



# Genomics in the Cloud

Book Club - Week 7

January 11, 2021

# Agenda

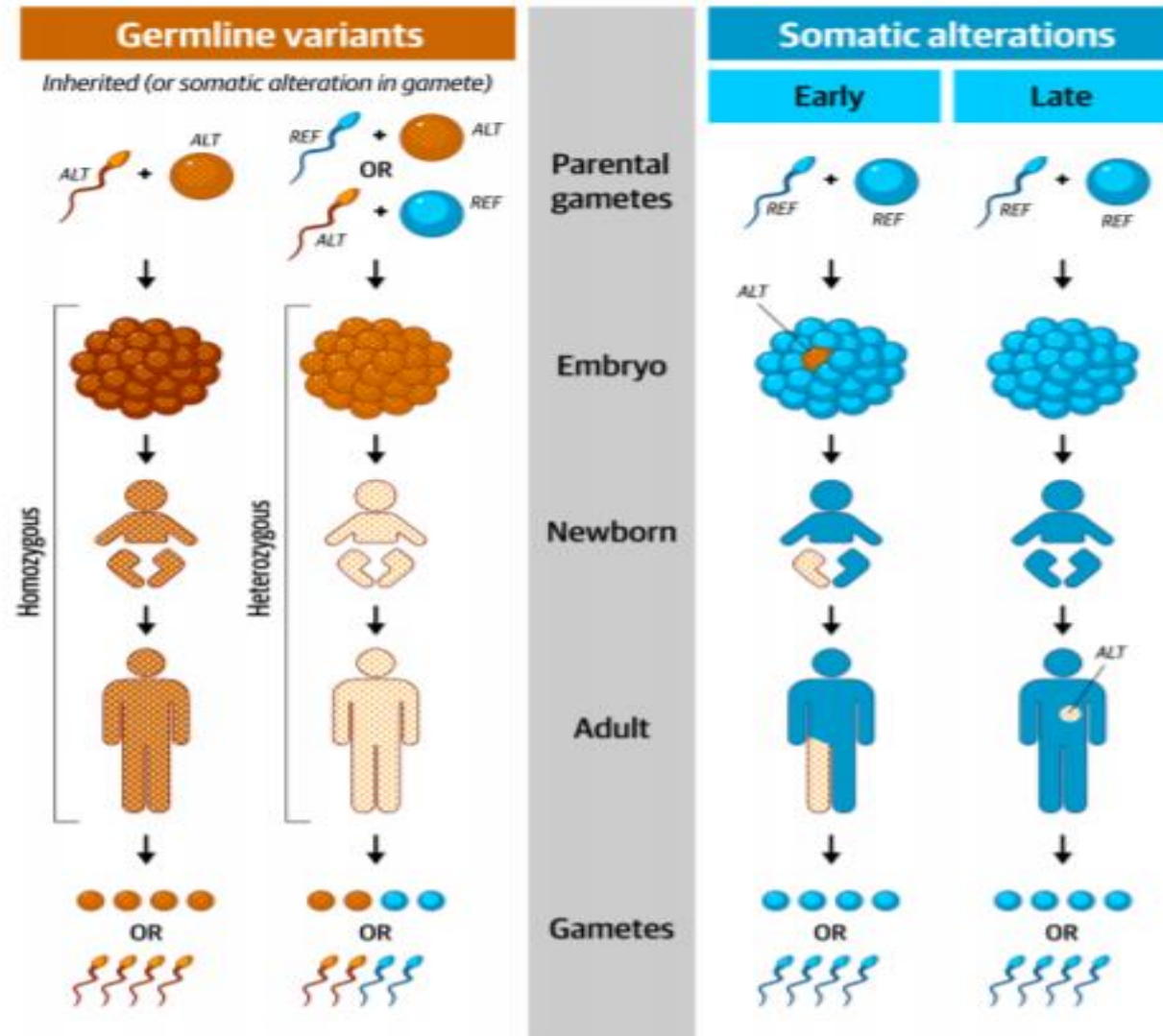
- Chapter 6: GATK Best Practices for Germline Short Variant Discovery
- Additional resources
- Open discussion



# Chapter 6: GATK Best Practices for Germline Short Variant Discovery

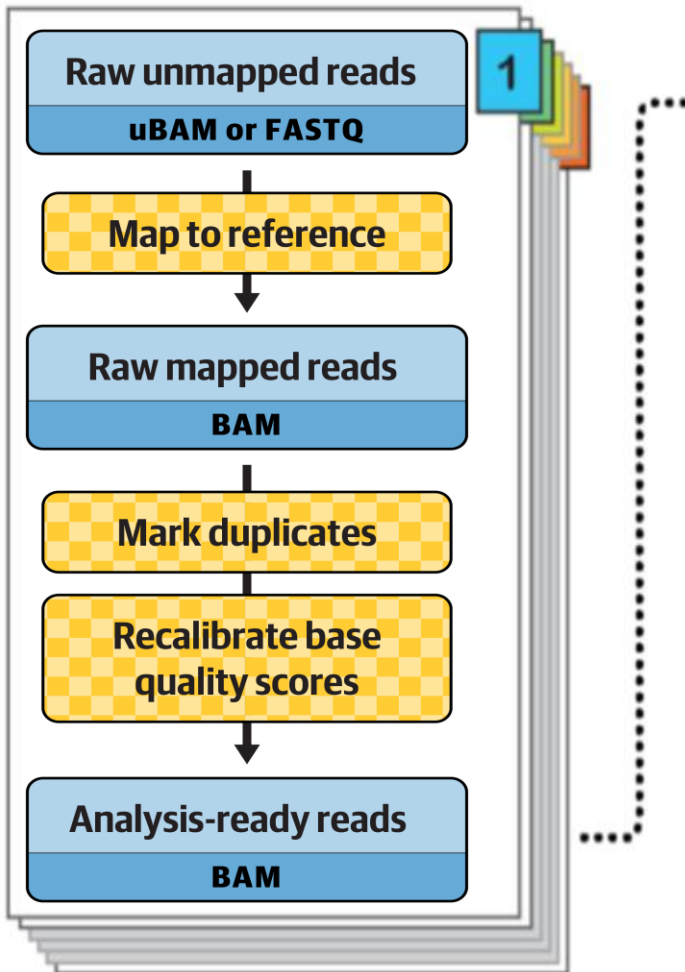
*Genomics in the Cloud* by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.

