Introduction to Bioinformatics

Alexandros C. Dimopoulos
alexdem@di.uoa.gr

Master of Science
"Data Science and Information Technologies"
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens

2024-25

Variant Calling I

- Variants: differences between two genomes
- It is now feasible (technical and financial wise) to sequence human samples at large scale for medical and genetic studies
- Major projects, e.g.:
  - 1000 Genomes project (http://www.internationalgenome.org/)
  - The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/)

Variant Calling II

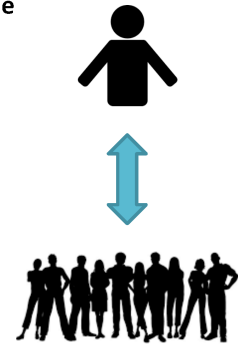
- Clarify the full spectrum of human genetic diversity
- Study the complete genetic architecture of human diseases
- Find mutations that hide links to Mendelian diseases
- Find mutations for which no mapping data is available, e.g.
  - somatic mutations in cancer
  - de novo mutations in autism and schizophrenia

Variant Calling III

- Mapping raw reads (fastq file) into a genome (fasta file)
  - creation of a bam file
- Search (per base) for differences between the bam file and the genome and create a vcf (variant call format) file
  - misaligned reads e.g. because of a low quality read
  - SNP (Single Nucleotide Polymorphism): different nucleotide in just one position
  - INDEL (INsertion/DELetion): a small number of nucleotides has been inserted or deleted
  - CNV (Copy Number Variation): repetition or deletion of larger blocks of nucleotides
- It is hard to distinguish a real polymorphism from artifacts

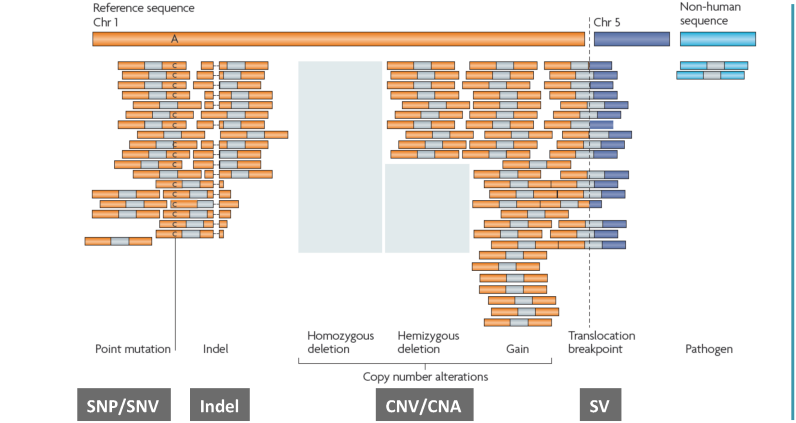
Variant discovery I

- Genetic changes in individuals **relative to a reference genome**
  - Germline (inherited)
  - Somatic (cancer)
- **Reference genome** = a standardized genomic sequence
- Human genome reference sequence
  - Current standard: hg19 / b37
  - New standard (on the rise): hg38
- Other organisms
  - Many have a fully assembled reference available
  - Many still do not -> SOL



https://software.broadinstitute.org/gatk/documentation/presentations

Variant discovery II



https://software.broadinstitute.org/gatk/documentation/presentations

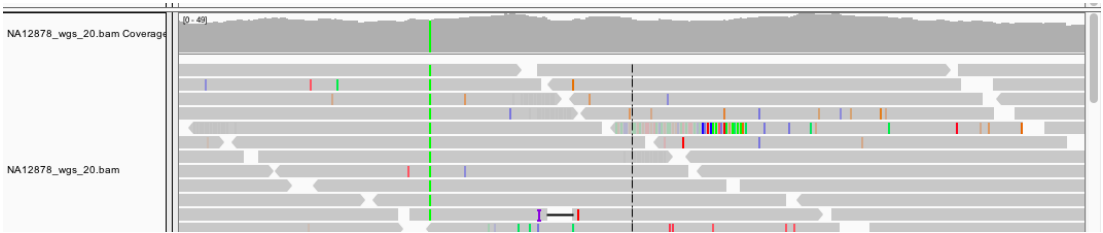
IGV I

Integrative Genomics Viewer - Variant Calling

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

`http://software.broadinstitute.org/software/igv/`

IGV II



Various options for Variant Calling

- Samtools mpileup
- Freebayes
- VarScan
- Atlas2
- **GATK**
- …

GATK

Genome Analysis Toolkit - GATK

A collection of command-line tools for analyzing high-throughput sequencing (HTS) data in formats such as SAM/BAM/CRAM and VCF, with a focus on variant discovery.

GATK

Genome Analysis Toolkit - GATK

A collection of command-line tools for analyzing high-throughput sequencing (HTS) data in formats such as SAM/BAM/CRAM and VCF, with a focus on variant discovery.

