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

2025-26

Variant Calling I

- It is now feasible (technical and financial wise) to sequence human samples at large scale for medical and genetic studies
- **High-Throughput Sequencing/Next-Generation Sequencing (NGS)** allows millions of short DNA fragments (reads) to be sequenced rapidly and affordably
- **Read Alignment:** Sequenced reads are mapped to a known **Reference Genome**
- **Coverage (C):** The average number of reads overlapping a specific position in the genome
  - High coverage ($C > 30\times$) is crucial for confident variant calling
- Major projects, e.g.:
  - 1000 Genomes project (http://www.internationalgenome.org/)
  - The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/)

Variant Calling II

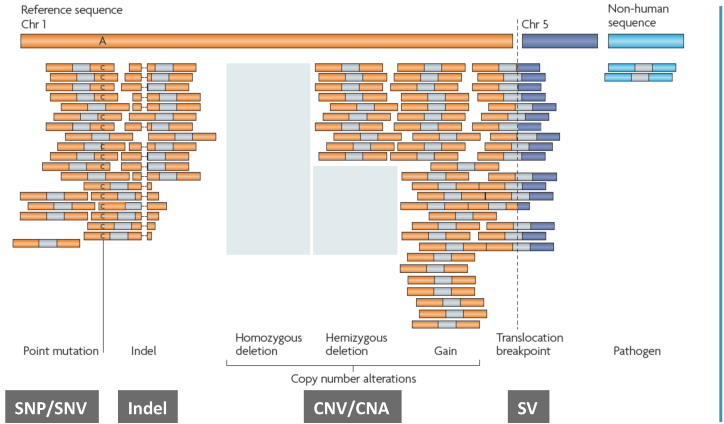
What is Genetic Variation?

- **Definition:** Differences in the DNA sequence among individuals or populations
- **Importance:** The foundation of biological diversity, disease susceptibility, and evolution

Major Types of Genetic Variation

- **SNPs (Single Nucleotide Polymorphisms):** Single base-pair changes (A $\rightarrow$ G)
- **Indels (Insertions/Deletions):** Small additions or removals of 1 to 50 base pairs
- **Structural Variants (SVs):** Large-scale changes ($>$ 50 bp), e.g. CNV (Copy Number Variations), inversions

Variant Calling III



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

Variant Calling IV

- Clarify the full spectrum of human genetic diversity
- Identify disease-associated mutations
- Find mutations for which no mapping data is available, e.g.
  - somatic mutations in cancer
  - de novo mutations in autism and schizophrenia
- Population genomics and evolutionary studies
- Personalized medicine and pharmacogenomics
- Downstream analyses: GWAS, functional annotation

Variant Calling V

- 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
- **Goal:** To accurately identify differences (variants) between the aligned sequence data of a sample and the established reference genome
- **Output:** A list of genomic positions where the sample's DNA differs from the reference, along with the confidence/quality score of the call

Variant Calling VI

Why is it a Statistical Problem?

The process must distinguish true biological variation from noise introduced by the sequencing process, which includes random errors, alignment ambiguities, and systematic biases. This is achieved through statistical modeling (e.g. Bayesian inference)

The Standard Output: VCF

- **VCF (Variant Call Format):** The universal text format for storing gene sequence variations
- **Key Fields:**
  - CHROM: Chromosome name
  - POS: Position of the variant
  - REF: Reference allele
  - ALT: Alternative allele(s)
  - QUAL: Phred-scaled quality score for the assertion that one or more ALT alleles exist
  - FILTER: Indicates if the variant passed the quality filters
  - INFO: Additional information (e.g. allele frequency)

NGS Data Pre-processing (I): Quality Control

1. **Initial Quality Control (QC):**
   - Assessing raw FASTQ read quality using tools like FastQC
   - Examining per-base quality scores (Phred scores) and adapter content
2. **Trimming and Filtering:**
   - Removing low-quality bases from the ends of reads
   - Eliminating sequencing adapter sequences

Alignment: Reads are mapped to the reference genome (e.g. using $BWA - MEM$). The output is a SAM/BAM (Sequence Alignment Map) file

Core Concept: Pileup and Genotype Likelihood

- **Pileup:** A visualization/data structure showing all aligned bases covering a single genomic position
- **Allele Counting:** The variant caller counts the reference and alternative alleles in the pileup
- **Genotype Likelihood ($\mathcal{L}$):** The core of robust calling
  - Uses statistical models to calculate the probability of observing the sequence reads (D) given a specific genotype (G): $P(D|G)$
  - The most common approach uses **Bayesian Inference** to estimate $P(G|D)$, the probability of the genotype given the data

$$P(G|D) \propto P(D|G) \cdot P(G)$$

Challenges in Variant Calling

- **Sequencing Errors:** Can be mistaken for a true alternative allele (False Positive)
- **Low Coverage:** Makes it difficult to distinguish a true heterozygous state (A/T) from sequencing noise, leading to False Negatives or incorrect Genotyping
- **Repetitive Regions:** Reads from these regions may map to multiple locations (ambiguous mapping), leading to spurious calls
- **Indels:** Errors near small insertions/deletions can cause misalignment, requiring complex **Local Realignment** (now handled by Haplotype − based models)

