Constrained Variant Detection with SPaRC: Sparsity, Parental Relatedness, and Coverage

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) are rearrangements of DNA sequences such as inversions, deletions, insertions and translocations. The common method for detecting SVs has been to sequence data from a test genome and map it to a reference genome. More recently, DNA sequencing studies may consist of hundreds, or even thousands of individuals, some of which may be related. In order to improve our ability to identify SVs, we boost the true SV signals by simultaneously analyzing parent and child genomes. Our algorithmic formulation – SPaRC – employs realistic criteria such as sparsity of SVs, relatedness between individuals and variable sequencing coverage throughout the genome.**

I. Introduction

The genome of an individual consists in sequences of nucleotides (A,C,G,T) that can range in length from millions of letters (for a bacteria) or billions of letters (for a mammalian genome). Structural variants (SVs) represent a rearrangement within an individual's DNA sequence – as compared to a reference. Once thought to be primarily associated with genetic diseases like cancer [1], there are many types of SVs, like inversions, deletions and duplications that have been identified in the genomes of health individuals [2], [3]. As such, SVs represent an important part of understanding our recent population history [4], [5].

Recent advances in high-throughput sequencing have further enabled the efforts to understand human population history by making it possible to study hundreds or even thousands of individuals, such as in the recent 1000 Genomes Project [2]. In most SV detection studies, fragments from test genomes are sequenced and mapped to a reference. SVs are identified by statistically significant deviations from expected patterns of paired-reads [6]. However, due to errors in the sequencing and mapping process itself and by relying on data from only one genome at a time, false predictions may be made and true variants may be missed.

We provide an SV detection pipeline that distinguishes itself from traditional methods, in several ways. First, most SV methods consider variants in isolation without enforcing global constraints such as a the expected sparsity of true SVs. Second, despite the fact that many studies contain multiple (often related) individuals, SV predictions are typically made without using the relatedness between individuals. Finally, many DNA sequencing methods are biased by the GC-content – percentage of G/C as opposed to A/T – which varies throughout the genome [7], [8], [9] (see Figure 1). As such, computational methods must be able to accommodate variable coverage throughout the genome in accordance with this bias. In this paper we attempt to address each of these criteria by expanding our maximum likelihood approach [10], [11] for variant detection to incorporate coverage bias in concordant regions by considering the GC-content in the neighborhood of each SV for each sequenced individual. Specifically, we provide an algorithm developed for parent-child trios to incorporate sparsity, parental relatedness and varying sequencing coverage, or SPaRC. We present numerical results on both low-coverage simulated and real sequencing data variant detection. We show that by considering variable sequencing coverage and relatedness between individuals improves the ability to predict true SVs.

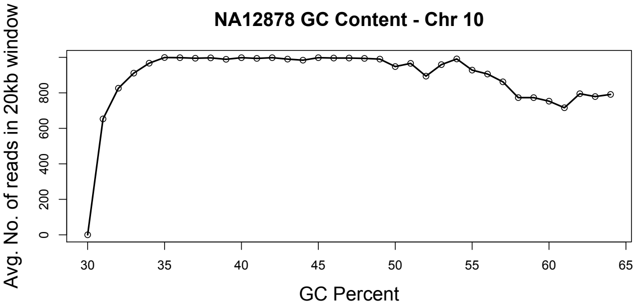

Fig. 1. GC-content vs read coverage in Chromosome 10 in 20 kb windows of the NA12878, which was aligned to March 2006 human reference sequence (NCBI Build 36.1).

II. Method

Here we extend the general framework presented in [11] for detecting structural variants (SVs) given sequencing data from one father-mother-child trio (p_1, p_2 and c). We consider each individual to have only one copy of each chromosome (haploid) for simplicity and n represents the number of locations where a variant can be present. If the individual does not have a variant at a given location, then the true signal is 0. Otherwise, if an SV is present, the signal is 1. The observations $\vec{y}_{p_1}, \vec{y}_{p_2}, \vec{y}_c \in \mathbb{R}^n$ are the number of DNA fragments supporting a possible SV for parents and child. Moreover, we assume that the data follow a Poisson distribution [6]:

$$\vec{y}_i \sim \text{Poisson}(\vec{\sigma}_i[(k_i - \epsilon)\vec{f}_i^*] + \vec{\sigma}_i \epsilon), \tag{1}$$

