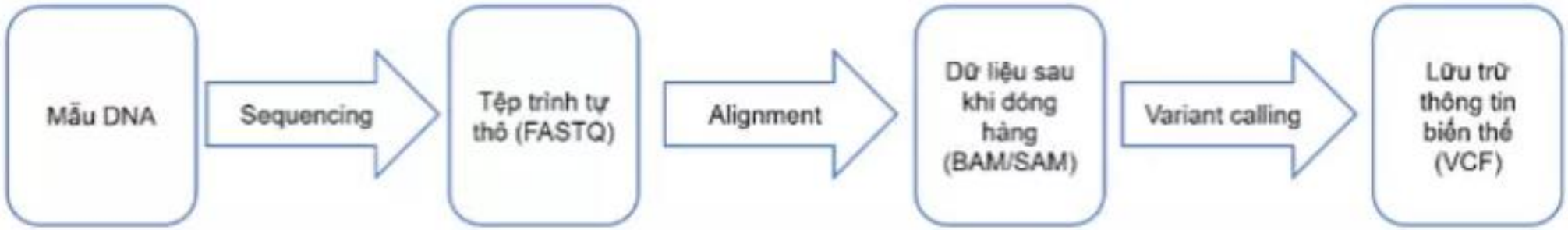
DEEPVARIANT

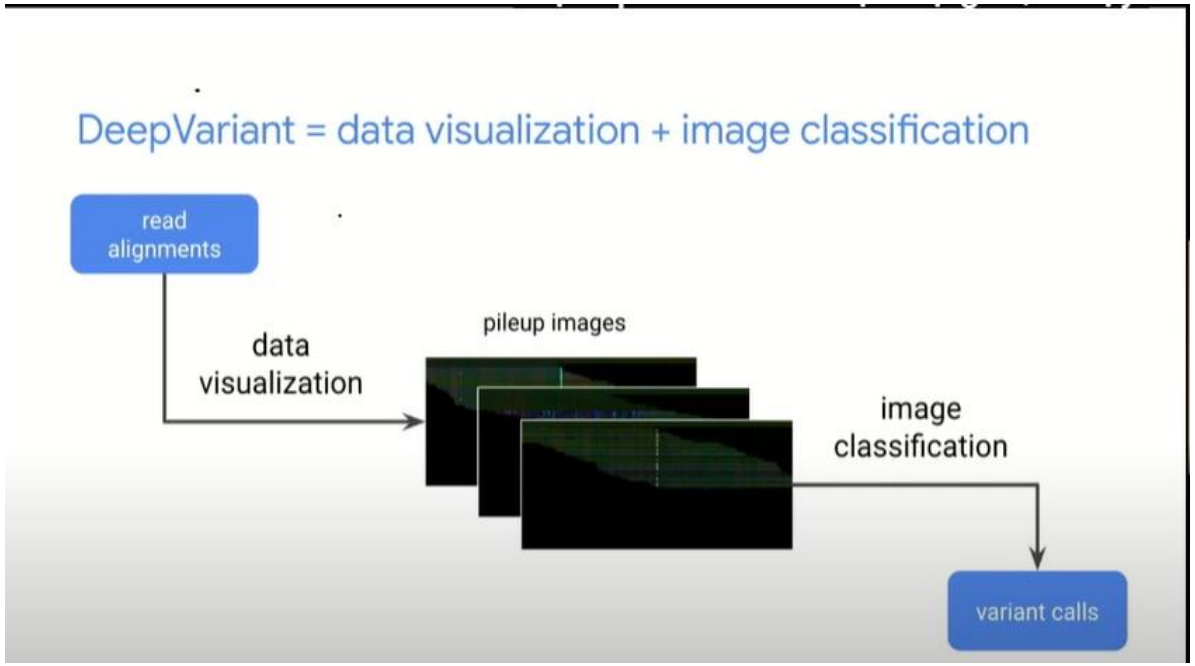
Phạm Xuân Trường

Nguyễn Duy Long

Abstract

