NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

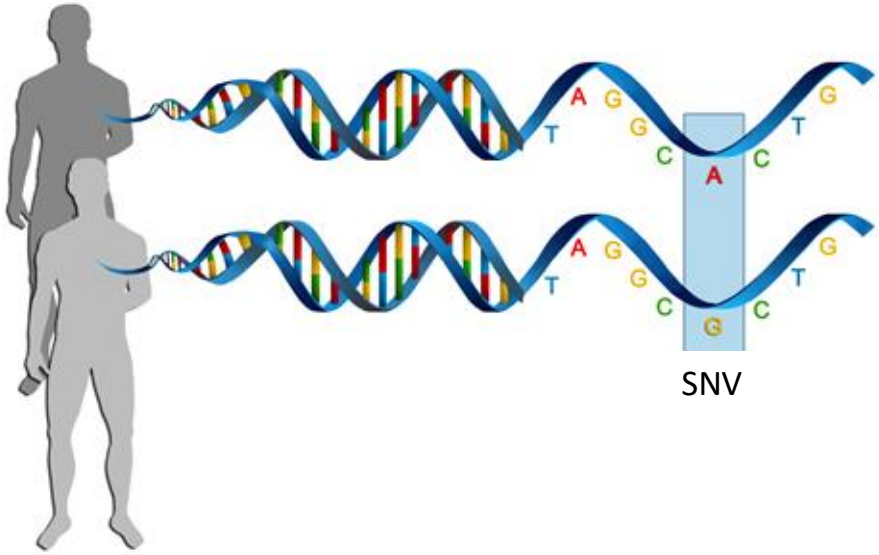
# Variant calling

# Genetic variation

**Genetic variation**
differences in DNA among individuals of the same species:

- Single nucleotide variants (SNVs)
- Small insertions and deletions (indels)
- Structural variants (SVs)
- Copy Number Variants (CNVs)

**Variant calling**
Identify genetic variations compared with a reference sequence in NGS data



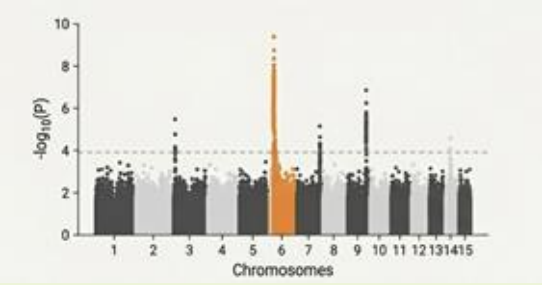
SNV

# Variant calling applications

Finding variants allows us to connect genetics to biology,

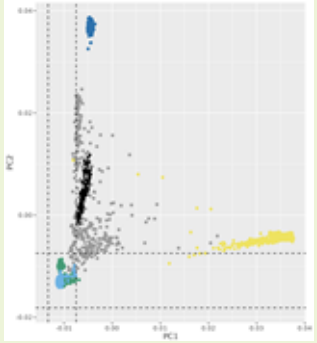
from diagnosing disease to understanding population history

**Disease & Trait association**
Identify variants linked to mendelian or complex diseases, cancer (somatic variants), or favorable traits (e.g. in agriculture)




GWAS/manhattan plot

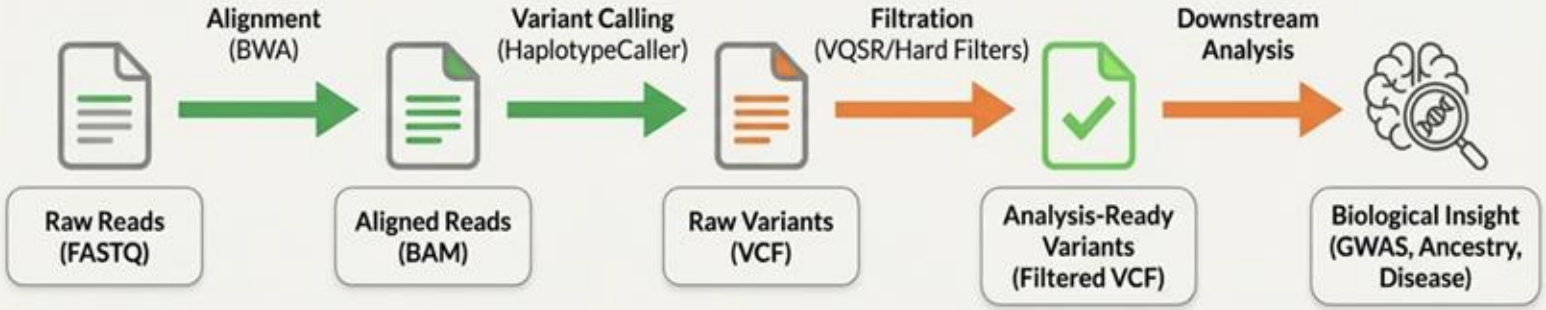
**Evolution & Population Genomics**
Uncover patterns of ancestry, migration and biodiversity



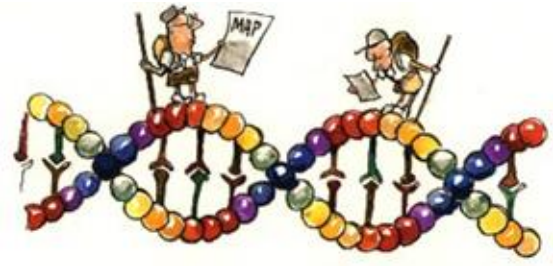
PCA/Ancestry

# Variant calling workflow

The variant calling workflow transforms millions of short DNA reads into a list of genetic variants



# The reference genome




A reference genome is a haploid sequence that represents the genome of a species

- Necessary to have a common reference so that loci and annotations can be compared between studies
- Human genome created as a consensus of a number of individuals and has evolved from GRCh37 (hg19) with 250 gaps, to GRCh38 (hg38) and the “gapless” T2T (2022)
- Choosing a reference version depends on available annotations and a need to be compatible with earlier studies, so the latest version is not always the best choice.

