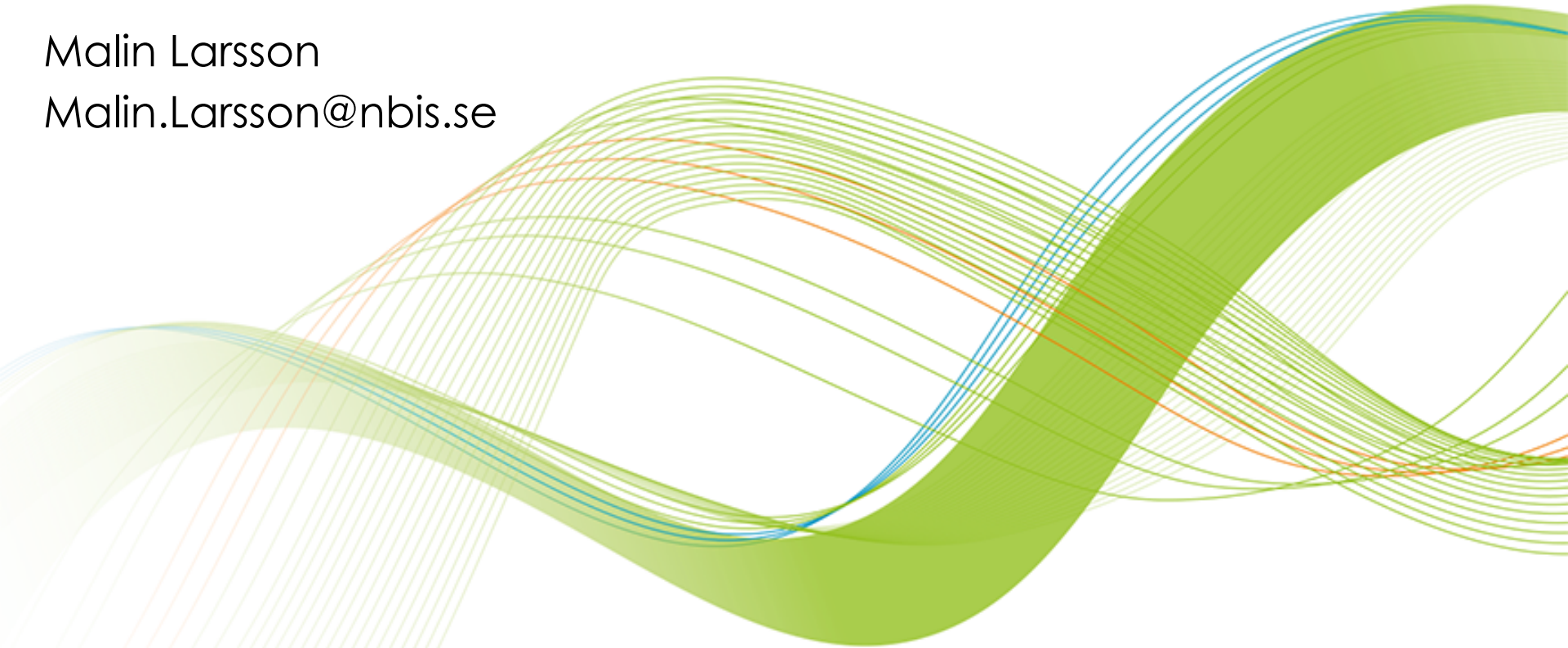
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Variant Calling Workflows

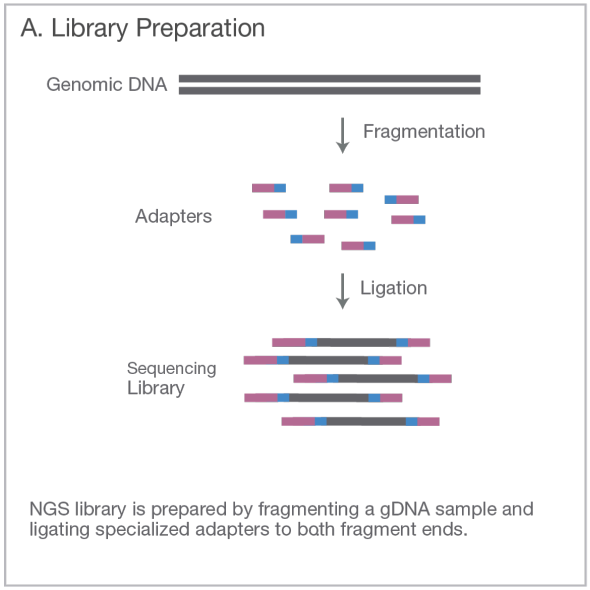
Malin Larsson
Malin.Larsson@nbis.se

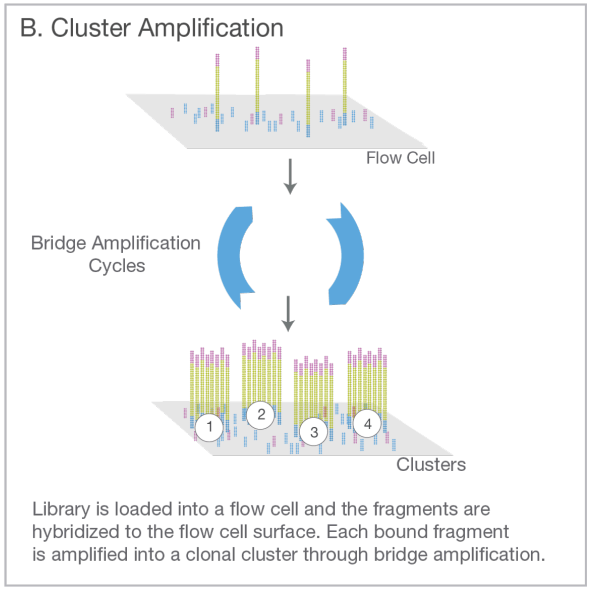


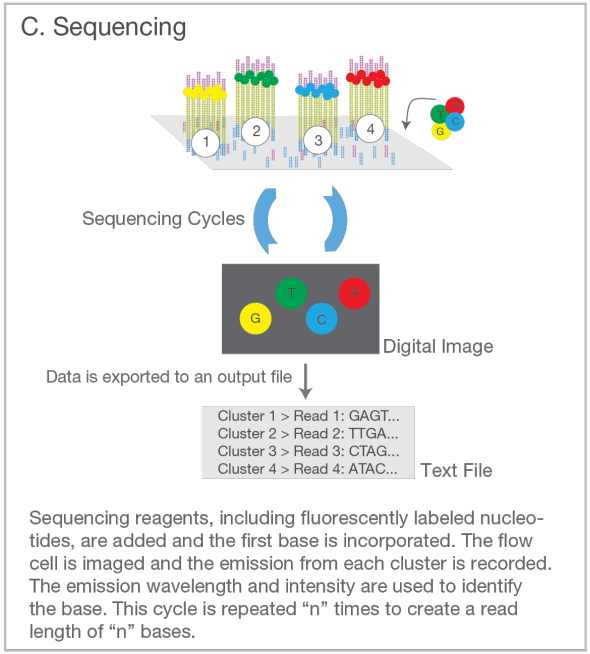
# Talk Overview

- The reference genome
- Genetic variation
- Workflows
- Basic variant calling in one sample
- Basic variant calling in cohort
- GATK Best practices

# Illumina Sequencing



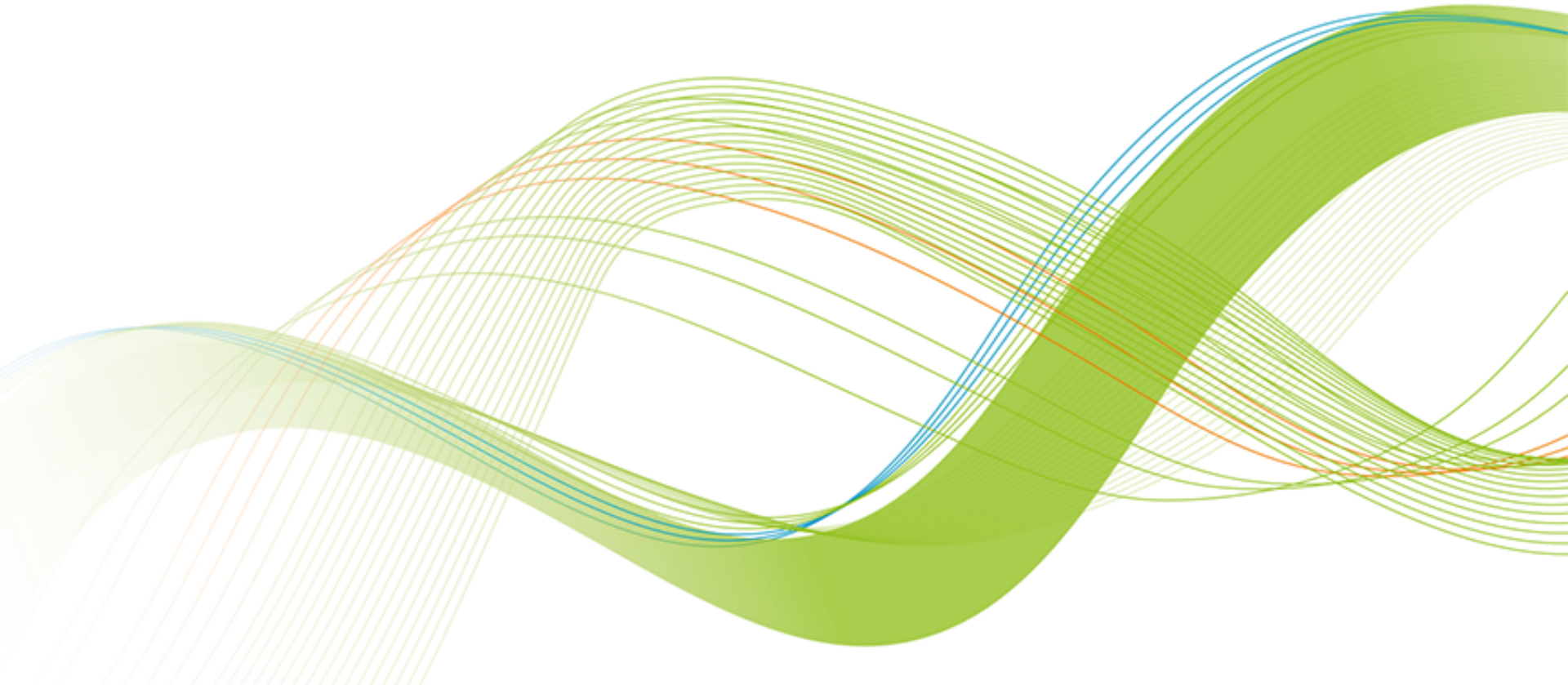


https://www.youtube.com/watch?v=fCd6B5HRaZ8

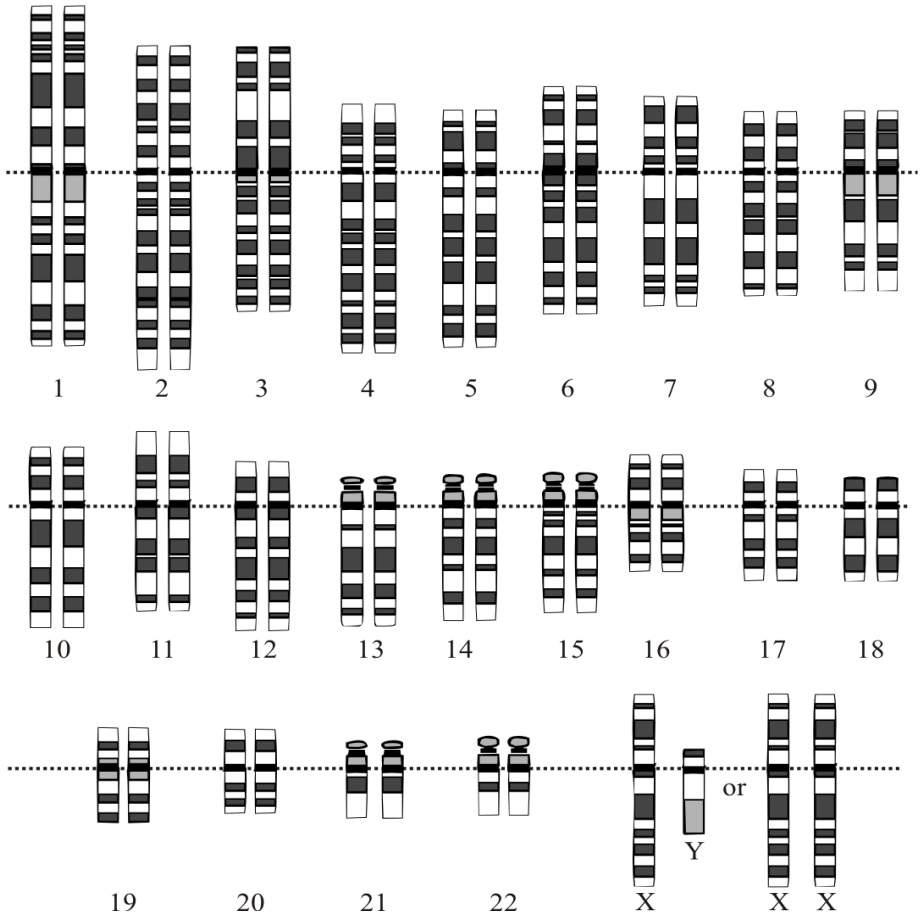
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# The reference genome sequence



# Each chromosome…



# …represented by a sequence

```
>chr1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTG
GGGGGTGTGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTC
TTTGATTCCTGCCTCATTCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACCTAC
TAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGCACAGCCGC
TTTCCACACAGACATCATAACAAAAAATTTCCACCAAACCCCCCCCTCCCCCCGCTTCTGGCCACA
GCACTTAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTC
AAATTTTATCTTTAGGCGGTATGCACTTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCT
CCCACTCCCATACTACTAATCTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCG
CTGCTAACCCCATACCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACC
TCCTCAAAGCAATACACTGAAAATGTTTAGACGGGCTCACATCACCCCATAAACAAATAGGTTTGG
TCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGTTCAC
CCTCTAAATCACCACGATCAAAAGAGGCGGTATGCACTTTTAACAGTCACCCCCAGGCGGTATGCA
```

# The reference genome


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

In 2001: The International Human Genome Sequencing Consortium published the first draft of the human genome sequence.
It contained 150,000 gaps.