Advanced Algorithmic Approaches

- **Pileup/Base-Counting Models (e.g. Samtools):** Simple and fast. Directly use base counts and quality scores to estimate genotype likelihoods
- **Haplotype-based Models (e.g. GATK HaplotypeCaller):**
  - **The Gold Standard:** Considers the phasing of variants (i.e. which variants appear together on the same chromosome copy)
  - Performs **local de novo assembly** of reads around a variant site to define the best possible local haplotype
  - Dramatically improves accuracy for Indels and complex regions

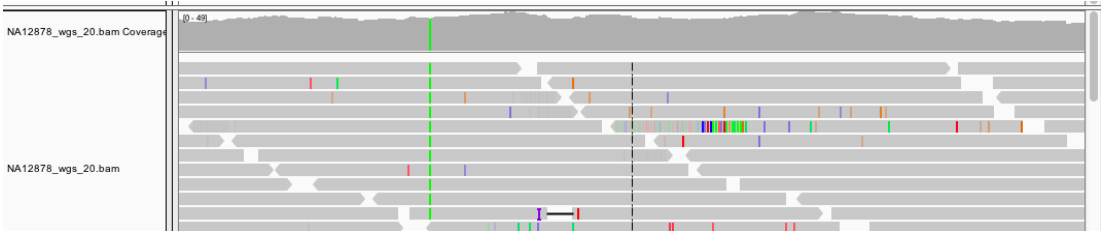
IGV I

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

Tool	Key Method / Focus	Typical Use Case
GATK	Haplotype-based modeling	Gold standard for Germline Variants
Samtools/BCFtools	Pileup-based, highly efficient	Fast filtering and manipulation of VCF/BAM
FreeBayes	Bayesian genetic variant detector	Population studies, non-model organisms
DeepVariant	Deep Neural Networks (DNN)	High accuracy, requires TPU/GPU acceleration

Focus of this presentation: GATK, the Broad Institute's Best Practices

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

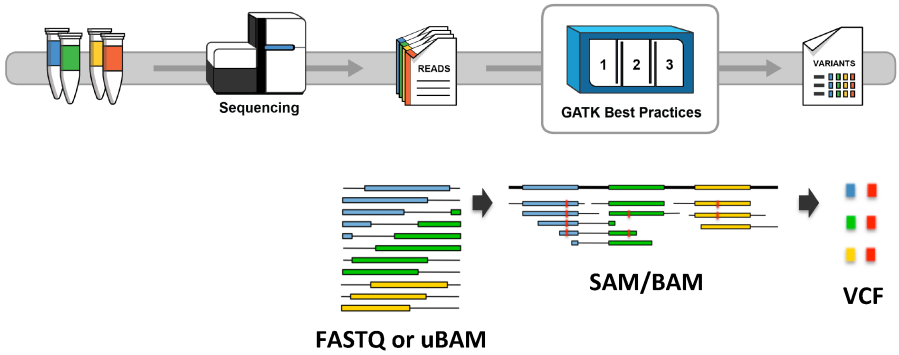
- **Origin:** Developed by the Broad Institute (Cambridge, MA)
- **Status:** Widely adopted as the **gold standard** for variant discovery from NGS data, especially for short germline variants
- **Design Philosophy:** To provide a robust, consistent, and computationally efficient set of tools based on state-of-the-art statistical models

GATK Best Practices Workflow

A multi-step procedure divided into 3 parts:

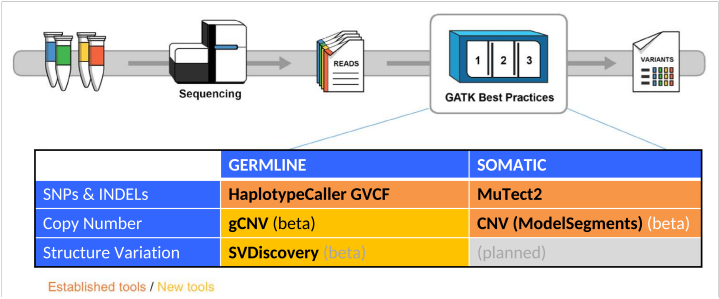
1. **Pre-processing:** Data refinement to correct systematic errors
2. **Variant Calling:** Identifying variants using the HaplotypeCaller
3. **Post-Calling Filtering:** Recalibrating and filtering low-quality calls

