A Neural Network Approach for Anomaly Detection in Genomic Signals

Erica Sawyer*, Mario Banuelos*, Roummel F. Marcia†, and Suzanne Sindi†
* California State University, Fresno, Department of Mathematics, Fresno, CA, USA
E-mail: mbanuelos22@csufresno.edu
† University of California, Merced, Department of Applied Mathematics, Merced, CA, USA
E-mail: rmarcia@ucmerced.edu

***Abstract*—Structural variants (SVs) are observed differences between the sequenced genome of an individual as compared to a reference genome for that species. These differences include deletions, inversions, insertions, and duplications. Since some variations are associated with certain diseases, our work focuses on developing methods to detect such genomic anomalies. Current DNA sequencing methods may be costly and existing SV-detection techniques often rely on high quality data. We present a deep learning method to identify deletions in DNA based on genomic information of related individuals.**

In this paper, we implement neural networks to predict SVs as a means to reduce the false positive rates of existing methods. A neural network - a sequence of linear and nonlinear transformations - takes in training data and uses that information to learn how to classify corresponding test data. Our preliminary model incorporates the observed genomic information of two parents and an offspring to predict locations of SVs in the genome of the child. We also investigate the performance of this model under different neural network architectures using various performance metrics. With these limited features and low-quality data, we propose a generalization of our model that allows for the simultaneous prediction of SVs in all three individuals.

***Index Terms*—Structural Variation, Deep Learning, Computational Genomics, Biomedical Signal Processing**

I. Introduction

Structural variants (SVs) are observed differences, longer than one base pair, between the sequenced genome of an individual as compared to a reference genome for that species. These differences are generally greater than 50 base pairs in length and include deletions, inversions, insertions, and duplications [1], [2]. Since some variations are associated with disease, our work focuses on developing methods to detect such genomic changes [3].

Because of their relevance to disease and their record of human evolutionary history, detecting SVs has become a major scientific question. The common method for SV identification is to sample fragments from an individual's genome and compare these fragments to a high-quality reference. Regions where these fragments are consistent with the reference suggest the absence of an SV. This process of comparison is known as mapping, and the precise characteristics of consistency depend on the specific sequencing technology used [4], [5]. Computational methods for structural variation detection began to appear in the early 21st century when the cost of next-generation sequencing decreased to the point that the genomes of many individuals could be sequenced [6], [7], [8], [9]. In recent years, SV detection algorithms have emerged to cope with ever-changing genomic technologies [10], [11]. However, the majority of SV detection algorithms perform poorly when the average sequencing depth is low. It is also challenging to separate the signal of a true SV from erroneous mappings and most existing algorithms have difficulties resolving boundaries of SVs [12], [13]. One successful approach to improving SV detection is to use multiple SV detection methods and report a consensus series of predictions [14]. In this case the weak signal is "boosted" by combining SV predictors. We take an alternate approach by leveraging the knowledge that since SVs are shared by closely related individuals, we can boost the signal of a true SV by simultaneously predicting SVs in multiple related individuals at the same time.

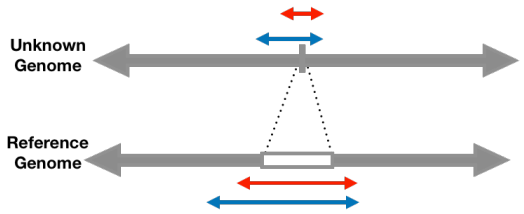

Fig. 1. **Example Deletion Structural Variant.** The unknown genome (top) has a deletion (missing segment) relative to a reference or known genome (bottom). The signal for a deletion SV comes from multiple fragments (blue and red arrows) which are sampled from the test genome and whose ends map to a longer distance than expected in the reference.

In this work, we use neural networks to predict the location of SVs by simultaneously considering related individuals. For simplicity, we focus only on the deletion SV. (A *deletion* structural variant occurs when a portion of the reference genome is not present in the sample genome of an individual. See Figure 1). Our first model considers two parents and their shared offspring, predicting deletions in the child's genome based on the observed presence of DNA fragments in all three individuals. We then expand the model to simultaneously predict deletions in both parents and the child from the same input of genomic information. Further, we compare the performance of each model to preexisting techniques of deletion calling.

II. Method

We consider a framework for reducing false positive predictions in structural variation (SV) genomic signals in a two parent-one child family. As such, we have data from two parents (p_1 and p_2) and one offspring c, for n potential variation locations. For simplicity, we consider haploid signals (either a variant is present or not), where the truth for individual i takes a binary value $\vec{f}_i \in \{0,1\}^n$. In the case of $m = 3$ individuals, this results in a total of 8 possible combinations, or classes, describing variant presence.

