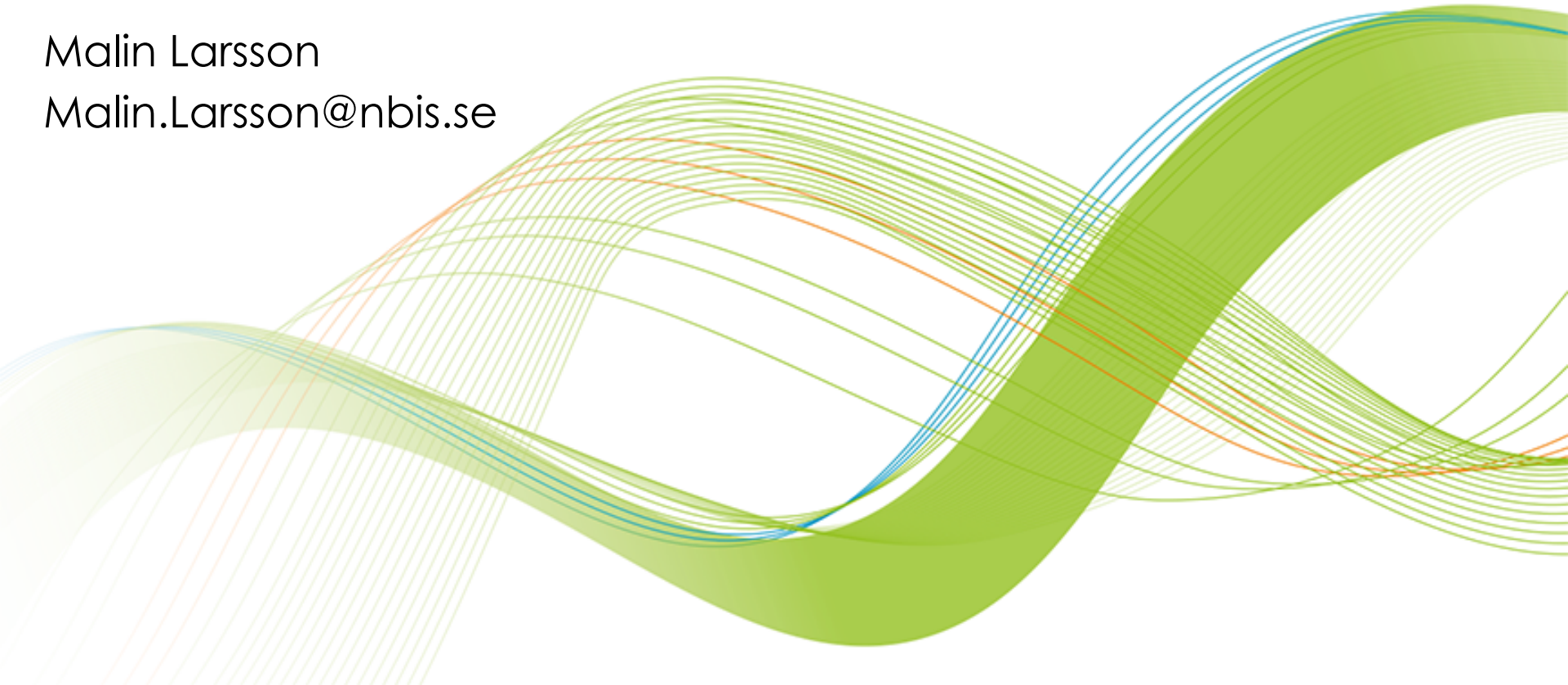
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Variant Calling Workflow

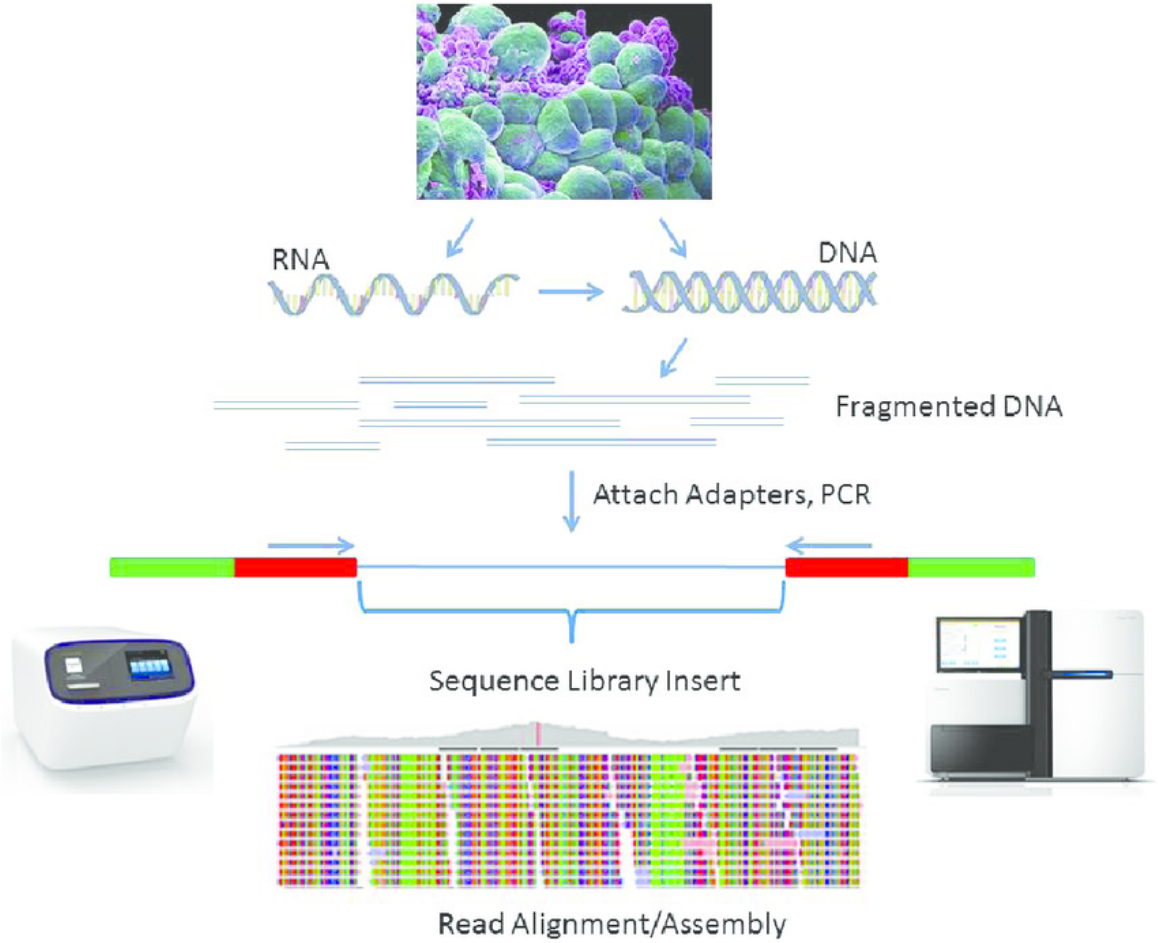
Malin Larsson
Malin.Larsson@nbis.se



Talk Overview

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

NGS overview



Library construction for next-generation sequencing: Overviews and challenges, Steven Head et al.

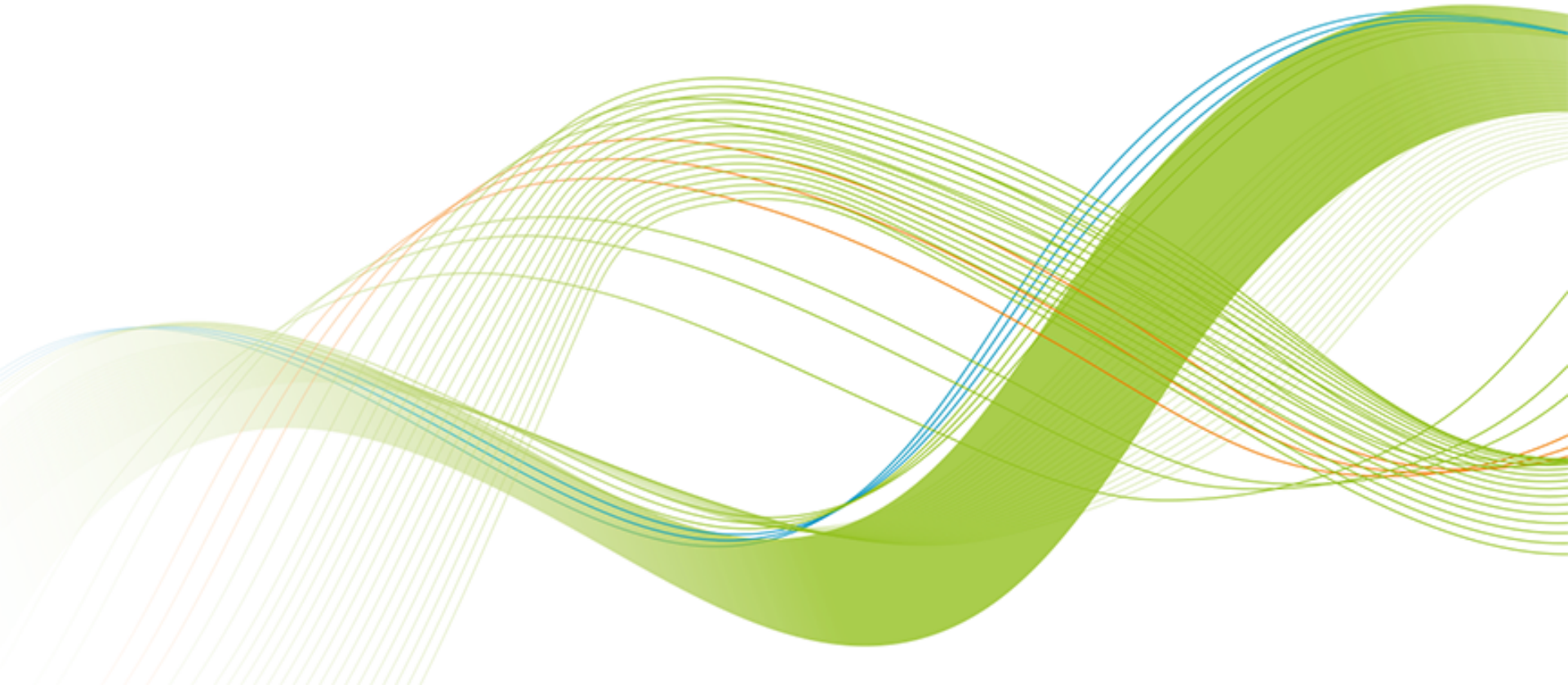
Illumina sequencing

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

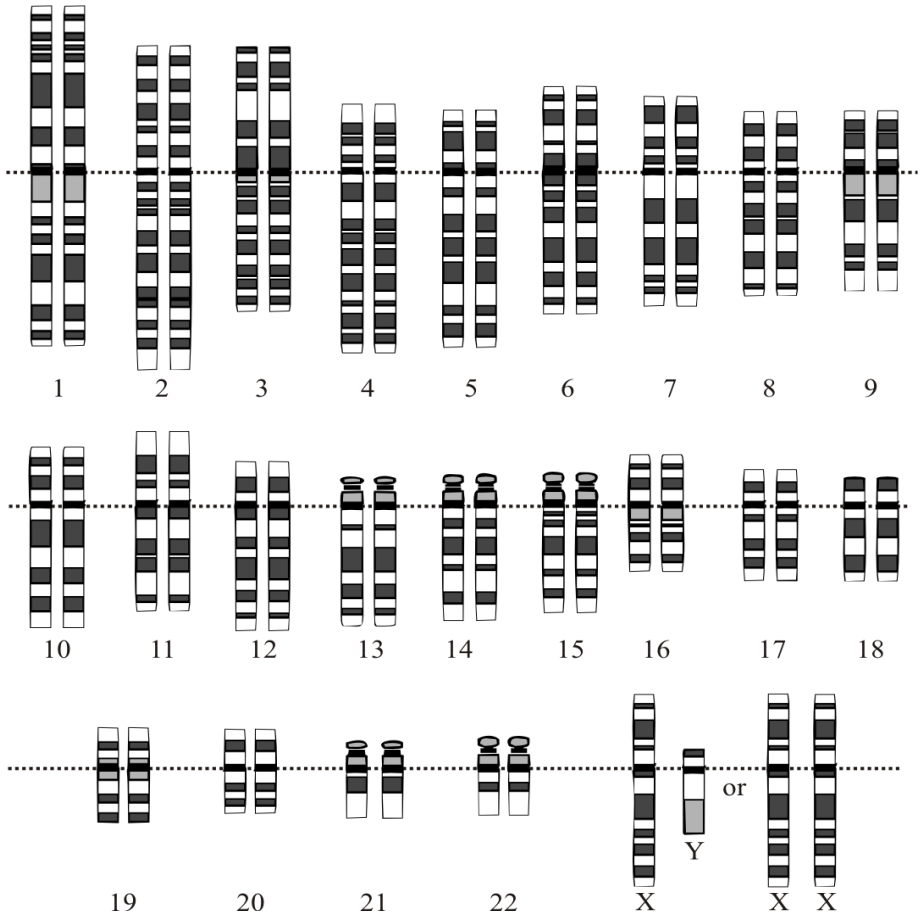
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

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

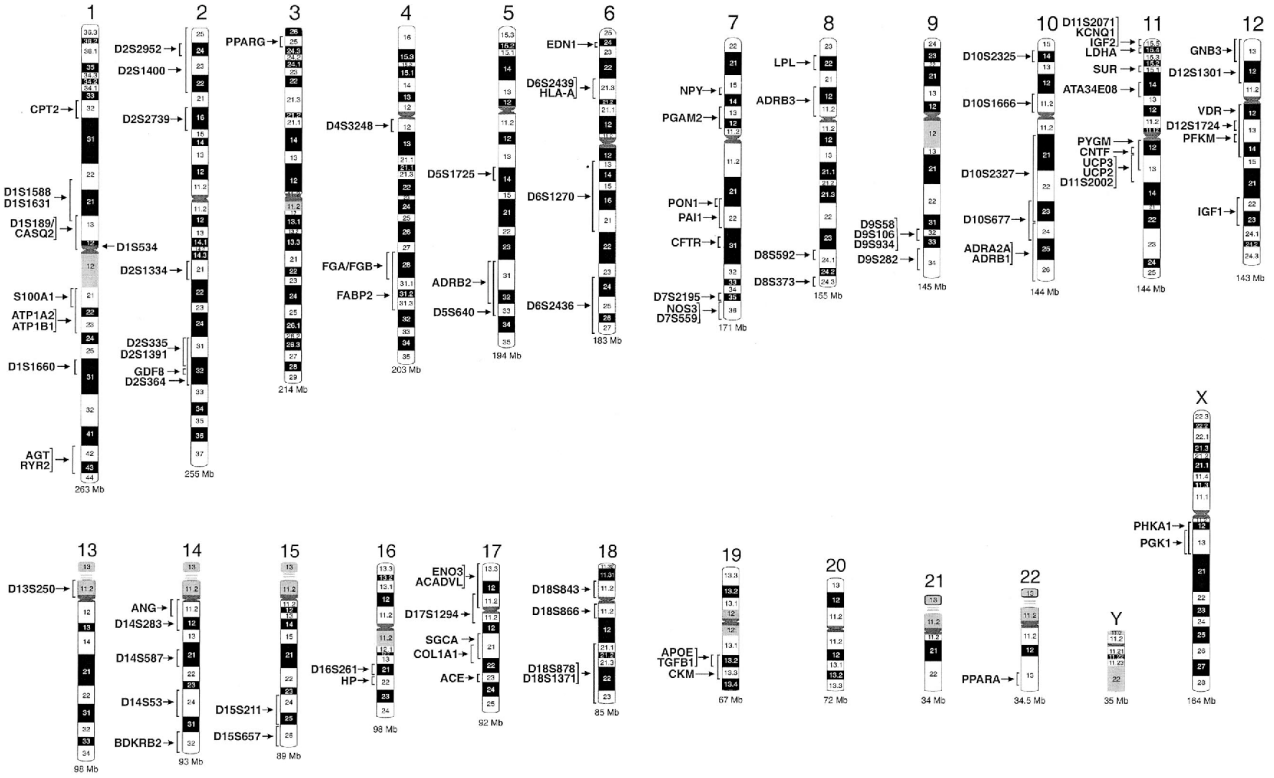
A reference genome is a haploid nucleic acid sequence which represents a species genome.

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.

Provides a map of all genes



Genome Reference Consortium

Genome Reference Consortium

GRC Home | Data | Help | Report an Issue | Contact Us | Credits | Curators Only

Human | Mouse | Zebrafish | Chicken | Paper Supplemental Data

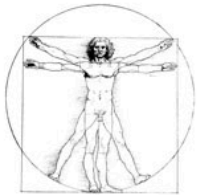
Genome Assemblies

The GRC has built tools to facilitate the curation of genome assemblies based on the sequence overlaps of long, high quality sequences (clones and PCR products, not short sequence reads). The GRC currently supports production of assemblies for human, mouse or zebrafish. If your assembly data fits this model and you are interested in using these tools, please contact us. Subscribe to the grc-announce email list to receive email notification for all GRC assembly updates.

GRC News

GRCm39: the new mouse reference genome assembly	*Jul 22, 2020*
ZFIN and the GRC: Supporting the zebrafish reference genome assembly	*Jun 08, 2020*

see all



Human

The human genome assembly was produced as part of the Human Genome Project (HGP). The previous assembly (NCBI36) was the last one produced by the HGP and was described in 2004 (PMID: 15496913); this was the starting point for the GRC. The assembly is based largely on assembling overlapping clone sequences.

Human assembly information

Current major assembly	GRCh38
Regions with alternate loci	178
Assembly N50	67,794,873 bp
Remaining gaps	875
Patch release version	p13
Patches released	FIX: 113, NOVEL: 72

More human assembly statistics...

