Detecting Mosaic Mutations with Deep Learning

Poster Session

Mentor: Dr Yufeng Shen
Team: Mingfang Chang, Karan Rao, Yadin Rozov

Problem definition

- Mosaicism: mutations in DNA that arise after fertilization
- Can lead to autism, congenital heart disease
- Need to separate genuine mutations from errors/artifacts
  - Artifacts arise from improper sequence reading and alignment
- Current methods are manual; we aim to use deep learning
  - Training set preparation is a major task as we lack large labeled datasets (for true mosaics)
  - Need to generate trainable representations of our data as well
  - Modify and choose parts of existing frameworks (such as DeepVariant)

Mosaics, artifacts, germline variants

- Mosaics, artifacts have a low variant allele fraction (VAF)
  - VAF: proportion of non-reference reads
  - One is a mutation and one is an rare error, hence low VAF
- Germline variants: mutations in germ cells (which give rise to gametes), passed to offspring
  - VAF is 50% (heterozygous) or 100% (homozygous)
  - Either one parent is different or both are
- Our dataset simulates mosaics (positives) and uses real artifacts (negatives)
  - Heterozygous germline variant reads are downsampled so their VAFs decrease to match the mosaic VAF distribution
  - Mendelian errors become our negatives, and are in fact mostly artifacts