A. Observational Model

The observed data represent the number of DNA fragments supporting a potential SV at particular genomic loci. We assume the observed data, x_i, for each individual i follows a Poisson distribution [4], i.e.,

$$\vec{x}_i \sim \text{Poisson}\left((\lambda_i - \epsilon)\vec{f}_i + \epsilon\right), \quad (1)$$

where λ is the expected DNA sequencing coverage and ϵ represents the sequencing and alignment errors. When λ increases, $\vec{x}_i$ can be approximated by a normal distribution and methods exists to classify such signals (see e.g., [15]). Related work (see [16] and the references therein) based on sparse optimization techniques [17] infer $\vec{f}_i$ from $\vec{x}_i$ using the maximum likelihood principle while leveraging parent-child relationships formulated as optimization constraints. Machine learning-based approaches have also been previously used for detecting SVs [18], [19], [20], [21]. Our proposed method differs from these existing method in two ways: 1) we consider familial relationships simultaneously and 2) we differentiate between types of inheritance with no pre-processing and limited features in related individuals. In turn, this allows us to develop the following framework without relying on convolutional layers in our proposed architectures.

B. Classification Frameworks

For simplicity, we focus on classifying deletion structural variants. Given this noisy data, we consider the following two classification approaches:

Predicting Offspring Variation. We first developed a model for a specific case of our problem, where we use the two parents and child observational data to predict the presence of a variant in the offspring. This simplification of detecting SV presence in the child results in a binary classification problem for reconstructing the truth signal f_c.

Simultaneous Trio Prediction. To further use the relatedness of the individuals, we expand our classification framework to simultaneously predict the location of deletions in all trio members. We enumerate and describe all variant possibilities in Table I. We reduce the 8 categories to 5 possible classes, according to relatedness information. Since we only consider the presence of an SV and do not account for the number of copies, we note that the Class 4 contains the possibility of a variant in both parents which is absent in the child (i.e., $f_{p_1} = f_{p_2} = 1$ and $f_c = 0$).

c	p_1	p_2	**Class**	**Description**
0	0	0	0	No SV
1	0	0	1	*De novo* (novel) SV
1	0	1	2	Inherited SV
1	1	0		
0	0	1	3	Non-inherited SV
0	1	0		
1	1	1	4	Parent SV (present in both parents)
0	1	1		

TABLE I
CLASS DEFINITIONS AND DESCRIPTIONS FOR SIMULTANEOUS PREDICTION FOR c-p_1-p_2 TRIO.

C. Neural Network Approach

We use fully-connected neural network models which take the genomic information of related individuals and output predicted deletion locations for those individuals. In particular, we consider neural networks with exactly two hidden layers where the first hidden layer results from a linear transformation of the input values and the second hidden layer consists of linear transformations on the first hidden layer. After a nonlinear transformation on the second hidden layer and the application of the sigmoid function, the model produces the desired output as class probability. Additionally, we allow each layer to have between 2 and 5 nodes, providing a total of 16 different models to explore (see Figure 2 for the proposed neural network architectures).

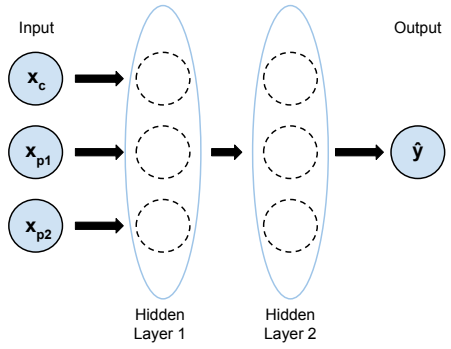

Fig. 2. Neural network architecture with inputs x_c, x_{p1} and x_{p2} and output $\hat{y}$ with variable widths of Hidden Layers 1 and 2.

We use the `nn.CrossEntropyLoss()` loss and encode the respective classes as outlined above [22]. We refer to specific models as having 3-l-m-1 architectures, where the 3 denotes the number of inputs (the genomic information of all trio individuals), l represents the width of layer 1, m represents the width of layer 2, and 1 represents the number of outputs (where outputs are either binary for the child prediction models or one of 5 classes in the simultaneous prediction case).

III. Numerical Experiments

Next, we consider both simulated and real genomic data of related individuals for our models. We implemented the subsequent models in Python using the open source machine

learning framework PyTorch. Models were trained on 40-60, 50-50, and 60-40 training-testing splits on a commodity machine with 8 GB of RAM and an Intel i5 processor. The neural networks were trained using the default stochastic gradient descent optimizer Adam, 1000 epochs, with learning rate = 0.01 [23].

