NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Variant-calling Workflow

Malin Larsson

Malin.Larsson@nbis.se

# Overview

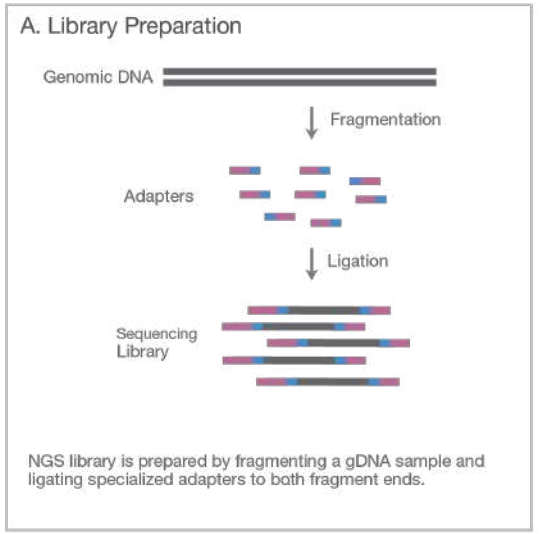
- Workflows
- Basic variant calling in one sample
- Basic variant calling in cohort
- Introduction to exercise

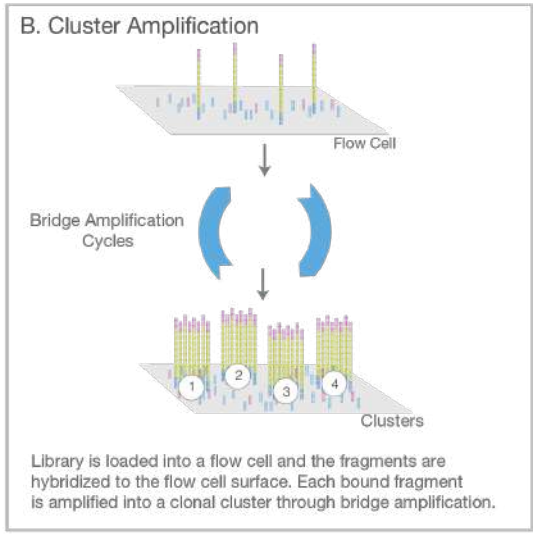
In separate talk Thursday at 9:

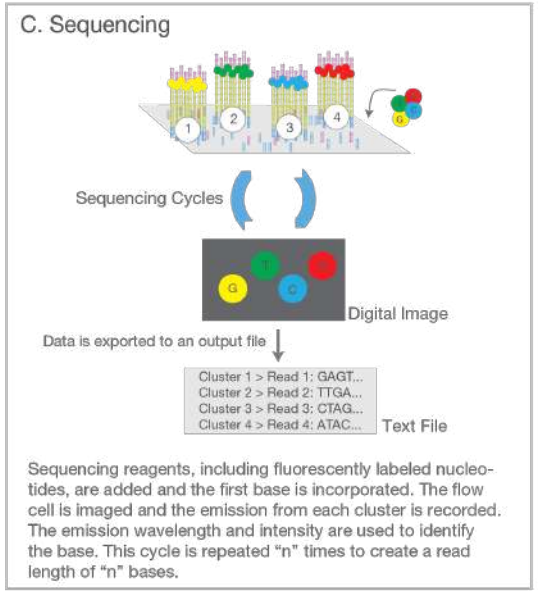
- GATK's Best practices

# Illumina Sequencing

https://www.youtube.com/watch?v=fCd6B5HRaZ8



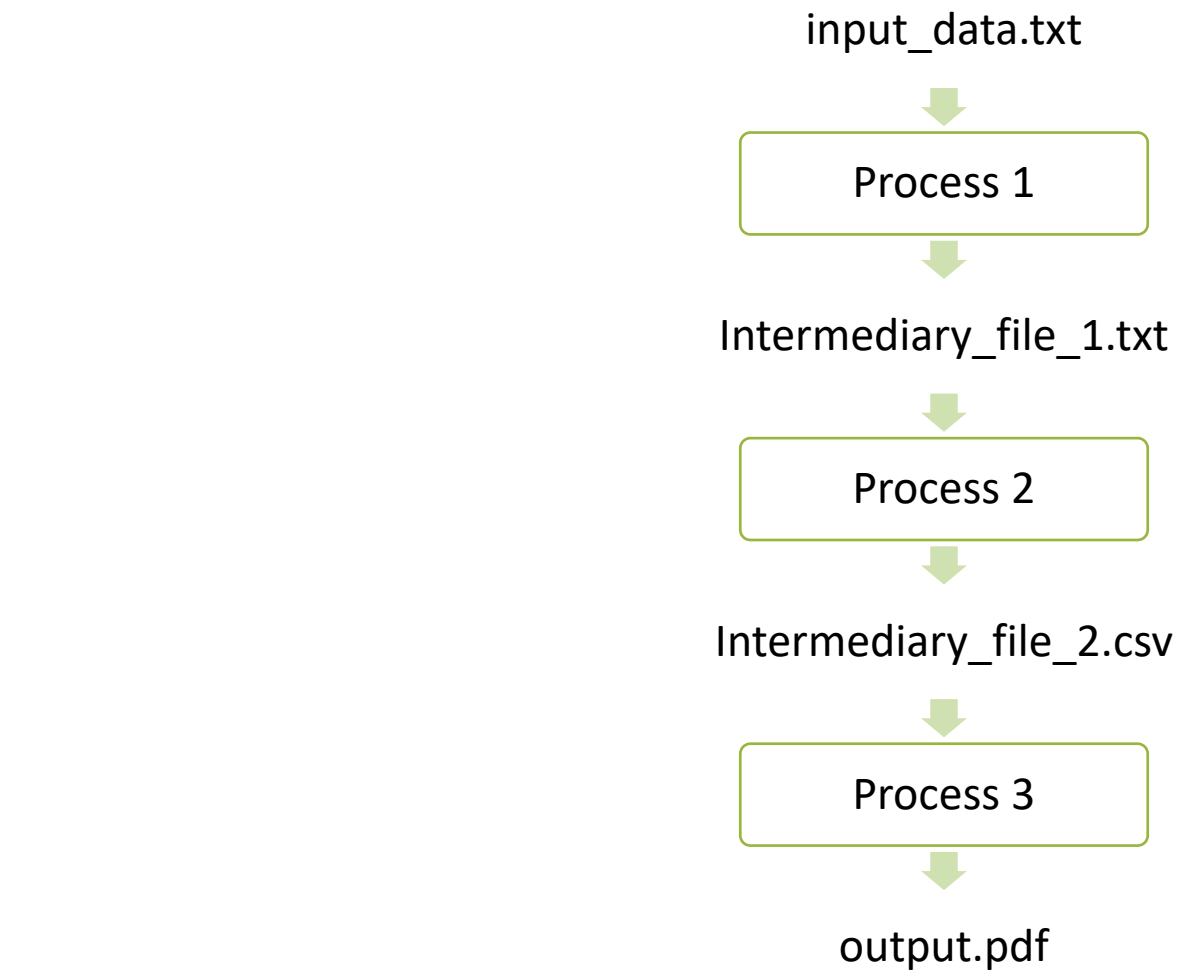




NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Workflows

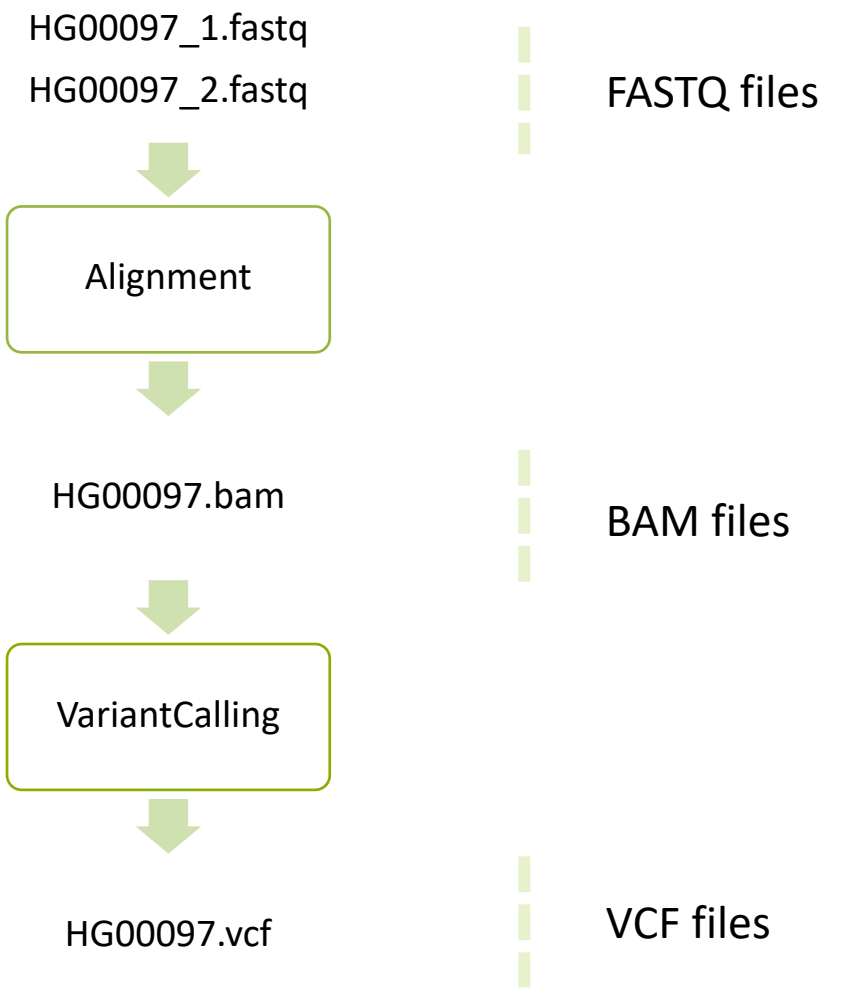
# What is a workflow



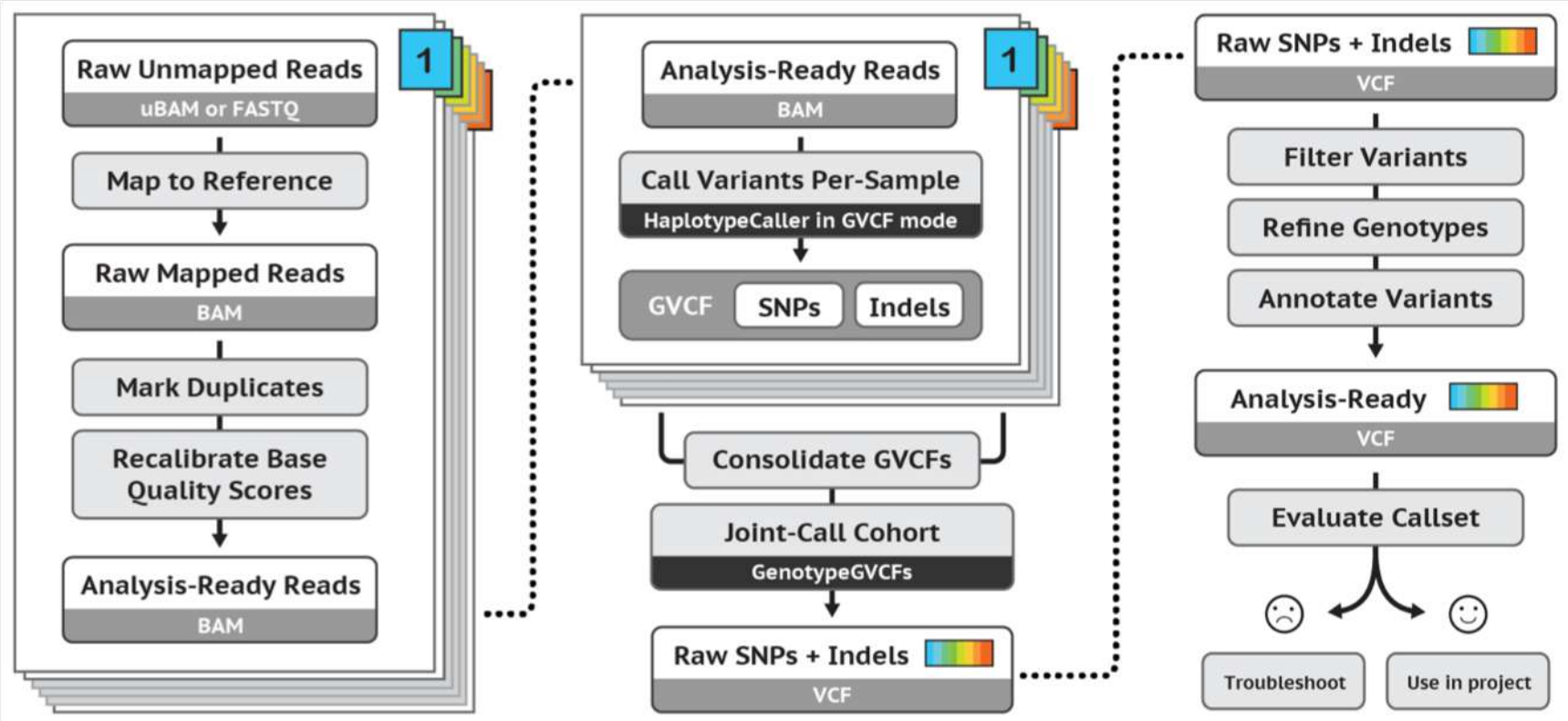
# Workflow conventions

- Create a new output file in each process – don't ower write the input file
- Use informative file names
- Include information of the process in output file name

# Example: Basic variant calling in one sample



# GATK's best practices workflow for germline short variant discovery



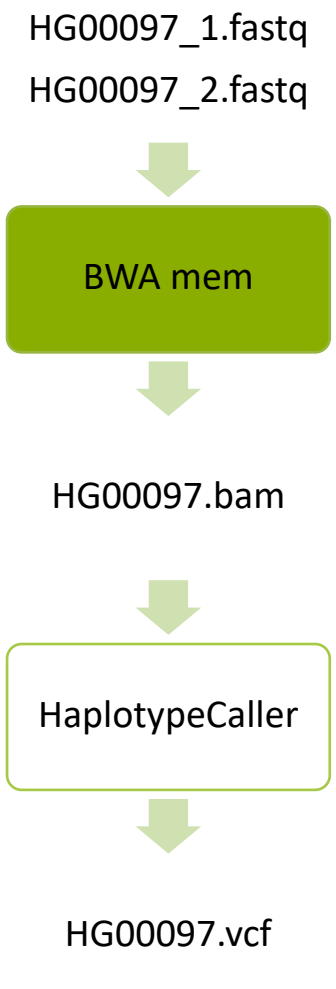

https://software.broadinstitute.org/gatk/best-practices/

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Basic Variant Calling in one sample

# Alignment

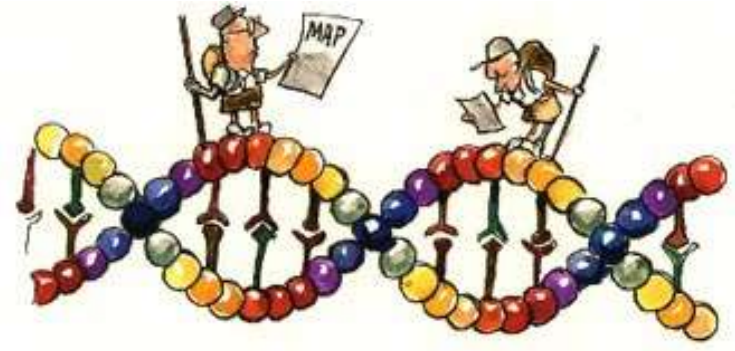



FASTQ files

BAM files

VCF files

# The reference genome




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

The first draft of the human genome contained 150,000 gaps.

GRCh37: 250 gaps

GRCh38 is the latest version of the human reference genome, but we will work with GRCh37 in the lab.

# Keep track of the reference version!

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

- mapping
- variant calling
- annotation

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.

# File indices

- Most large files we work with, such as the reference genome, need an index
- Allows efficient access to the file
- Different indices for different file-types
- Bwa index = Burrows-Wheeler transform of reference genome (several files)
- Needs index: fasta, bam vcf files

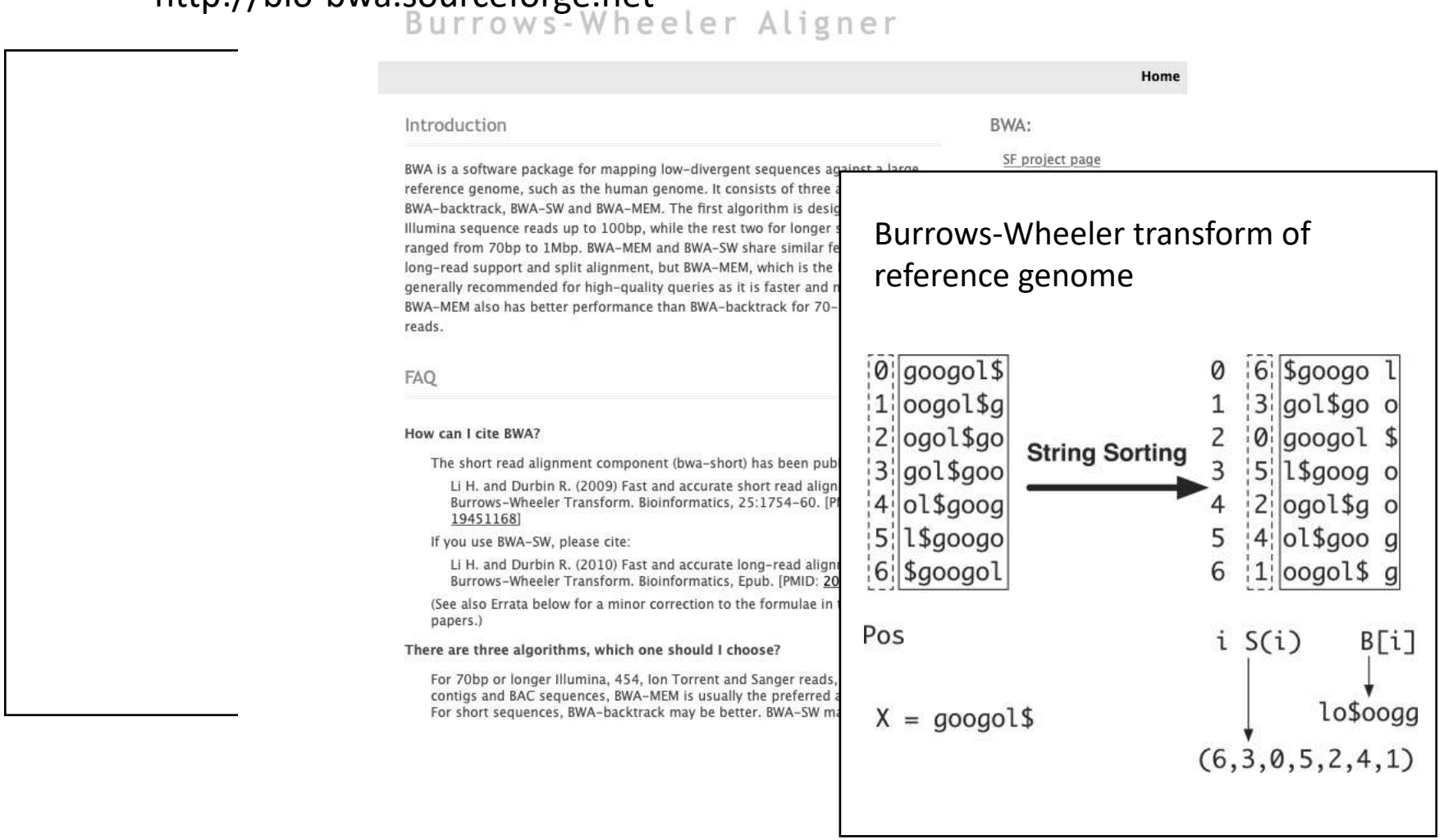
# Alignment

```
module load bwa
```

```
AACAGGTATATCTTCCCCGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTACCCTCTTCCTTAGGGACTGTAC
                                                GCTAGCTAGCTAGCTACCCT
```

