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

- not sensible to alignment method
- needs high sensitity and low specificity
- can work without realignment
- can work on different sequencers (using provided BAM)
- no need for VQSR
- works accross species and can be trained on species with more data (human and mice)
- works accross sequencers (need ground truth but that's all)
- opens door to transfer learning

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

Table of Contents

Table of Contents