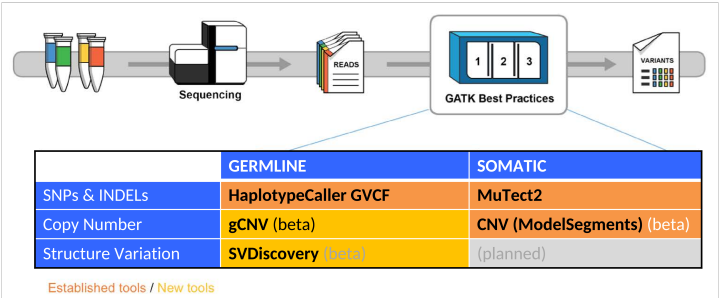
A multi-step procedure divided into 3 parts:

- Pre-processing
- Variant discovery
- Callset refinement

GATK Overview I

https://software.broadinstitute.org/gatk/documentation/presentations

GATK Overview II

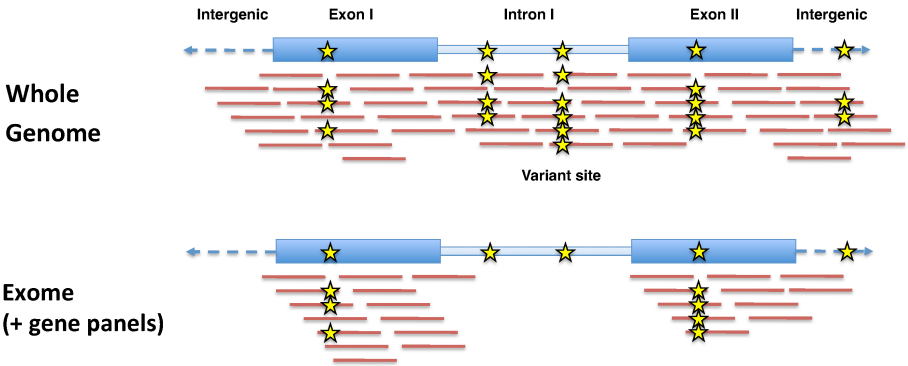


	GERMLINE	SOMATIC
SNPs & INDELs	**HaplotypeCaller GVCF**	**MuTect2**
Copy Number	**gCNV** (beta)	**CNV (ModelSegments)** (beta)
Structure Variation	**SVDiscovery** (beta)	(planned)

Established tools / New tools

https://software.broadinstitute.org/gatk/documentation/presentations

GATK Overview III



https://software.broadinstitute.org/gatk/documentation/presentations

GATK - Technical details

- Java wrapper

```
gatk --version
```

```
java -Dsamjdk.use_async_io_read_samtools=false
-Dsamjdk.use_async_io_write_samtools=true
-Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2
-jar /opt/gatk-4.4.0.0/gatk-package-4.4.0.0-local.jar --version
```

- Collection of various tools

```
gatk --java-options "-Xmx4G" ToolName [tool arguments]
```

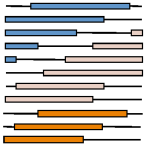
```
gatk HaplotypeCaller -R reference.fasta -I sample1.bam -O
variants.vcf
```

- The jar file is compiled for POSIX systems (i.e. non-Windows)

Pre-processing I

https://software.broadinstitute.org/gatk/documentation/presentations

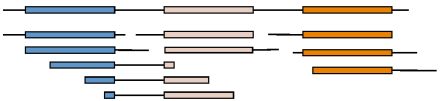
Pre-processing II



Enormous pile of short reads from HTS

Mapping and alignment algorithms

- **BWA for DNA**
- **STAR for RNAseq**



Reference genome

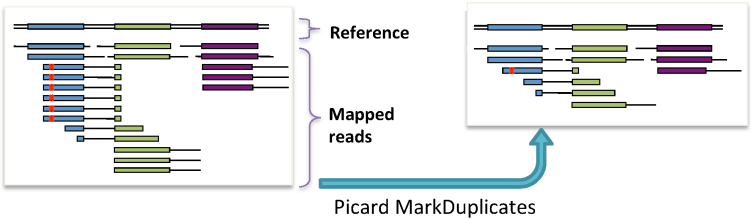
Reads mapped to reference

https://software.broadinstitute.org/gatk/documentation/presentations

Mark-Duplicates I

Duplicates = **non-independent measurements** of a sequence fragment

-> Must be removed to assess support for alleles correctly

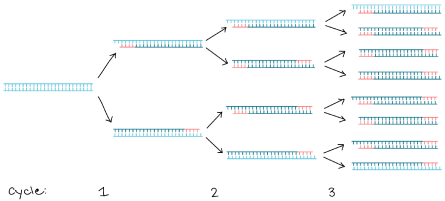


✖ = sequencing error propagated in duplicates

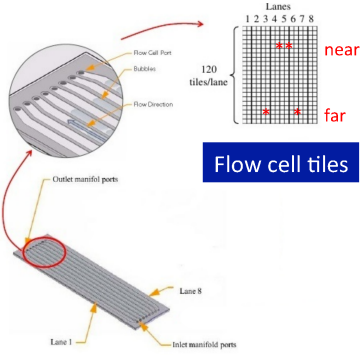
https://software.broadinstitute.org/gatk/documentation/presentations

Mark-Duplicates II

- **LIBRARY DUPLICATES**
  - **Increases with PCR cycles**
- **OPTICAL DUPLICATES**
  - **Are nearby clusters on a flow cell lane**