Mouse

The GRC has produced an updated assembly (GRCm38). This is an update of the last MGSC assembly (MGSCv37) which was described in 2009 (PMID: 19468303). The primary assembly is based on assembling overlapping BAC clones derived from the C57BL/6J strain and several loci have sequence available from other strains.

Mouse assembly information

Current major assembly	GRCm39
Regions with alternate loci	0
Assembly N50	106,145,001 bp
Remaining gaps	347
Patch release version	None
Patches released	FIX: 0, NOVEL: 0

More mouse assembly statistics...

Zebrafish

The zebrafish genome assembly was produced at the Wellcome Sanger Institute. The last assembly produced from the original project was Zv9 and was described in 2013 (PMID: 23594743). This assembly is the starting point for the GRC. The

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

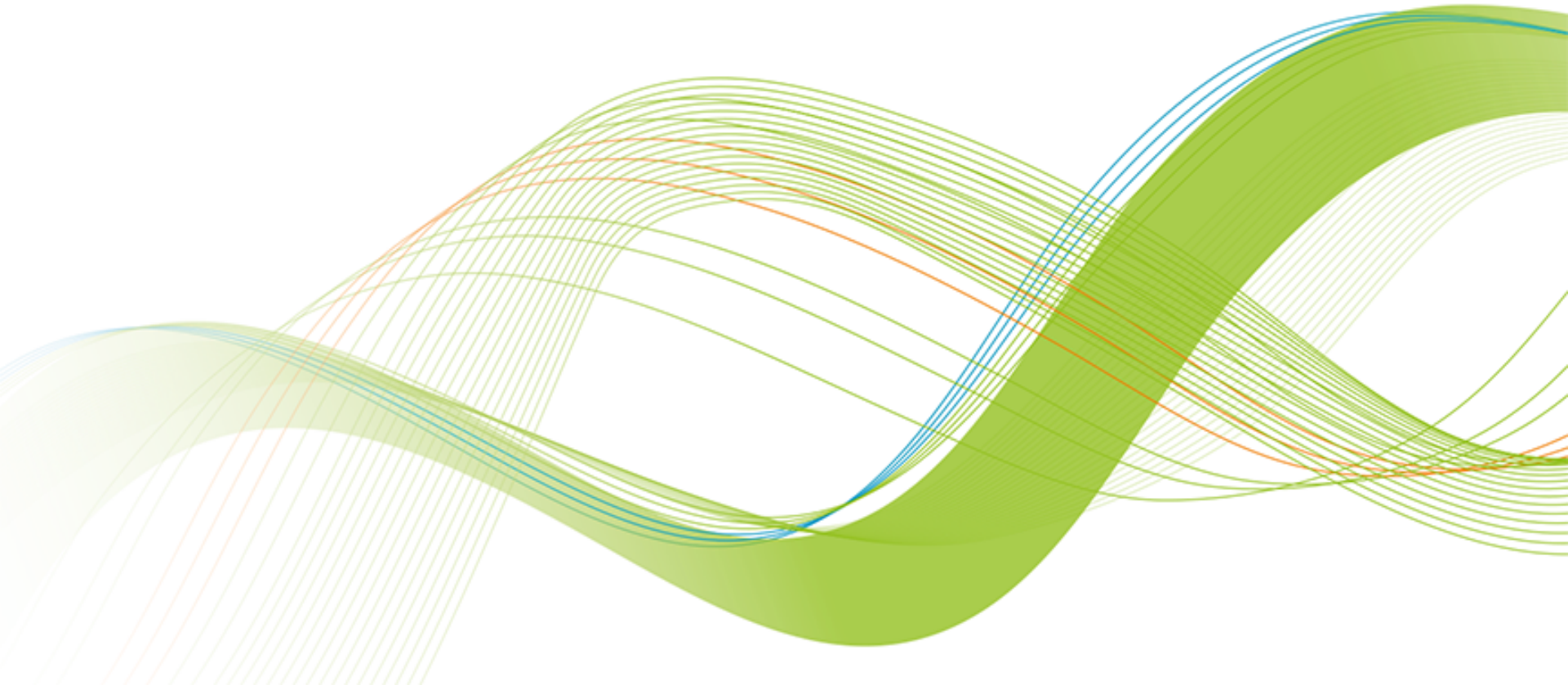
You must keep track of which version of the reference genome your data was mapped to.

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

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

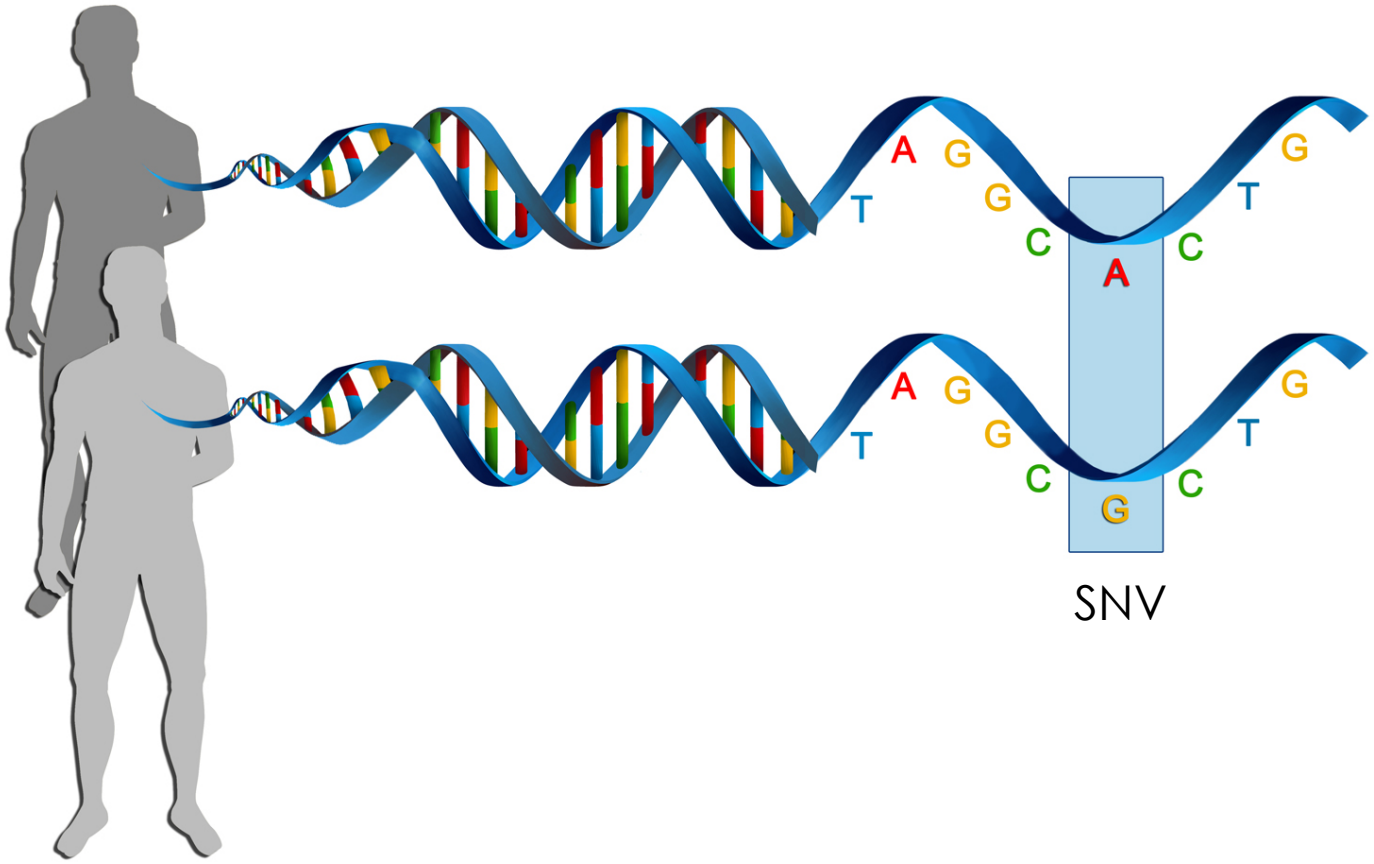
Genetic variation



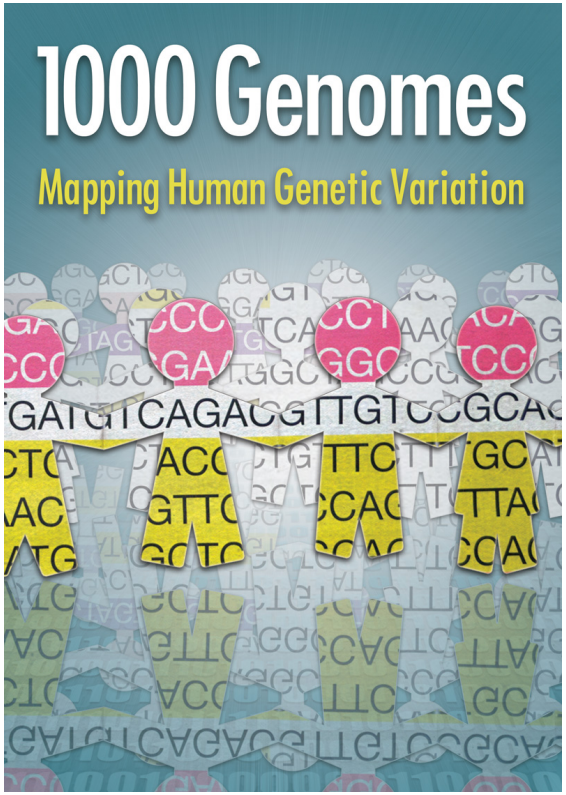
Genetic Variation

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

Single Nucleotide Variants (SNVs)



Population based variant projects



SweGen

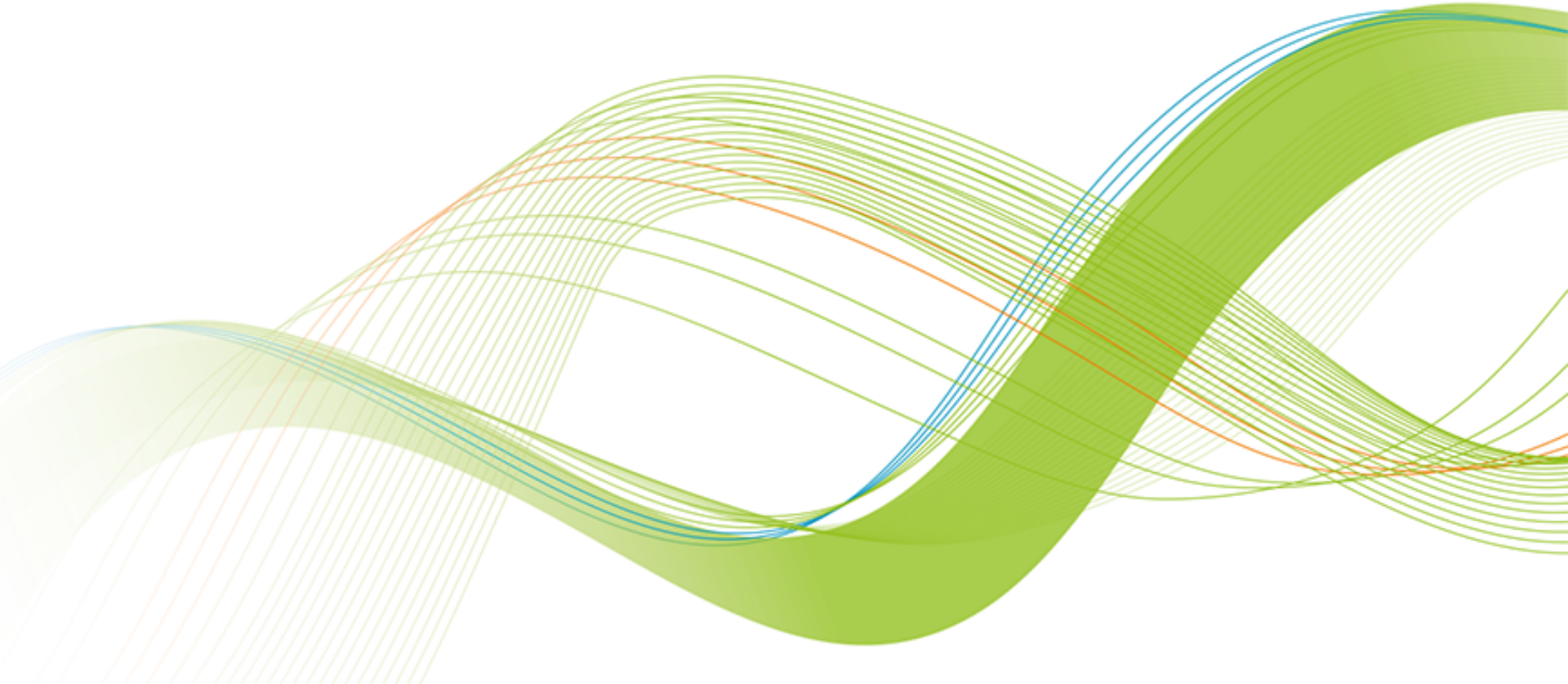
UK 10K

RARE GENETIC VARIANTS IN HEALTH AND DISEASE

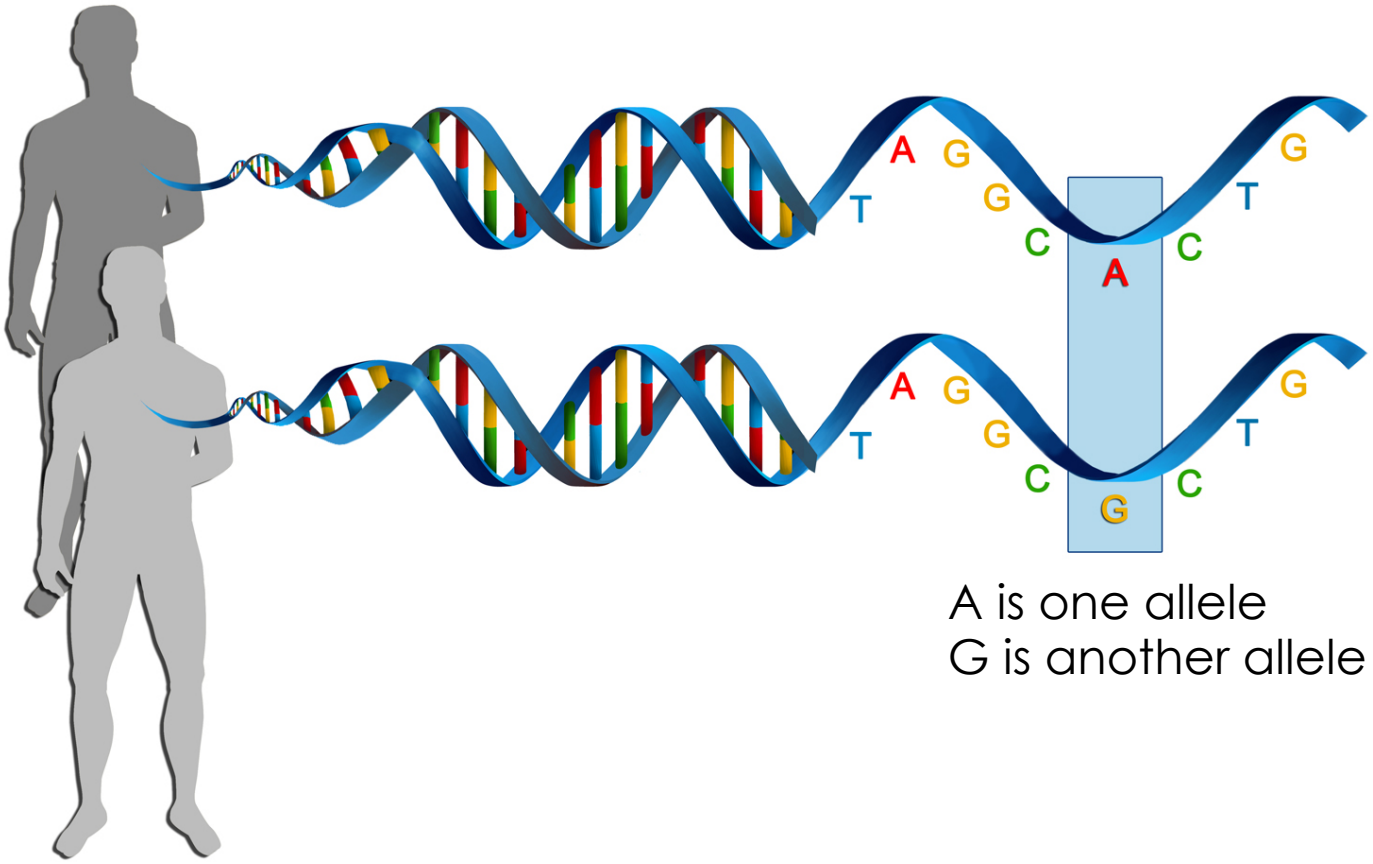
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Reference- and Alternative Alleles