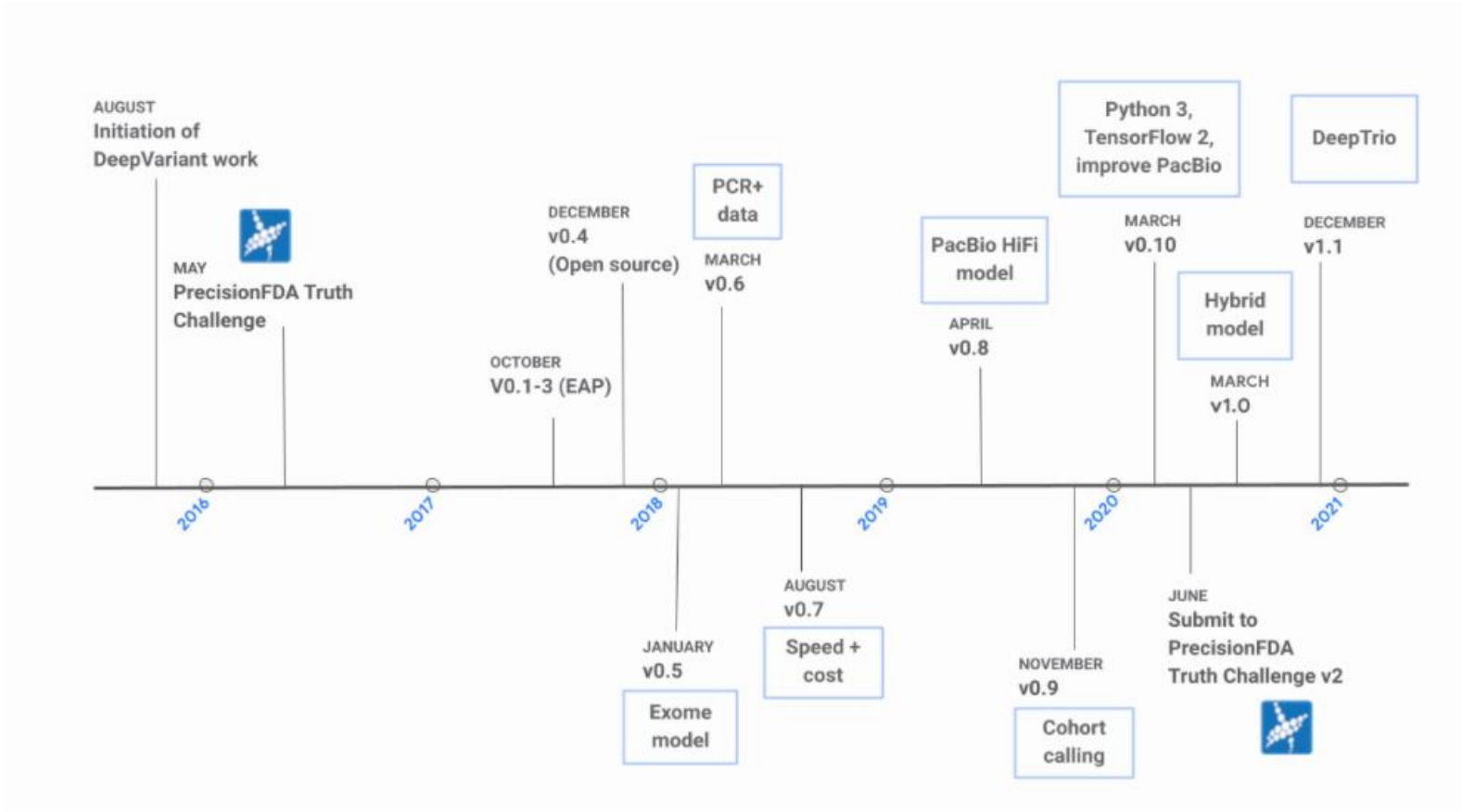

Mẫu DNA
Sequencing
Tệp trình tự thô (FASTQ)
Alignment
Dữ liệu sau khi dóng hàng (BAM/SAM)
Variant calling
Lưu trữ thông tin biến thể (VCF)

