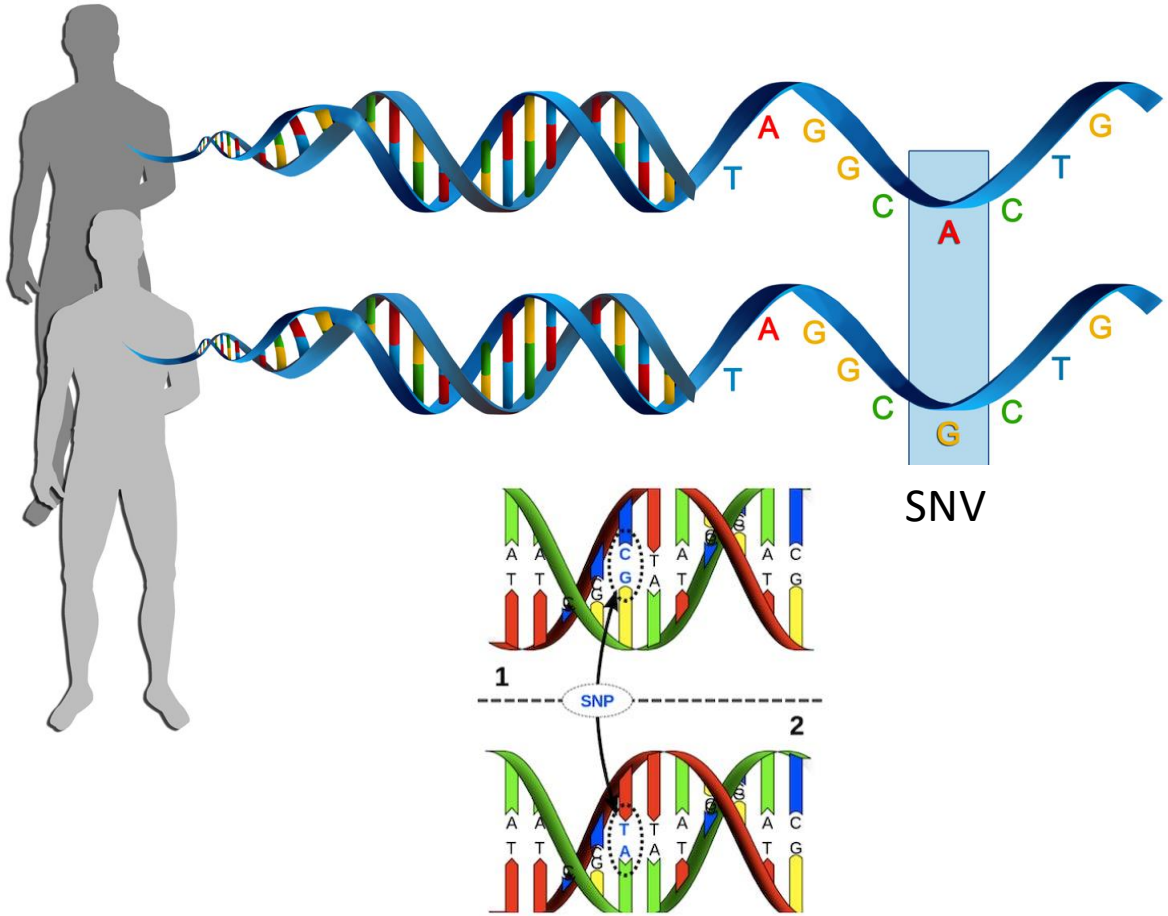
Variant calling

Genetic variation

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

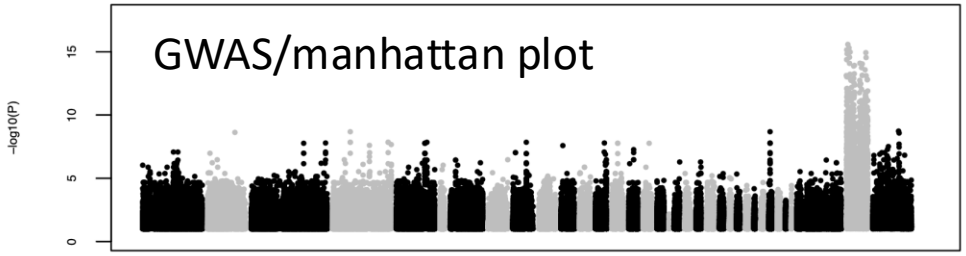
- Single nucleotide variants (SNVs)
- Small insertions and deletions (indels)
- Structural variants (SVs)
- Copy Number Variants (CNVs)

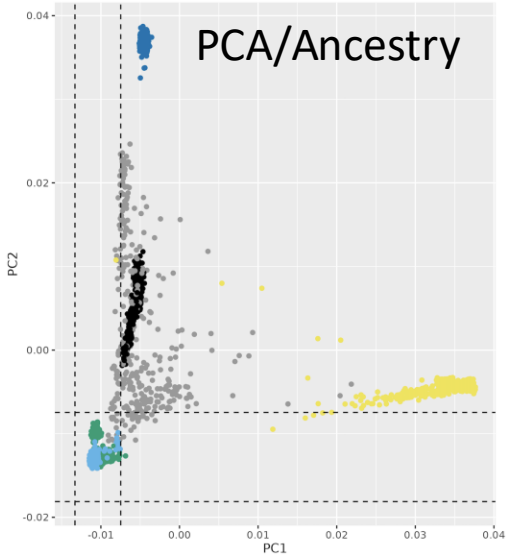
Variant calling
Identify genetic variations compared with a reference sequence in NGS data



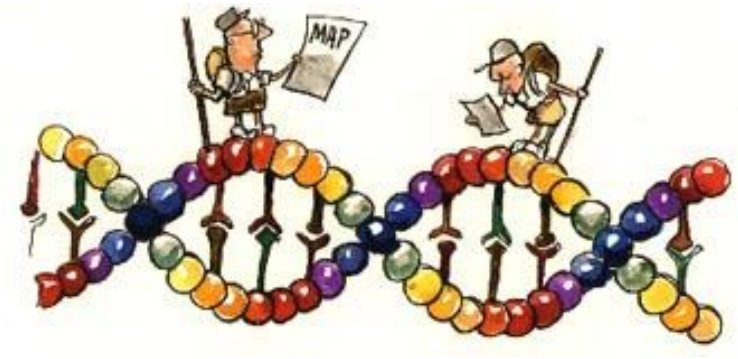
Variant calling applications

- Disease/trait associated variation, e.g.
  - Mendelian diseases/rare diseases
  - GWAS/complex diseases
  - Cancer (somatic variants)
  - Favorable traits (e.g. in plants)
- Evolution/population genomics, e.g:
  - Ancestry, migration
  - Biodiversity, studies of endangered species





The reference genome



A reference genome is a haploid nucleic acid sequence which represents a species genome

Human genome versions

- GRCh37 (hg19): 250 gaps
- GRCh38 (hg38) - 2013
- T2T (“gapless”) - 2022

The reference genome sequence is used as input in many bioinformatics applications for NGS data

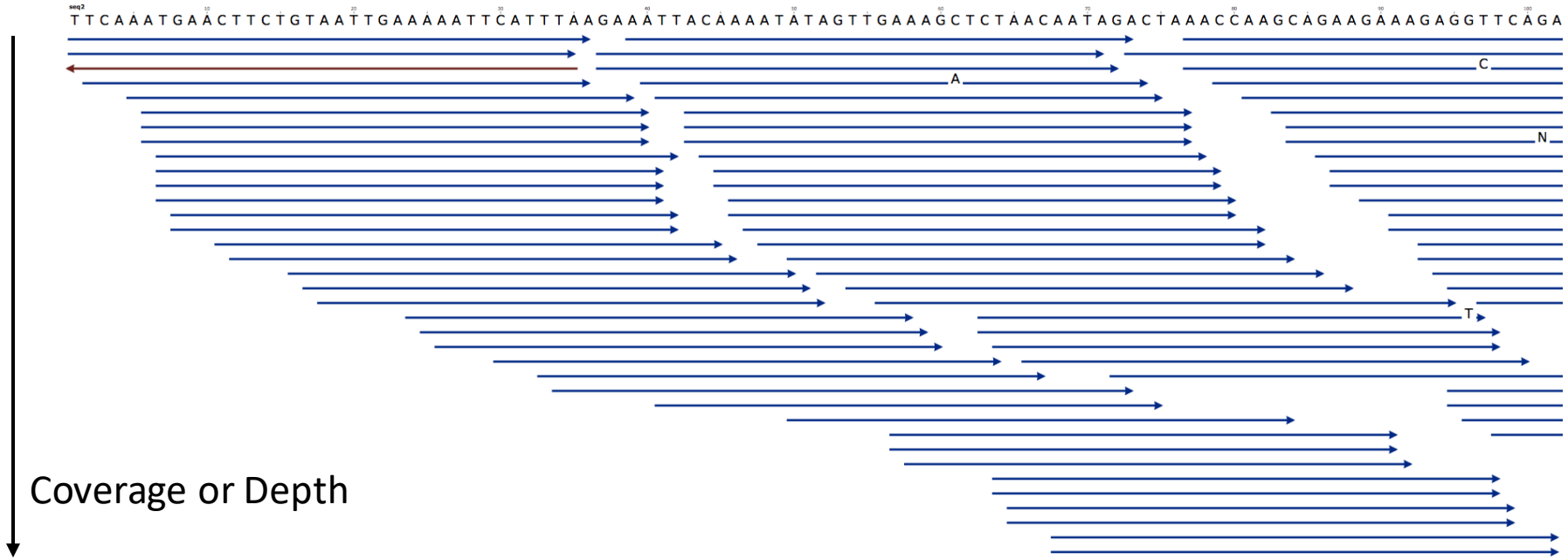
Some annotations may only be available for specific versions, so it is not always best to use the latest version. An older version may also be used to ensure compatibility with previous studies.