HG19: 250 gaps

HG38 is the latest version of the human reference genome, but we will work with HG19.

# Keep track of the Reference version

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

- mapping
- visualizing
- variant calling
- annotation
- etc

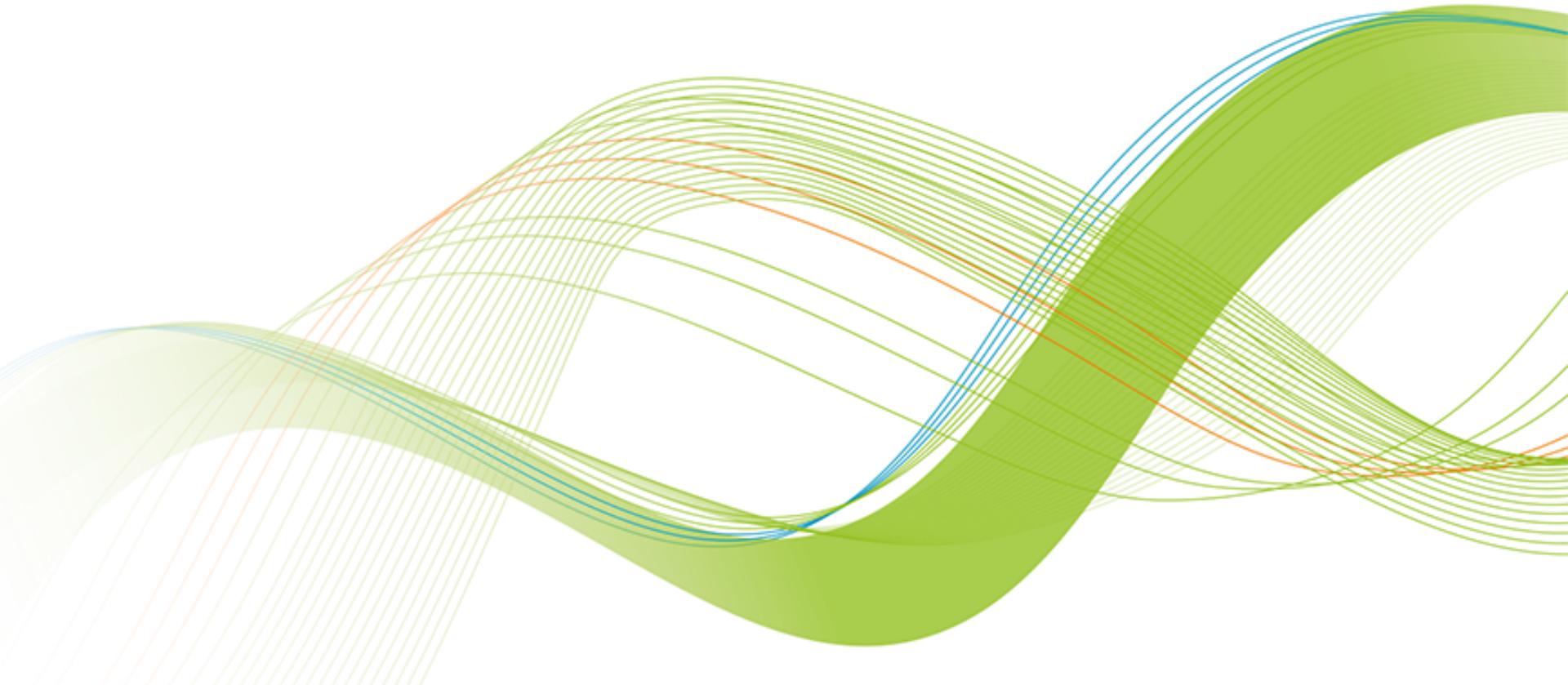
You must keep track of which version of the reference genome your data was mapped to.

The same reference sequence must be used in all downstream analyses.