# Germline vs Somatic Variants

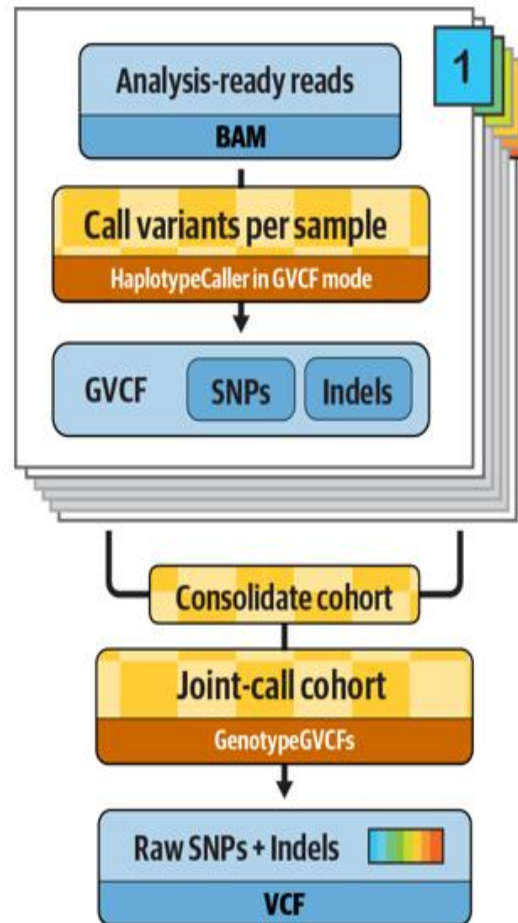


# Germline Best Practices Overview

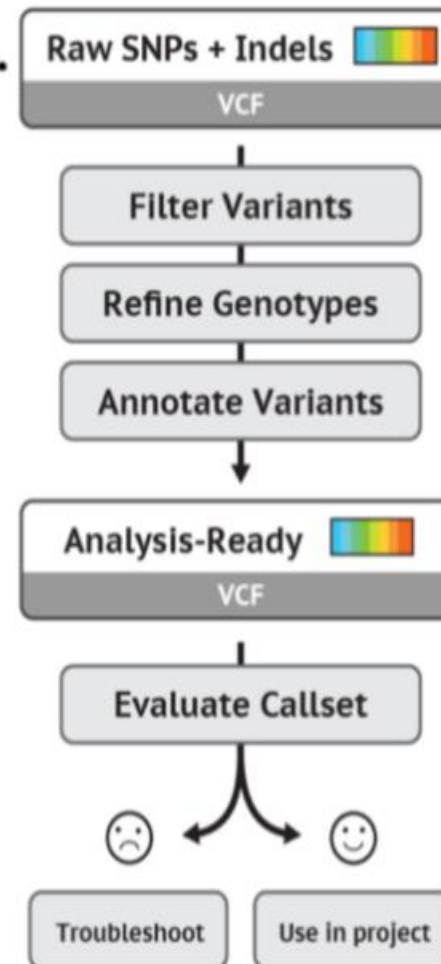
## Data Preprocessing



## Joint Variant Discovery

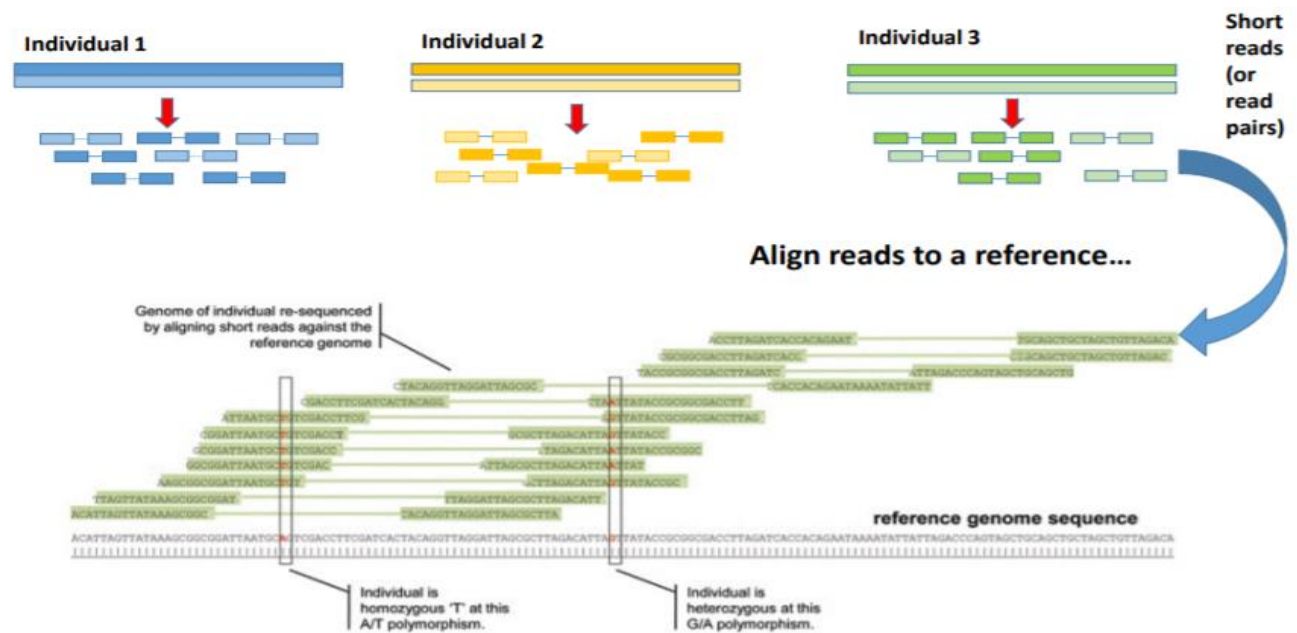
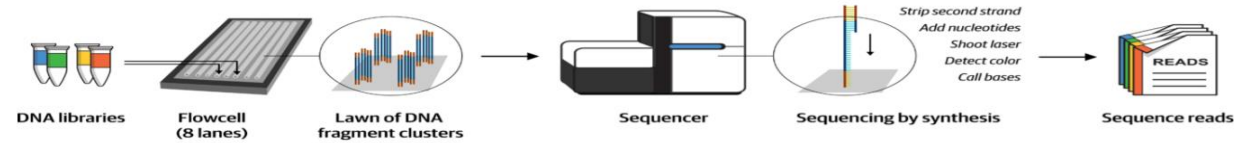
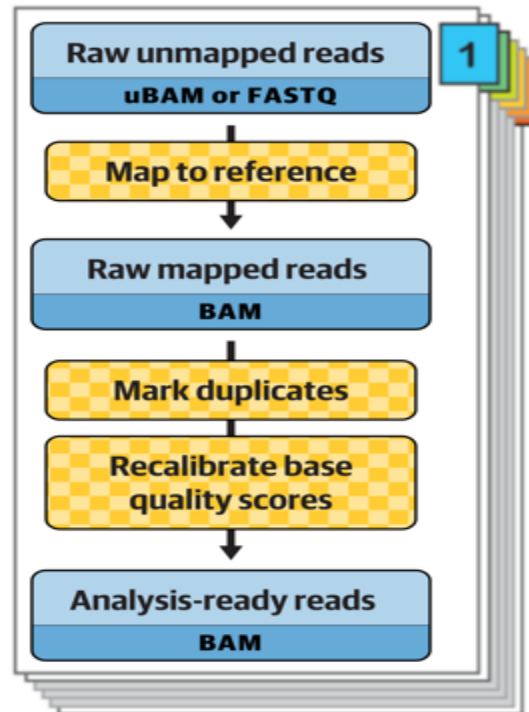


## Variant Filtration



# Step 1 – Map to Reference

## Data Preprocessing



```
bwa mem -M -t 7 -p reference.fasta unmapped_reads.fq > mapped_reads.sam
```

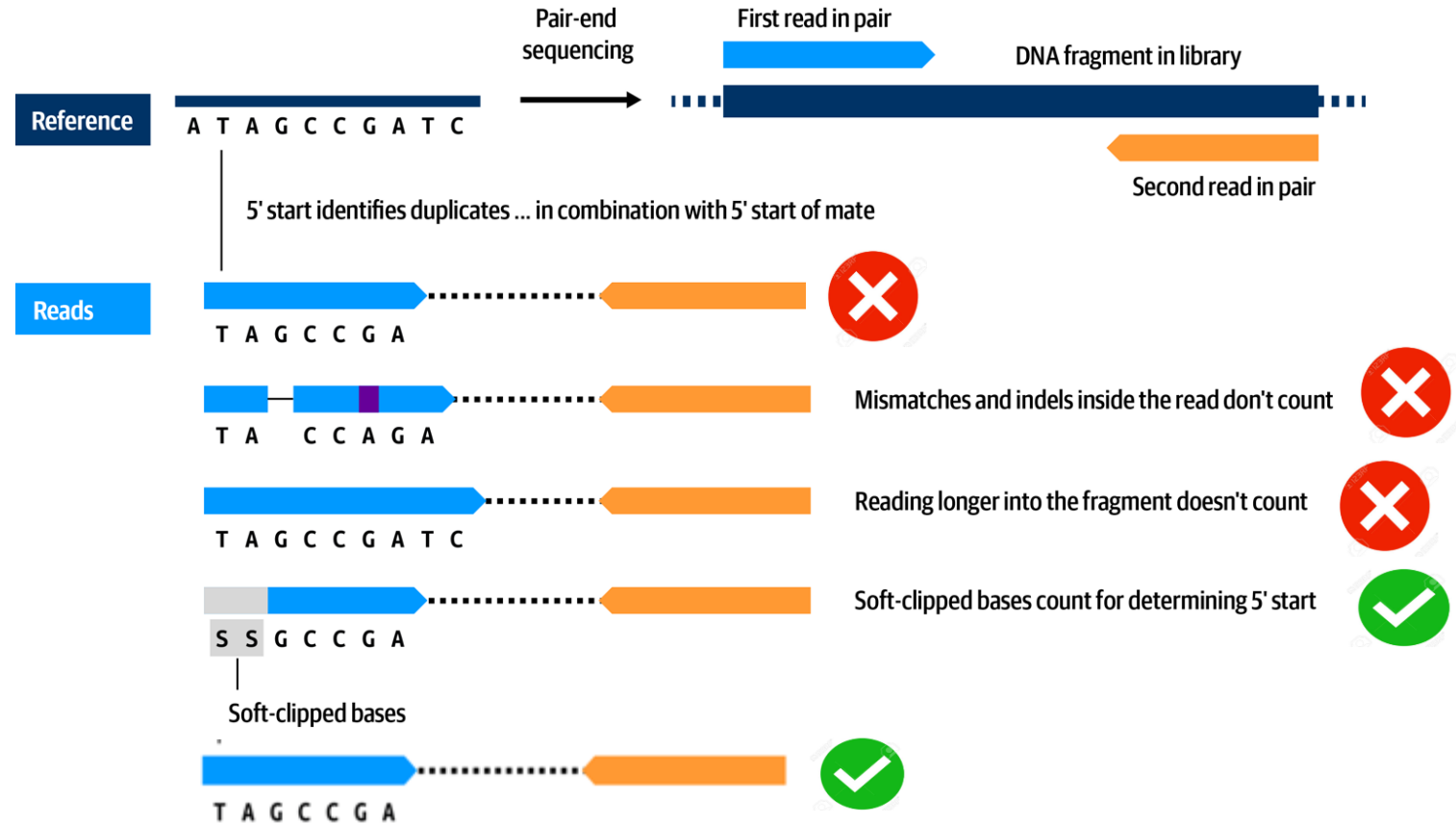
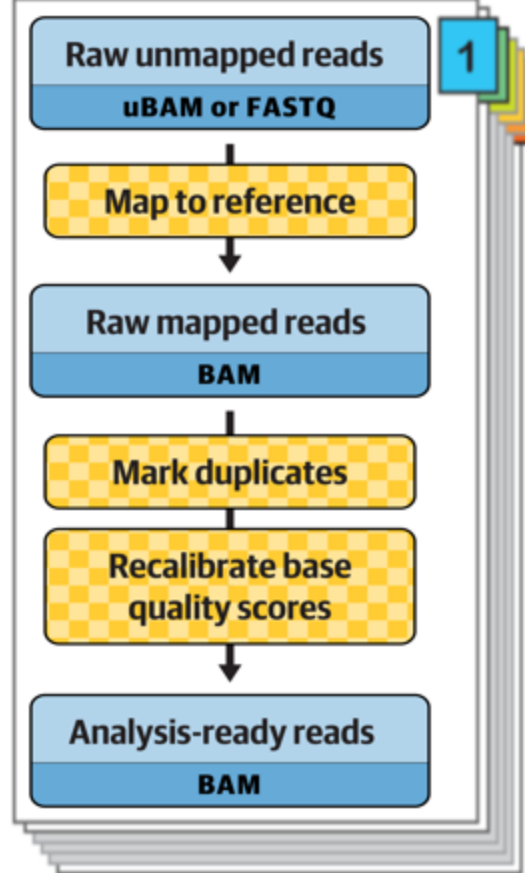
```
samtools view -S -b mapped_reads.sam > mapped_reads.bam
```





