Deep Variant implementation and algorithms

A short story

F. Raimundo[1]

[1]CEDAR
École Polytechnique

21th September 2017

DeepVariant Paper

Original Paper

"Creating a universal SNP and small indel variant caller with deep neural networks"

Ryan Poplin, Dan Newburger, Jojo Dijamco, Nam Nguyen, Dion Loy, Sam Gross, Cory Y. McLean, Mark A. DePristo

DOI: https://doi.org/10.1101/092890

Table of Contents

Table of Contents

FDA precision challenge

Objectives

- Created with the Genome in a Bottle Consortium.
- Only one dataset with high quality annotations available (NA12878).
- Quality tested on new datasets.
- Evaluation of FScore, recall and precision for SNPs and Indels.

Deep Variant results

- Best FScore for SNPs, honorable mention for precision and recall.
- First method to use Deep Learning (DL).
- Proof of concept that DL is a promising method.

Table of Contents

Haplotype-aware realignment of reads

- Reads are previously mapped (method unspecified).
- Candidates windows (size unspecified) are chosen based on mismatches and soft clips.
- Creation of De-Bruijn graphs for kmers of size 20 to 75 (increment of 5) for the reference and all overlapping reads in the window.
- Edges are weighted according to their number of occurences.

Haplotype-aware realignment of reads (cont)

- Edges with weight ower than 3 are trimmed (except for reference).
- Candidates haplotype are selected by traversing the graph, the two most likel are selected (evaluated with HMM).
- Reads are realigned with Smith-Waterson with affine gap penality.
- Position and CIGAR strings are updated in the reads.

Finding candidate variants

- Each position in the genome is evaluated.
- Collect all reads overlaping that position and aligned.
- Each possible allele is considered.
- If it is not reference, is present at least a number of time and represents a certain fraction of alleles it is emitted as a candidate.

Comparative with GATK

- No VQSR.
- No first pass of HaplotypeCaller.
- Mark duplicates used, but not described.

Prepocessing conclusion

- The realignment can be skipped (with lower results, only done for not illumina).
- The candidates are emitted with high sensitivity and low specificity on purpose.
- This whole step can be skipped (by using provided candidates).

Table of Contents

Property of the image

- An 221x100px image is created for each candidate variant.
- First 5 rows are for the reference genome.
- Each row below is used for an overlapping read.
- Each column encodes for the base pair at that position (relative to the ref) in the row of the read.
- Reads are thus 221bp long and there is at most 95 reads.
- Center column is assumed (by me) to be the position of the candidate.

Pixel encoding

- Red: encodes the base color (A: 250, G: 180, T: 100, C: 30)
- Green: encodes the quality (intensity linear in the quality).
- Blue: direction of the strand (70 if positive, 240 otherwise).
- Alpha: encodes if the read is equal to the ref and if there is an alternative allele.

