From raw reads to variants

Nina Norgren, NBIS

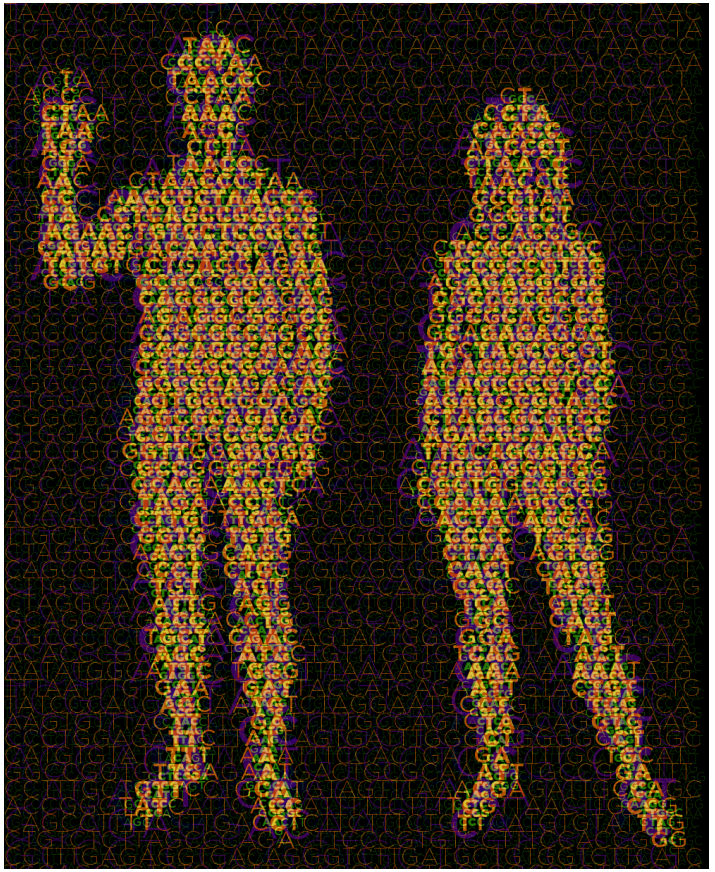
Göteborg, May 2019

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

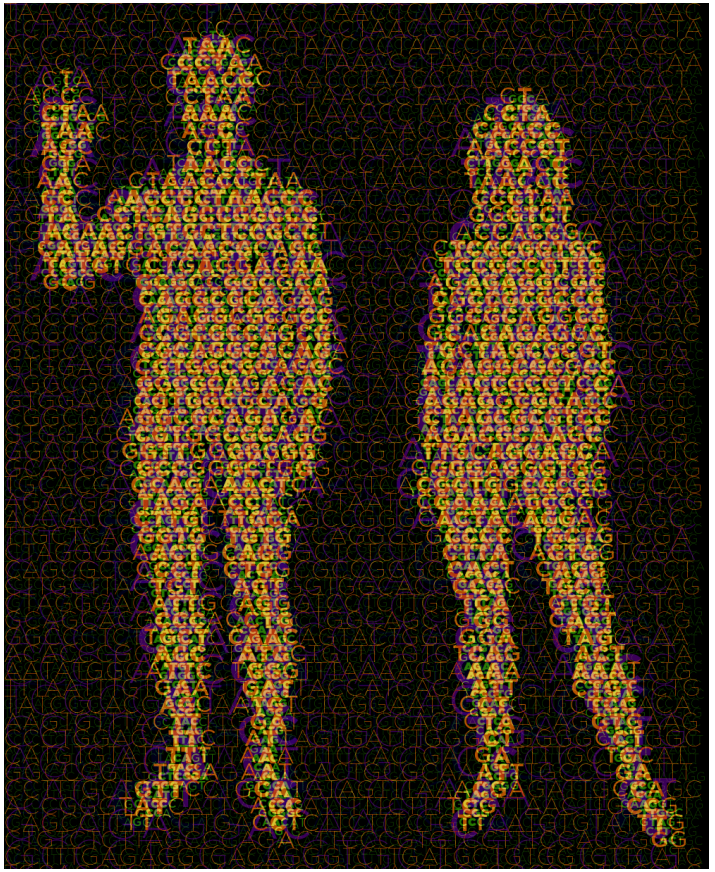