The different genetic variants at a position are called Alleles



A is one allele
G is another allele

Reference- and Alternatve Alleles

GGCTTTTCCAACAGGTATATCTTCCCCGCTAGCTAGCTAGCTACTTCAAAT

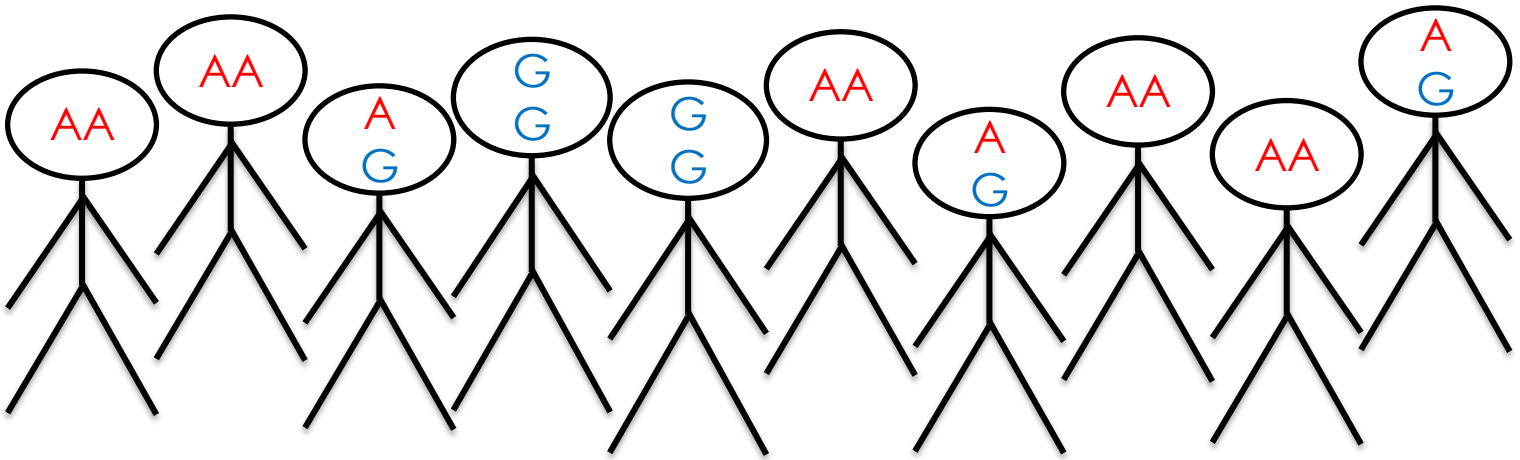
Reference allele AGCTAGCTA

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

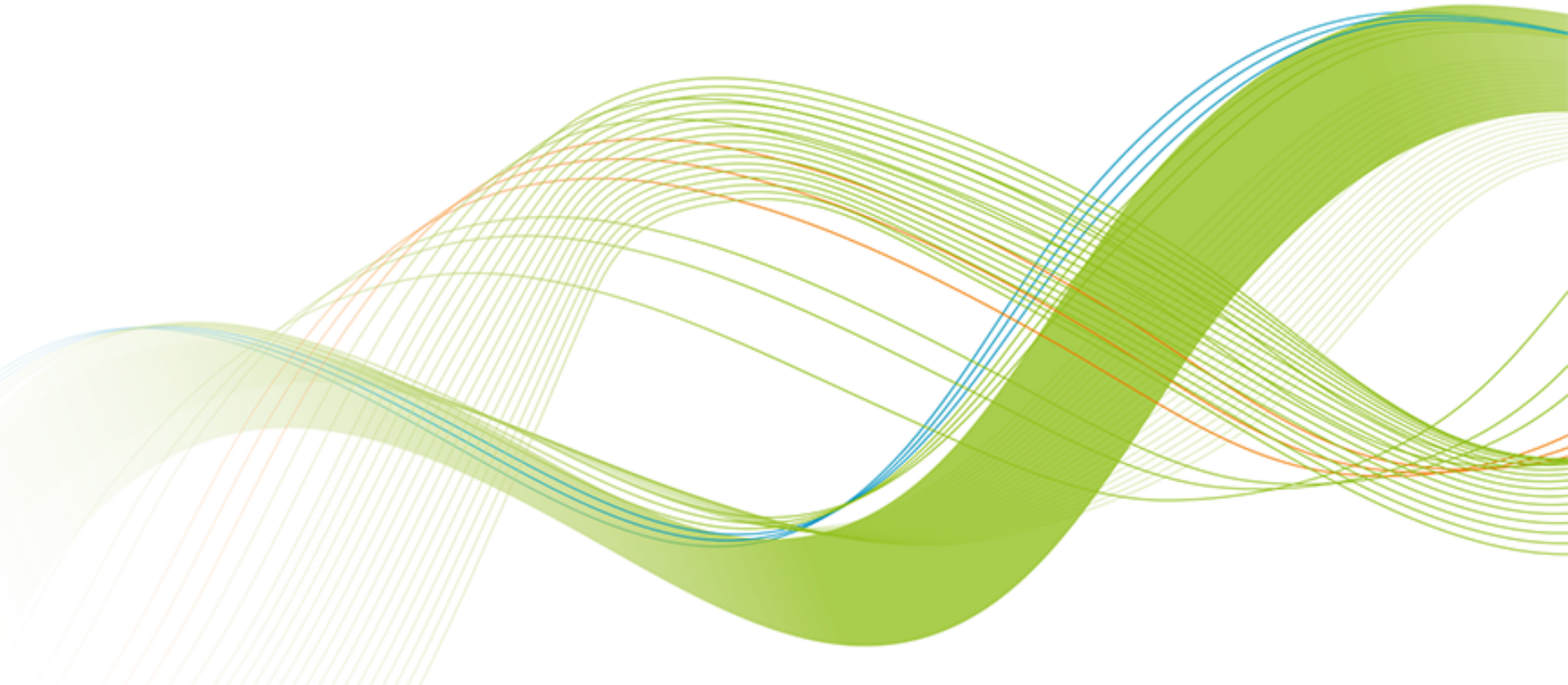


Genotypes:	AA	AG	GG
Genotype frequency:	5/10	3/10	2/10
Frequency of allele A:	13/20 = 0.65		
Frequency of allele G:	7/20 = 0.35		

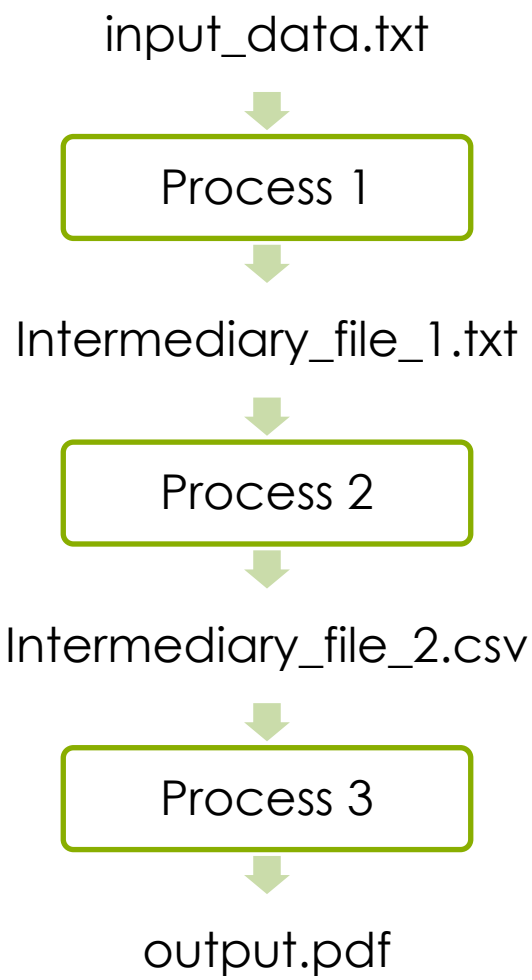
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

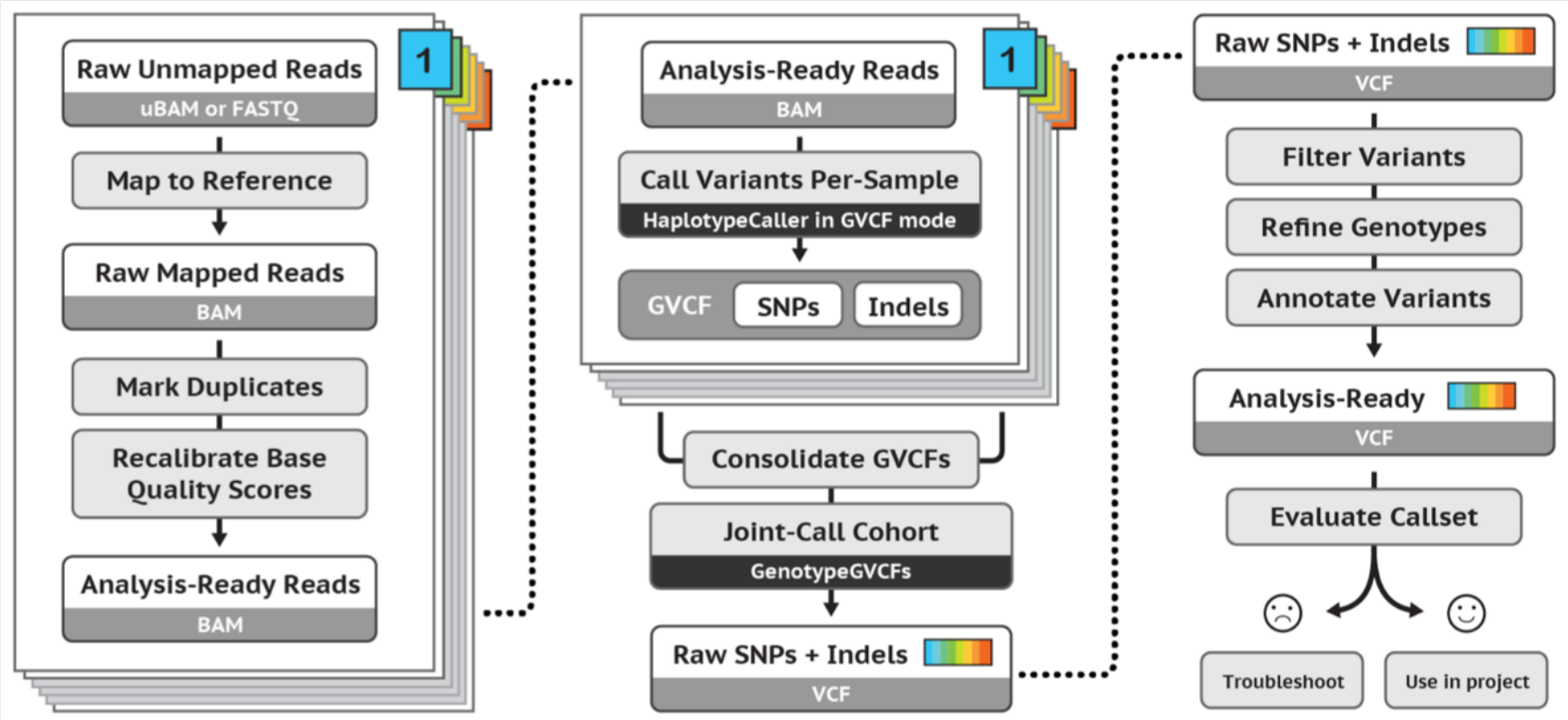
Introduciton to workflows



A bioinformatics workflow



GATK best practices workflow for variant discovery

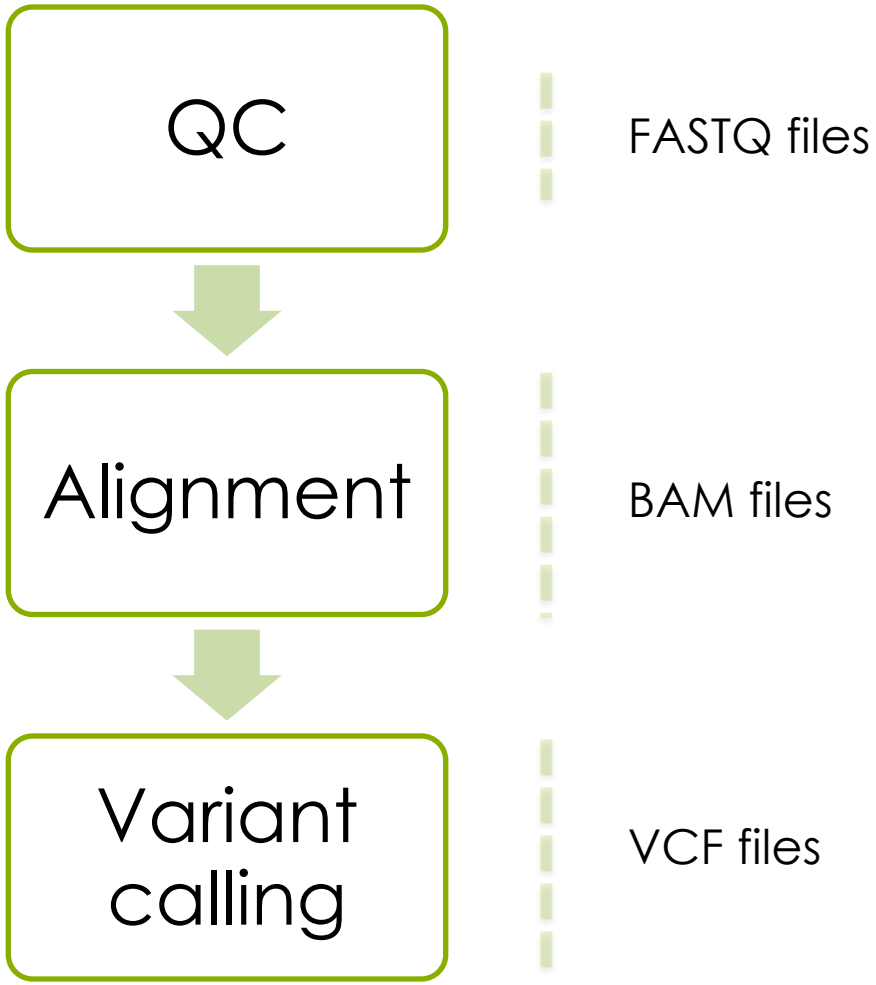


https://software.broadinstitute.org/gatk/best-practices/

Workflow conventions

- create a new output file in each process
- Use informative file names

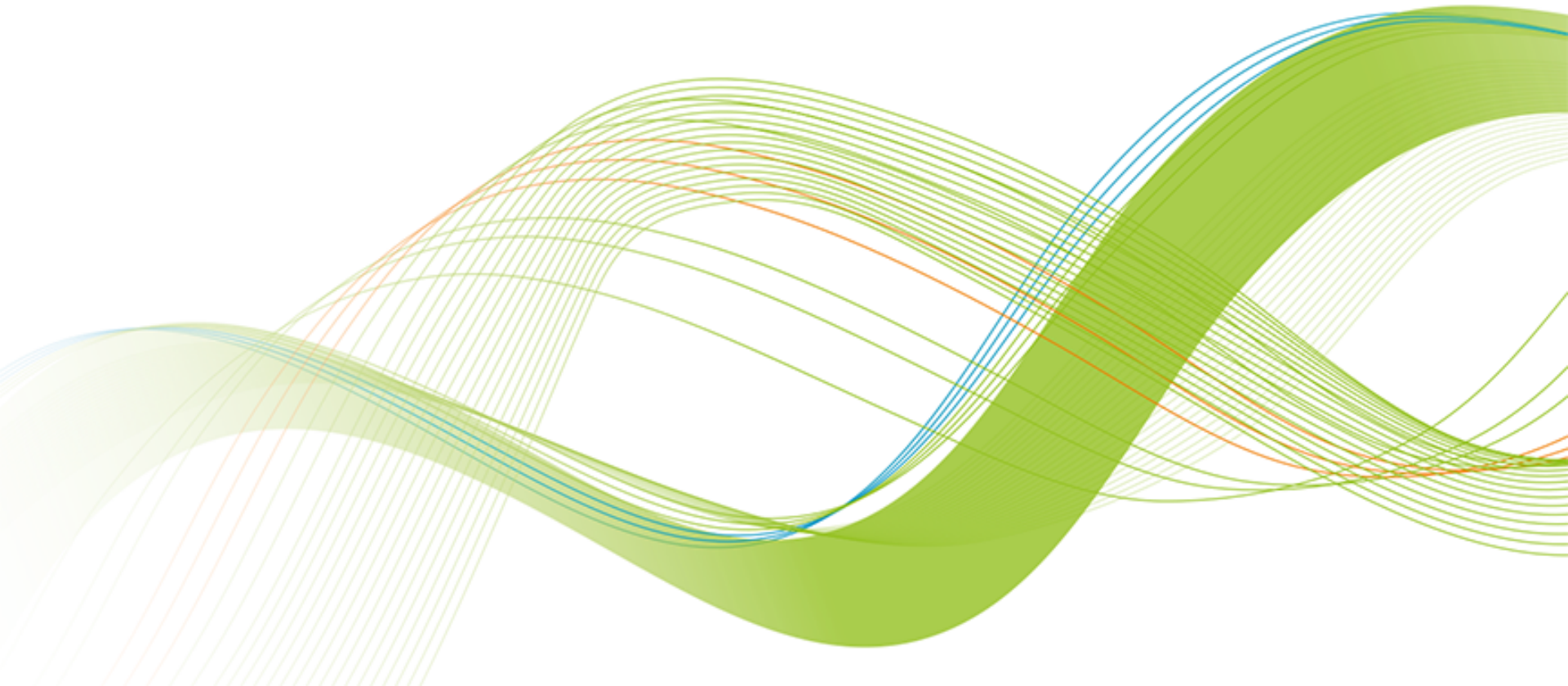
Example: Basic variant calling in one sample



NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Quality control

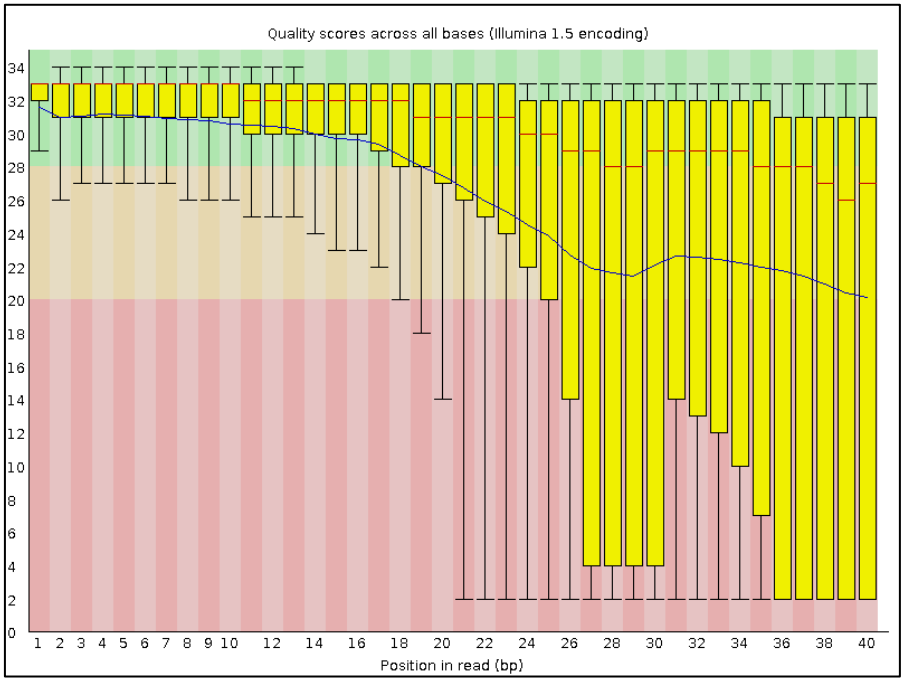


Basic variant calling in one sample

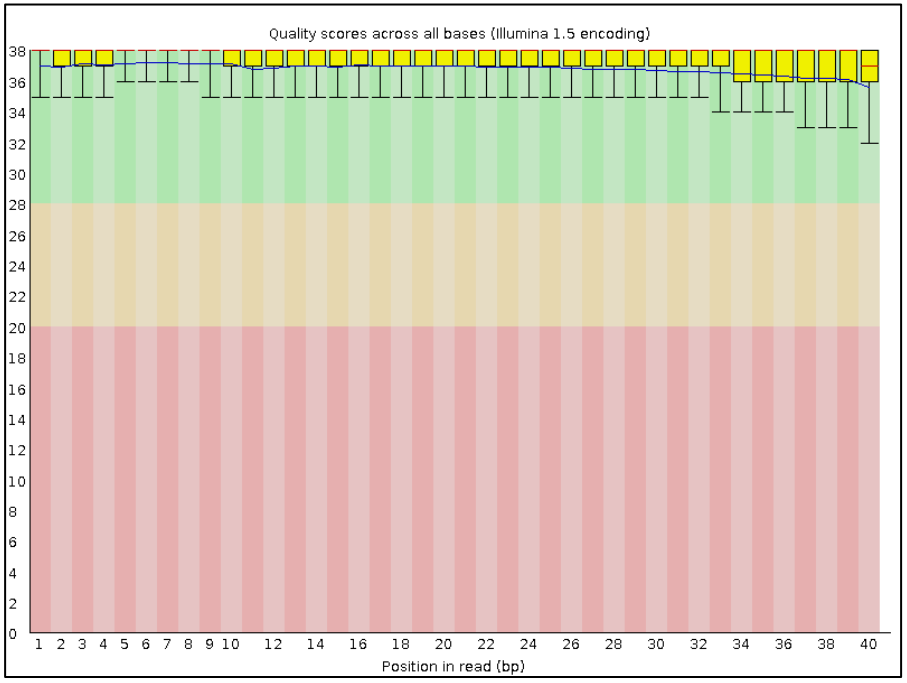
Quality control

```
module load FastQC
```

Bad qualities:



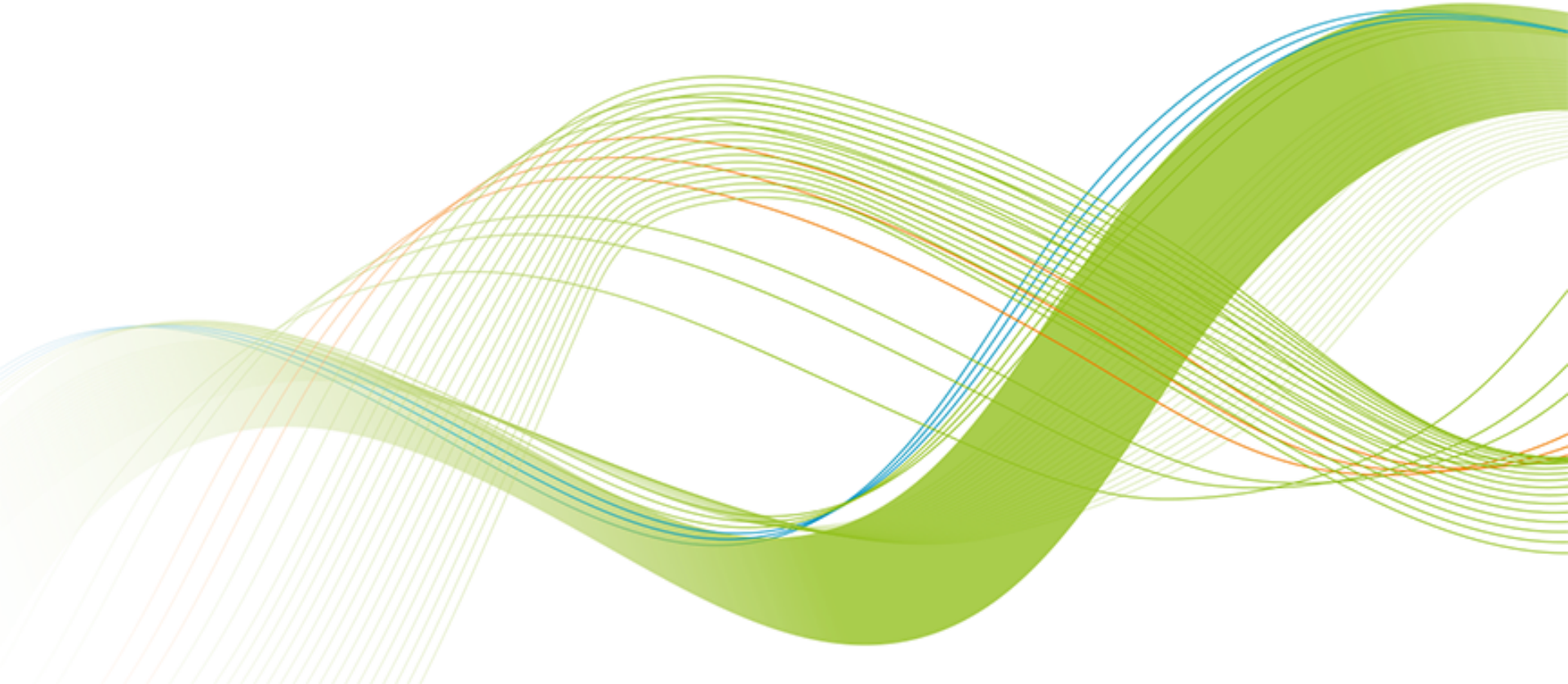
Good qualities:



NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Map to reference for variant calling



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

Alignment

```
module load bwa
```

Alignment

```
module load bwa
```

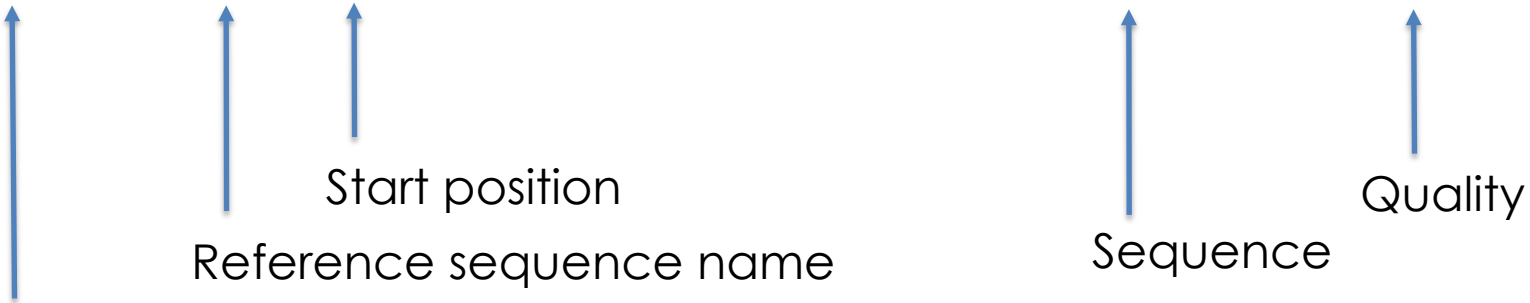
Output from mapping - Sam format

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

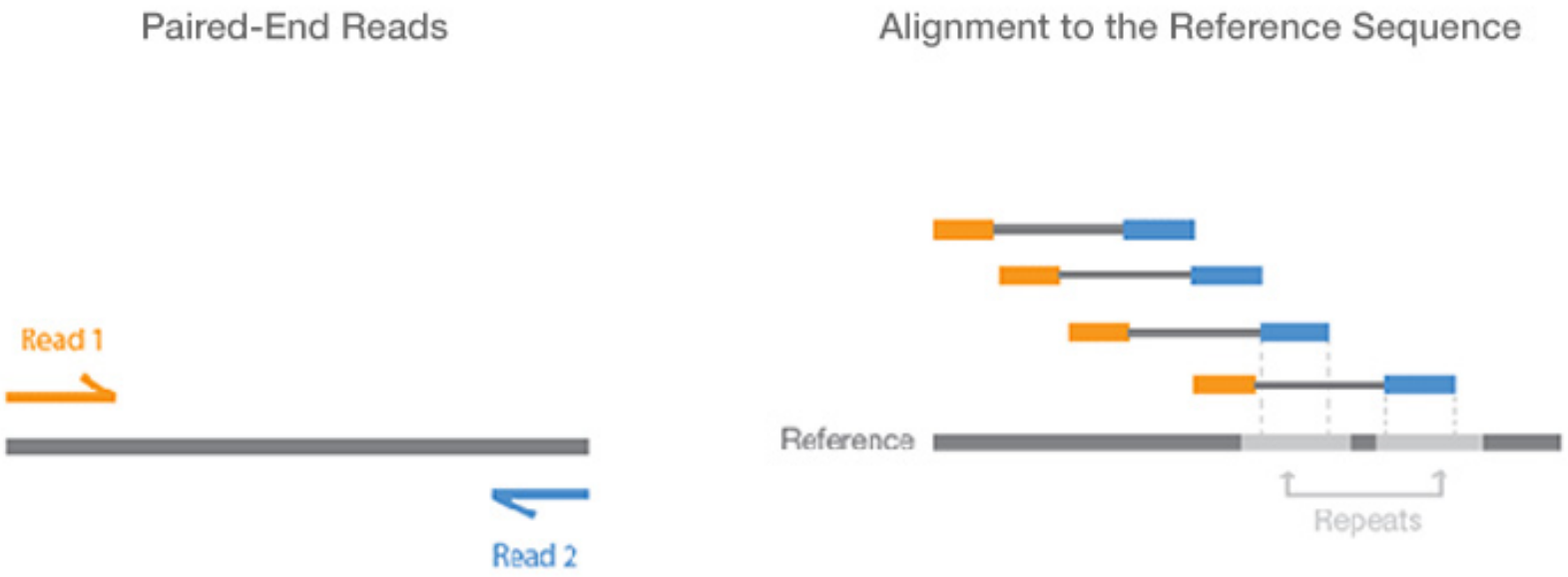
- Link *sample id*, *library prep*, *flowcell* and *sequencing run* to the reads.
- Good for error tracking!
- Often needed for variant calling
- Detailed description in tutorial or https://gatkforums.broadinstitute.org/gatk/discussion/6472/read-groups

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
- Different index for different file-types
- Bwa index creates one set of index for the reference that it needs for performing alignment
- Other programs like samtools produce other types of index for .fasta and .bam files needed by other programs