# Step 2 – Marking Duplicates

## Data Preprocessing



```
gatk MarkDuplicates -R reference.fasta -I mapped_reads.bam \
-O sample_markdups.bam
```



# Visualizing duplicate reads in IGV

Showing duplicate reads



Hiding duplicate reads



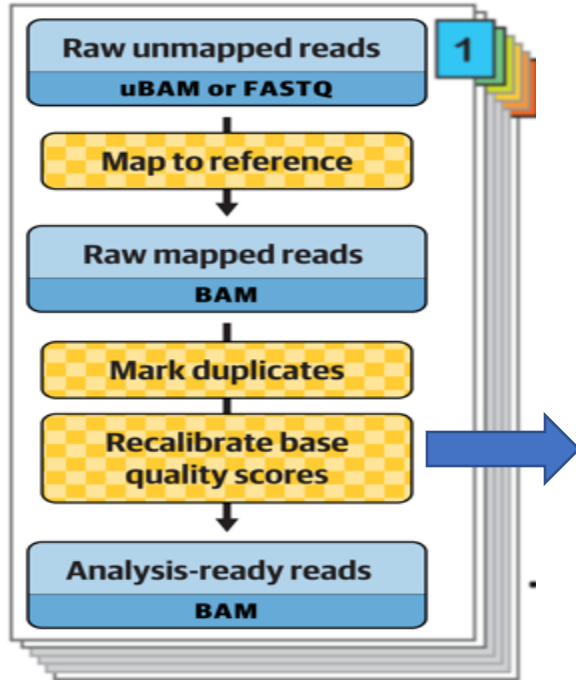
*Reads marked as duplicates are still in the file but are ignored by default by many tools*



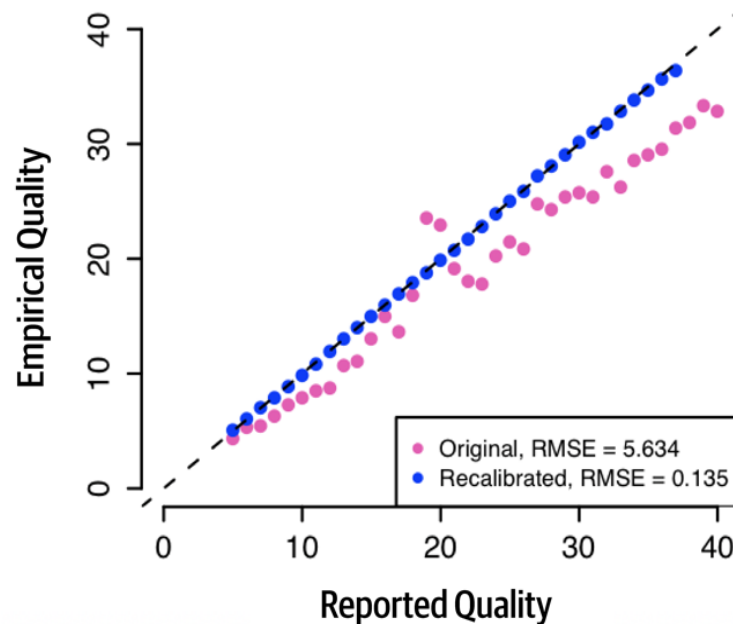


# Step 3 – Base Quality Score Recalibration

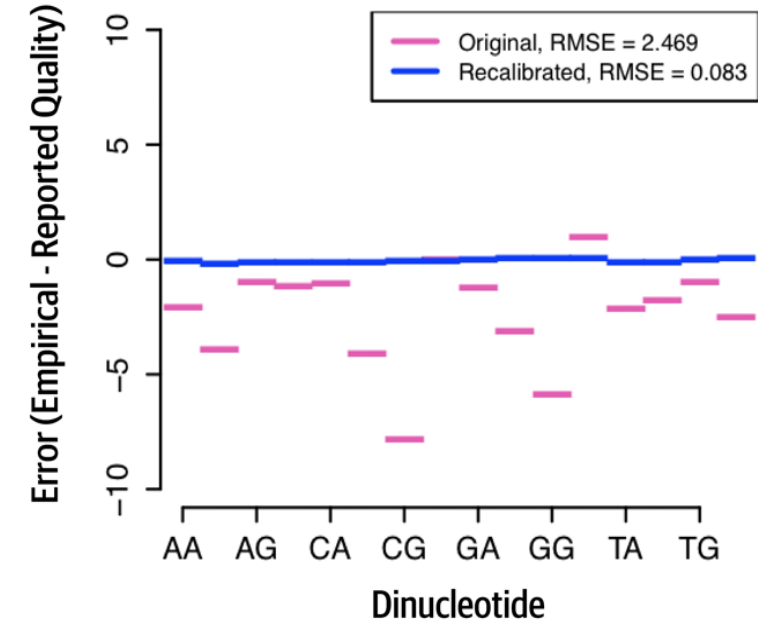
## Data Preprocessing



Overall effect of bias



Bias on nucleotide context

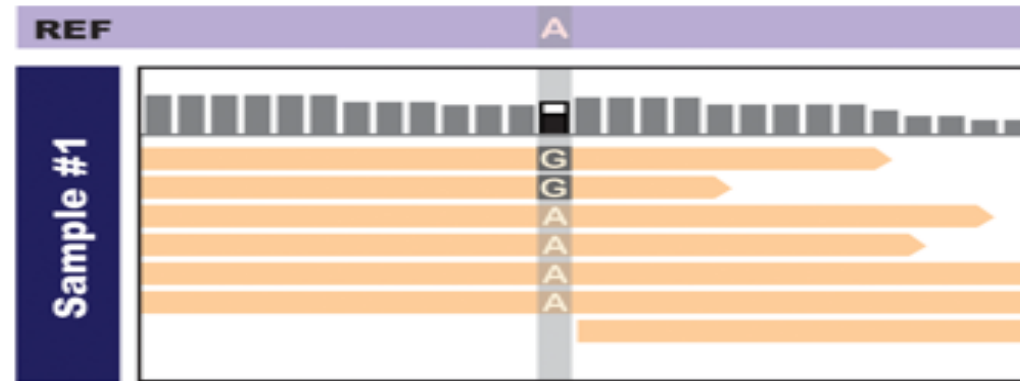
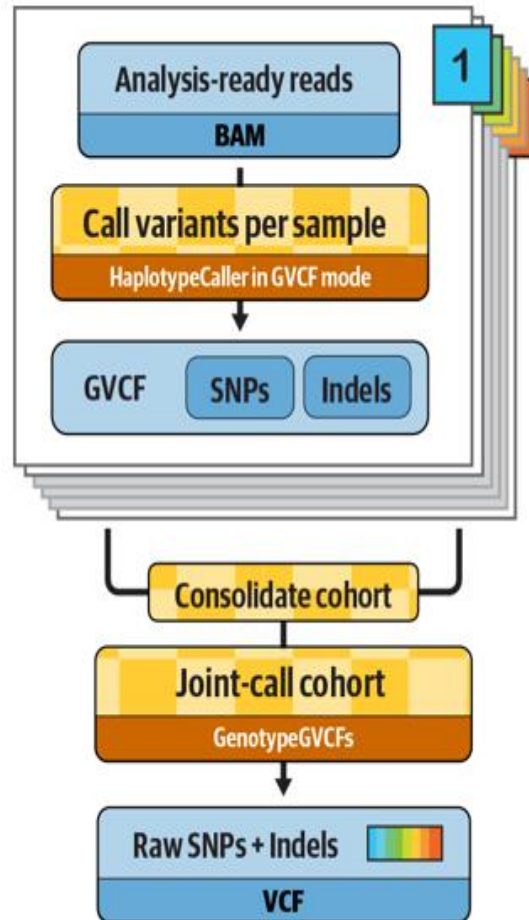


```
gatk BaseRecalibrator -R reference.fasta -I sample_markdups.bam \
    -known-sites known_variation.vcf -O recal_data.tables
```

```
gatk ApplyBQSR -R reference.fasta -I sample_markdups.bam \
    --bqsr-recal-file recal_data.tables -O sample_markdups_recal.bam
```