Reference genome

- 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

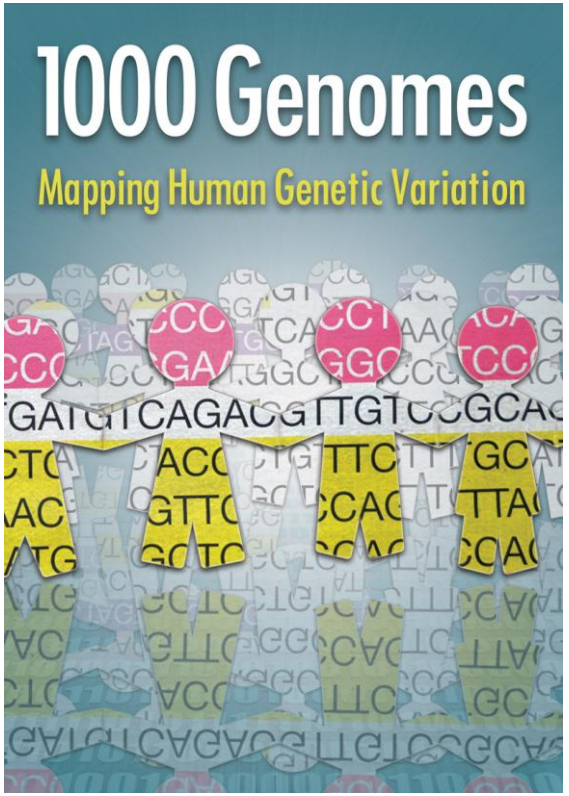
Reference: ...GTGCGTAGACTGCTAGATCGAAGA...
Sample: ...GTGCGTAGACTGATAGATCGAAGA...

