NEGATIVE BINOMIAL OPTIMIZATION FOR BIOMEDICAL STRUCTURAL VARIANT SIGNAL RECONSTRUCTION

Mario Banuelos, Suzanne Sindi, and Roummel F. Marcia

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

ABSTRACT

Structural variants (SVs) – novel adjacencies in an individual's genome – lead to genomic diversity across all organisms. When DNA fragments of an unknown genome are compared to a reference genome, errors in sequencing and mapping obscure true genomic rearrangements. When the sequencing coverage is low, this may lead to high false positive rates in predicted SVs. In this paper, we propose a novel maximum likelihood approach to SV prediction incorporating low-coverage sequencing data and coverage distribution. Specifically, we address mean and variance assumptions proposed by Poisson models and develop a Negative Binomial framework which reflects a more accurate representation of DNA fragments in an individual's genome. We incorporate both sparsity and inheritance in our model with an ℓ_1 penalty and linear constraints, respectively. We validate our model on both simulated and real genomic data of related individuals. Moreover, our results indicate an improvement on thresholding observations of candidate variants.

Index Terms— Sparse signal recovery, nonconvex optimization, structural variants, computational genomics

1. INTRODUCTION

Structural variations (SVs) – genomic rearrangements longer than one basepair long – account for a large portion of genetic diversity in all organisms (see Fig. 1). In humans, SVs have been associated with genomic diseases, but the vast majority appear to be harmless variants passed from parents to offspring [1, 2]. The dominant method for detecting SVs in an individual's genome is to sequence their genome, which results in many short DNA sequence reads, and then map (or align) these reads to a high-quality reference genome. Regions of the sequenced genome that differ from the reference correspond to SVs in the sequenced genome. The locations of SVs in the sequenced genome are thus computationally determined by identifying clusters of reads in discordant arrangements [3, 4]. As DNA sequencing continues to advance, and produce ever longer DNA reads, most methods to detect SVs still suffer from high error rates associated with the sequencing and mapping process [3]. While one solution would be to sequence individuals to extremely high coverage, this comes at increased financial and computational cost. Moreover, portable sequencing technology provides the opportunity to sequence many individuals at low coverage relatively quickly [5].

During the sequencing process, if genomic fragments are randomly chosen from the genome, then the Poisson distribution describes the number of reads covering any genomic locus [6]. The Poisson assumption with a mean represented by the *coverage* also assumes the same variance. However, sequencing technologies are known to be biased, resulting in large variation of coverage depth. This is particularly true in low-coverage settings [7, 8, 9]. In this regime, studies suggest that the two parameter negative binomial distribution may be more accurate in describing the distribution of fragments [10, 11].

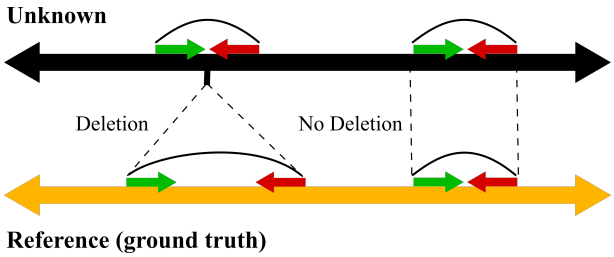

Fig. 1. Illustration of regions in sequenced genome where there is a deletion (*left*) and no deletion (*right*) relative to a reference genome (ground truth). When sequenced fragments of the unknown genome do not map concordantly to the reference genome, we consider this a signal for a potential deletion or other structural variants (SVs). Note that for a deletion, the fragment from the individual maps to a larger than expected region in the reference. Fragments aligning to the reference in a concordant fashion indicate there is no genomic variation.

We note that many computational methods exist for processing mapped fragments and predicting SVs [12, 13, 14, 15, 16]; however most are based on only the mapped fragments and do not utilize other information about SVs if available. For example, SVs are relatively rare in an individual's

This work was supported by NSF Grant IIS-1741490 and the UC Merced Fletcher Jones Fellowship.

genome, but most methods do not attempt to rank or prioritize predictions by how likely they are. This results in many false positive predictions because fragments that have been mapped to incorrect locations in the genome are likely to be mistaken as an SV [17, 18, 19]. In addition, when analyzing related individuals, who should share SVs, variant detection methods only use relatedness to filter calls as a post-processing step [12, 14, 15]. While some computational methods utilize the probability of arrangements of fragments, allowing them to estimate the probability a prediction is false or to rank their predictions by likelihood, most methods rely on the assumption of Poisson coverage [20]. Overall, most computational methods suffer from high false positive rates, but high-coverage and high quality data tend to resolve many false calls [3, 21].