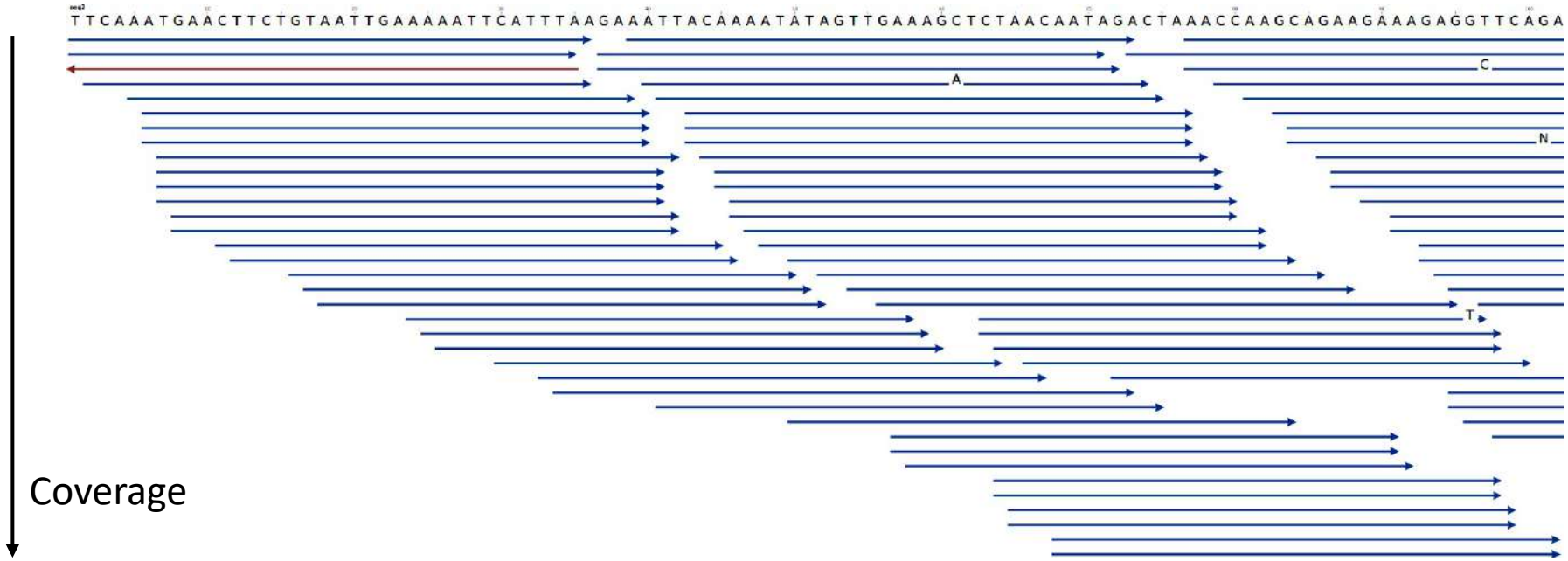
# Burrows-Wheeler Aligner

http://bio-bwa.sourceforge.net



# Alignment

```
module load bwa
```



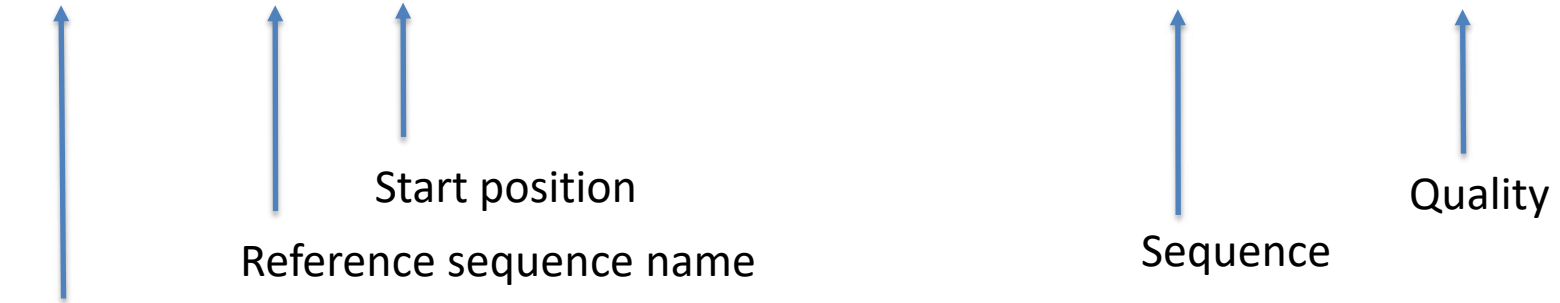
# Output from mapping - Sam format

## 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
```



# Convert to Bam

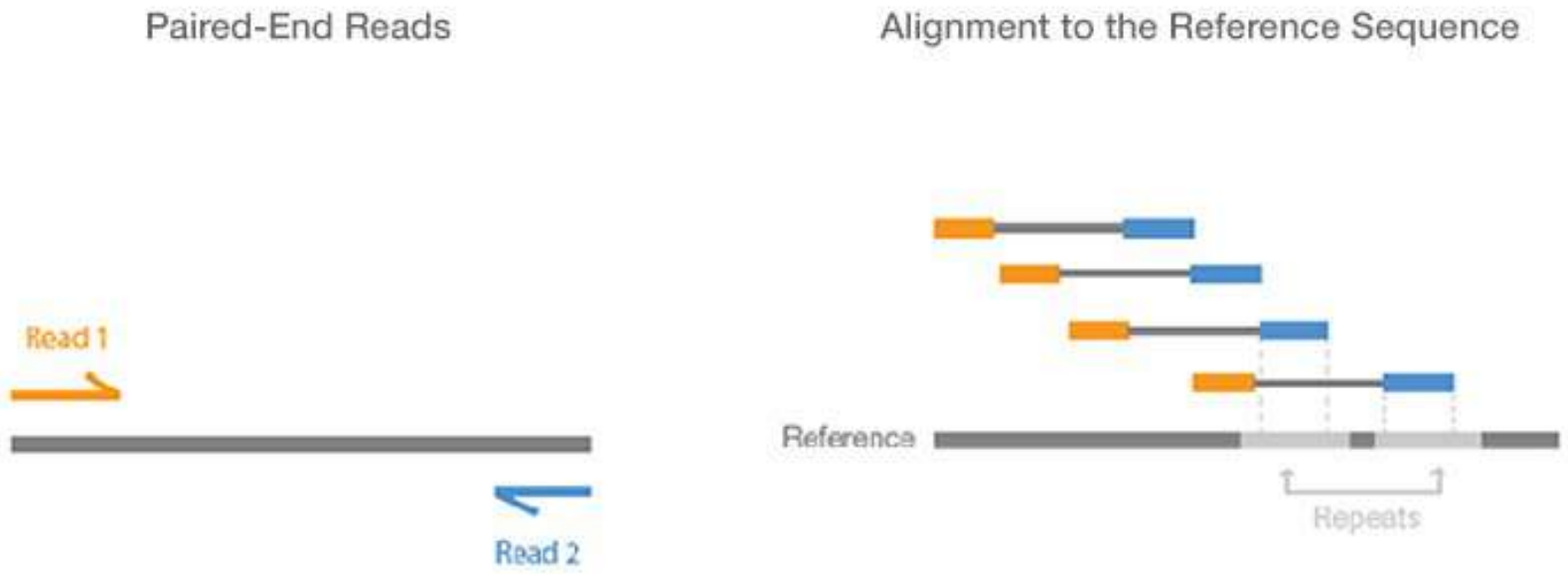
Bam file is a binary representation of the Sam file

# Read groups

- Link *sample id*, *library prep*, *flowcell* and *sequencing run* to the reads.
- Good for error tracking!
- Often needed for variant calling
- Detailed description at https://gatkforums.broadinstitute.org/gatk/discussion/6472/read-groups

**RGID** = *combination of the sample id and run id*
**RGLB** = Library prep
**RGPL** = Platform (for us ILLUMINA)
**RGPU** = Run identifier *usually barcode of flowcell*
**RGSM** = Sample name

# Paired-End data




Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

# Paired-end data

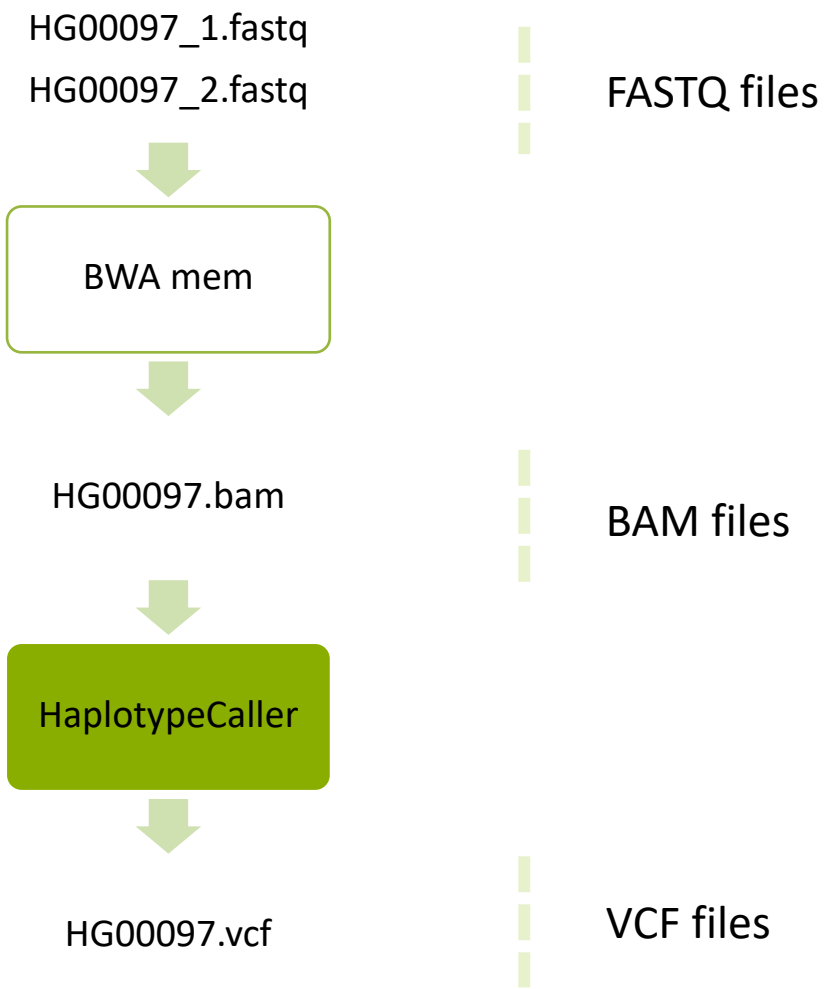
ID_R1_001.fastq