Variants

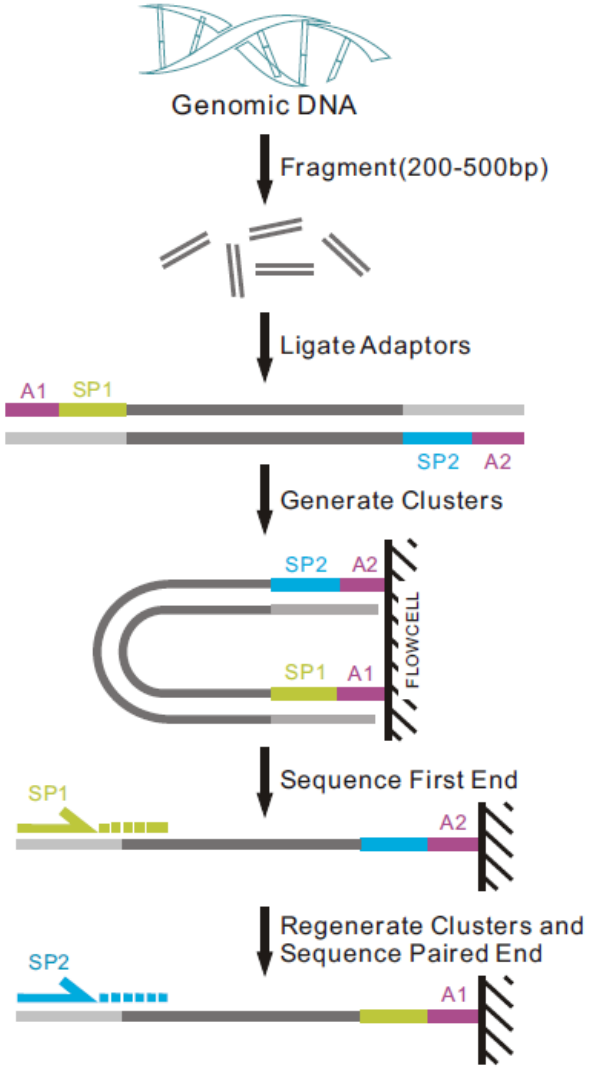
Population based variant projects

UK 10K

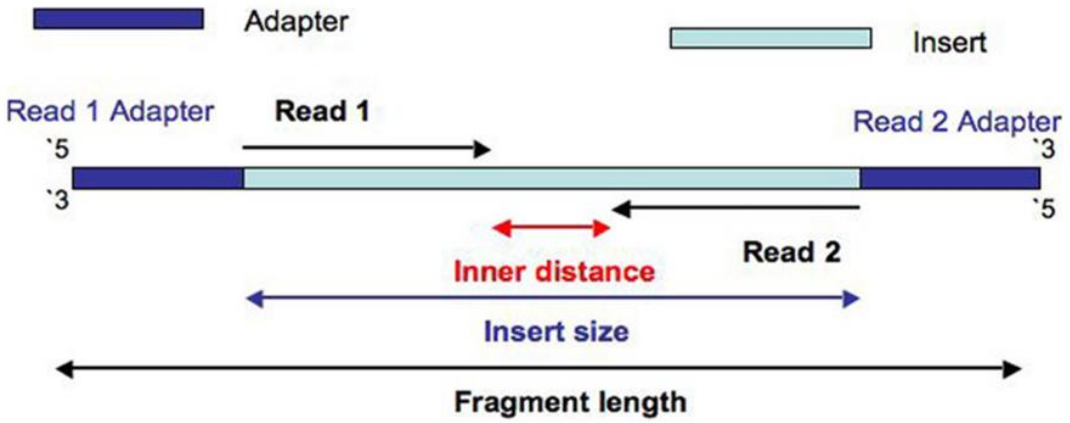
RARE GENETIC VARIANTS IN HEALTH AND DISEASE