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

Alignment

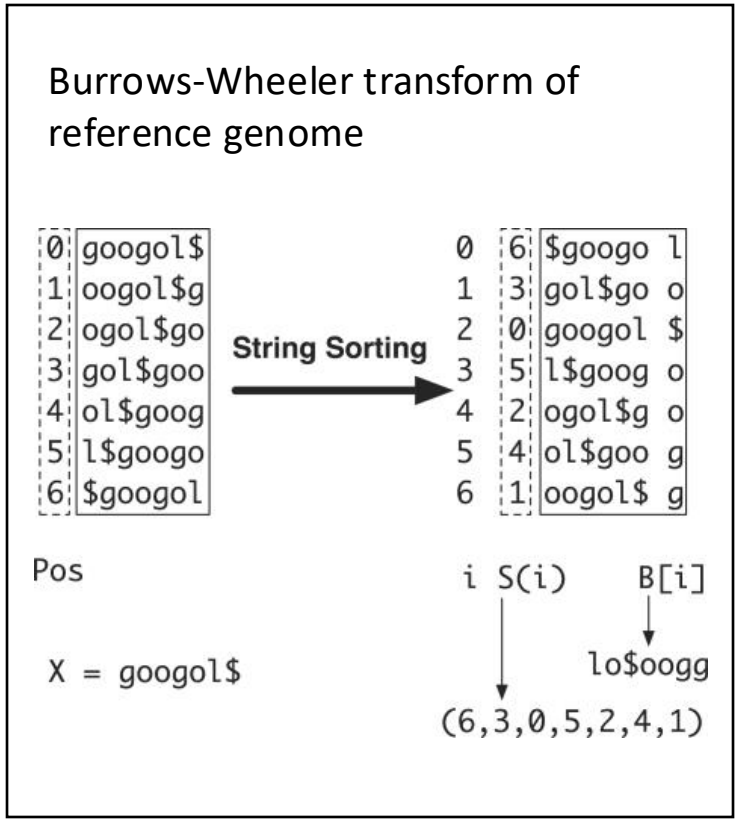
```
REFERENCE          CCCCGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTACCCTCTTCCTTAGGGACTGTAC
SEQUENCING READ                          GCTAGCTAGCTAGCTACCCT
```

Alignment


seq2
TTCAAATGAACTTCTGTAATTGAAAAATTCATTTAAGAAATTACAAAATATAGTTGAAAGCTCTAACAATAGACTAAACCAAGCAGAAGAAAGAGGTTCAGA
A
C
N
T
Coverage or Depth

Alignment tools

- Short reads (often Illumina)
  - BWA (Burrows-Wheeler Aligner)
  - Bowtie2
- Long reads (e.g. PacBio, Nanopore)
  - minimap2



http://bio-bwa.sourceforge.net

Output from mapping - Sam format

HEADER SECTION

```
@HD   VN:1.6SO:coordinate
@SQ   SN:2  LN:243199373
@PG   ID:bwaPN:bwaVN: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:bwaVN: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    TTTGGTTCCATATGAACTTT    0F<BFB<FFFBFBFFFBFBB
Read_001    147   2    3843625    0    101M    =    3843448    -278   TTATTTCATTGAGCAGTGGT    FBBI7IIFIB<BBBB<BBFF
Read_002    163   2    4210055    0    101M    =    4210377    423    TGGTACCAAAACAGAGATAT    0IIFBFFFIIIFFIFFFBBF
Read_003    99    2    4210066    0    101M    =    4210317    352    CAGAGATATAGATCAATGGA    0IIFFFIFFFIFIFIIIIIF
```

Start position

Reference sequence name

Sequence

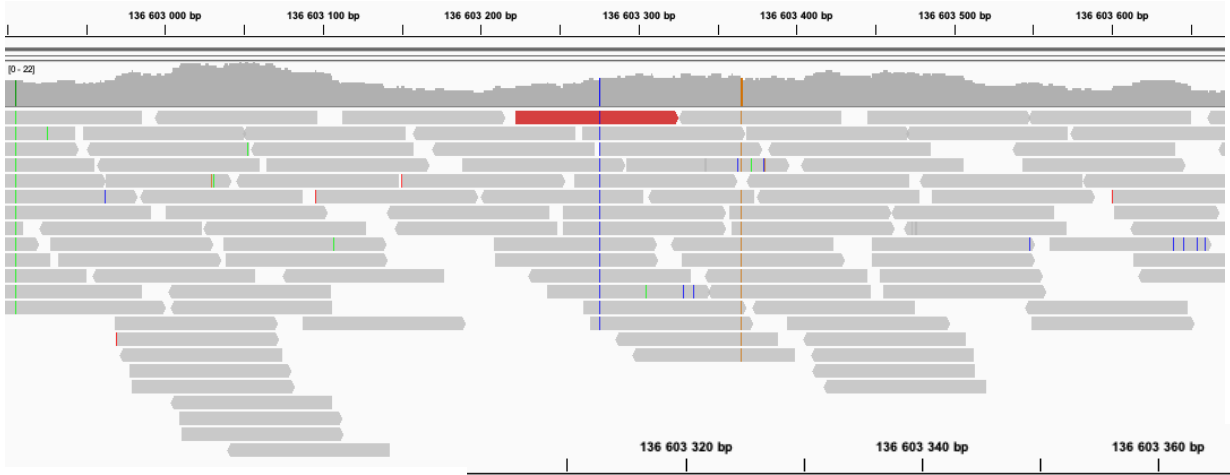
Quality

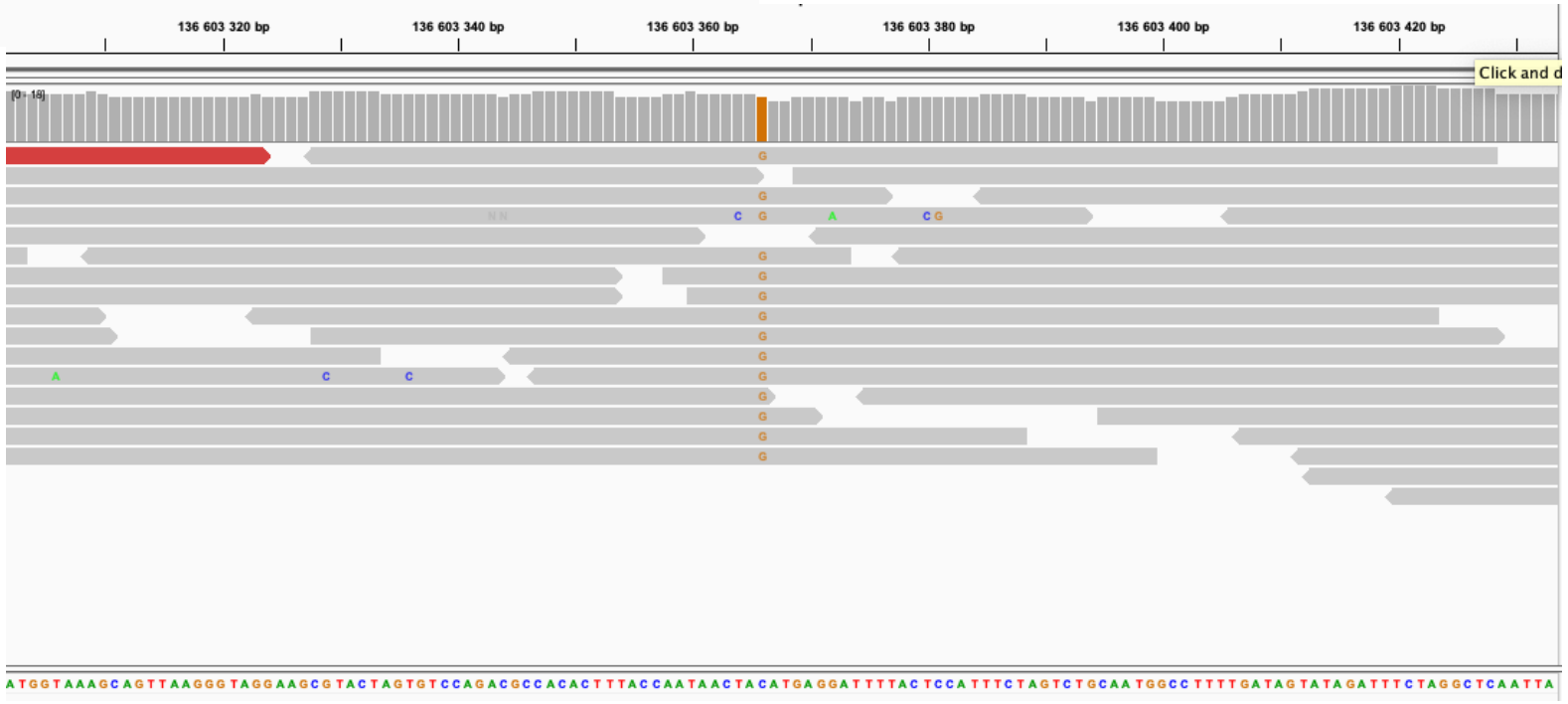
Read name
(usually more complicated)

Detecting variants

```
Reference:      ...GTGCGTAGACTGCTAGATCGAAGA...
Sample:         ...GTGCGTAGACTGATAGATCGAAGA...
                ...GTGCGTAGACTGATAGATCGAAGA...
                ...GTGCGTAGACTGCTAGATCGAAGA...
                ...GTGCGTAGACTGCTAGATCGAAGA...
                ...GTGCGTAGACTGATAGATCGAAGA...
                ...GTGCGTAGACTGATAGATCGAAGA...
                ...GTGCGTAGACTGCTAGATCGAAGA...
                ...GTGCGTAGACTGATAGATCGAAGA...
```