A. Simulated Data

We simulated truth signals $\vec{f}_c, \vec{f}_{p_1}, \vec{f}_{p_2}$, along with the corresponding observed data $\vec{x}_c, \vec{x}_{p_1}, \vec{x}_{p_2}$ with $n = 10^5$ potential deletion locations. For all individuals, the expected coverage was set $\lambda = 4$ and the error term was defined to be $\epsilon = 0.01$. In each individual, only 500 true variants are present.

B. 1000 Genomes Data

We train and test our models previously sequenced data from the 1000 Genomes project [24]. Specifically, we use the CEU trio which consists of three Utah residents with European ancestry (namely two parents and their child). After calling deletions with GASV, the data input to our method consists of the total number of fragments supporting a potential deletion. For each of the three individuals, we consider 57,078 genomic locations. Of these positions in the observed data, approximately 2% were experimentally validated as true deletions.

C. Data Imbalance

Due to the imbalanced nature of the data, we also discuss models trained on an upsampled training set, where minority classes are oversampled, so that the models are trained on a balanced dataset across classes. We note that Class 0 comprises approximately 98% of all the data, Class 4 makes up 1.5%, and the remaining classes comprise the rest of the data. To evaluate the performance of our models, we will explore AUC (Area Under the Receiver Operator Curve), test loss, Top-1 accuracy, and Top-3 accuracy.

IV. Results

For both of our approaches, we observed high accuracy in signal reconstruction for the simulated data. We report an AUC of 0.99 for predicting offspring deletions and a Top-1 accuracy of 0.99 (results not displayed). As such, we focus on the results for CEU trio from the 1000 Genomes Consortium. For our first approach, we predict child SV locations and found that the best model in terms of AUC was the neural network with $l = 5$ and $m = 4$. This model was created using a 50/50 train/test split. We note that this model architecture produces an AUC 0.09 higher than the GASV model proposed in [25] (see Figure 3).

For simultaneously predicting deletions in all trio individuals, we measure performance using Top-1 test accuracy, since we are unable to calculate the AUC (as previously discussed) for this multi-class classification problem. The highest Top-1 test accuracy is 0.8718 and is produced by the 3-3-4-1 neural network model which was trained on 60% of the full dataset. Figure 4 shows the top 1 test accuracies for all 16 models created with a 60/40 split.

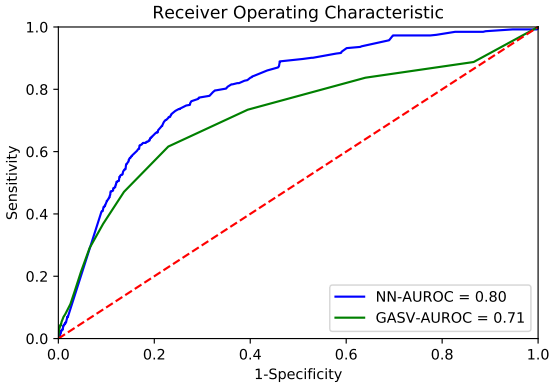

Fig. 3. ROC and Area Under the Curve (AUC) for 3-5-4-1 model (blue) compared to GASV ROC (green) when predicting CEU offspring deletion locations.

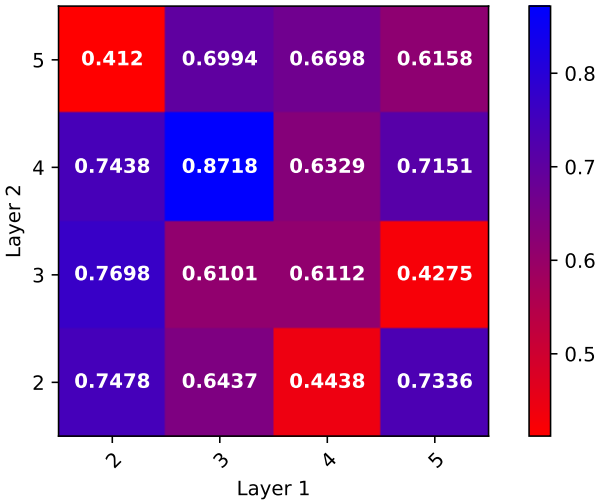

Fig. 4. Top-1 accuracy across all classes for the CEU 60/40 train-test split, for all explored neural network architectures.

Further, we can evaluate the Top-1 accuracies of this model by class, as shown in Table II. This model performs very well for classes 0 and 3, and Top-3 accuracies indicate potential to correctly classify other SV inheritance patterns. When compared to multinomial logistic regression, we see an 8% improvement in class 4 predictions. We also observe architectures (i.e. 3-2-3-1) which perform better at detecting *de novo* deletions, with a Top-1 Accuracy of 83.3% for class 1. These results further warrant more exploration of using a limited set of features in a population to better predict the distribution of types of structural variants in related individuals.