# Joint Calling – Why call jointly?

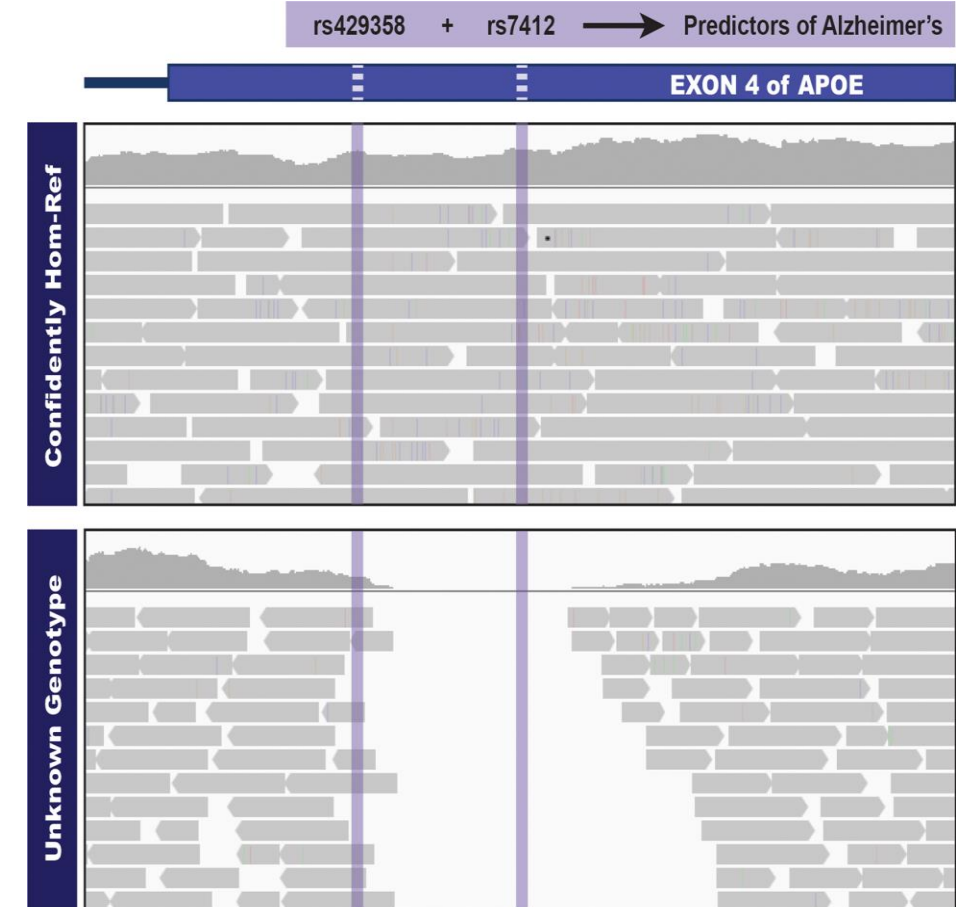
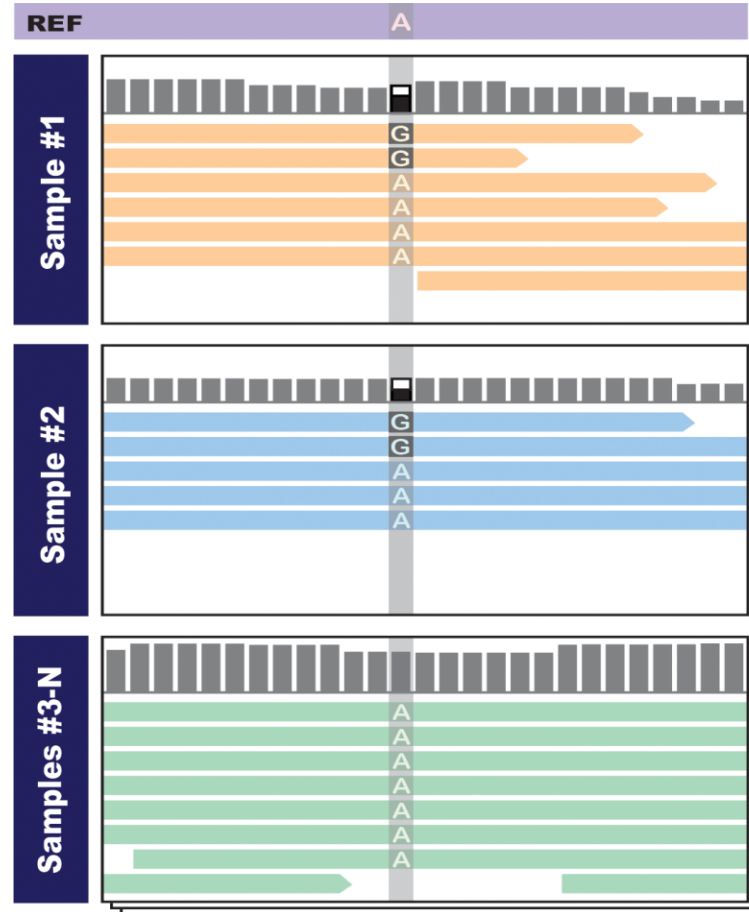
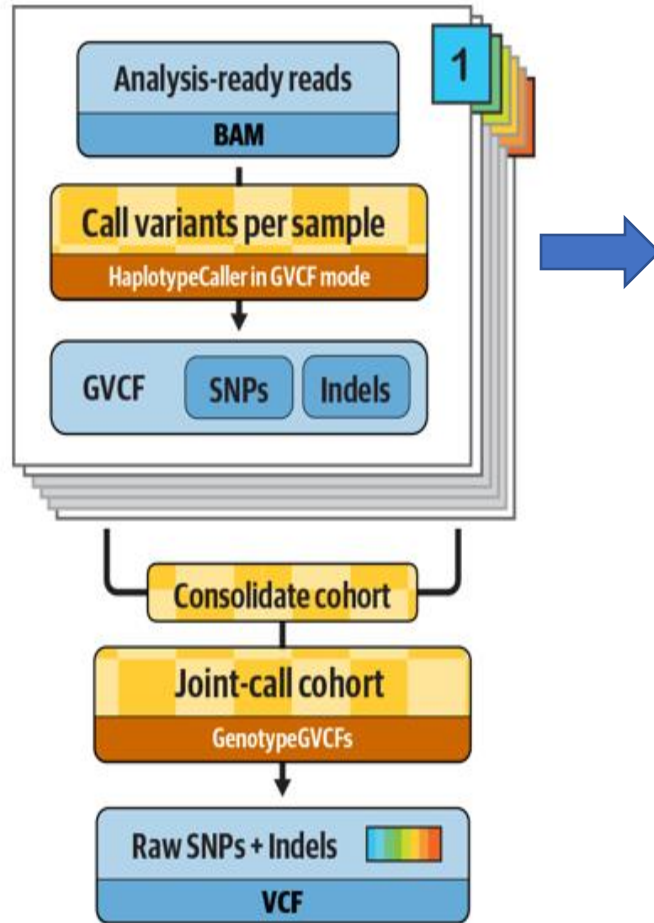


Low confidence variant calls can get filtered out in downstream variant filtering processes

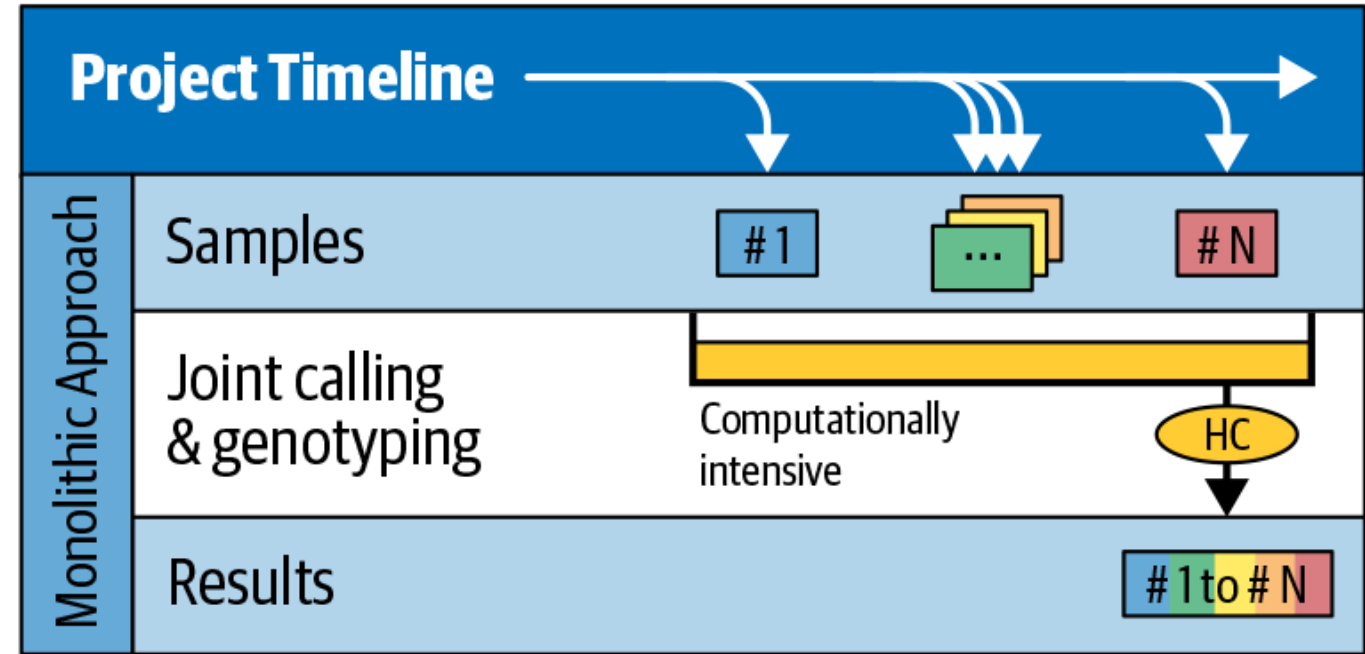
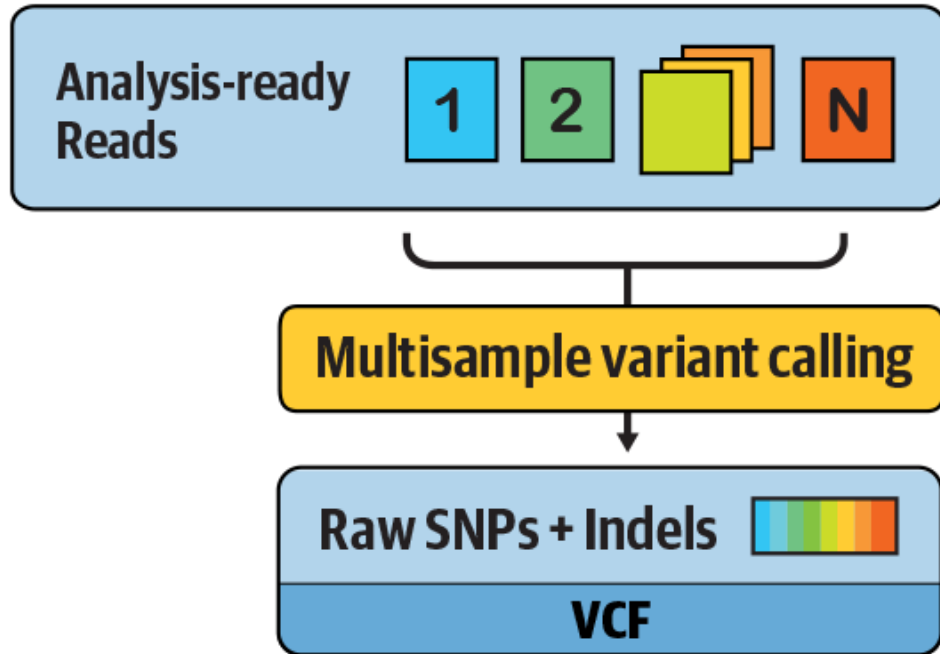