GATK Overview I



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

GATK Overview II

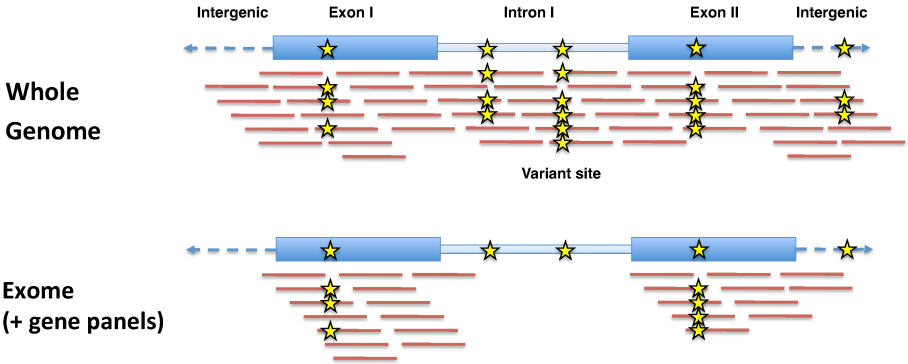


	GERMLINE	SOMATIC
SNPs & INDELs	**HaplotypeCaller GVCF**	**MuTect2**
Copy Number	**gCNV** (beta)	**CNV (ModelSegments)** (beta)
Structure Variation	**SVDiscovery** (beta)	(planned)

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

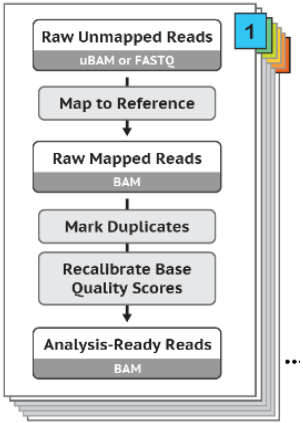
```
gatk --version
    java -Dsamjdk.use_async_io_read_samtools=false
    -Dsamjdk.use_async_io_write_samtools=true
    -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2
    -jar /opt/gatk-4.4.0.0/gatk-package-4.4.0.0-local.jar --version
```

- Collection of various tools

```
gatk --java-options "-Xmx4G" ToolName [tool arguments]
    gatk HaplotypeCaller -R reference.fasta -I sample1.bam -O
    variants.vcf
```

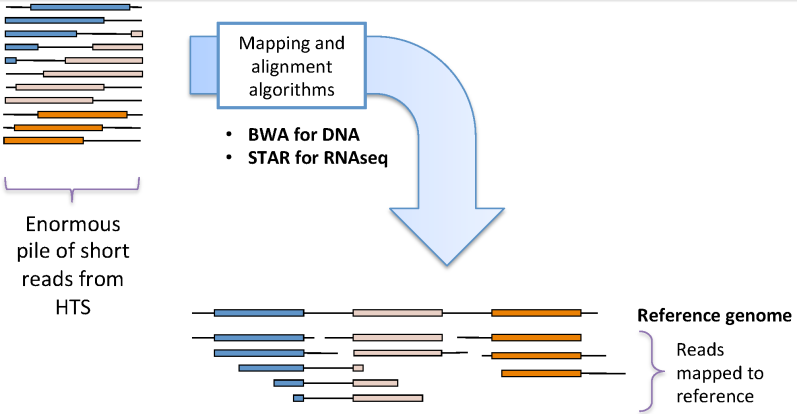
- The jar file is compiled for POSIX systems (i.e. non-Windows)

Pre-processing I



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

Pre-processing II

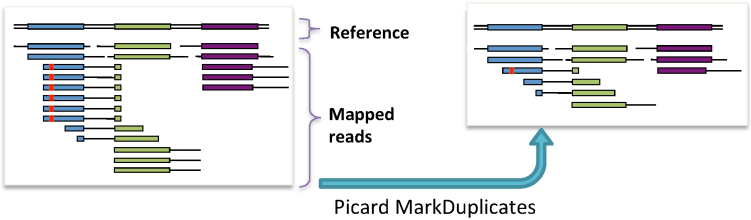


https://software.broadinstitute.org/gatk/documentation/presentations

Mark-Duplicates I

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

-> Must be removed to assess support for alleles correctly

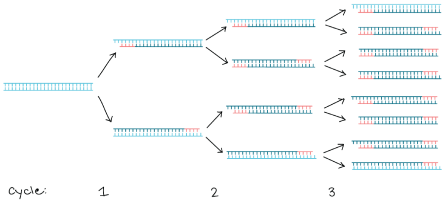


✖ = sequencing error propagated in duplicates

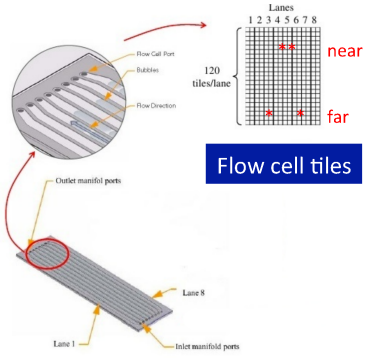
https://software.broadinstitute.org/gatk/documentation/presentations

Mark-Duplicates II

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

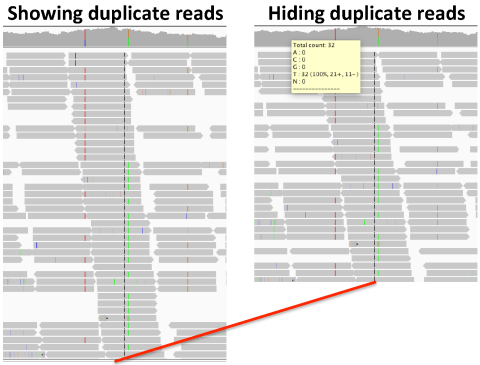


https://www.khanacademy.org/science/biology/biotech-dna-technology/dna-sequencing-pcr-electrophoresis/a/polymerase-chain-reaction-pcr



