & f_{p_1} + f_{p_2} - 1 \leq f_c \leq f_{p_1} + f_{p_2}, \\ & 0 \leq f_c, f_{p_1}, f_{p_2} \leq 1 \end{aligned} \tag{5}$$

where $c = s_c - \lambda$, $p_1 = s_{p_1} - \lambda$, and $p_2 = s_{p_2} - \lambda$. The feasible solution to (5) is obtained by orthogonally projecting the solution (c, p_1, p_2) to a three-dimensional feasible region (see Fig. 2). In particular, there are 27 regions to be considered in the f_c-f_{p_1}-f_{p_2} tri-dimensional space according to the constraints of (5). If (c, p_1, p_2) satisfy the constraints of (5), then the minimizer for the subproblem (5) corresponds to the minimizer given in Table 1 (Interior). Otherwise, the subproblem solution is projected to a vertex, edge, or surface of Figure 2. Tables 1 and 2 summarize a few of the regions and projections considered.

3. RESULTS

In this section, we evaluate the effectiveness of the proposed familially-constrained two-parent approach on both simulated and real genomic data. The proposed method is implemented in MATLAB by modifying the existing SPIRAL approach [12] to solve subproblems (5). The algorithm is initialized by $\hat{A}^T\vec{y}$ and terminates if the relative difference between consecutive iterates converged to $\|\vec{f}^{k+1} - \vec{f}^k\|_2 / \|\vec{f}^k\|_2 \leq 10^{-8}$. We compare the results with the regular constrained (i.e., only nonnegativity constrained) method and familially-constrained one-parent method [8]. The regularization parameters (τ) for all experiments are optimized to get the minimum RMSE $(\%) = 100 \cdot \|\hat{f} - \vec{f}^*\|_2 / \|\vec{f}^*\|_2$.

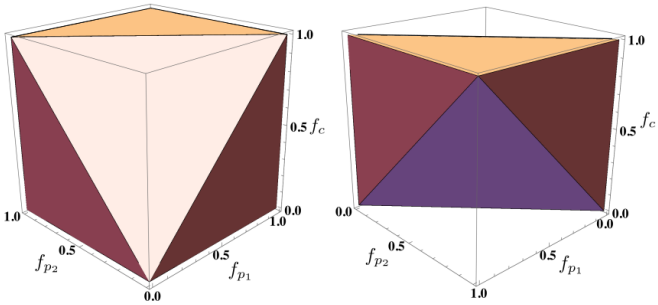

Fig. 2: The three-dimensional feasible region of the minimization problem (5) on the $f_c - f_{p_1} - f_{p_2}$ axis. Subproblem minimizers not satisfying the constraints are projected onto the region. Left : Front view. Right : Back view.

Table 1: Examples of solutions on the feasible region

	Minimizer	c	p_1	p_2
Interior	(c, p_1, p_2)	$0 \le c \le 1$, $c \le p_2 + p_1$, $c \ge p_2 + p_1 - 1$	$0 \le p_1 \le 1$	$0 \le p_2 \le 1$
Vertex	$(0, 0, 0)$	$c \le -p_2$, $c \le -p_1$	$p_1 \le 0$	$p_2 \le 0$
	$(0, 0, 1)$	$c \le 0$, $c \le -p_1$	$p_1 \le p_2 - 1$	$p_2 \ge 1$
	$(1, 1, 0)$	$c \ge 1$, $c \ge -p_1 + 2$	$p_1 \ge p_2 + 1$	$p_2 \le 0$
	$(1, 1, 1)$	$c \ge -p_1 + 2$, $c \ge -p_2 + 2$	$p_1 \ge 1$	$p_2 \ge 1$
Edge	$(0, s_1, t_1)$	$c < \frac{1}{2}(1 - p_1 - p_2)$	$p_1 < p_2 + 1$, $p_1 > p_2 - 1$	$p_2 > 1 - p_1$
	$(1, s_2, t_2)$	$c > \frac{1}{2}(3 - p_1 - p_2)$	$p_1 < p_2 + 1$, $p_1 > p_2 - 1$	$p_2 < 1 - p_1$
	$(r_3, 0, t_3)$	$c > -p_2$, $c > p_2$	$p_1 < \frac{1}{2}(p_2 - c)$	$p_2 < 2 - c$
	$(r_4, s_4, 0)$	$c > -p_1$, $c > p_1$	$p_1 < 2 - c$	$p_2 < \frac{1}{2}(p_1 - c)$
Surface	(r_7, s_7, t_7)	$c \ge p_2 + p_1$, $c \le \frac{3}{2} - \frac{1}{2}(p_1 + p_2)$	$p_1 \ge \frac{1}{2}(p_2 - c)$	$p_2 \ge \frac{1}{2}(p_1 - c)$
	(r_8, s_8, t_8)	$c \le p_2 + p_1 - 1$, $c \ge \frac{1}{2}(1 - p_2 - p_1)$	$p_1 \le \frac{1}{2}(p_2 - c + 2)$	$p_2 \le \frac{1}{2}(p_1 - c + 2)$