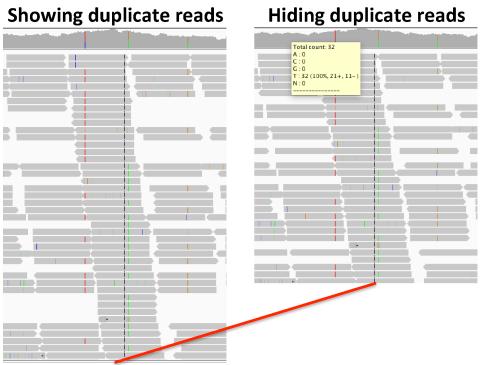


https://www.khanacademy.org/science/biology/biotech-dna-technology/dna-sequencing-pcr-electrophoresis/a/polymerase-chain-reaction-pcr



http://www.slideshare.net/jandot/next-generation-sequencing-course-part-2-sequence-mapping
http://www.slideshare.net/cosentia/illumina-gaiix-for-high-throughput-sequencing

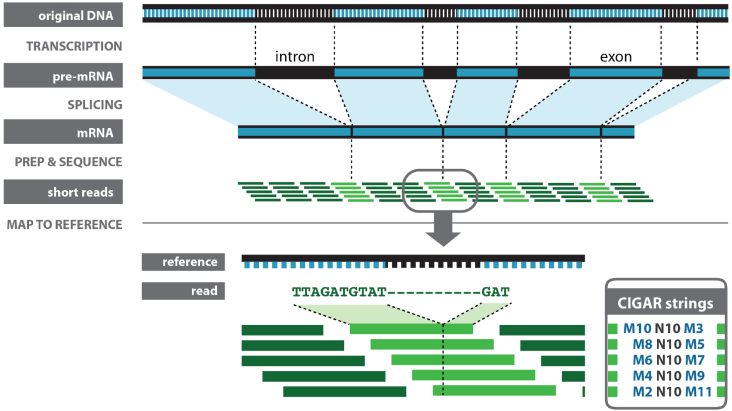
Mark-Duplicates III



- Duplicate status is indicated in SAM flag
- Duplicates are **not removed**, just tagged (unless you request removal)
- Downstream tools can read the tag and choose to ignore those reads
- Most GATK tools ignore duplicates by default

https://software.broadinstitute.org/gatk/documentation/presentations

Special handling for RNAseq splice junctions



https://software.broadinstitute.org/gatk/documentation/presentations

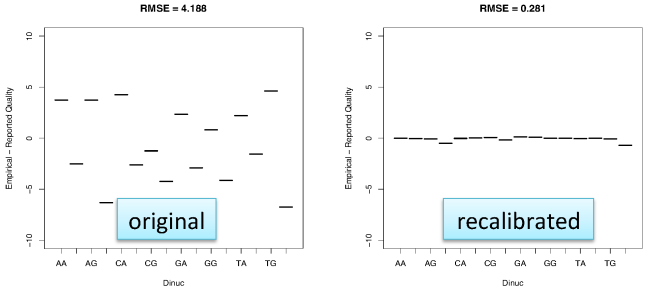
How-to map and clean up short read sequence data efficiently

- (How to) Map and clean up short read sequence data efficiently
- (How to) Fix a badly formatted BAM

Base Recalibration (BQSR) I

- Sequencers make systematic errors in base quality scores
- BQSR corrects the quality scores (not the bases)

Example of bias: qualities reported depending on nucleotide context

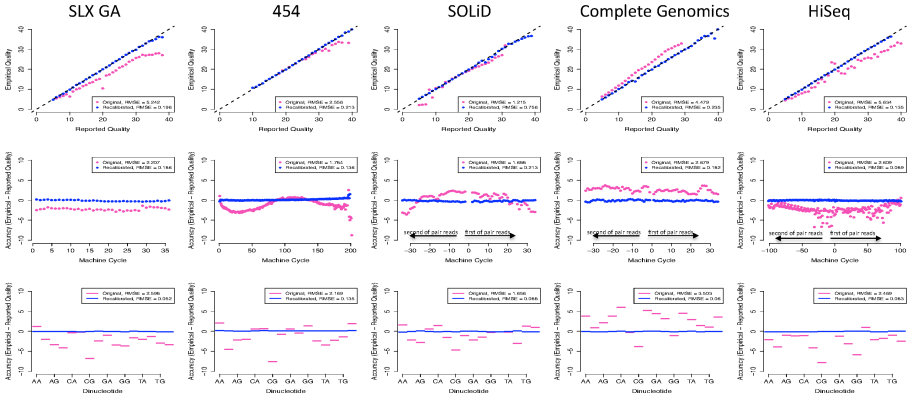


https://software.broadinstitute.org/gatk/documentation/presentations

Base Recalibration (BQSR) II

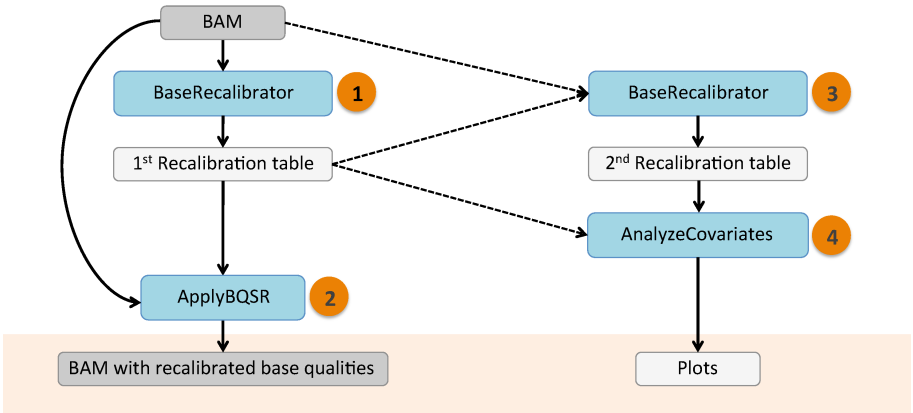
https://software.broadinstitute.org/gatk/documentation/presentations

