Relation Augmentation: A Gradient Boosting Approach for Detecting Genomic Anomalies

Anonymous Author(s)
Affiliation
Address
email

Abstract

Genomic anomalies, or variations, are often shared between members of the same species. Although rare, these changes may result in disease or an increase in host fitness. Most approaches for detecting structural variation rely on high quality data and are typically limited to one type of structural variant such as deletions or inversions. We propose a new data augmentation method to mitigate errors in low quality DNA sequencing data by leveraging offspring DNA information to predict genomic variants in their associated parent. To our knowledge, this is the first time such an approach has been proposed to detect multiple structural variation classes, including complex variants, using related individual data. The author(s) of this work identify as Latinx.

1 Introduction - Biological Motivation

Advances in DNA sequencing have increased the number of large sequencing studies, with the goal of quantifying genomic variation and its influence on both genotypes and phenotypes in species [1, 2, 3]. These genomic anomalies may appear as a single basepair change (SNP) or as a rearrangement of a larger region. Detecting these larger regions, known as structural variants (SVs), remains a challenging problem. This is particularly true when incorporating low quality DNA sequencing data [4, 5, 6]. The ability to detect such genomic variation remains an important area of study, as these changes have applications in detecting negative and positive outcomes (e.g. cancer susceptibility and increased fitness) [7, 8, 9, 10].

Of particular interest is being able to detect complex structural variants, including deletions, inversion, translocations, or duplications [11]. This typically involves aligning sample DNA to a known reference genome and looking for differences. In certain cases, these variations may occur simultaneously, resulting in complex variants (see Fig. 1) [12]. Our work builds upon previous methods by using related individuals – in this case, offspring data – as a data augmentation method [13, 14, 15, 16, 17].

2 Method

Our method was designed with the intention of identifying the various structural variants from features provided from aligned sequencing data using `samtools` [18]. Rather than identifying all of the associated structural variants using a single model, each model was trained to identify each variant individually. This required converting a multi-class problem into a series of binary classification problems. By focusing each model on a single class, the capability of the model is more robust to the task and will not suffer because of the severe class imbalances present in other classes. After testing on various ensemble methods, the optimal method for classification across all classes was gradient boosting. Gradient boosting is an additive method that seeks to minimize a given cost function by incorporating new decision trees to compensate for the shortcomings of previous decision trees. New

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

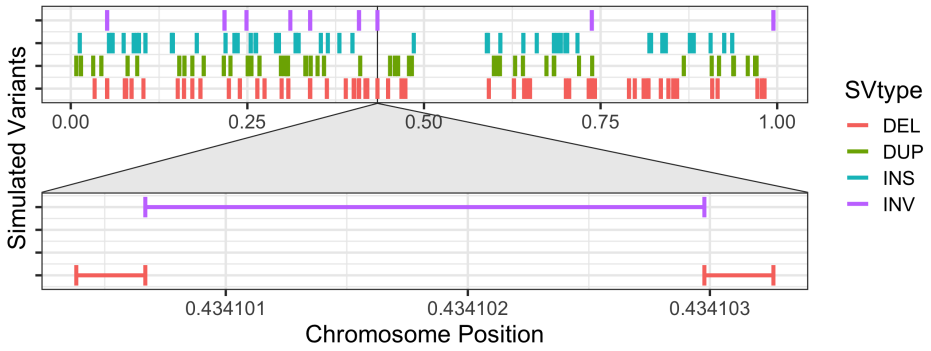

Figure 1: *Top.* We simulate the parent signal by introducing a set of 169 structural variants (deletions, duplications, insertions, and inversions) in Chromosome 1 of the human genome reference (relative position shown). *Bottom.* We illustrate inversion-deletions – known as complex structural variants – where variation occurs simultaneously and an inversion event is flanked by two small deletion events.

trees are added to the ensemble using a form of functional gradient descent where the new tree's parameters are chosen to minimize the loss of the cost function [19].

2.1 Simulated Data

We first simulate a number of structural variants in an individual and then try to predict the structural variants introduced in the simulated data. Moreover, we wish to correctly identify when there are complex rearrangements introduced with low-quality data. The parent signal was simulated using SURVIVOR with 50 (of each) duplications, insertions, and deletions, 4 inversions, and 5 inversion-deletion events [20]. This tool was also used to simulate 8 offspring based on this parent.