```
@HISEQ:100:C3MG8ACXX:5:1101:1160:2
197 1:N:0:ATCACG
CAGTTGCGATGAGAGCGTTGAGAAGTATAATAGG
AGTTAAACTGAGTAACAGGATAAGAAATAGTGAG
ATATGGAAACGTTGTGGTCTGAAAGAAGATGT
+
B@CFFFFFHHHHHGJJJJJJJJJJFHHIIIIJJ
JIHGIIJJJJIJIJIJJJJIIJJJJJIIEIHHIJ
HGHHHHHDFFFEDDDDDCDDDCDDDDDDDCDC
```

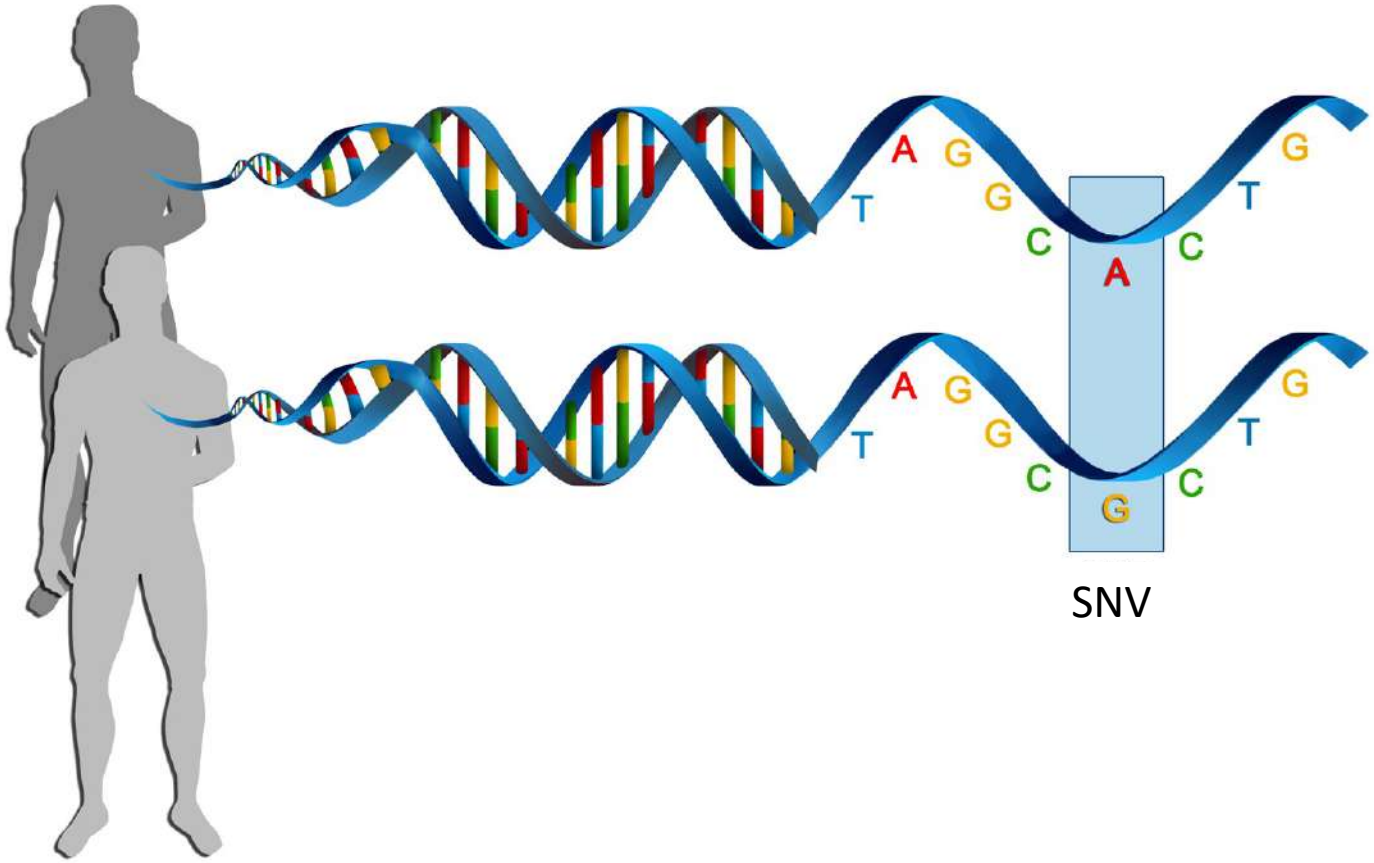
ID_R2_001.fastq

```
@HISEQ:100:C3MG8ACXX:5:1101:1160:
2197 2:N:0:ATCACG
CTTCGTCCACTTTCATTATTCCTTTCATACATG
CTCTCCGGTTTAGGGTACTCTTGACCTGGCCTT
TTTTCAAGACGTCCCTGACTTGATCTTGAAACG
+
CCCFFFFFHHHHHJJJJIJJJJJJJJJJJJJJ
JJJJJJJIJIJGIJHBGHHIIIJIJJJJJJJJI
JJJHFFFFFFDDDDDDDDDDDDDDDDEDCCDDDD
```

# Variant calling



# Genetic variation




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

# Detecting variants in reads

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

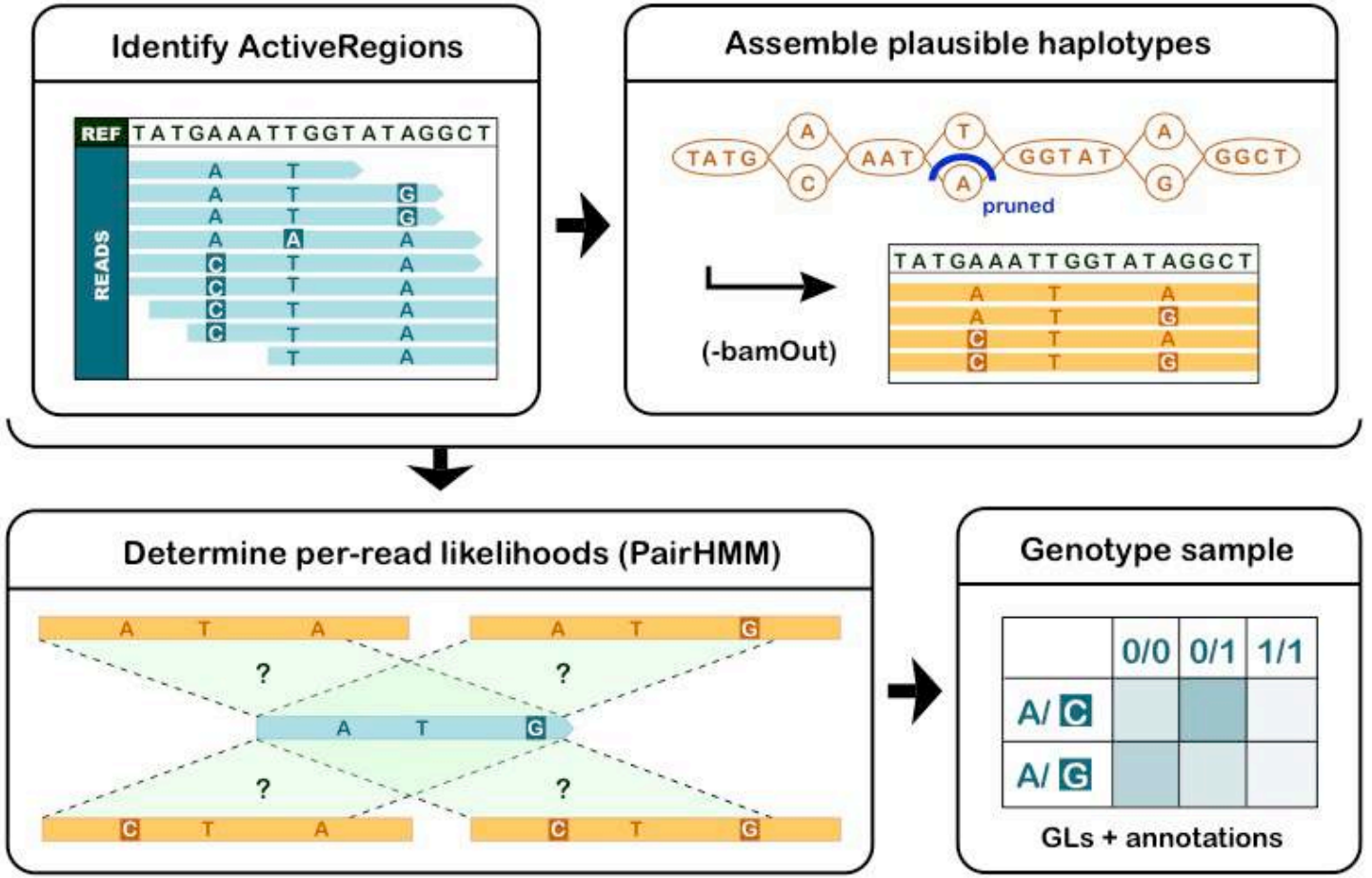
# Reference- and alternatve alleles

TGGGCTTTTCCAACAGGTATATCTTCCCCGCTAGCTAGCTAGCTACTTCAAATTCCT

| | |
|---|---|
| **Reference allele** | AGCTAGCTA |
| **Alternative allele** | AGCTGGCTA |

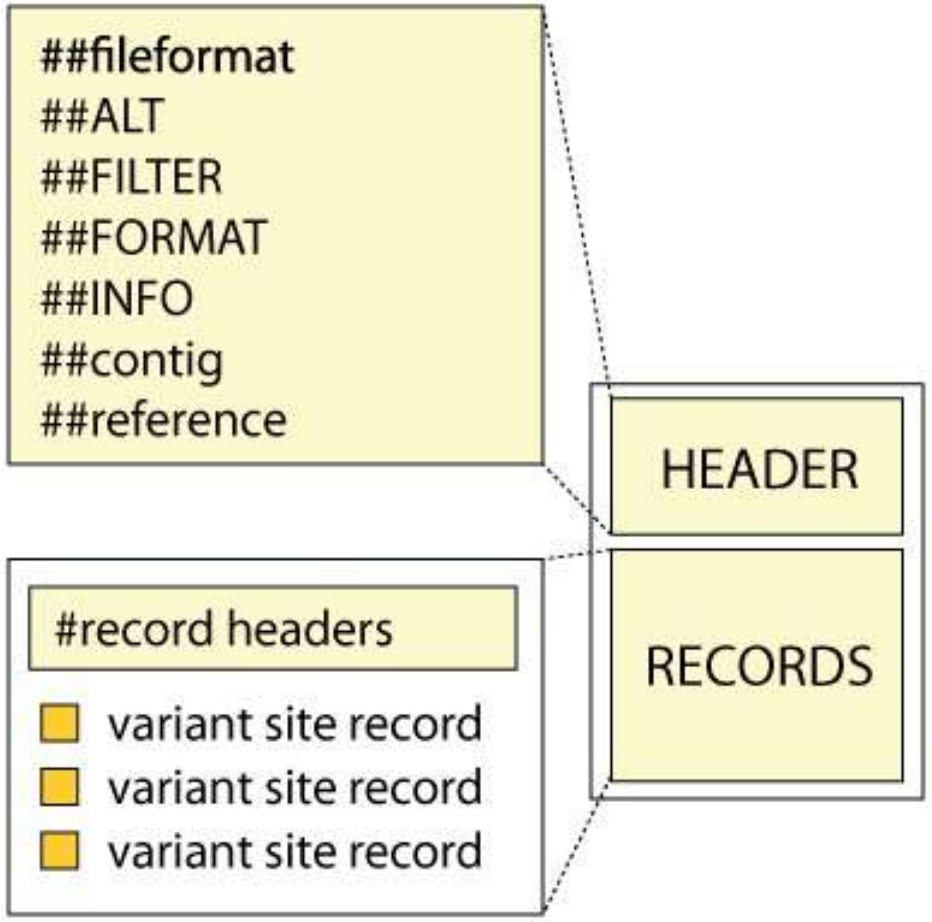
**Reference allele** = the allele in the refence genome
**Alternative allele** = the allele NOT in the refence genome

# Variant Calling HaplotypeCaller




For more info: https://www.youtube.com/watch?v=NQHGkVGICpY

# Variant Call Format (VCF)



# 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
2       136263277       .       T       A       148.60  .       AC=1;AF=0.500;AN=2  GT:AD:DP    0/1:8,5:13
…
…
```