Variants in genome viewer





Reference- and alternative alleles

TATATCTTCCCCGCTAGCTCGCTAGCTACTTCAAAT

Reference allele AGCTCGCTA

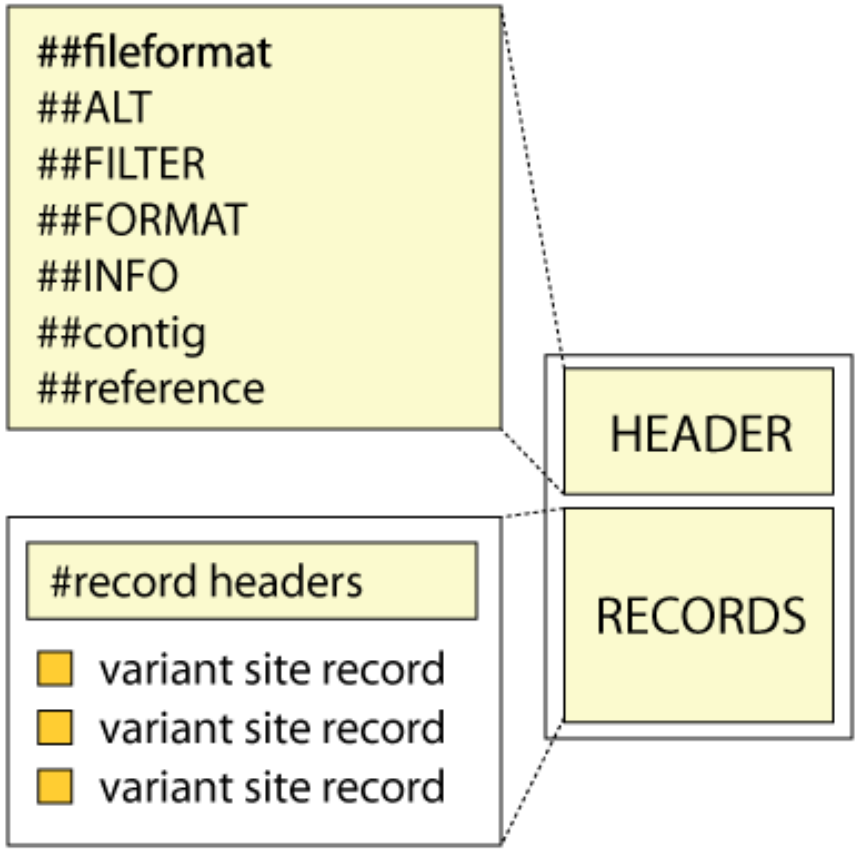
Alternative allele AGCTAGCTA

Reference allele = the allele in the refence genome
Alternative allele = the allele NOT in the refence genome

Variant Call Format (VCF)

```
##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 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=HaplotypeCaller
#CHROM  POS          ID   REF   ALT   QUAL    FILTER   INFO              FORMAT   HG00097
2       136220992    .    G     GT    30.64   .        AC=1;AF=0.500;AN=2  GT:AD:DP  0/1:3,2:5
2       136226814    .    GAC   G     44.60   .        AC=1;AF=0.500;AN=2  GT:AD:DP  0/1:4,2:6
2       136234279    .    C     T     102.60  .        AC=1;AF=0.500;AN=2  GT:AD:DP  0/1:3,4:7
2       136234284    .    C     T     102.60  .        AC=1;AF=0.500;AN=2  GT:AD:DP  0/1:3,4:7
…
```

Variant Call Format (VCF)

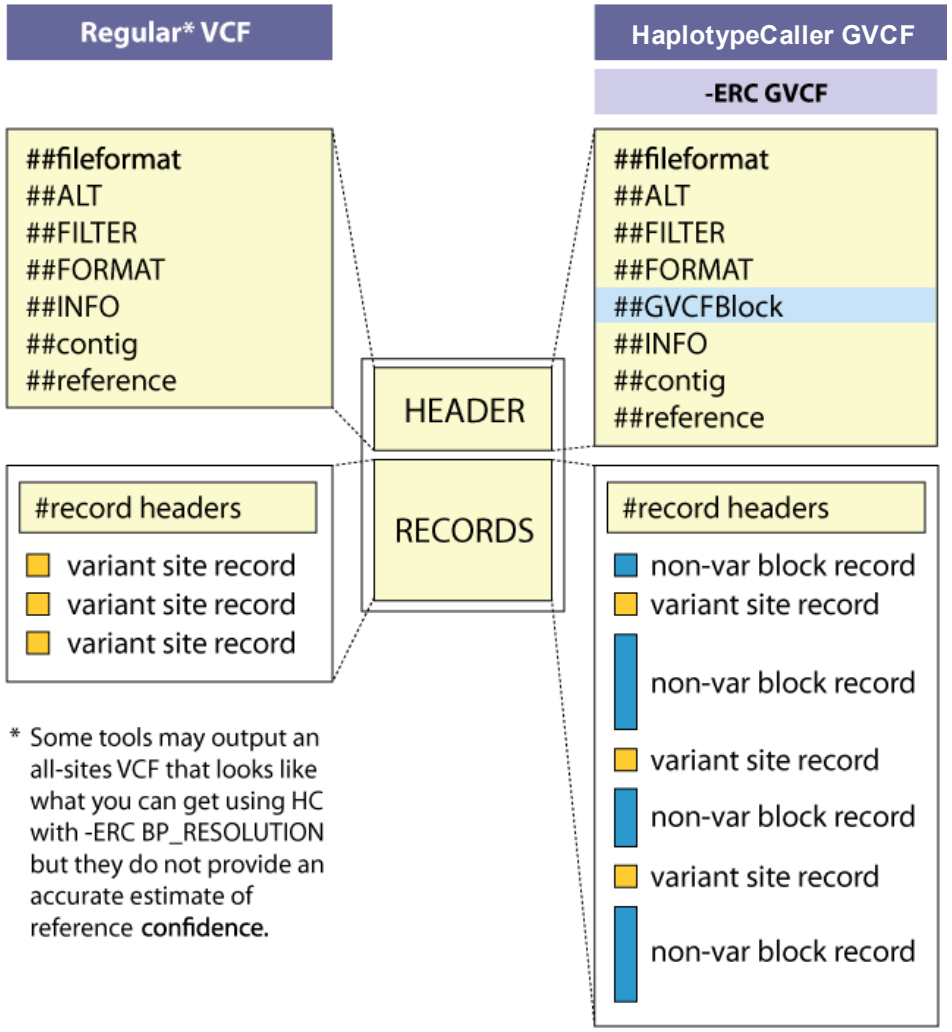



```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##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  .    G    GT     30.64 .       AC=1;AF=0.500;AN=2  GT:AD:DP 0/1:3,2:5
2       136226814  .    GAC  G      44.60 .       AC=1;AF=0.500;AN=2  GT:AD:DP 0/1:4,2:6
2       136234279  .    C    T      102.60 .      AC=1;AF=0.500;AN=2  GT:AD:DP 0/1:3,4:7
2       136234284  .    C    T      102.60 .      AC=1;AF=0.500;AN=2  GT:AD:DP 0/1:3,4:7
```

Multi-sample VCF

```
##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   HG00097     HG00100     HG00101
2       136045826    .    G    A    167.26  .       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    .    CGT  C    129.27  .       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    186.27  .       AC=1;AF=0.167;AN=6  GT:AD:DP 0/0:6,0:6   0/0:16,0:16 0/1:4,6:10
2       136048649    .    C    G    127.26  .       AC=1;AF=0.167;AN=6  GT:AD:DP 0/0:13,0:13  0/0:9,0:9    0/1:1,4:5
```

GVCF Files



- GVCF has records for all sites, whether there is a variant call there or not
- The records include an estimation of how confident we are in that the sites are homozygous-reference or not
- Adjacent non-variant sites merged into blocks

Variant callers

Selection of tools

- Germline variants
  - HaplotypeCaller, FreeBayes, BCFtools, DeepVariant
- Somatic variants
  - Mutect2, Strelka2

Which tool should I use?

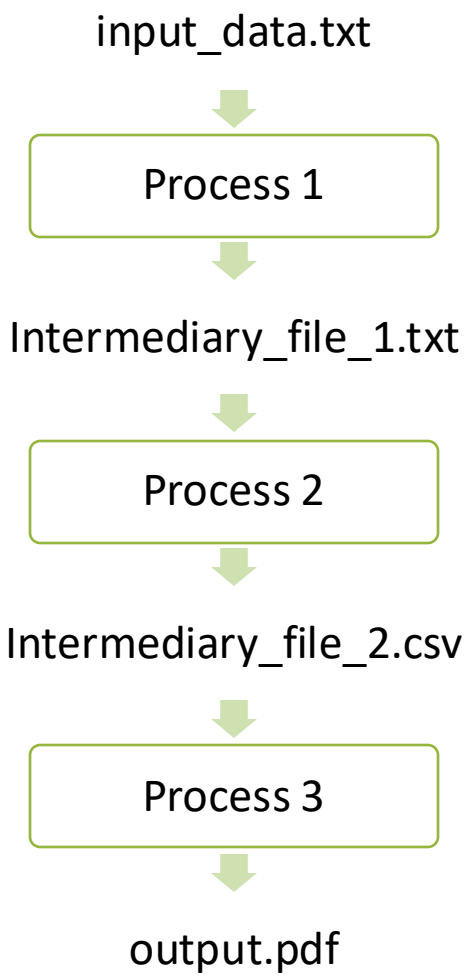
- GATK HaplotypeCaller is used in the lab – commonly used for variant calling in human
- Non-model organisms?