Tổng quan

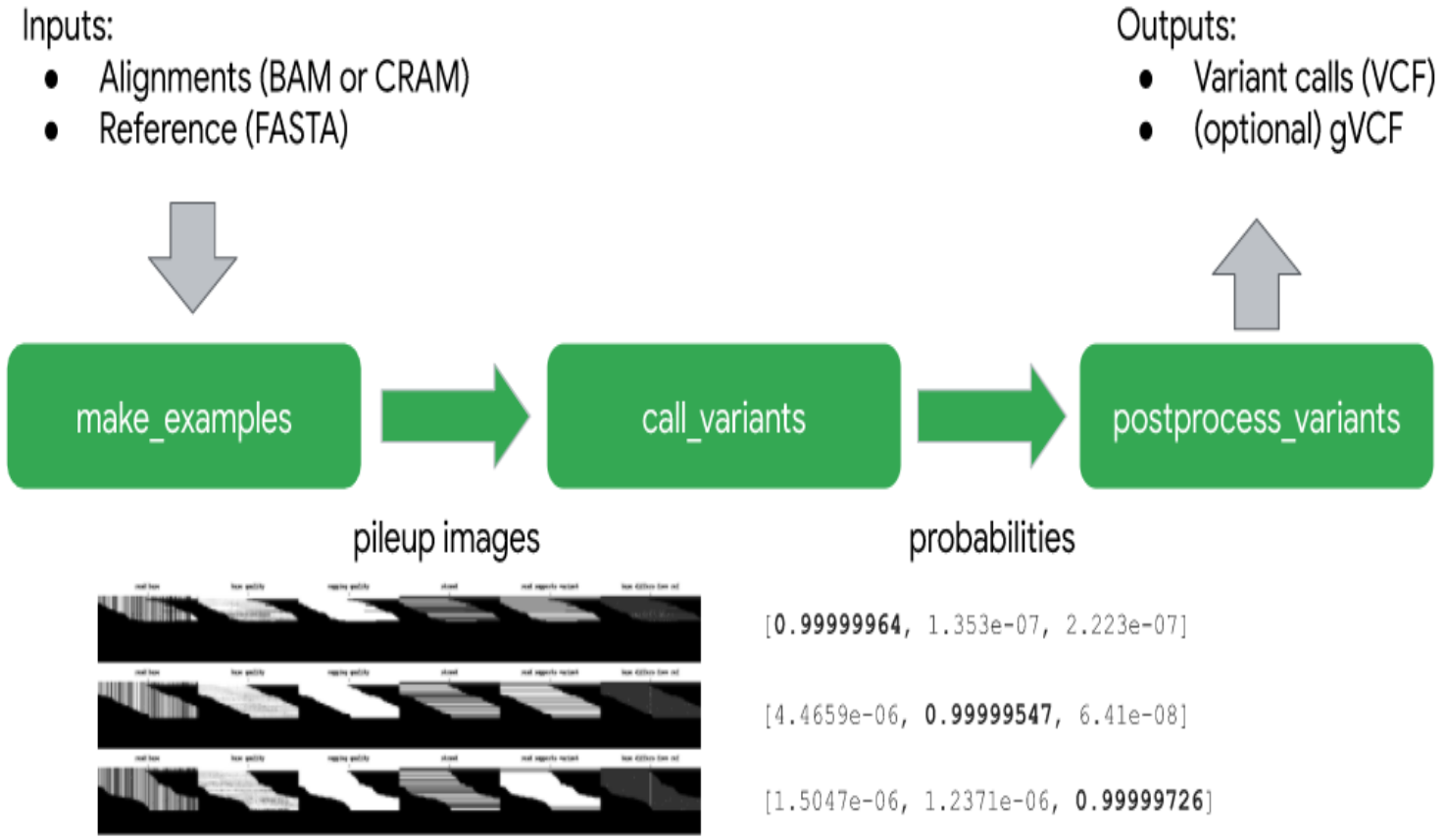


- Mục đích: Nhận dạng và phân loại biến thể gen đầu vào dựa trên bộ gen tham chiếu
- Mô hình sử dụng:CNN