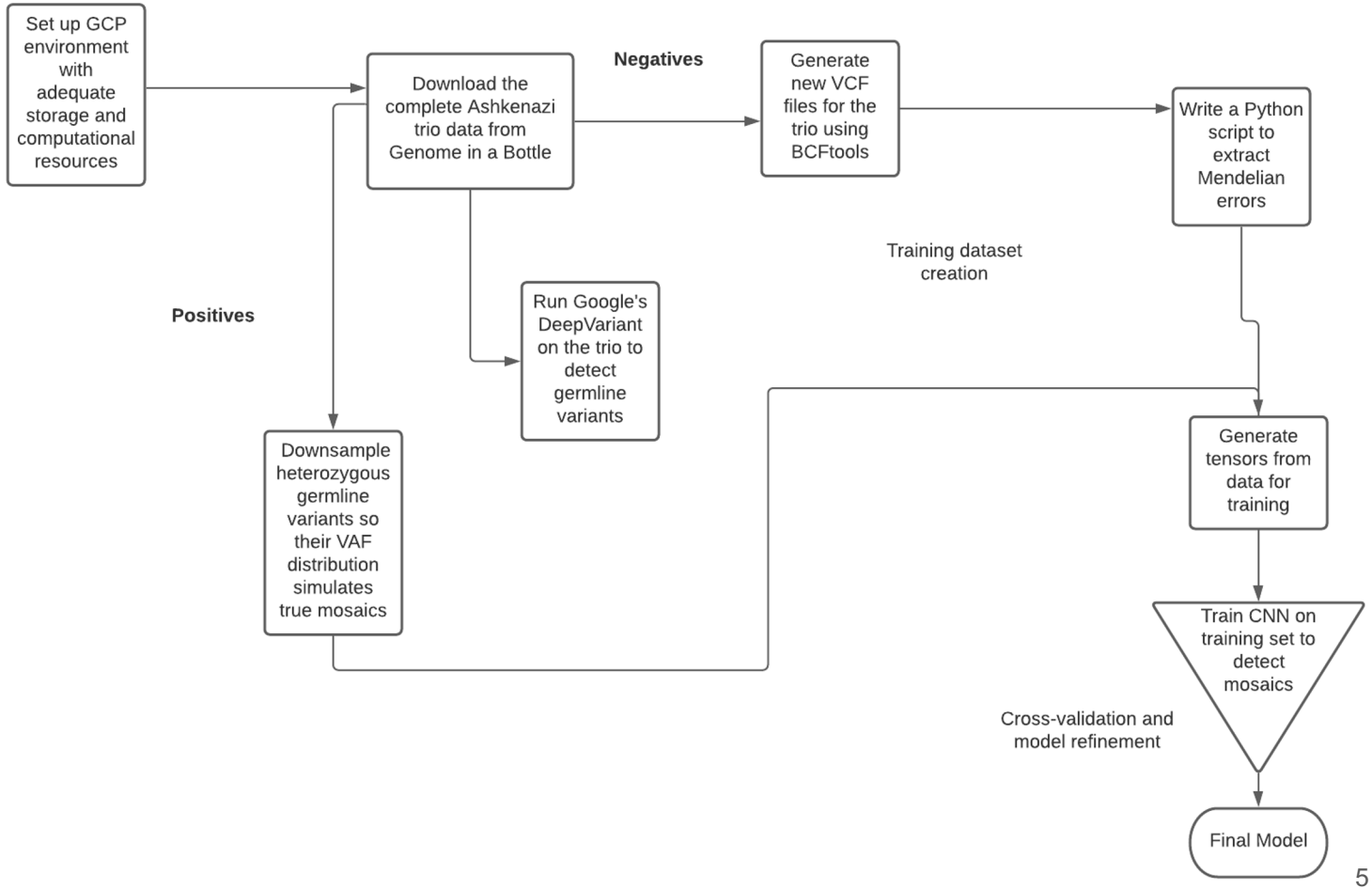
Data and genomics file formats

BAM (SAM, CRAM): binary Sequence Alignment File, stores which read segments are mapped to which sections of the genome (and how well)

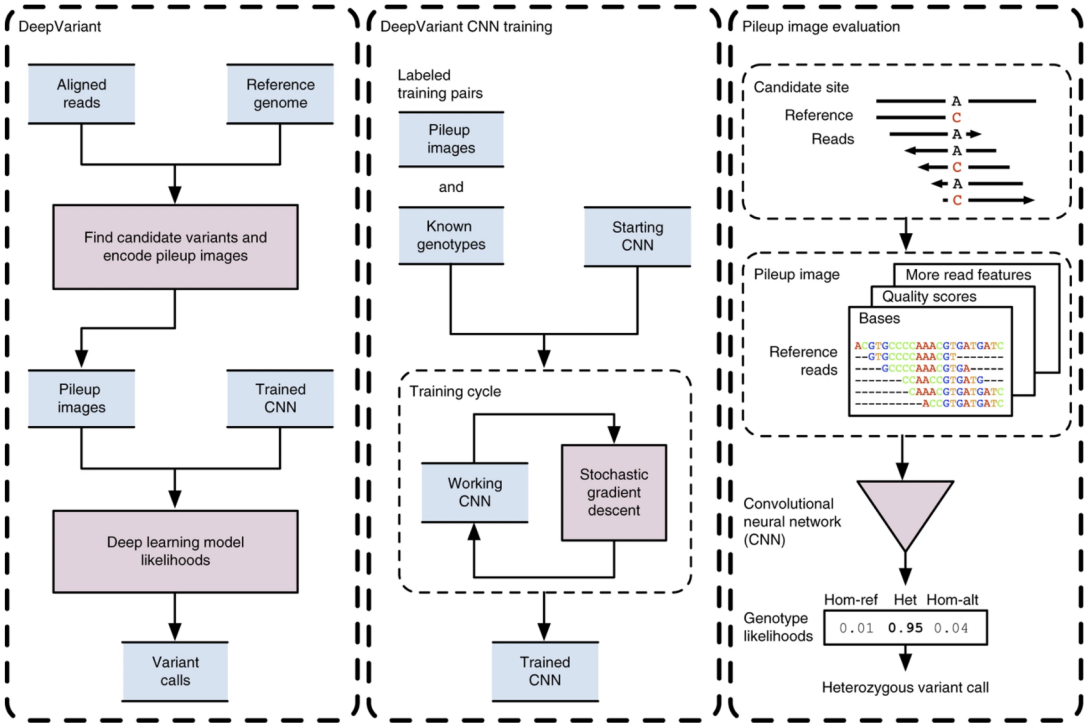
VCF: Variant Call Format, contains the sites where the sequence differs from the reference

We worked with the Ashkenazi Jewish trio of father, mother, and son. Data from Genome in a Bottle (GIAB) https://www.nist.gov/programs-projects/genome-bottle

Workflow


Set up GCP environment with adequate storage and computational resources
Download the complete Ashkenazi trio data from Genome in a Bottle
Negatives
Generate new VCF files for the trio using BCFtools
Write a Python script to extract Mendelian errors
Training dataset creation
Positives
Run Google's DeepVariant on the trio to detect germline variants
Downsample heterozygous germline variants so their VAF distribution simulates true mosaics
Generate tensors from data for training
Train CNN on training set to detect mosaics
Cross-validation and model refinement
Final Model