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
```

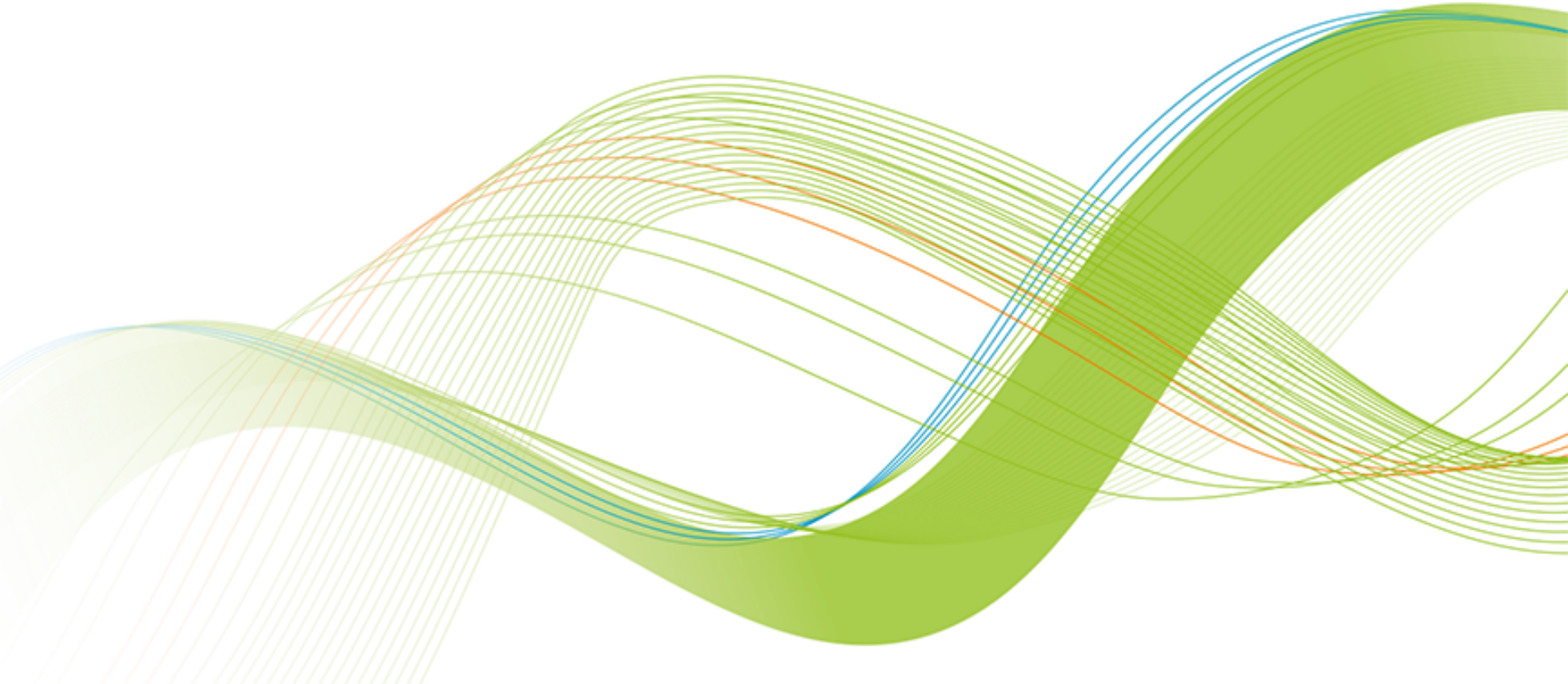
ID_R2_001.fastq

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

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

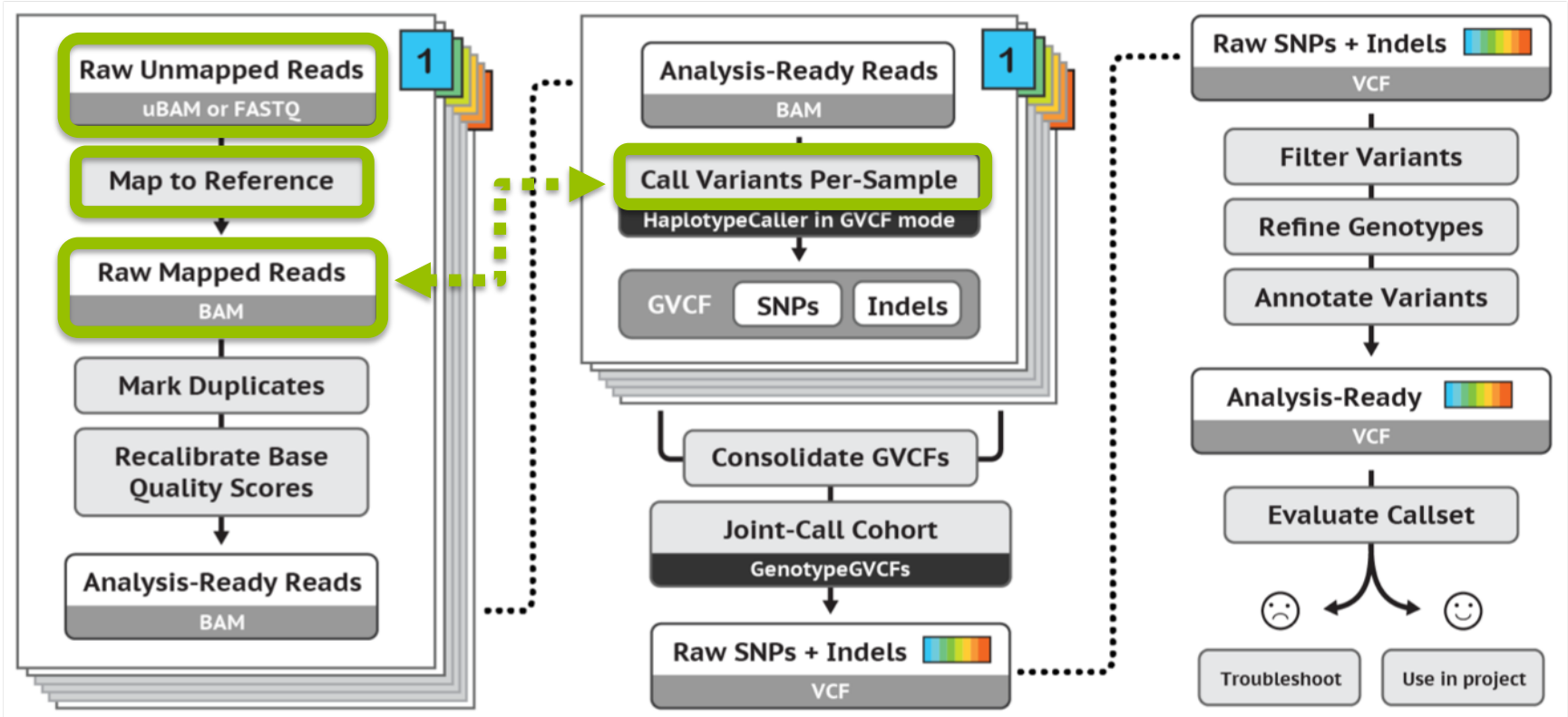
Variant calling in one sample



Basic variant calling in one sample

Basic variant calling workflow

- For learning the concepts!
- The minimum steps needed to go from fastq files to vcf file for one sample



Variants

A position where sample sequence does not agree with reference genome sequence

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

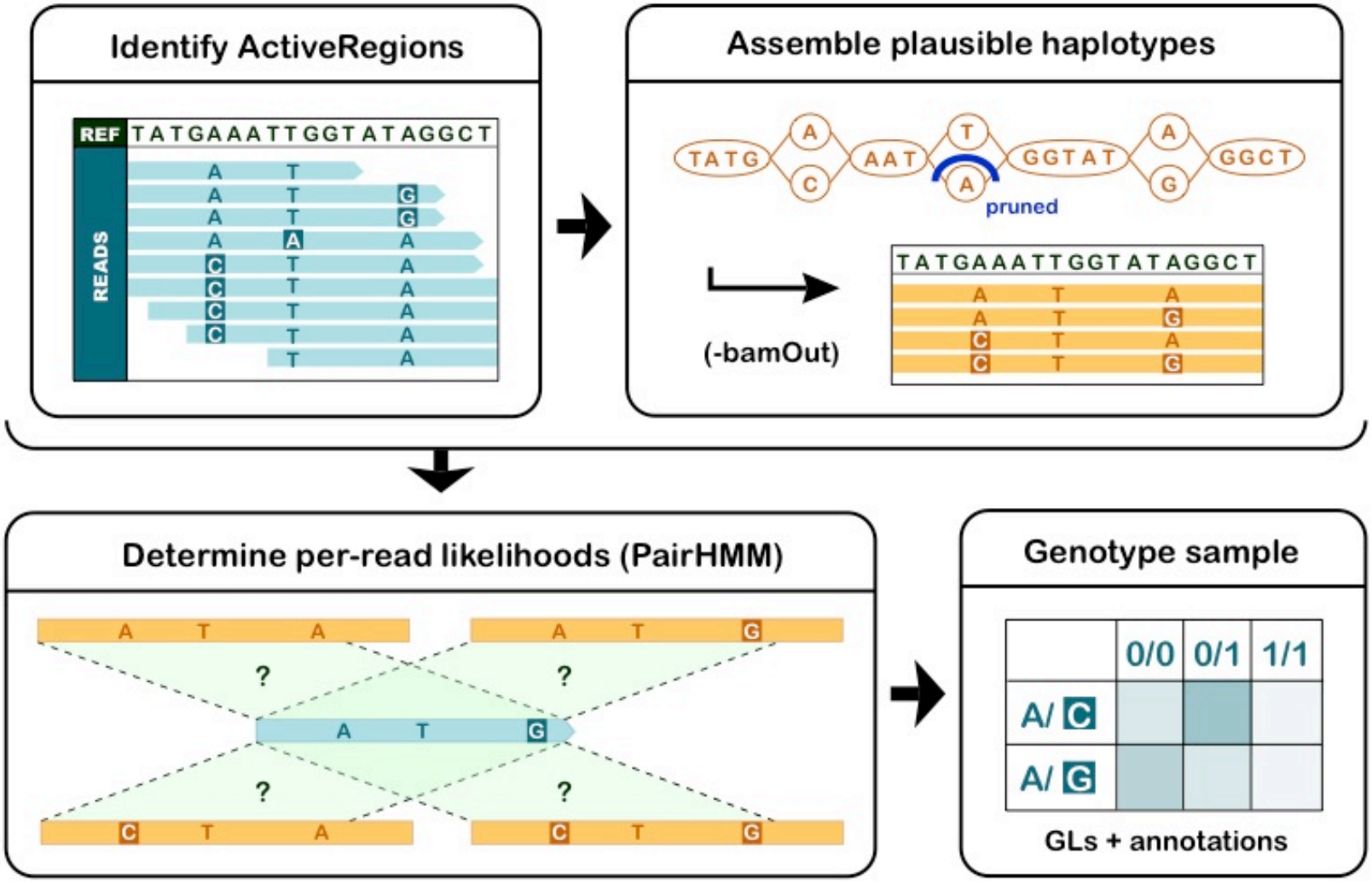
Variant calling

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...

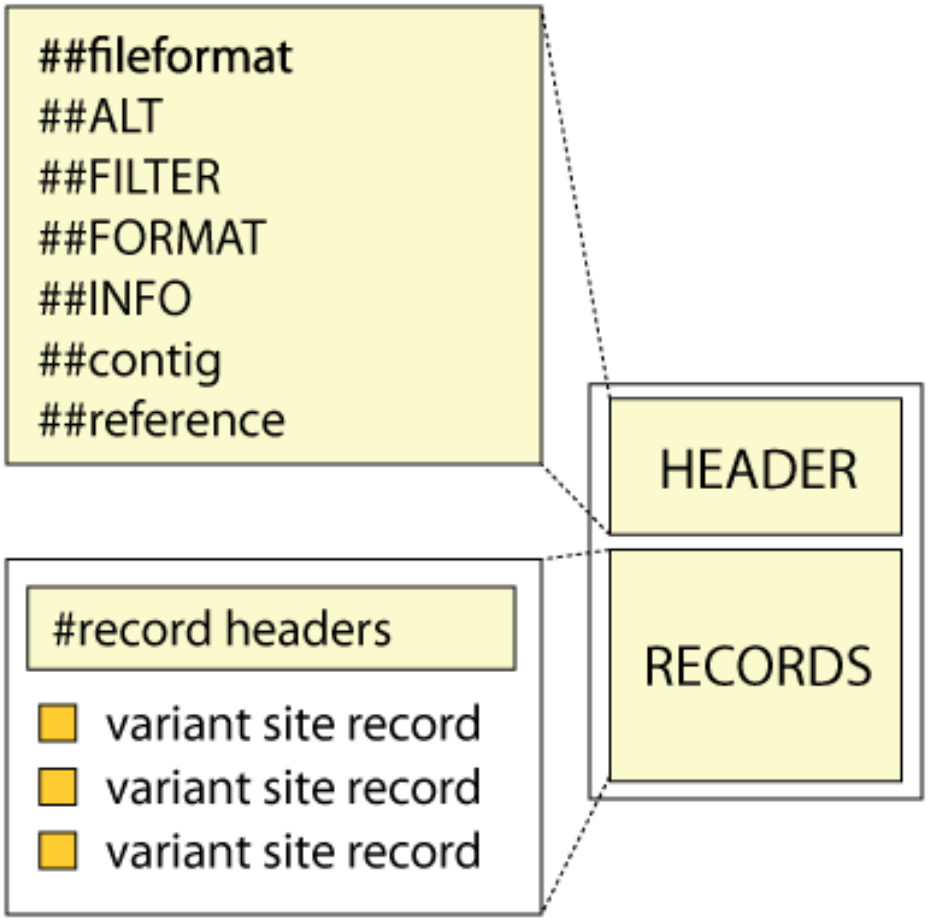
Sample:
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...

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

Variant Calling
HaplotypeCaller



Variant Call Format (VCF)



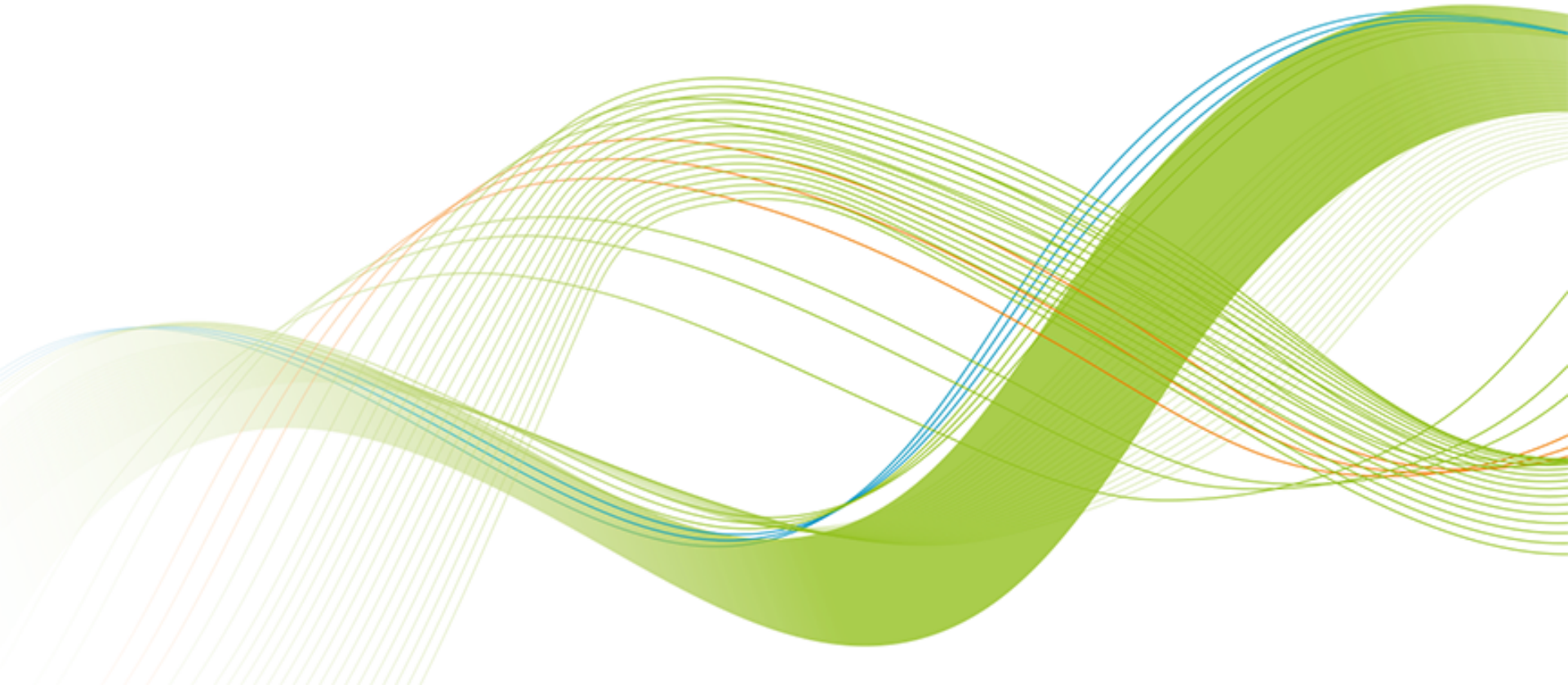
Variant Call Format (VCF)