Important to **keep track of which *version*** of the reference genome your data was mapped to - the same version must be used in all downstream analyses

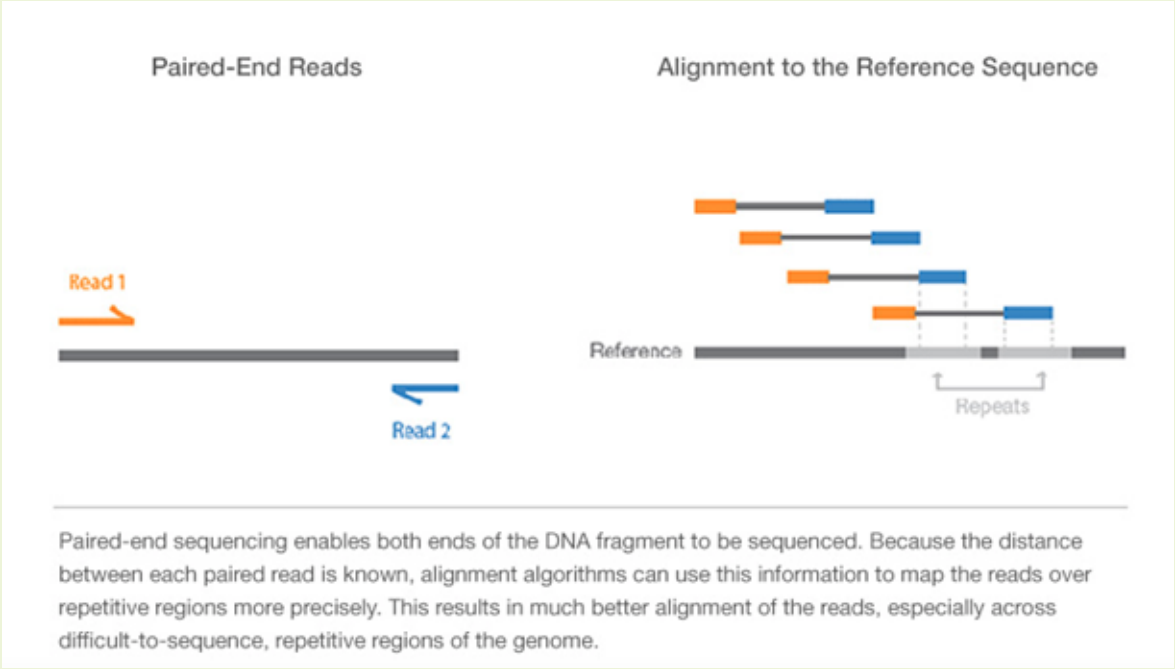
# Read alignment to reference

Alignment (or mapping) is the process of finding the best matching location of the sequencing read in the reference genome


**Alignment tools:** Mapping millions of reads requires specialized software, e.g. bwa mem, bowtie2 (short reads) or minimap2 (long reads)

**Uniquely mapped** reads align to one (best) location in the genome
**Multi-mapping read**s have many best matches (e.g. repeats)

# Paired-end reads




https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html

# Read Depth (coverage)

Mapping millions of reads results in many reads covering the same position


The number of reads that align to (cover) a given position in the reference is called Coverage or Depth

A mean coverage of 30 (30X) is common in variant calling

NotebookLM

# Mapping output - SAM format

HEADER SECTION

```
@HD     VN:1.6          SO:coordinate
@SQ     SN:2            LN:243199373
@PG     ID:bwa          PN:bwa          VN:0.7.17-r1188   CL:bwa mem -t 1 human_g1k_v37_chr2.fasta HG00097_1.fq  HG00097_2.fq
@PG     ID:samtools     PN:samtools     PP:bwa              VN:1.10           CL:samtools sort
@PG     ID:samtools.1   PN:samtools     PP:samtools         VN:1.10           CL:samtools view -H HG00097.bam
```

ALIGNMENT SECTION

```
Read_001  99   2  3843448   0 101M   = 3843625   278   TTTGGTTCCATATGAAC   0F<BFB<FFFBFBFFFB
Read_001  147  2  3843625   0 101M   = 3843448   -278  TTATTTCATTGAGCAG    FBBI7IIFIB<BBBB<B
Read_002  163  2  4210055   0 101M   = 4210377   425   TGGTACCAAAACAGAGA   0IIFBFFFIIIFFIFFF
```

Read name

Reference sequence name (chromosome/contig)

Start position

Sequence

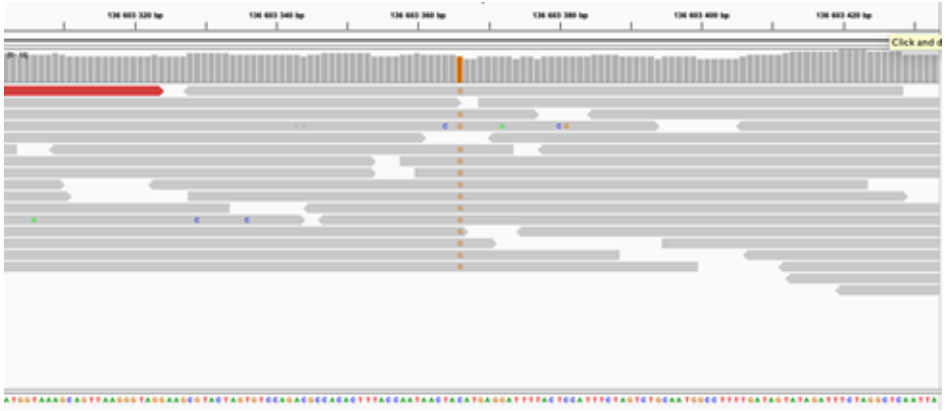
Quality

SAM: Sequence Alignment/Map format

BAM: binary SAM – used by most (all?) variant callers and many other tools

# Detecting variants

```
Reference:    GTAGACTGCTAGATCGA
Sample reads: TAGACTGATAGA
               AGACTGATAGATC
              TAGACTGCTAG
```



**Reference allele** (here C)
**Alternative allele** (here A): allele that differs from reference

A variant calling tool scans all reads for differences between the sequencing reads and the reference