V. Conclusions

We present a neural network framework to detect SVs in DNA sequencing data from parent-child trios. This method makes use of relatedness between the individuals to improve signal reconstruction of low quality data. Moreover, this

Class	0	1	2	3	4
Top 1 Accuracy	88.6	0	2.4	52.6	13.2
Top 3 Accuracy	91.4	0	97.6	73.7	93.4

TABLE II
TOP-1 AND TOP-3 TEST ACCURACIES (%) BY CLASS OF 3-3-4-1 MODEL ON REAL DATA, WHERE OUR METHOD IMPROVES UPON DETECTION NON-INHERITED SVS PRESENT IN A PARENT. WE NOTE THAT A MAJOR IMPROVEMENT IN THE TOP-3 ACCURACY FOR CLASS 2, 3, AND 4.

work aims to fill a gap in the currently available research by incorporating familial data into machine learning models and simultaneously calling SVs in multiple individuals. We present results for both simulated and real data from the 1000 Genomes Project and our framework is adaptable to individual and simultaneous predictions. In future work, we intend to incorporate more individuals of a population.

REFERENCES

[1] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85–97, 2006.

[2] D. L. Cameron, L. D. Stefano, and A. T. Papenfuss, "Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software," *Nature Communications*, vol. 10, 2019.

[3] J. Weischenfeldt, F. Symmons, O.and Spitz, and J. Korbel, "Phenotypic impact of genomic structural variation: insights from and for human disease," *Nature Reviews Genetics*, vol. 14, no. 2, pp. 125–138, 2013.

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

[5] P. Guan and W.-K. Sung, "Structural variation detection using next-generation sequencing data: a comparative technical review," *Methods*, vol. 102, pp. 36–49, 2016.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] M. C. Stancu, M. J. Van Roosmalen, I. Renkens, M. M. Nieboer, S. Middelkamp, J. De Ligt, G. Pregno, D. Giachino, G. Mandrile, J. E. Valle-Inclan, et al., "Mapping and phasing of structural variation in patient genomes using nanopore sequencing," *Nature communications*, vol. 8, no. 1, pp. 1–13, 2017.

[11] S. Chan, E. Lam, M. Saghbini, S. Bocklandt, A. Hastie, H. Cao, E. Holmlin, and M. Borodkin, "Structural variation detection and analysis using bionano optical mapping," in *Copy Number Variants*, pp. 193–203. Springer, 2018.

[12] M. Banuelos, *Developing Statistical Models for the Analysis of Genomic Variants*, Ph.D. thesis, UC Merced, 2018.

[13] P. Guan and W. K. Sung, "Structural variation detection using next-generation sequencing data: A comparative technical review.," *Methods*, vol. 102, pp. 36–49, 2016.

[14] Y. Xia, Y. Liu, M. Deng, and R. Xi, "Svmine improves structural variation detection by integrative mining of predictions from multiple algorithms," *Bioinformatics*, vol. 33, no. 21, pp. 3348–3354, 2017.

[15] J. O. Korbel, A. Abyzov, X. J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. B. Gerstein, "Pemer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data," *Genome Biology*, vol. 10, no. 2, pp. R23, 2009.

[16] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi, "Detecting inherited and novel structural variants in low-coverage parent-child sequencing data," *Methods*, vol. 173, pp. 61–68, 2020.

[17] 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.

[18] G. H. Lubke, C. Laurin, R. Walters, N. Eriksson, P. Hysi, T. D. Spector, G. W. Montgomery, N. G. Martin, S. E. Medland, and D. I. Boomsma, "Gradient boosting as a snp filter: An evaluation using simulated and hair morphology data," *Journal of Data Mining in Genomics & Proteomics*, vol. 4, 2013.

[19] H. Park, S. Chun, J. Shim, J. Oh, E. J. Cho, H. S. Hwang, J. Lee, D. Kim, S. J. Jang, S. J. Nam, et al., "Detection of chromosome structural variation by targeted next-generation sequencing and a deep learning application," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[20] D. Antaki, W. M. Brandler, and J. Sebat, "SV^2: accurate structural variation genotyping and de novo mutation detection from whole genomes," *Bioinformatics*, vol. 34, no. 10, pp. 1774–1777, 2018.

[21] E. Alzaid and A. E. Allali, "Postsv: A post–processing approach for filtering structural variations," *Bioinformatics and Biology Insights*, vol. 14, pp. 1177932219892957, 2020.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds., pp. 8024–8035. Curran Associates, Inc., 2019.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] 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.

[25] S. S. Sindi, E. Helman, A. Bashir, and B. J. Raphael, "A geometric approach for classification and comparison of structural variants," *Bioinformatics*, vol. 25, pp. i222 – i230, 2009.