where $i \in \{c, p_1, p_2\}$ and the constants k_c, k_{p_1}, and k_{p_2} represent the sequencing coverage for each individual. To model coverage bias in different regions in the genome, σ_i scales k_i by accounting for the GC-content in a given window. Finally, we assume that the error in measurement $\epsilon > 0$ is the same for all observations. Letting $A_i = \vec{\sigma}_i(k_i - \epsilon)\mathbb{I} \in \mathbb{R}^{n \times n}$, where $\mathbb{I}$ is the $n \times n$ identity matrix, represent the linear projection of the true genomic variants $\vec{f}_i^* \in \mathbb{R}^n$ to the observation $\vec{y}_i$, and stacking the true variant signals and observations in the form $\vec{f}^* = [\vec{f}_c^*; \vec{f}_{p_1}^*; \vec{f}_{p_2}^*]$ and $\vec{y} = [\vec{y}_c; \vec{y}_{p_1}; \vec{y}_{p_2}]$ respectively, the general observation model can be expressed as

$$\vec{y} \sim \text{Poisson}(\hat{A}\vec{f}^*), \tag{2}$$

where $\hat{A} \in \mathbb{R}^{3n \times 3n}$ is a block-diagonal matrix with $\hat{A} = \text{diag}(A_c, A_{p_1}, A_{p_2})$. Under the model (2), the probability of observing $\vec{y}$ is given by

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

where $\vec{e}_i$ is the i-th column of the $3n \times 3n$ identity matrix. We maximize the probability of observing $\vec{y}$ in (3) using a maximum likelihood approach to determine the true signal $\vec{f}^*$.

Continuous relaxation. The true signal $\vec{f}^*$ is a discrete binary-valued vector. Thus, maximizing $p(\vec{y}\,|\hat{A}\vec{f}^*)$ in (3) will be a combinatorial optimization problem, which is generally very difficult to solve. To overcome this challenge, we relax the space of admissible solutions to include vectors with continuous values, i.e., $\vec{f} \in \mathbb{R}^{3n}$. Consequently, this relaxation allows us to apply gradient-based optimization techniques.

Familial constraints. Since the true signal $\vec{f}^*$ can only take the values $\{0, 1\}$, we require the continuous approximation $\vec{f}$ to be within this interval (i.e., $0 \leq \vec{f} \leq 1$). Further, we impose the element-wise constraints that if both parents have the SV, then the child must also, i.e., $\vec{f}_{p_1} + \vec{f}_{p_2} - 1 \leq \vec{f}_c$. Finally, if neither parent has the SV, then the SV cannot be present in the child, i.e., $\vec{f}_c \leq \vec{f}_{p_1} + \vec{f}_{p_2}$.

Gradient-based optimization. We can reformulate variant detection as the following constrained optimization problem:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{3n}}{\text{minimize}} \quad & \phi(\vec{f}) \equiv F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} \quad & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \leq \vec{f}_c \leq \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & 0 \leq \vec{f} \leq 1 \end{aligned} \tag{4}$$

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

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

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

To reflect the rarity of true variants in the genome, we use the sparsity-promoting ℓ_1 norm, $\|\vec{f}\|_1$ (see [12]) as the penalty function in (4). Following the SPIRAL framework for sparse Poisson reconstruction [13], [14], [15], we solve (4) by minimizing quadratic approximations to $F(\vec{f})$. The Hessian of $F(\vec{f})$ in this approximation is replaced by the scaled identity matrix $\alpha_k I$ ($\alpha_k > 0$) at each iterate $\vec{f}^k$ (for details, see [16], [17], [18]). With $\lambda = \frac{\tau}{\alpha_k}$, this approximation can be simplified to a subproblem of the form:

$$\begin{aligned} \vec{f}^{k+1} = \underset{\vec{f} \in \mathbb{R}^{3n}}{\text{arg min}} \quad & \tfrac{1}{2}\|\vec{f} - \vec{s}^{\,k}\|_2^2 + \lambda\|\vec{f}\|_1 \\ \text{subject to} \quad & \vec{f}_{p_1} + \vec{f}_{p_2} - \mathbf{1} \leq \vec{f}_c \leq \vec{f}_{p_1} + \vec{f}_{p_2}, \\ & 0 \leq \vec{f}_c, \vec{f}_{p_1}, \vec{f}_{p_2} \leq 1, \end{aligned} \tag{5}$$