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Variant filtration

- Variant calling with e.g. HaplotypeCaller is designed to be sensitive
- Important to apply filters to limit false positives
- Examples of filters on information in the VCF file
  - QUAL < 30.0
  - Depth < 10
- VQSR is a filtering method based on machine learning
  - Recommended when there is a lot of data and
  - a list of known variants from the species

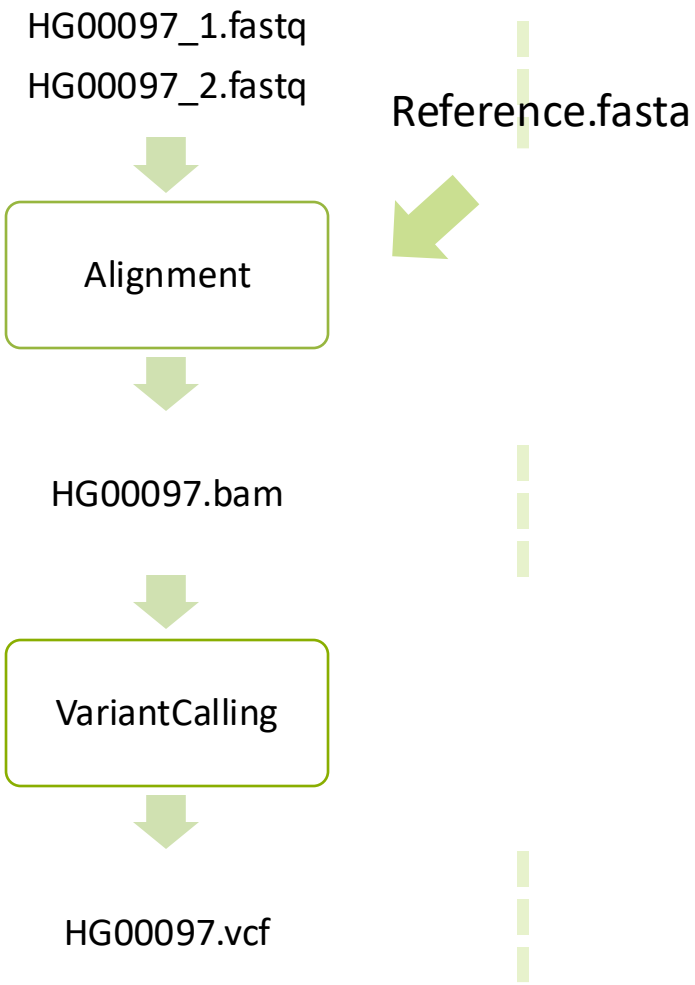
Workflows


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

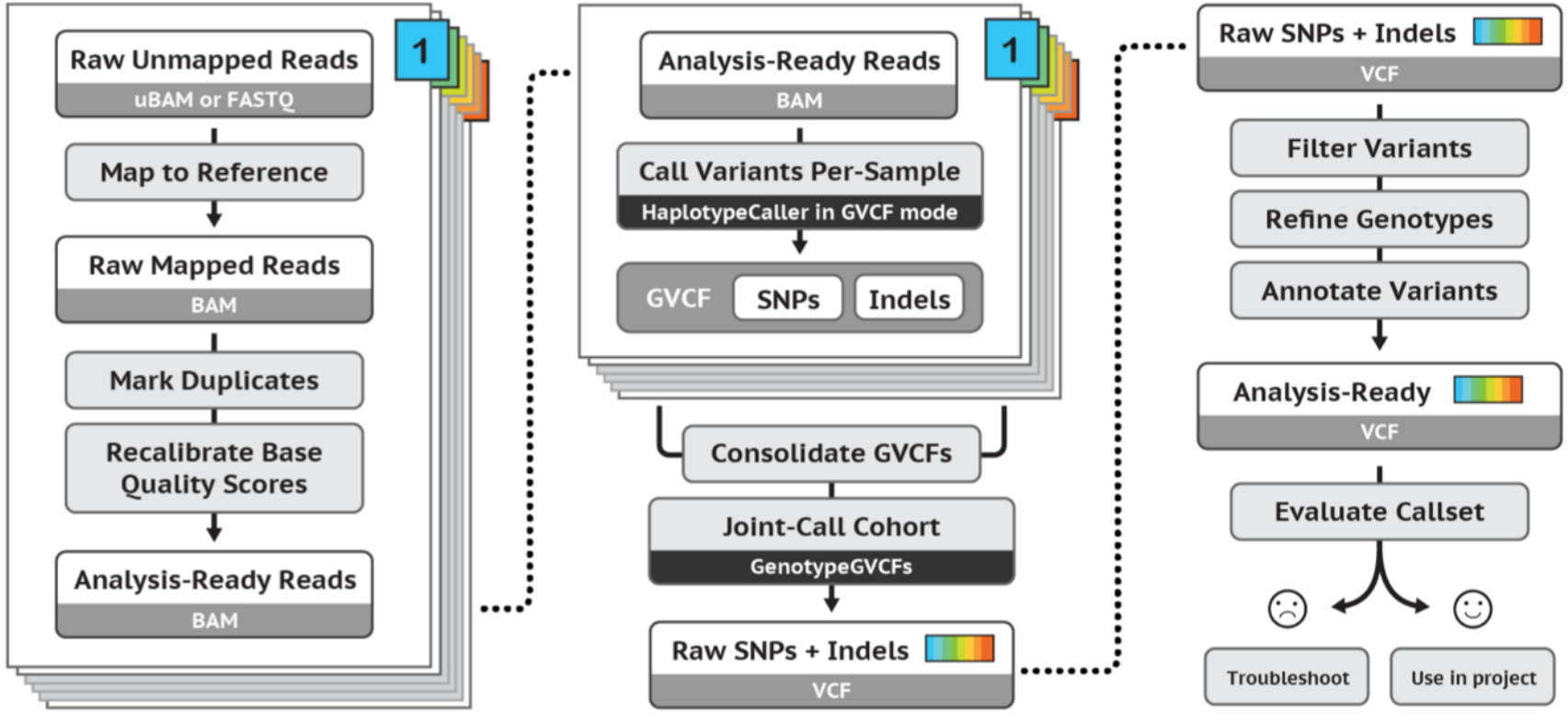
Workflow conventions

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

Example: Basic workflow, one sample

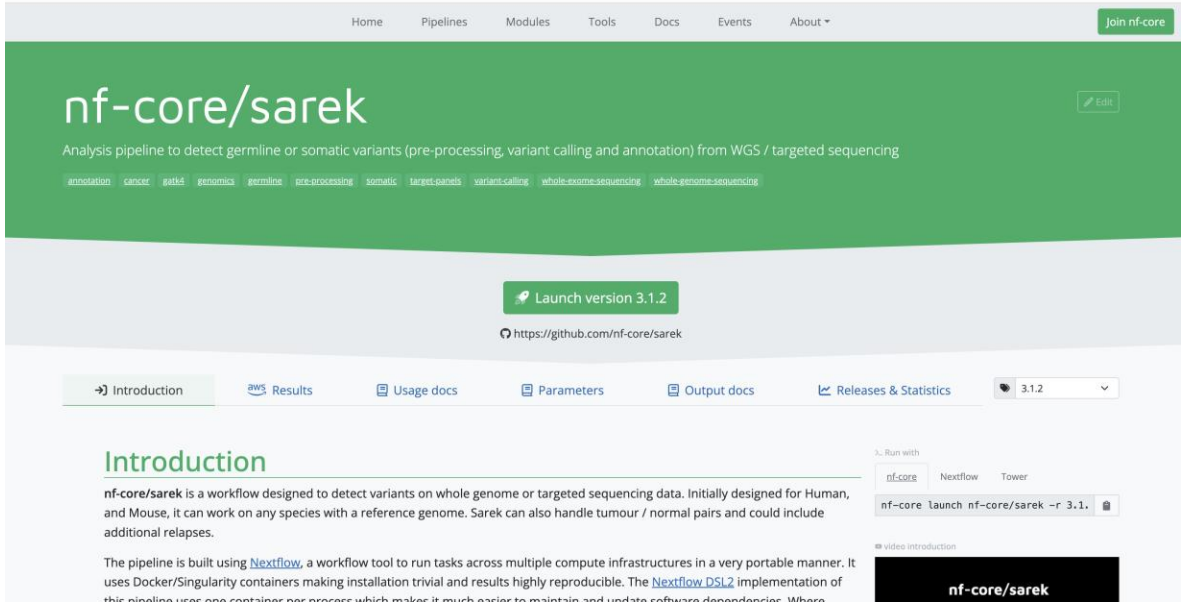


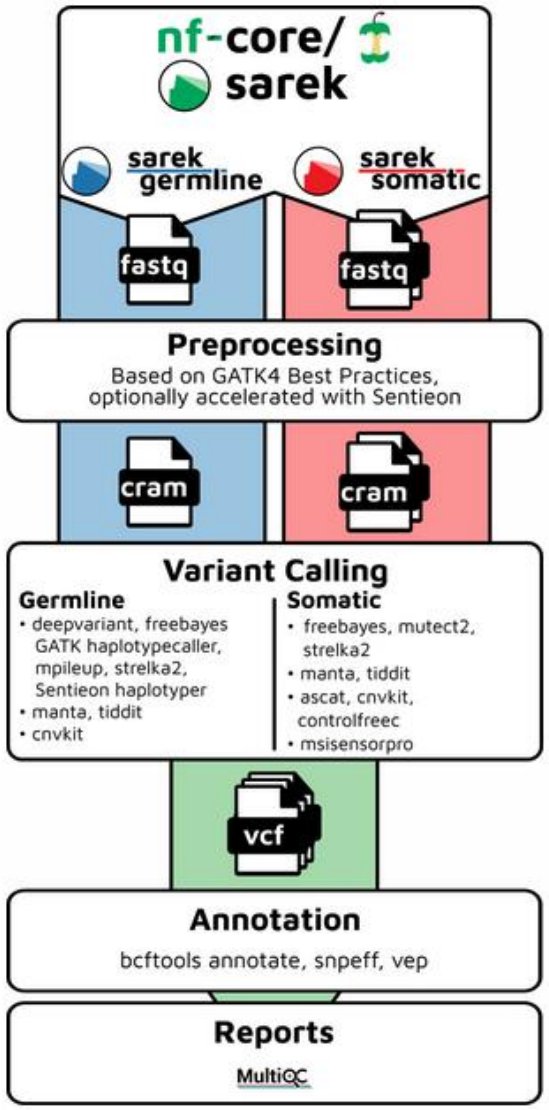
Refined workflow:
GATK's best practices workflow for germline short variant discovery