**Variant callers**
**Germline callers** (for inherited variants)**:** HaplotypeCaller, FreeBayes, Bcftools, Deepvariant
**Somatic callers** (for acquired variants, e.g. in tumors): Mutect2, Strelka2

# Variant Call Format (VCF)

Standard file format for storing genetic variants –
output from most variant callers and input to many downstream analysis tools

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##contig=<ID=2,length=243199373>
##source=HaplotypeCaller

#CHROM  POS        ID      REF   ALT   QUAL  FILTER  INFO              FORMAT      HG00097
2       136220992  rs111   G     T     30    PASS    AC=1;AF=1.0;AN=2  GT:AD:DP    1/1:3,2:5
2       136226814  rs222   GAC   G     44    .       AC=1;AF=0.5;AN=2  GT:AD:DP    0/1:4,2:6
2       136234279  .       C     T     102   .       AC=1;AF=0.5;AN=2  GT:AD:DP    0/1:3,4:7
```


**Header section** - Metadata defining the file content, including definitions of the INFO and FORMAT fields
**Records** (one line per variant) - Each line represents one genetic variant

# Multi-sample VCF

**Multi-sample vcf** files have one column per sample - necessary for downstream comparisons

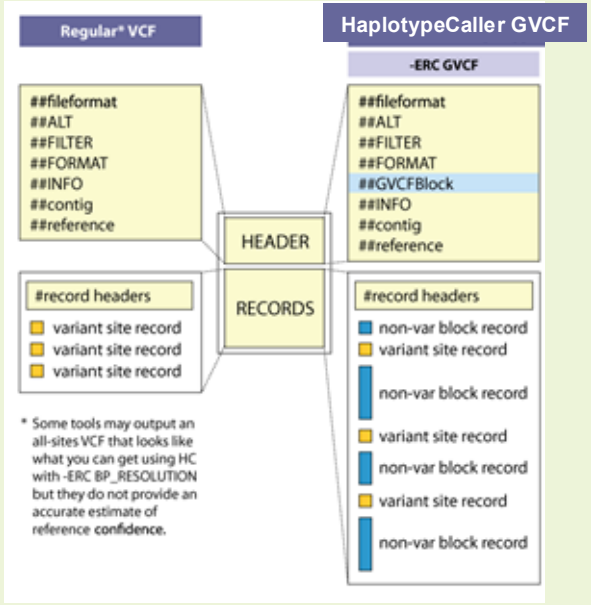
```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##contig=<ID=2,length=243199373>
##source=CombineGVCFs
##source=GenotypeGVCFs
##source=HaplotypeCaller
```

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | SAMPLE1 | SAMPLE2 | aSAMPLE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 136045826 | . | G | A | 167 | . | AC=1;AF=0.167;AN=6 | GT:AD:DP | 0/0:8,0:8 | 0/0:13,0:13 | 0/1:1,5:6 |