# Joint Variant Discovery – The “why?” part!

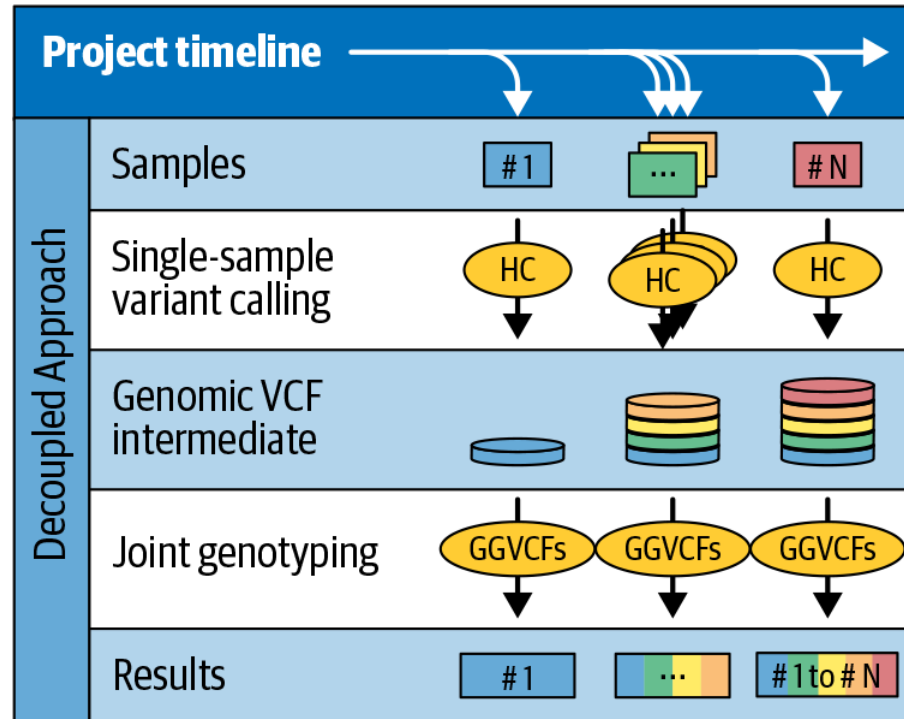
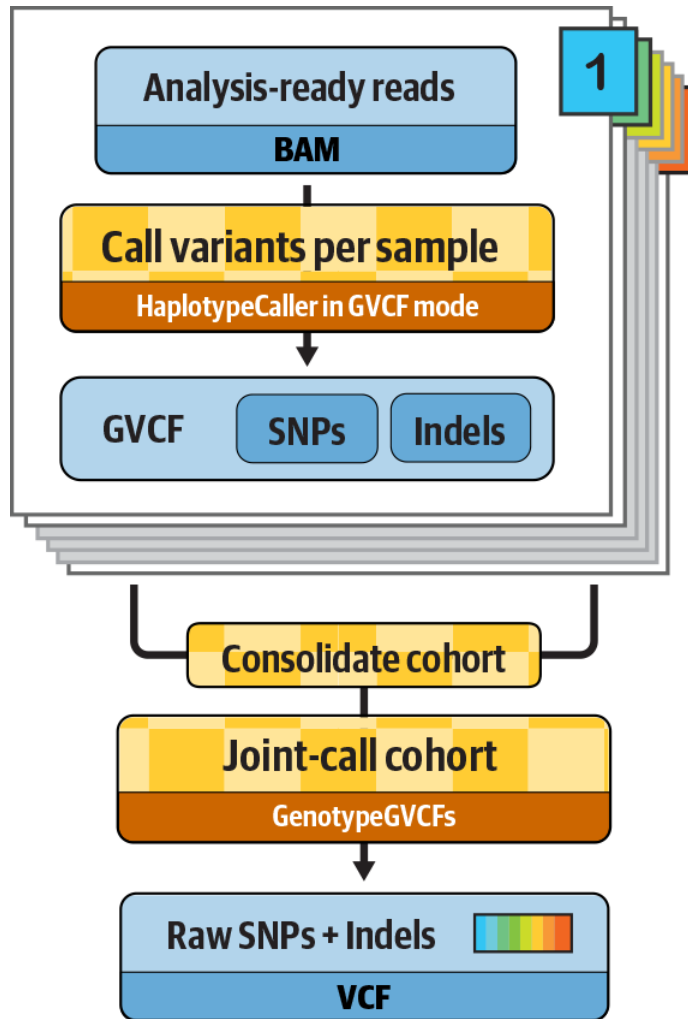
## Joint Variant Discovery



# Prior to GATK3 (The N+1 problem)



# Using GVCF to improve scalability

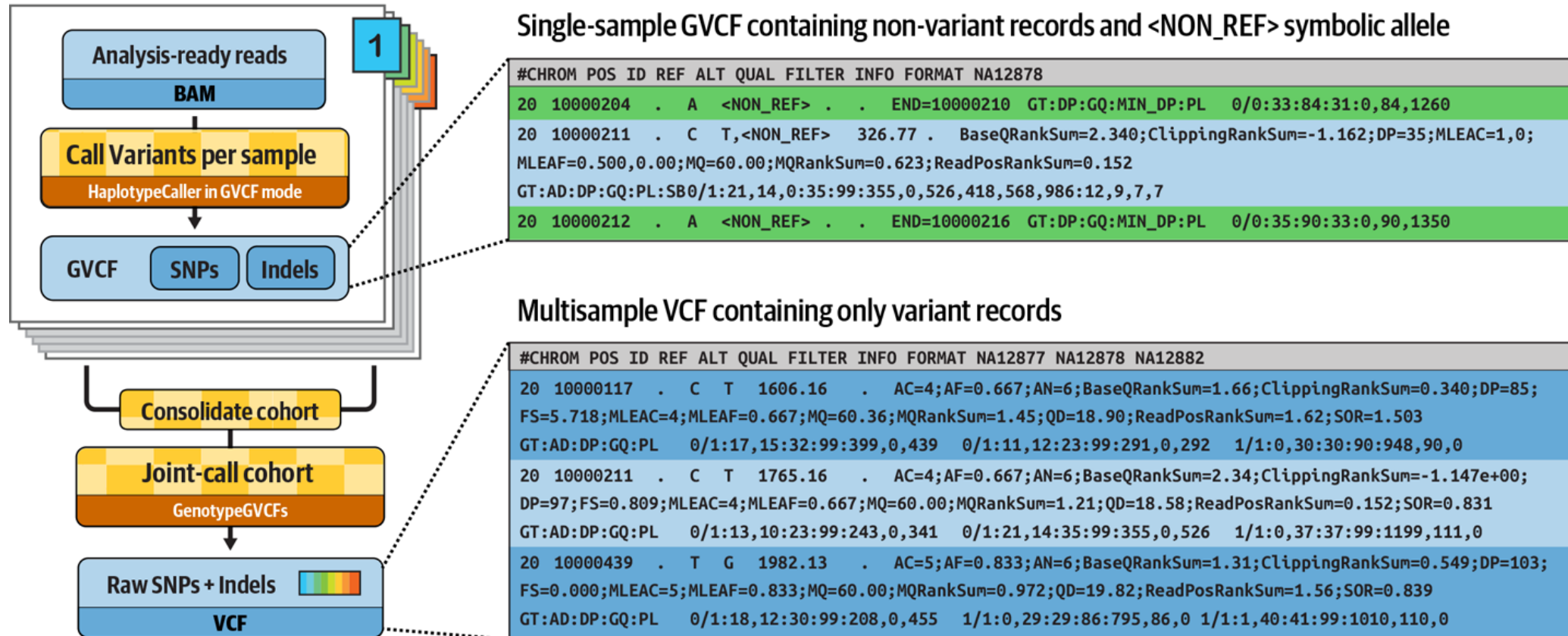


Scales linearly with number of samples.

Want to add a new sample?  
Make a GVCF for that sample then re-call the cohort.