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

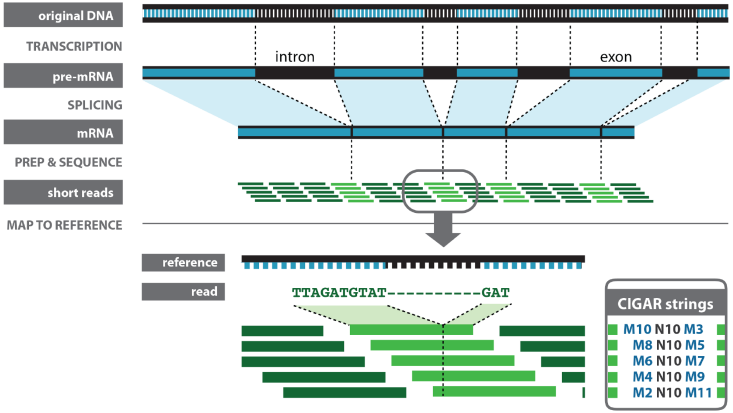
Mark-Duplicates III



- 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

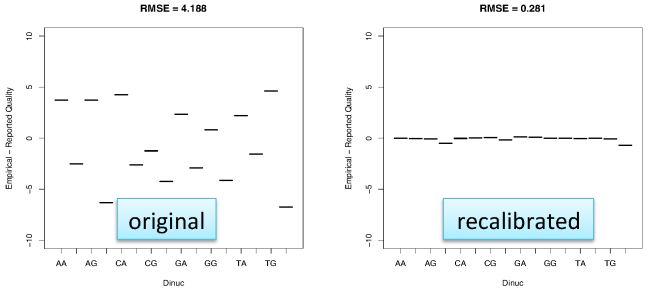
How-to map and clean up short read sequence data efficiently

- (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

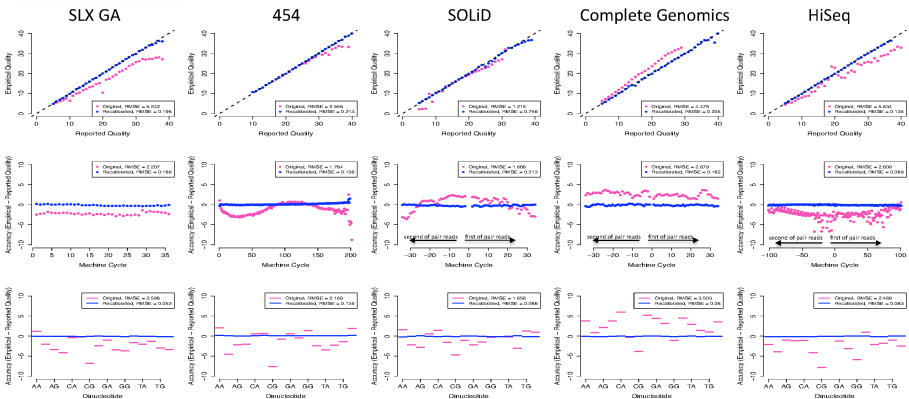


https://software.broadinstitute.org/gatk/documentation/presentations

Base Recalibration (BQSR) II

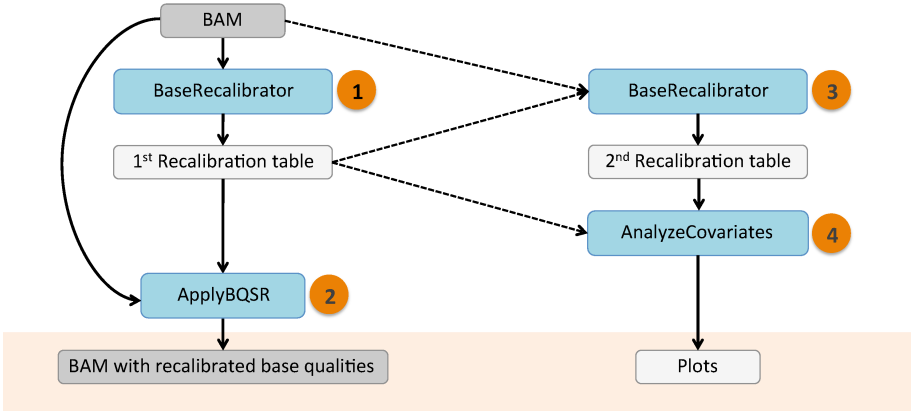
https://software.broadinstitute.org/gatk/documentation/presentations

Base Recalibration (BQSR) III



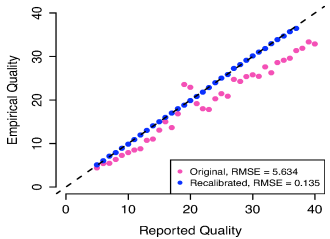
https://software.broadinstitute.org/gatk/documentation/presentations