3 Results

Four models were trained on aligned, low quality* DNA sequencing offspring data and then tested on the parent sequencing data. Fig. 2 demonstrates the effectiveness of our method and we note that as the number of offspring increases, the area under the curve (AUC) increases with the number of offspring. Although our method only uses discordant data, sequencing data that did not match to the reference, our model was able to correctly classify 13 insertion and 6 inversion (3 regular, 3 complex) events, while Lumpy detected 0 insertions and the same 6 inversions. For future work, we plan on incorporating more features with concordant and split-read sequencing data for all individuals. We also plan on taking advantage of stacking (an ensemble learning technique) to improve classification.

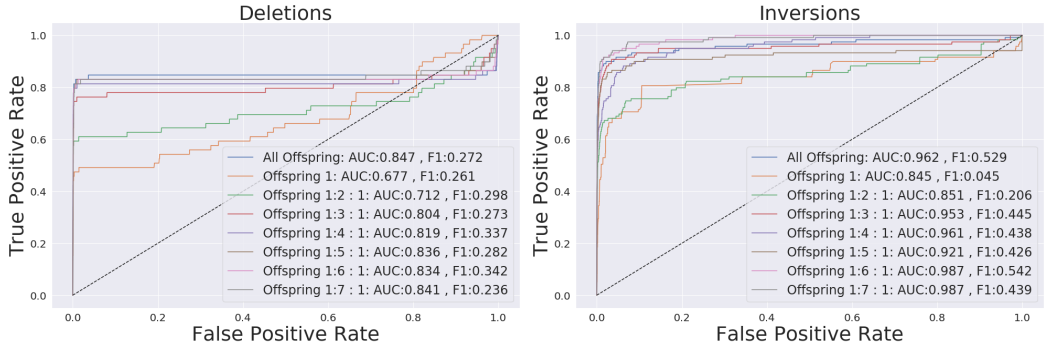

Figure 2: *Left.* Receiver operating characteristics (ROC) curve for deletions for the parent signal using 1 through 8 offspring sequencing data. *Right.* ROC curve for inversions for the parent signal training our model with 1 through 8 offspring sequencing data. Both figures include the AUC and F1 score as performance metrics.

References

[1] . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[2] 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.

[3] Z. R. Chalmers, C. F. Connelly, D. Fabrizio, L. Gay, S. M. Ali, R. Ennis, A. Schrock, B. Campbell, A. Shlien, J. Chmielecki, *et al.*, "Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden," *Genome medicine*, vol. 9, no. 1, p. 34, 2017.

[4] E. P. Consortium *et al.*, "Identification and analysis of functional elements in 1% of the human genome by the encode pilot project," *Nature*, vol. 447, no. 7146, p. 799, 2007.

[5] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping," *Nature Reviews Genetics*, vol. 12, no. 5, p. 363, 2011.

[6] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, "The database of genomic variants: a curated collection of structural variation in the human genome," *Nucleic acids research*, vol. 42, no. D1, pp. D986–D992, 2013.

[7] 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, p. 125, 2013.

[8] I. Martincorena and P. J. Campbell, "Somatic mutation in cancer and normal cells," *Science*, vol. 349, no. 6255, pp. 1483–1489, 2015.

[9] C. Jeong, D. B. Witonsky, B. Basnyat, M. Neupane, C. M. Beall, G. Childs, S. R. Craig, J. Novembre, and A. Di Rienzo, "Detecting past and ongoing natural selection among ethnically tibetan women at high altitude in nepal," *PLOS Genetics*, vol. 14, pp. 1–30, 09 2018.

[10] G. A. Gnecchi-Ruscone, P. Abondio, S. De Fanti, S. Sarno, M. G. Sherpa, P. T. Sherpa, G. Marinelli, L. Natali, M. Di Marcello, D. Peluzzi, *et al.*, "Evidence of polygenic adaptation to high altitude from tibetan and sherpa genomes," *Genome biology and evolution*, vol. 10, no. 11, pp. 2919–2930, 2018.

[11] 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.

[12] Z. Stephens, C. Wang, R. K. Iyer, and J.-P. Kocher, "Detection and visualization of complex structural variants from long reads," *BMC bioinformatics*, vol. 19, no. 20, p. 508, 2018.

[13] 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.

[14] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall, "Lumpy: a probabilistic framework for structural variant discovery," *Genome biology*, vol. 15, no. 6, p. R84, 2014.

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

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

[17] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi, "Detecting inherited and novel structural variants in low-coverage parent-child sequencing data," *Methods*, 2019.

[18] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[20] D. C. Jeffares, C. Jolly, M. Hoti, D. Speed, L. Shaw, C. Rallis, F. Balloux, C. Dessimoz, J. Bähler, and F. J. Sedlazeck, "Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast," *Nature communications*, vol. 8, p. 14061, 2017.