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

Nf-core variant calling workflow: Sarek





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

Overview of lab

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

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

The Lactase enzyme

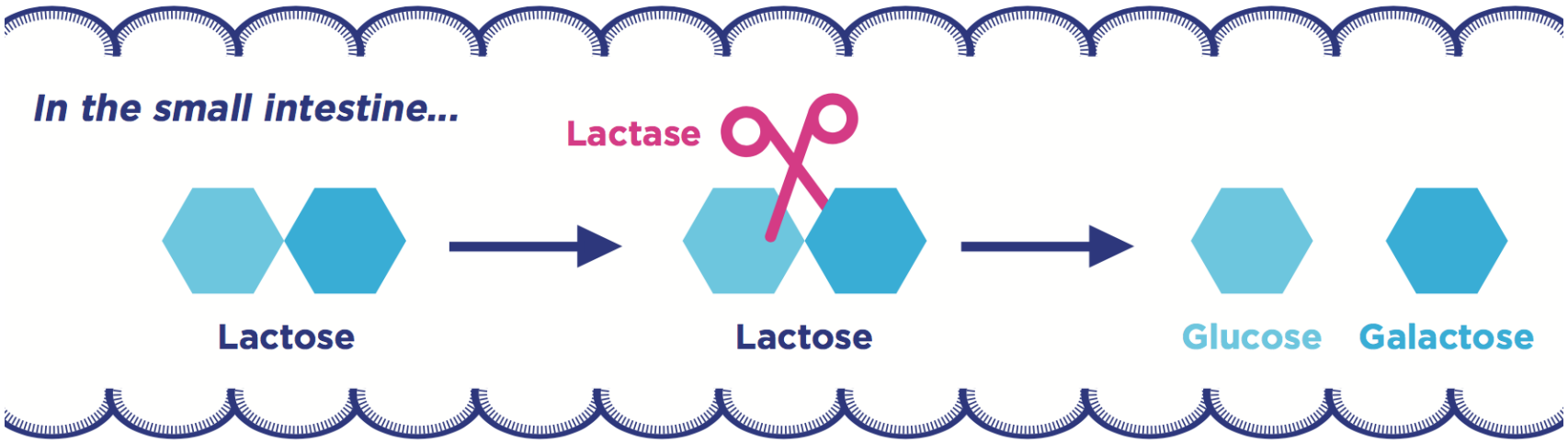
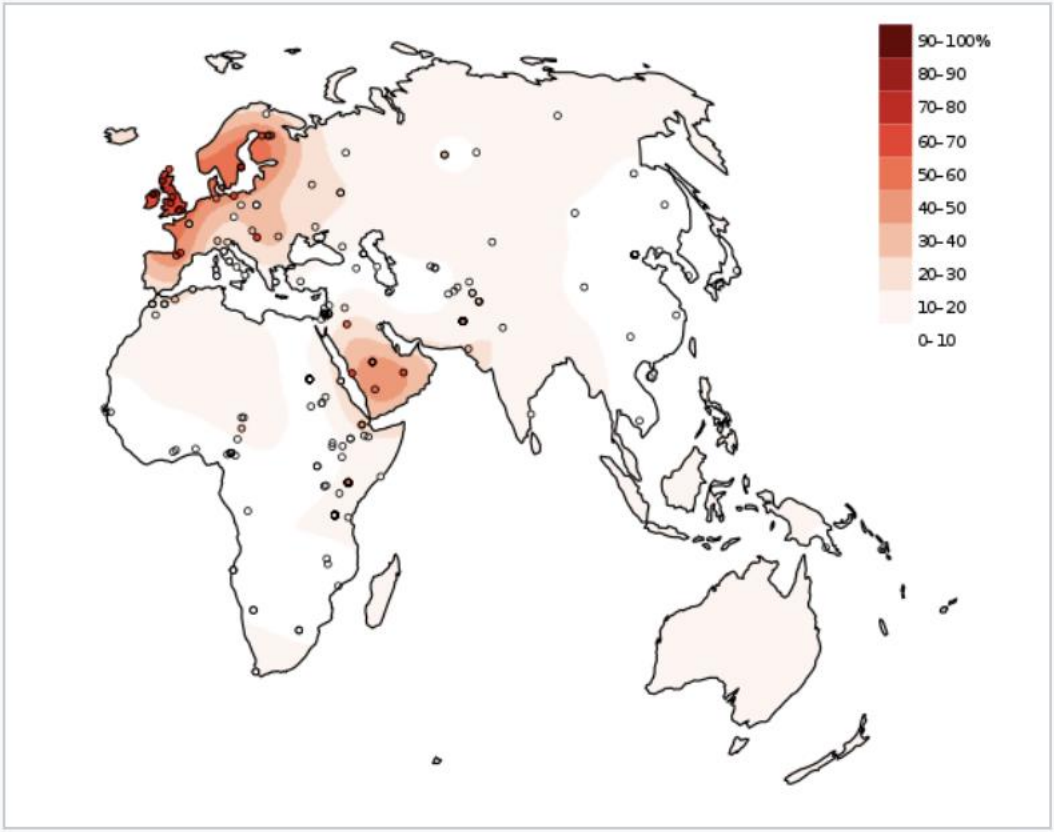


Figure 2. Lactose digestion in the intestine.

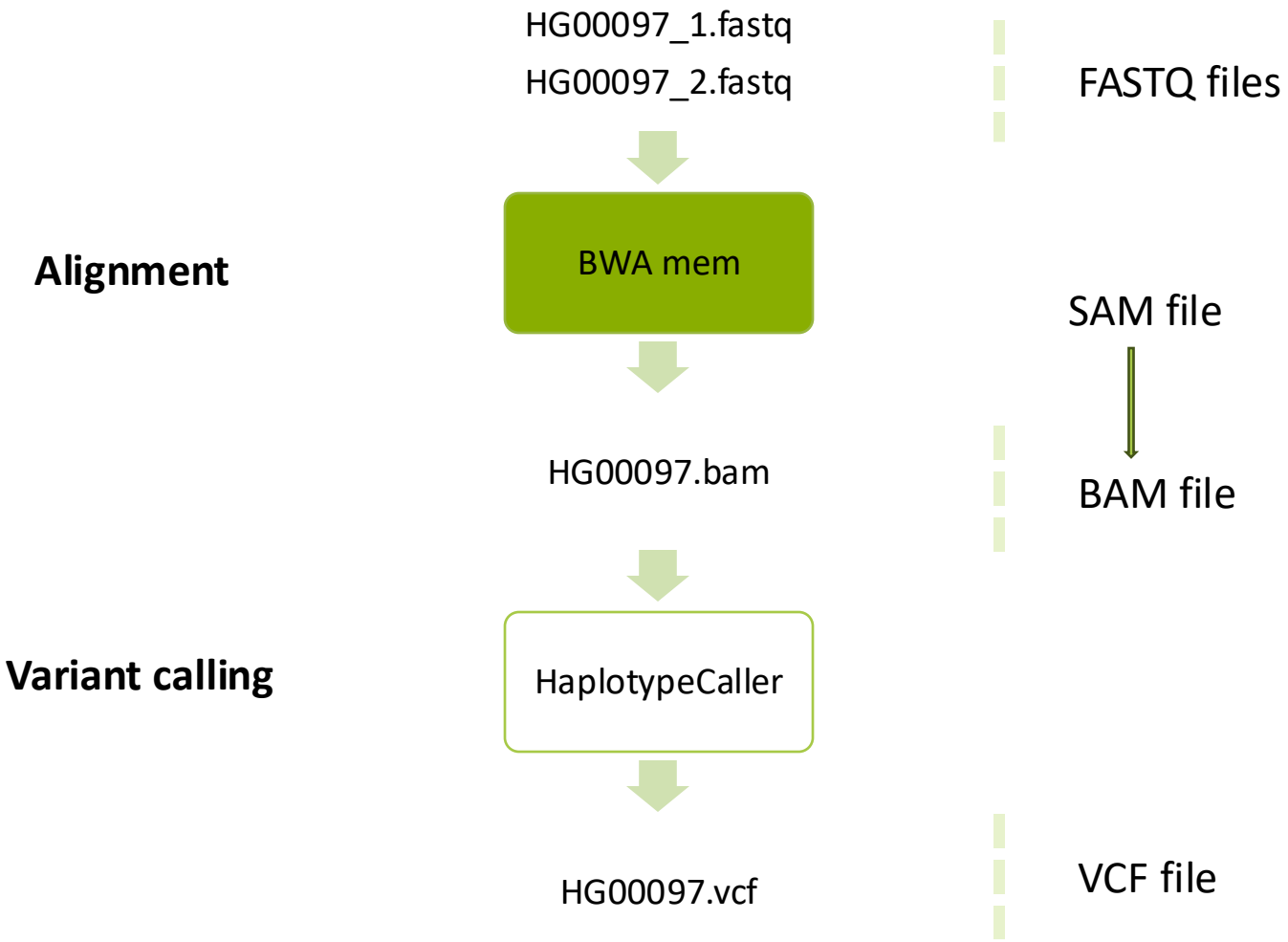
- All mammals produce lactase as infants
- Some human produce lactase in adulthood
- Genetic variation upstream of the *LCT* gene cause the lactase persistent phenotype (lactose tolerance)