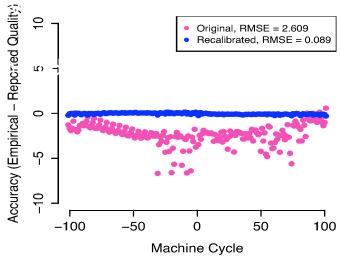
Base Recalibration (BQSR) IV



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

Base Recalibration (BQSR) V

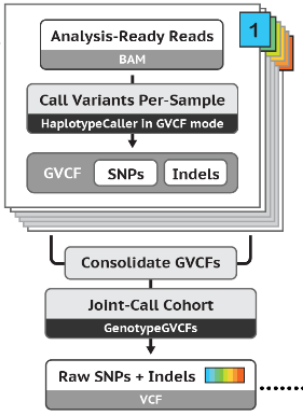




https://software.broadinstitute.org/gatk/documentation/presentations

Base Quality Score Recalibration (BQSR)

GATK - Variant discovery



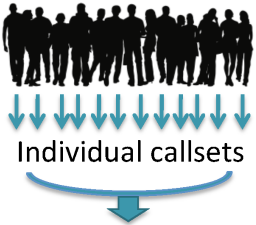
https://software.broadinstitute.org/gatk/documentation/presentations

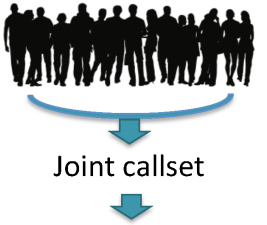
Variant discovery I

- 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

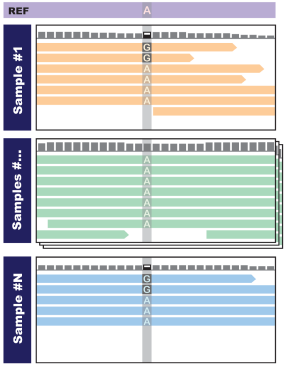
Variant discovery II





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

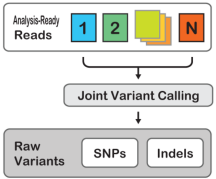
Variant discovery III



- 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**

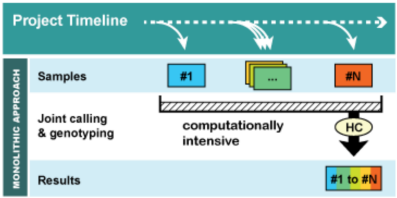
https://software.broadinstitute.org/gatk/documentation/presentations

Variant discovery - UnifiedGenotyper



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

It gives us the right answers, but…

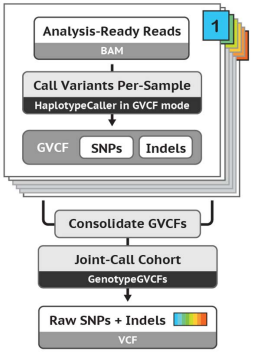


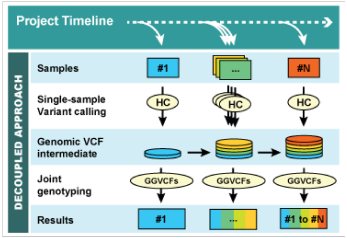
Want to add new samples?

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!

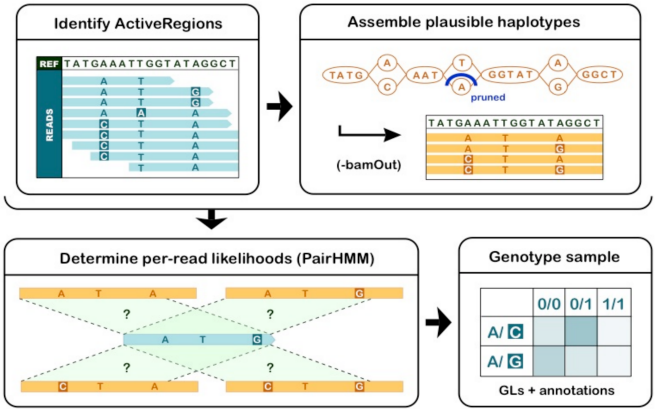
https://software.broadinstitute.org/gatk/documentation/presentations

The HaplotypeCaller Algorithm

1. **Active Region Determination:** Identifies regions that contain evidence of variation
2. **Local De Novo Assembly:** In these active regions, it re-assembles the reads to construct a set of candidate haplotypes
3. **Pair** HMM **Scoring:** Uses a Pair Hidden Markov Model to score the likelihood of the observed reads given each candidate haplotype
4. **Genotype Likelihoods:** Calculates the likelihoods for all possible genotypes (e.g. Ref/Ref, Ref/Alt, Alt/Alt) based on the HMM scores