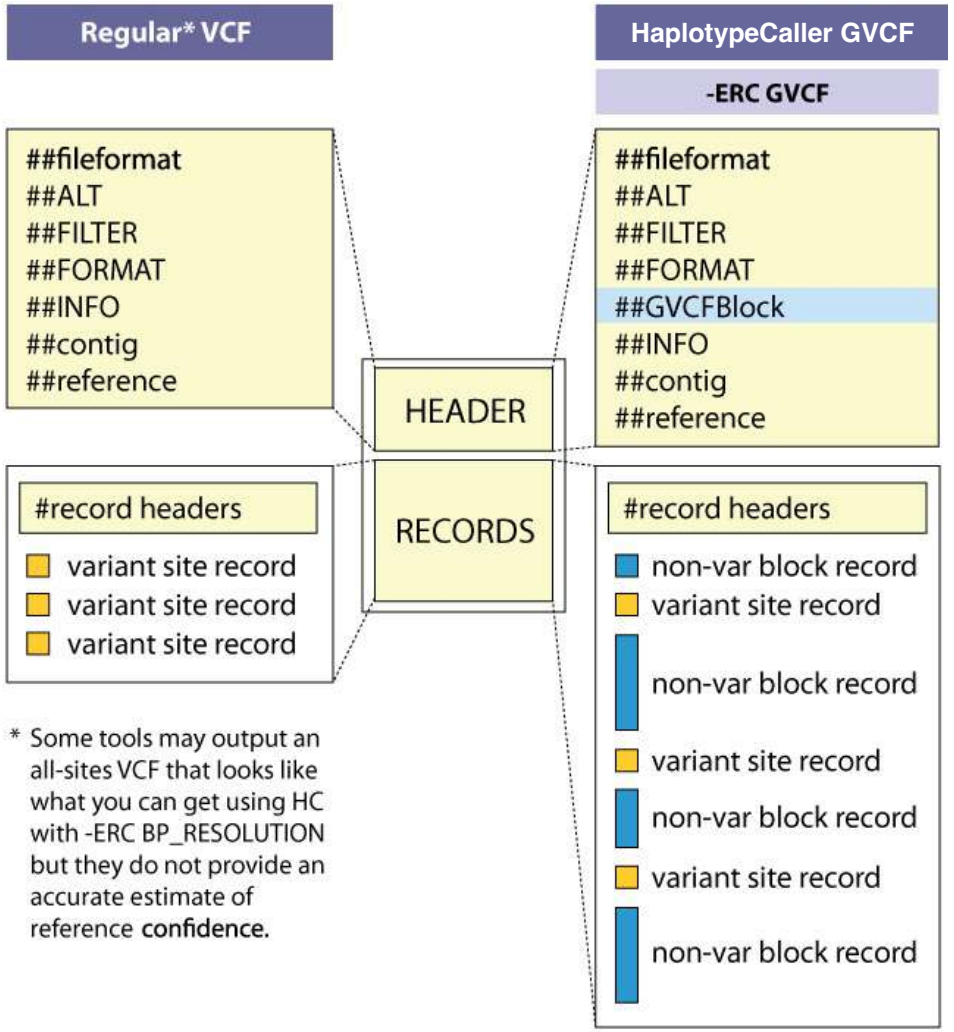
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Variant calling in cohort

# Basic variant calling in cohort

# GVCF Files are valid VCFs with extra information




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

# Basic variant calling in cohort

# Variant Call Format (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 |
| 2 | 136052318 | . | C | T | 107.26 | . | AC=1;AF=0.167;AN=6 | GT:AD:DP | 0/0:7,0:7 | 0/0:13,0:13 | 0/1:3,3:6 |

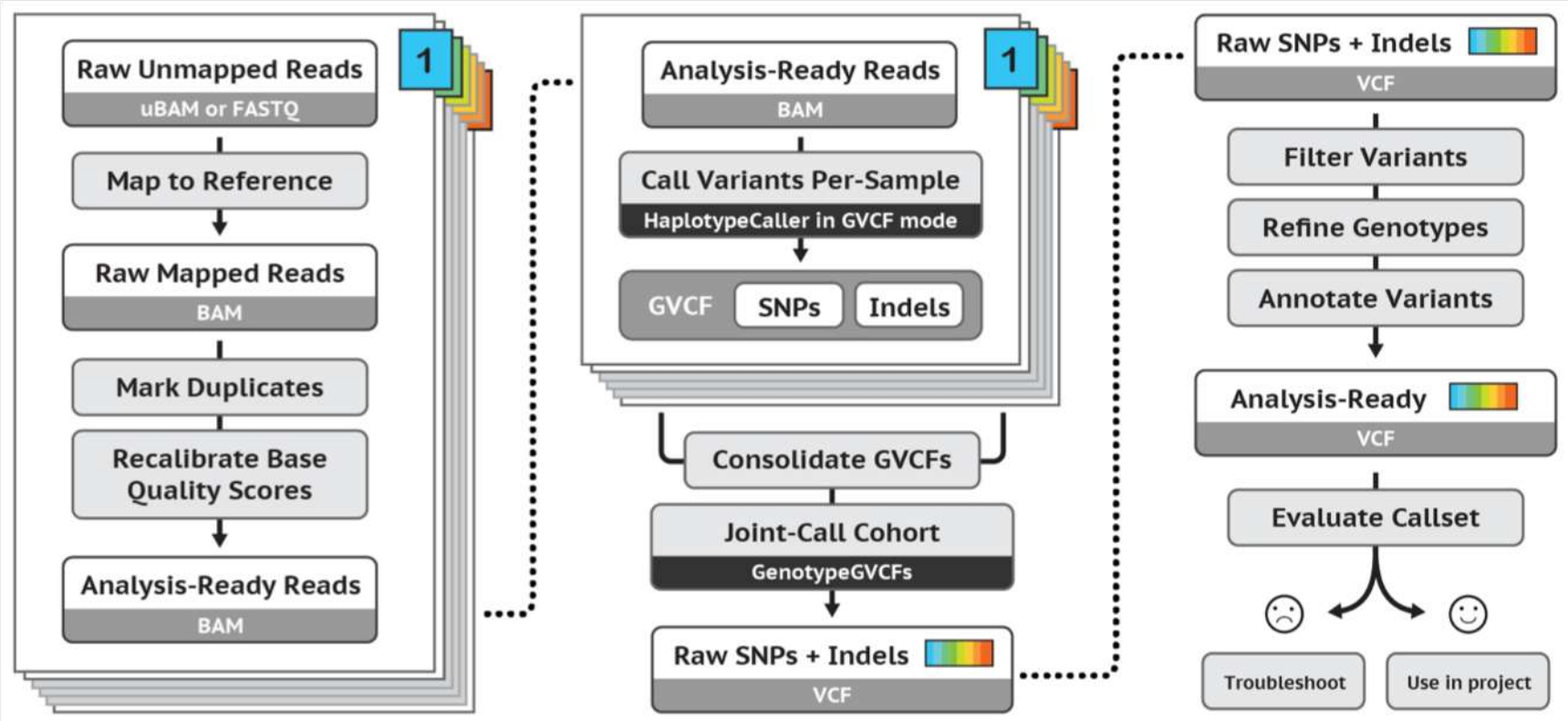
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# GATK's best practices for germline short variant discovery

