From raw reads to variants

Anna Johansson, NBIS

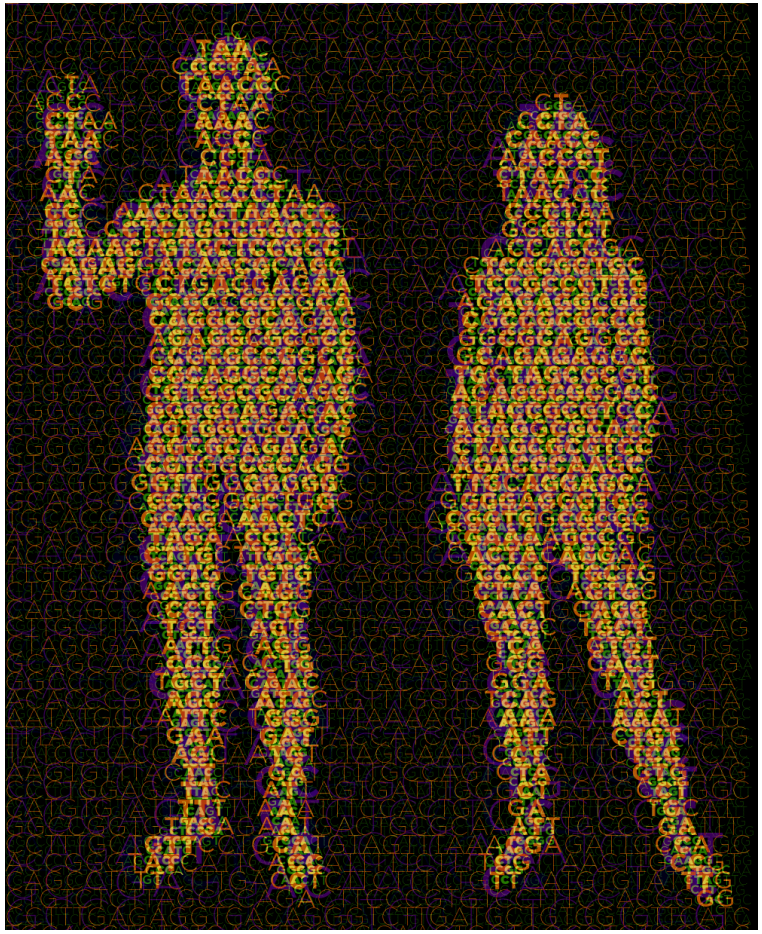
Uppsala, September 2018

Talk Overview

- Concepts
  - Reference genome
  - Variants
  - Paired-end data
- NGS Workflow
  - Quality control & Trimming
  - Alignment
  - Local realignment
  - PCR duplicates & removal
  - Base Quality Score Recalibration
  - Variant calling
- VCF files
- Joint genotyping & gVCF files
- Annotation & Filtering

Reference genome

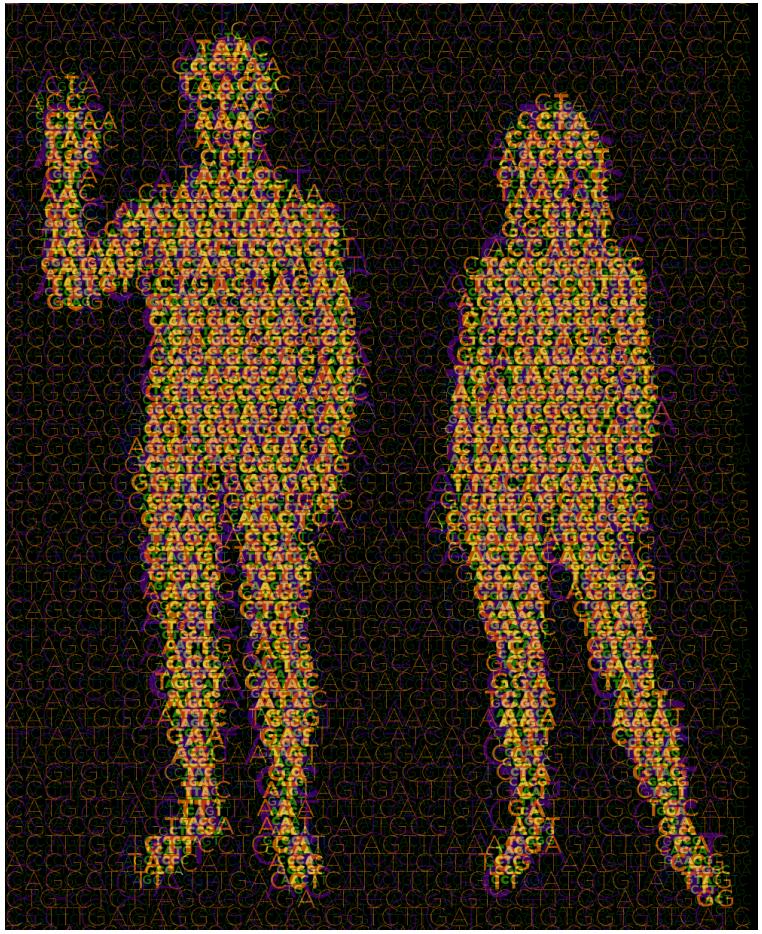
- Genome Reference Consortium
- A mosaic nucleic acid sequence
  - ...GTGCGTAGACTGCTAGATCGAAGA...



Reference genome

- Genome Reference Consortium
- A mosaic nucleic acid sequence
  - ...GTGCGTAGACTGCTAGATCGAAGA...

- What changes between versions?
  - First version: 150,000 gaps
  - HG19: 250 gaps



Variants

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

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...

Variants

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

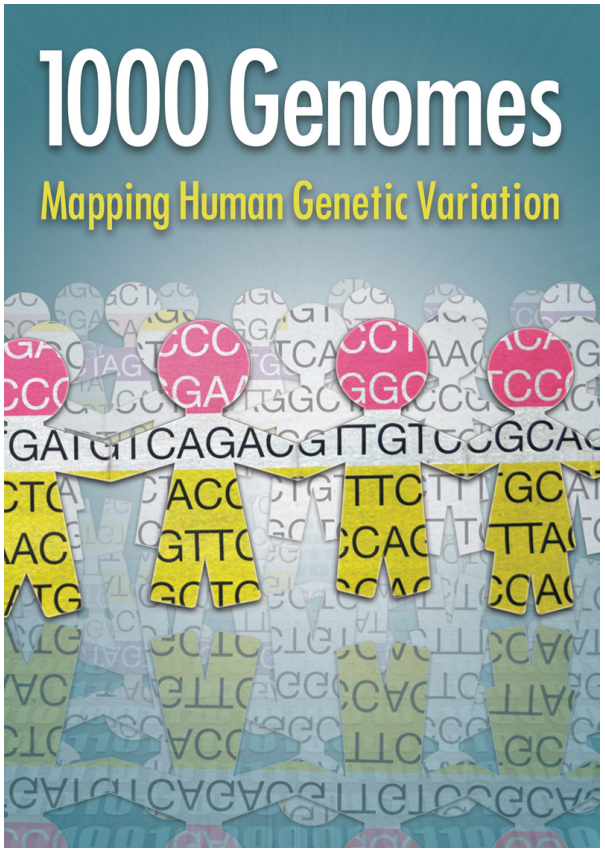
Reference: ...GTGCGTAGACTGCTAGATCGAAGA...
Sample: ...GTGCGTAGACTGATAGATCGAAGA...

Variants

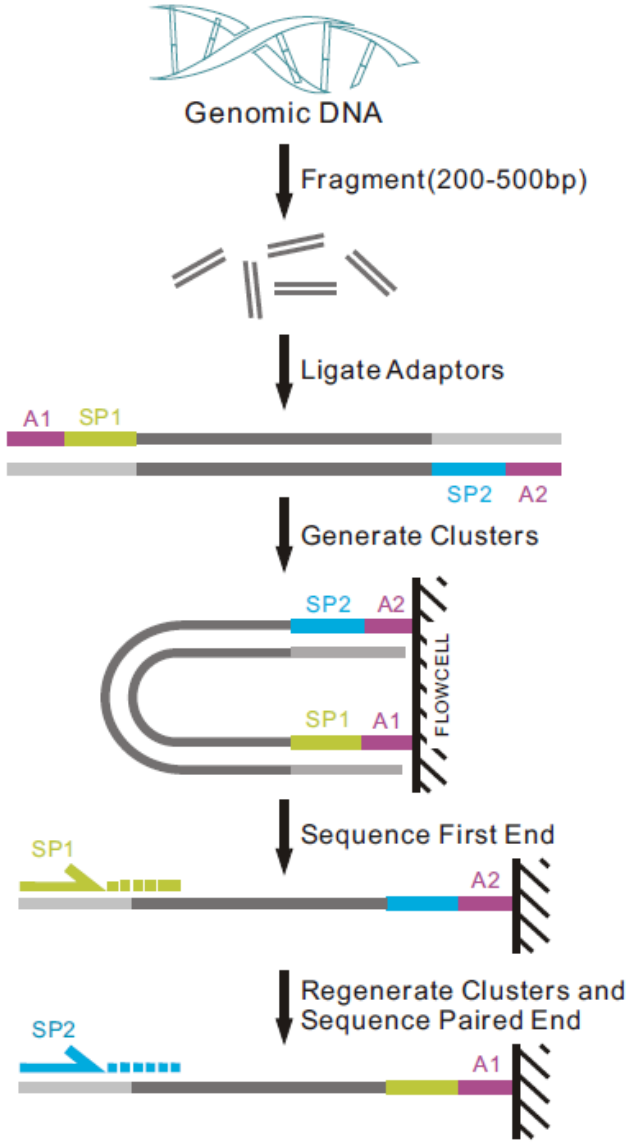
Population based variant projects

UK 10K

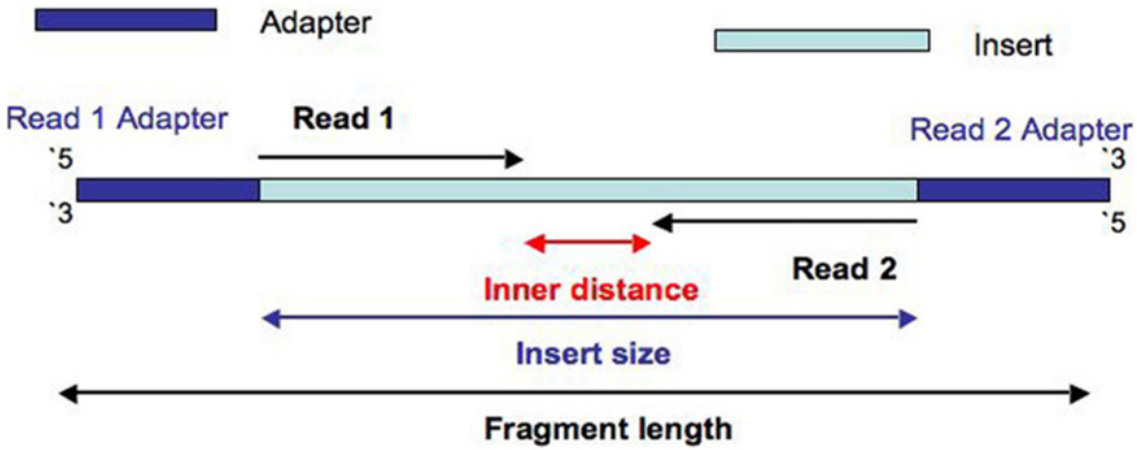
RARE GENETIC VARIANTS IN HEALTH AND DISEASE



Paired-end sequencing



Paired-end data



Illumina sequencing

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

Paired-end data

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

ID_R1_001.fastq

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

ID_R2_001.fastq

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

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.