Base Recalibration (BQSR) III



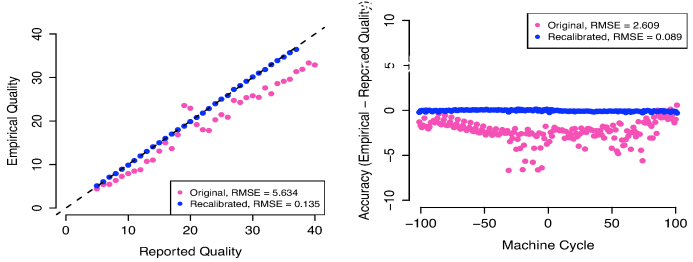
https://software.broadinstitute.org/gatk/documentation/presentations

Base Recalibration (BQSR) IV



https://software.broadinstitute.org/gatk/documentation/presentations

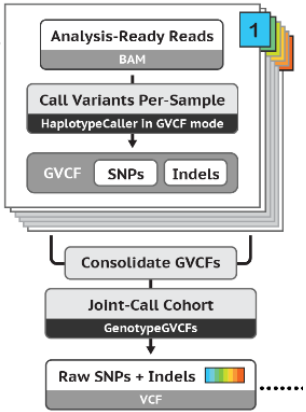
Base Recalibration (BQSR) V



https://software.broadinstitute.org/gatk/documentation/presentations

▸ Base Quality Score Recalibration (BQSR)

GATK - Variant discovery



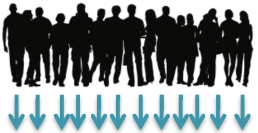
https://software.broadinstitute.org/gatk/documentation/presentations

Variant discovery I

- Single genome in isolation: almost never useful
- Family or population data add valuable information
  - rarity of variants
  - *de novo* mutations
  - ethnic background

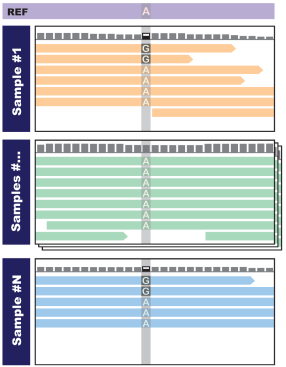
https://software.broadinstitute.org/gatk/documentation/presentations

Variant discovery II



https://software.broadinstitute.org/gatk/documentation/presentations

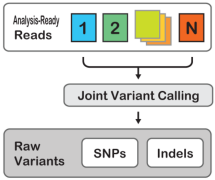
Variant discovery III



- Sample #1 or Sample #N alone:
  - **weak evidence for variant**
  - **may miss calling the variant**
- Both samples seen together:
  - **unlikely to be artifact**
  - **call the variant more confidently**

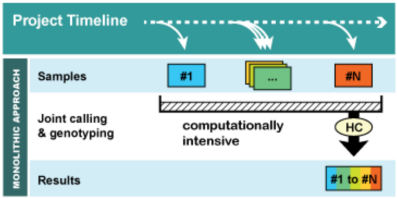
https://software.broadinstitute.org/gatk/documentation/presentations

Variant discovery - UnifiedGenotyper



Compute requirements scale very badly with number of samples!!!

It gives us the right answers, but…

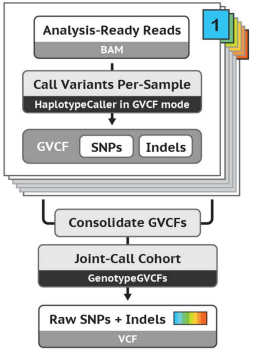


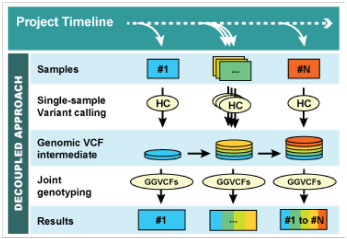
Want to add new samples?

Got to re-run pipeline from scratch! The N+1 problem!

https://software.broadinstitute.org/gatk/documentation/presentations

Variant discovery - HaplotypeCaller



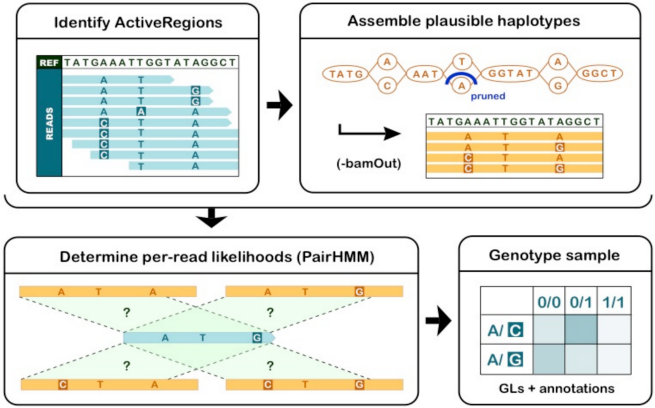


Scales linearly with number of samples!