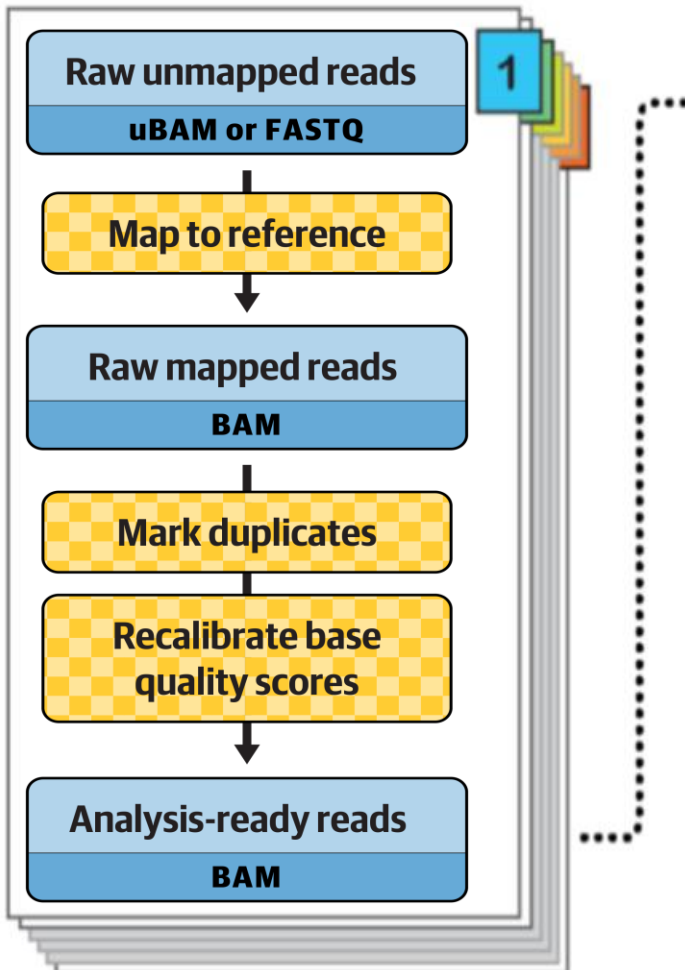
# Progressing to the final cohort VCF



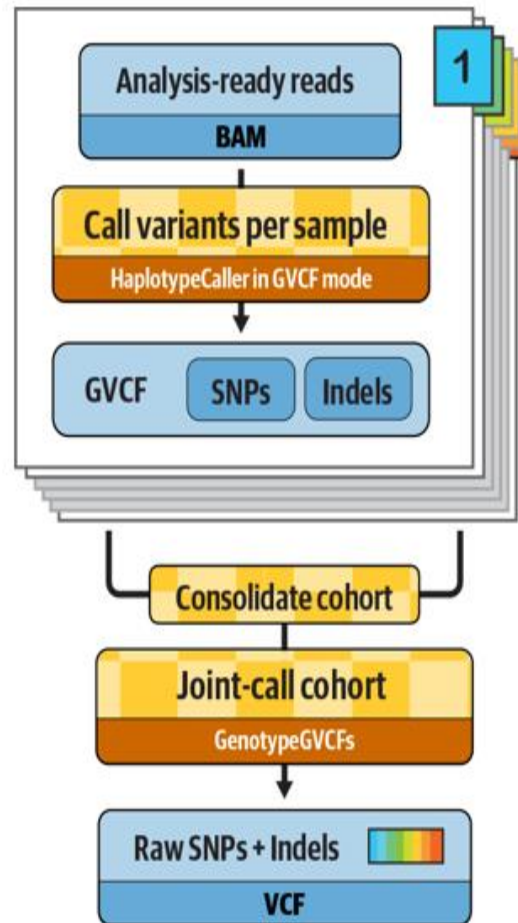


# Germline Best Practices Overview

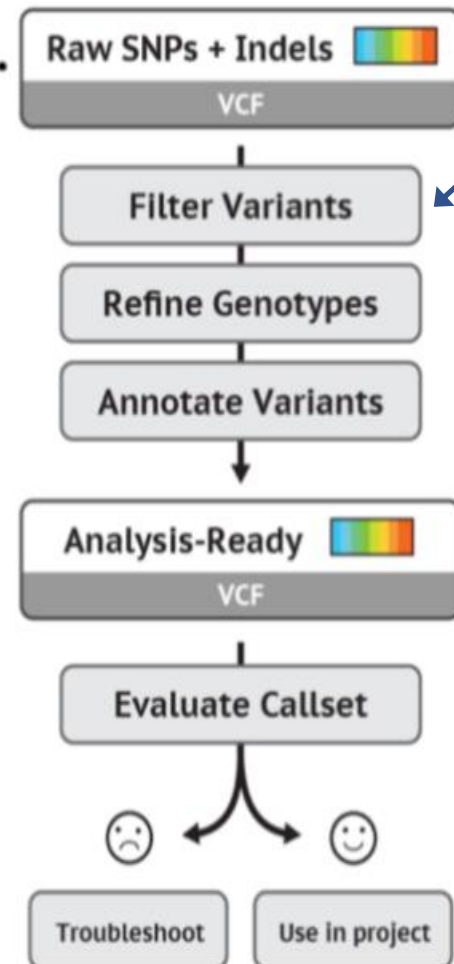
## Data Preprocessing



## Joint Variant Discovery



## Variant Filtration



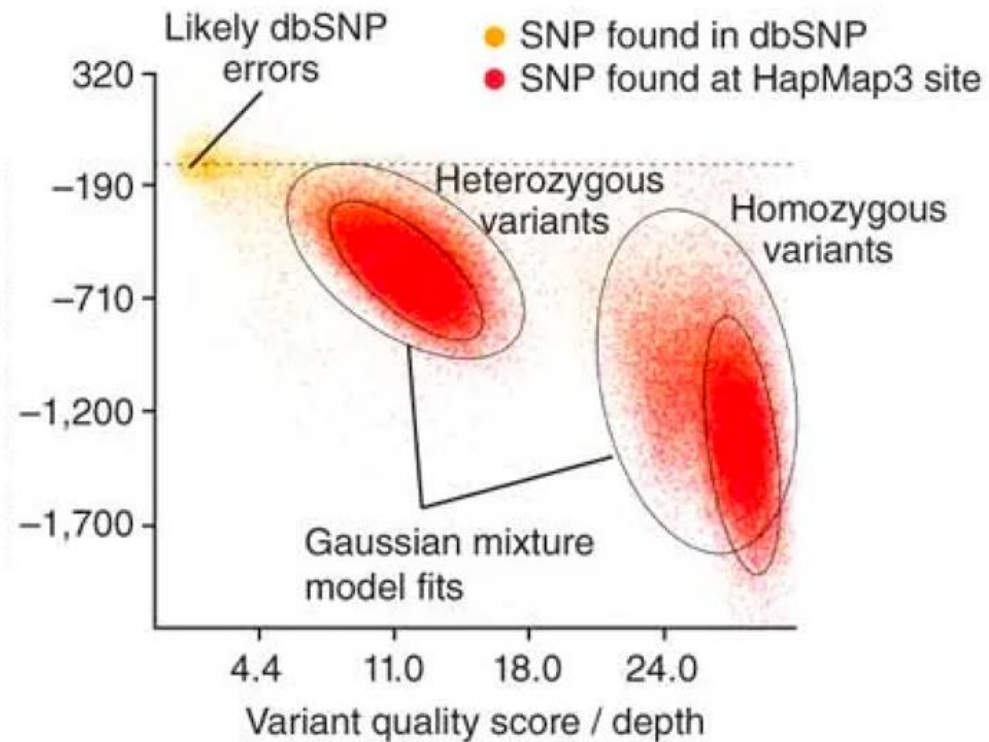
You are here



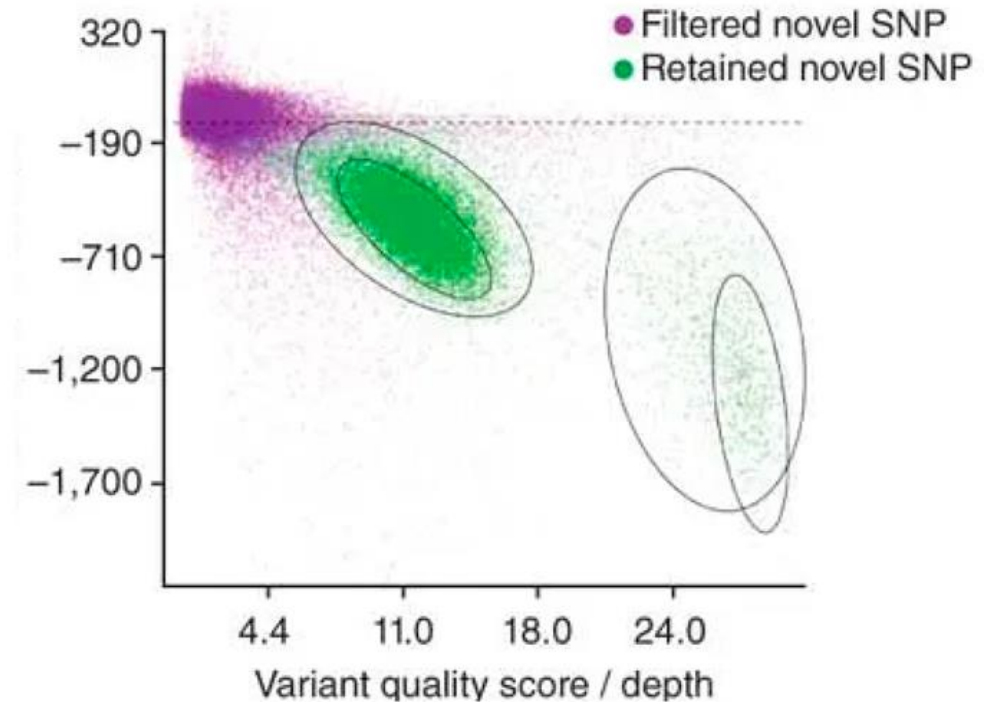


# Filtering with cluster analysis

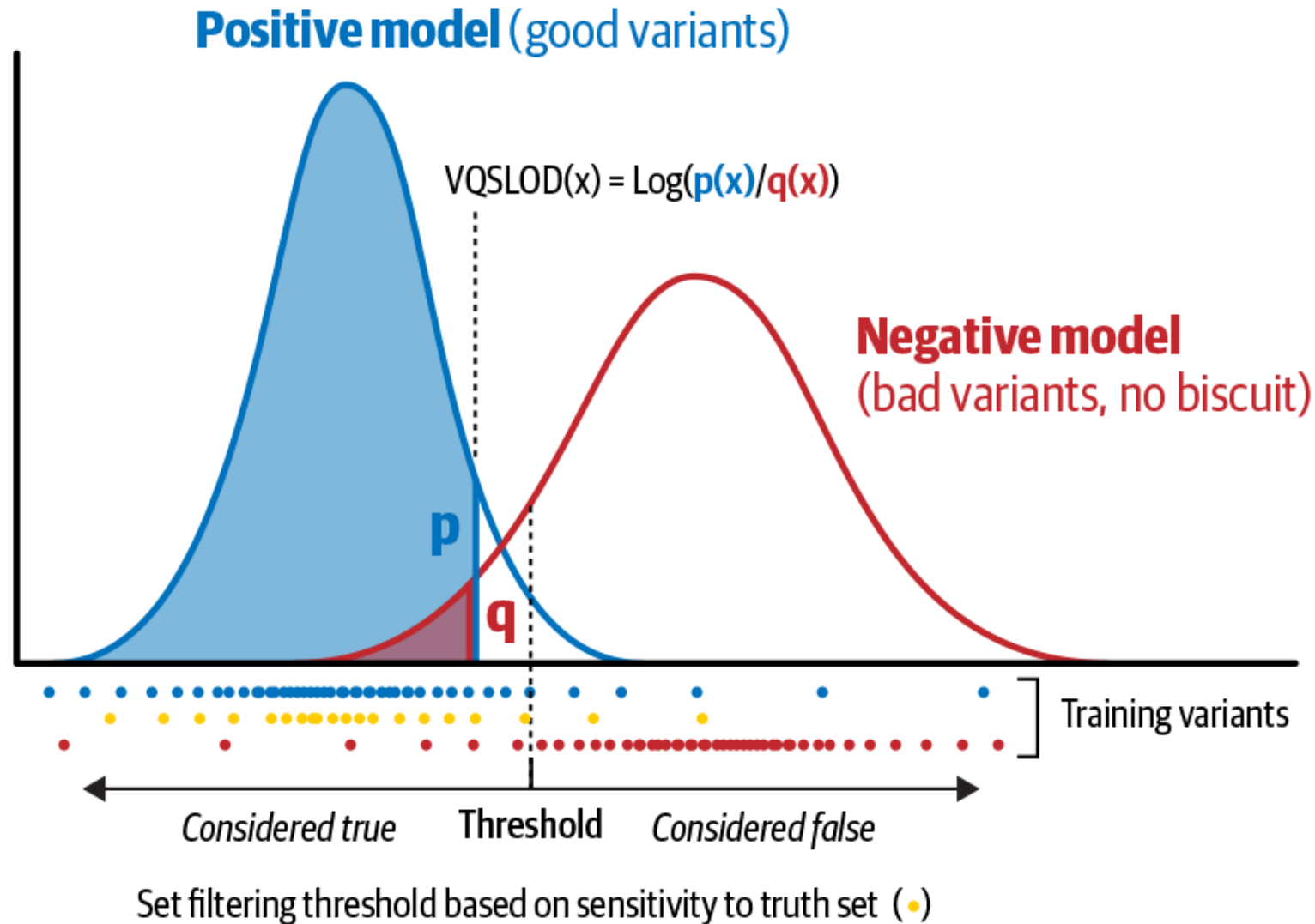
**Model trained on HapMap**



**Model applied to novel SNPs**



# Calculating the VQSLOD score (Variant Quality Score Log Odds Ratio)



# Calculating VQSLOD

```
# gatk VariantRecalibrator \  
  -R ref/ref.fasta \  
  -V jointcalls_hc.vcf.gz \  
  --resource:hapmap,known=false,training=true,truth=true,prior=15.0 \  
  hapmap_sites.vcf.gz \  
  --resource:omni,known=false,training=true,truth=false,prior=12.0 \  
  1000G_omni2.5.sites.vcf.gz \  
  --resource:1000G,known=false,training=true,truth=false,prior=10.0 \  
  1000G_phase1.snps.high_conf.vcf.gz \  
  --resource:dbsnp,known=true,training=false,truth=false,prior=2.0 \  
  dbsnp.vcf.gz \  