```
##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        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
```

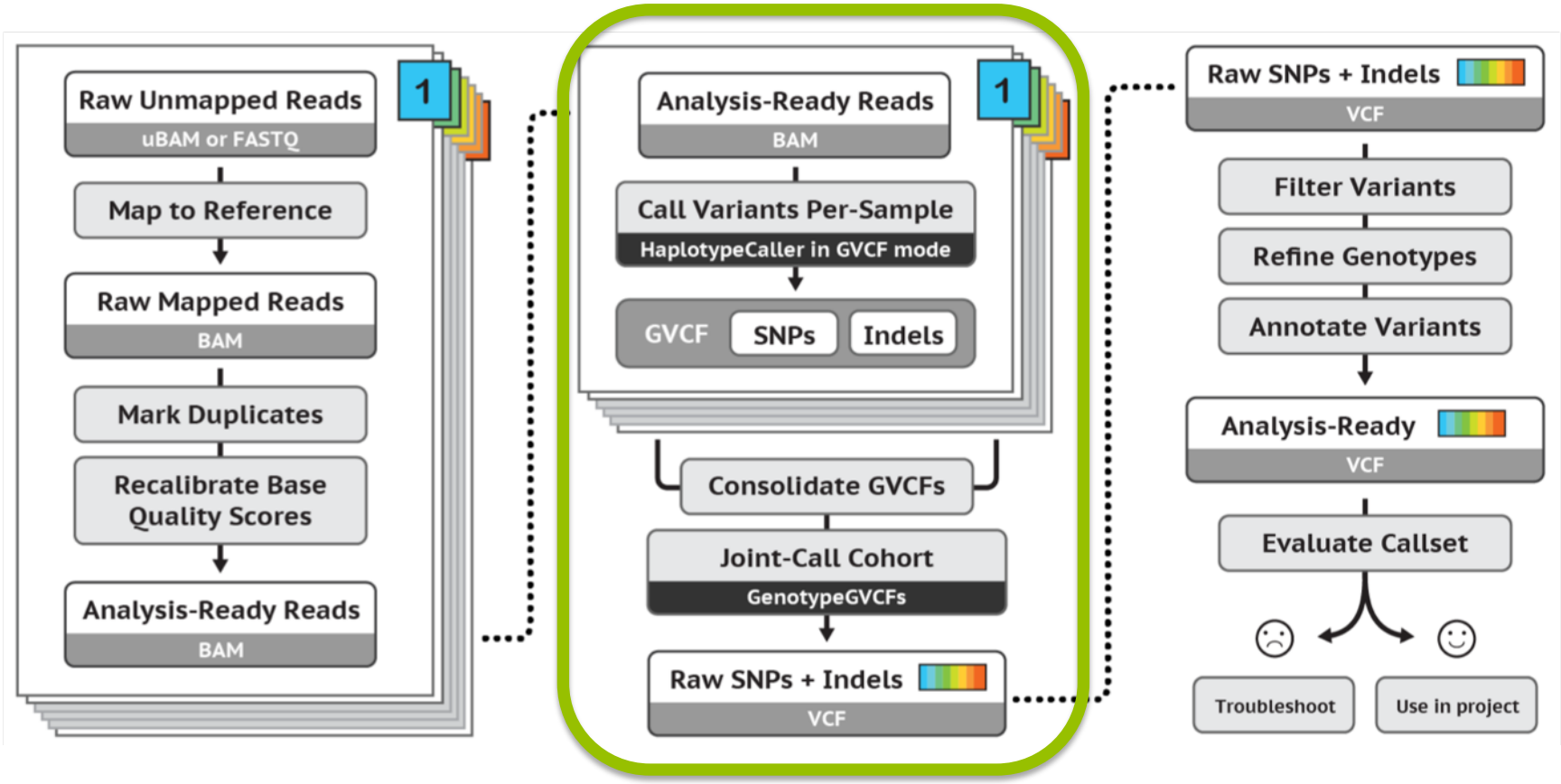
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

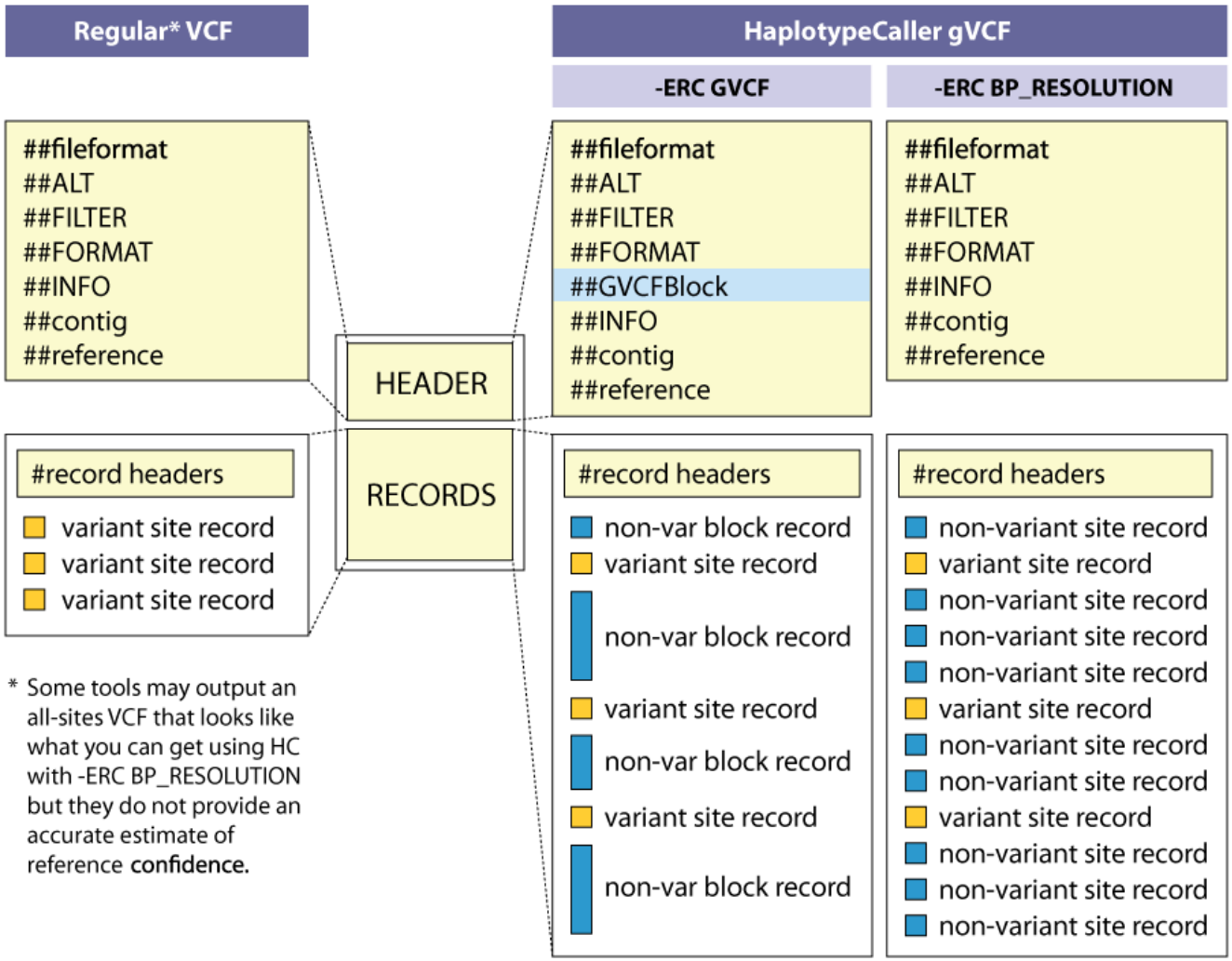
Variant calling in cohort



Joint genotyping of cohort



gVCF Files

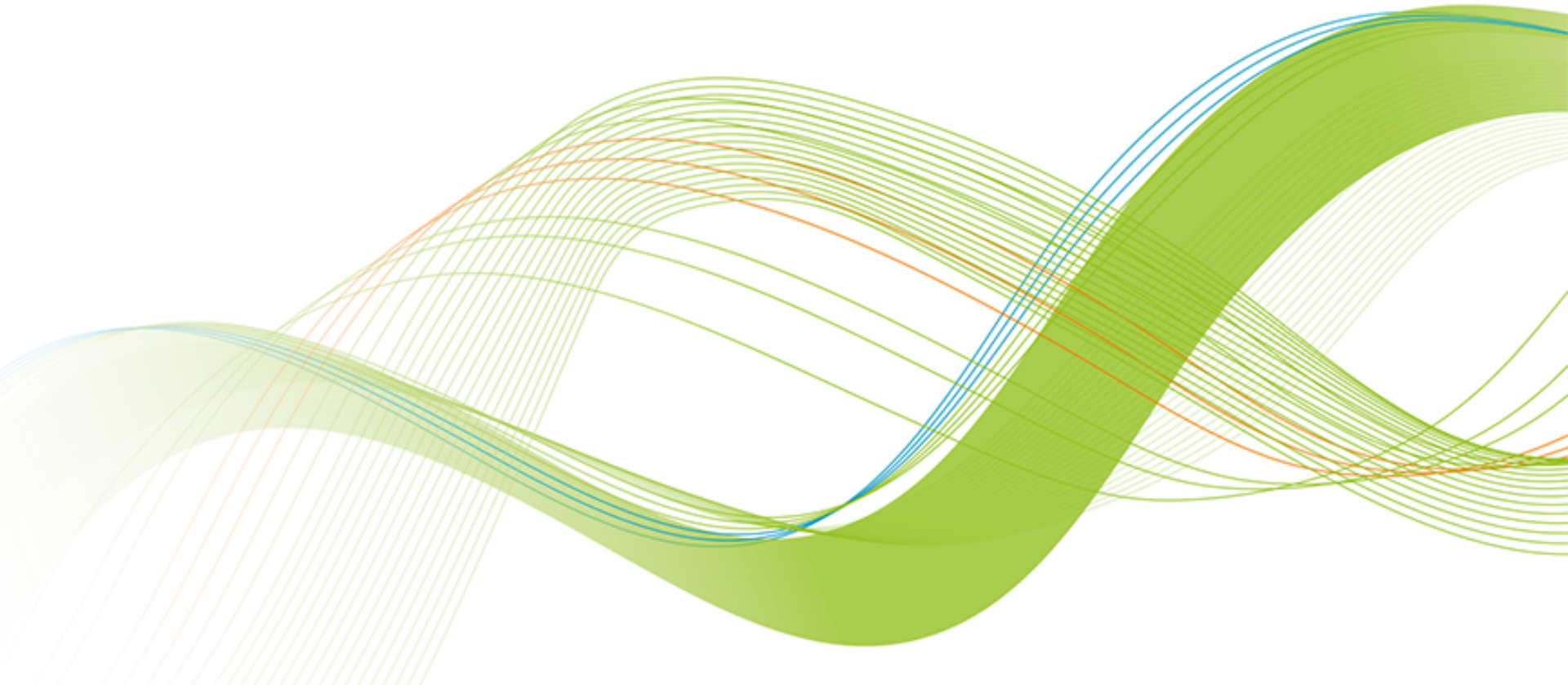

Regular* VCF
HaplotypeCaller gVCF
-ERC GVCF
-ERC BP_RESOLUTION
##fileformat
##ALT
##FILTER
##FORMAT
##INFO
##contig
##reference
##fileformat
##ALT
##FILTER
##FORMAT
##GVCFBlock
##INFO
##contig
##reference
##fileformat
##ALT
##FILTER
##FORMAT
##INFO
##contig
##reference
HEADER
RECORDS
#record headers
variant site record
variant site record
variant site record
#record headers
non-var block record
variant site record
non-var block record
variant site record
non-var block record
variant site record
non-var block record
#record headers
non-variant site record
variant site record
non-variant site record
non-variant site record
non-variant site record
variant site record
non-variant site record
non-variant site record
variant site record
non-variant site record
non-variant site record
non-variant site record
* Some tools may output an all-sites VCF that looks like what you can get using HC with -ERC BP_RESOLUTION but they do not provide an accurate estimate of reference confidence.

Joint variant calling workflow

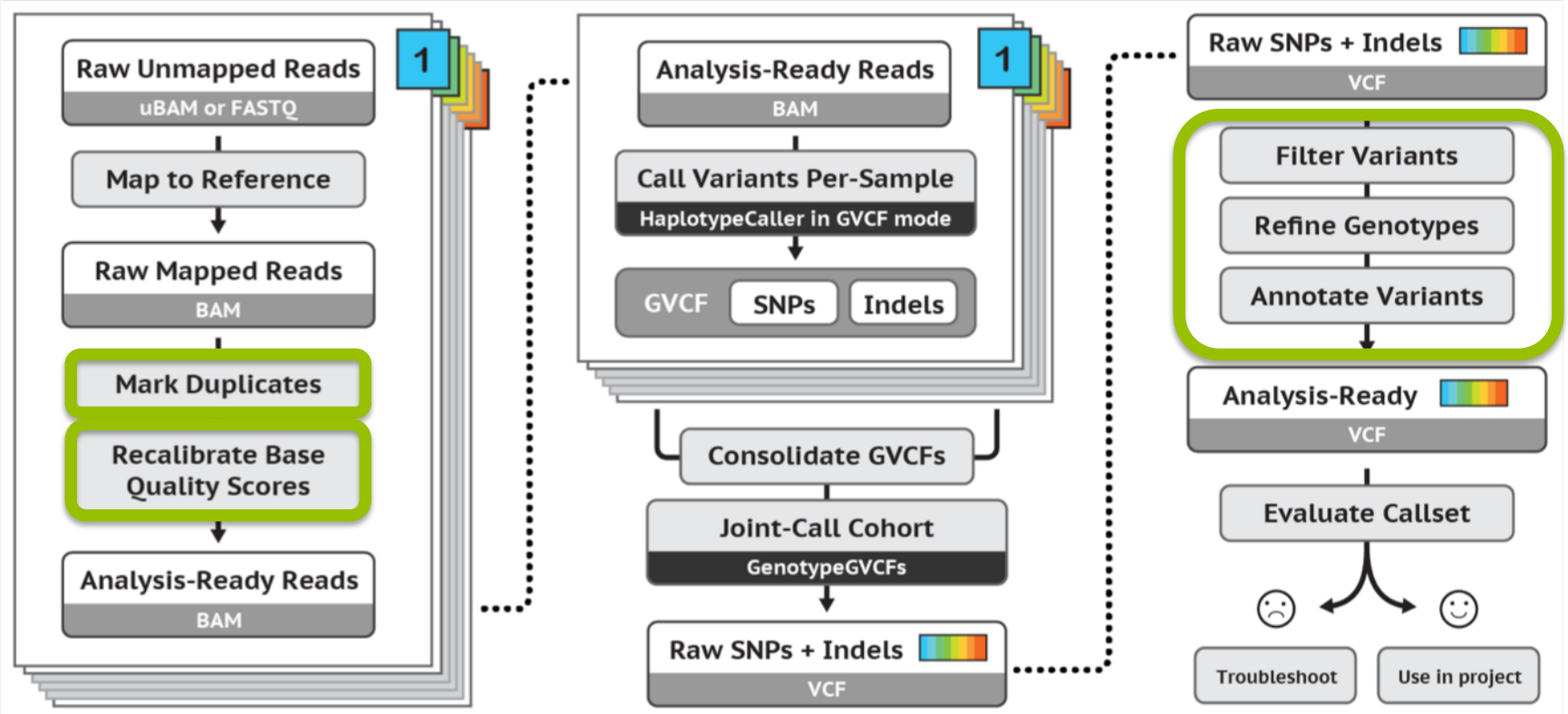
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

GATK's best practices



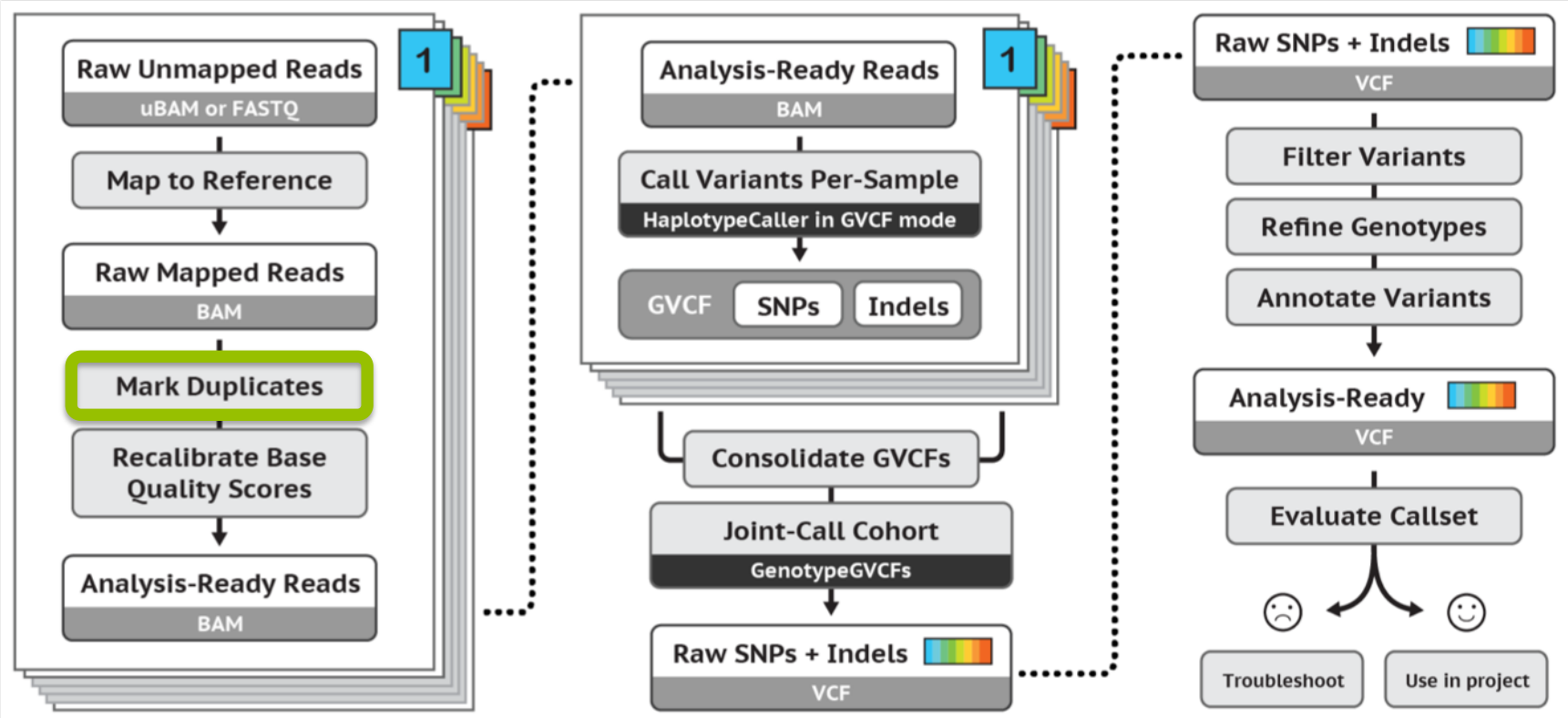
GATK
Best Practice Variant Calling Workflow



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

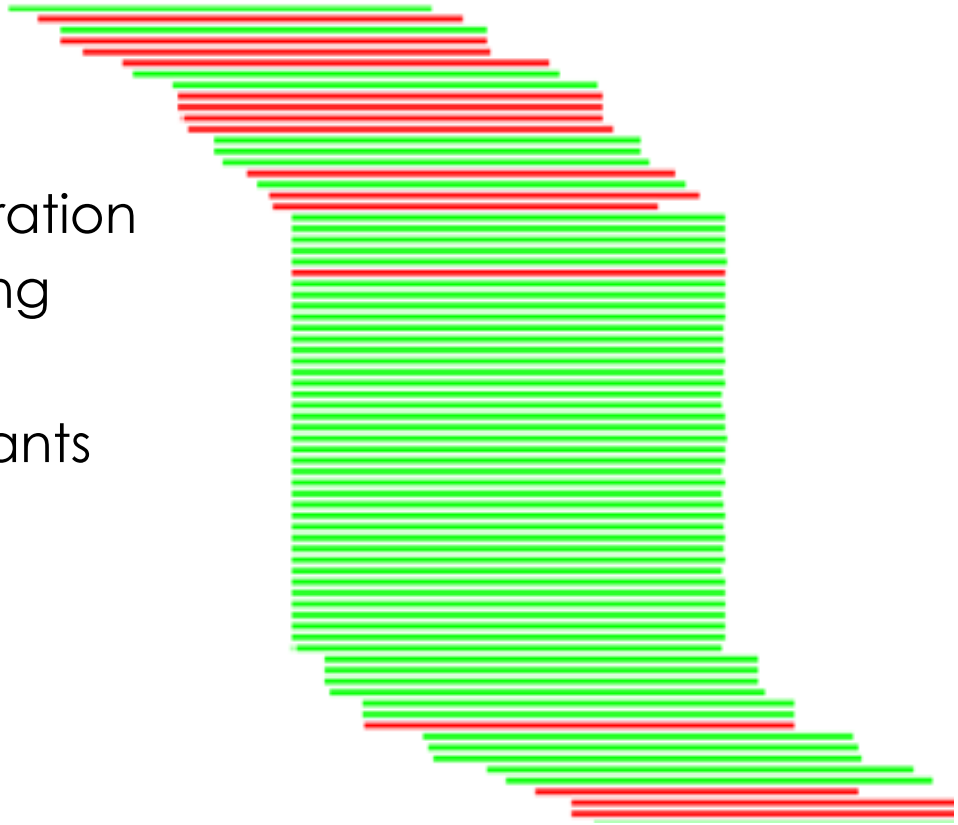
Germline short variant discovery (SNPs + Indels)

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

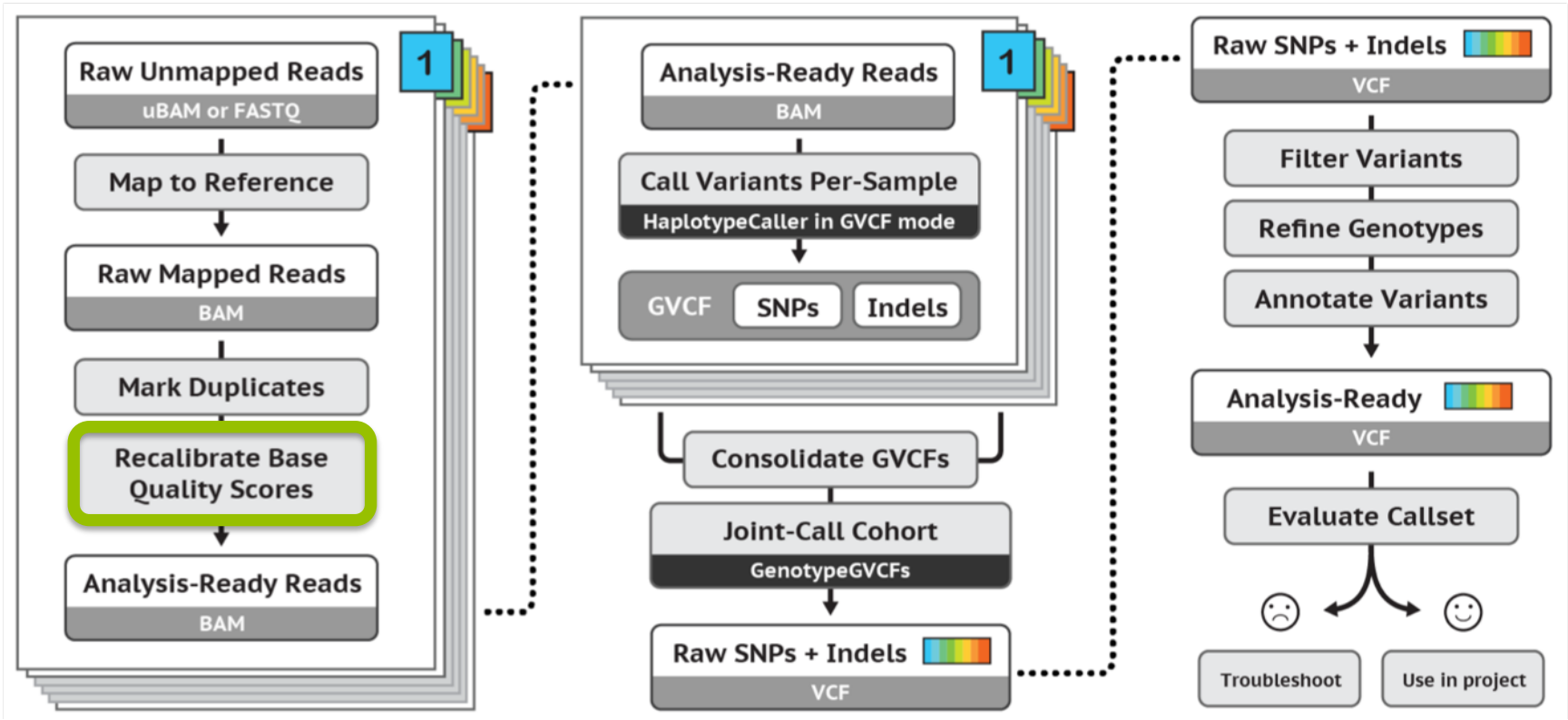


Picard MarkDuplicates