Table of Contents

Overview

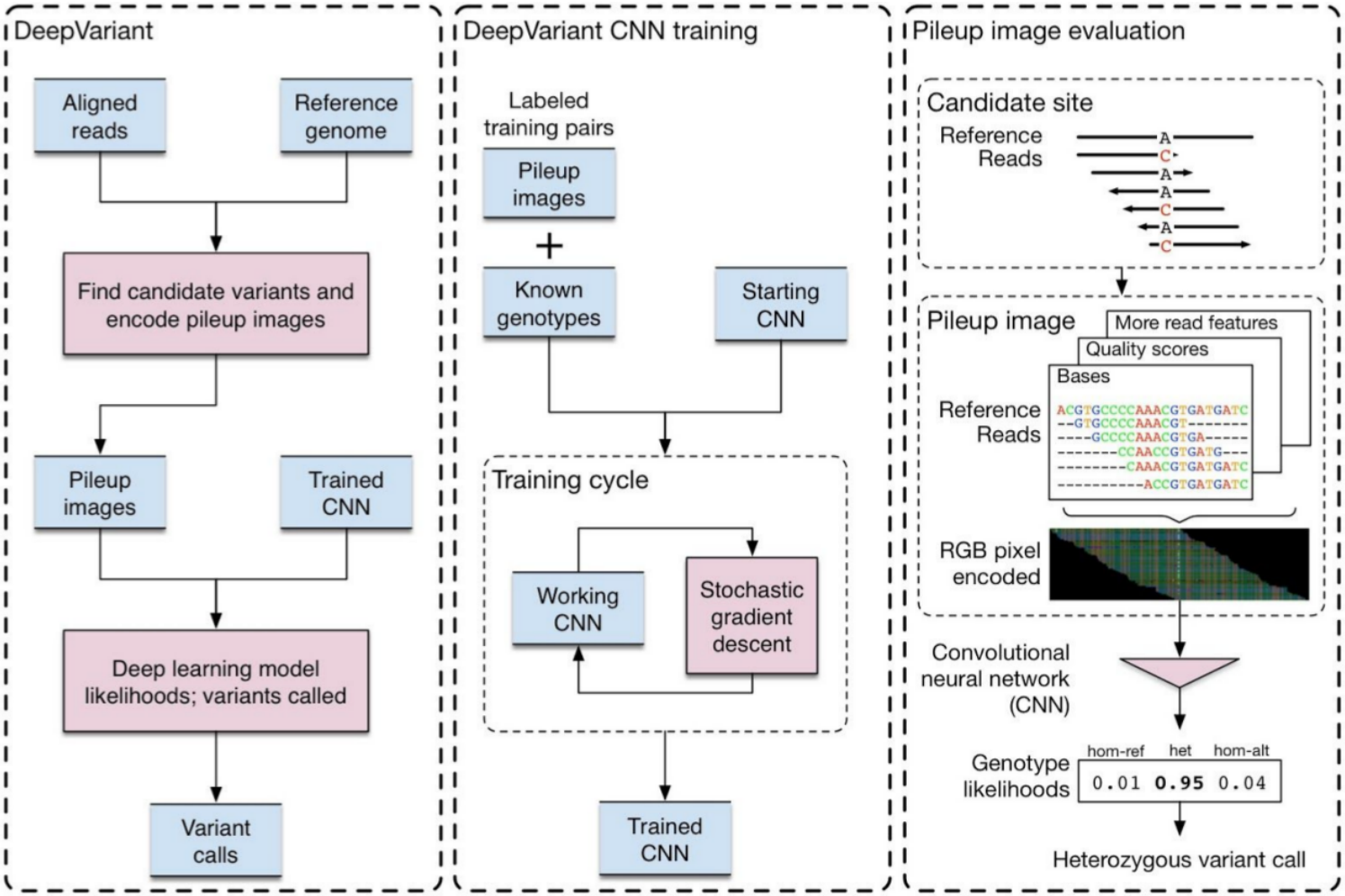

Figure: DeepVariant overview

Problem statement

- Supervised.
- Classification into "hom-ref", "het", "hom-alt".
- Ground truth comes from NA12878.
- Trained on chr1-18, hyperparam tuned on chr19, test on chr20-22.

Neural architecture

- Inception v2.
- Pretrained on ImageNet.
- Used DistBelief framework (to be replaced with TensorFlow).

Training

- 9 partitions.
- Last layer initialised with gaussian random weights.
- SGD with 32 images per batch and 8 replicates.
- Training stoped after 80 hours or 250.000 epochs or training accuracy convergence.

Evaluation against GATK

- GATK implemented following Best practices and VQSR for all chr.
- GATK implemented following Best practices and VQSR for chr1-18.
- Beats both.

Evaluation on Mice

- Used MGP data.
- Beats state of the art (F1: 98.29% vs 97.84%).
- Gets better results when trained on NA12878 than mouse genome.
- Implies transfer learning accross species.

Evaluation on other sequencers

Consistently beats other methods on

- Ion AmpliSeq exome
- Illumina TruSeq Genome
- 10x Chromium 75x WGS
- PacBio raw reads 40x WGS
- SOLID 85x

Evaluation on other sequencers (cont)

- Used realigned data from GiaB (local realignment was tuned for Illumina WGS).
- Retrained the CNN on the data from the new sequencers.
- Sugests that DeepVariant can learn on all sequencers.
- Sugests that DeepVariant is not too sensible to realignment method.

Table of Contents

Proof of concept for Deep Learning

- DeepVariant got best Fscore on FDA truth challenge.
- Only team to use Deep Learning and beat state of the Art.
- Suggests Deep Learning is an appropriate tool.
- Model had no prior on genomic data and representation was suboptimal.

Opportunities

- Transfer accross species opens the door for a unique caller for all species trained on all species.
- Retraining on different sequencing means that expert parameter tuning for each sequencer might be a thing of the past.
- Success in small variant calling migth be pushed to structural variants where state of the art is more art than science.

Population genomics

- Training a model on multiple samples teaches priors, which are the definition of population genomics.
- Model can be trained online with new incoming data, promising better performances as time goes on.

Latency

- Computation cost of the trained model are only dependent on the size of the image (here 299x299px).
- Deep Learning training methods are easily distributed (see DistBelief).
- Main bottleneck are preprocessing and image creation (assumed).
- Smaller models can be learned (see Distill).