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

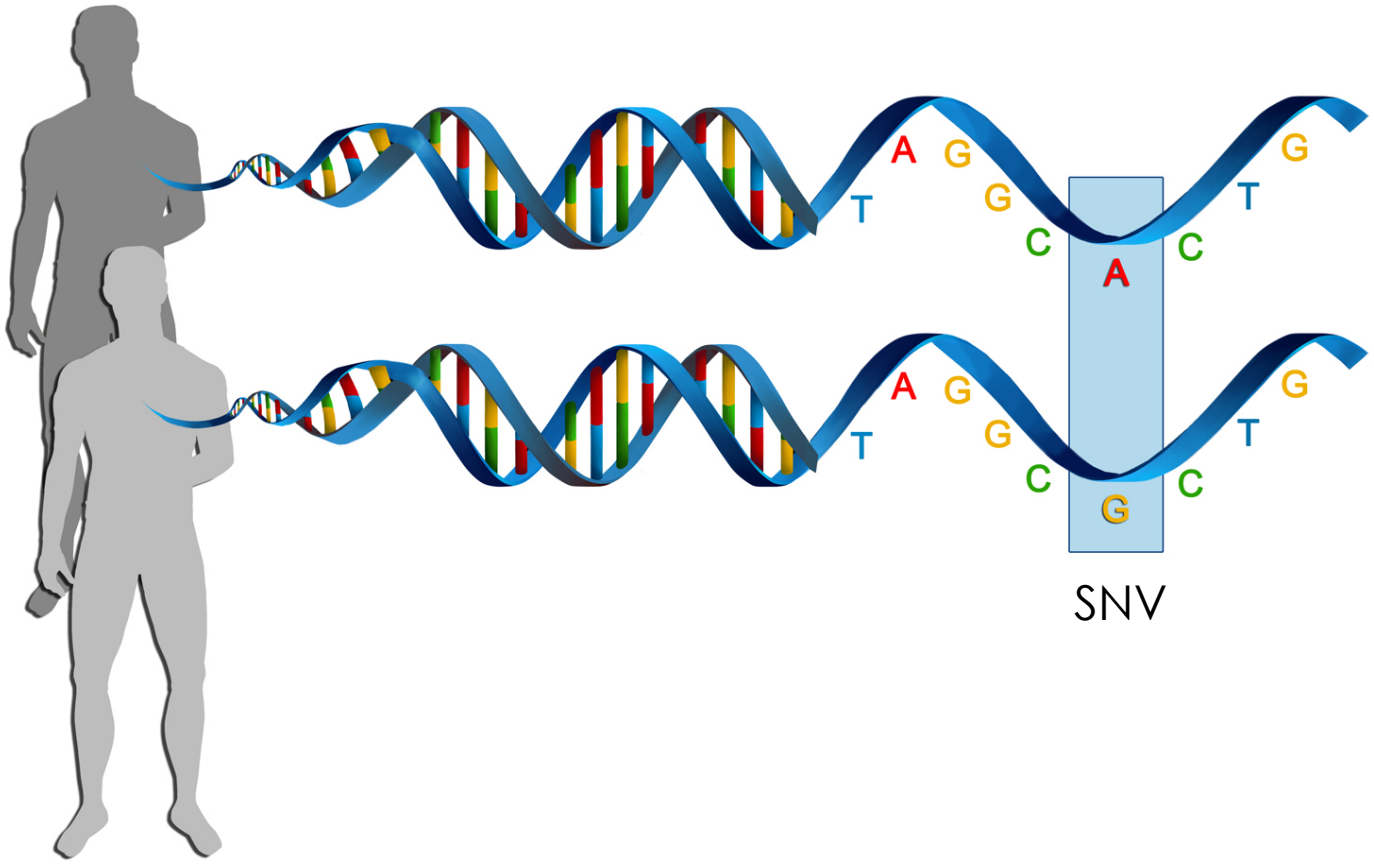
# Genetic variation



# Genetic Variation

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

# Single Nucleotide Variants (SNVs)



# Reference- and Alternatve Alleles

GGCTTTTCCAACAGGTATATCTTCCCCGCTAGCTAGCTAGCTACTTCAAAT

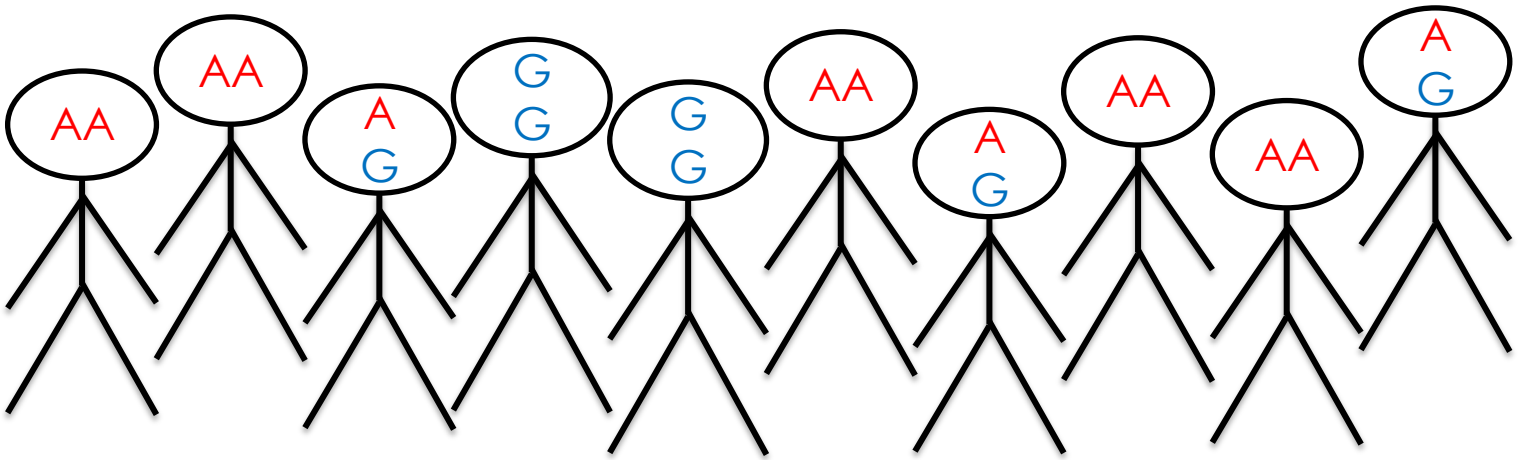
**Reference allele** AGCTAGCTA

**Alternative allele** AGCTGGCTA

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

# Allele frequency in a population

AACAGGTATATCTTCCCCGCTAGCT**A**GCTAGCTACTTCCTTAGGGACTGTA
AGCT**G**GCTA

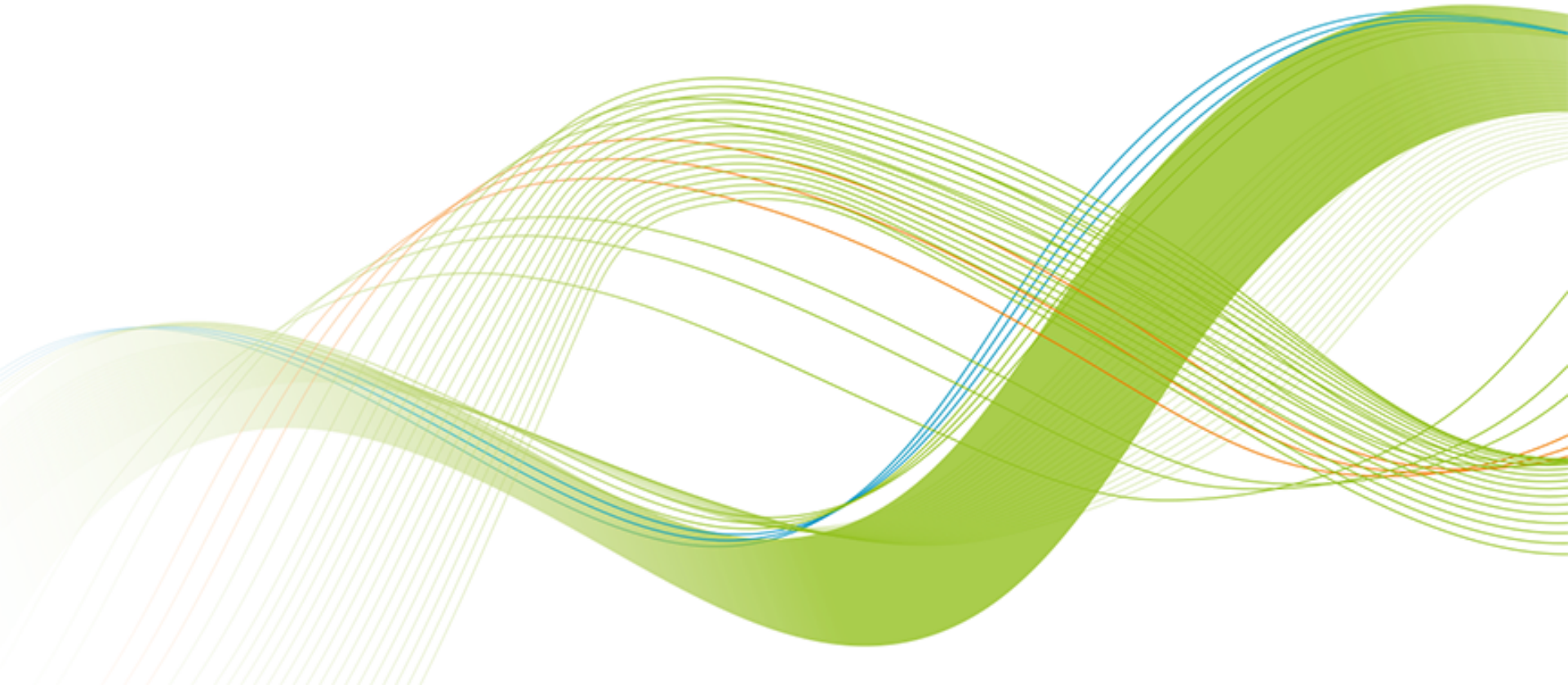



| Genotypes: | AA | AG | GG |
|---|---|---|---|
| Genotype frequency: | 5/10 | 3/10 | 2/10 |
| Frequency of allele A: | 13/20 = 0.65 | | |
| Frequency of allele G: | 7/20 = 0.35 | | |

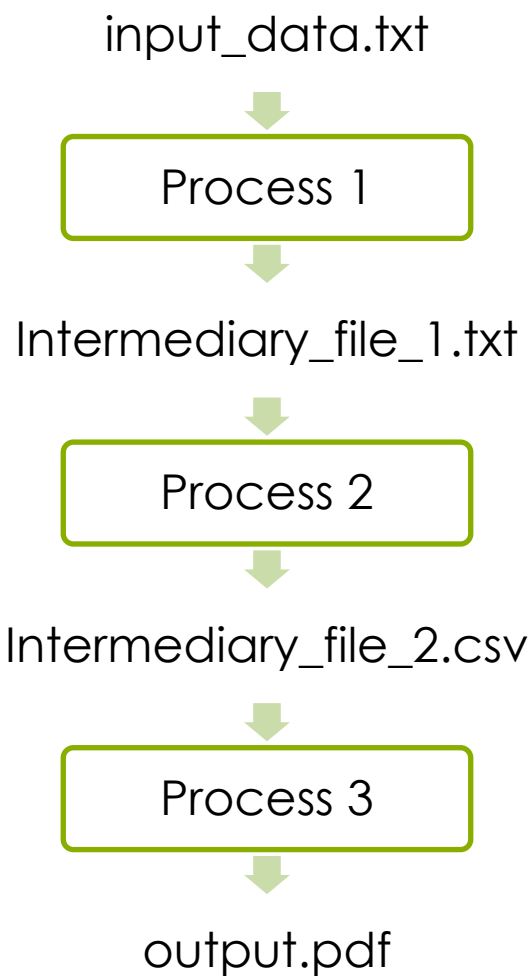
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Introduciton to workflows



# A bioinformatics workflow



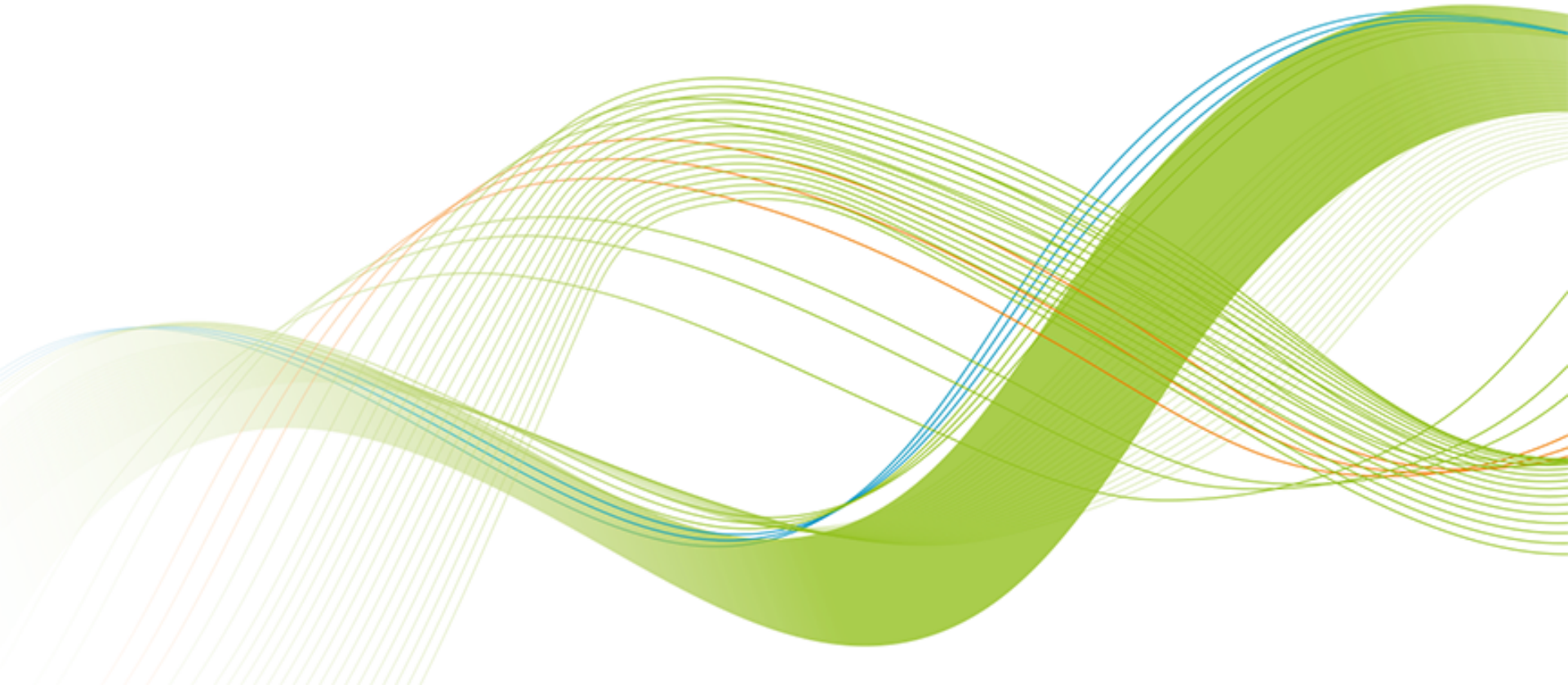
# Workflow conventions

- Each process has an input and output file(s)
- Create a new output file in each process – don't owerwrite the input file
- Use informative file names
- Include information of the process in output file name

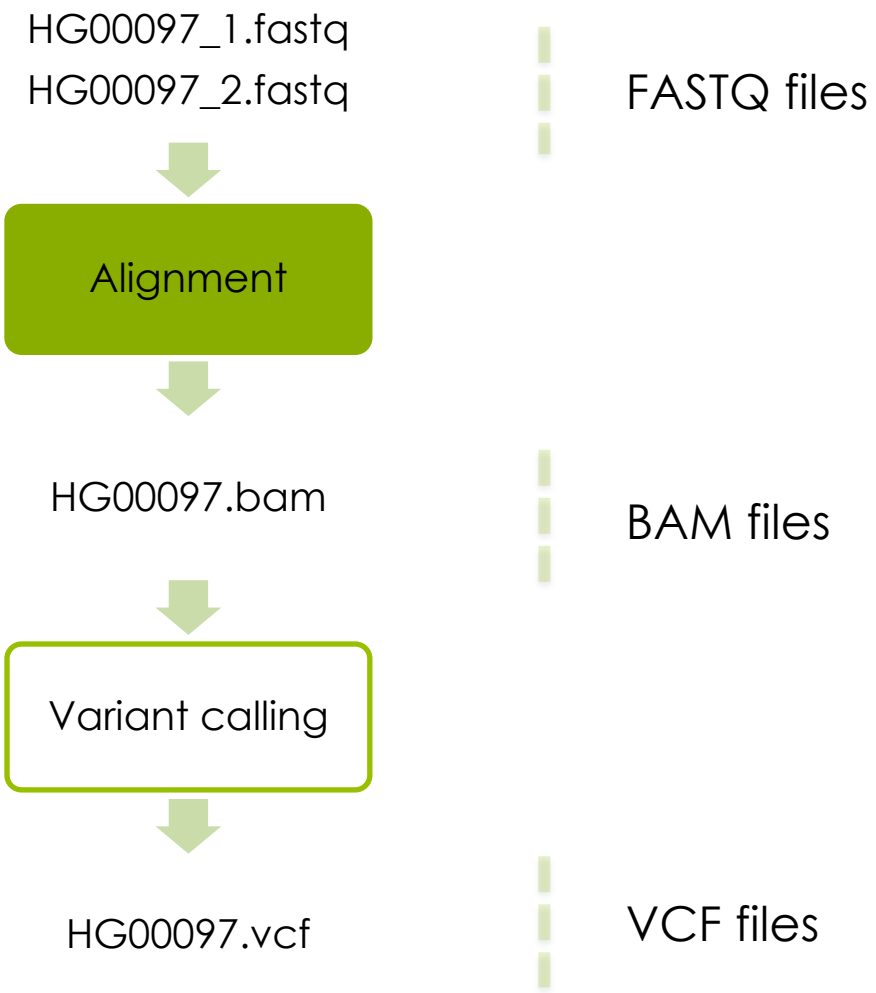
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Basic Variant Calling in one sample