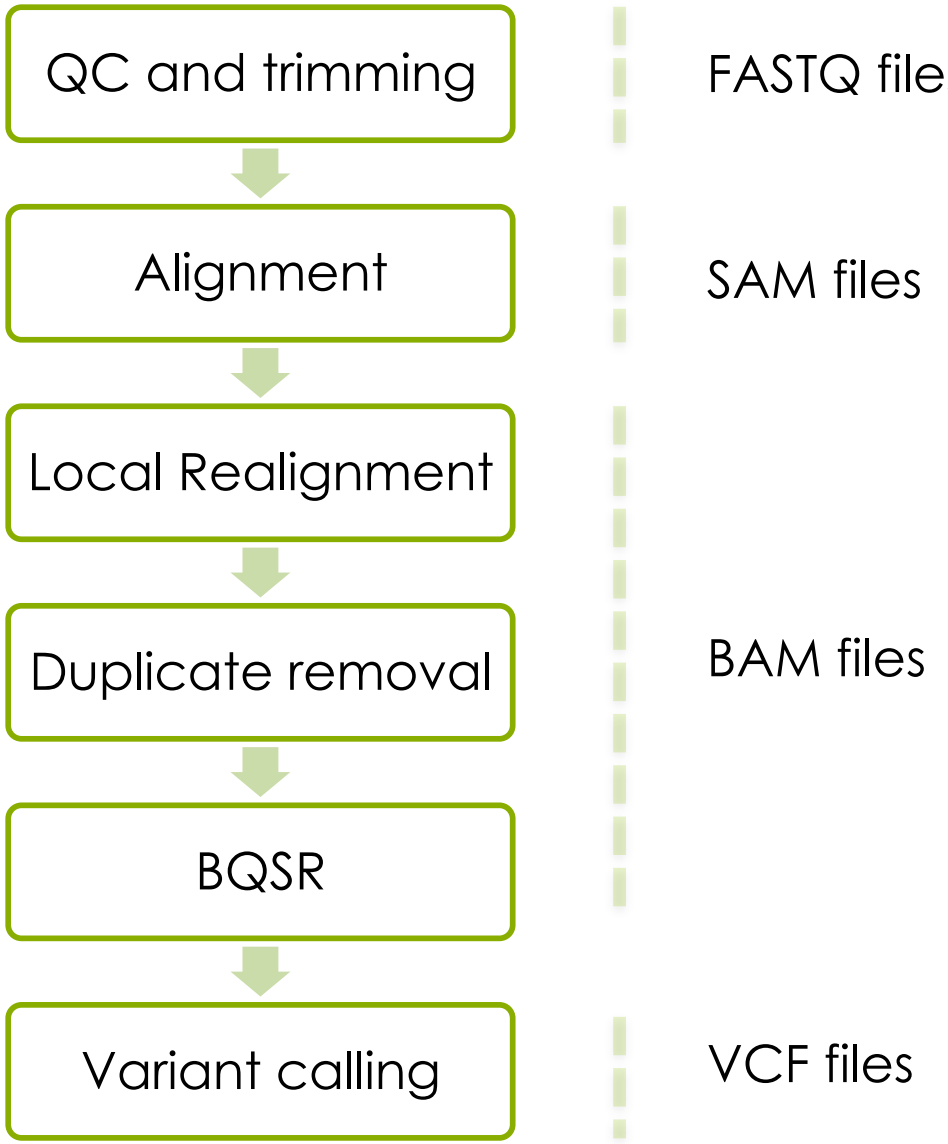
ID_R1_001.fastq

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

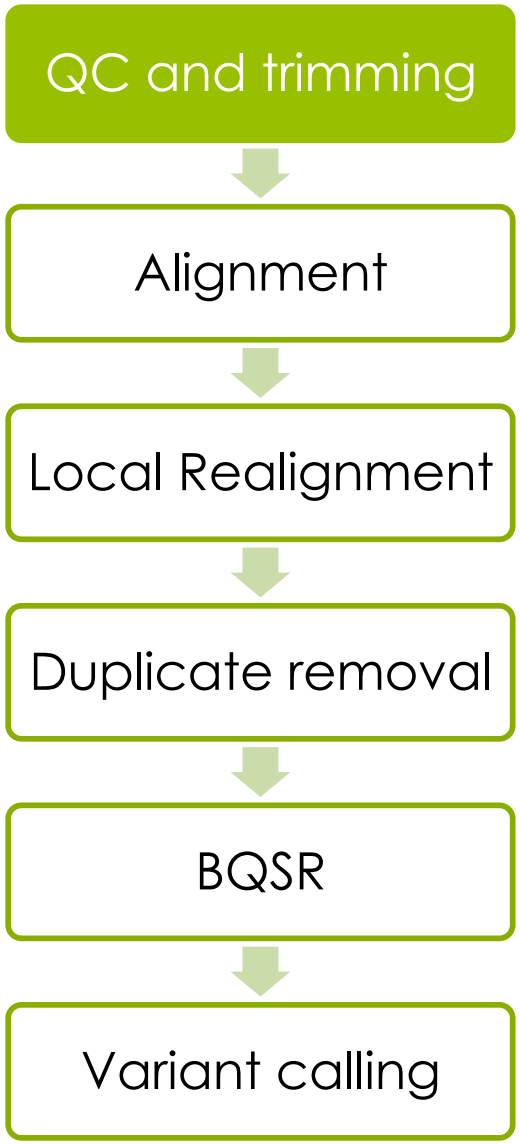
ID_R2_001.fastq

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

NGS workflow



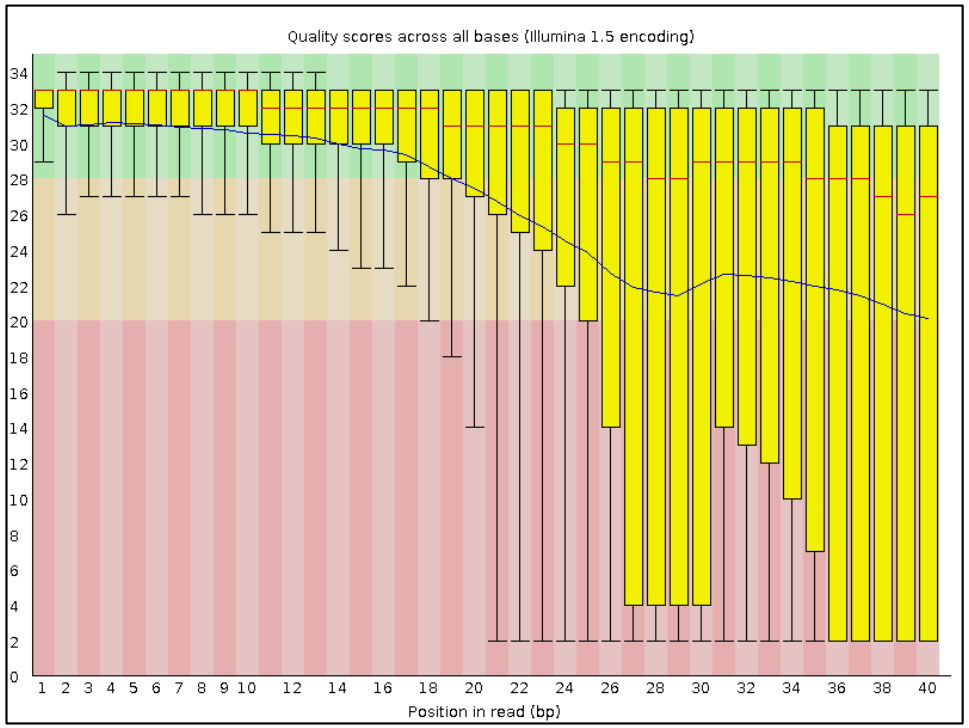
NGS workflow



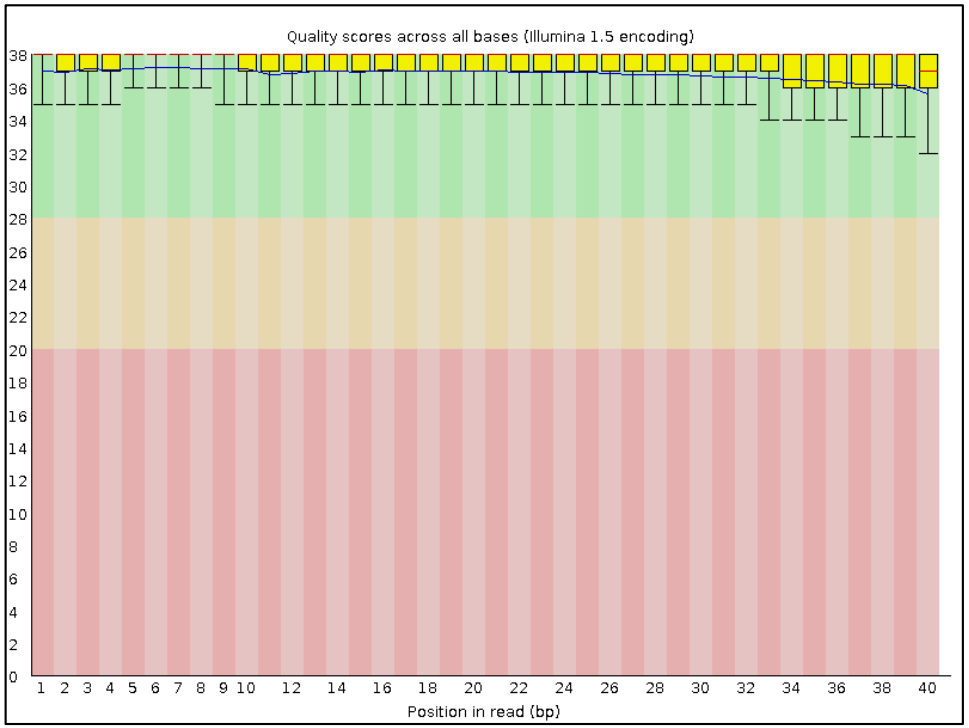
Quality control

```
module load FastQC
```