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 (5) can be separated into scalar minimization problems (see [15] for details). Completing the squares and ignoring constant terms, we have

$$\begin{aligned} \underset{f_c, f_{p_1}, f_{p_2} \in \mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}(f_c - c)^2 + \tfrac{1}{2}(f_{p_1} - p_1)^2 + \tfrac{1}{2}(f_{p_2} - p_2)^2 \\ \text{subject to} \quad & f_{p_1} + f_{p_2} - 1 \leq f_c \leq f_{p_1} + f_{p_2}, \\ & 0 \leq f_c, f_{p_1}, f_{p_2} \leq 1 \end{aligned} \tag{6}$$

where $(c, p_1, p_2) = (s_c - \lambda, s_{p_1} - \lambda, s_{p_2} - \lambda)$. The feasible solution to (6) 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 c-p_1-p_2 tri-dimensional space according to the constraints of (6) (see [11] for more details).

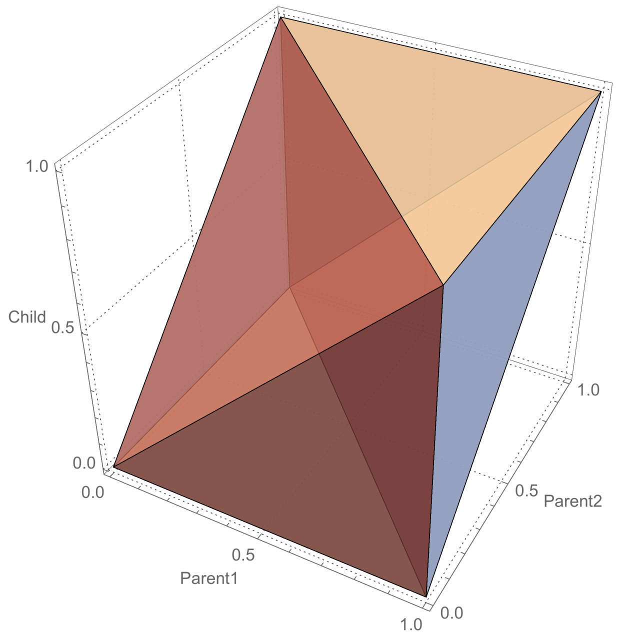

Fig. 2. The three-dimensional feasible region of the minimization problem (6) representing the familial constraints. Note for example that if both f_{p_1} and f_{p_2} are 1, i.e., both parents have the SV, then the child must have the SV as well, i.e., f_c is also 1. Similarly, if neither parents have the SV, i.e., f_{p_1} and f_{p_2} are 0, then f_c must also be 0.

III. RESULTS

We implemented our proposed scaled familially-constrained two-parent method in MATLAB by modifying and incorporating constraints to the SPIRAL framework [19] and performed reconstructions on both simulated and real genomic data. The algorithm is initialized by $\hat{A}^T\vec{y}$ and terminates if the relative difference between consecutive iterates converged to $\|\vec{f}^{k+1} - \vec{f}^k\|_2/\|\vec{f}^k\|_2 \leq 10^{-8}$. The results are compared to the unscaled familially-constrained two-parent method [11] with constant coverage values. The regularization parameters (τ) for all experiments are optimized to get the minimum RMSE $(\%) = 100 \cdot \|\hat{f} - \vec{f}^*\|_2/\|\vec{f}^*\|_2$.

A. Simulated Data

For the simulated data experiments, we generated two true parent variant signals $\vec{f}^*_{p_1}$ and $\vec{f}^*_{p_2}$ of size 10^5 with the level of sparsity 0.5%, i.e., each signal has only 500 variants. These two parents $\vec{f}^*_{p_1}$ and $\vec{f}^*_{p_2}$ are chosen to share a percentage of variants ranging from 0% to 100%, in 25% increments. In addition, another true signal of size 10^5 is also generated as the true child signal $\vec{f}^*_c$, where the value at each point of the child signal is chosen from either parent with equal probability. The scaling vector for the sequencing coverage of each individual $\vec{\sigma}_i$ is chosen to be uniformly distributed in the interval $(0.5, 1)$, i.e., $\vec{\sigma}_i \sim \mathrm{U}(0.5, 1)$. Sequencing coverage values for each individual in unscaled familially-constrained two-parent method are set to $k_i\ \mathbb{E}(\vec{\sigma}_i)$, where $\mathbb{E}(\vec{\sigma}_i)$ is the expected value of the scaling vector $\vec{\sigma}_i$. The error term ϵ in both methods is set to 0.01.