Timelines

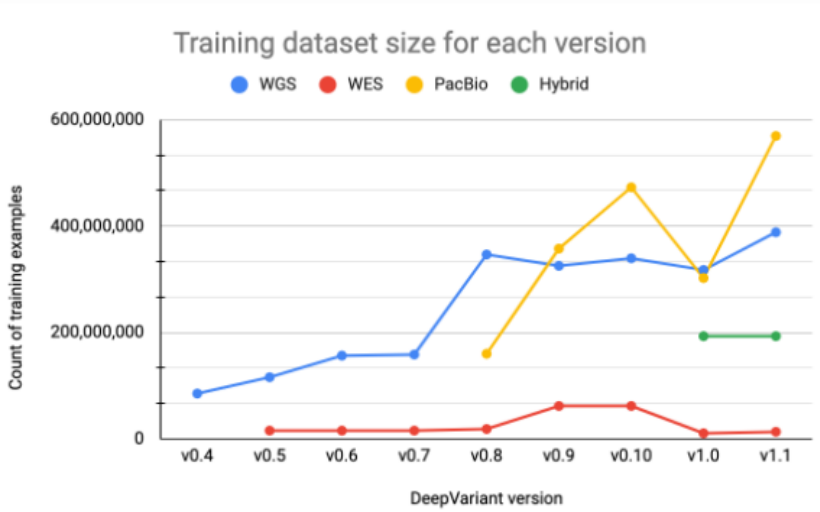

AUGUST
Initiation of DeepVariant work
MAY
PrecisionFDA Truth Challenge
OCTOBER
V0.1-3 (EAP)
DECEMBER
v0.4
(Open source)
MARCH
v0.6
PCR+ data
JANUARY
v0.5
Exome model
AUGUST
v0.7
Speed + cost
APRIL
v0.8
PacBio HiFi model
NOVEMBER
v0.9
Cohort calling
MARCH
v0.10
Python 3, TensorFlow 2, improve PacBio
MARCH
v1.0
Hybrid model
JUNE
Submit to PrecisionFDA Truth Challenge v2
DECEMBER
v1.1