| 2 | 136046443 | . | CG | C | 129 | . | AC=3;AF=0.500;AN=6 | GT:AD:DP | 0/0:8,0:8 | 0/1:3,1:4 | 1/1:0,4:4 |
| 2 | 136047387 | . | T | C | 186 | . | AC=1;AF=0.167;AN=6 | GT:AD:DP | 0/0:6,0:6 | ./.:.,.:. | 0/1:4,6:10 |

**Limitation:** Individual vcf files have information only for variant sites - not possible to distinguish non-variant (reference) positions and positions with missing information

**Solution:** Creating gVCF files and do joint variant calling in all samples (cohort) simultaneously

# Genomic VCF (gVCF) Files




- VCF files have information only about variant sites
- gVCF files have records for *all* sites
- Adjacent homozygous-reference sites are merged into blocks
- The gVCF files can be used to generate a multi-sample VCF through *joint genotyping* without loss of information

Some tools can call variants from all samples simultaneously (outputting multi-sample VCFs directly)
Creating gVCF files first saves computation time in large cohorts or if samples are added at a later time

# Variant filtration

Variant calling with HaplotypeCaller is designed to be sensitive –

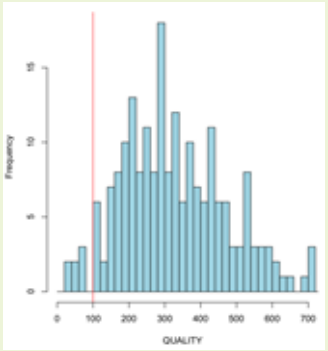
Apply filters to limit false positives

**Advanced filtration (VQSR)**

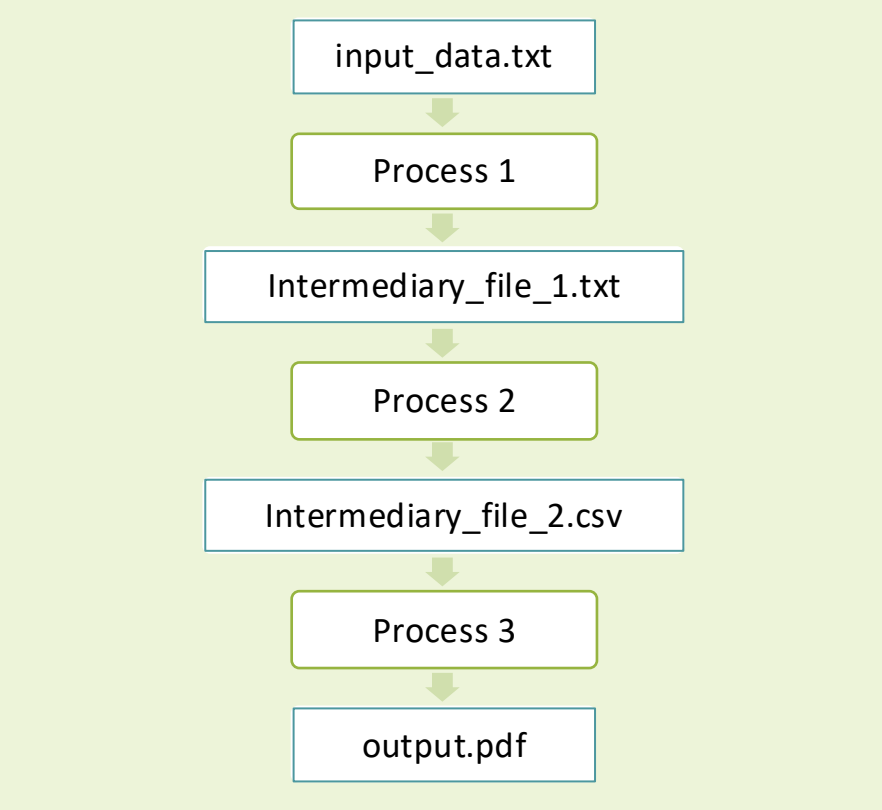
Machine Learning approach
Need well-curated list of known variants
Requires large datasets

**Hard filters**

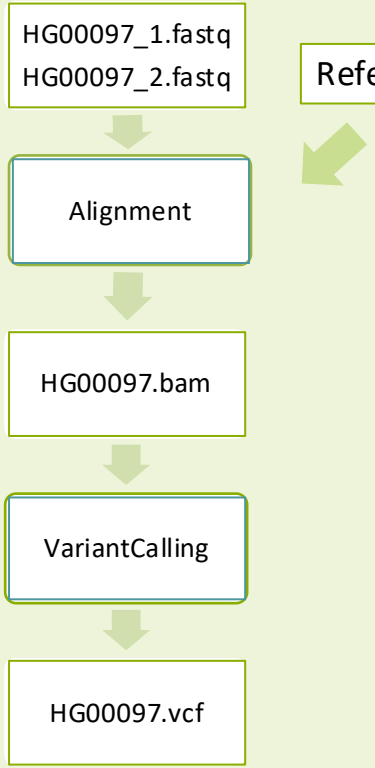