# Basic variant calling in one sample



# Alignment

```
module load bwa
```

| | |
|---|---|
| Read | TCGATCC |
| Reference | GACCTCATCGATCCCACTG |

# Alignment

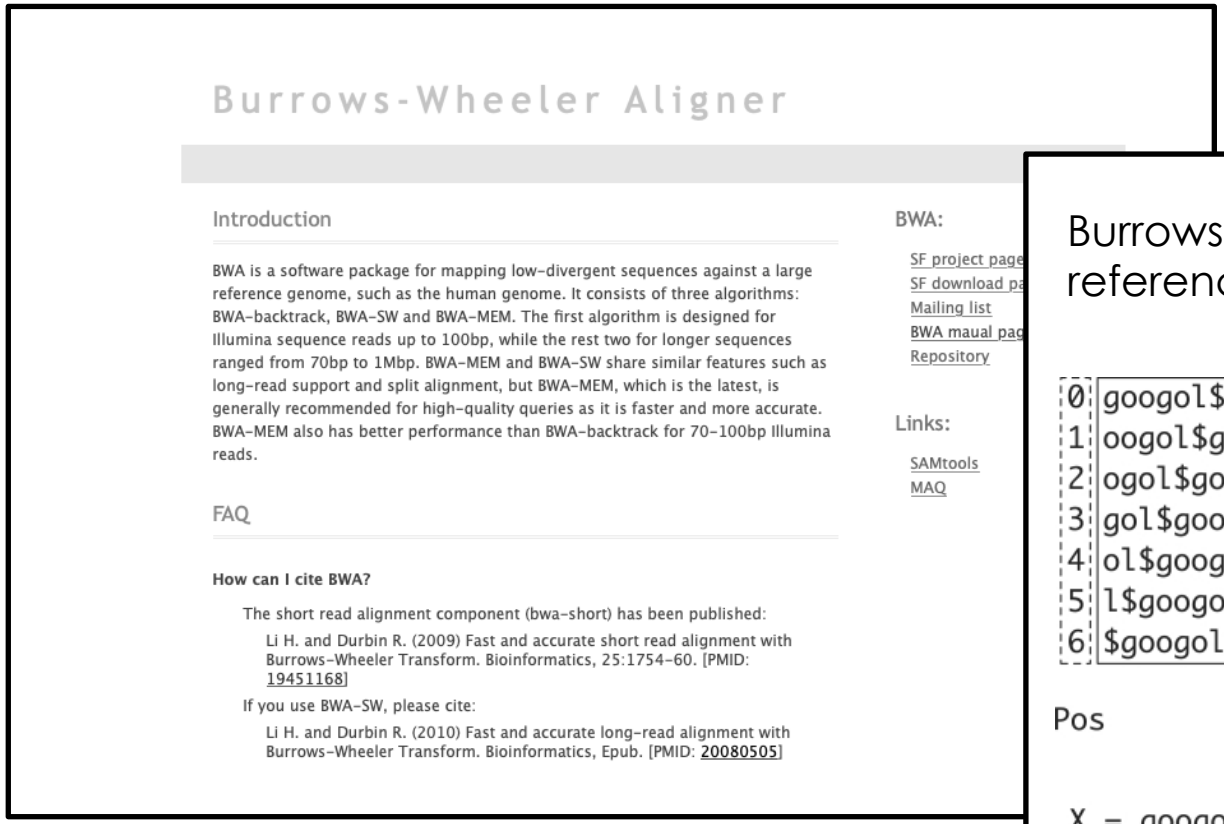
```
module load bwa
```

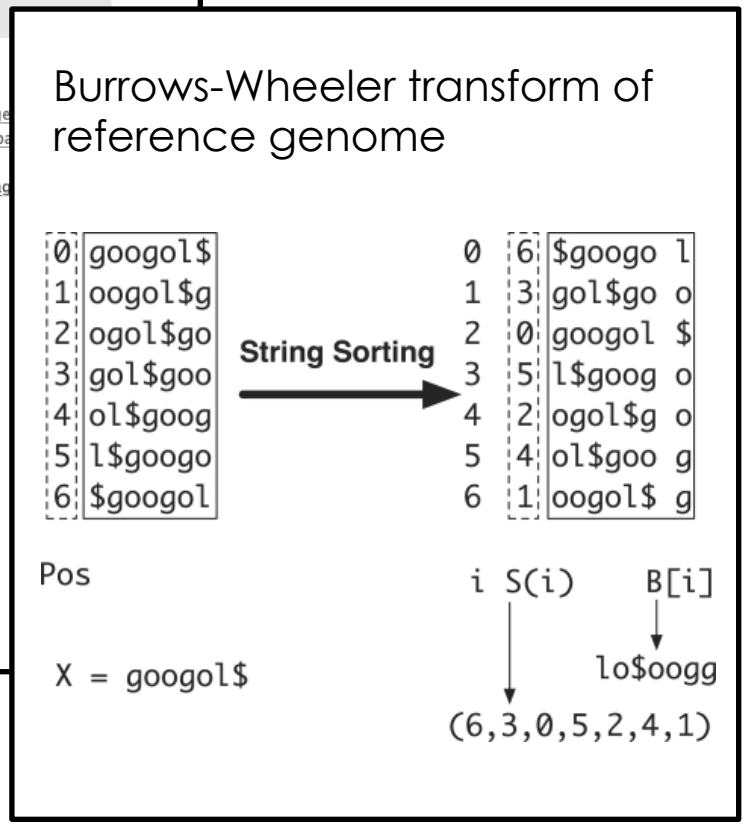
Read TCGATCC
Reference GACCTCATCGATCCCACTG

Read TCGATCC
Reference GACCTCATCGATCCCACTG

# Burrows-Wheeler Aligner

http://bio-bwa.sourceforge.net



# Alignment

```
module load bwa
```

# Alignment

```
module load bwa
```

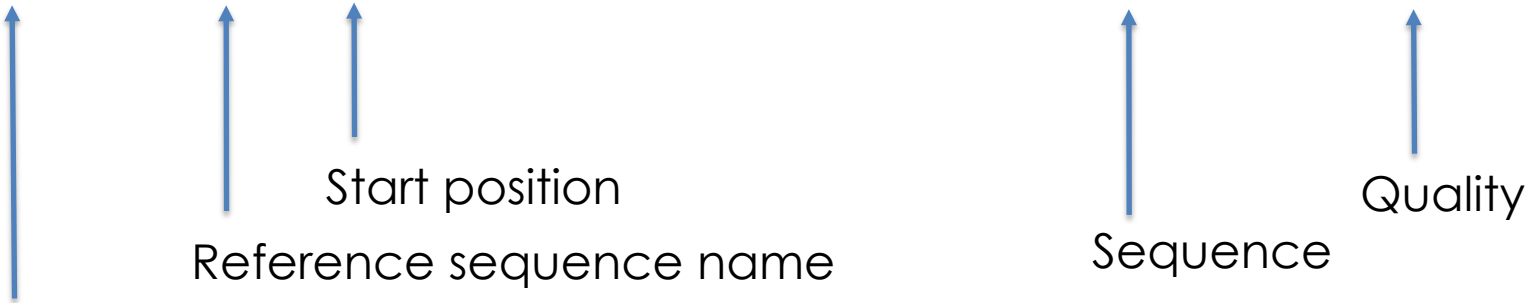
# Output from mapping - Sam format

## 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

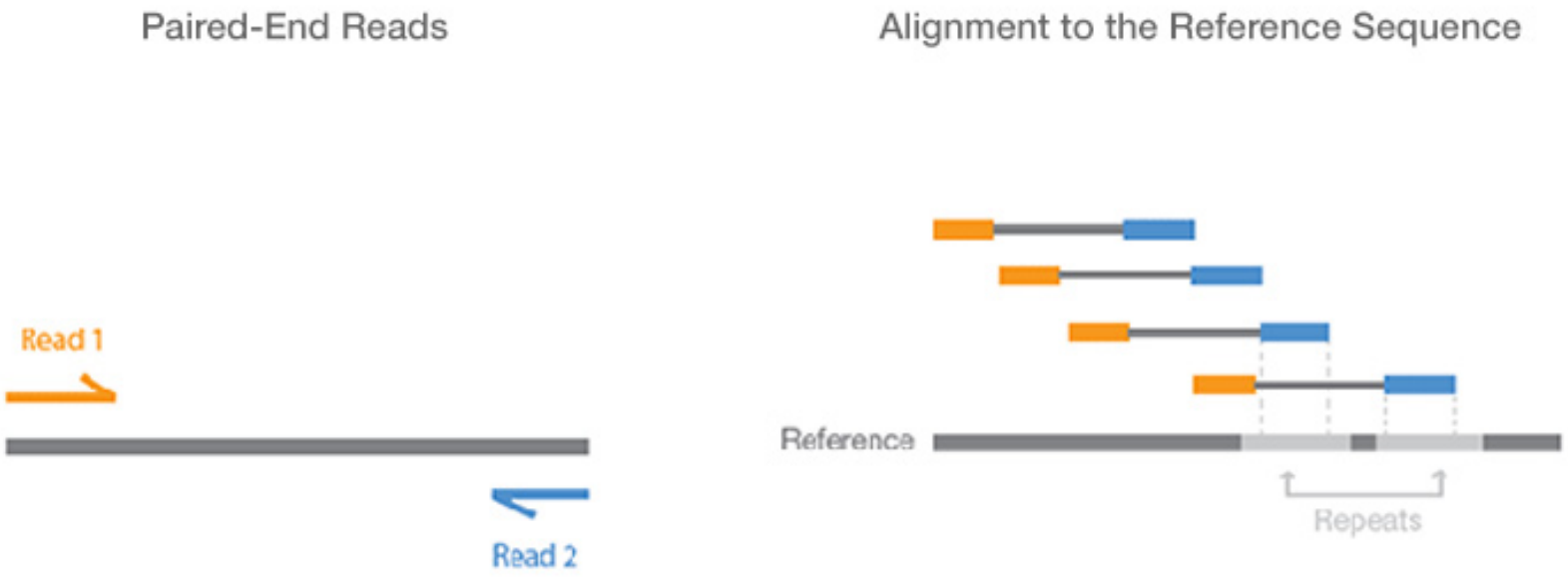
- Link *sample id*, *library prep*, *flowcell* and *sequencing run* to the reads.
- Good for error tracking!
- Often needed for variant calling
- Detailed description in tutorial or https://gatkforums.broadinstitute.org/gatk/discussion/6472/read-groups

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

# File Indices

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

# 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

The forward and reverse reads are stored in two fastq files.

The order of pairs and naming is identical, except the designation of forward and reverse.

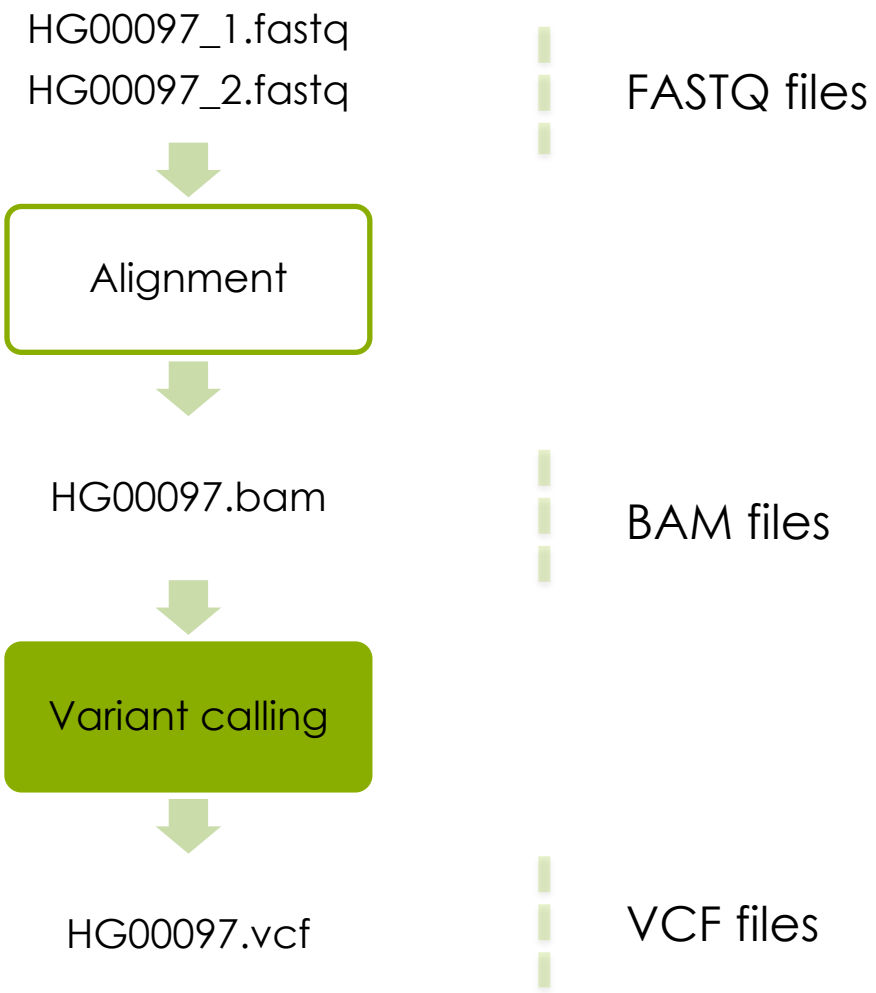
ID_R1_001.fastq

```
@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
JJJHFFFFFFDDDDDDDDDDDDDDDEDCCDDDD
```