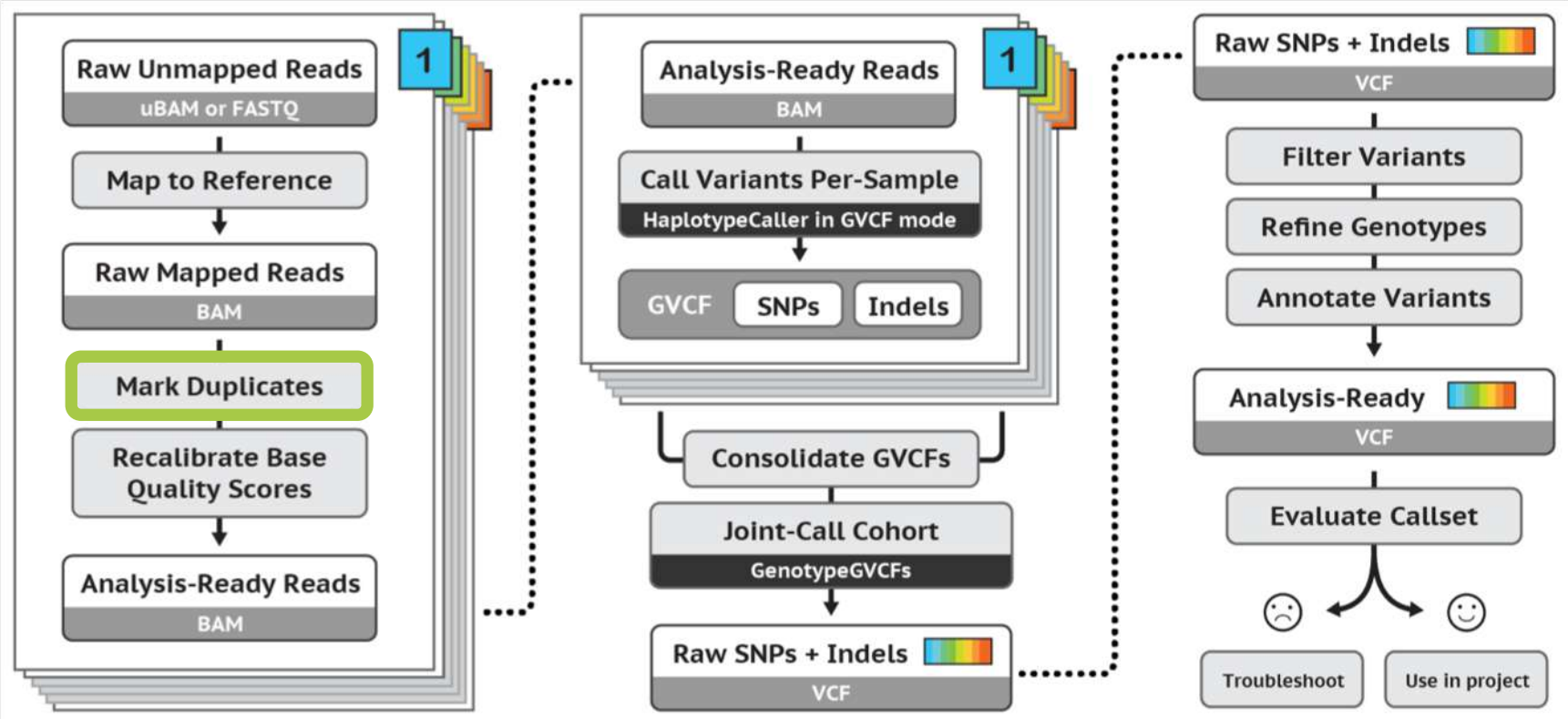
https://gatk.broadinstitute.org

# GATK's best practices workflow for germline short variant discovery



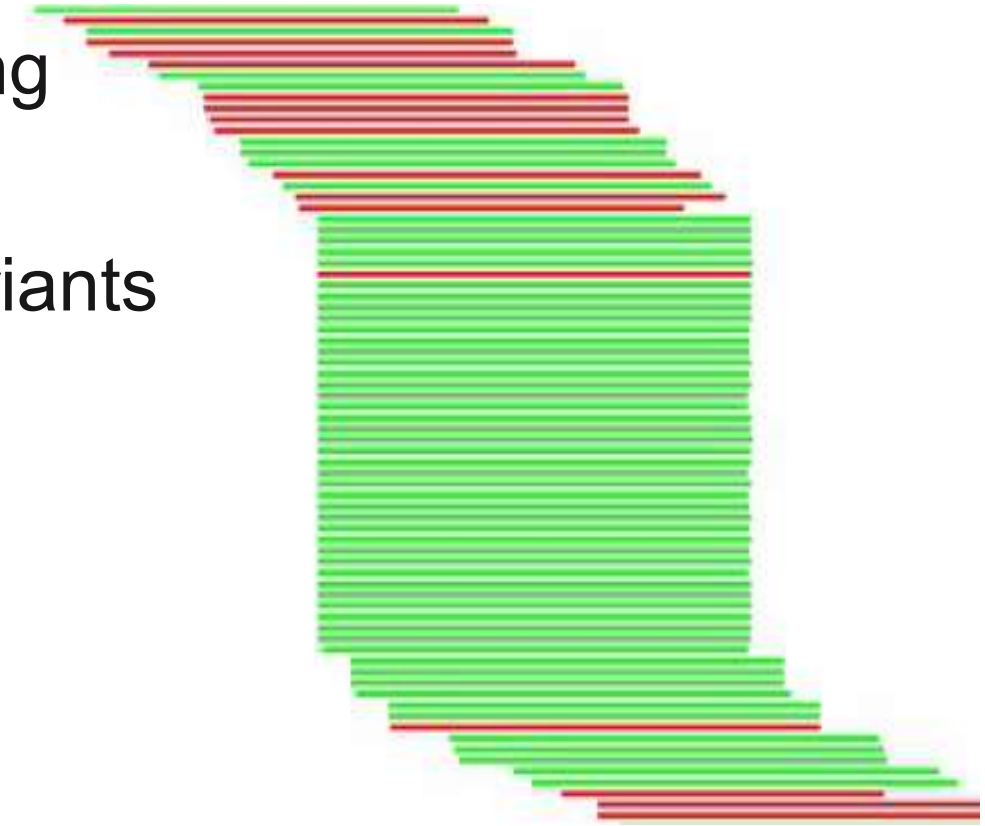

https://software.broadinstitute.org/gatk/best-practices/

# Mark Duplicates



# Duplicate reads

- PCR duplicates - library preparation
- Optical duplicates - sequencing
- Don’t add unique information
- Gives false allelic ratios of variants
- Should be removed/marked