```
Module load Picard

java -Xmx7g -jar $PICARD_HOME/picard.jar MarkDuplicates
INPUT=sample.bam
OUTPUT=sample.md.bam
METRICS_FILE=sample.md.metrix.txt
READ_NAME_REGEX=null
```

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)

Module load GATK

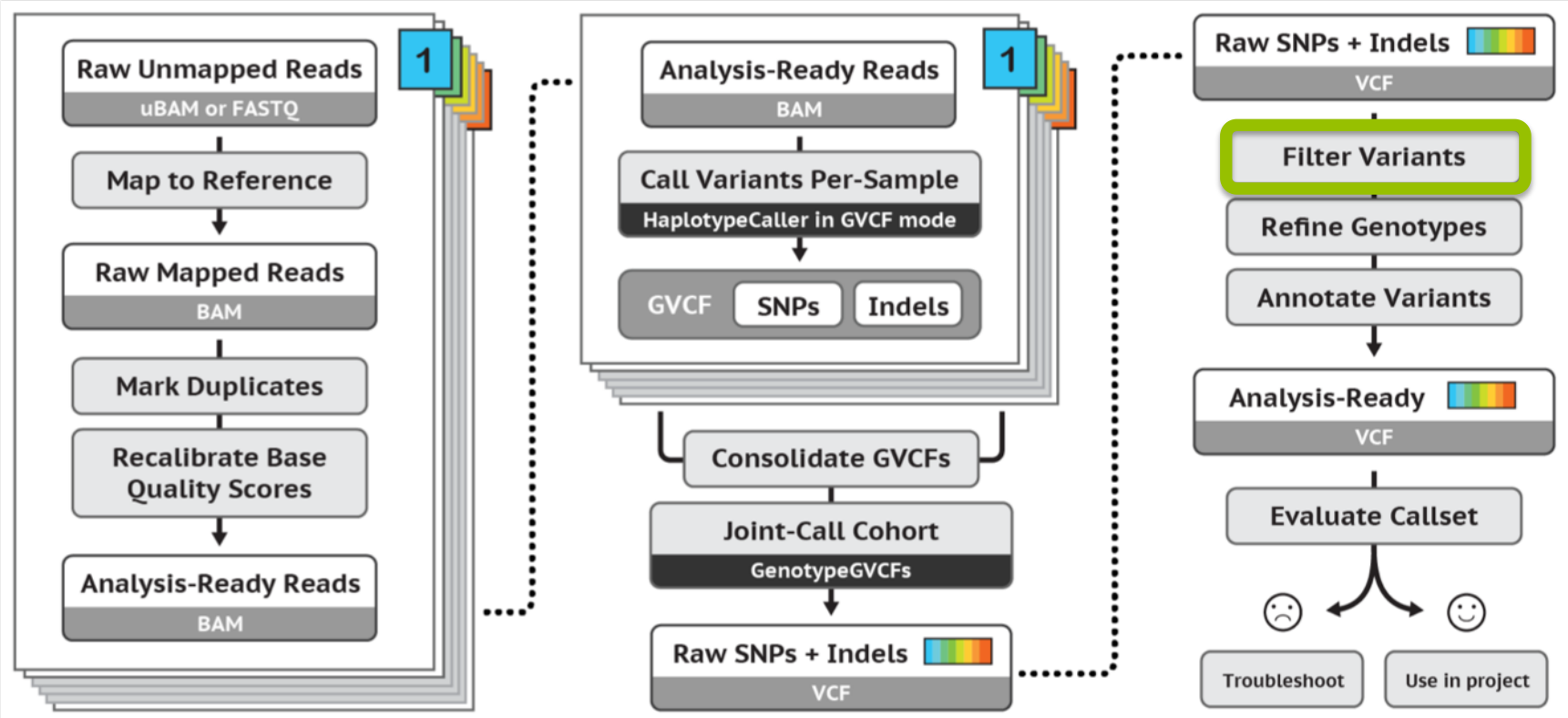
```
java -Xmx7g -jar $GATK_HOME/GenomeAnalysisTK.jar
-T BaseRecalibrator
-R Human_genome.fasta
-I sample.bam
-o sample_calibrationtable.txt
-knownSites high_confidence_SNPs.vcf
```

Generates a recalibration table

```
java -Xmx7g -jar $GATK_HOME/GenomeAnalysisTK.jar
-T PrintReads
-R Human_genome.fasta
-BQSR sample_calibrationtable.txt
-I sample.bam
-o sample.bsqr.bam
```

Applies the recalibration table on the bam file. Generates new recalibrated bam.

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

Filtering SNPs

```
module load GATK

java -Xmx7g -jar $GATK_HOME/GenomeAnalysisTK.jar
-T SelectVariants
-R Human_genome.fasta
-V cohort.vcf
-selectType SNP
-o snps.vcf

java -Xmx7g –jar $GATK_HOME/GenomeAnalysisTK.jar
-T VariantFiltration
-R Human_genome.fasta
-V snps.vcf
-o snps.filtered.vcf
--filterExpression "QD < 2.0"
--filterName QDfilter
--filterExpression "MQ < 40.0"
--filterName MQfilter
--filterExpression "FS > 60.0"
--filterName FSfilter
```

Filtering INDELs