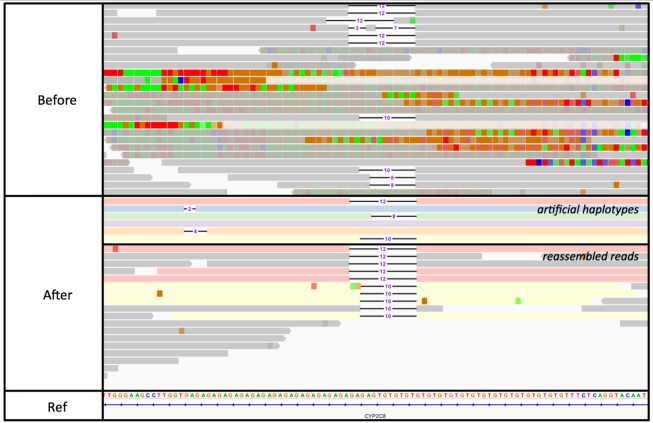
Output: GVCF (Genomic VCF) file for single-sample calling, storing likelihoods for all sites, not just variants

HaplotypeCaller I



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

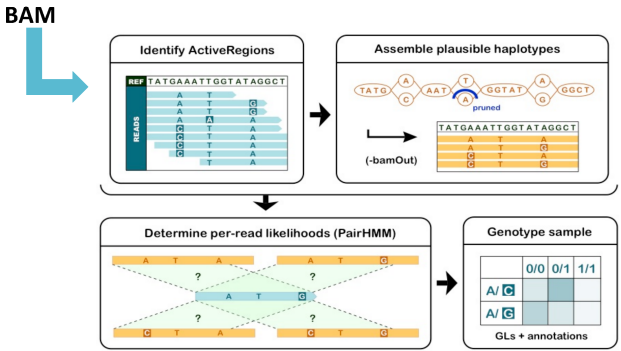
HaplotypeCaller II



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

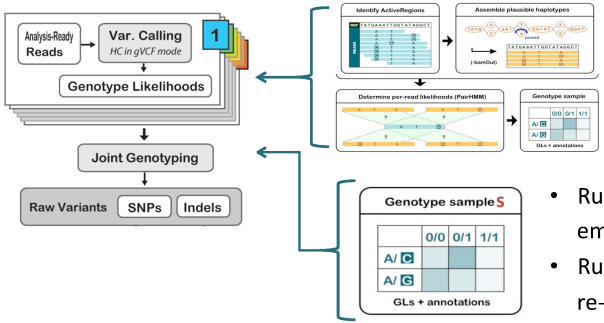
HaplotypeCaller III



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

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

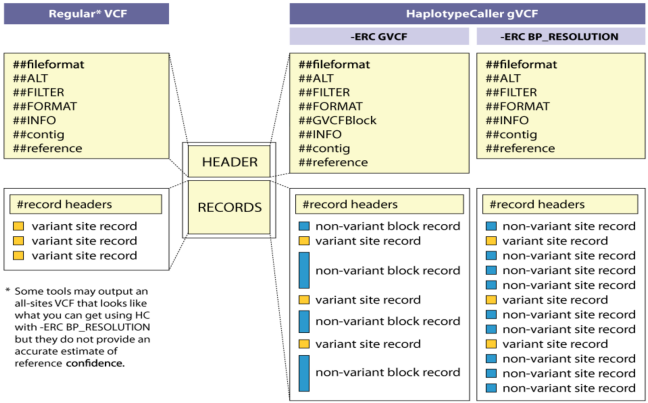
HaplotypeCaller IV



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

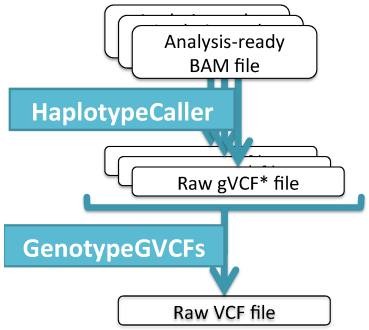
https://software.broadinstitute.org/gatk/documentation/presentations

HaplotypeCaller V



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

HaplotypeCaller VI



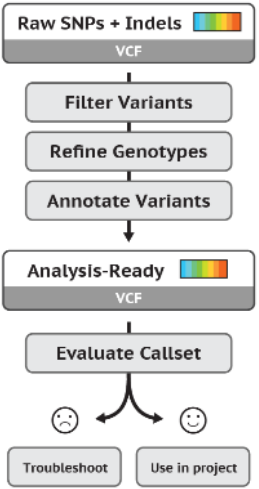
https://software.broadinstitute.org/gatk/documentation/presentations

```
gatk HaplotypeCaller \
    -R reference.fasta \
    -I sample.bam \
    -O sample.g.vcf.gz \
    -ERC GVCF

gatk CombineGVCFs \
    -R reference.fasta \
    -V sample1.g.vcf.gz \
    -V sample2.g.vcf.gz \
    -O cohort.g.vcf.gz

gatk GenotypeGVCFs \
    -R reference.fasta \
    -V cohort.g.vcf.gz \
    -O cohort.vcf.gz
```