DeepTrio
2016
2017
2018
2019
2020
2021

Workflow


Inputs:
Alignments (BAM or CRAM)
Reference (FASTA)
Outputs:
Variant calls (VCF)
(optional) gVCF
make_examples
call_variants
postprocess_variants
pileup images
probabilities
[0.99999964, 1.353e-07, 2.223e-07]
[4.4659e-06, 0.99999547, 6.41e-08]
[1.5047e-06, 1.2371e-06, 0.99999726]

Input

Training Data



Input

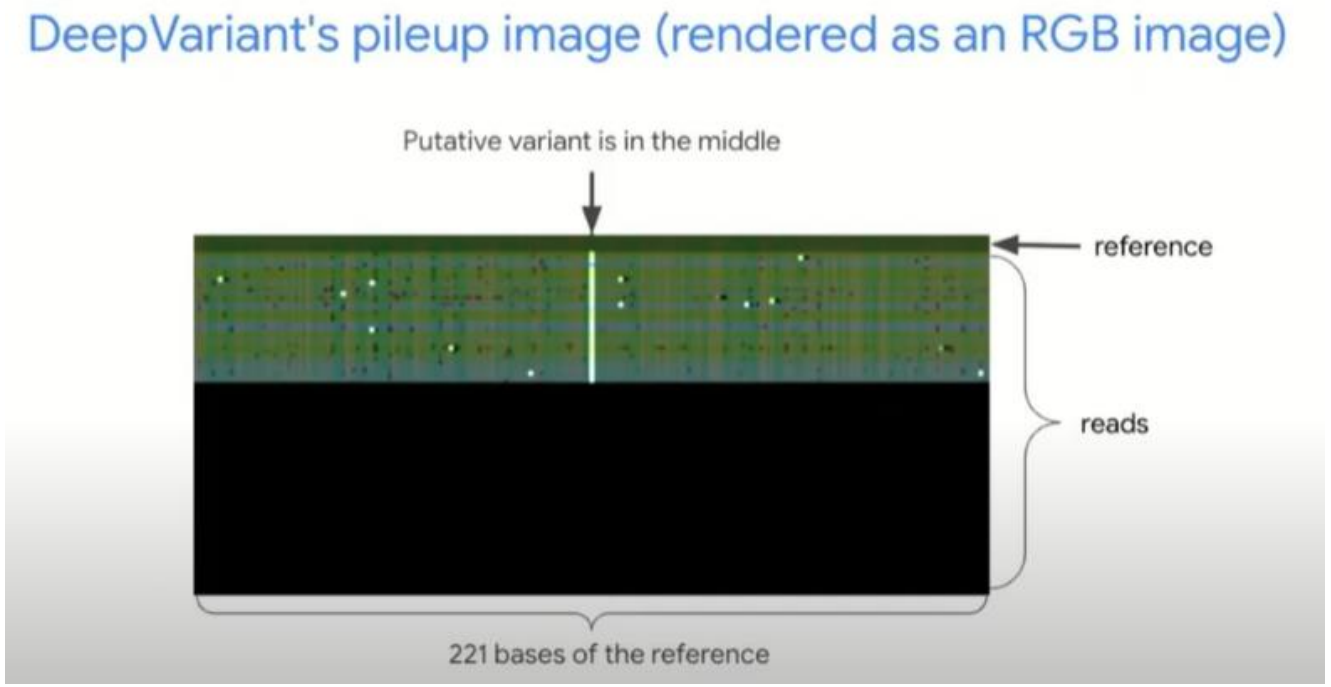
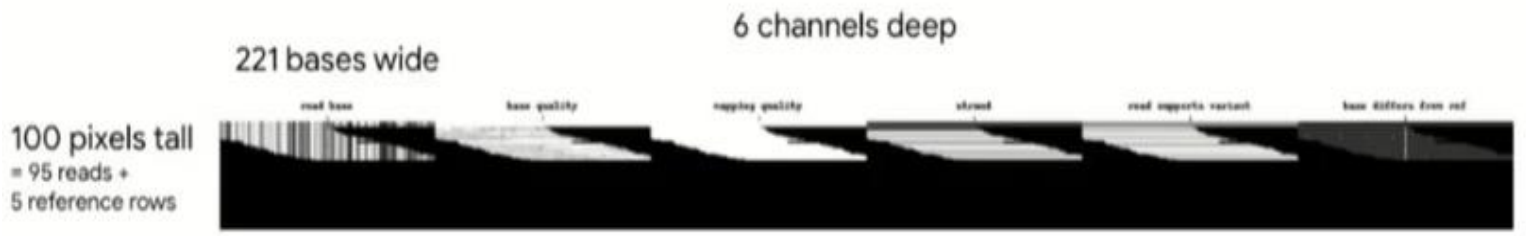