```
module load GATK
```

```
java -Xmx7g -jar $GATK_HOME/GenomeAnalysisTK.jar
-T SelectVariants
-R Human_genome.fasta
-V cohort.vcf
-selectType INDEL
-o indels.vcf
```

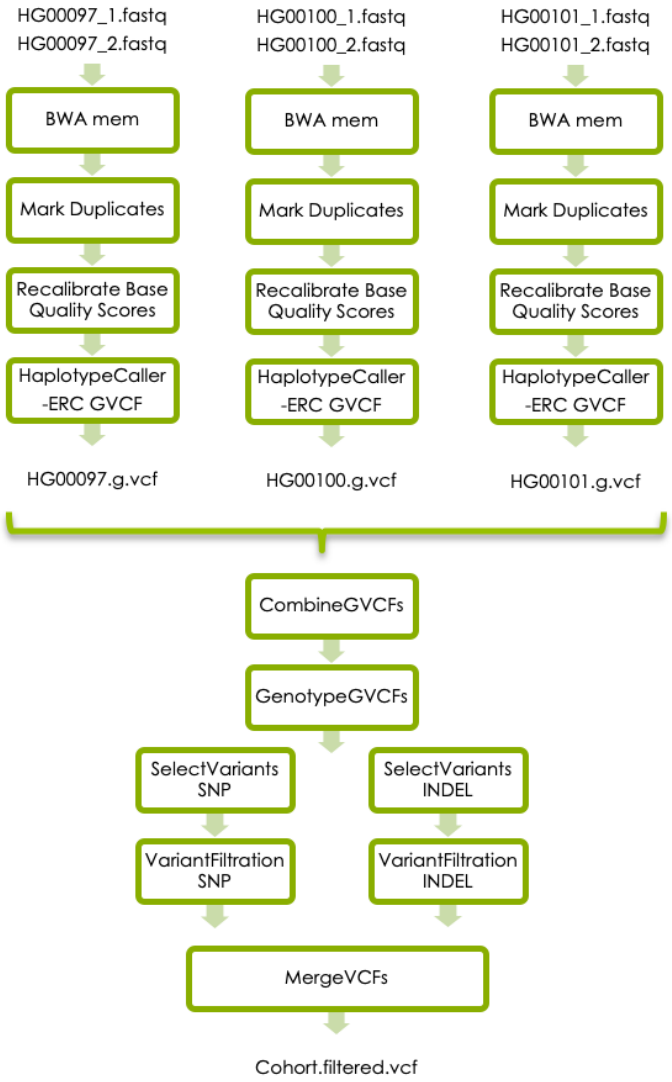
```
java -Xmx7g –jar $GATK_HOME/GenomeAnalysisTK.jar
-T VariantFiltration
-R Human_genome.fasta
-V indels.vcf
-o indels.filtered.vcf
--filterExpression "QD < 2.0"
--filterName QDfilter
--filterExpression "FS > 200.0"
--filterName FSfilter
```

Merge filtered SNPs and INDELs

- Module load GATK

- java -Xmx7g -jar $GATK_HOME/GenomeAnalysisTK.jar
- -T CombineVariants
- -R Human_genome.fasta
- --variant:snp snps.filtered.vcf
- --variant:indel indels.filtered.vcf
- -o snps.indels.filtered.vcf
- -genotypeMergeOptions PRIORITIZE
- -priority snp,indel

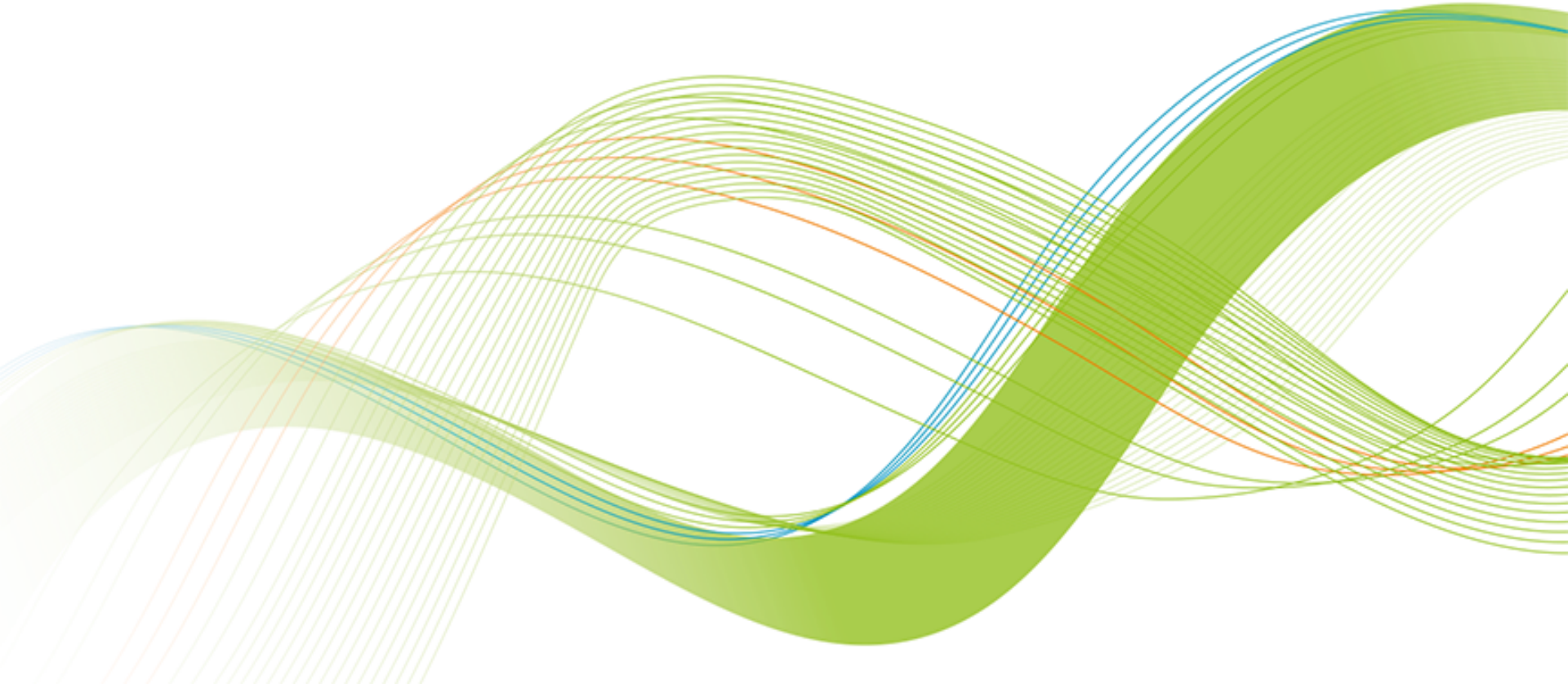
GATK best practises



NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Introduction to workshop



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

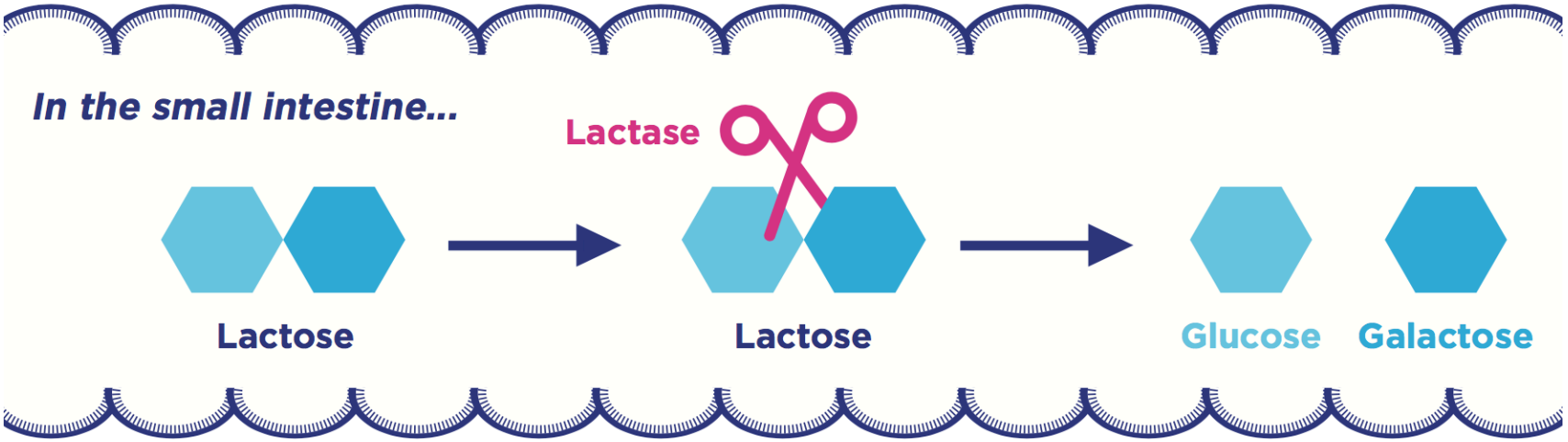
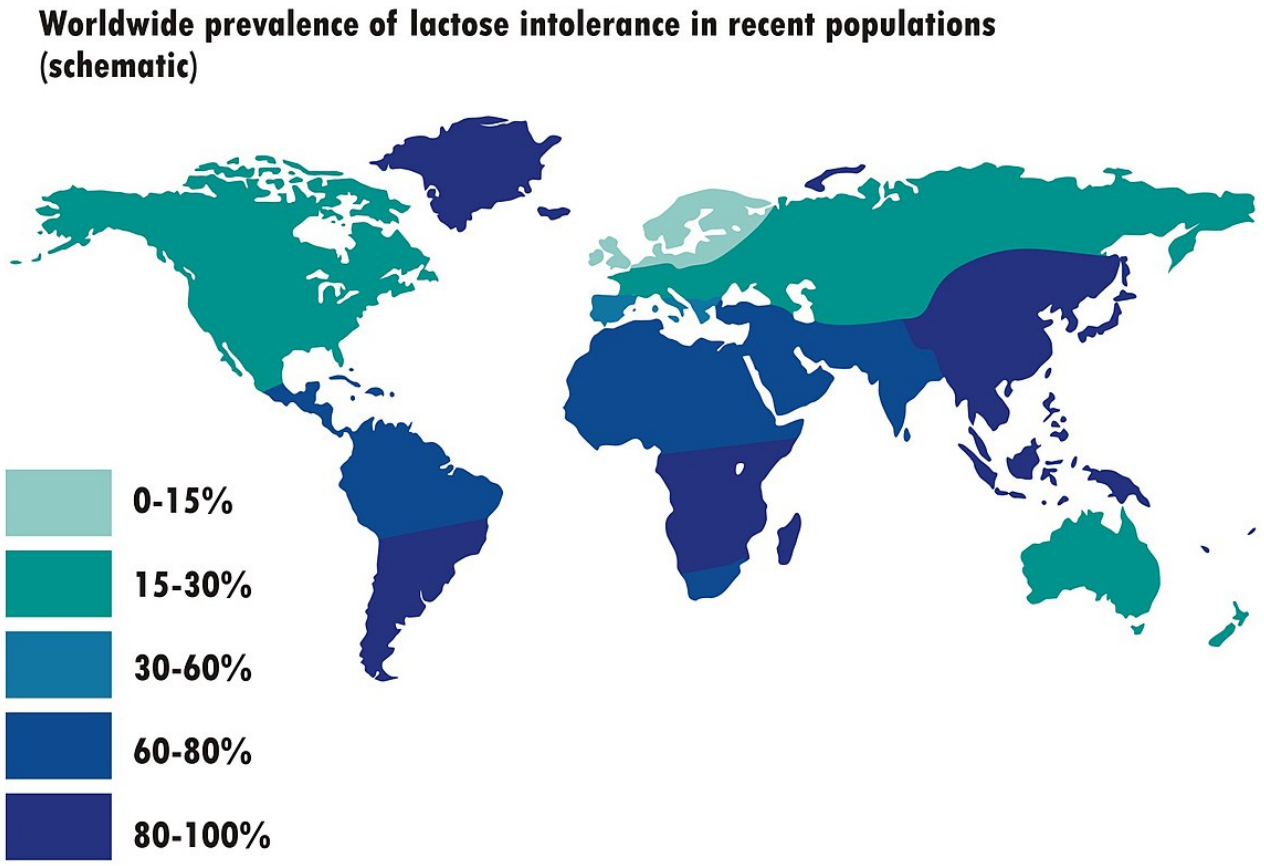


Figure 2. Lactose digestion in the intestine.

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

Prevalence of lactose intolerance



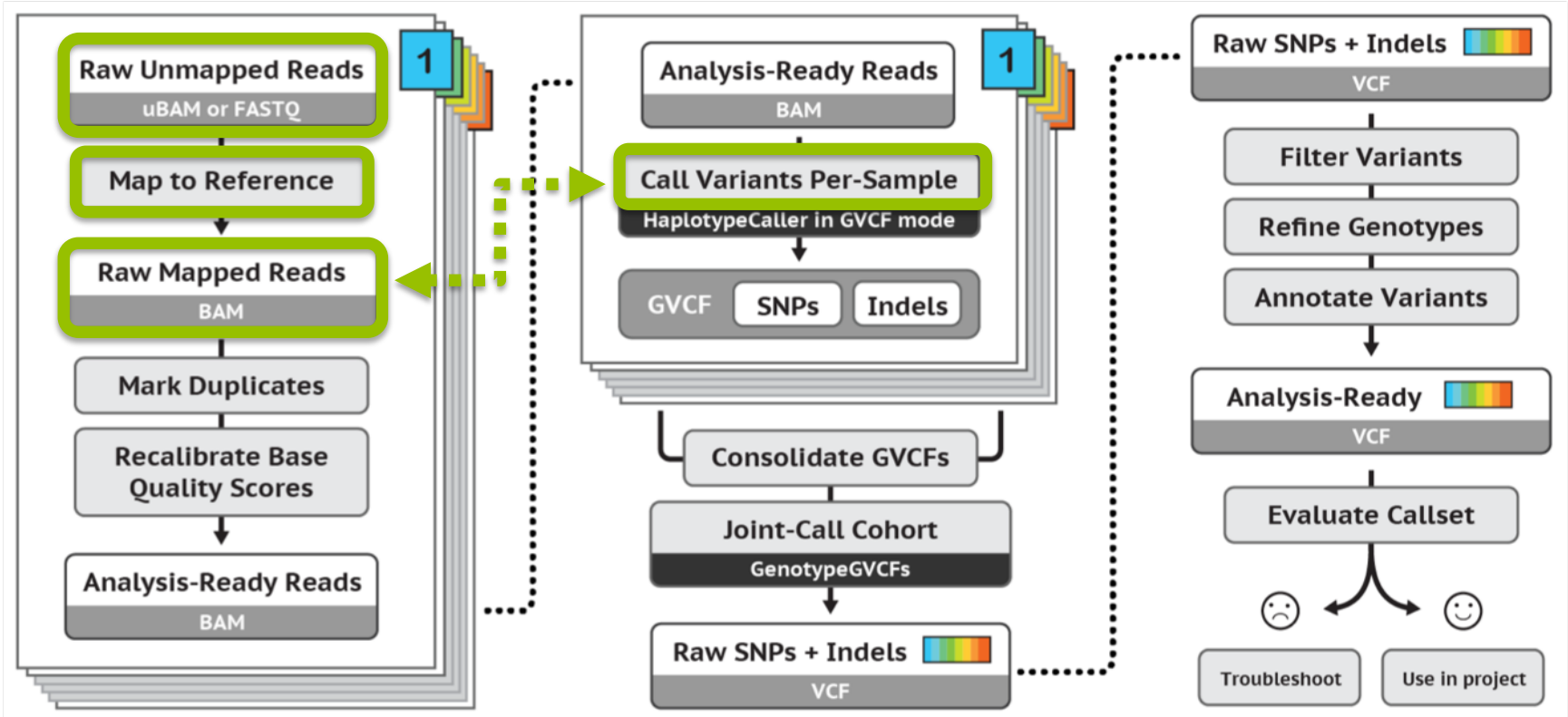
NBIS SciLifeLab

part one:

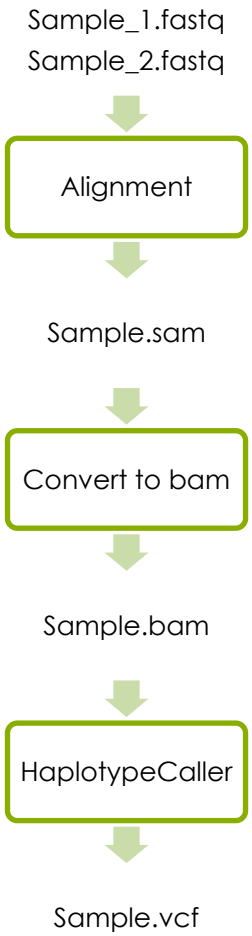
variant calling in one sample

Basic variant calling workflow

- For learning the concepts!
- The minimum steps needed to go from fastq files to vcf file for one sample



Workshop part one

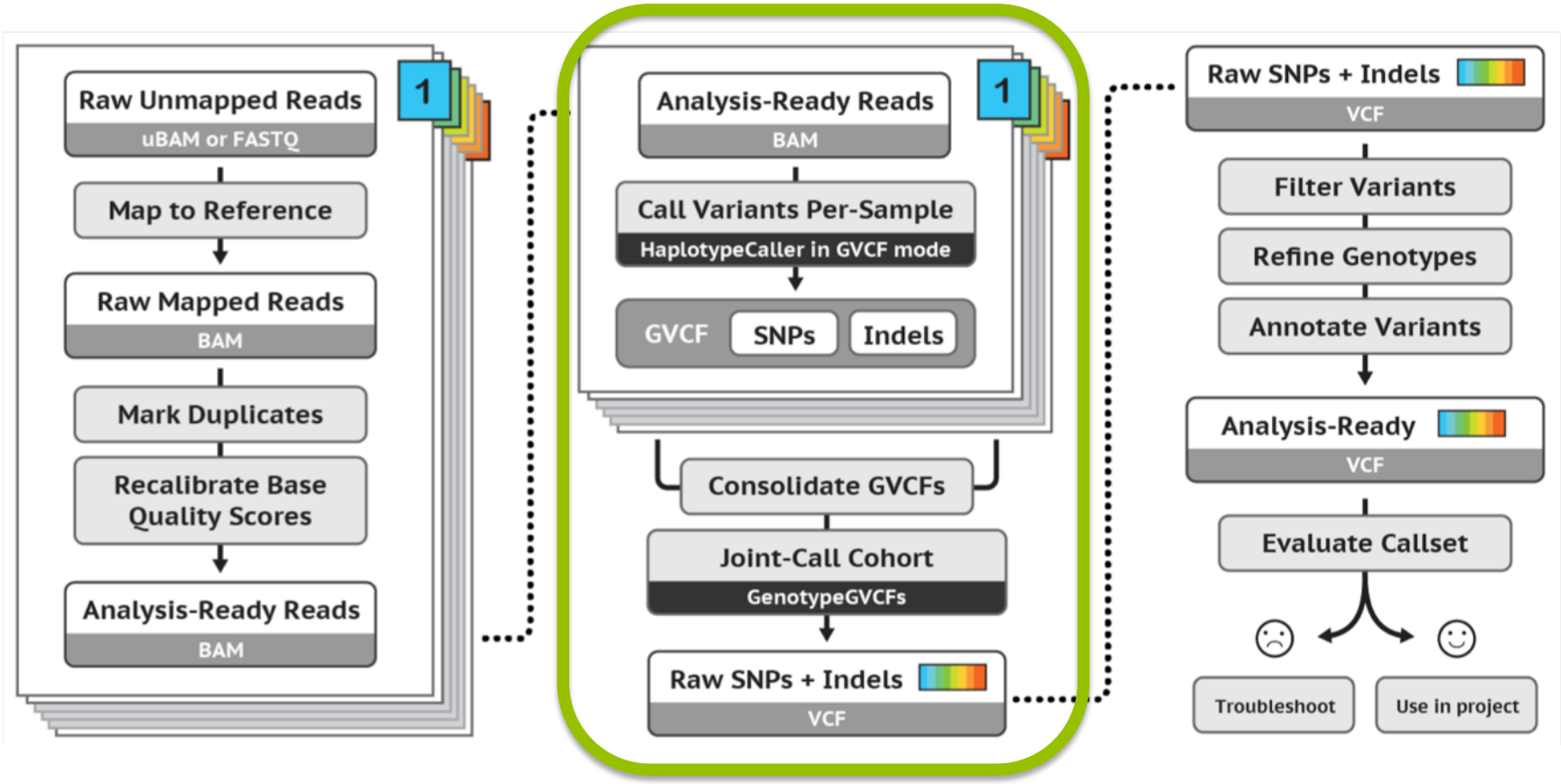


NBIS SciLifeLab

Part two (if you have time):

variant calling in cohort

Joint genotyping of cohort



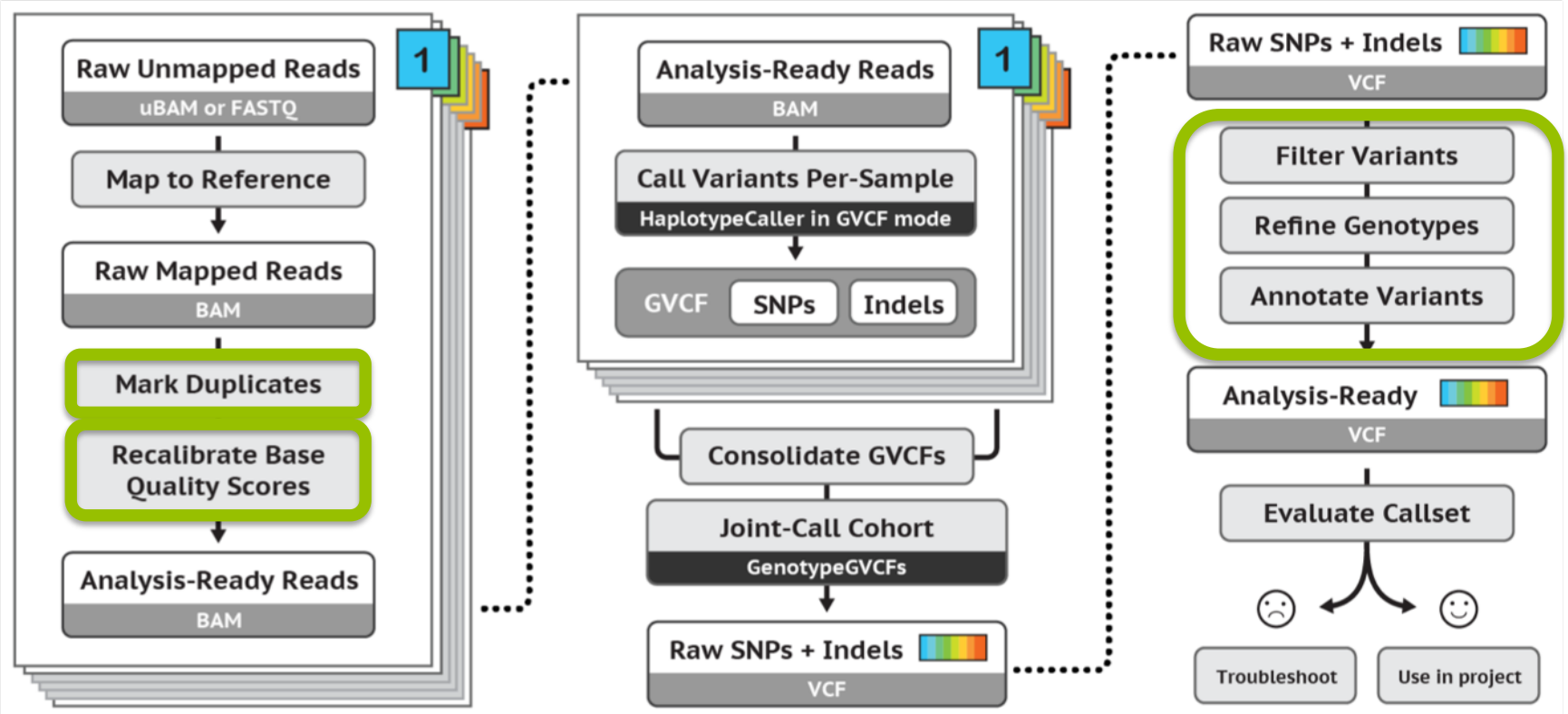
Joint variant calling workflow

NBIS SciLifeLab

Part three (if you have time):

Follow GATK best practices for short variant discovery

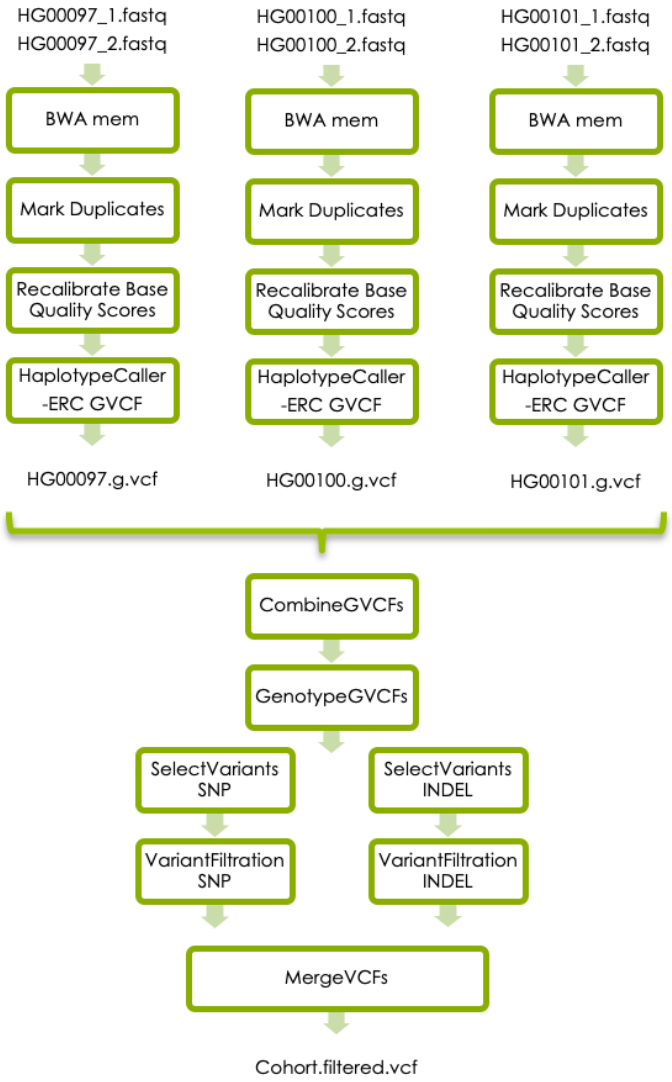
GATK
Best Practice Variant Calling Workflow



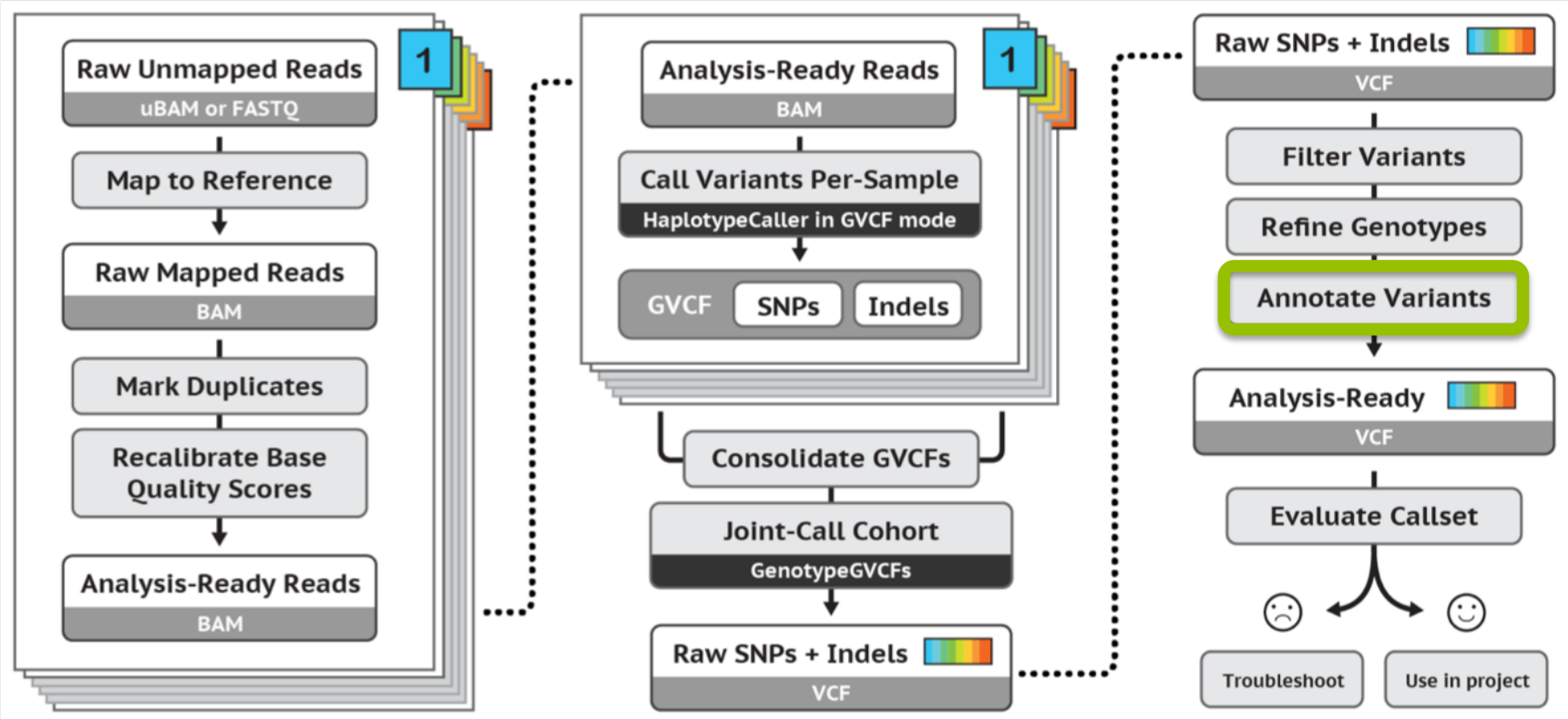
https://software.broadinstitute.org/gatk/best-practices/

Germline short variant discovery (SNPs + Indels)

GATK best practises



Not includer in workshop: Annotation!



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

Germline short variant discovery (SNPs + Indels)

NBIS SciLifeLab

Questions?