Paired-end sequencing



Paired-end data



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@CFFFFFHHHHHGJJJJJJJJJJFHHIIIIJJ
JIHGIIJJJJIJIJIJJJJIIJJJJJIIEIHHIJ
HGHHHHHDFFFEDDDDDCDDDCDDDDDDDCDC
```

ID_R2_001.fastq

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

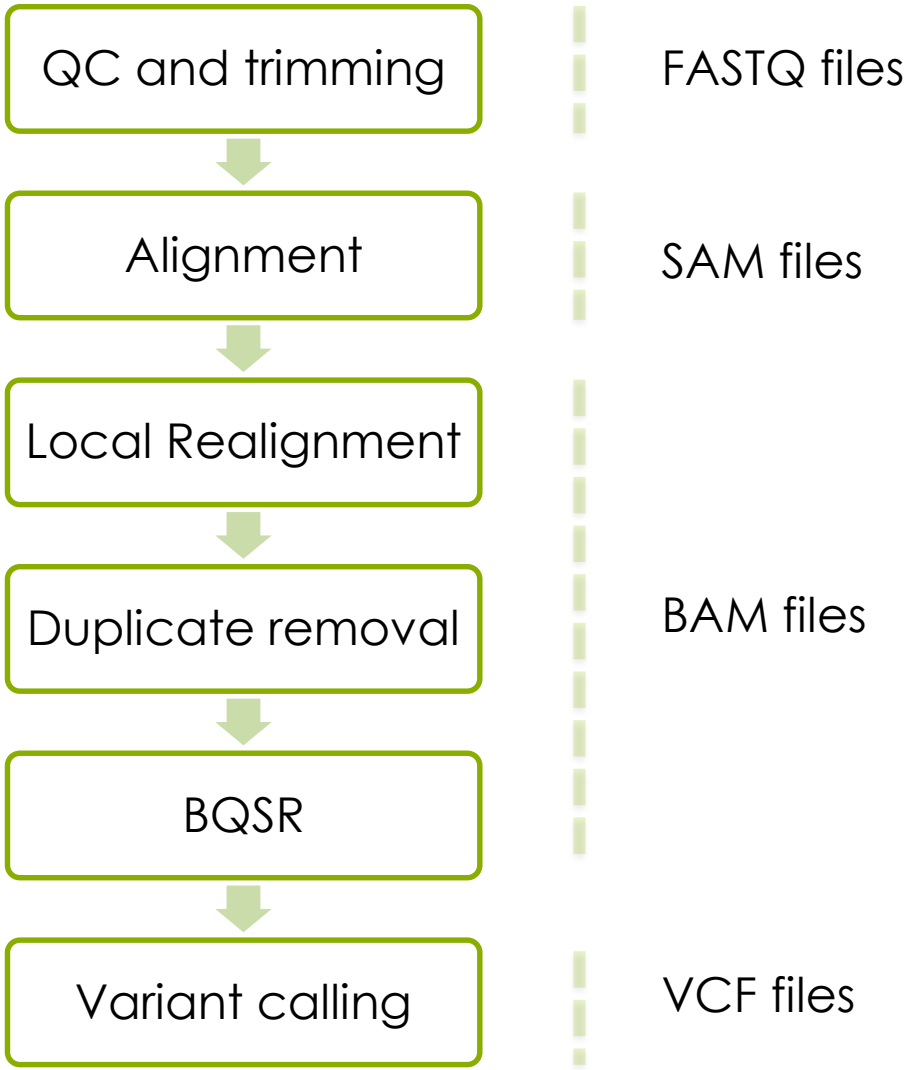
ID_R1_001.fastq

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

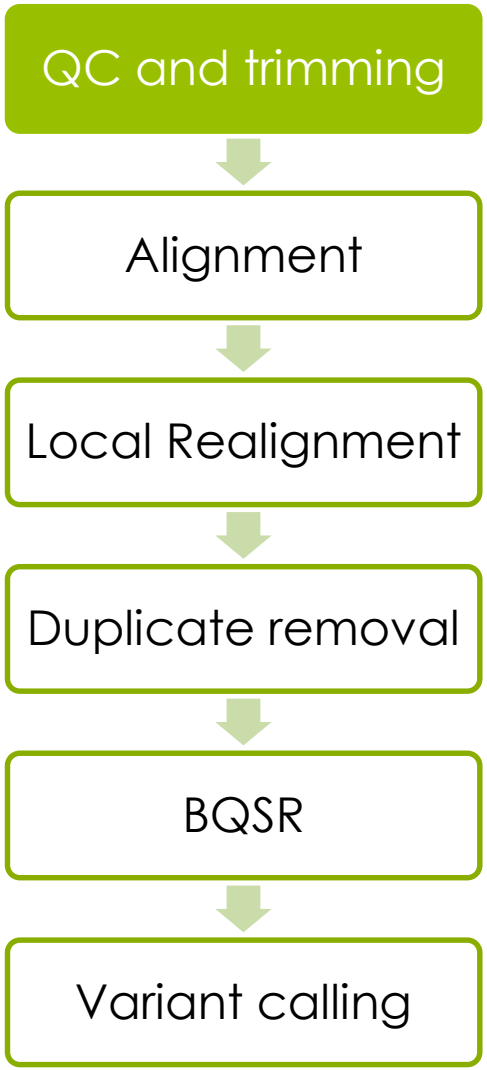
ID_R2_001.fastq

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

NGS workflow



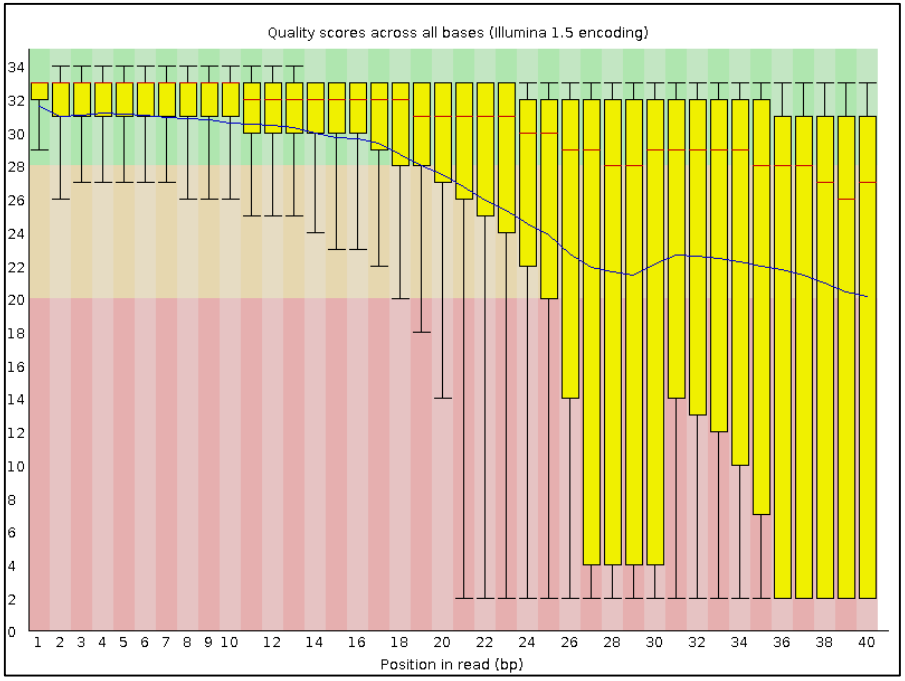