DeepVariant Workflow



Generating positives

- Identify heterozygous germline variants with Python package ‘scikit-allel’
- Use Python package Pysam to extract reads from BAM
- Downsample reads to match the VAF distribution of mosaics
  - Het germline variant VAF: binomial with mean 0.5
  - Mosaic VAF: binomial with mean VAF << 0.5, square of mean VAF modeled as a Beta distribution
  - Control which reads are included
- Need to match VAFs otherwise neural networks will use this trivial difference to distinguish them

Generating negatives

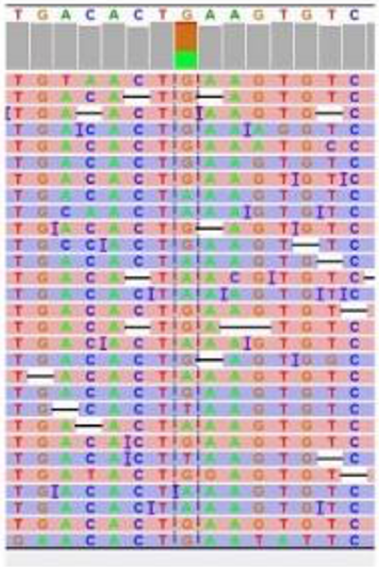
- Generate new VCFs which include low-VAF variants with special BCFtools parameters
- Search for Mendelian errors - variants in child not explained by either parent
  - Work with VCF files of the trio
  - Look for VAF < 0.4
- Most Mendelian errors found with this method will actually be artifacts
  - This is what we want!
- Found chromosomes and positions where variants existed and exported them into a TRUTH_VCF
  - Input to make_examples()

Generating pileup tensors

- DeepVariant: analysis pipeline to call germline variants from sequences
- make_examples() generates pileup “images” or tensors with 6 channels, and adds labels (which we need to modify)

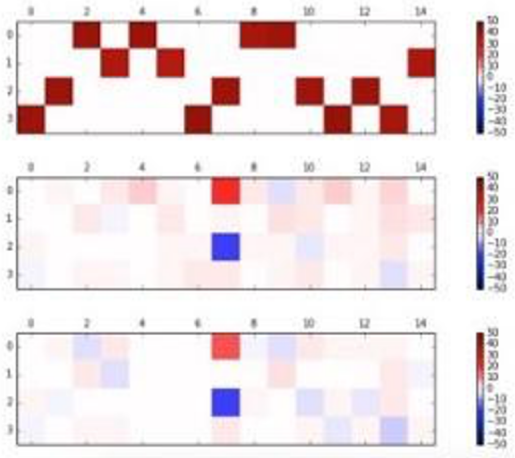
- Jason Chin’s VariantNET generates 15x4x3 tensors
  - For the reference: 7 bases flanking the variant on each side, one-hot for the presence of 4 bases (hence 4 rows)
  - 2 other tensors track differences between reference and sequence

Sequence alignments



Each candidate +/- 7 bp

Alignment Tensor

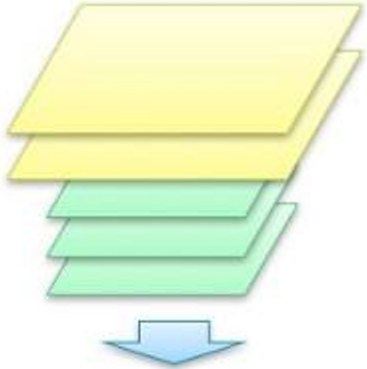


Encode all alignments to a 15 x 4 x 3 tensor

VariantNET

2 convolution layers

3 full connected layers



Genotype

[0.5, 0.0, 0.5, 0.0]

A C G T

Softmax output

[0.98, 0, 0, 0.02]

het none

hom complex

Training

Plan was to train on Inception (CNN) through DeepVariant

Currently working on Jason Chin’s VariantNET, a simpler CNN

Challenges faced

- Low-VAF VCF generation takes ~15 hours for a 50GB BAM
  - Child BAM was corrupted; alternative file is 600GB and has the necessary read depth, but time required is prohibitive
  - Proceeded with packaged VCFs though they do not have low VAFs
- DeepVariant has few (and specific) parameters; could not adapt for our purposes
  - Attempted to make tensors with TRUTH_VCF, standard and generated TRUTH_BEDs, BAM with both positives and negatives, only positives, and all reads
  - 2 main errors with no solutions/documentation; generated a new .bai as directed for one error, to no effect