In this work, we aim to improve upon past SV prediction methods in primarily three ways. Whereas previous work assumed mapped reads follow a Poisson distribution, we incorporate a negative binomial distribution to model the distribution of fragments [10, 22, 23, 24]. Instead of assuming equal mean and variance, we estimate both from the data and the negative binomial model captures the large variability in the sequencing coverage. Fig. 2, for example, provides empirical examples of this phenomena from the 1000 Genomes Project [17]. Secondly, we incorporate low-coverage data instead of relying on high-quality genomic data. Finally, we concurrently consider sequencing data of related individuals and enforce inheritance of variants through inequality constraints.

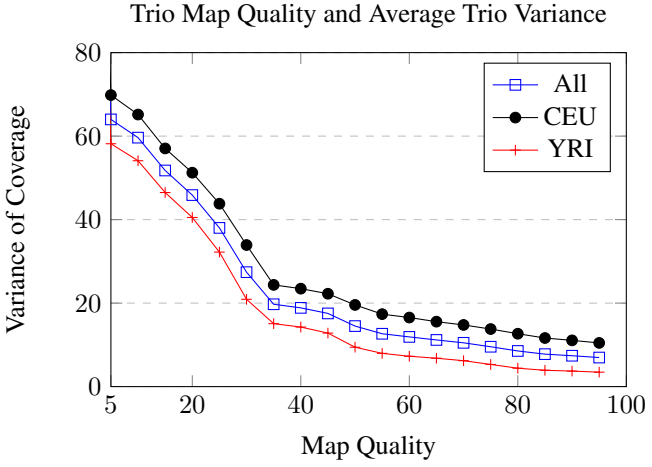

Fig. 2. Plot of the map quality vs depth of coverage variance (mean per trio reported) for European (CEU) trio, Yoruba (YRI) trio, and both trios (father-mother-child) genomes from the 1000 Genomes Project. Varying the minimum map quality of reads, we calculate the depth of coverage for each genomic locus. The data show a much higher variance than the expected coverage of ≈ 4X.

2. NEGATIVE BINOMIAL LOG-LIKELIHOOD OPTIMIZATION

We consider the true signal $\vec{f}^* \in \{0,1\}^n$ to be a binary vector indicating the presence of a genetic variant, with $\vec{f}^*_j = 1$ if a variant is present at location j and 0 otherwise [22, 24, 25]. Thus, the corresponding parent $\vec{y}_p$ and child $\vec{y}_c$ observations are given by

$$\vec{y}_p \sim \text{NegBin}(\vec{\mu}_p, \vec{\sigma}^2_p) \quad \text{and} \quad \vec{y}_c \sim \text{NegBin}(\vec{\mu}_c, \vec{\sigma}^2_c), \quad (1)$$

where mean μ_i and variance $\sigma^2_i, (i \in \{p, c\})$ of depth of coverage will be determined by the sequencing data of each respective individual. We consider the stacked child-parent signal $\vec{y} = [\ \vec{y}_p^T \quad \vec{y}_c^T\ ]^T$ and corresponding mean and variance vectors, $\vec{\mu}$ and $\vec{\sigma}^2$. (Here, the notation $\vec{\sigma}^2$ is to be understood component-wise.) In particular, we have the following expressions for the components of $\vec{\mu}$ and $\vec{\sigma}^2$:

$$(\mu)_j = \left(A\vec{f}^*\right)_j \quad \text{and} \quad (\sigma)^2_j = \left(A\vec{f}^*\right)_j + \frac{1}{r}\left(A\vec{f}^*\right)^2_j,$$

where A, representing expected sequencing coverage, linearly projects the true signal $\vec{f}^*$ onto the n-dimensional set of observations, and r is the dispersion parameter of the negative binomial distribution.

Problem Formulation. When $r \to \infty$, we have $\sigma = \mu$ and this reduces to the Poisson case. If we choose to estimate these parameters from the sample, then we must observe a variance higher than the mean. Under this model, the probability of observing $\vec{y}$ is given by the following:

$$p(\vec{y}) = \prod_{j=1}^{n} \binom{y_j + \frac{\mu_j^2}{\sigma_j^2 - \mu_j} - 1}{y_j} \left(\frac{\mu_j}{\sigma_j^2}\right)^{\frac{\mu_j^2}{\sigma_j^2 - \mu_j}} \left(1 - \frac{\mu_j}{\sigma_j^2}\right)^{y_j}. \quad (2)$$