# MarkDuplicates (Picard) Follow

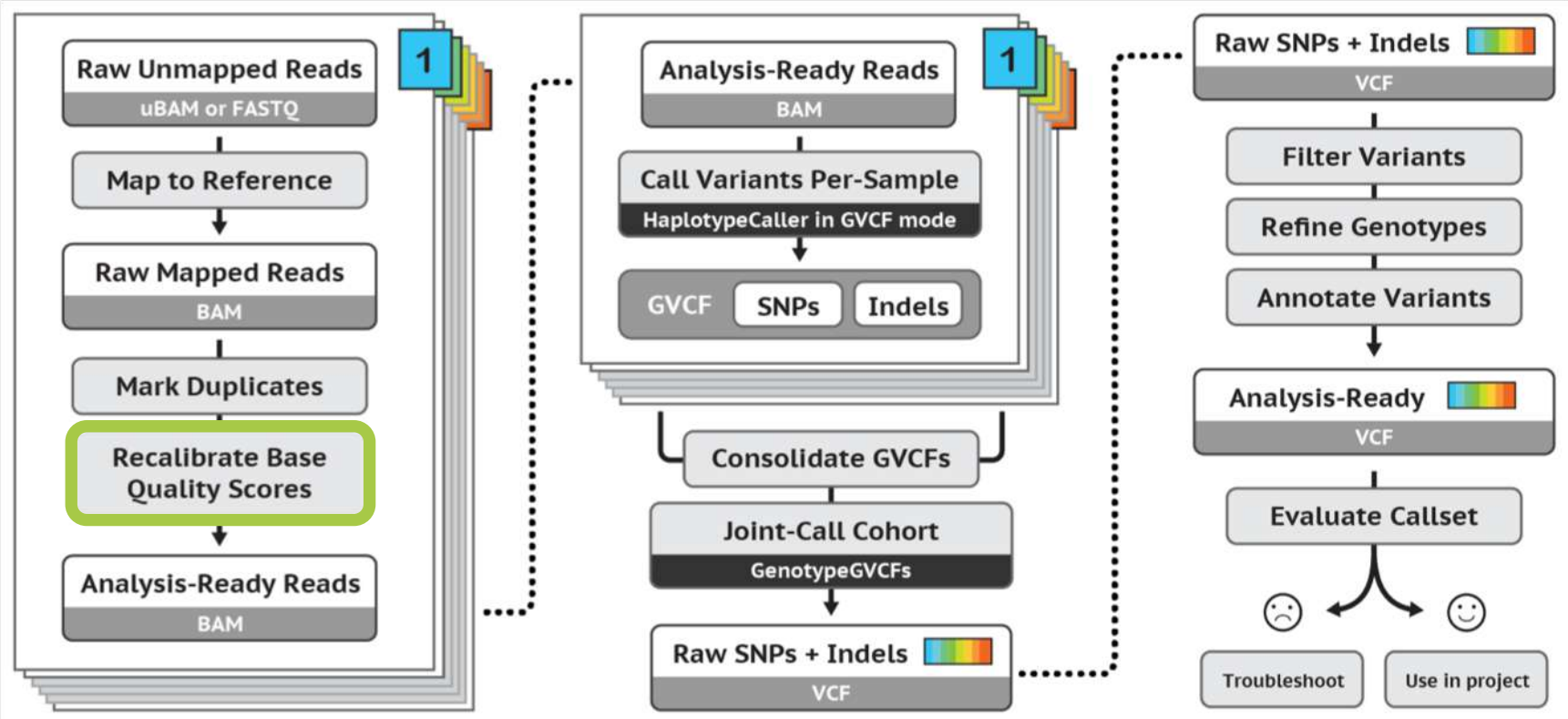
GATK Team
10 months ago · Updated

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
```

# Base Quality Score Recalibration (BQSR)



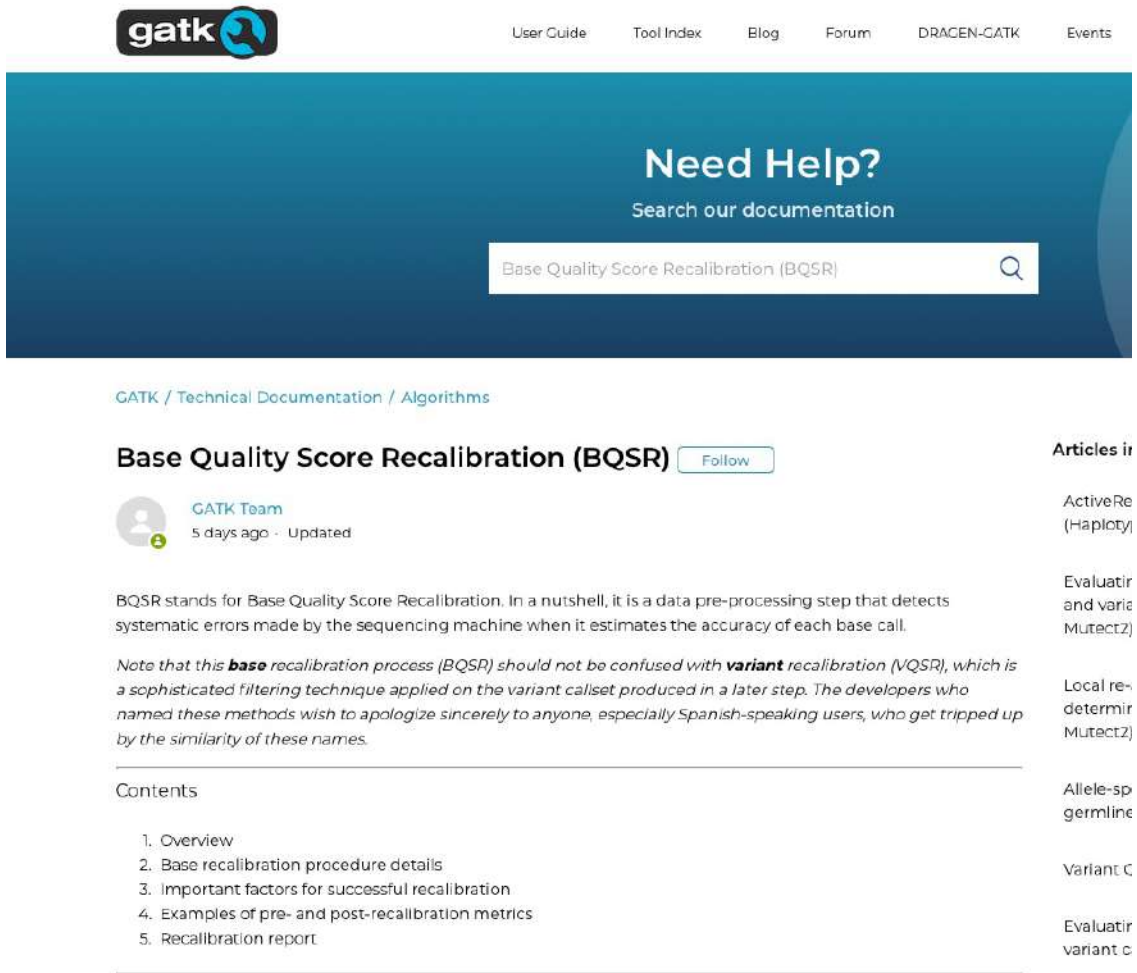
# Base Quality Score Recalibration (BQSR)

- During base calling, the sequencer estimates a quality score for each base. This is the quality scores present in the fastq files.

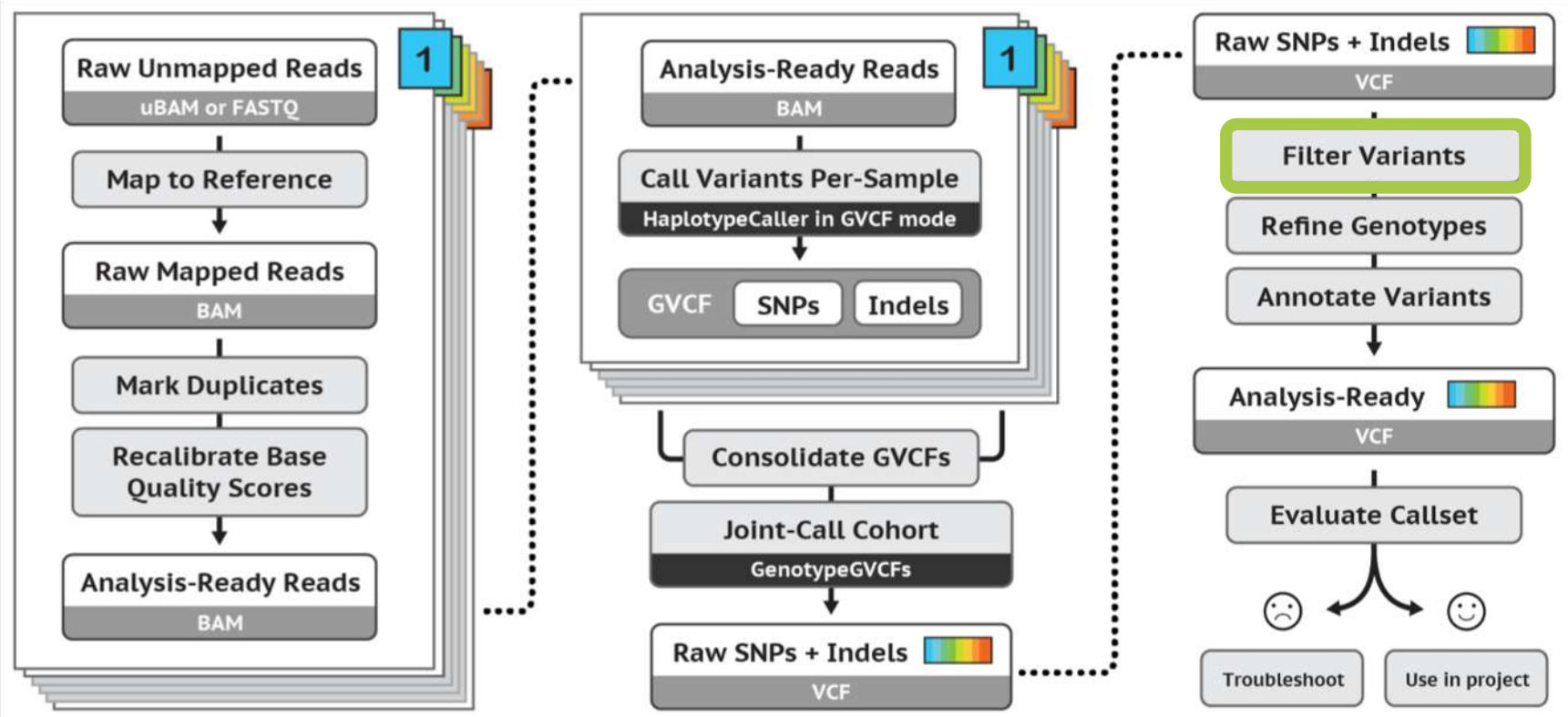
- Systematic (non-random) errors in the base quality score estimation can occur.
  - due to the physics or chemistry of the sequencing reaction
  - manufacturing flaws in the equipment
  - etc

- Can cause bias in variant calling

- **Base Qualtiy Score Recalibration** helps to calibrate the scores so that they correspond to the real per-base sequencing error rate (phred scores)


gatk
User Guide
Tool Index
Blog
Forum
DRAGEN-GATK
Events
Need Help?
Search our documentation
Base Quality Score Recalibration (BQSR)
GATK / Technical Documentation / Algorithms
Base Quality Score Recalibration (BQSR)
Follow
GATK Team
5 days ago · Updated
BQSR stands for Base Quality Score Recalibration. In a nutshell, it is a data pre-processing step that detects systematic errors made by the sequencing machine when it estimates the accuracy of each base call.
Note that this base recalibration process (BQSR) should not be confused with variant recalibration (VQSR), which is a sophisticated filtering technique applied on the variant callset produced in a later step. The developers who named these methods wish to apologize sincerely to anyone, especially Spanish-speaking users, who get tripped up by the similarity of these names.
Contents
1. Overview
2. Base recalibration procedure details
3. Important factors for successful recalibration
4. Examples of pre- and post-recalibration metrics
5. Recalibration report