Table 1. Examples of minimizers to the subproblem (5) above are obtained by projecting the minimum of the unconstrained problem (c, p_1, p_2) onto the feasible region illustrated by (2). The f_c-f_{p_1}-f_{p_2} tri-dimensional space is partitioned into 27 regions depending on c, p_1, and p_2. Projections onto edges and surfaces are represented as linear combinations of c, p_1, and p_2 in Table 2.

3.1. Simulated Data

For this experiment, we simulated two parent signals $\vec{f}_{p_1}$ and $\vec{f}_{p_2}$ and a child signal $\vec{f}_c$ of size 10^5. Both true parent signals have 500 variants, i.e., the sparsity of each signal is 0.5%, while they share 50% of variant similarities. The value of the child signal at each point is chosen from either parent with equal probability. The error term ϵ is set to 0.01 in obtaining measurements from the forward model.

Table 2: Projections on the feasible region

	Minimizer	r	s	t
Edge	$(0, s_1, t_1)$	0	$\frac{1}{2}(p_1 - p_2 + 1)$	$\frac{1}{2}(p_2 - p_1 + 1)$
	$(1, s_2, t_2)$	1	$\frac{1}{2}(p_1 - p_2 + 1)$	$\frac{1}{2}(p_2 - p_1 + 1)$
	$(r_3, 0, t_3)$	$\frac{1}{2}(c + p_2)$	0	$\frac{1}{2}(c + p_2)$
	$(r_4, s_4, 0)$	$\frac{1}{2}(c + p_1)$	$\frac{1}{2}(c + p_1)$	0
	$(r_5, s_5, 1)$	$\frac{1}{2}(c + p_1)$	$\frac{1}{2}(c + p_1)$	1
Surface	(r_7, s_7, t_7)	$\frac{2c + p_1 + p_2}{3}$	$\frac{c + 2p_1 - p_2}{3}$	$\frac{c - p_1 + 2p_2}{3}$
	(r_8, s_8, t_8)	$\frac{1}{3}(2c + p_1 + p_2 - 1)$	$\frac{1}{3}(c + 2p_1 - p_2 + 1)$	$\frac{1}{3}(c - p_1 + 2p_2 + 1)$

Table 2. Examples of projections (r, s, t) of (c, p_1, p_2) to the surfaces and edges of (2) are obtained by first considering the region in f_c-f_{p_1}-f_{p_2} tri-dimensional space as described in Table 1. All projections to the feasible region are calculated and represented as linear combinations of c, p_1, and p_2.

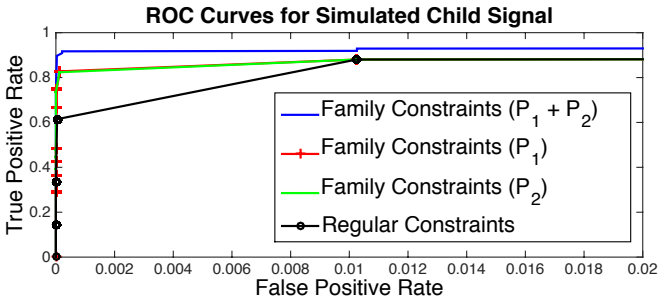

Fig. 3: ROC curves illustrate the false positive rate vs true positive rate for the child signal reconstruction in the simulated data with $k_{p_1} = 5$, $k_{p_2} = 5$, $k_c = 2$, and 50% similarity of variants between parents using the three methods.

Analysis. We first examine the child signal reconstruction. Fig. 3 illustrates the false positive rate vs. true positive rate for the child signal reconstruction with coverages $k_{p_1} = 5$, $k_{p_2} = 5$, $k_c = 2$, and 50% similarity of variants between parents. (Since for false positive rate values > 0.02 no significant information could be discerned, the false positive rate axis was shortened in order to provide a more detailed view of the curves.) We can clearly observe the improvement in true predictions over false predictions obtained by the pro-

posed familially-constrained two-parent model (see the blue curve in Fig. 3) over the two existing methods (i.e., regular constrained method and familially-constrained one-parent method). Specifically, the child signal reconstruction is most improved with our proposed method when the parent share more variants in common. With regards to the parent signal reconstructions, however, we did not observe this type of improvement in signal reconstruction.