The Lactase enzyme



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

Part1: Basic Variant Calling in one sample



File indices

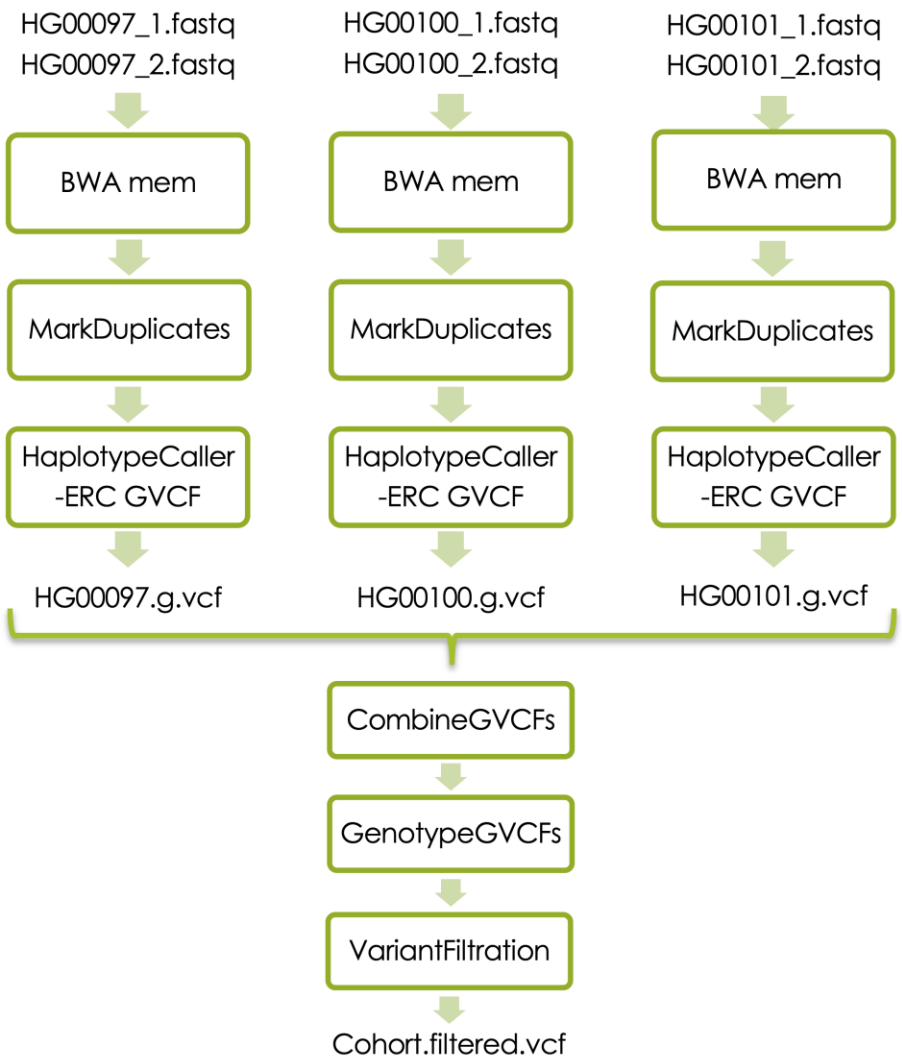
- Most large files we work with need indices
  - reference genome (.**fasta**)
  - aligned reads (.**bam**)
  - Variants **(.vcf**)
- Allows efficient access to the large file
- The index is stored in a file (or several files)
- Different indices for different file types
  - .fasta.fai, .bam.bai, .vcf.idx, etc
  - BWA index (several files)

ReadGroups

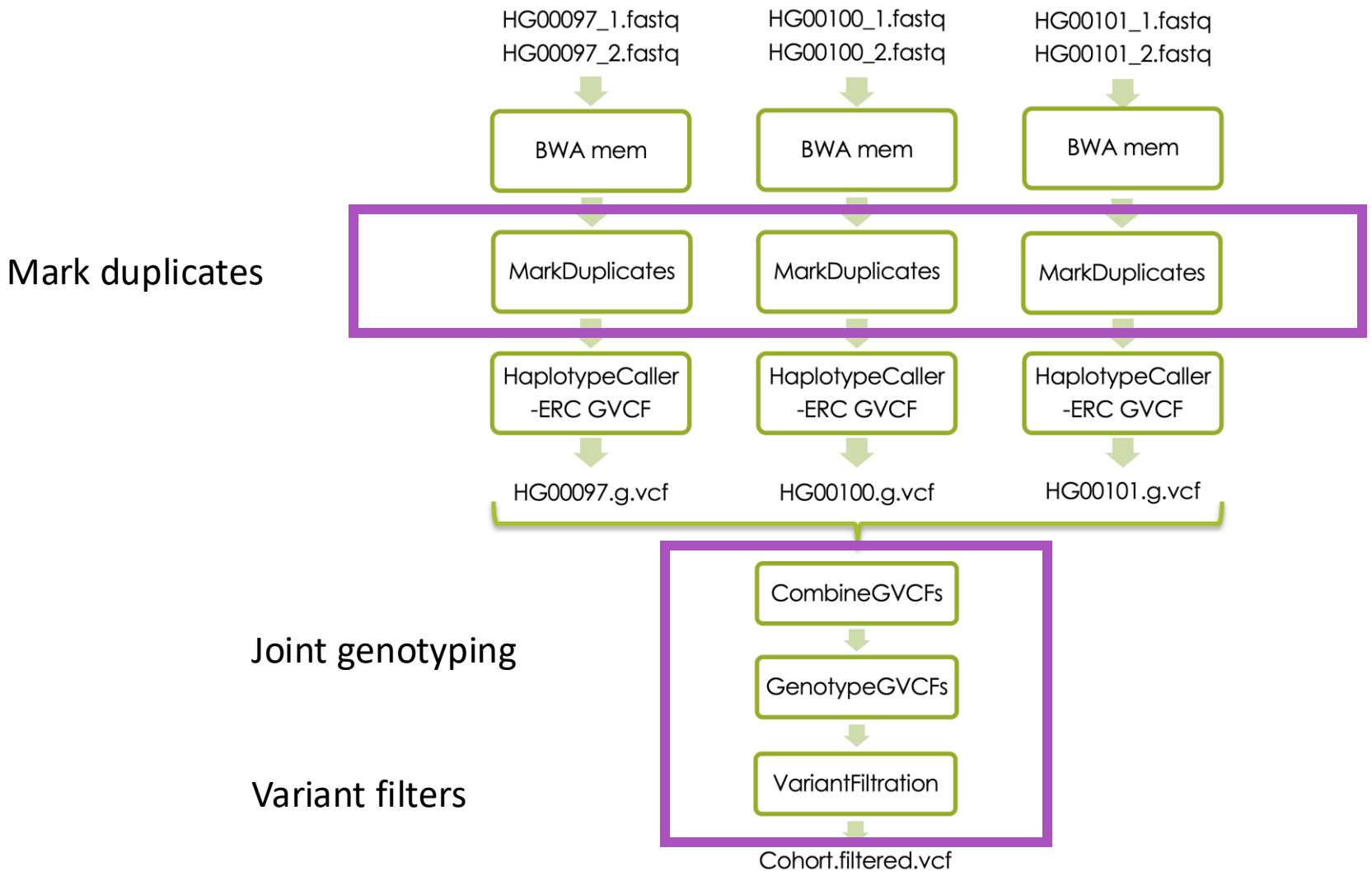
- Tags that mark which sample and sequencing run each read comes from
  - RGID: unique identifier of sequencing run
  - RGSM: sample name used in the vcf file
  - RGLB – library id, used to get correct duplicate marking (for example)

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

Part2: Basic variant calling in cohort

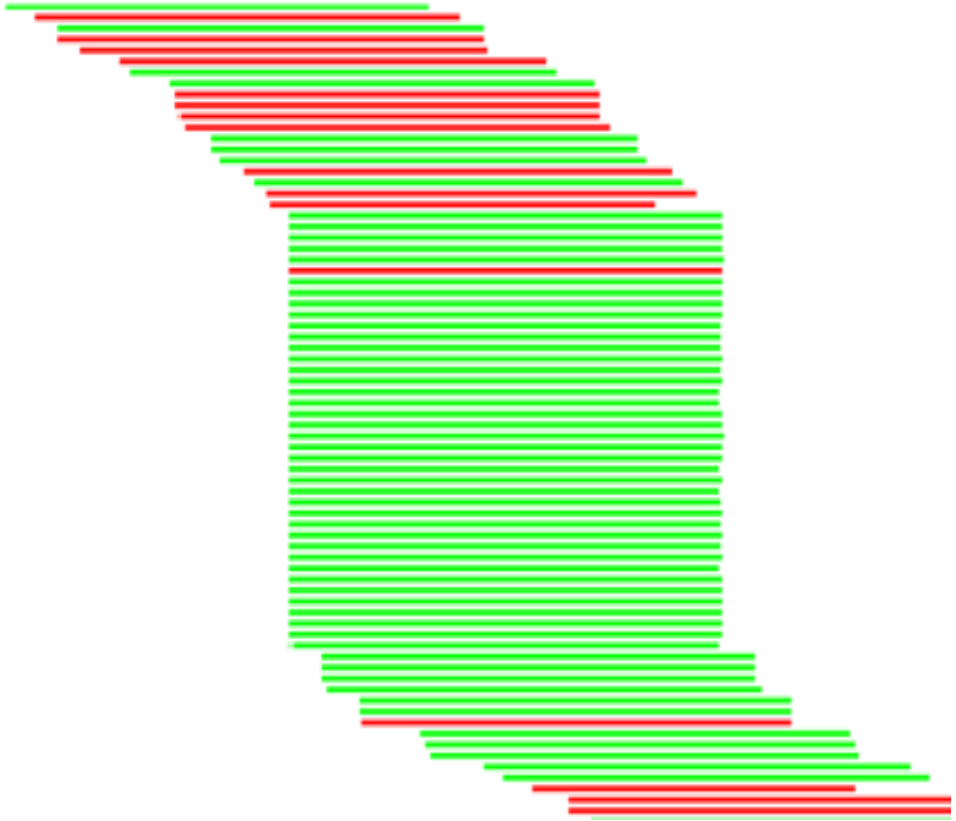


Basic variant calling in cohort



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



NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

gatk

User Guide Tool Index Blog Forum DRAGEN-GATK

Need Help?