Select cutoffs, e.g. quality and depth
Cutoffs selected after viewing score distributions
Works for small datasets and non-model organisms



# Workflows

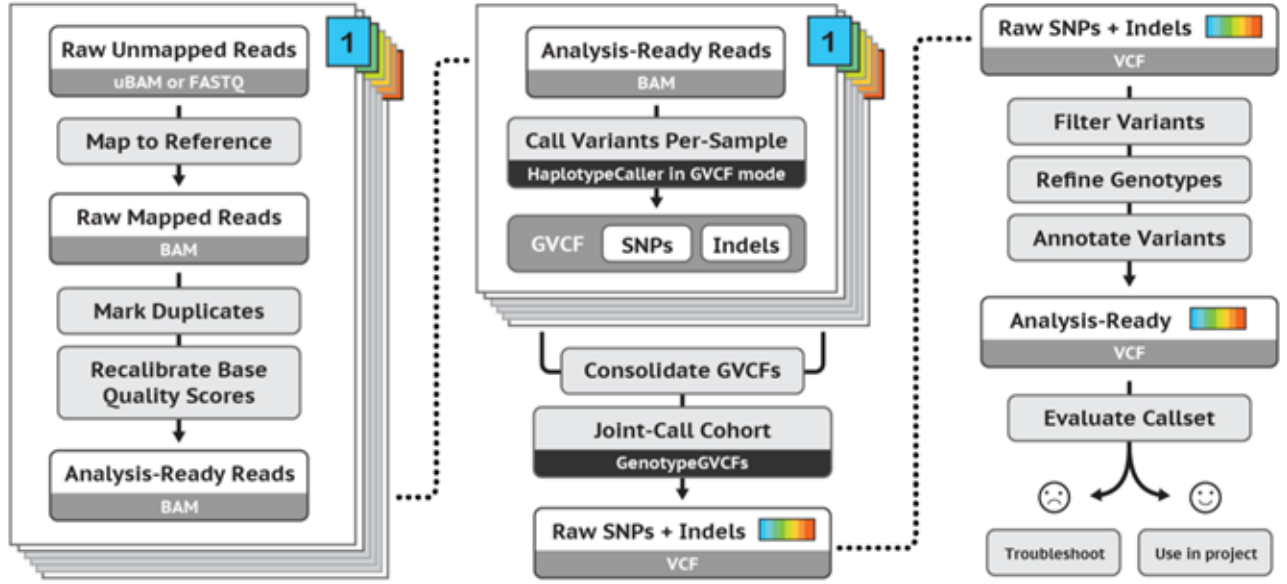

input_data.txt
Process 1
Intermediary_file_1.txt
Process 2
Intermediary_file_2.csv
Process 3
output.pdf

# Example: Basic workflow, one sample




- Create a new output file in each process
  - Do not overwrite the input file
- Use informative file names
  - Include information about the process + sample
- Correct name extension (e.g. .bam, .vcf)

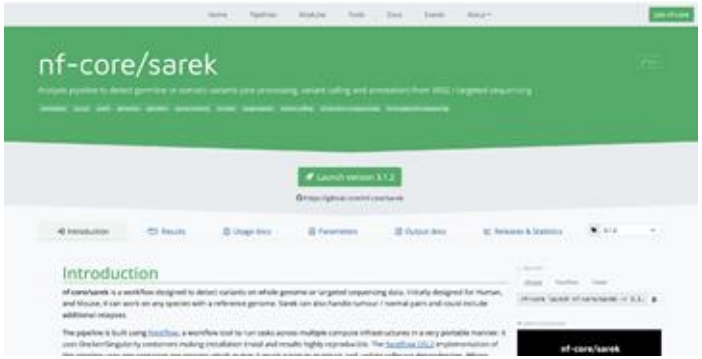
# Refined workflow: GATK best practices workflow for germline short variant discovery



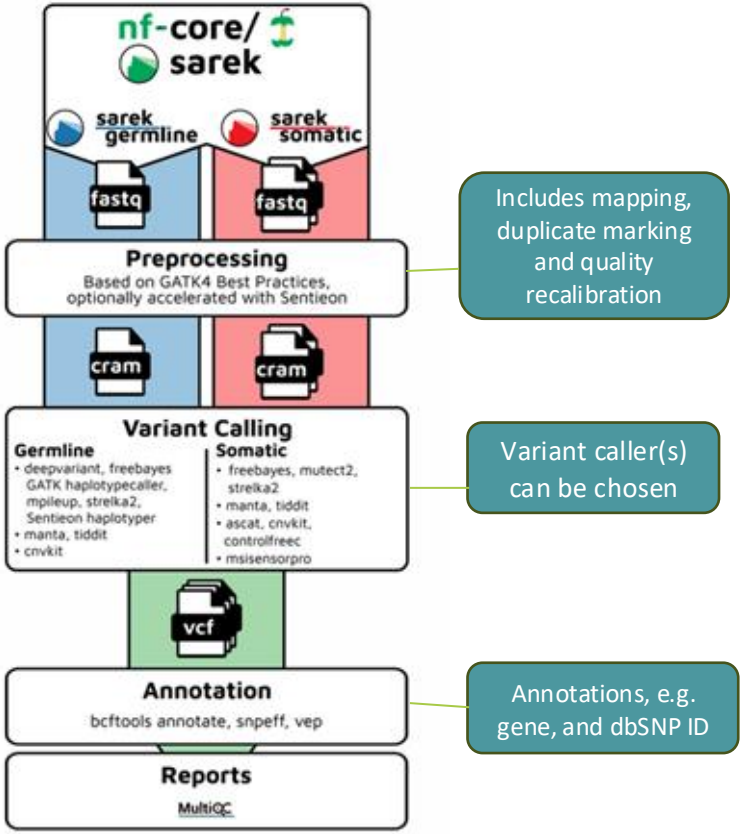

https://software.broadinstitute.org/gatk/best-practices/

# Nf-core variant calling workflow: Sarek

All steps in the GATK best-practises (germline variant calling) have been combined in the commonly used pipeline nf-core/sarek



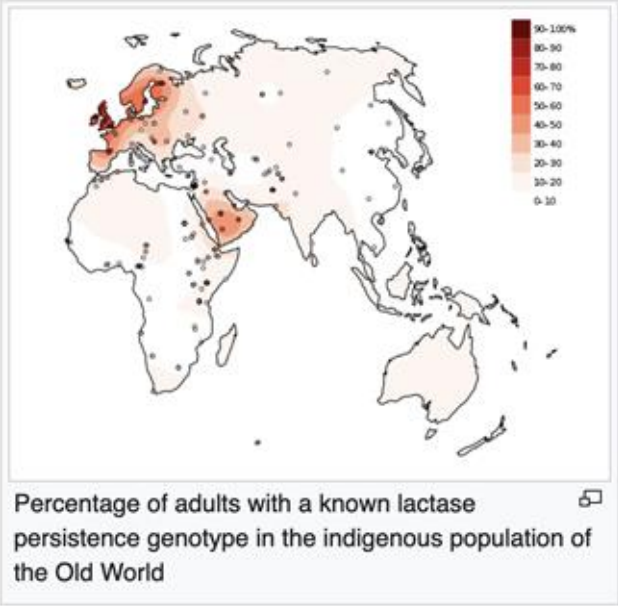

https://nf-co.re/sarek/3.4.2/



# The lab: Lactose tolerance

Figure 2. Lactose digestion in the intestine.

- All mammals produce lactase as infants, but some humans produce lactase also in adulthood
- A genetic variant upstream of the *LCT* gene leads to the lactase persistent phenotype (lactose tolerance)



Percentage of adults with a known lactase persistence genotype in the indigenous population of the Old World

# Data

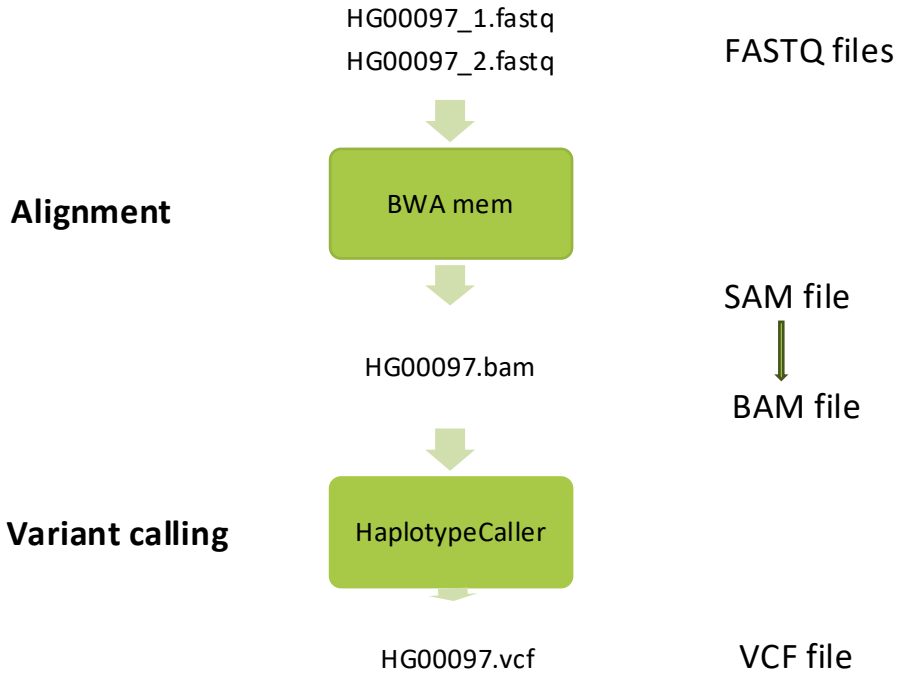
- 3 samples (from 1000 genomes)
- Low coverage WGS data