Analysis. The false positive rate vs. true positive rate for the child signal reconstruction with two different set of coverages, $k_{p_1} = 5$, $k_{p_2} = 5$, $k_c = 2$ and $k_{p_1} = 4$, $k_{p_2} = 4$, $k_c = 4$, both with 75% similarity of variants between parents is represented in Fig. 3. Notice the amelioration in predictive power achieved by the scaled familially-constrained two-parent method (see the red curves in Fig. 3) over the unscaled familially-constrained two-parent method (see the blue dashed curves in Fig. 3). Furthermore, we observe a higher improvement in the child signal reconstruction with the scaled method when the parents share more variants in common. We further observe a similar improvement in signal reconstruction for the case of the parent signal reconstructions. However, parent reconstructions show improvement to a smaller extent when compared to the child and do not follow the same increasing pattern.

B. 1000 Genomes Project Trio Data

We next apply our method to 1000 Genomes Project [20] father-mother-daugher CEU trio data (NA12891, NA12892, NA12878). All genomes were aligned to NCBI36 and sequence at low coverage ($\approx 4\times$) in Pilot 1 of the study. We obtain observations of possible variants using GASV with WRITE_CONCORDANT option on CEU trio data. To obtain the bias of coverage, we consider a sliding window of size 20kb. Using the kent tool hgGcPercent from [21] on the March 2006 NCBI 36 release of the human genome (hg18.2bit from http://hgdownload.soe.ucsc.edu), we calculated the GC content for this reference [22]. We aggregate the total number of concordant regions within the window using interval trees to determine if either a start or end position lies within this window. The biased coverage λ_i is calculated by accounting for the median number of reads in a given window in a similar framework as [8]. As such, we have $\lambda_i = \lambda_{\text{expect.}} \frac{M}{M_{GC}}$, where M represents the median read counts in all the windows, M_{GC} represents the median counts with the same GC percentage, and λ_{expect} is the expected coverage for the entire genome of the individual. Reconstructions are compared against reported experimentally validated deletions longer than 250bp. Observations marked *LowQual* as well as regions near centromeres or telomeres were also filtered from validated deletions. The resulting true signal had a sparsity level of 1.77%, consistent with rarity of SV assumption.

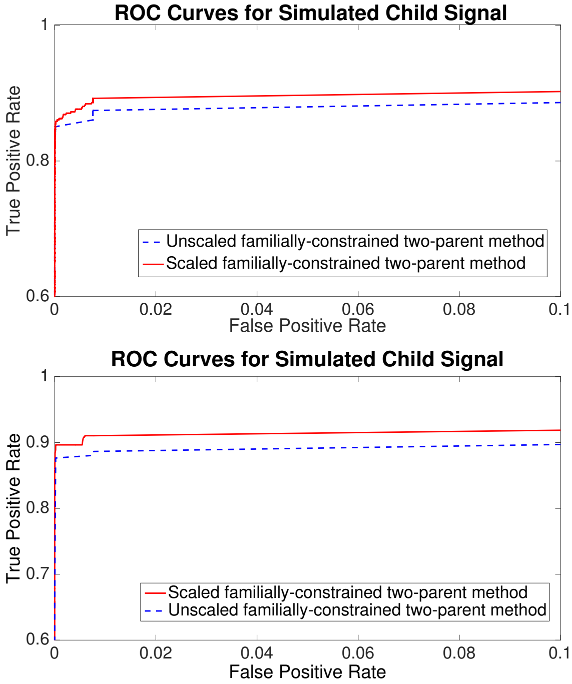

Fig. 3. ROC curves illustrating the true positive rate vs. false positive rate for the child signal reconstruction. From top to bottom: ROC curves for child signal reconstruction with $k_{p_1} = 5$, $k_{p_2} = 5$, $k_c = 2$; and ROC curves for child signal reconstruction with $k_{p_1} = 4$, $k_{p_2} = 4$, $k_c = 4$. Both cases have a 75% similarity of variants between parents for each scaled and unscaled two-parent familially-constrained methods. Notice the scaled two-parent method recovers the simulated child signal with higher true positive rate than the existing unscaled method.