Image: 100*221

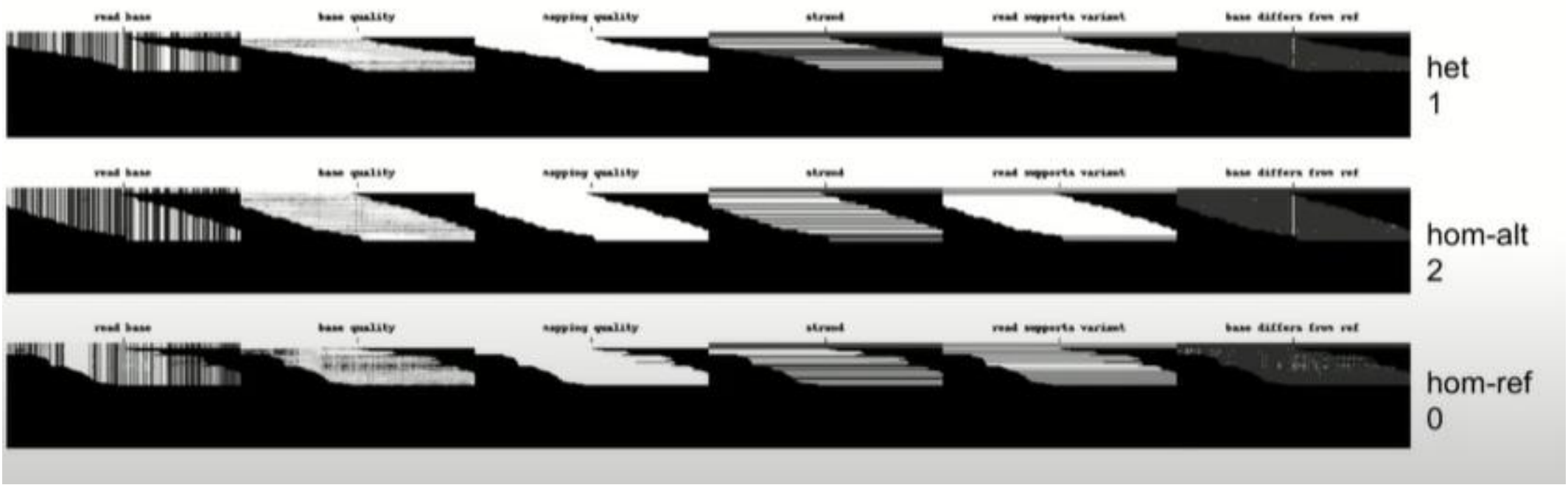
Input

DeepVariant's pileup image - more than RGB channels



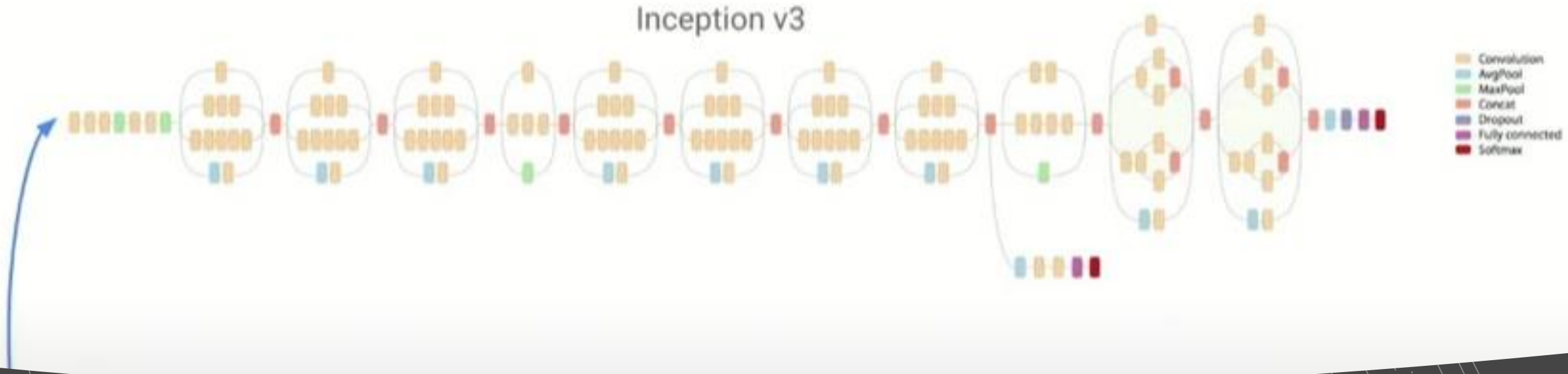
Tensor of shape
(Height, Width, Channels)
=(100, 221, 6)

1. Read base: different intensities represent A, C, G, and T bases in the read.
2. Base quality: set by the sequencing machine. White is higher quality.
3. Mapping quality: set by the aligner. White is higher quality.
4. Strand of alignment: Black is forward; white is reverse.
5. Read supports variant: White means the read supports the given alternate allele, grey means it does not.
6. Base differs from ref: White means the base is different from the reference, dark grey means the base matches the reference.



Example

Passing the pileup images through the convolutional neural network (CNN)



Model

VCF FILES

Example [edit]

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS      ID         REF   ALT    QUAL FILTER   INFO                              FORMAT      NA00001         NA00002         NA00003
20     14370    rs6054257  G     A      29   PASS     NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
20     17330    .          T     A      3    q10      NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50  0|1:3:5:65,3    0/0:41:3
20     1110696  rs6040355  A     G,T    67   PASS     NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27  2|1:2:0:18,2    2/2:35:4
20     1230237  .          T     .      47   PASS     NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60  0|0:48:4:51,51  0/0:61:2
20     1234567  microsat1  GTC   G,GTCT 50   PASS     NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4        0/2:17:2        1/1:40:3
```

Common FORMAT fields [edit]

Name	Brief description
AD	Read depth for each allele
ADF	Read depth for each allele on the forward strand
ADR	Read depth for each allele on the reverse strand
DP	Read depth
EC	Expected alternate allele counts
FT	Filter indicating if this genotype was “called”
GL	Genotype likelihoods
GP	Genotype posterior probabilities
GQ	Conditional genotype quality
GT	Genotype
HQ	Haplotype quality
MQ	RMS mapping quality
PL	Phred-scaled genotype likelihoods rounded to the closest integer
PQ	Phasing quality
PS	Phase set

Any other format fields are defined in the .vcf header.

Output

Quick start

- Environment:Unix-like
- Tool:Docker
- Reference genome:Hg19-NST 20
- Genome input:`NA12878_S1-NST20`