- Small region on chromosome 2

About the samples:
https://www.internationalgenome.org/data-portal/sample

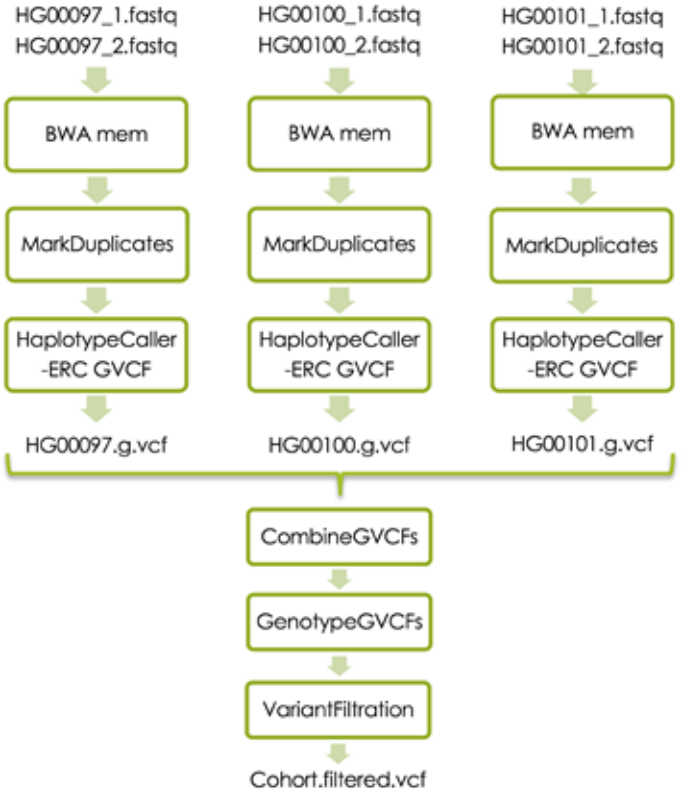
# Overview of lab

- Part1: Basic variant calling for one sample
- Part2: Variant calling in cohort (multiple samples)
- Part3: Write a bash script
- Extra material (part4): GATK Best practices

# Part1: Basic Variant Calling in one sample



# Part2: Basic variant calling in cohort

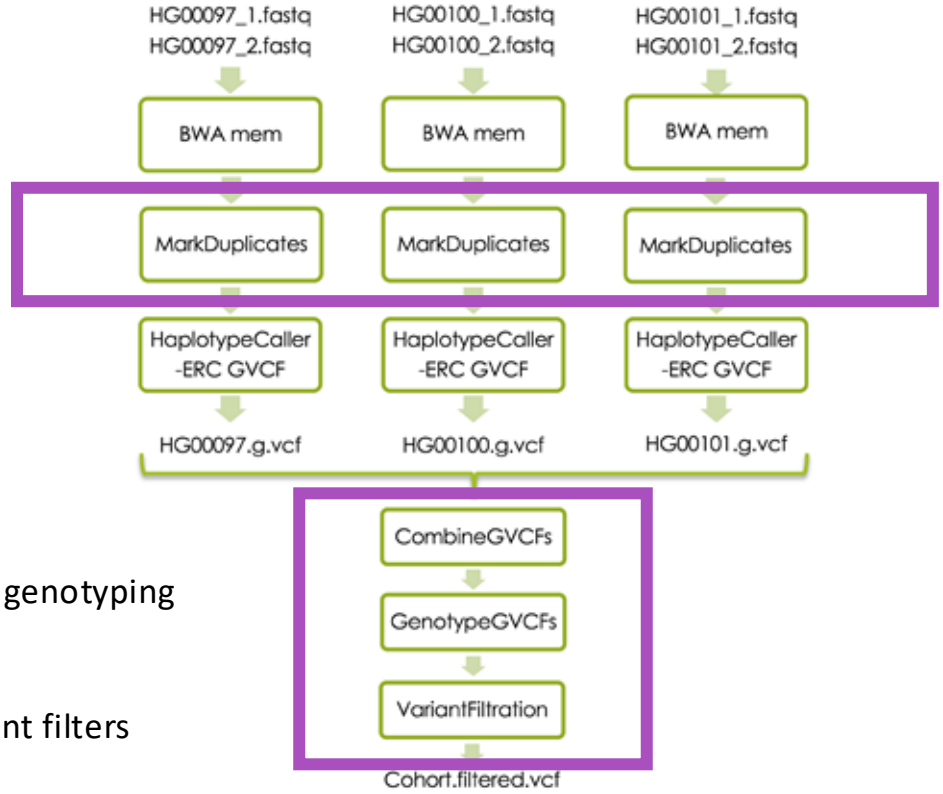


# Basic variant calling in cohort

Mark duplicates

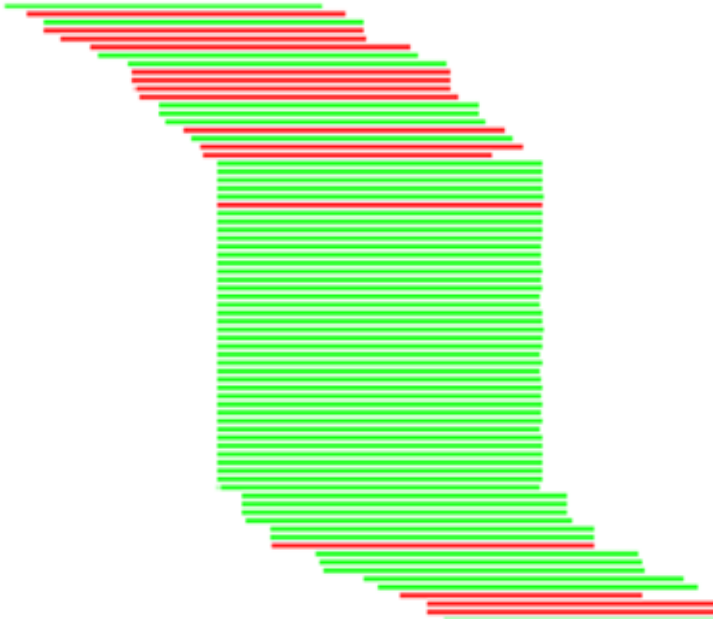
Joint genotyping

Variant filters



# Duplicate reads

- PCR duplicates - library preparation
- Optical duplicates - sequencing
- Can give false allelic ratios of variants
- Should often be removed/marked
  - Picard MarkDuplicates
  - Samtools dedup



# File indices

- Most large files that we work with need indices - Allows efficient access
- Different indices for different file types
- The index is stored in a file with specific ending
- Most tools will find the index file automatically as long as the ending is correct

File endings
Reference genome (**.fasta.fai/fasta.dict**)
Aligned reads (**.bam.bai**)
Variants **(.vcf.idx**)

# ReadGroups

Tags that mark which *sample* and *sequencing run* each read comes from

**REQUIRED TAGS**

RGID: unique identifier of sequencing run
RGLB: library id, used to get correct duplicate marking
RGPL: platform (here Illumina)
RGPU: platform unit
**RGSM:** sample name used in the vcf file