Articles i
ActiveRe
(Haploty
Evaluatir
and varia
Mutect2)
Local re-
determir
Mutect2)
Allele-sp
germline
Variant C
Evaluatir
variant c

# Filter variants




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

Germline short variant discovery (SNPs + Indels)

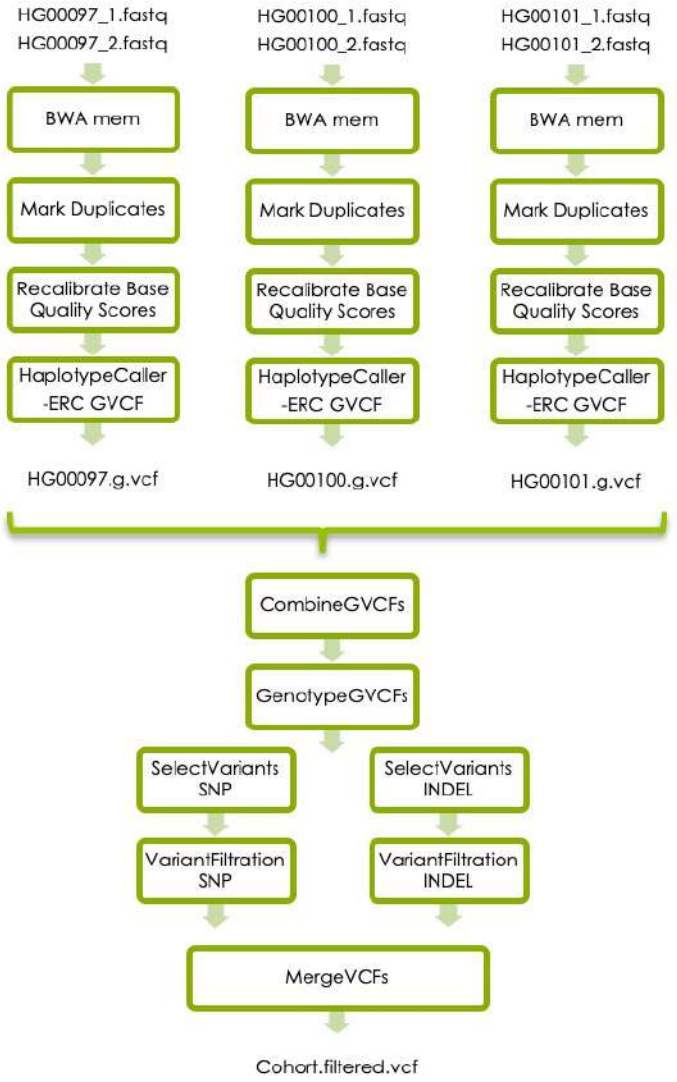
# Filtering

- Remove low quality variants

- Variant quality score recalibration (VQSR):
  - For large data sets ( >1 WGS or >30WES samples)
  - GATK has a machine learning algorithm that can be trained to recognise ”likely false” variants
  - **We do recommend to use VQSR when possible!**

- Hard filters:
  - For smaller data sets
  - Hard filters on information in the VCF file
  - For example: Flag variants with ”QD < 2” and ”MQ< 40.0”
  - GATK discussion on hard filters: https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set

# GATK's best practices workflow




More details and links to GATK for each step is found in the lab instructions.

# Today's lab

# 1000 Genomes data

- Low coverage WGS data
- 3 samples
- Small region on chromosome 2

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

# The Lactase enzyme

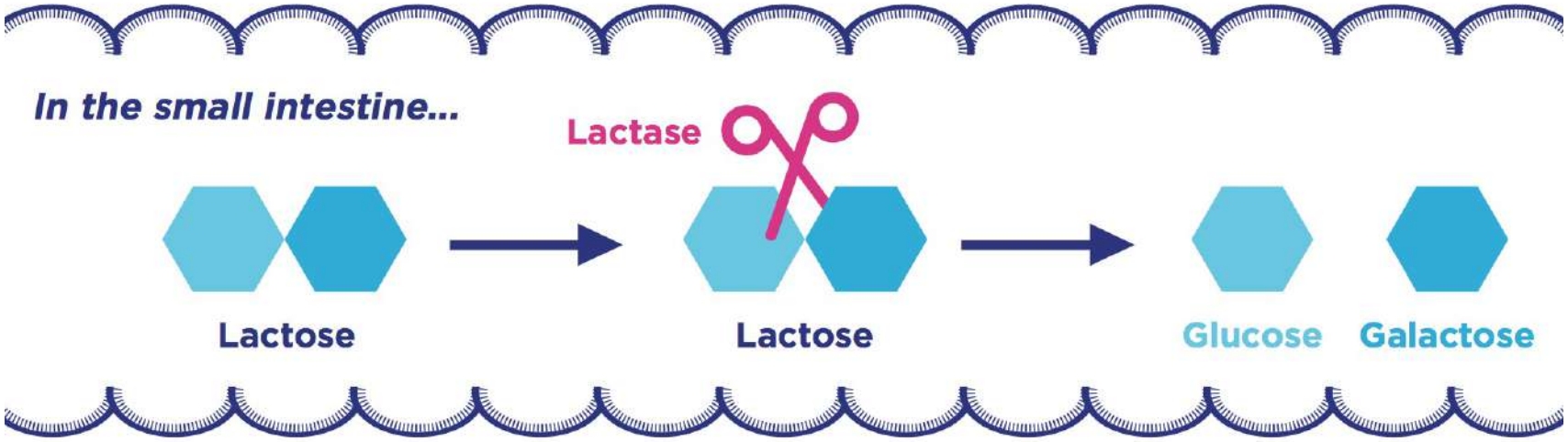



Figure 2. Lactose digestion in the intestine.

- 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)

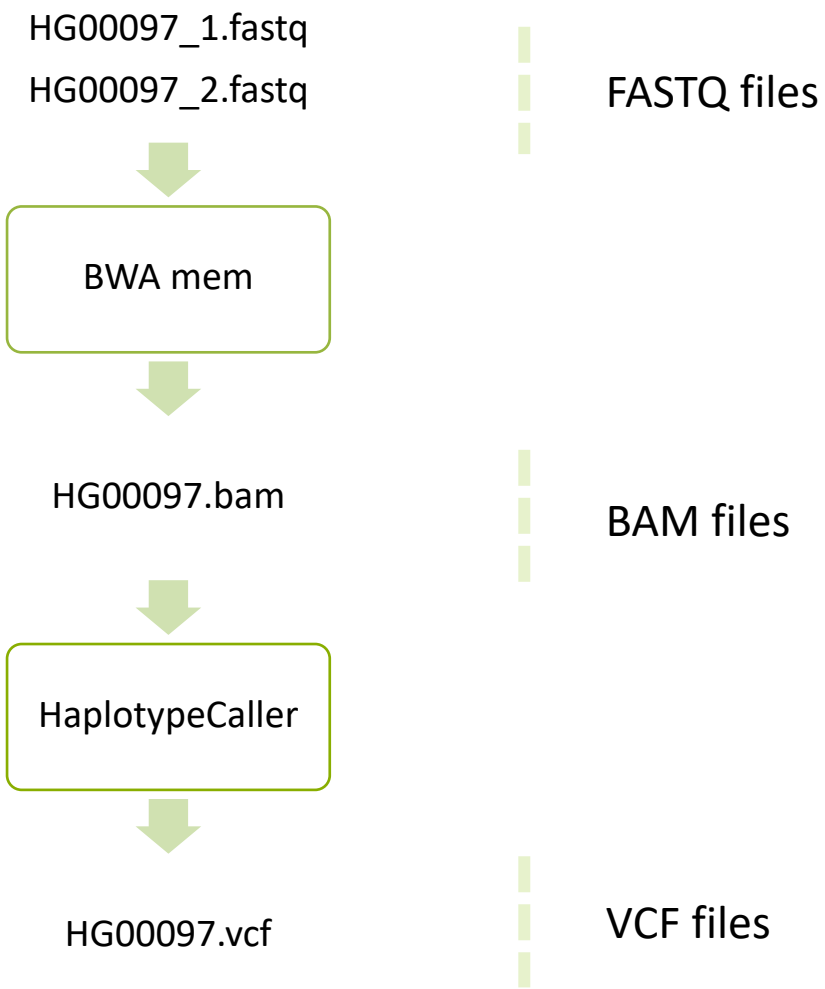
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# part one:

# variant calling in one sample

# Basic variant calling in one sample



# Part two (if you have time):

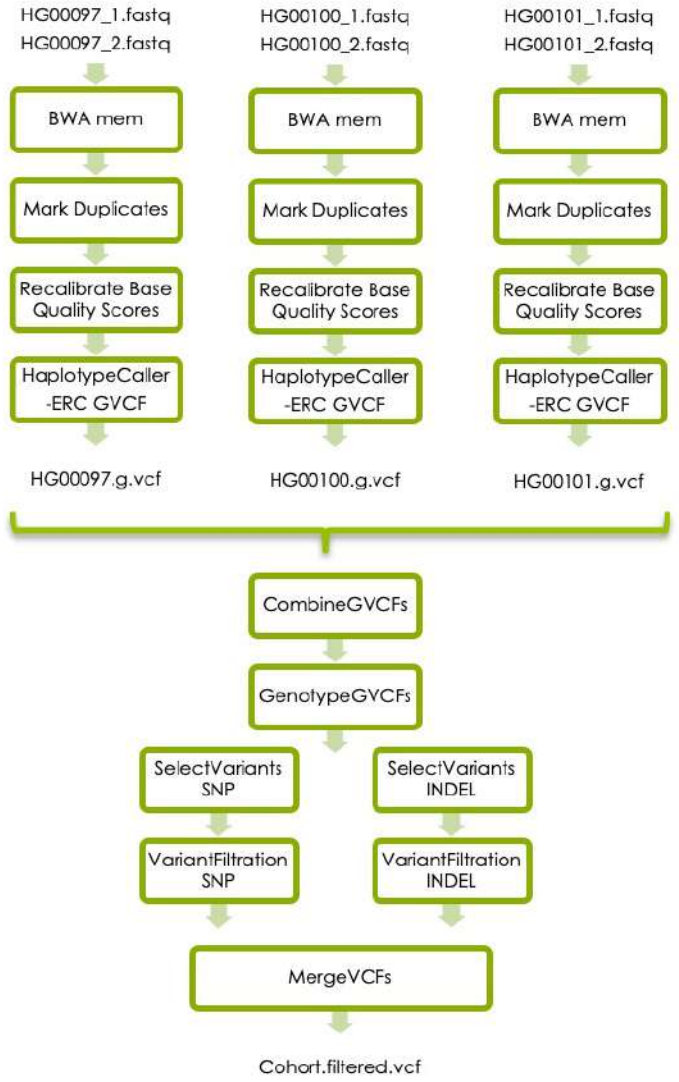
# variant calling in cohort

# Joint variant calling workflow

# Part three (if you have time):

# Follow GATK best practices for short variant discovery

# GATK's best practises




First look at video about this linked from schedule!

https://gatk.broadinstitute.org

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Questions?