Ignoring constant terms, the negative log-likelihood term, $F(\mu, \sigma^2)$, becomes

$$F(\mu, \sigma^2) \equiv \sum_{j=1}^{n} -\log\left(\frac{\left[y_j + \frac{\mu_j^2}{\sigma_j^2+\mu_j} - 1\right]!}{(y_j)!\left[\frac{\mu_j^2}{\sigma_j^2+\mu_j} - 1\right]!}\right) - \frac{\mu_j^2}{\sigma_j^2 - \mu_j} \log\left(\frac{1}{\sigma_j^2}\mu_j\right) - y_j \log\left(1 - \frac{1}{\sigma_j^2}\mu_j\right). \quad (3)$$

Maximizing variance. Without reverting to the use of Gamma functions for $r \in \mathbb{R}$, we assume $r \in \mathbb{Z}^+$ and we know $\sigma_j^2 = \mu_j + \frac{1}{r}\mu_j^2$, where σ_j^2 is maximized when $r = 1$. Thus, we can rewrite the probability (2) of observing $\vec{y}$ as

$$p(\vec{y}) = \prod_{j=1}^{n} \left(\frac{1}{1+\mu_j}\right)\left(\frac{\mu_j}{1+\mu_j}\right)^{y_j}, \quad (4)$$

with associated negative log-likelihood,

$$F \equiv \sum_{j=1}^{n} (y_j + 1)\log(1 + \mu_j) - y_j \log(\mu_j).$$

However, we know the mean $\mu_j = e_i^T Af$. Then, adding the small parameter ε to represent sequencing or mapping error, we have

$$F(f) \equiv \sum_{j=1}^{n} (y_j+1) \log\left(1 + e_i^T Af + \varepsilon\right) - y_j \log\left(e_i^T Af + \varepsilon\right), \tag{5}$$

with gradient

$$\nabla F(f) = \sum_{j=1}^{n} \frac{y_j + 1}{1 + e_i^T Af + \varepsilon} A^T e_i - \frac{y_j}{e_i^T Af + \varepsilon} A^T e_i. \tag{6}$$

Continuous Relaxation. To apply calculus of variations approaches in this classification problem, we allow for f to take on continuous values in $[0, 1]$. Otherwise, the combinatorial optimization problem may be intractable with a maximum-likelihood approach. As such, the negative binomial reconstruction algorithm takes the following form of the following constrained optimization problem for a one-parent and one-child (P, C) model:

$$\begin{aligned} \underset{\vec{f} \in \mathbb{R}^{2n}}{\text{minimize}} \quad & \psi(\vec{f}) \equiv F(\vec{f}) + \tau \, \|\vec{f}\|_1 \\ \text{subject to} \quad & 0 \leq \vec{f_c} \leq \vec{f_p} \leq \mathbf{1}, \end{aligned} \tag{7}$$

where $\vec{f} = \left[\, \vec{f_p}^T \;\; \vec{f_c}^T \,\right]^T$, $\mathbf{1}$ is a vector of ones, and τ is a regularization parameter. We assume that a child will have an SV at a certain location only if the parent also has the SV at the same location. We enforce this through the linear constraint $0 \leq \vec{f_c} \leq \vec{f_p} \leq 1$. Using a gradient-descent approach, the next iterate in our estimation is given by

$$\vec{f}^{k+1} = \left[\vec{f}^k - \alpha_k \nabla F(\vec{f}^k) + \tau \mathbf{1}\right]_{P,C}, \tag{8}$$

with step size (learning rate) α_k and the operation $[\cdot]_{P,C}$ is a projection onto the feasible set defined by the linear constraints in (7) (see [22] for further details).

3. RESULTS

We evaluate the effectiveness of the proposed method on both simulated and real genomic data and compare our reconstructions with thresholding observations $\vec{y}$ and previous Poisson models. The proposed method is implemented in Python 3.6. We explored ten logarithmically-spaced regularization parameters τ from a 10^{-2} to 10^2 grid and chose the value yielding the largest average maximum area under curve for the receiver operating characteristic (ROC) using 5-fold cross-validation. To determine the number of true and false positives, we threshold the reconstructed signal – thereby un-relaxing our continuous assumption. For all experiments, we set $\alpha = 0.01$. The algorithm terminates if the relative difference between consecutive iterates $\|\vec{f}^{k+1} - \vec{f}^k\| / \|\vec{f}^k\|_2 \leq 10^{-6}$ or exceeds the maximum number of iterates.