```
RGID=4 \
RGLB=lib1 \
RGPL=ILLUMINA \
RGPU=unit1 \
RGSM=20
```

# Part3: Bash script for variant calling

```
#!/bin/bash
#SBATCH -A naiss2025-xx-xxx
#SBATCH -p shared
#SBATCH -c 8
#SBATCH -t 1:00:00
#SBATCH -J JointVariantCalling

module load bioinfo-tools
module load bwa/0.7.17
module load samtools/1.20
module load gatk/4.5.0.0

## loop through the samples:
for sample in HG00097 HG00100 HG00101;
do
  echo "Now analyzing: "${sample}
  #Fill in the code for running bwa-mem for each sample here
  #Fill in the code for samtools index for each sample here
  #Fill in the code for MarkDuplicates here
  #Fill in the code for HaplotypeCaller for each sample here
done
#Fill in the code for CombineGVCFs for all samples here
#Fill in the code for GenotypeGVCFs here
```

# General advice

**Copying** commands to terminal can introduce errors

- Type commands in terminal or
- Copy and edit in text editor

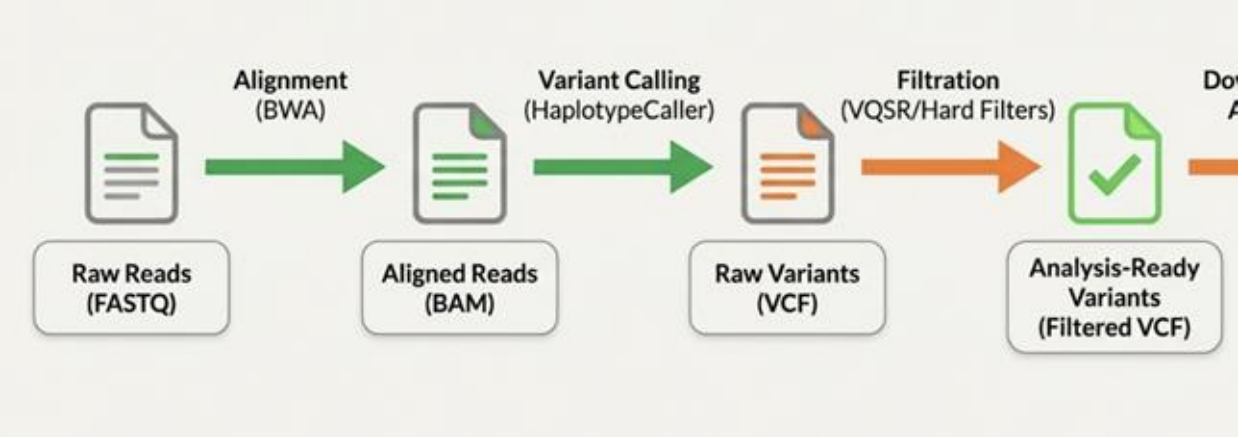
**Backslash “\”**

- the command continues on the next line, but should not be included when the command is copied

**Pipe “|”**

- Bwa mem ref.fa fastq.fq | samtools sort > output.bam

# Start practising!

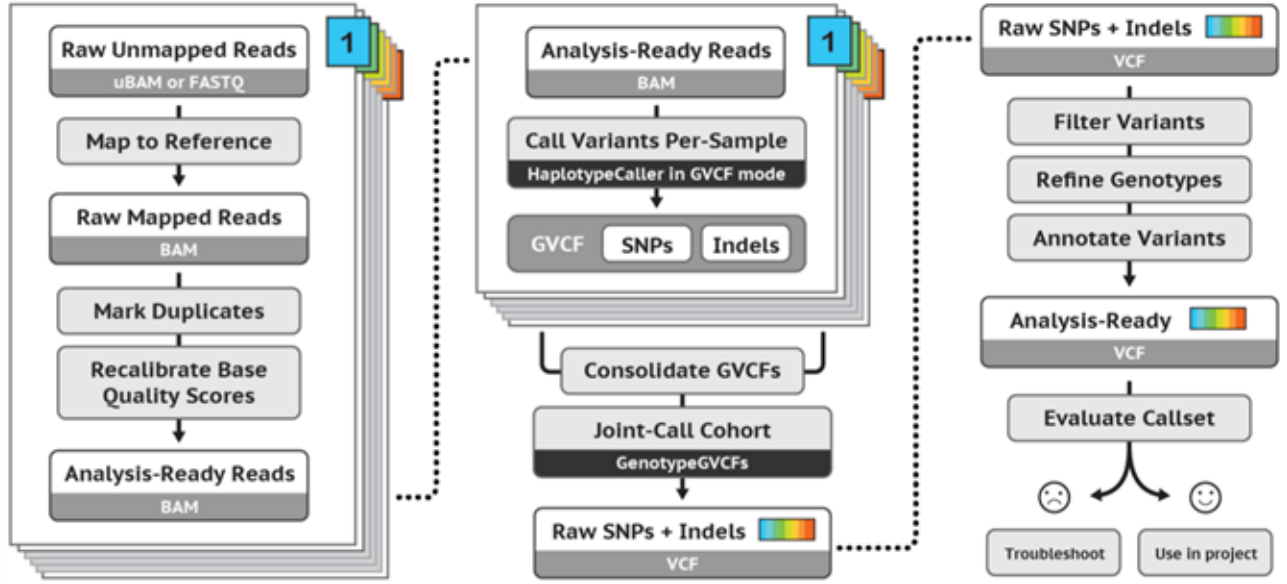


# Extra lab (Part4): GATK's best practises

# GATK best practices for short variant discovery

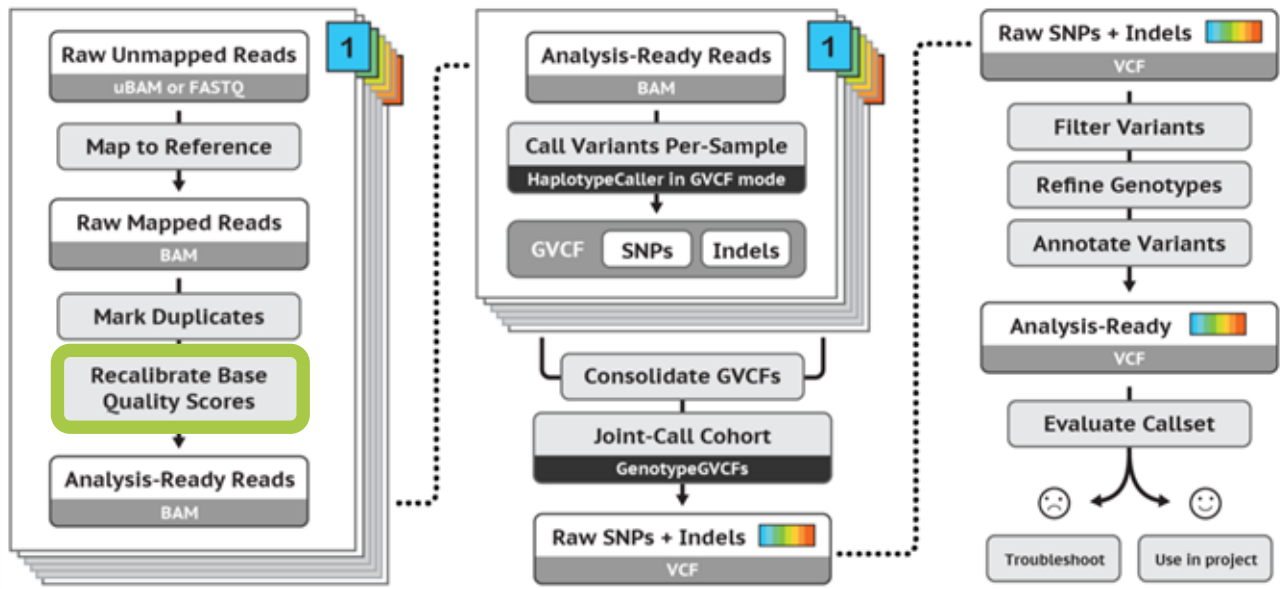