# Basic variant calling in one sample



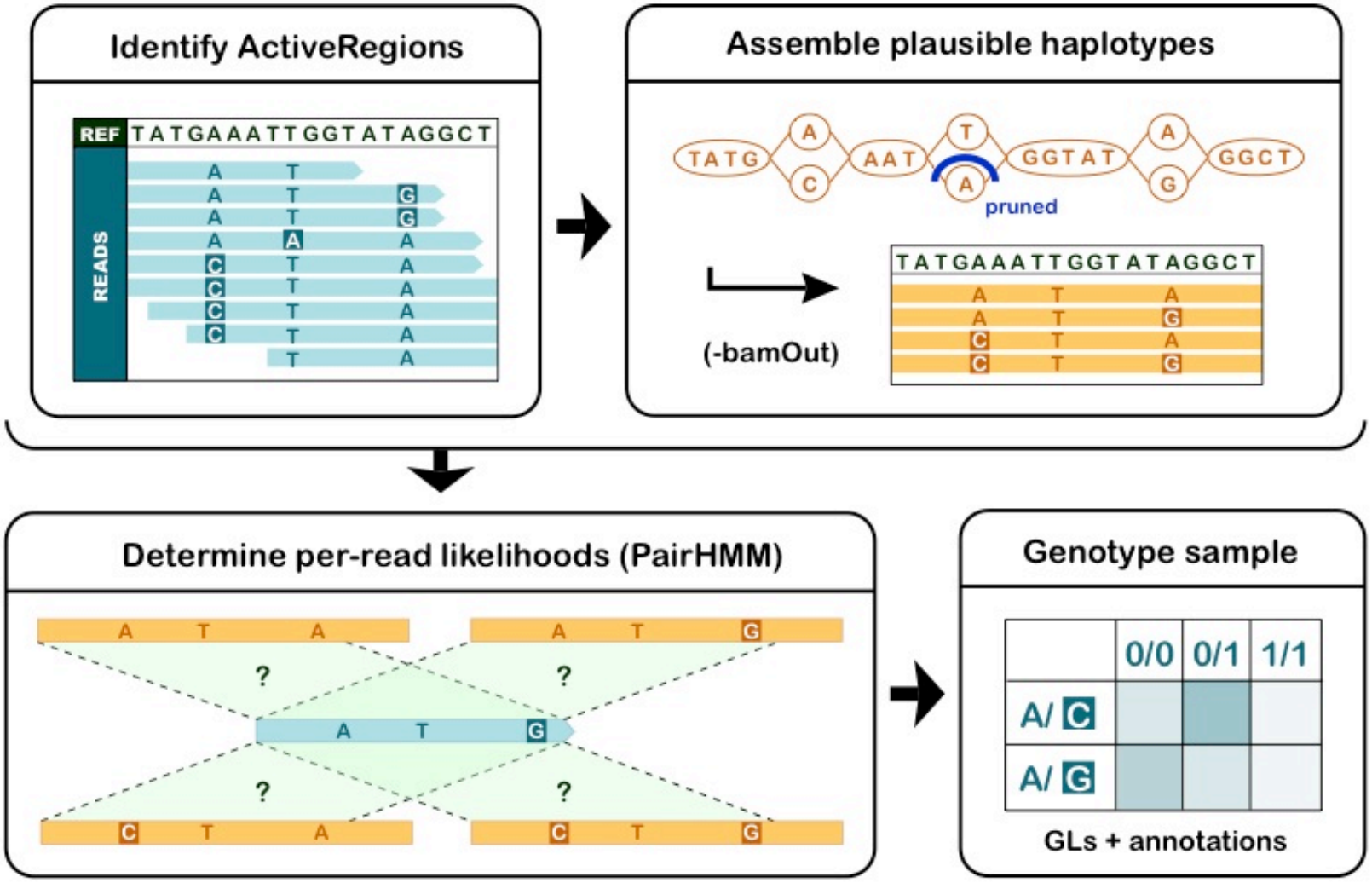
# Detecting variants in reads

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...

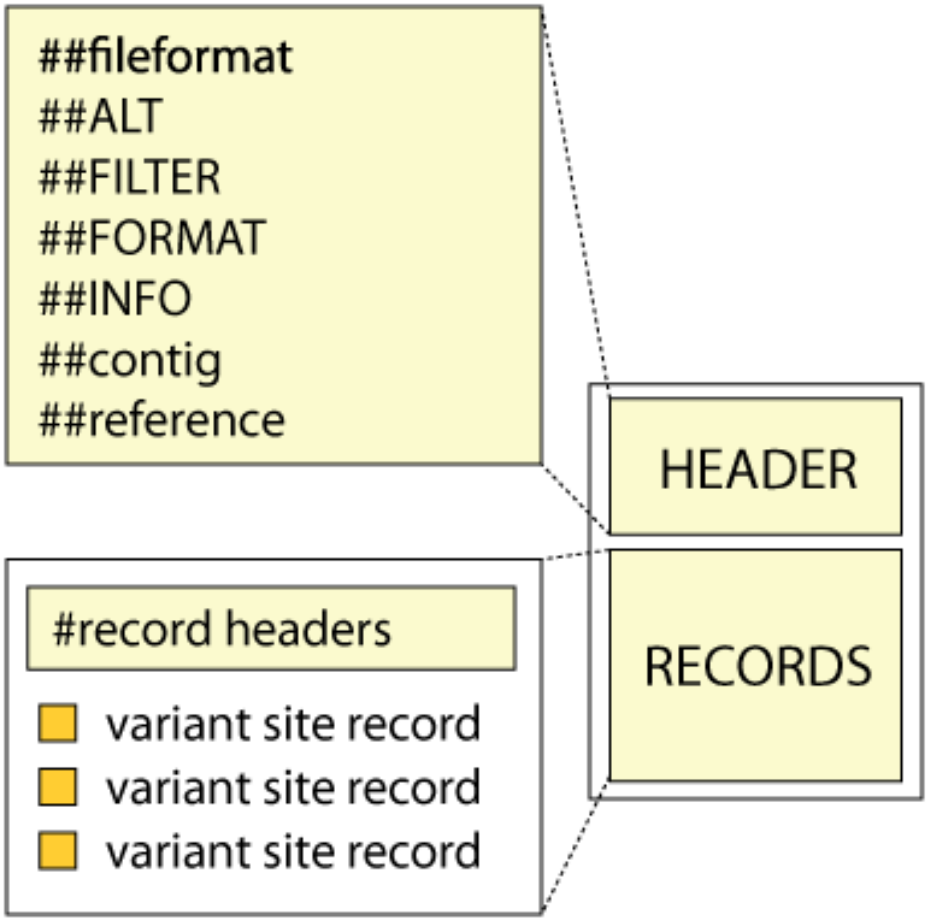
Sample:
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...

$$Allelic\ depths:$$
$$\#referece\ allels\ in\ a\ position$$
$$\#alternative\ allels\ in\ a\ position$$

# Variant Calling
# HaplotypeCaller



# Variant Call Format (VCF)



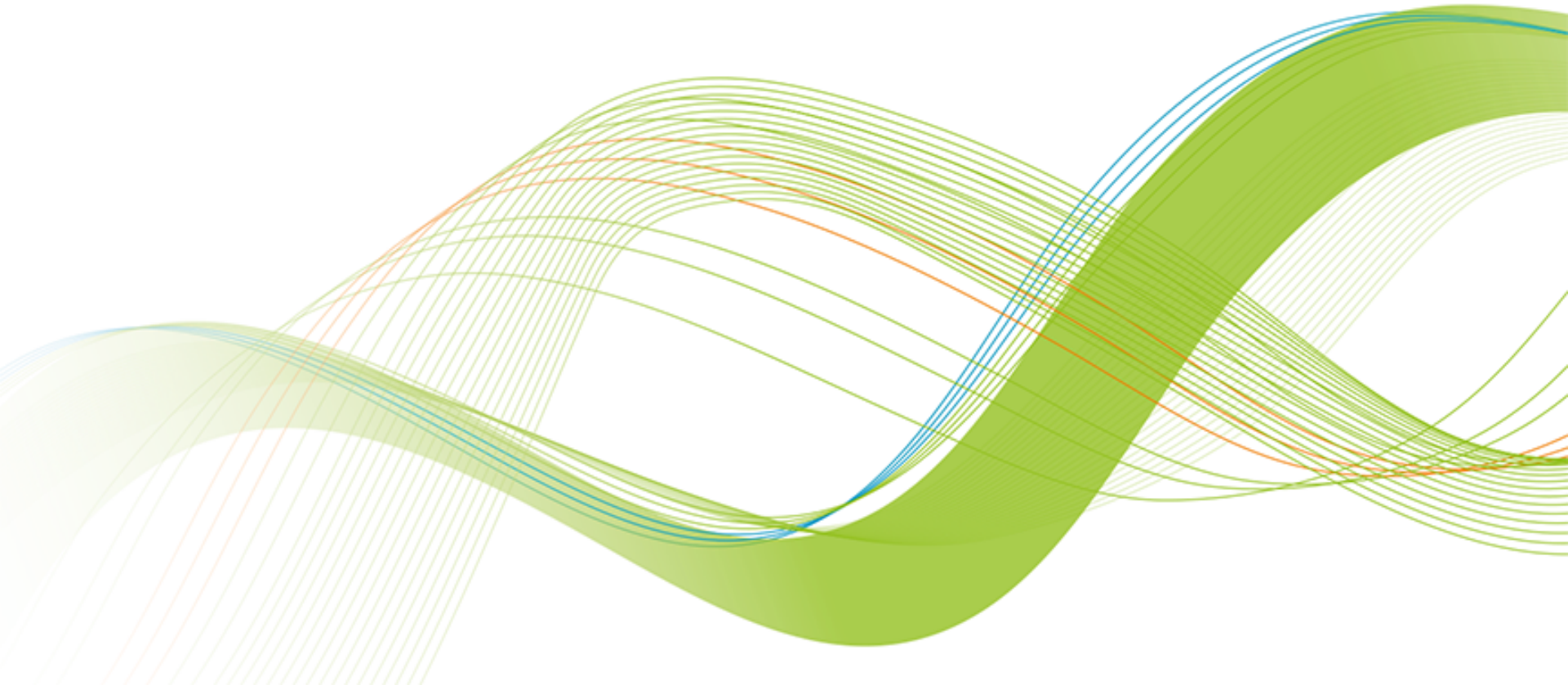
# Variant Call Format (VCF)