Bad qualities:



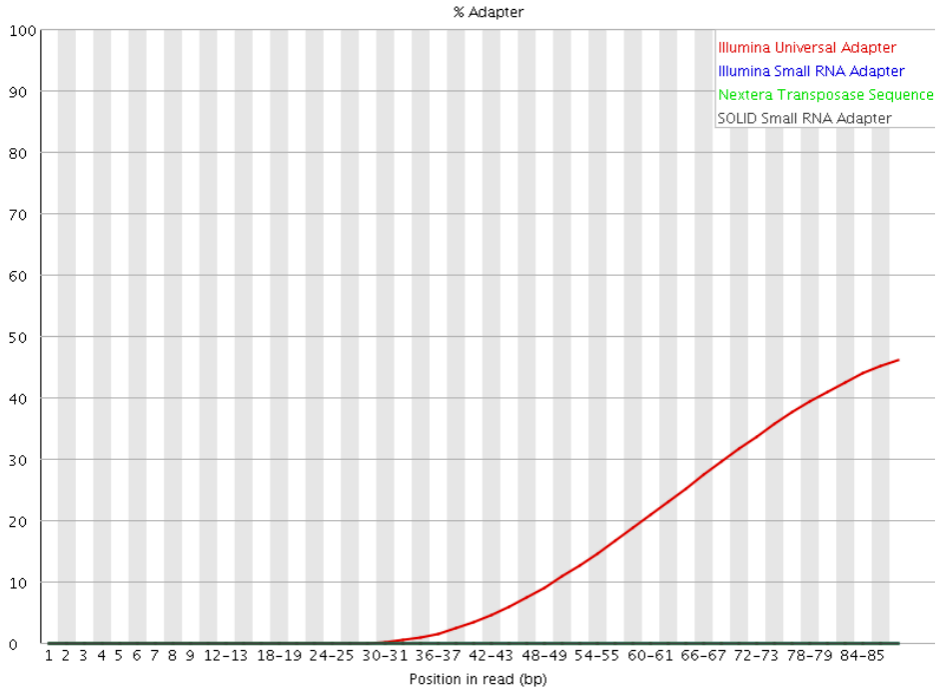
Good qualities:



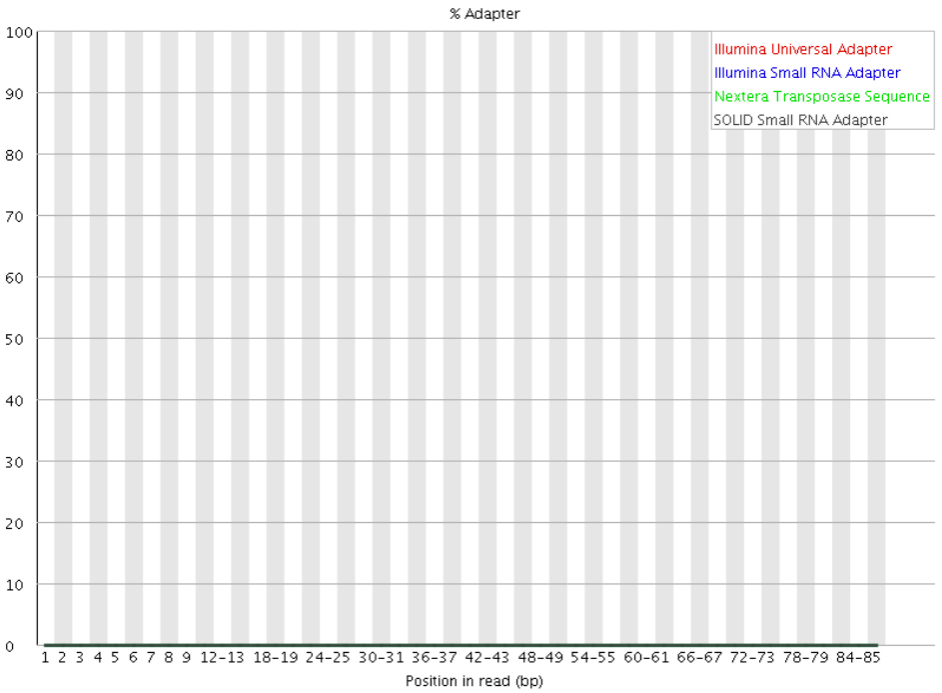
Quality control

```
module load FastQC
```

Adapters present:

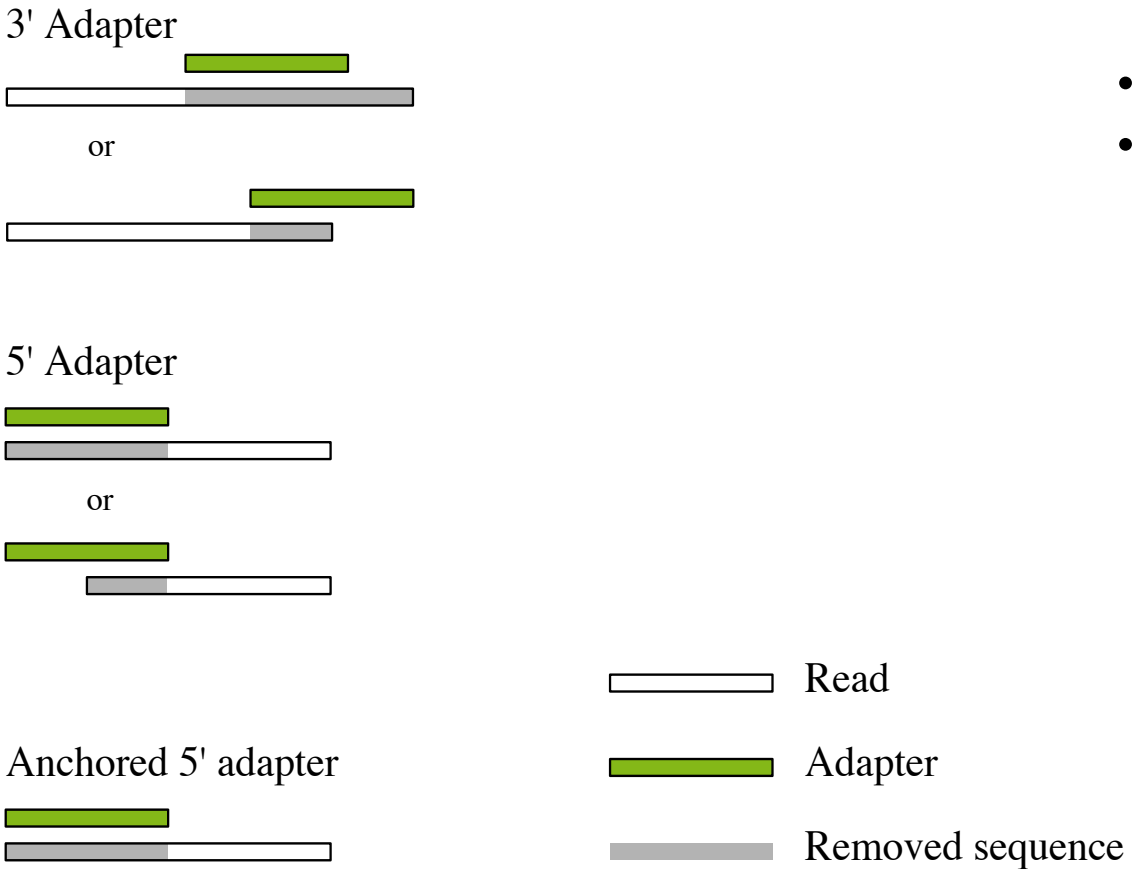


Adapters Absent:



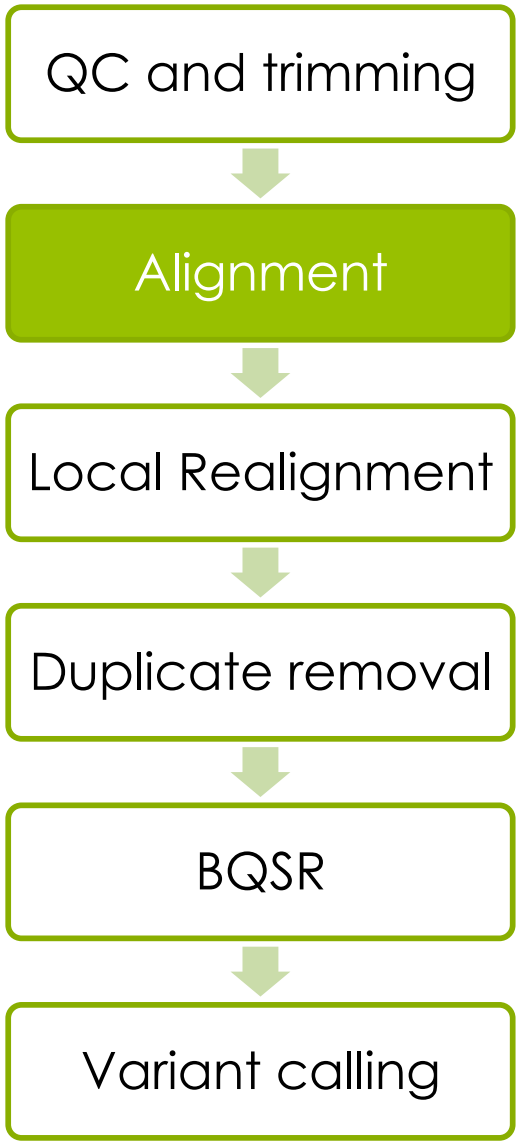
Trimming

`module load cutadapt / TrimGalore / trimmomatic`

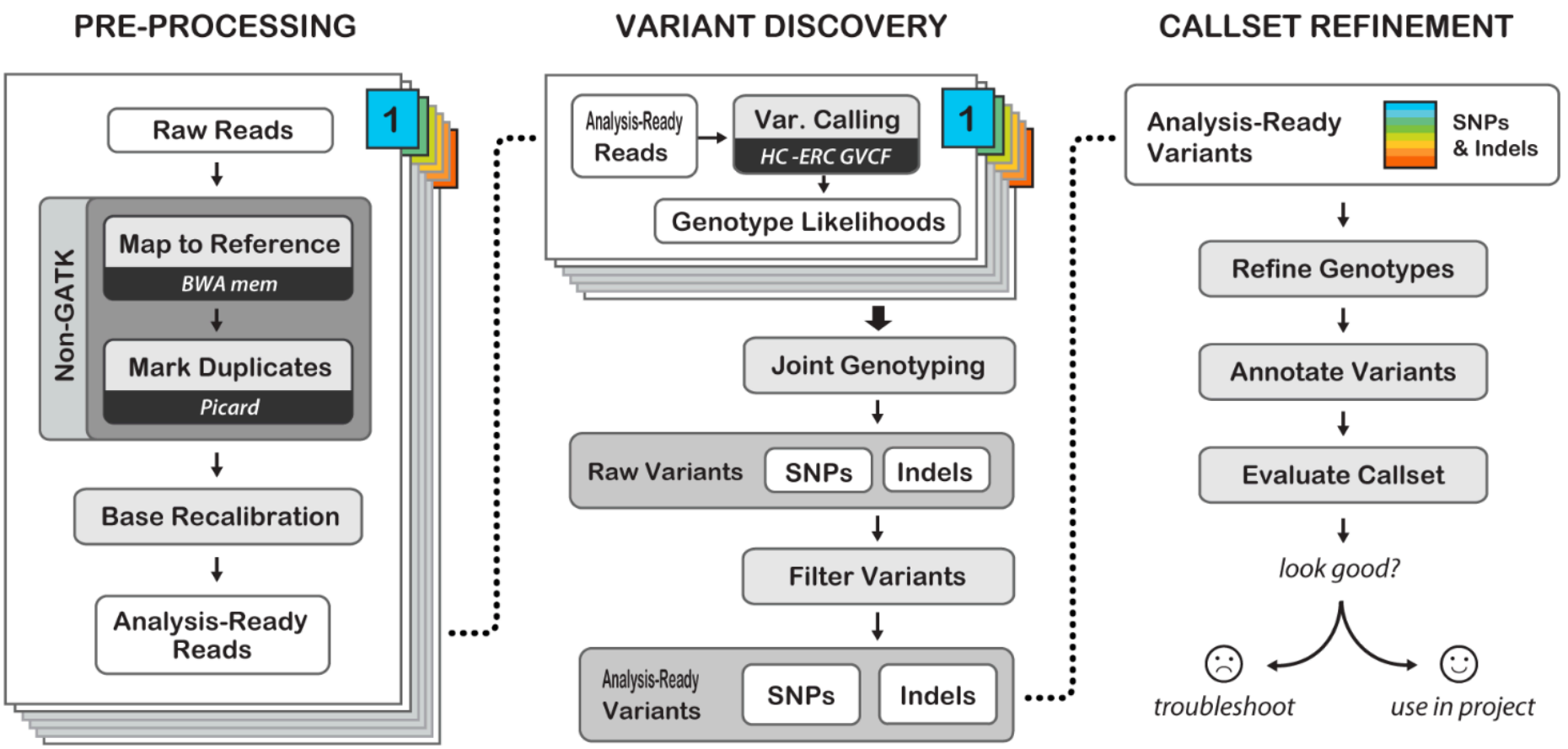


- Remove bad quality reads
- Remove adapters

NGS workflow



GATK Best Practices



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

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

Alignment

```
module load bwa
```

Read	TCGATCC
Reference	GACCTCATCGATCCCACTG

Alignment

```
module load bwa
```

Read TCGATCC

Reference GACCTCATCGATCCCACTG

Read TCGATCC

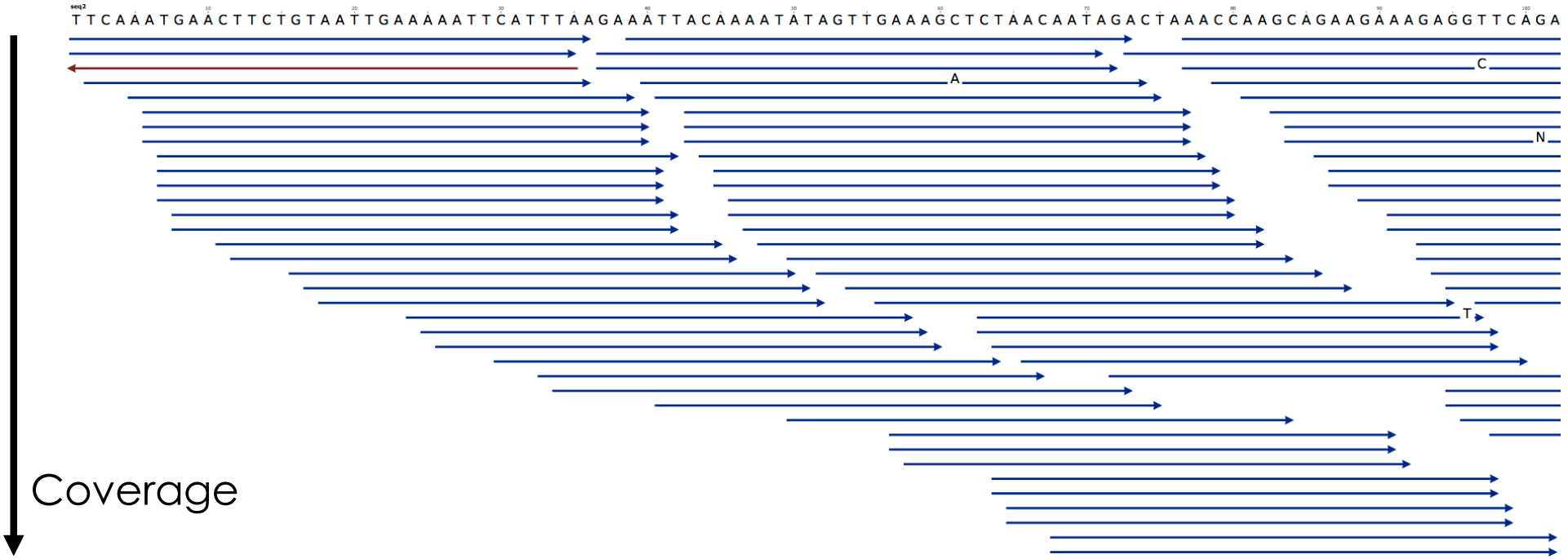
Reference GACCTCATCGATCCCACTG

Alignment

```
module load bwa
```

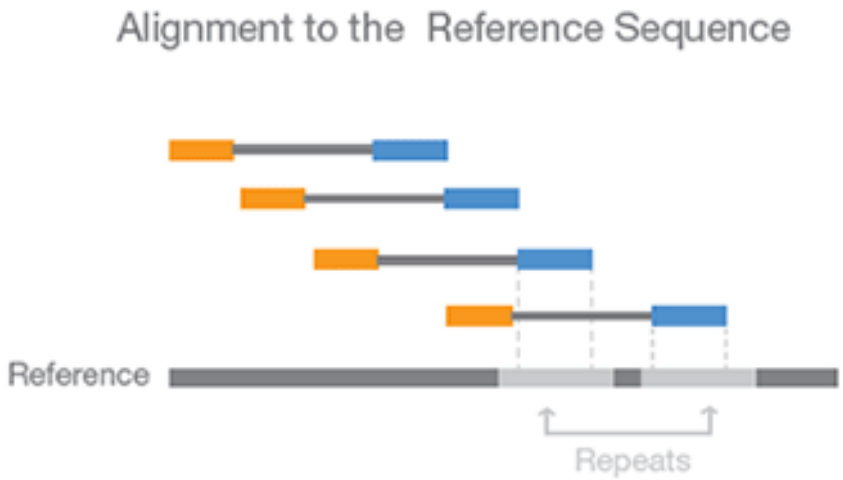
Alignment

```
module load bwa
```



Paired-end data & Alignment

The known distance between paired reads allows improved mapping over repeat regions



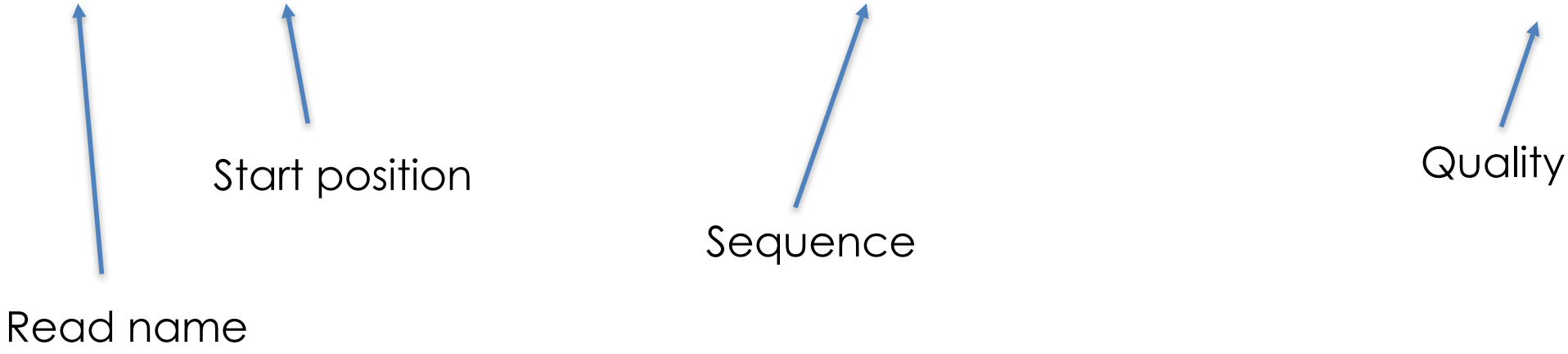
Output from mapping - Sam format

HEADER SECTION

```
@HD VN:1.0 SO:coordinate
@SQ SN:1 LN:249250621 AS:NCBI37 UR:file:/data/local/ref/GATK/human_g1k_v37.fasta M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ SN:2 LN:243199373 AS:NCBI37 UR:file:/data/local/ref/GATK/human_g1k_v37.fasta M5:a0d9851da00400dec1098a9255ac712e
@SQ SN:3 LN:198022430 AS:NCBI37 UR:file:/data/local/ref/GATK/human_g1k_v37.fasta M5:fdfd811849cc2fadebc929bb925902e5
@RG ID:UM0098:1 PL:ILLUMINA PU:HWUSI-EAS1707-615LHAAXX-L001 LB:80 DT:2010-05-05T20:00:00-0400 SM:SD37743 CN:UMCORE
@RG ID:UM0098:2 PL:ILLUMINA PU:HWUSI-EAS1707-615LHAAXX-L002 LB:80 DT:2010-05-05T20:00:00-0400 SM:SD37743 CN:UMCORE
@PG ID:bwa VN:0.5.4
```