https://gatk.broadinstitute.org

# GATK best practices workflow for germline short variant discovery



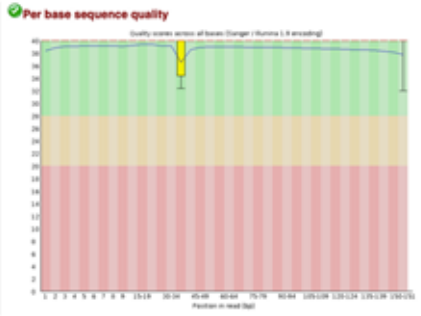

https://software.broadinstitute.org/gatk/best-practices/

# Base Quality Score Recalibration (BQSR)



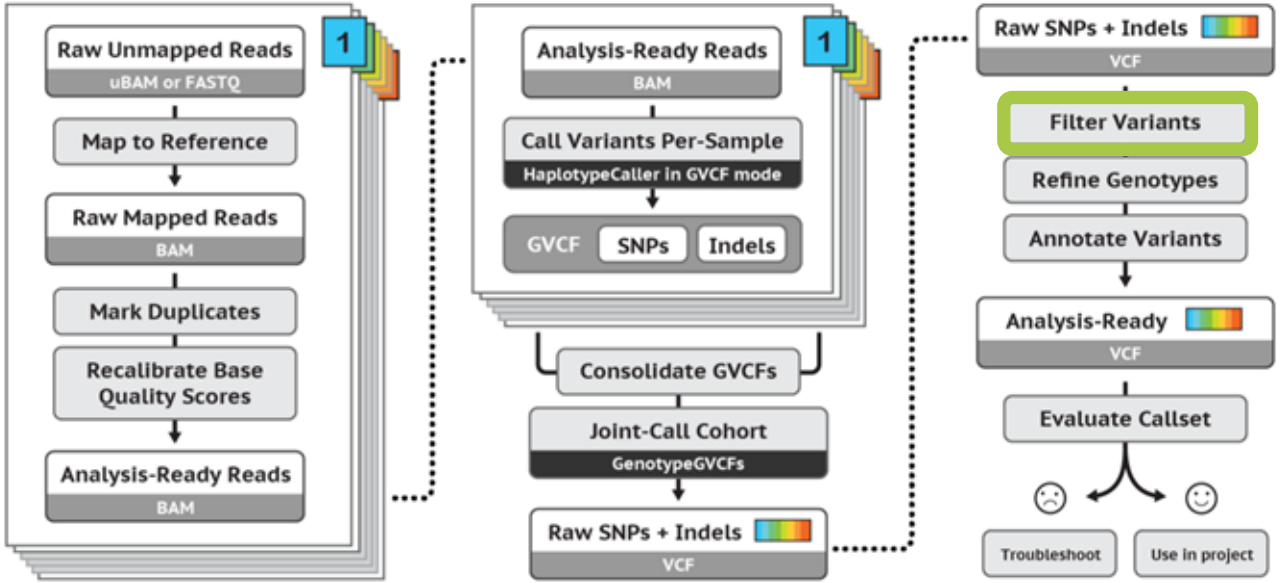
# Base Quality Score Recalibration (BQSR)

1. During base calling, the sequencer estimates a quality score for each base. These are the quality scores present in the fastq files
2. Systematic (non-random) errors in base quality score estimation can occur due to for example
   - physics or chemistry of the sequencing reaction
   - manufacturing flaws in the equipment
3. Can cause biases in variant calling
4. **Base Quality Score Recalibration** helps to calibrate the scores so that they correspond to the real per-base sequencing error rate (phred scores)




Only R1 pos 37 affected

# Filter variants




https://software.broadinstitute.org/gatk/best-practices/

Germline short variant discovery (SNPs + Indels)

# Hard Filters

- Small data set used in this lab
- Filters on information in the VCF file using set cutoffs
- SNPs and indels can have slightly different filter values
- For example: Flag variants with
  - QD < 2.0 – Quality by Depth (variant QUAL divided by the depth of non-hom-ref samples)
  - FS > 60 – Fisher Strand Bias