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

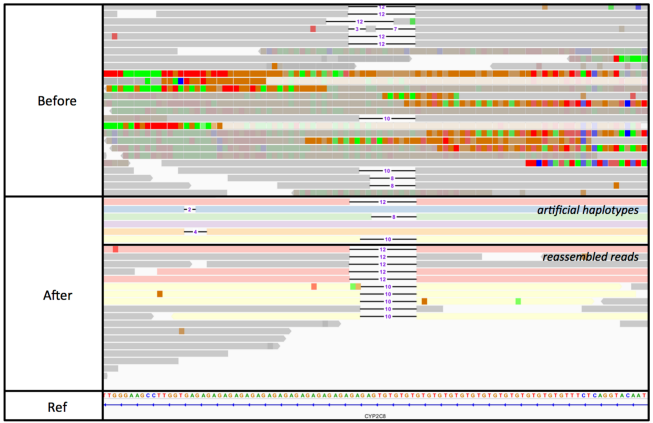
https://software.broadinstitute.org/gatk/documentation/presentations

HaplotypeCaller I



https://software.broadinstitute.org/gatk/documentation/presentations

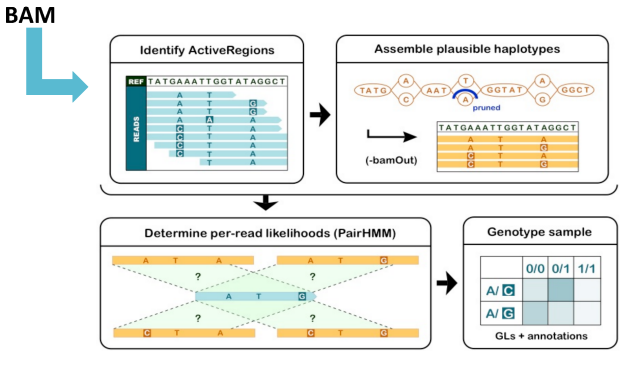
HaplotypeCaller II



Showing 100bp region starting at 10:96,825,862 for NA12878. IGV is a snapshot version from 2017/8/28

https://software.broadinstitute.org/gatk/documentation/presentations

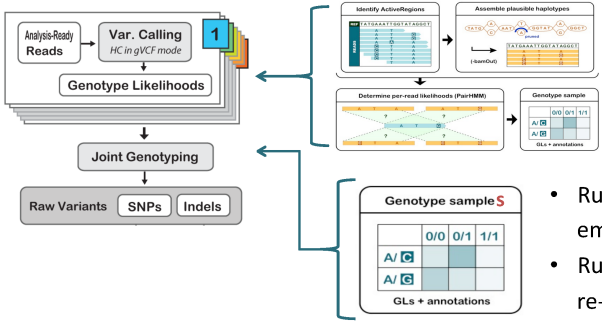
HaplotypeCaller III



This is all you need for a **single sample** or **traditional multi-sample** analysis

https://software.broadinstitute.org/gatk/documentation/presentations

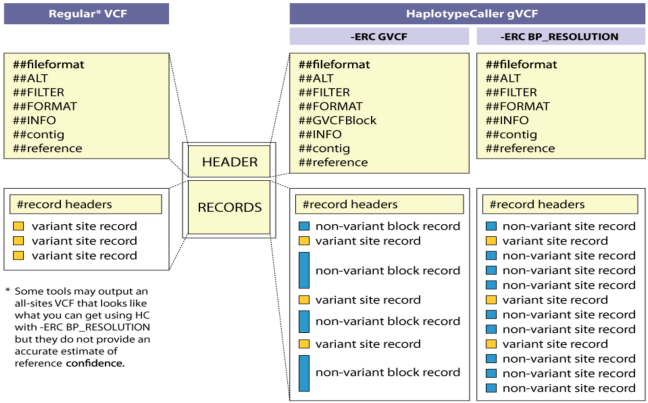
HaplotypeCaller IV



- Run HC in **GVCF mode** to emit GVCF
- Run GenotypeGVCFs to re-genotype samples with **multi-sample model**

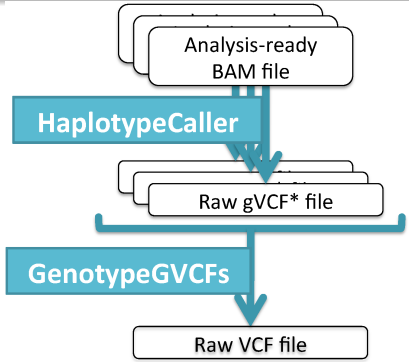
https://software.broadinstitute.org/gatk/documentation/presentations