ALIGNMENT SECTION

```
8_96_444_1622 73 scaffold00005 155754 255 54M * 0 0 ATGTAAAGTATTTCCATGGTACACAGCTTGGTCGTAATGTGATTGCTGAGCCAG C@B5)5CBBCCBCCCBC@@7
8_80_1315_464 81 scaffold00005 155760 255 54M = 154948 0 AGTACCTCCCTGGTACACAGCTTGGTAAAAATGTGATTGCTGAGCCAGACCTTC B?@?BA=>@>>7;AB
8_17_1222_1577 73 scaffold00005 155783 255 40M1116N10M * 0 0 GGTAAAAATGTGATTGCTGAGCCAGACCTTCATCATGCAGTGAGAGACGC BB@BA??>CCBA2AA
8_43_1211_347 73 scaffold00005 155800 255 23M1116N27M * 0 0 TGAGCCAGACCTTCATCATGCAGTGAGAGACGCAAACATGCTGGTATTTG #>8<=<@6/:@9';@7
8_32_1091_284 161 scaffold00005 156946 255 54M = 157071 0 CGCAAACATGCTGGTAGCTGTGACACCACATCAACAGCTTGACTATGTTTGTAA BBBBB@AABACBCA
```



Read groups

- Link information of *sample id*, *library prep*, *flowcell* and *sequencing runs* to each read.
- Good for error tracking!
- 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

Convert to Bam

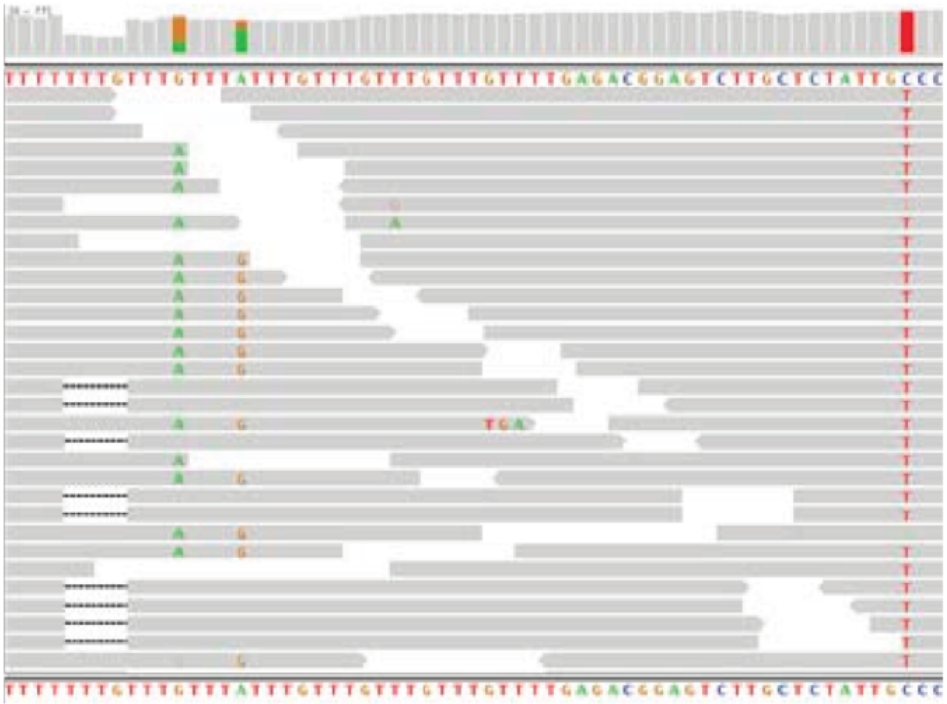
Bam file is a binary representation of the Sam file

NGS workflow

Local realignment

Problem: Reads are mapped **one** read at a time, this sometimes leads to single variants being split into multiple variants

Solution: Realign such a region taking **all** reads into account



HiSeq data, raw BWA alignments

HiSeq data, after MSA

Local realignment

```
module load GATK
```

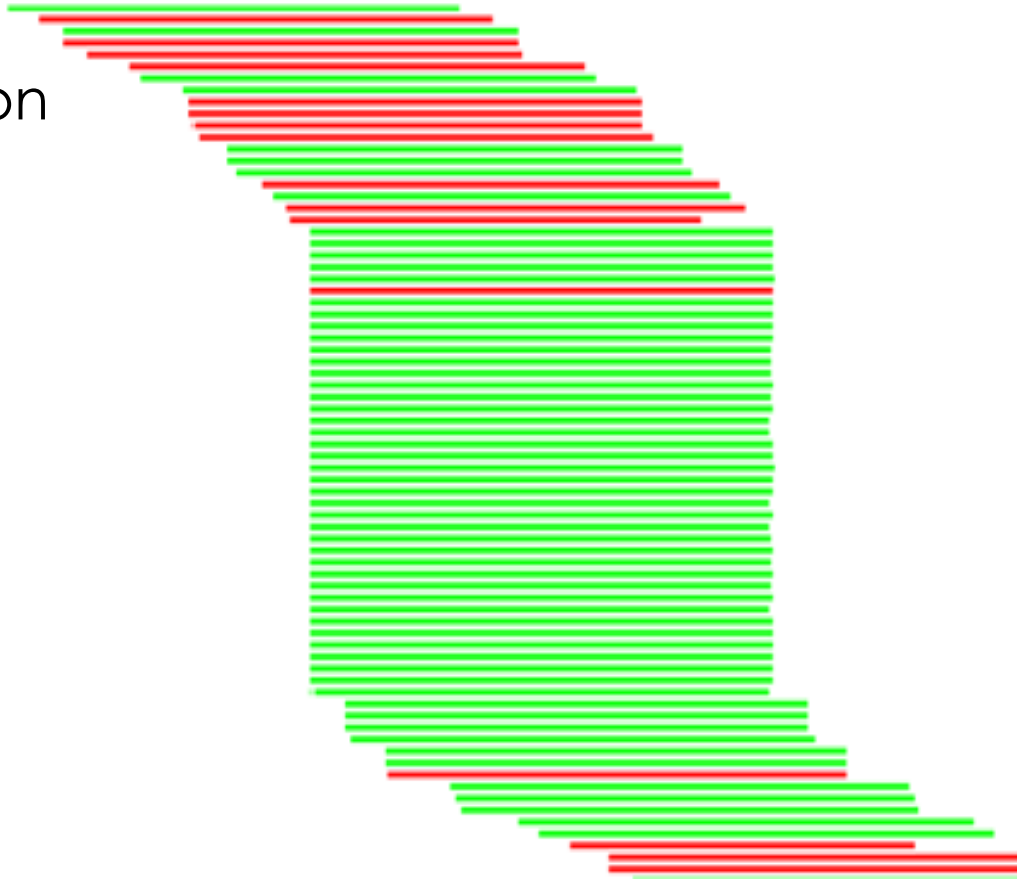
- `Genome Analysis ToolKit`
  - RealignerTargetCreator
  - IndelRealigner
- Local realignment, still needed?
  - HaplotypeCaller (HC)
  - Mutect2

NGS workflow

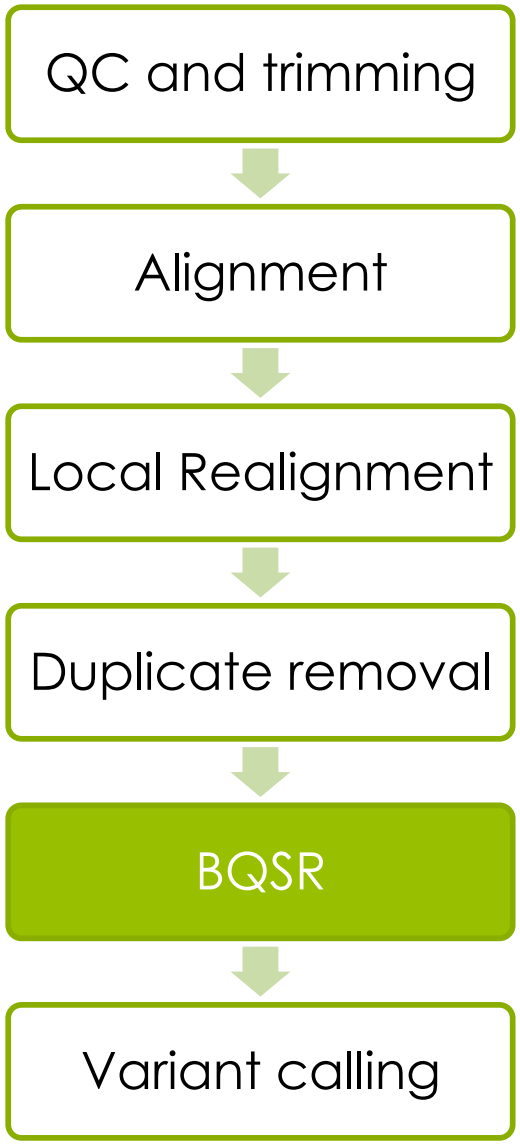
PCR duplicates & removal

```
module load picard
```