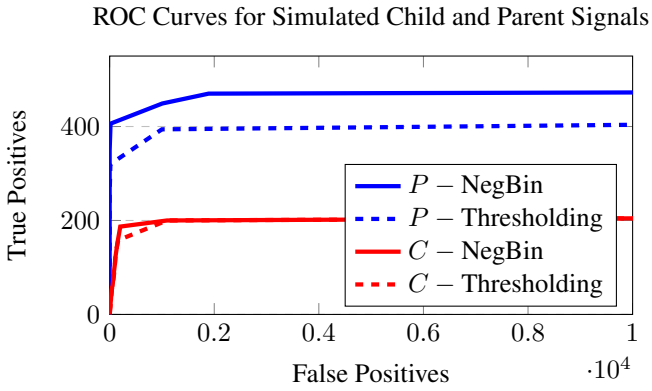

Fig. 3. ROC curves illustrating the number of false positives vs the number of true positives for both the parent and child signal reconstruction with $\mu_p = 4$, $\mu_c = 4$, $\varepsilon = 0.01$, and 50% similarity. In both reconstructions, we set $\tau = 1.6681$ based on 5-fold cross validation. We observe more true positives using our proposed method when compared to thresholding the signal. This is particularly true in the first thousand predictions.

3.1. Simulated Data

We simulated two signals $\vec{f_p}$ and $\vec{f_c}$, representing the parent and child signals respectively. Each candidate set of SVs were drawn from a negative binomial distribution with dispersion parameter $r = 1$ and mean $\mu_p = \mu_c = 4$. We observe $n = 10^5$ potential SV candidates for each individual, with 500 true variants for $\vec{f_p}$, and 250 inherited variants for $\vec{f_c}$. This reflects a 50% similarity level and we set $\varepsilon = 0.01$ to represent the mapping and sequencing error in the forward model.

Analysis. We first examine the parent signal reconstruction. Fig. 3 presents the number of false positive vs true positives for the reconstruction of the parent and child signals with mean coverage $\mu_p = \mu_c = 4$, $r = 1$, and $\varepsilon = 0.01$. Although $n = 10^5$, we focus on a more detailed view in the ROC curve to discern differences in prediction. Based on AUC measurements, we immediately observe an improvement in the number of true predictions over thresholding with our proposed model. For the 1 Parent-1 Child model, we expect parental reconstructions to be more informed by the child signal [22].

For the reconstructed child signal, we observe a marginal improvement when the number of false positives is relatively low. We note, however, that both the parent and child reconstructions incorporate a penalty of $\tau = 1$. This is an improvement on our previous methods, which typically resulted in tuning τ for each individual [26, 27].

3.2. 1000 Genomes Project Data

We applied our method to both sequenced genomes of the father-mother-daughter trios from European (CEU) and

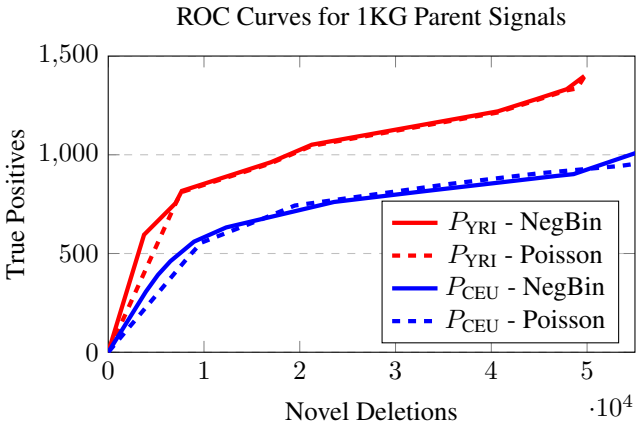

Fig. 4. ROC curves illustrating the number of false positives vs the number of true positives for parent signal reconstruction with $\mu_p = \mu_c = 4, \varepsilon = 0.01, \tau = 0.01$ for both CEU and YRI populations. We observe an improvement in true predictions across both signals of interest.

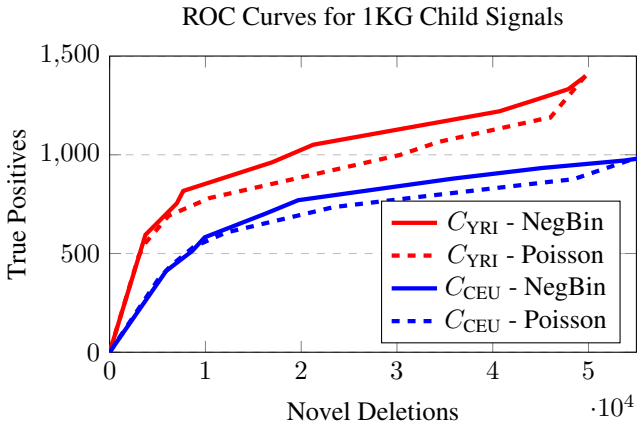

Fig. 5. ROC curves illustrating the number of false positives vs the number of true positives for child signal reconstructions with $\mu_p = \mu_c = 4, \varepsilon = 0.01, \tau = 0.01$ for both CEU and YRI populations. For a fixed number of novel deletions, we report a higher number of true positives.