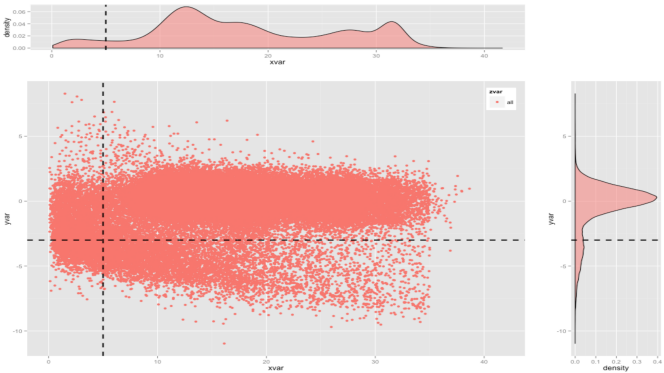
VCF Filtering



Post-Calling Filtering

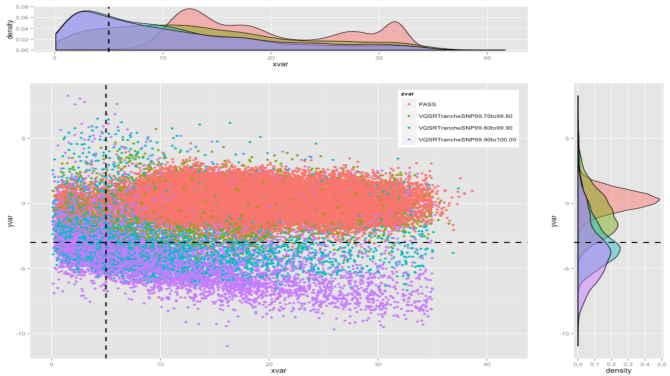
1. **Variant Quality Score Recalibration (**VQSR**):** (Preferred Method)
   - Uses a machine learning approach (Gaussian Mixture Model) to build a probability model of what a true variant looks like
   - Clusters variants based on annotation features (QD, MQ, FS, etc.) with respect to known, validated variants (Truth Sets)
   - Assigns a Tranche Sensitivity score to each variant
2. **Hard-Filtering:** (Alternative)
   - Applying fixed, empirical thresholds on quality metrics (e.g. QUAL > 30)

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

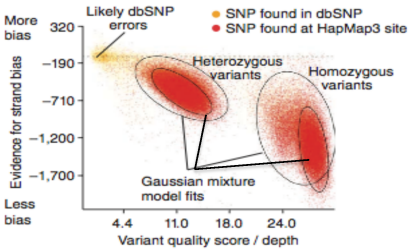
Train on high-confidence known sites to determine the probability that other sites are true or false

- 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**

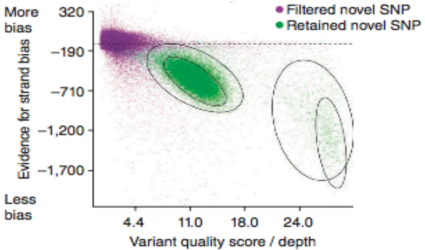
https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration III

Model trained on HapMap



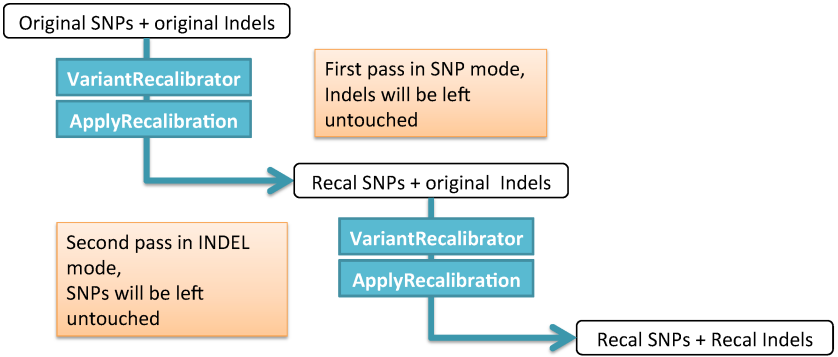
Model applied to new SNPs



Modified from DePristo et al. Nature Genetics. 2011

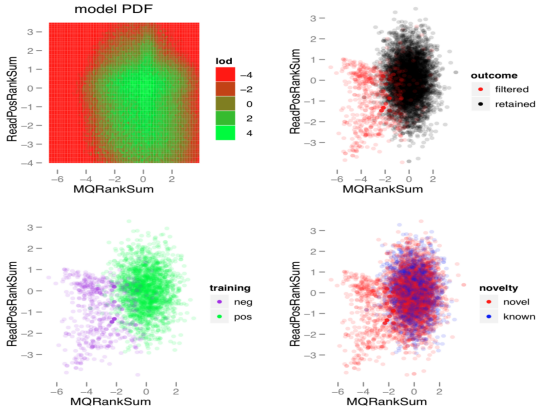
https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration IV



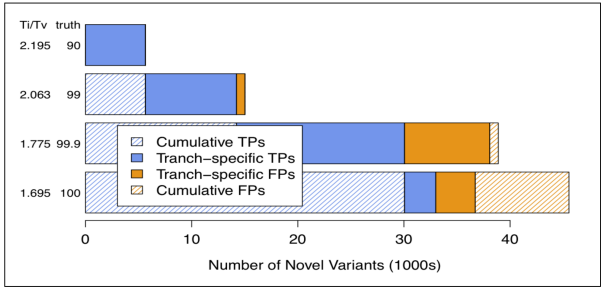
https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration V