HaplotypeCaller V



https://software.broadinstitute.org/gatk/documentation/presentations

HaplotypeCaller VI




```
java -jar GenomeAnalysisTK.jar \
     -T HaplotypeCaller \
     -R human.fasta \
     -I sample1.bam \
     -o sample1.g.vcf \
       [ -L exome_targets.intervals \ ]
     -ERC GVCF
```

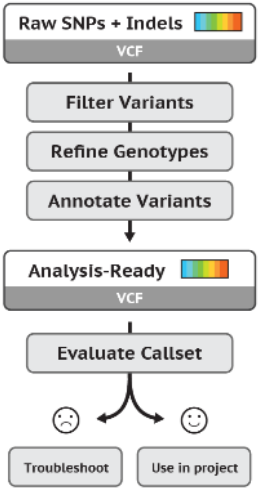
```
java -jar GenomeAnalysisTK.jar \
     -T GenotypeGVCFs \
     -R human.fasta \
     -V sample1.g.vcf \
     -V sample2.g.vcf \
     -V sampleN.g.vcf \
     -o output.vcf
```

If >200 samples, combine in batches first using CombineGVCFs

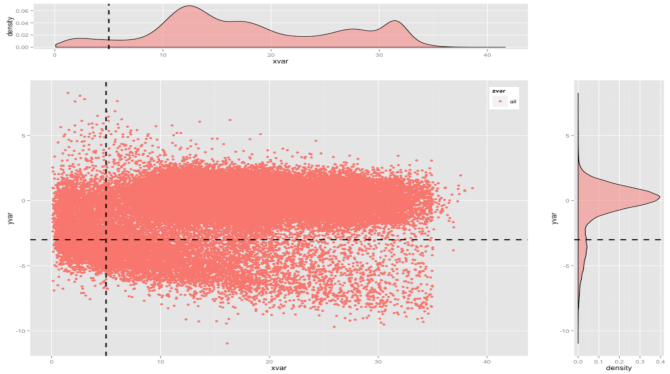
https://software.broadinstitute.org/gatk/documentation/presentations

▸ Germline short variant discovery (SNPs + Indels)

VCF Filtering

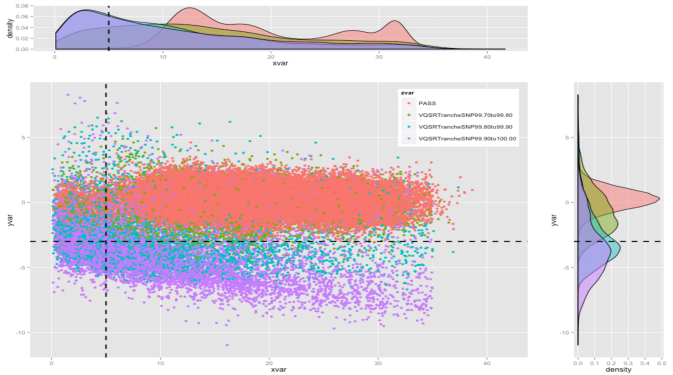


VCF Filtering - Hard filter



https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration I



https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration II

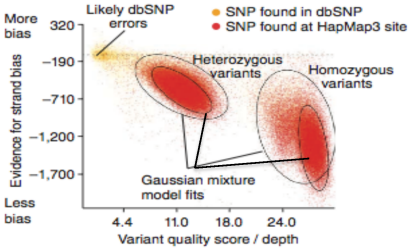
Train on high-confidence known sites to determine the probability that other sites are true or false

- Assume annotations tend to form **Gaussian clusters**
- Build a "Gaussian mixture model" from annotations of **known variants** in our dataset
- Score **all variants** by where their annotations lie relative to these clusters
- Filter base on **sensitivity to truth set**

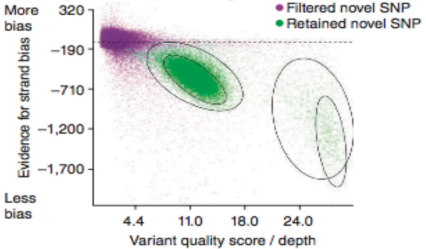
https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration III

Model trained on HapMap



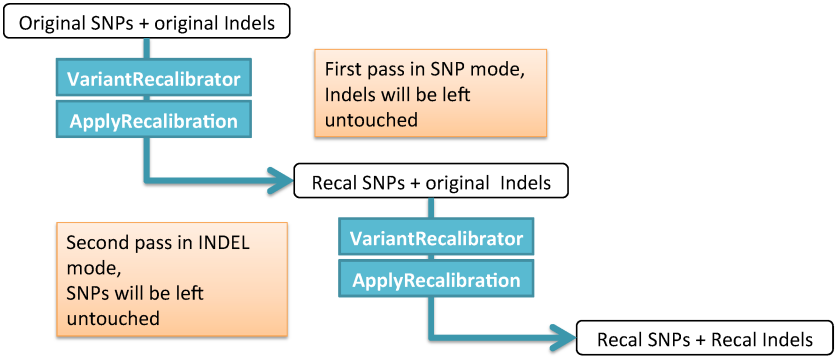
Model applied to new SNPs



Modified from DePristo et al. Nature Genetics. 2011

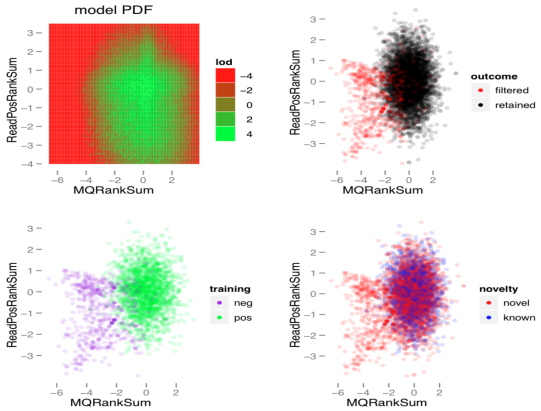
https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration IV