Yoruba (YRI) populations. All six individuals from the 1000 Genomes Project were sequenced to $\approx$ 4X in Pilot 1 and aligned to NCBI36 [17]. We consider experimentally validated deletions meeting the following criteria as the true deletions: longer than 250bp, not *LowQual*, and non-overlapping with centromere and telomere regions.

We implemented GASV [13] on this data to obtain the candidate variant set. The intersection between the candidate SVs and true deletions results in the true signal $\vec{f}^*$. We observe high variability in expected coverage in Fig. 2 and thus threshold the minimum map quality at 10 for all individuals.

Analysis. We note a higher area under the curve in ROCs of the reconstructed signals for both CEU and YRI populations in Fig. 4 and 5 in comparison to the previous Poisson model. Additionally, this Fig. 4 depicts the number of novel deletions vs. true (experimentally validated) deletions since not the true set may be incomplete. Although not pictured, we observe similar trends for p_2 in CEU and YRI populations. Next, we consider the reconstruction for the child signals for both CEU and YRI populations. Fig. 5 illustrates a small but measurable difference in true predictions for both with the same τ across all individuals.

4. CONCLUSIONS

We propose a novel optimization method to detect structural variants from sequencing data of related individuals. Our method addresses mean and variance assumptions of previous methods and incorporates both relatedness and sparsity into the signal reconstruction. In future studies, we will relax the integer assumption on the dispersion parameter r to generalize our method and accommodate higher variances in sequencing data. We also intend to incorporate other optimization approaches for this nonconvex formulation.

5. REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, et al., "The potential and challenges of nanopore sequencing," *Nature biotechnology*, vol. 26, no. 10, pp. 1146–1153, 2008.

[6] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, 1988.

[7] 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.

[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] 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.

[10] J. Sampson, K. Jacobs, M. Yeager, S. Chanock, and N. Chatterjee, "Efficient study design for next generation sequencing," *Genetic epidemiology*, vol. 35, no. 4, pp. 269–277, 2011.

[11] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, "Sequencing depth and coverage: key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 121–132, 2014.

[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. 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.

[14] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, "Delly: structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics*, vol. 28, no. 18, pp. i333–i339, 2012.

[15] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, et al., "Breakdancer: an algorithm for high-resolution mapping of genomic structural variation," *Nature methods*, vol. 6, no. 9, pp. 677–681, 2009.

[16] F. Hormozdiari, C. Alkan, E. E. Eichler, and S. C. Sahinalp, "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes," *Genome research*, vol. 19, no. 7, pp. 1270–1278, 2009.

[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] D. F. Gudbjartsson, H. Helgason, S. A. Gudjonsson, and et al., "Large-scale whole-genome sequencing of the Icelandic population," *Nature Genetics*, vol. 47, no. 5, pp. 435 – 444, 2015.

[19] J. Huddleston and E. E. Eichler, "An incomplete understanding of human genetic variation," *Genetics*, vol. 202, no. 4, pp. 1251–1254, 2016.

[20] S. S. Sindi, S. Önal, L. C. Peng, H. Wu, and B. J. Raphael, "An integrative probabilistic model for identification of structural variation in sequencing data," *Genome biology*, vol. 13, no. 3, pp. R22, 2012.

[21] S. Shetty, *Structural Variant Detection*, Ph.D. thesis, Arizona State University, 2014.

[22] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*.

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

[24] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, and S. Sindi, "Constrained variant detection with sparc: Sparsity, parental relatedness, and coverage," 2016, Proceedings of *International Conference of the IEEE Engineering in Medicine and Biology Society*.

[25] M. Banuelos, L. Adhikari, R. Almanza, A. Fujikawa, J. Sahagún, K. Sanderson, M. Spence, S. Sindi, and R. F. Marcia, "Sparse diploid spatial biosignal recovery for genomic variation detection," in *Medical Measurements and Applications (MeMeA), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 275–280.

[26] M. Banuelos, L. Adhikari, R. Almanza, 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.

[27] 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.