Search our documentation

MarkDuplicates

GATK / Tool Index / 4.0.1.1

MarkDuplicates (Picard)

Follow

GATK Team
10 months ago · Updated

Identifies duplicate reads.

This tool locates and tags duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA. Duplicates can arise during sample preparation e.g. library construction using PCR. See also EstimateLibraryComplexity for additional notes on PCR duplication artifacts. Duplicate reads can also result from a single amplification cluster, incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument. These duplication artifacts are referred to as optical duplicates.

```
gatk --java-options -Xmx7g MarkDuplicates \
      -I input.bam \
      -O marked_duplicates.bam \
      -M marked_dup_metrics.txt
```

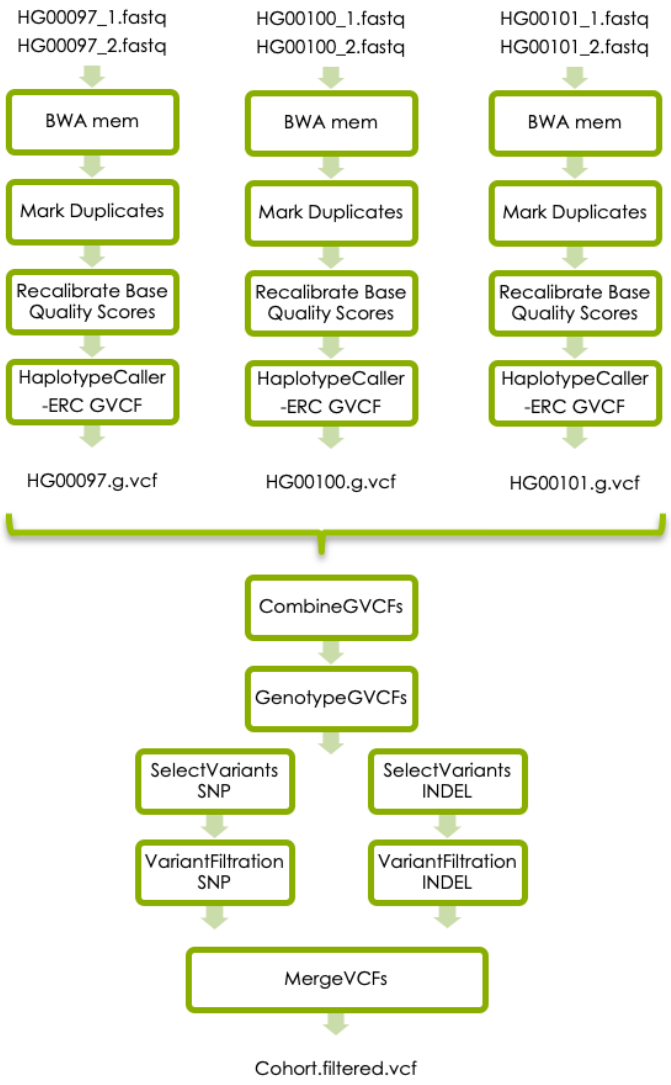
Variant filtration

- Variant calling with HaplotypeCaller is designed to be sensitive
  - Apply filters to limit false positives
  - Advanced filtration (VQSR) requires more data than we have here
  - Here: “Hard filters” - select cutoffs on e.g. quality or read coverage
    - Cutoffs can be selected after viewing quality score distributions

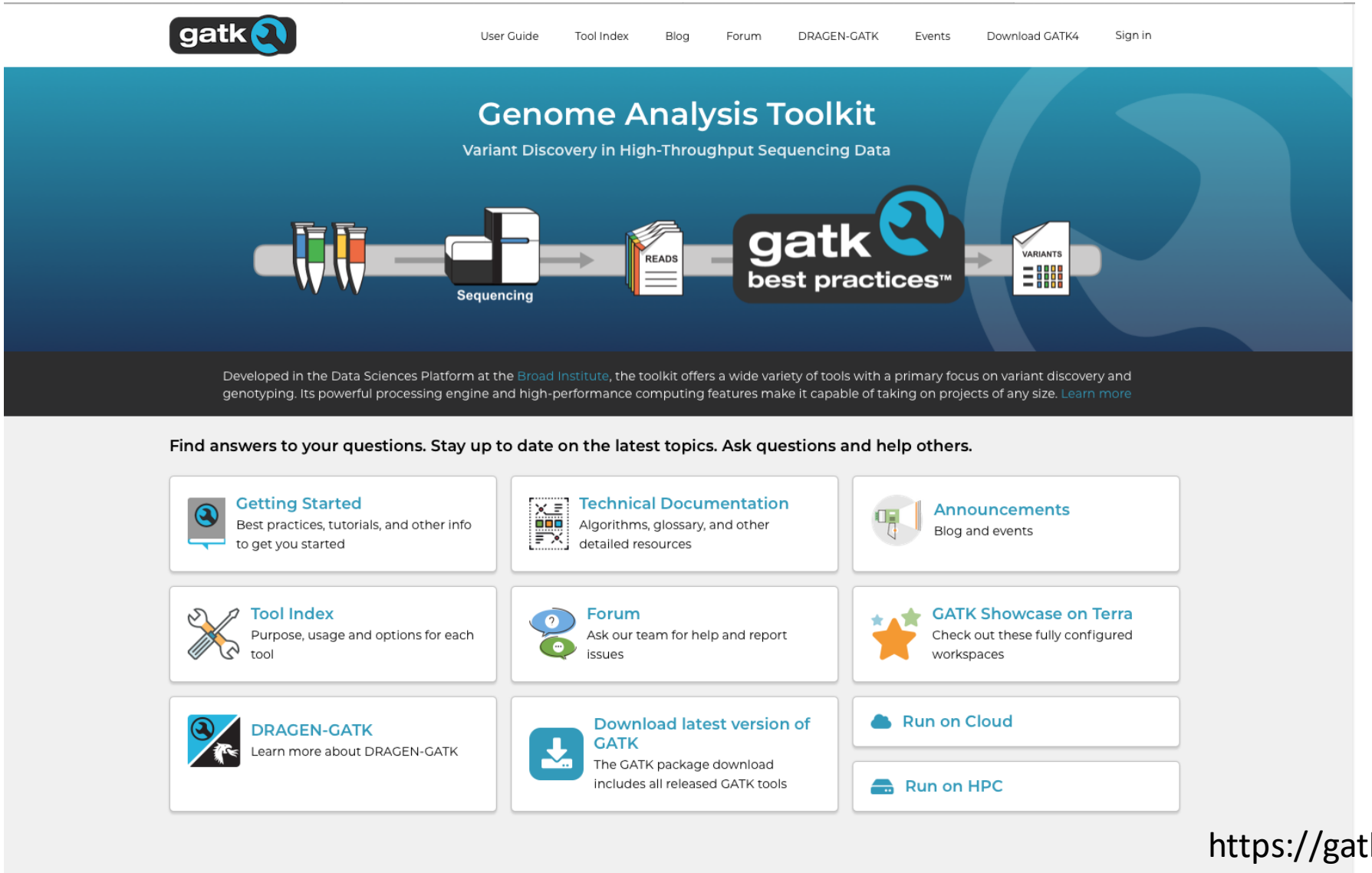
Part3: Bash script for variant calling

```bash
#!/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
```