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens”...
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G   A      29   PASS   NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP  0|0:48:1
20     17330   .         T   A      3    q10    NS=3;DP=11;AF=0.017     GT:GQ:DP  0|0:49:3
20     1230237 .         T   .      47   PASS   NS=3;DP=13;AA=T         GT:GQ:DP  0|0:54:7
20     1234567 microsat1 GTC G,GTCT 50   PASS   NS=3;DP=9;AA=G          GT:GQ:DP  0|1:35:4
```

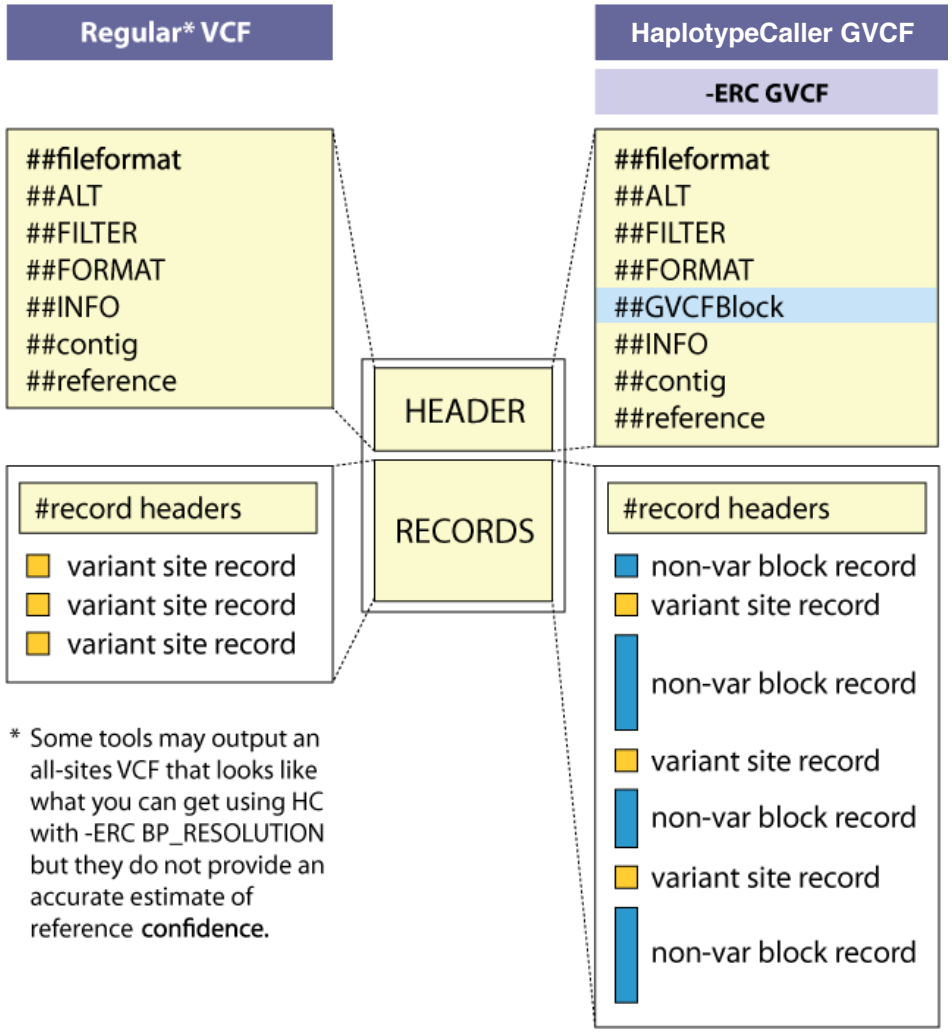
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# 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)

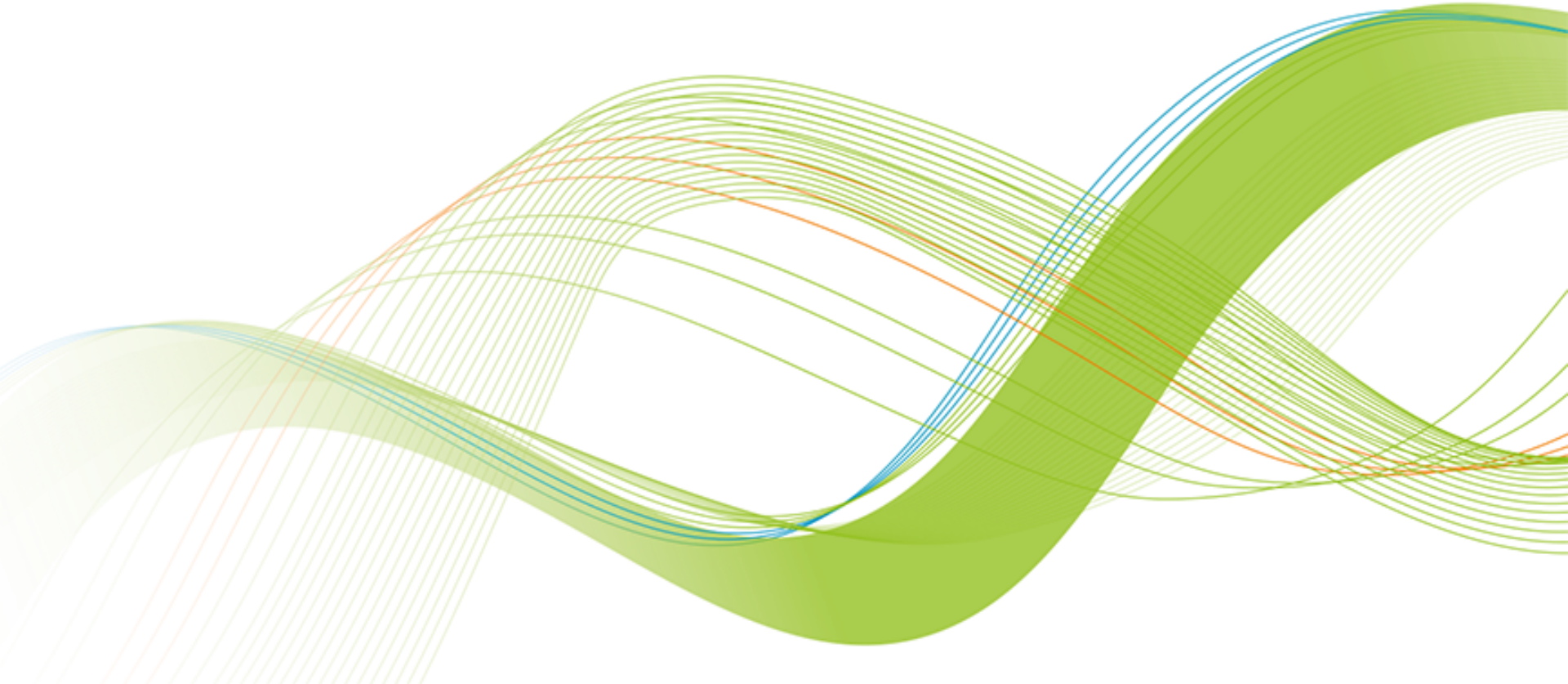
```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens”...
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
```

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | NA00001 | NA00002 | NA00003 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 14370 | rs6054257 | G | A | 29 | PASS | NS=3;DP=14;AF=0.5;DB;H2 | GT:GQ:DP | 0\|0:48:1 | 1\|0:48:8 | 1\|1:43:5 |
| 20 | 17330 | . | T | A | 3 | q10 | NS=3;DP=11;AF=0.017 | GT:GQ:DP | 0\|0:49:3 | 0\|1:3:5 | 0\|0:41:3 |