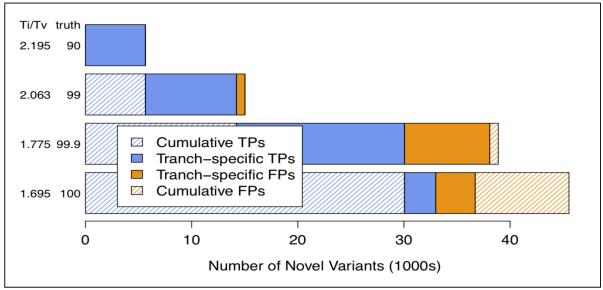
https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration V



https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration VI

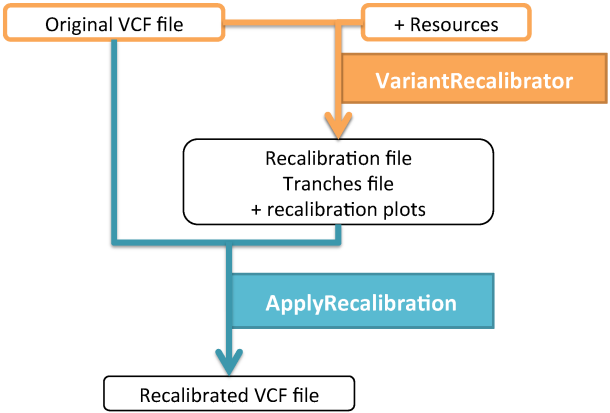


Estimation is based on Ti/Tv ratio of novel variants

Default target Ti/Tv is for WGS and must be adapted for exomes

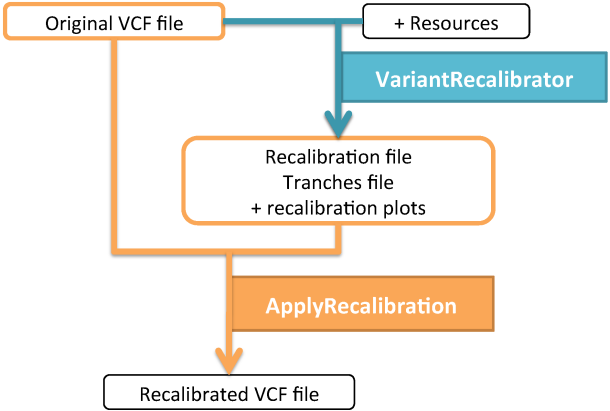
https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration VII



https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration VIII



https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration IX

- Variant Quality Score Recalibration (VQSR)

VCF Filtering - Variant recalibration X

- Before VQSR (input vcf):

#CHROM	POS	FILTER	INFO
1	10146	.	AC=1;DP=32;FS=9.208; MQ=31.96;MQRankSum=0.085;...
1	10403	.	AC=1;DP=64;FS=1.645;MQ=41.86;MQRankSum=1.87;...
1	234313	.	AC=1;DP=239;FS=12.675;MQ=38.19;MQRankSum=-0.122;...

- After VQSR (output vcf):

#CHROM	POS	FILTER	INFO
1	10146	VQSRTrancheINDEL99.30to99.50	AC=1...;NEGATIVE_TRAIN_SITE;VQSLOD=-1.328;culprit=SOR
1	10403	PASS	AC=1;...;QD=0.60; VQSLOD=0.794;culprit=QD
1	234313	VQSRTrancheSNP99.90to100.00	AC=1;...;POSITIVE_TRAIN_SITE;VQSLOD=-5.356;culprit=MQ

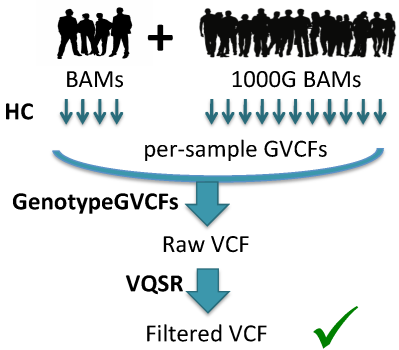
- Hard filtered vcf:

#CHROM	POS	FILTER	INFO
1	10146	PASS	AC=1;DP=32;FS=9.208; MQ=31.96;MQRankSum=0.085;...
1	10403	INDEL_Filter	AC=1;DP=64;FS=1.645;MQ=41.86;MQRankSum=1.87;...
1	234313	SNP_Filter	AC=1;DP=239;FS=12.675;MQ=38.19;MQRankSum=-0.122;...