Extra lab (Part4): GATK's best practises

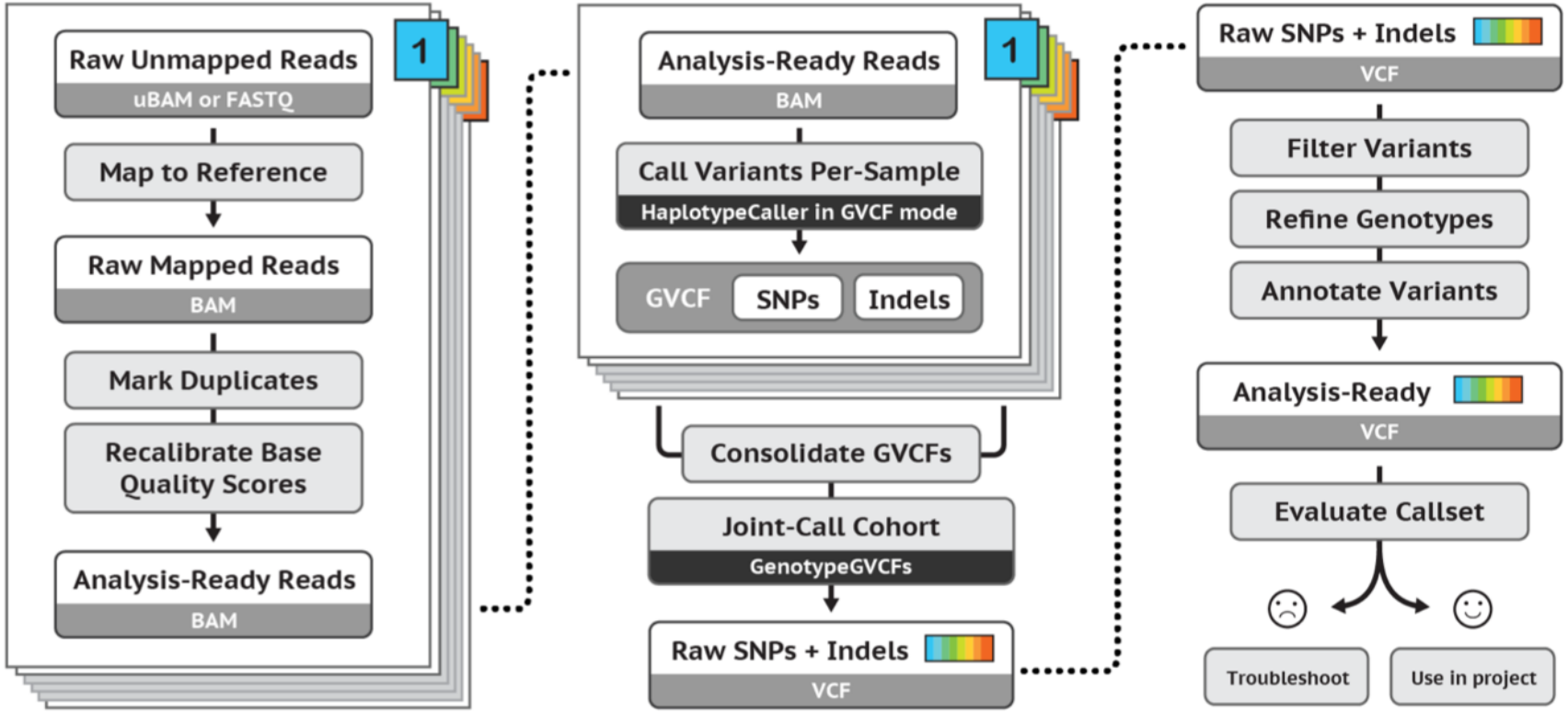


GATK best practices for short variant discovery



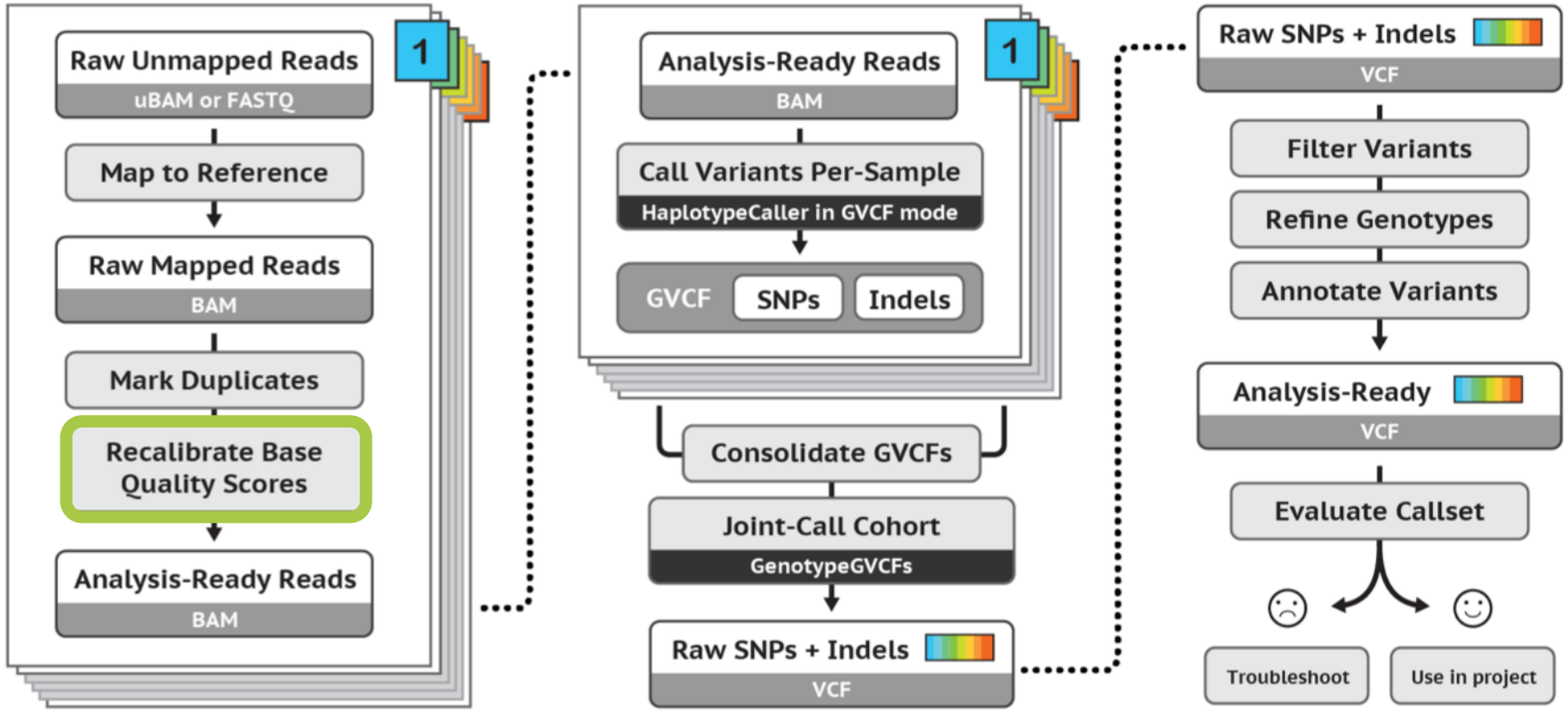
https://gatk.broadinstitute.org

GATK's best practices workflow for germline short variant discovery



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

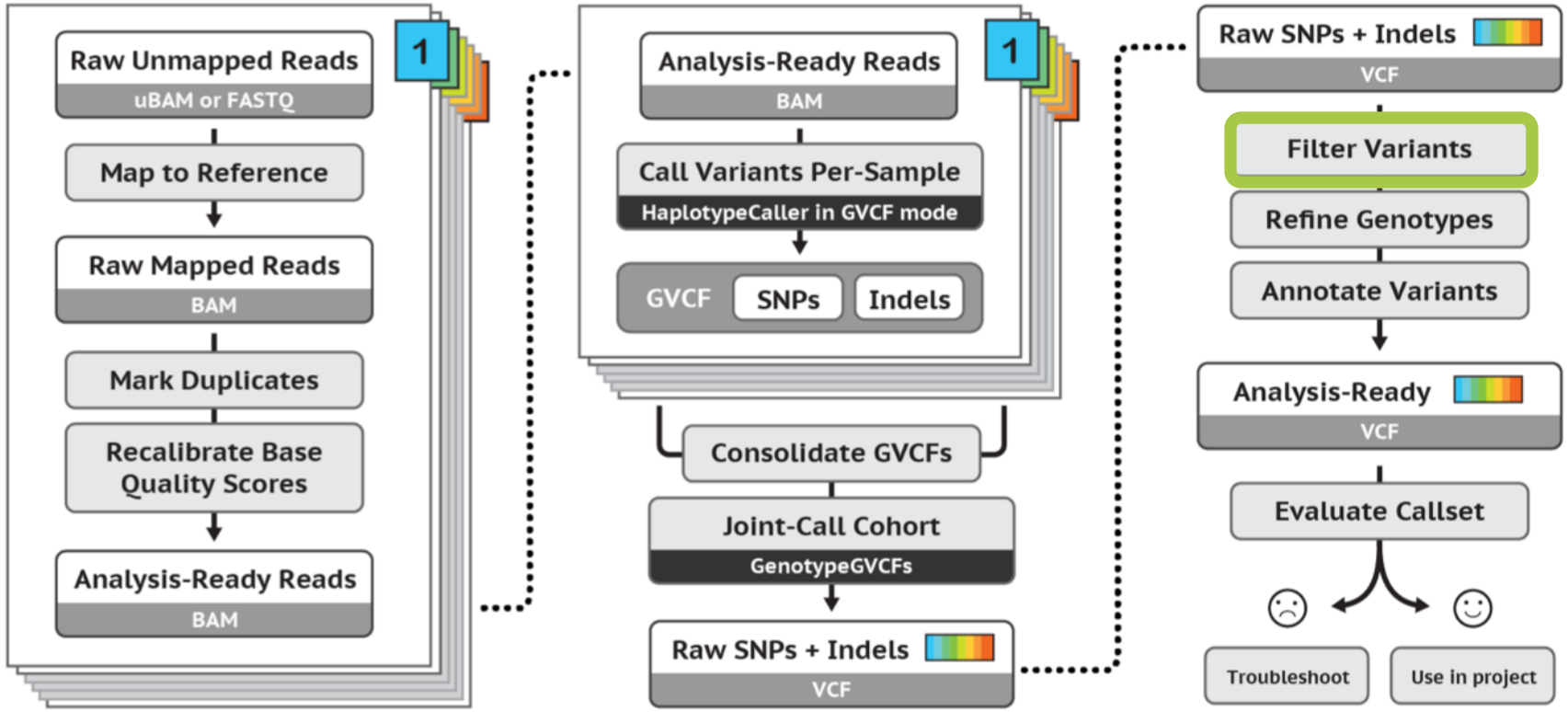
Base Quality Score Recalibration (BQSR)



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 the physics or chemistry of the sequencing reaction
   - manufacturing flaws in the equipment
   - etc
3. Can cause biases in variant calling
4. **Base Qualtiy Score Recalibration** helps to calibrate the scores so that they correspond to the real per-base sequencing error rate (phred scores)

Filter variants



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

Germline short variant discovery (SNPs + Indels)

Filter variants

Variant quality score recalibration (VQSR):

- Machine learning algorithm trained to recognise ”likely false” variants
- For large data sets (>1 WGS or >30WES samples) - **Use VQSR when possible!**

“Hard” filters:

For smaller data sets

Filters on information in the VCF file using set cutoffs

For example: Flag variants with ”QD < 2.0”