  -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR \  
  -mode SNP \  
  -O output.recal \  
  --tranches-file output.tranches
```

# Apply VQSR (Variant Quality Score Recalibrator)

```
# gatk ApplyVQSR \  
  -R ref/ref.fasta \  
  -V jointcalls_hc.vcf.gz \  
  -O jointcall_filtered.vcf.gz \  
  --truth-sensitivity-filter-level 99.9 \  
  --tranches-file output.tranches \  
  --recal-file output.recal \  
  - mode SNP
```

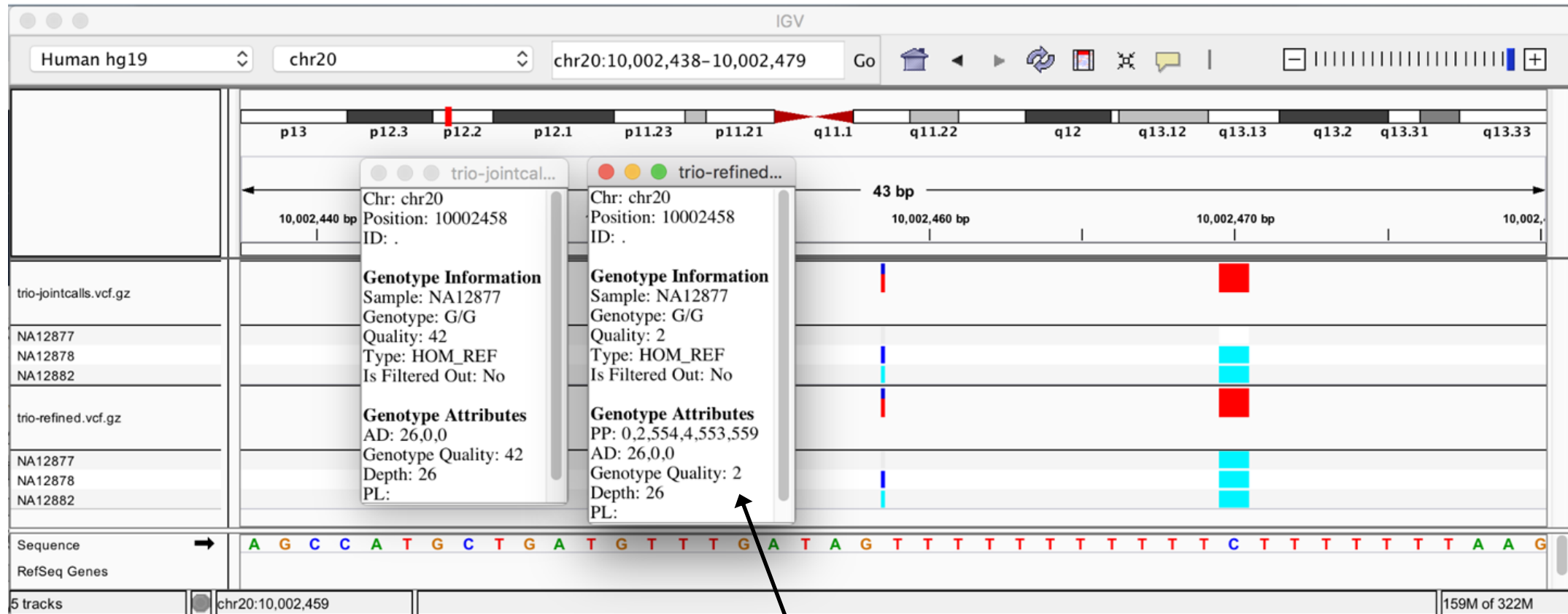


# Optional: Genotype refinement

```
# gatk CalculateGenotypePosteriors \  
  -V sandbox/trio-jointcalls.vcf.gz \  
  -ped resources/trio-pedigree.ped \  
  --supporting-callsets resources/af-only-gnomad.vcf.gz \  
  -O sandbox/trio-refined.vcf.gz
```



# Refining assignments and adjusting genotype confidence



Genotype Quality reduced from 42 to 2 after refinement



# CNN single-sample workflow

- Uses Convolutional Neural Networks
- Useful for filtering *short tandem repeats*
  - Example: TATATATATA
- Useful for filtering *homopolymers*
  - Example: AAAAAAAAAA
- GATK includes pre-computed models:
  - SynDip, GiaB, Platinum Genomes
  - Or use your own training model

