Genotyping Structural Variants with Deep Learning

Jacob Wallace ‘21 & Alderik van der Heyde ‘21, Advisor: Michael Linderman
Computer Science, Middlebury College

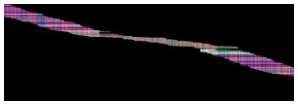
Goal: Accurately predict genotype (1 of 3 “classes”) for structural variants (SV), >50 nucleotides, in whole genome sequencing (WGS) data

Input: SV and WGS data

Classifier

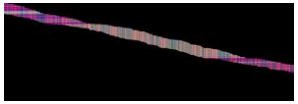
Output: 0 **1** 2

Real Data:



Labeled Synthetic Data:

Class 0



Class 1

Class 2

Siamese Network

Neural Network

Identical

Neural Network

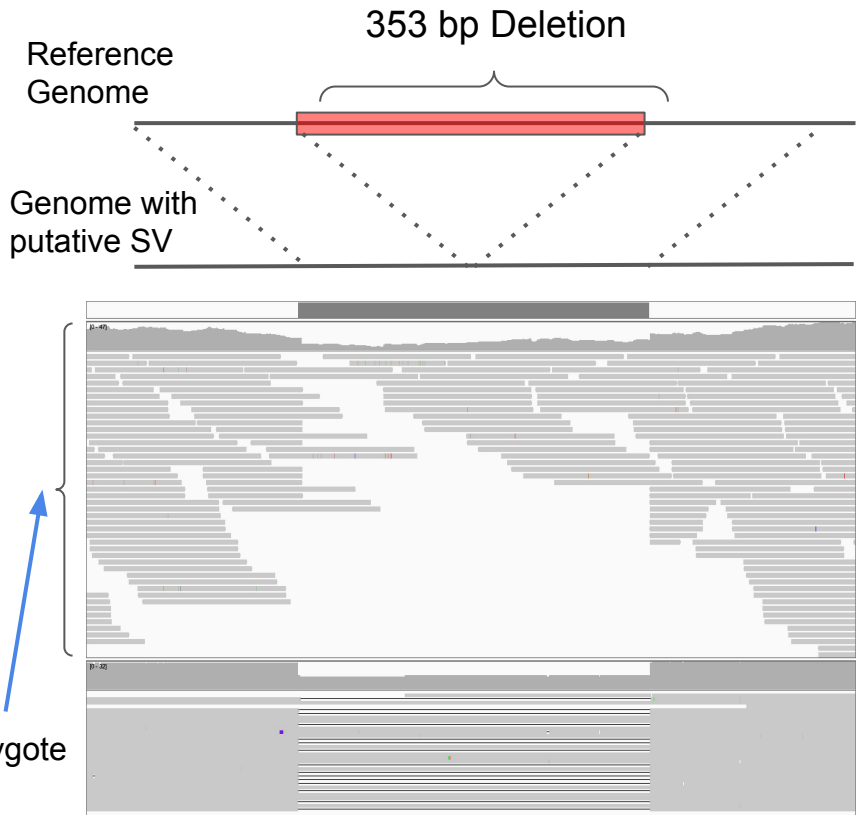
Similarity Function

Similarity Score

Classify based on max similarity

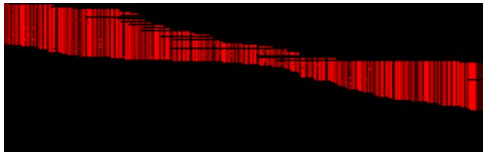
Problem - Genotyping Structural Variants (SVs)

- SVs play a causal role in many genetic diseases
- SVs are difficult to detect with short-read sequencing (NGS) because the variant is larger than the read length
- Genotyping determines the number of copies (0, 1, 2) or zygosity (*homozygous reference*, *heterozygous*, or *homozygous alternate)* of a putative SV

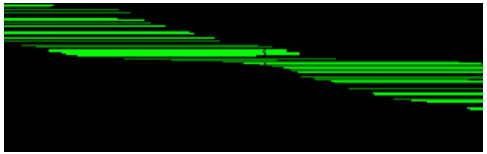


Pileup Image Generation

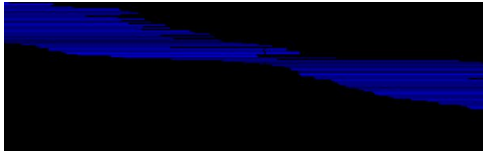
Red: Nucleotide Base (A,T,G,C)



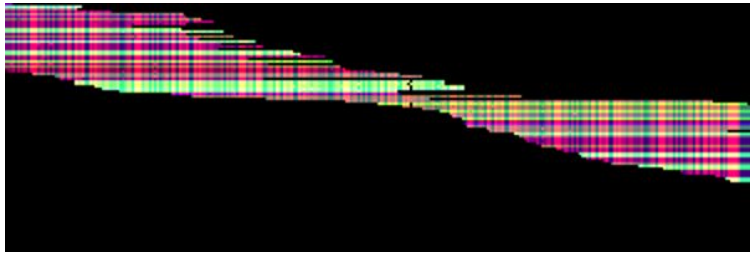
Green: Read Supports Reference or Alternate Allele



Blue: Insert Size (z-score)



Final Combined Image



Models and Results for GIAB "Tier 1" SV deletions[1]

Model	Paired Network	Similarity Function	Het. vs. Hom Alt. (SVs <200 bp)	Het. vs. Hom Alt. (SVs <900 bp)	All Genotype: (SVs <900 bp)
CNNSim[2]	CNN	Euclidean distance	87.9%	92.1%	74.6%
Kaggle[3]	CNN	FCN layers	88.3%*	91.1%*	77.5%
Omniglot[4]	CNN	Single FCN layer	88.7%	91.8%	79.4%
Xception[5]	Xception CNN	Single FCN layer	86.1%	89.5%	70.7%

In order of increasing model complexity

*Ensemble Model

Existing tools have genotype concordance of 60.8-87.4%

[1]Zook, J.M., Hansen, N.F., Olson, N.D. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnology* (2020)
[2]https://github.com/keras-team/keras/blob/master/examples/mnist_siamese.py and https://github.com/kerasteam/keras/blob/master/examples/mnist_cnn.py
[3]https://www.kaggle.com/kmader/image-similarity-with-siamese-networks
[4]https://towardsdatascience.com/one-shot-learning-with-siamese-networks-using-keras-17f34e75bb3d
[5]http://sujitpal.blogspot.com/2017/04/predicting-image-similarity-using.html

Discussion

- We can successfully treat SV genotyping as an image similarity problem
- Can accurately distinguish genotypes without labeled real variants using simulated training data
- Compressing large variants to smaller fixed image size doesn’t harm accuracy and may actually improve accuracy
- Ensemble prediction can improve accuracy for some models (e.g. “Kaggle”)
- Future work:
  - Experiment with different loss function
  - Tune hyperparameters
  - Incorporate real data into training
  - Incorporate more data into image (e.g. mapping quality)
  - Speedup model training to permit evaluating larger models over longer training durations