3.2. 1000 Genomes Project Trio Data

We next applied our method to the previously sequenced genomes of the father-mother–daughter CEU trio (NA12891, NA12892, NA12878) and YRI trio (NA19238, NA19239, NA19240) from the 1000 Genomes Project [1]. These genomes were sequenced at low coverage ($\approx 4\times$) in Pilot 1 of the study and were aligned to NCBI36. To obtain observations of possible variants, we used the GASV [13] method on the 1000 Genomes Project data. We compared our reconstructions against the reported validated set of low coverage deletions longer than 250bp. Moreover, the validated deletions were filtered by removing regions overlapping centromeres or telomores. Cases marked with *LowQual* for all three individuals were also removed.

Analysis. In both trios, parents shared the majority of validated variants. Algorithm run time for reconstructing parent and child signals on a commodity machine in MATLAB was 8.15 seconds and 6.81 seconds for CEU data and YRI data, respectively. Figures 4 and 5 represent ROC plots depicting true positives versus novel deletions since the validated set of deletions may be incomplete. Thus, our method may correctly identify true deletions not in the experimental validated set. The length of observations for each CEU genome is $n = 56,840$, and for each YRI genome, $n = 51,087$. Both Figures 4 and 5 illustrate improvement in identifying child variants on the regularly constrained method, but also improves on our one-parent familially-constrained model. Since SV-detecting biological algorithms are already used to process sequencing data, this method adds low computational cost while improving the detection of variants in the child signal. Moreover, predicting variants of large amounts of sequenced trios would be tractable with the proposed method.

4. CONCLUSIONS

We present a novel optimization method to detect SVs from next-generation sequencing data. Our method employs relatedness between samples – specifically a child and both parents – to improve the ability of signal reconstruction in the presence of noisy genomic data. By enforcing both sparsity of SVs – that they are rare within the genome – and the fact that any SV carried by a child must be present in one of the two parents, we considerably improve the sensitivity and specificity of detection on both simulated SV data and real

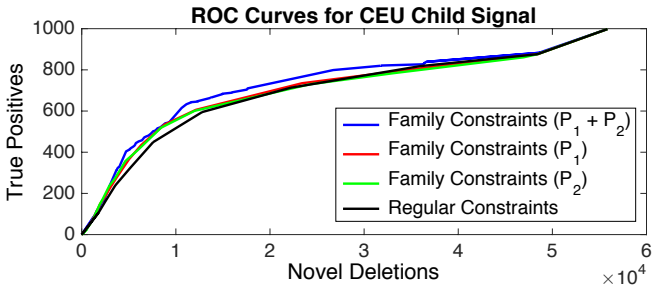

Fig. 4: ROC curves illustrating the novel deletions (validated set of deletions may be incomplete) vs. true positives in the signal of the child data from the parent-child trio of the CEU population studied in the 1000 Genomes Project Pilot 1. In this trio, both parents shared 92.5% of variants in common and $\tau = 5$.

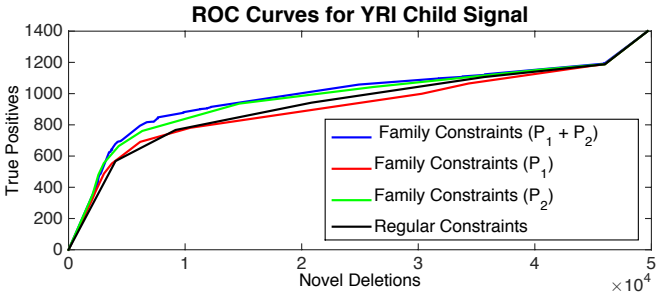

Fig. 5: ROC curves illustrating the novel deletions (validated set of deletions may be incomplete) vs. true positives in the signal of the child data from the parent-child trio of the YRI population studied in the 1000 Genomes Project Pilot 1. In this data, both parents shared 89.96% of variants in common and $\tau = 1$.

sequencing data from the 1000 Genomes Project. Our work suggests that the high-false positive rate of prediction suffered by most traditional SV algorithms can be considerably improved by enforcing the relatedness of individuals sampled in large-sequencing studies.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] M. Banuelos, R. Almanza, L. Adhikari, S. Sindi, and R. F. Marcia, "Sparse signal recovery methods for variant detection in next-generation sequencing data," Accepted *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[9] 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.

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

[11] 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.

[12] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "The Sparse Poisson Intensity Reconstruction ALgorithms (SPIRAL) Toolbox," `http://drz.ac/code/spiraltap/`.

[13] 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.