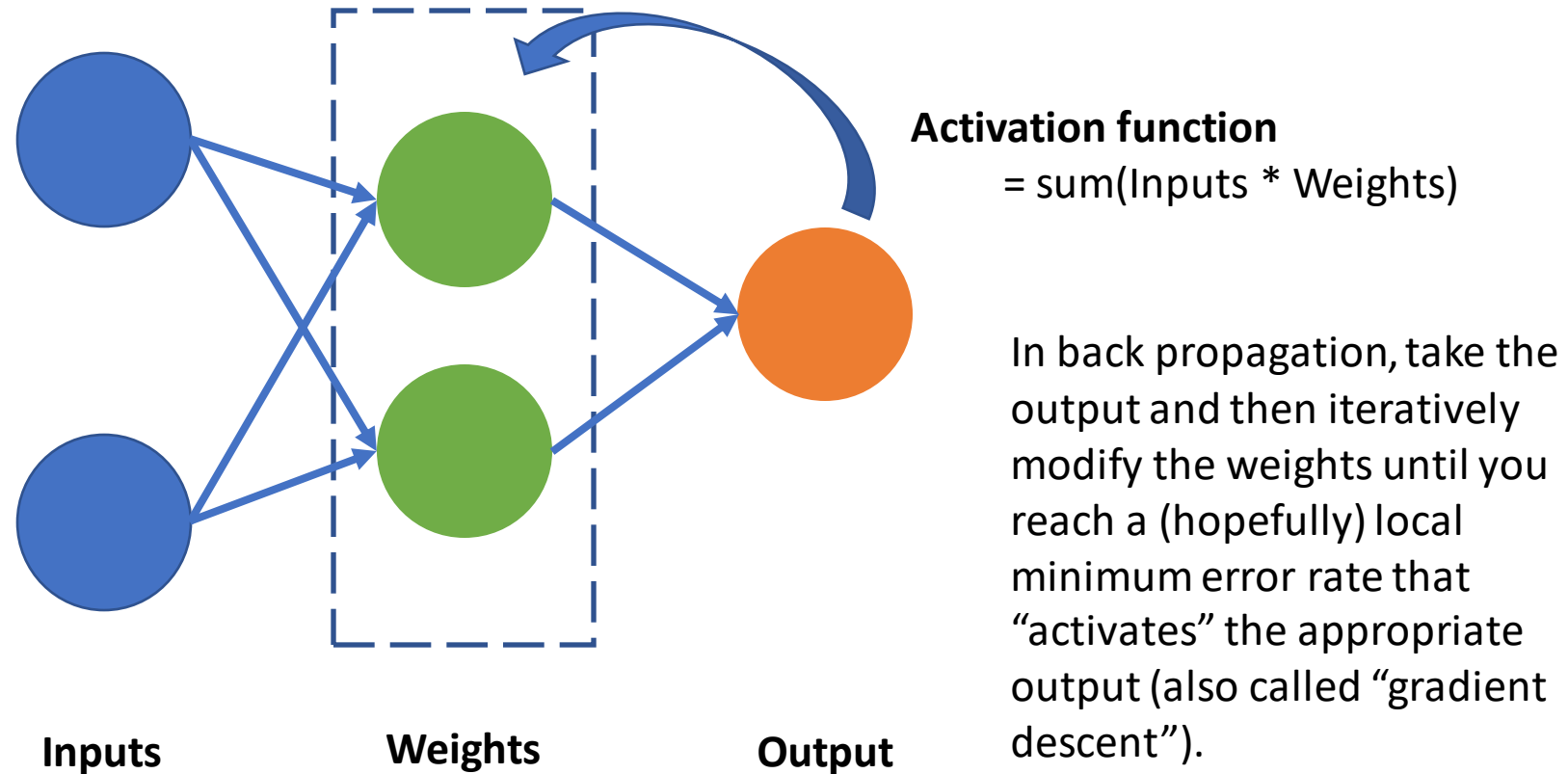


Chihuahua or Muffin?





# Neural networks in 60 seconds



# 1D CNN annotation

```
# cd ../cnn
# mkdir sandbox

# gatk CNNScoreVariants \
    -R ref/Homo_sapiens_assembly19.fasta \
    -V vcfs/g94982_b37_chr20_1m_15871.vcf.gz \
    -O sandbox/my_1d_cnn_scored.vcf
```



# 1D CNN filtering

```
# gatk FilterVariantTranches \  
  -V sandbox/my_1d_cnn_scored.vcf \  
  -O sandbox/my_1d_cnn_filtered.vcf \  
  --resource resources/1000G_omni2.5.b37.vcf.gz \  
  --resource resources/hapmap_3.3.b37.vcf.gz \  
  --info-key CNN_1D \  
  --snp-tranche 99.9 \  
  --indel-tranche 95.0
```



# 1D CNN output

```
# cat my_1d_cnn_filtered.vcf | grep -v '##' | head -3
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 1000072 rs6056638 A G 998.77 PASS
AC=2;AF=1.00;AN=2;CNN_1D=3.256;DB;DP=32;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF
=1.00;MQ=60.00;POSITIVE_TRAIN_SITE;QD=31.21;SOR=0.818;VQSLOD=20.79;culprit=MQ
GT:AD:DP:GQ:PL 1/1:0,32:32:96:1027,96,0
20 1000152 rs6056639 C T 678.77 PASS
AC=2;AF=1.00;AN=2;CNN_1D=2.313;DB;DP=28;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF
=1.00;MQ=60.00;POSITIVE_TRAIN_SITE;QD=24.24;SOR=0.693;VQSLOD=18.18;culprit=QD
GT:AD:DP:GQ:PL 1/1:0,28:28:81:707,81,0
```

## Candidate indel call filtered out by 1D CNN modeling

```
20 1012919 rs34579666 CT C 439.73 CNN_1D_INDEL_Tranche_95.00_100.00
AC=2;AF=1.00;AN=2;CNN_1D=-2.925;DB;DP=31;
ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=59.62;POSITIVE_TRAIN_SITE;QD=
19.12;SOR=1.708;VQSLOD=3.87;culprit=SOR
GT:AD:DP:GQ:PL 1/1:0,23:23:67:477,67,0
```



# 2D CNN annotation

```
# gatk CNNScoreVariants \  
  -R ref/Homo_sapiens_assembly19.fasta \  
  -I bams/g94982_chr20_1m_10m_bamout.bam \  
  -V vcfs/g94982_b37_chr20_1m_895.vcf \  
  -O sandbox/my_2d_cnn_scored.vcf \  
  --tensor-type read_tensor \  
  --transfer-batch-size 8 \  
  --inference-batch-size 8
```



# 2D CNN filtering

```
# gatk FilterVariantTranches \  
  -V sandbox/my_2d_cnn_scored.vcf \  
  -O sandbox/my_2d_cnn_filtered.vcf \  
  --resource resources/1000G_omni2.5.b37.vcf.gz \  
  --resource resources/hapmap_3.3.b37.vcf.gz \  
  --info-key CNN_2D \  
  --snp-tranche 99.9 \  
  --indel-tranche 95.0
```



# 2D CNN output

```
# cat my_2d_cnn_filtered.vcf | grep -v '##' | head -3
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 1000072 rs6056638 A G 998.77 PASS
AC=2;AF=1.00;AN=2;CNN_2D=7.687;DB;DP=32;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF
=1.00;MQ=60.00;POSITIVE_TRAIN_SITE;QD=31.21;SOR=0.818;VQSLOD=20.79;culprit=MQ
GT:AD:DP:GQ:PL 1/1:0,32:32:96:1027,96,0
20 1000152 rs6056639 C T 678.77 PASS
AC=2;AF=1.00;AN=2;CNN_2D=5.349;DB;DP=28;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF
=1.00;MQ=60.00;POSITIVE_TRAIN_SITE;QD=24.24;SOR=0.693;VQSLOD=18.18;culprit=QD
GT:AD:DP:GQ:PL 1/1:0,28:28:81:707,81,0
```

**The indel is still filtered out using 2D CNN modeling**

```
20 1012919 rs34579666 CT C 439.73 CNN_2D_INDEL_Tranche_95.00_100.00
AC=2;AF=1.00;AN=2;CNN_2D=5.576;DB;DP=31;
ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=59.62;POSITIVE_TRAIN_SITE;QD=
19.12;SOR=1.708;VQSLOD=3.87;culprit=SOR
GT:AD:DP:GQ:PL 1/1:0,23:23:67:477,67,0
```





# Calls made by 1D and 2D CNN models

1D CNN track

2D CNN track



# Additional resources

- About BWA-MEM
  - <https://arxiv.org/abs/1303.3997>
- About HaplotypeCaller
  - <https://www.biorxiv.org/content/10.1101/201178v3>
- GATK documentation: Germline short variant discovery (SNPs + Indels)
  - <https://gatk.broadinstitute.org/hc/en-us/articles/360035535932>
  - <https://github.com/gatk-workflows/gatk4-germline-snps-indels>



# Additional resources (continued)

- GATK Best Practices single-sample pipeline documentation (NEW!)
  - [https://broadinstitute.github.io/warp/documentation/Pipelines/Whole\\_Genome\\_Germline\\_Single\\_Sample\\_Pipeline/#set-up](https://broadinstitute.github.io/warp/documentation/Pipelines/Whole_Genome_Germline_Single_Sample_Pipeline/#set-up) (for WGS)
  - [https://broadinstitute.github.io/warp/documentation/Pipelines/Exome\\_Germline\\_Single\\_Sample\\_Pipeline/#set-up](https://broadinstitute.github.io/warp/documentation/Pipelines/Exome_Germline_Single_Sample_Pipeline/#set-up) (for WES)
- BroadE: GATK - Introduction to Germline Variant Discovery
  - [https://www.youtube.com/watch?v=F\\_U7ImMJkq4&t=326s](https://www.youtube.com/watch?v=F_U7ImMJkq4&t=326s)





# Thank you for joining us today!

Next week: Chapter 7

Next meeting: January 17, 2021