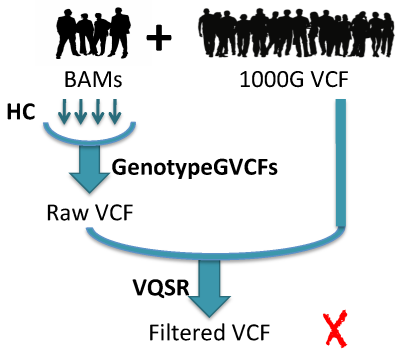
https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration XI

ALWAYS do this:

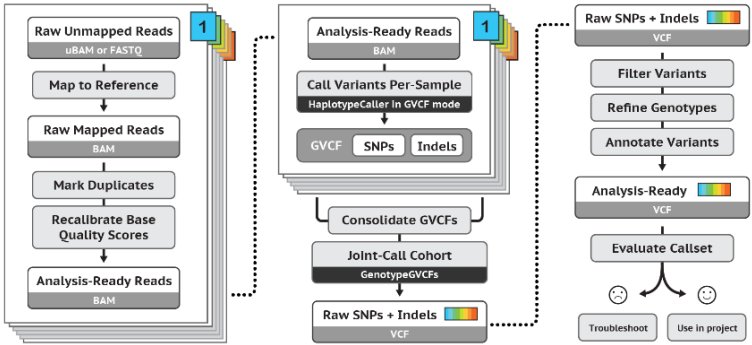


NEVER do this :



https://software.broadinstitute.org/gatk/documentation/presentations

Presented GATK pipeline



https://software.broadinstitute.org/gatk/documentation/presentations

Variant Annotation I

- Variant annotation is a very important step in the analysis
- Functional annotation can have a strong impact on the final conclusions of the studies
- Inaccurate or incorrect annotation can lead to the skipping of polymorphisms potentially responsible for a disease or to conceal interesting variations in a group of false positives

Variant Annotation II

Various tools for annotation:

- Funcotator (GATK)
- SnpEff
- Annovar
- VEP

Funcotator I

Funcotator

Funcotator (FUNCtional annOTATOR) analyzes given variants for their function (as retrieved from a set of data sources) and produces the analysis in a specified output file. This tool is a functional annotation tool that allows a user to add annotations to called variants based on a set of data sources, each with its own matching criteria.

Funcotator II

- For **somatic** data sources:

```
./gatk FuncotatorDataSourceDownloader --somatic --validate-integrity --extract-after-download
```

- For **germline** data sources:

```
./gatk FuncotatorDataSourceDownloader --germline --validate-integrity --extract-after-download
```

▸ Funcotator Information and Tutorial

SnpEff

SnpEff

SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes).

http://snpeff.sourceforge.net/SnpEff.html

SnpEff:Basic example

```
java -Xmx4g -jar snpEff.jar GRCh37.75 examples/test.chr22.vcf >
test.chr22.ann.vcf
```

SnpEff:Basic example

```
java -Xmx4g -jar snpEff.jar GRCh37.75 examples/test.chr22.vcf >
test.chr22.ann.vcf
```

SnpEff adds functional annotations in the ANN field (8^{th} column in the VCF file test.chr22.ann.vcf)

- Putative_impact: A simple estimation of putative impact / deleteriousness : HIGH, MODERATE, LOW, MODIFIER

  frameshift_variant, stop_gained, stop_lost, start_lost, . . .
- Gene Name: Common gene name (HGNC). Optional: use closest gene when the variant is “intergenic”
- Gene ID: Gene ID
- . . .

Annovar

ANNOVAR

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, hg38, as well as mouse, worm, fly, yeast and many others.

http://annovar.openbioinformatics.org/en/latest/

check also wANNOVAR

Variant Effect Predictor

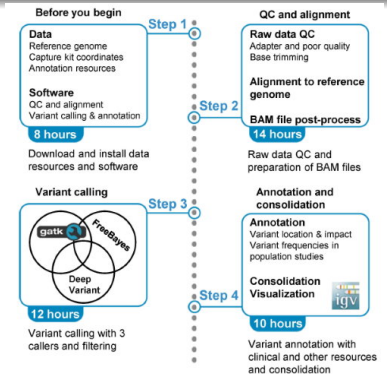
Variant Effect Predictor - VEP

VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

- Standalone perl script
- Web interface

https://www.ensembl.org/info/docs/tools/vep/index.html

Combining three variant callers (HaplotypeCaller, FreeBayes, and DeepVariant)



> STAR Protoc. 2022 May 30;3(2):101418. doi: 10.1016/j.xpro.2022.101418. eCollection 2022 Jun 17.

Protocol for unbiased, consolidated variant calling from whole exome sequencing data

Kleio-Maria Verrou [1], Georgios A Pavlopoulos [1 2], Panagiotis Moulos [1 2]

Affiliations — collapse

Affiliations

1 Center of New Biotechnologies & Precision Medicine, Medical School, National and Kapodistrian University of Athens, Athens, Greece.
2 Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center 'Alexander Fleming', Vari, Greece.

PMID: 35669050 PMCID: PMC9163752 DOI: 10.1016/j.xpro.2022.101418
Free PMC article

https://pubmed.ncbi.nlm.nih.gov/35669050/

Hands on

Lab Exercise 6 - GATK TUTORIAL :: Variant Discovery

All the necessary files are alreade stores at your home folder:
`~/GATK_tutorial/data`

Questions ?