| 20 | 1230237 | . | T | . | 47 | PASS | NS=3;DP=13;AA=T | GT:GQ:DP | 0\|0:54:7 | 0\|0:48:4 | 0\|0:61:2 |
| 20 | 1234567 | microsat1 | GTC | G,GTCT | 50 | PASS | NS=3;DP=9;AA=G | GT:GQ:DP | 0\|1:35:4 | 0\|2:17:2 | 1\|1:40:3 |

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

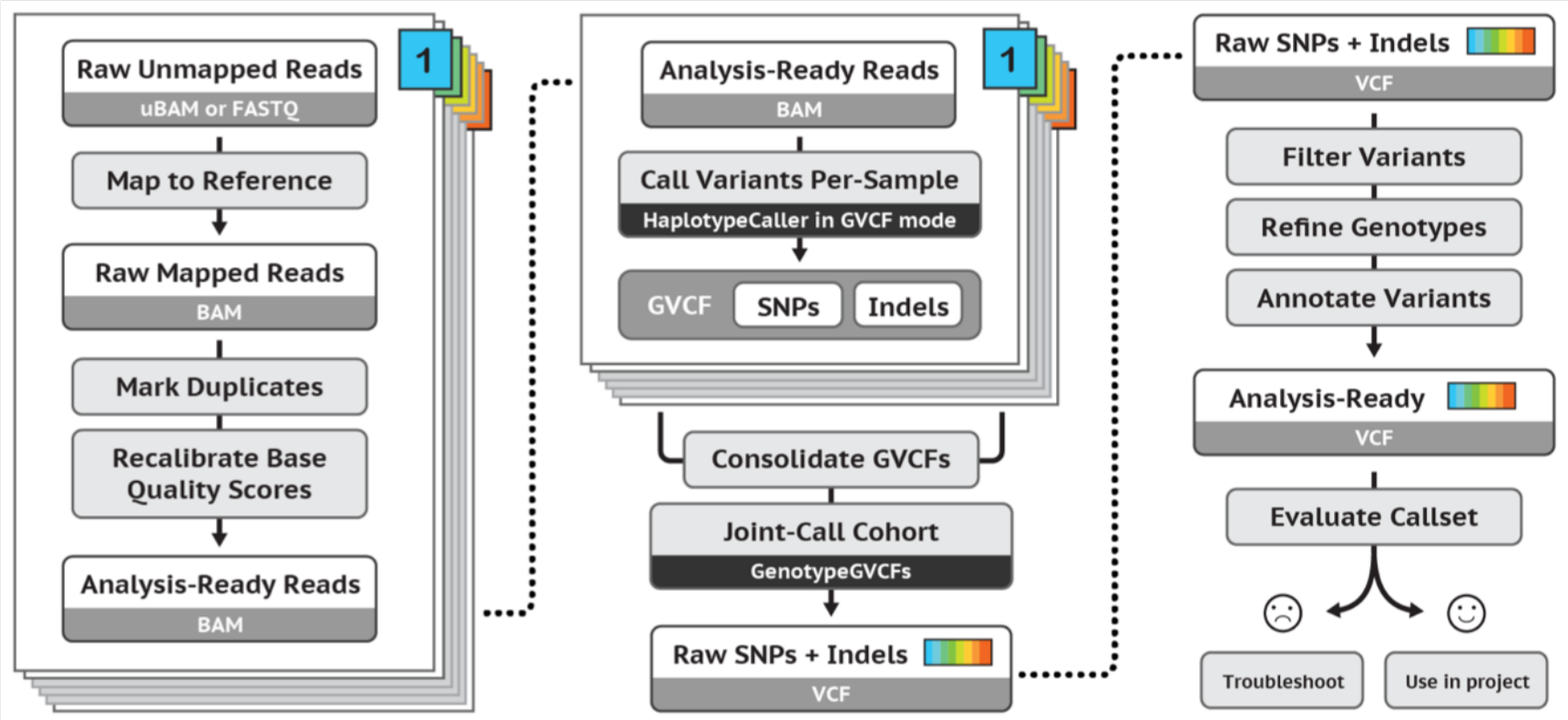
SciLifeLab

# Genome Analysis Toolkit (GATK) best practice variant discovery



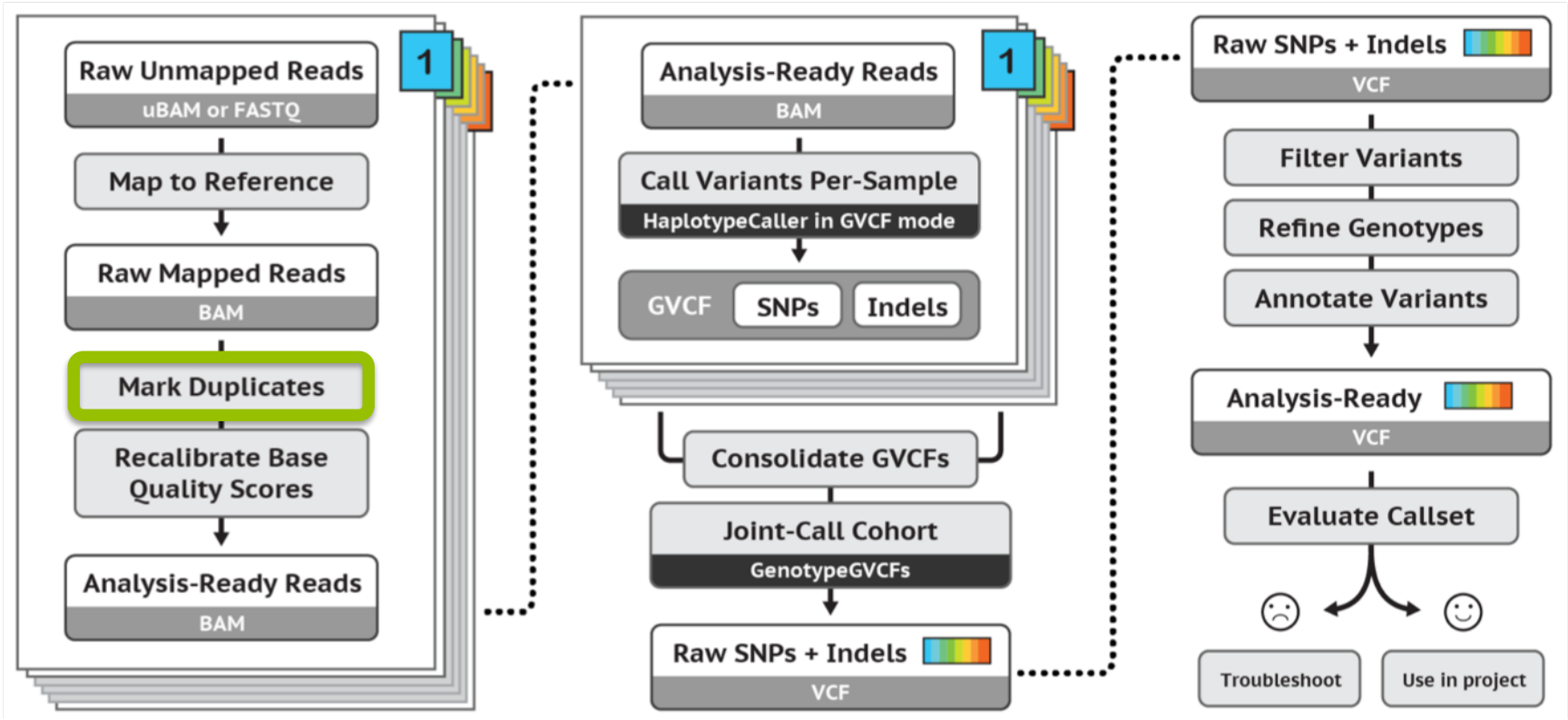
https://gatk.broadinstitute.org

# GATK best practices workflow for variant discovery



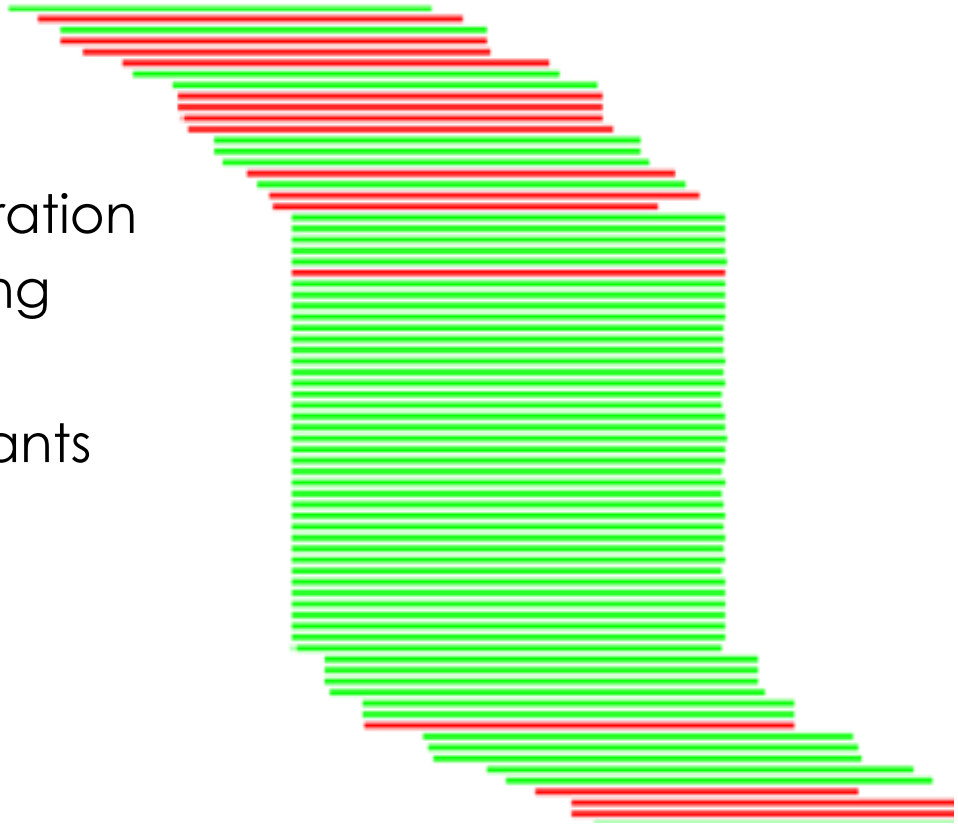

https://software.broadinstitute.org/gatk/best-practices/

# 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

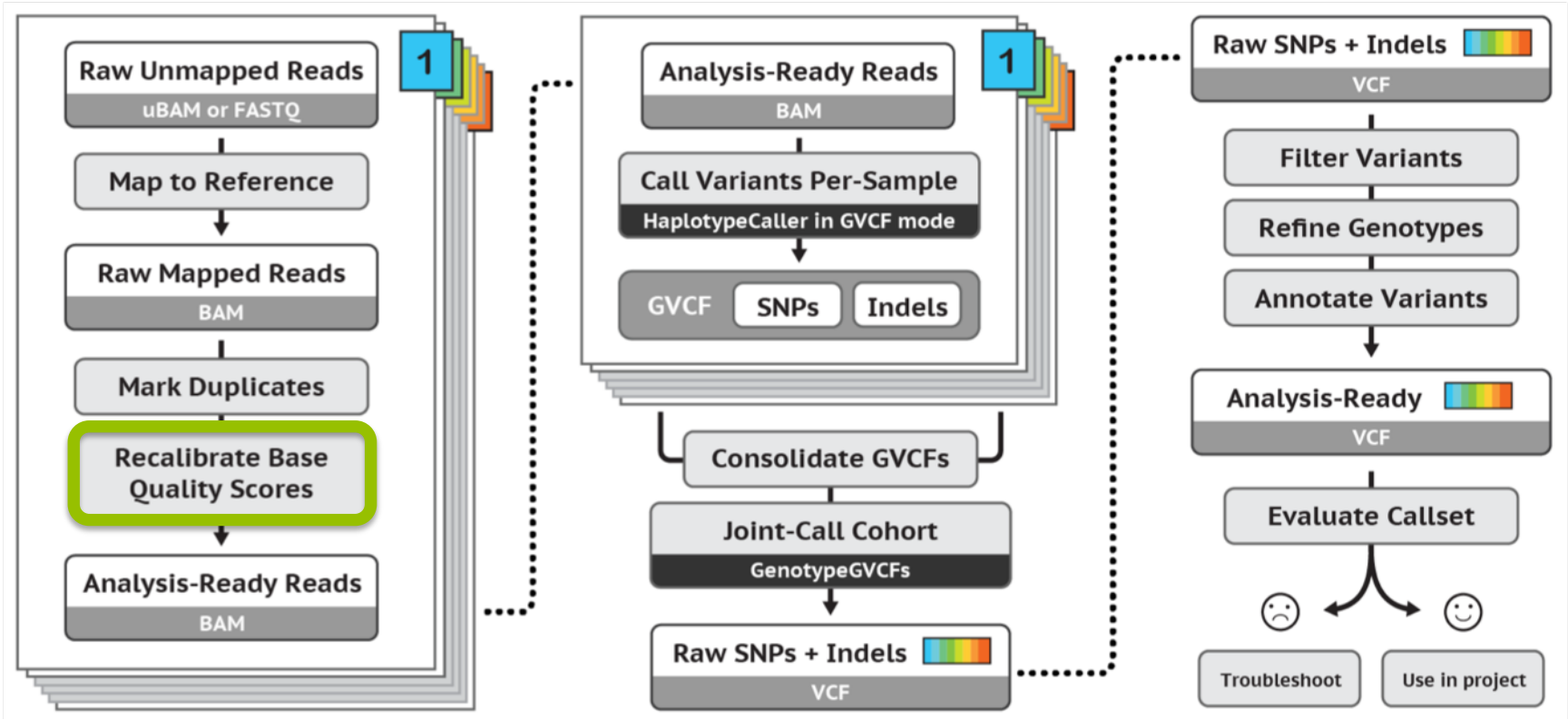
GATK Team
10 months ago · Updated

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)

# 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.*

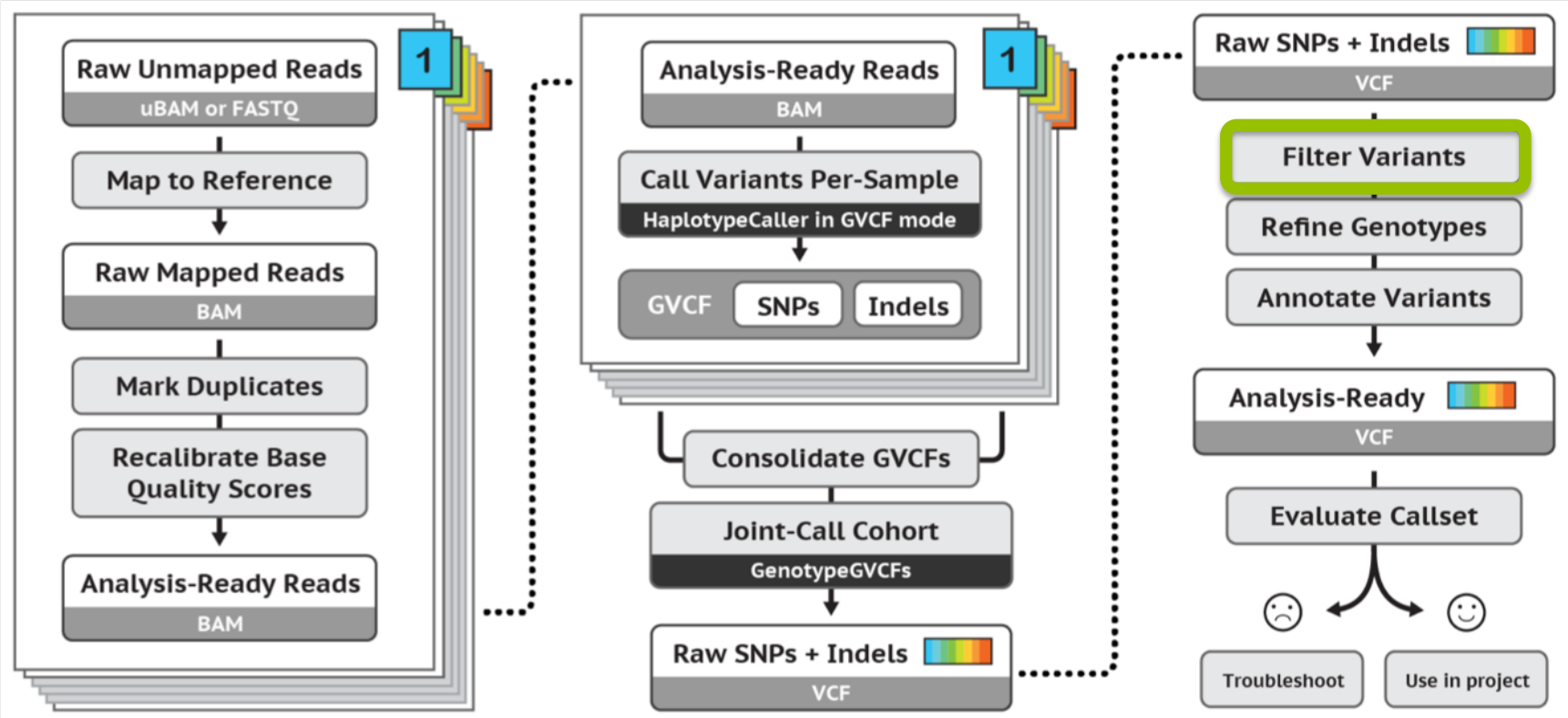
---

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

---

# Filter variants




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

Germline short variant discovery (SNPs + Indels)

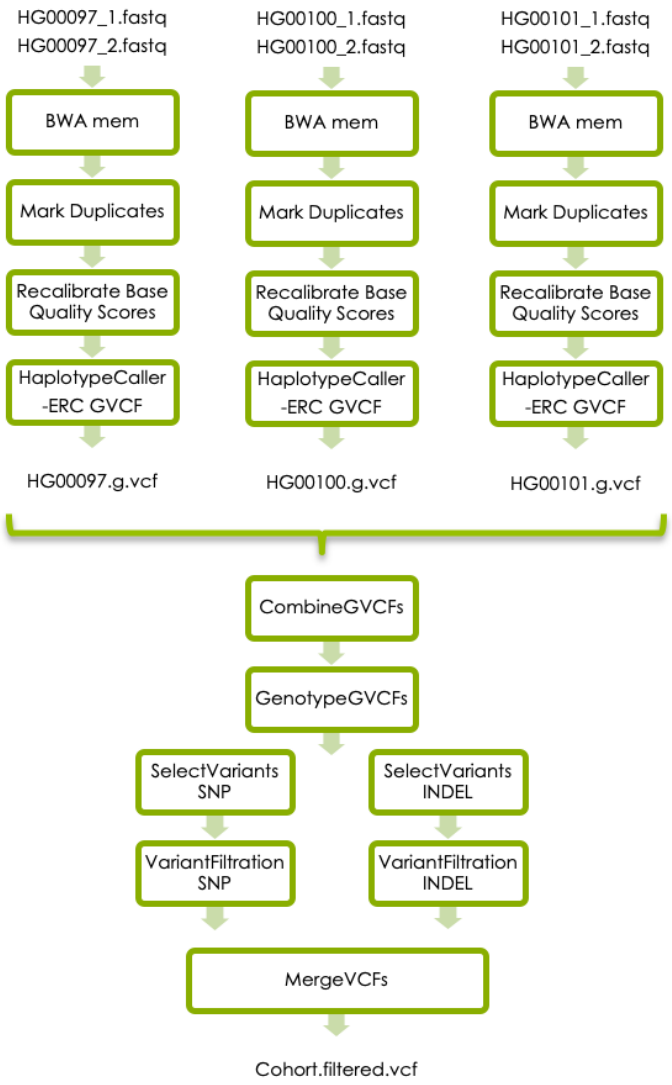