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

VCF Filtering - Variant recalibration VI

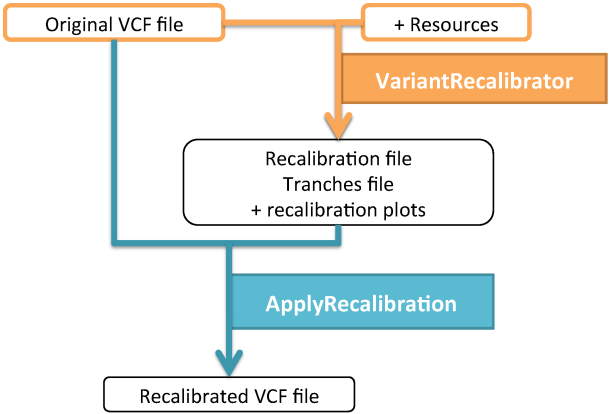


Estimation is based on Ti/Tv ratio of novel variants

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

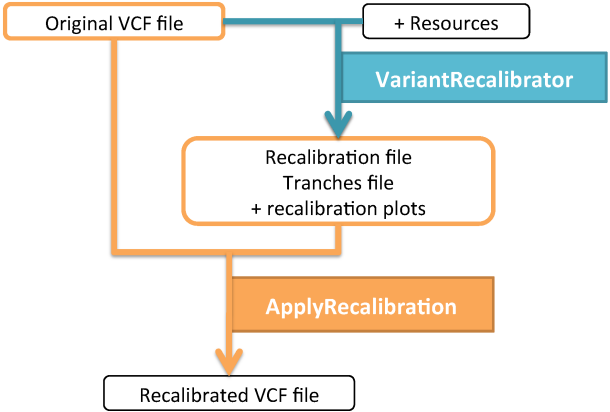
https://software.broadinstitute.org/gatk/documentation/presentations

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

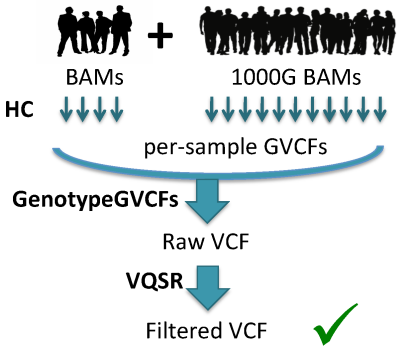
- Hard filtered vcf:

#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;...

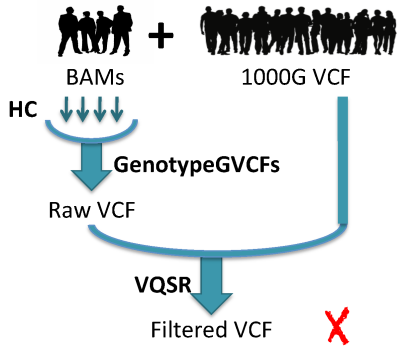
https://software.broadinstitute.org/gatk/documentation/presentations

VCF Filtering - Variant recalibration XI

ALWAYS do this:

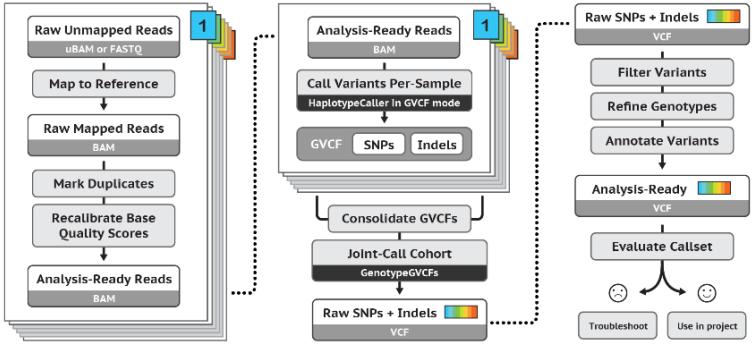


NEVER do this :



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

Presented GATK pipeline



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

Variant Annotation I

- **The Final Step:** Once high-confidence variants are called and filtered, they must be annotated to predict their functional consequences
- 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
- **Consequences Identified:**
  - Missense (changes amino acid)
  - Synonymous (no change)
  - Nonsense (introduces a stop codon)
  - Splice site variant

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

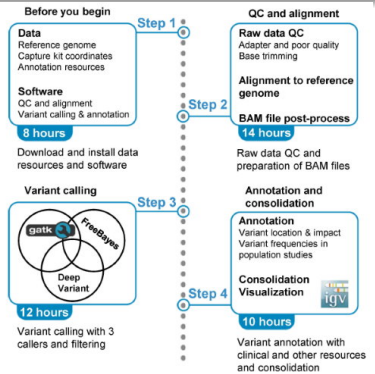
Variant Effect Predictor - VEP

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 ?