- Occur during library preparation
- Don't add unique information



NGS workflow



Base Quality Score Recalibration

NBIS SciLifeLab

```
module load GATK
```

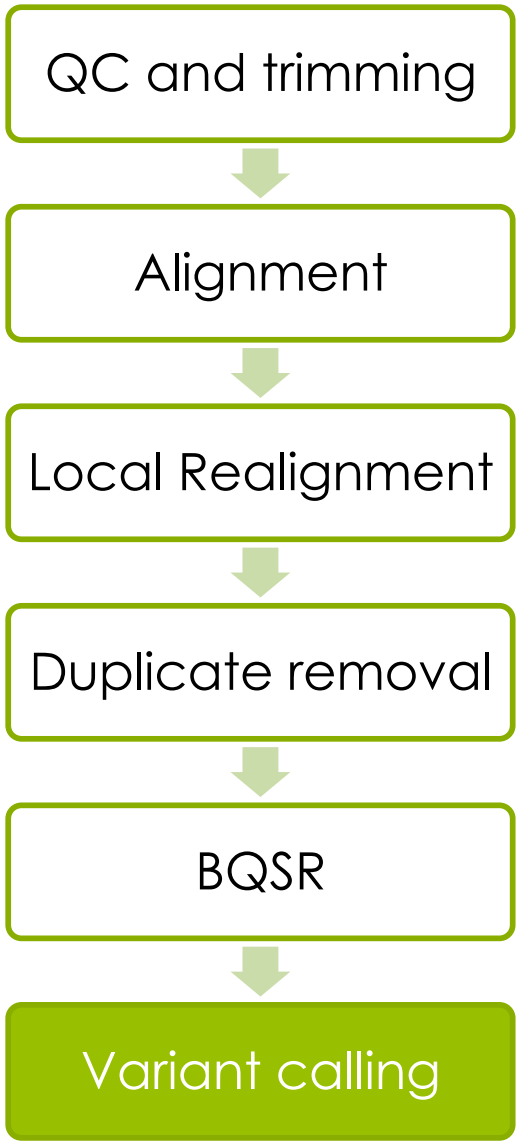
- **Identifies and corrects systematic (non-random) technical errors made by the sequencer when estimating the quality score of each base call**

- Correcting for over-/Underestimation of quality scores
  - Helps fight false positive variant calls
  - Rescues false negatives variant calls

- Some errors can be due to the physics or chemistry of the sequencing reaction, some to manufacturing flaws in the equipment

- Errors are identified over several covariates, mainly related to sequence context, position in read or machine cycle

NGS workflow



Variant calling

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...

Sample: ...GTGCGTAGACTGATAGATCGAAGA...

Variant calling

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...

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

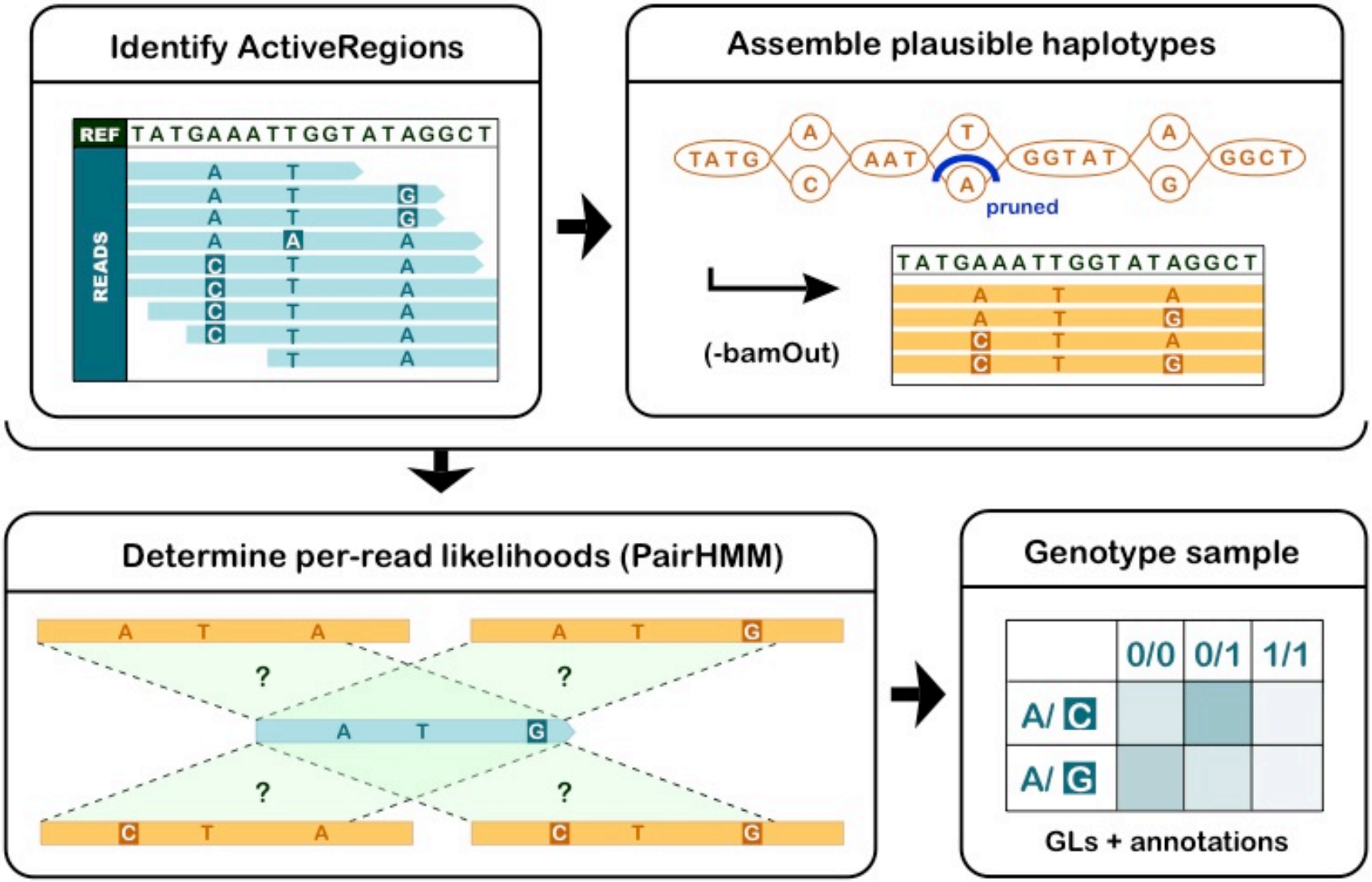
Variant calling

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...

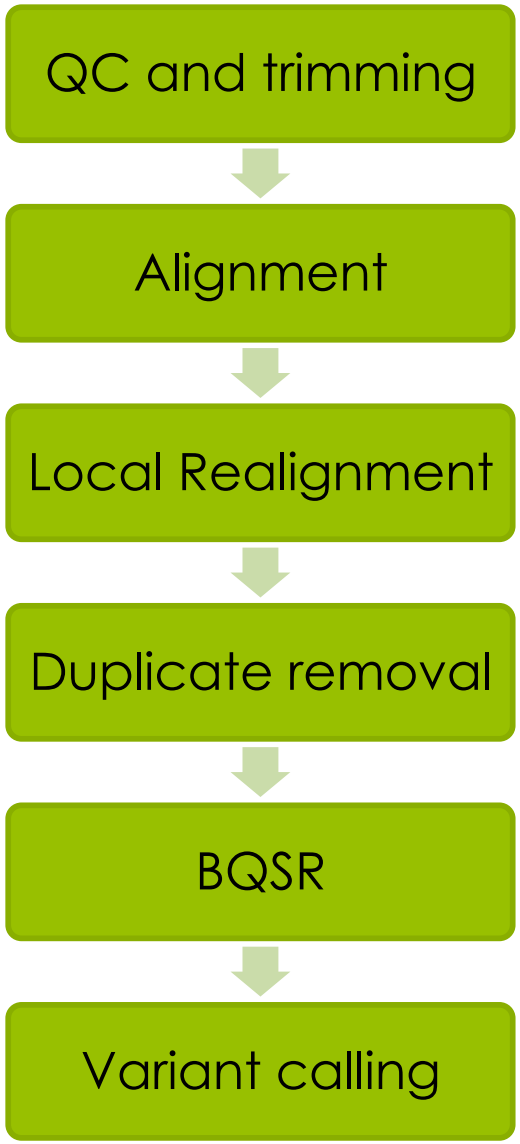
Sample:
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...

$$\frac{\#Variants\ in\ a\ position}{\#Reads\ in\ a\ position} = A\ variants\ allele\ frequency$$

Variant Calling
HaplotypeCaller



NGS workflow



VCF Files

```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##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=.,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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#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:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
```

VCF Files

```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##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=.,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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#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:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
```

VCF Files

```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##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=.,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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#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:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
```

VCF Files

```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##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=.,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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#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:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
```

VCF Files

```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##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=.,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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#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:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
```

VCF Files

```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##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=.,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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#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:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
```

VCF Files

```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##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=.,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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM  POS      ID          REF ALT   QUAL FILTER INFO
20      14370    rs6054257   G   A     29   PASS   NS=3;DP=14;AF=0.5;DB;H2
20      17330    .           T   A     3    q10    NS=3;DP=11;AF=0.017
20      1110696  rs6040355   A   G,T   67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB
```