# Filtering

- 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 recommendations on hard filters: https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set

# GATK best practices workflow

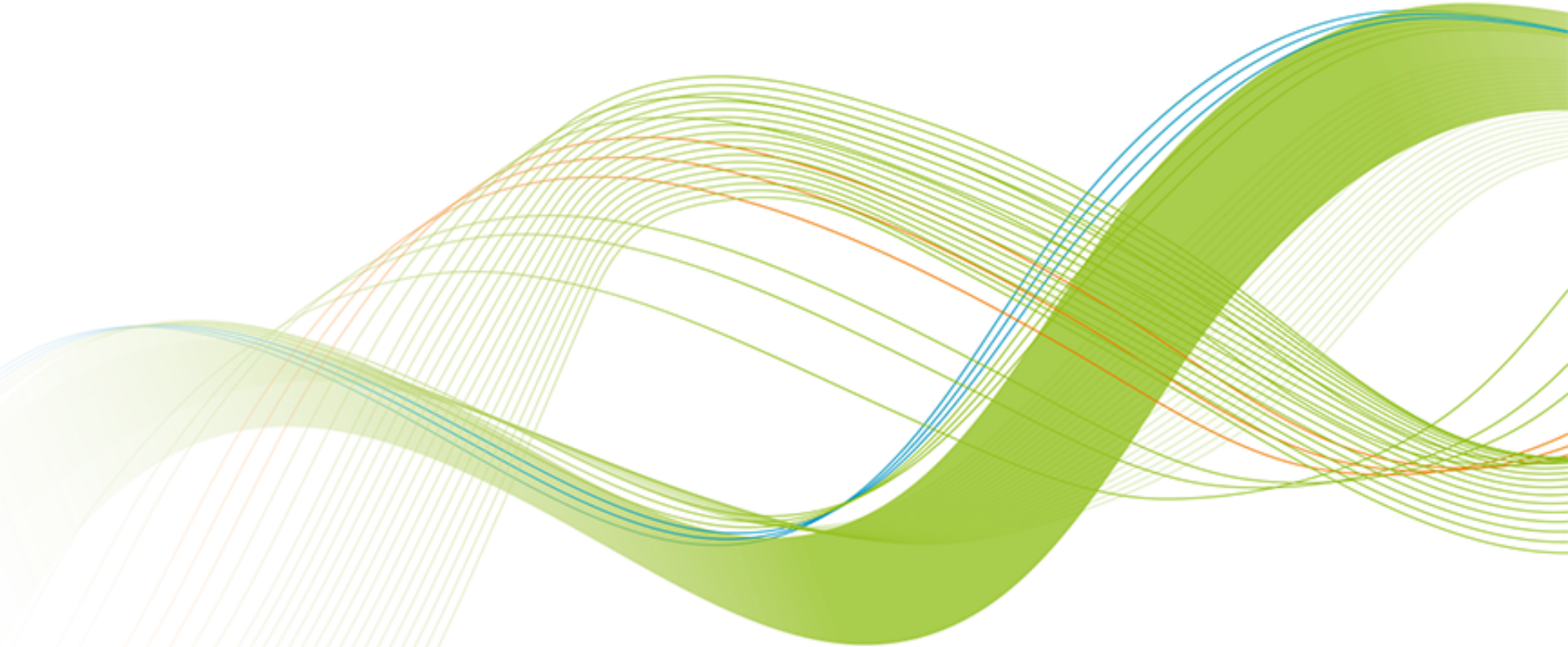



More details and links to GATK documentation is found in the lab instructions.

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

# Introduction to computational exercise



# 1000 Genomes data

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

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

# Lactose and Lactase

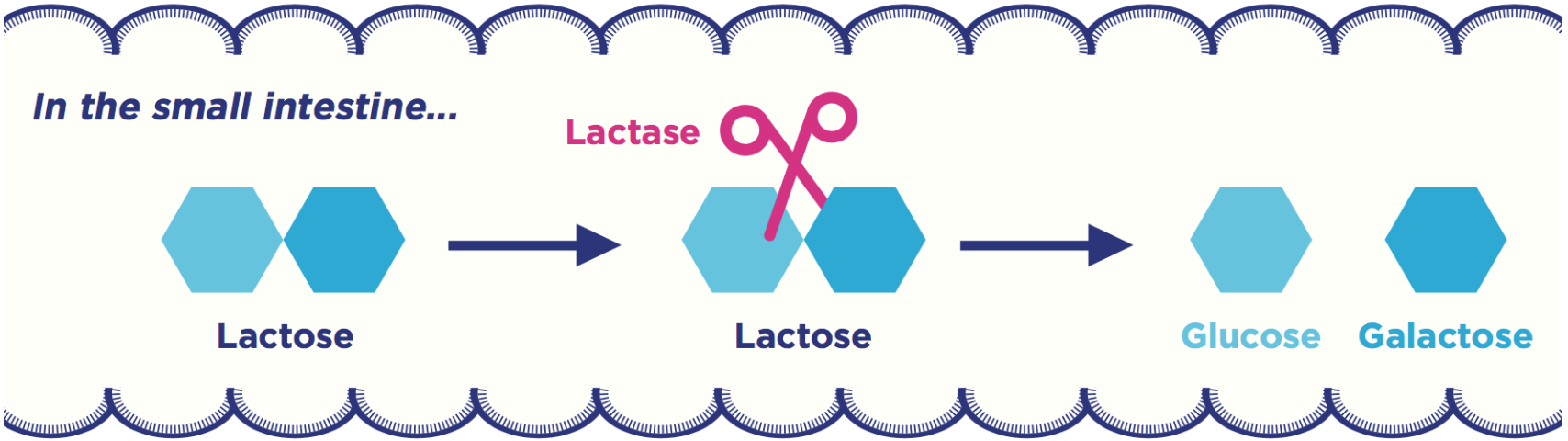
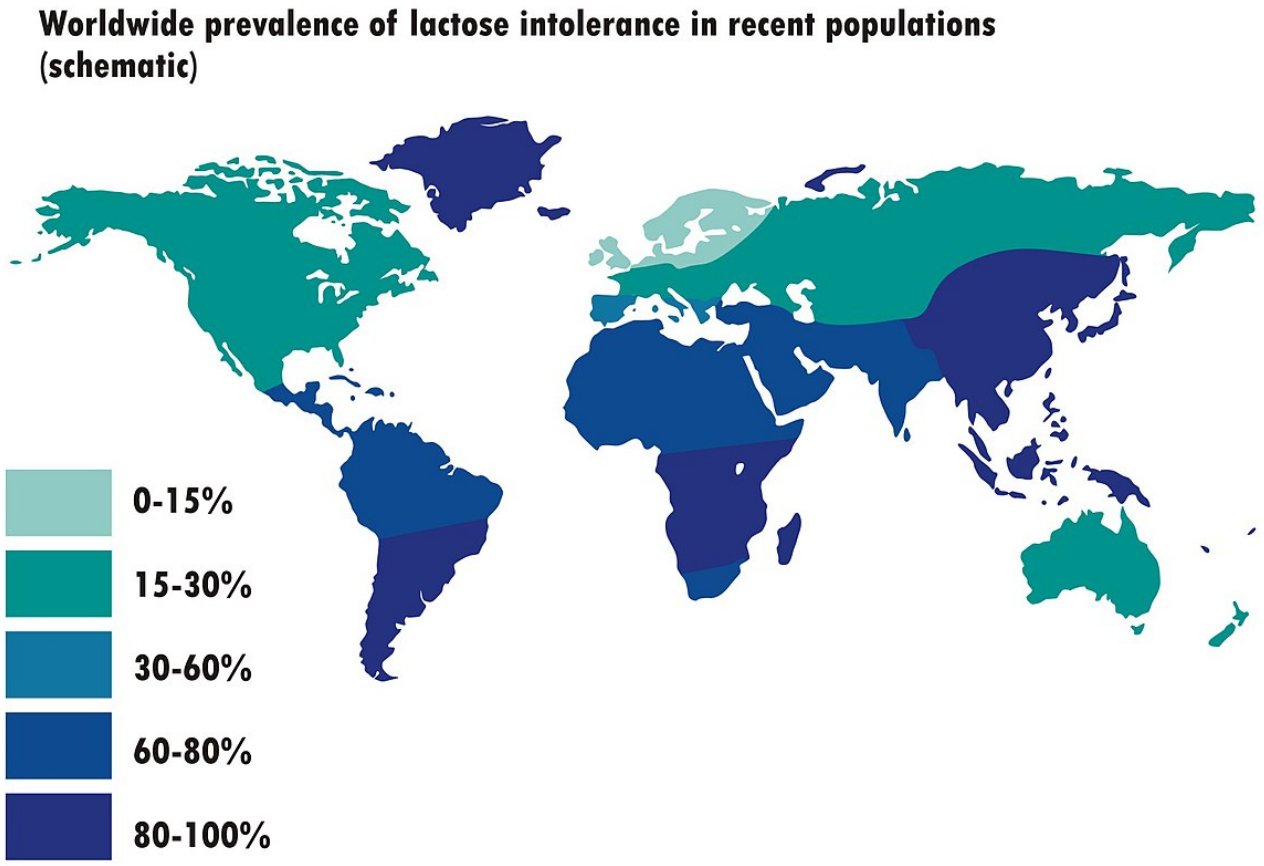



Figure 2. Lactose digestion in the intestine.

http://www.yogurtinnutrition.com/lactose-as-a-nutrient/

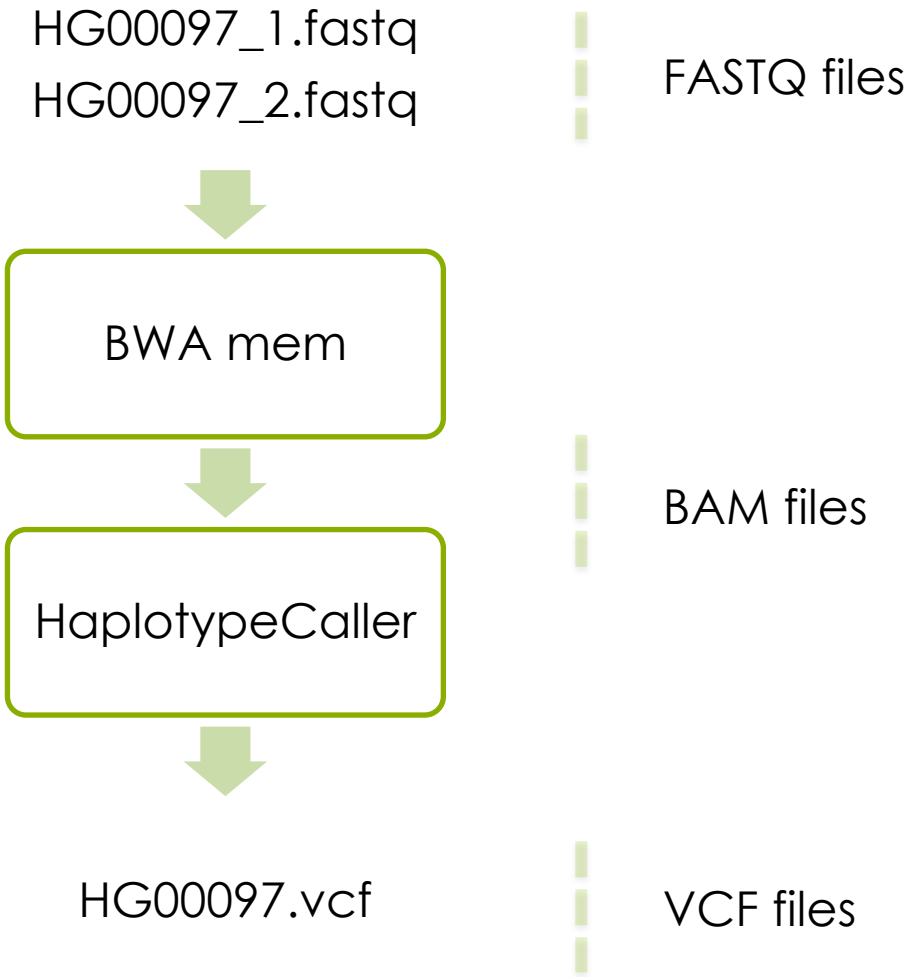
# Prevalence of lactose intolerance



# part one:

# variant calling in one sample

# Basic variant calling in one sample



# Part two (if you have time):

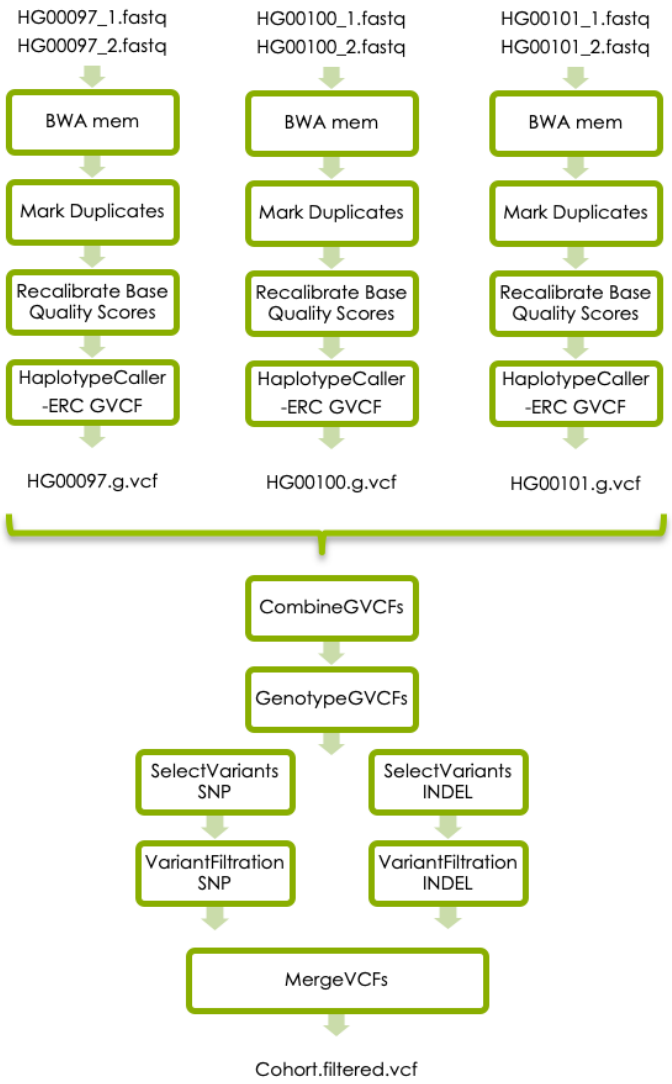
# variant calling in cohort

# Joint variant calling workflow

**Part three (if you have time):**

**Follow GATK best practices for short variant discovery**

# GATK best practises



# Questions?