Analysis. For the CEU trio, parents shared the majority of variants as well as similar coverage bias. Figure 4 depicts novel deletions versus true positives. Since the validated set of deletions may not be complete, we note that our method may correctly identify true deletions not in the experimentally validated set. Figure 4 also illustrates that the addition of coverage bias does not provide any advantage in the detection of SVs in the real data. We expect the rest of the CEU genomes to reflect similar results due to dominant deletion signals obscuring any affect of variable coverage. Moreover, the parental signal provides much better support for a variant in relation to the child signal prediction than coverage related to GC-content bias. Due to the relatively large window of 20kb and the heterogeneity of GC-content throughout even one chromosome in an individual, we may see an improvement with a more narrow sliding window to determine bias of coverage. Since *a priori* knowledge regarding concordant regions in the genome is required for this method to determine $\vec{\sigma}_i$, additional computation is required with respect to the previous familially-constrained method.

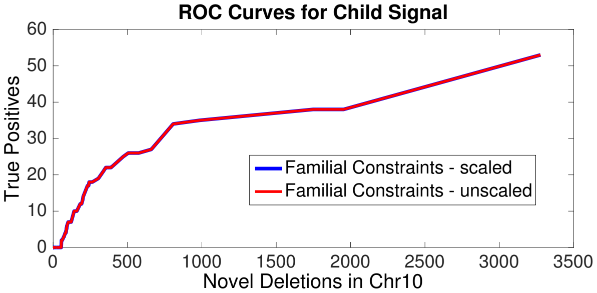

Fig. 4. ROC curves depicting novel deletions vs validated (true) deletions for NA12878 Chromosome 10 with $\tau = 10$ and $\epsilon = 0.01$. Coverage bias determined by GC-content did not yield any additional advantage and similar results are expected for remaining CEU trio genomes.

IV. Conclusions

We present a novel optimization method to detect SVs from next-generation sequencing data utilizing sparsity of SVs, relatedness between individuals and variable sequencing coverage throughout the genome. Our method incorporates relatedness of sequenced trios and considers GC-content bias coverage to improve variant detection in next generation sequencing data. For the simulated data, we improve on the specificity and sensitivity of the previous familially-constrained model. Although coverage bias, as determined by 20kb windows, did not improve SV prediction in the CEU data set from 1000 Genomes Project, we describe similar model performance in light of parental signals and deletion signal strength in comparison to GC-adjusted coverage.

By employing relatedness of individuals sampled in large-sequencing studies during the SV prediction phase – instead of post-processing – our method provides a means to reduce high-false positive rate of prediction. Future directions of this work focuses on comparing our results to other methods of SV detection. Including a sparsity-promoting penalty and GC-content bias in our predictions yields analytical solutions to our algorithm, making extending our work to analyze populations of low-coverage related individuals tractable.

References

[1] 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. 125–138, 2013.

[2] 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.

[3] J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, et al., "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, no. 7191, pp. 56–64, 2008.

[4] 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.

[5] H. Stefansson, A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. G. Gudnadottir, et al., "A common inversion under selection in europeans," *Nature genetics*, vol. 37, no. 2, pp. 129–137, 2005.

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

[7] V. Boeva, A. Zinovyev, K. Bleakley, J.-P. Vert, I. Janoueix-Lerosey, O. Delattre, and E. Barillot, "Control-free calling of copy number alterations in deep-sequencing data using gc-content normalization," *Bioinformatics*, vol. 27, no. 2, pp. 268–269, 2011.

[8] S. Yoon, V. Xuan, Z.and Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome research*, vol. 19, no. 9, pp. 1586–1592, 2009.

[9] D. Iakovishina, I. Janoueix-Lerosey, E. Barillot, M. Regnier, and V. Boeva, "Sv-bay: structural variant detection in cancer genomes using a bayesian approach with correction for gc-content and read map-pability," *Bioinformatics*, p. btv751, 2016.

[10] 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.

[11] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, and S. Sindi, "Sparse genomic structural variant detection: Exploiting parent-child relatedness for signal recovery," Accepted *IEEE Workshop on Statistical Signal Processing*, 2016.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[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, "Sparsity-regularized photon-limited imaging," in *Proceedings of IEEE International Symposium on Biomedical Imaging*, Rotterdam, The Netherlands, April 2010.

[15] 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.

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

[17] 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.

[18] S. J. Wright, R. D. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, 2009.

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

[20] 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.

[21] W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik, "Bigwig and bigbed: enabling browsing of large distributed datasets," *Bioinformatics*, vol. 26, no. 17, pp. 2204–2207, 2010.

[22] K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, et al., "The ucsc genome browser database: 2015 update," *Nucleic acids research*, vol. 43, no. D1, pp. D670–D681, 2015.