VCF Files

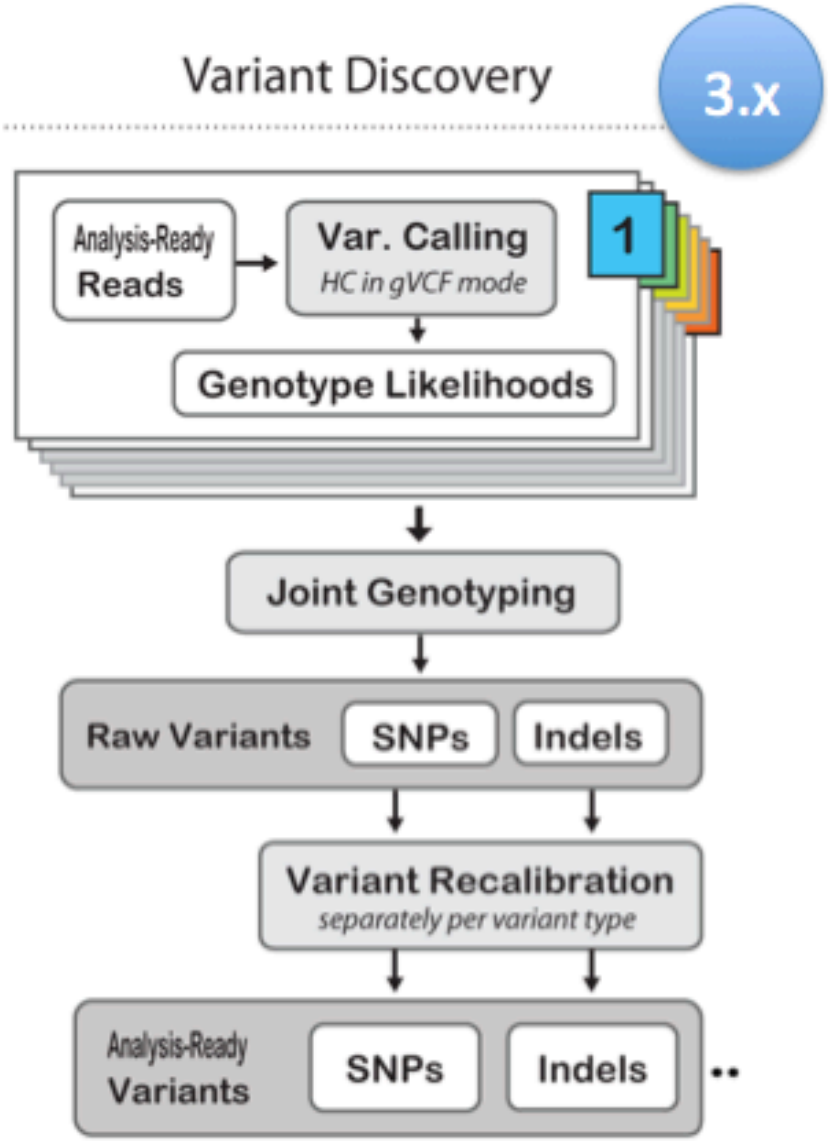
```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele 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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#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:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
```

VCF Files

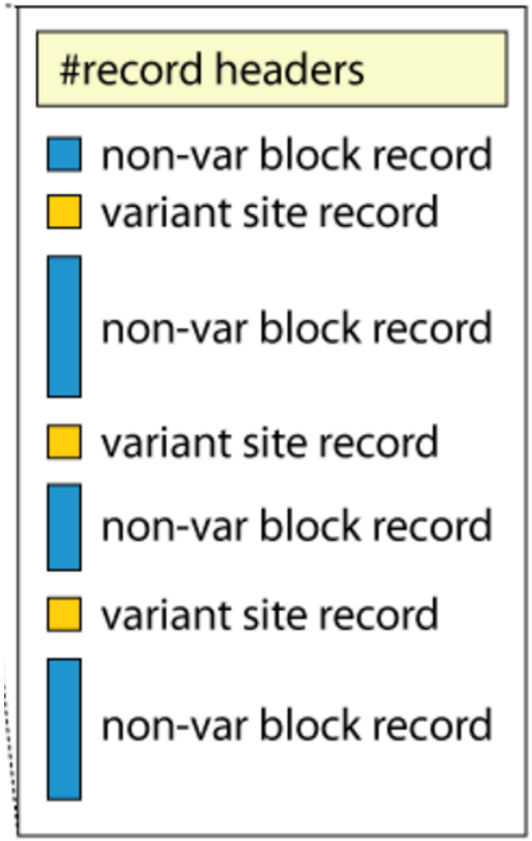
```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele 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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

#FORMAT	NA00001	NA00002	NA00003
GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4

Joint genotyping



gVCF Files




```
##GVCFBlock=minGQ=0(inclusive),maxGQ=5(exclusive)
##GVCFBlock=minGQ=20(inclusive),maxGQ=60(exclusive)
##GVCFBlock=minGQ=5(inclusive),maxGQ=20(exclusive)
```

Filtering

```
module load GATK
```

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ
```

VariantFiltration

--filterExpression "QUAL > 30"

--filterName QUAL_filter

--filterExpression "QUAL / DP < 10.0"

--filterName QUALDP_filter

Annotation

```
module load annovar /snpEff / vep
```

```
#CHROM POS ID REF ALT QUAL
20 14370 rs6054257 G A 29
```

- Gene-based
  - Non-synonymous/synonymous
- Region-based
  - CpG-islands
  - Conserved regions
  - Predicted transcription factor binding sites
- Filter-based
  - dbSNP
  - 1000G
  - COSMIC

Annotation

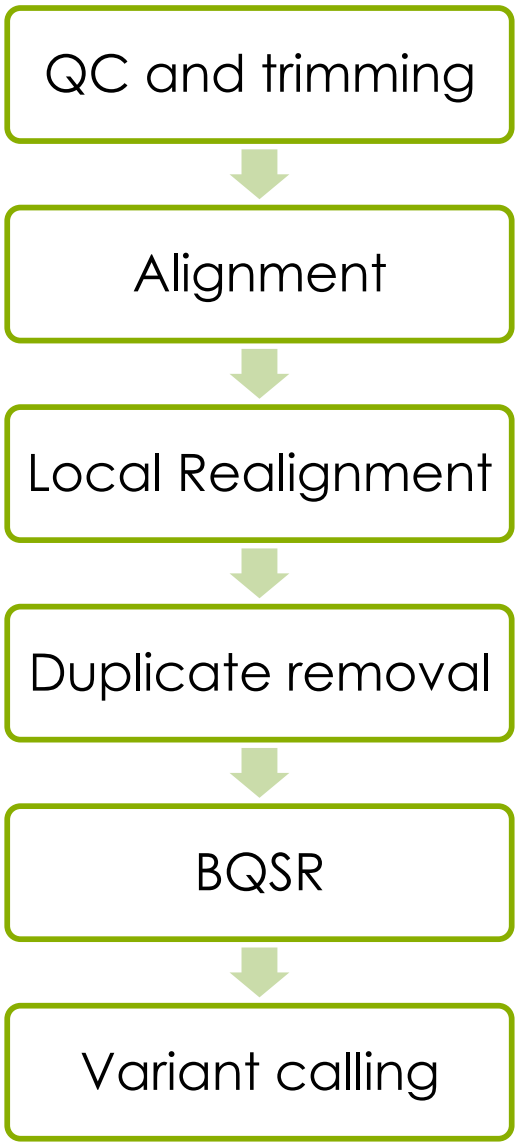
```
module load annovar /snpEff / vep
```

```
#CHROM POS ID REF ALT QUAL
20 14370 rs6054257 G A 29
```

- Gene-based
  - Non-synonymous/synonymous
- Region-based
  - CpG-islands
  - Conserved regions
  - Predicted transcription factor binding sites
- Filter-based
  - dbSNP
  - 1000G
  - COSMIC

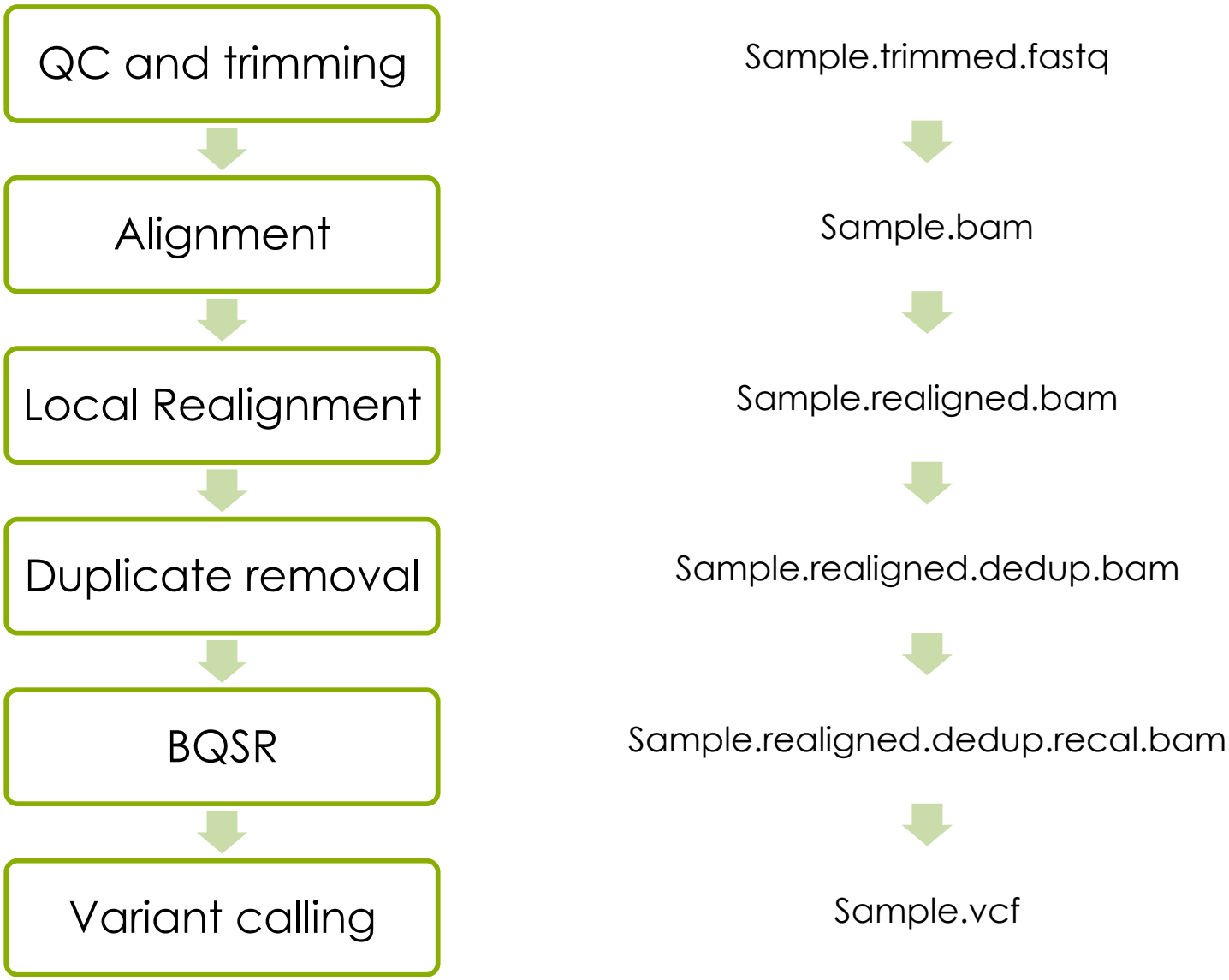
USE THE SAME REFERENCE!

File naming conventions



- Use informative file names
- create a new output file in each process
- Include description of process in output file name

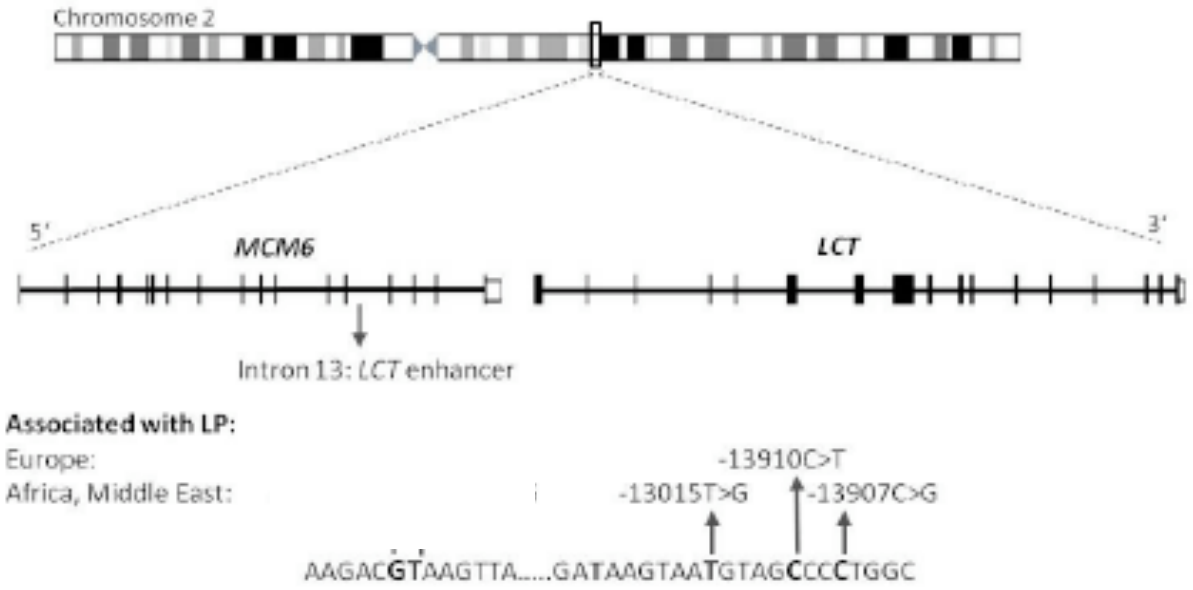
File naming conventions



Variant relating lactase persistence

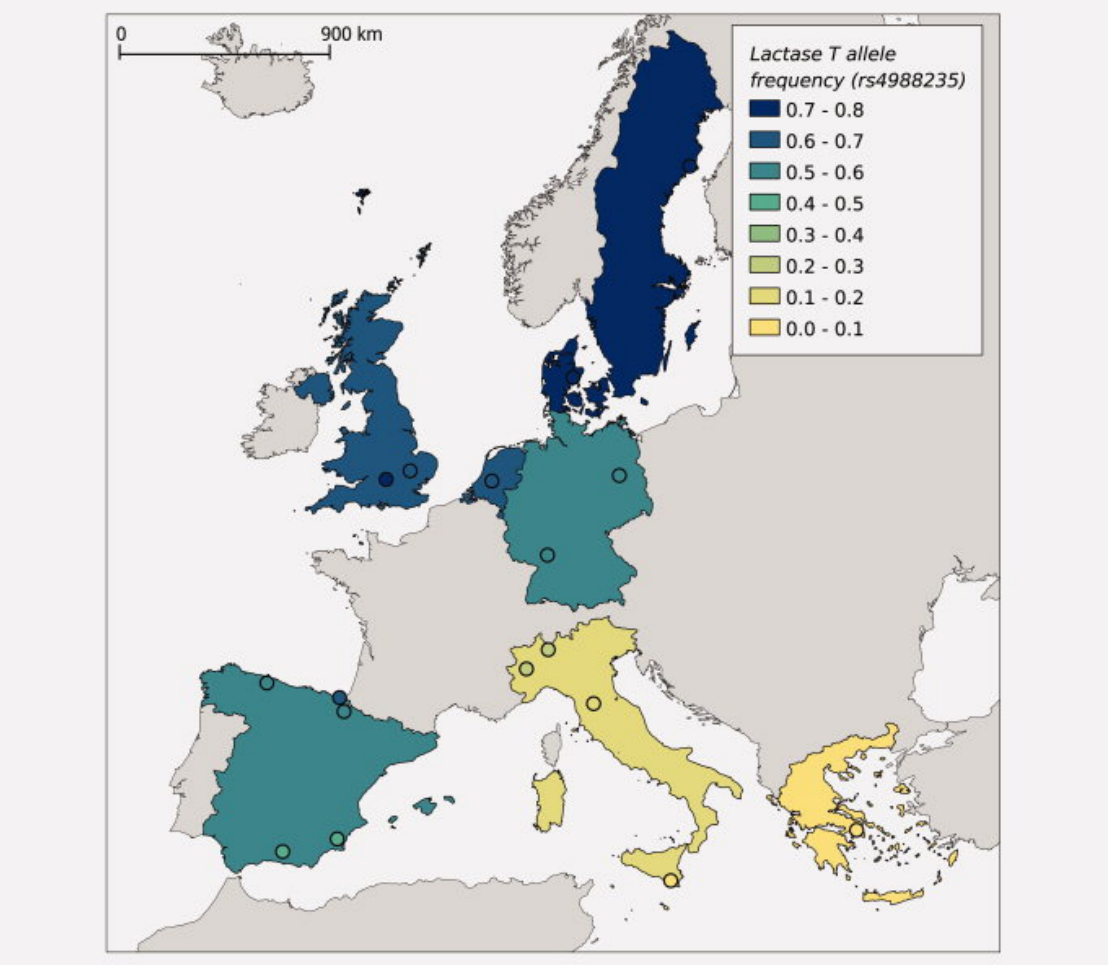
One single variant in the enhancer to the LCT gene
is associated with the ability to digest lactase as adults, e.g.
lactase Persistence

The variant location is LCT-13910C>T and it has dbSNP id is rs4988235

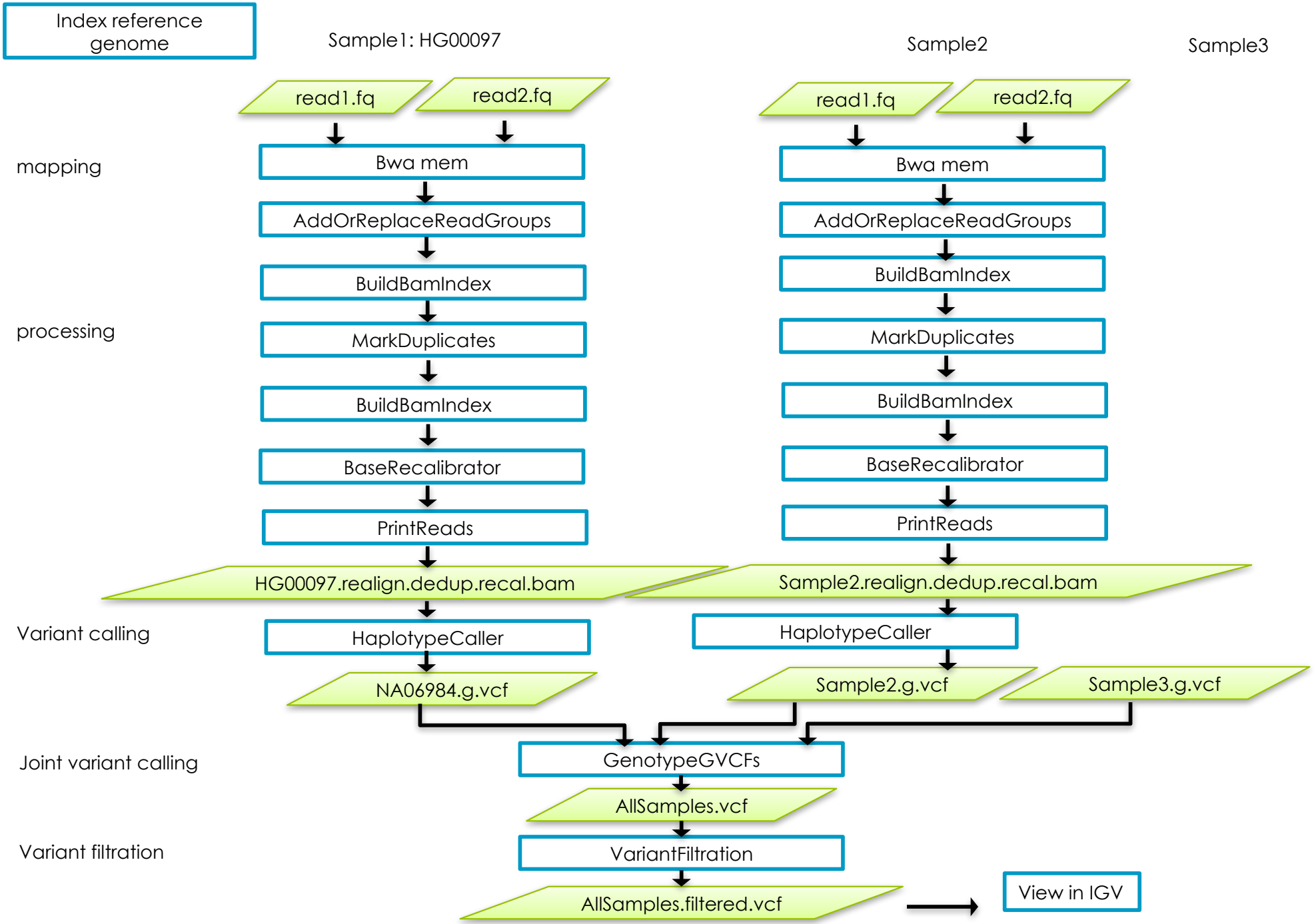


Varying allele frequencies

- The allele frequencies varies between different contries, for example 74% of the alternative allele in Sweden to 9% in Greece



Flowchart of lab



NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Questions?

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Questions?

Work like a professional bioinformatician – Google errors!