NGS workflow



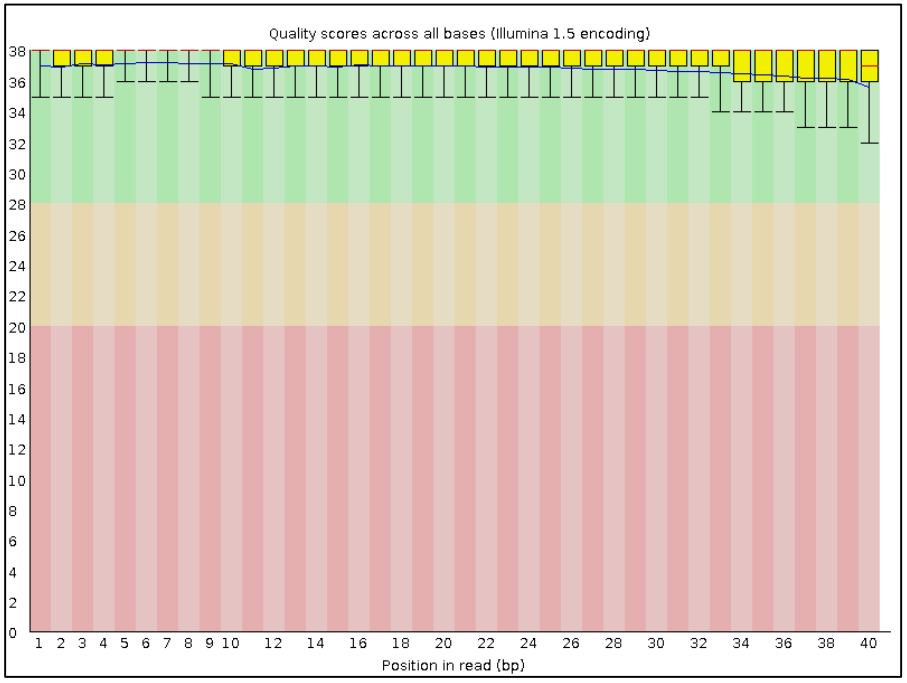
Quality control

module load FastQC

Bad qualities:



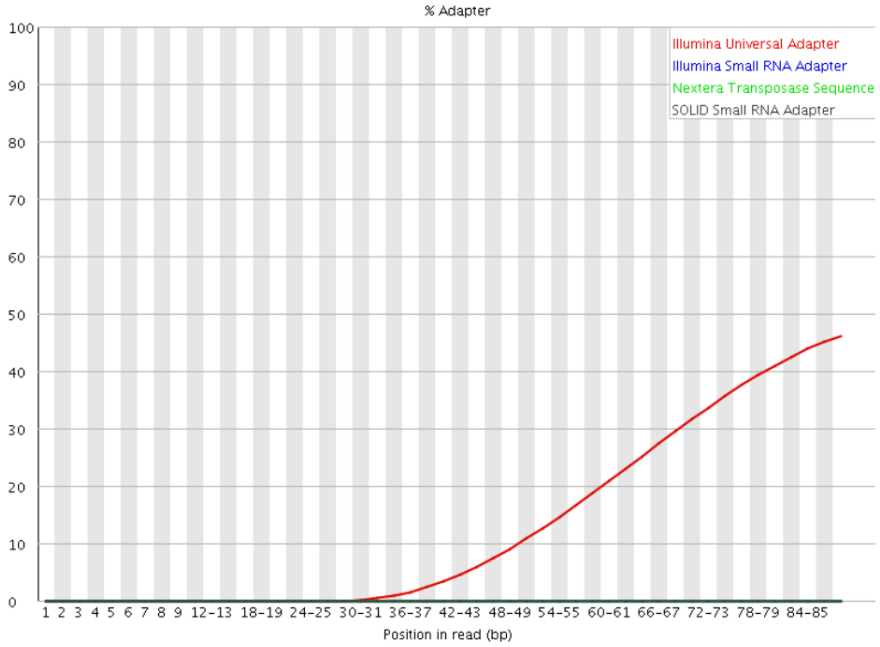
Good qualities:



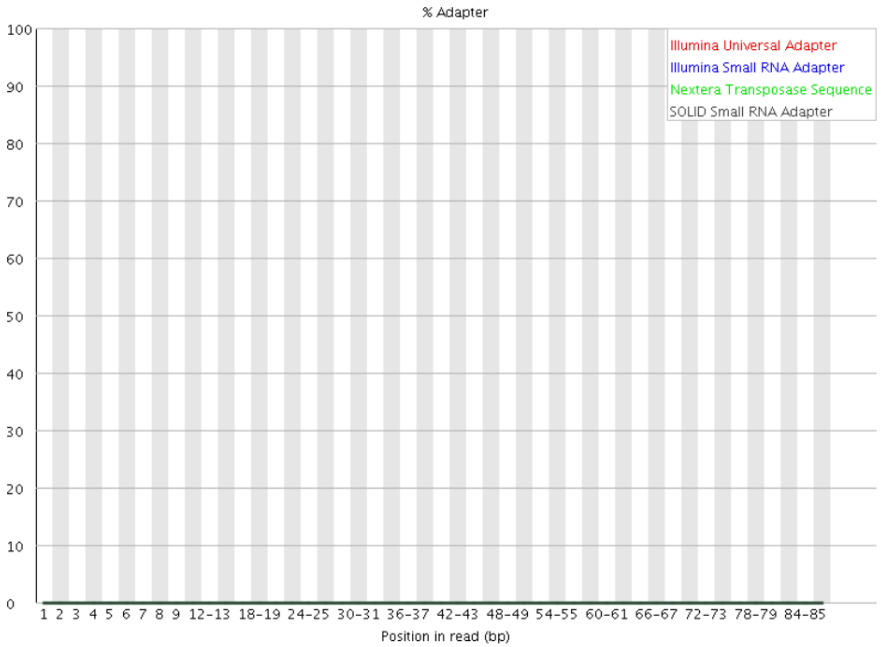
Quality control

module load FastQC

Adapters present:

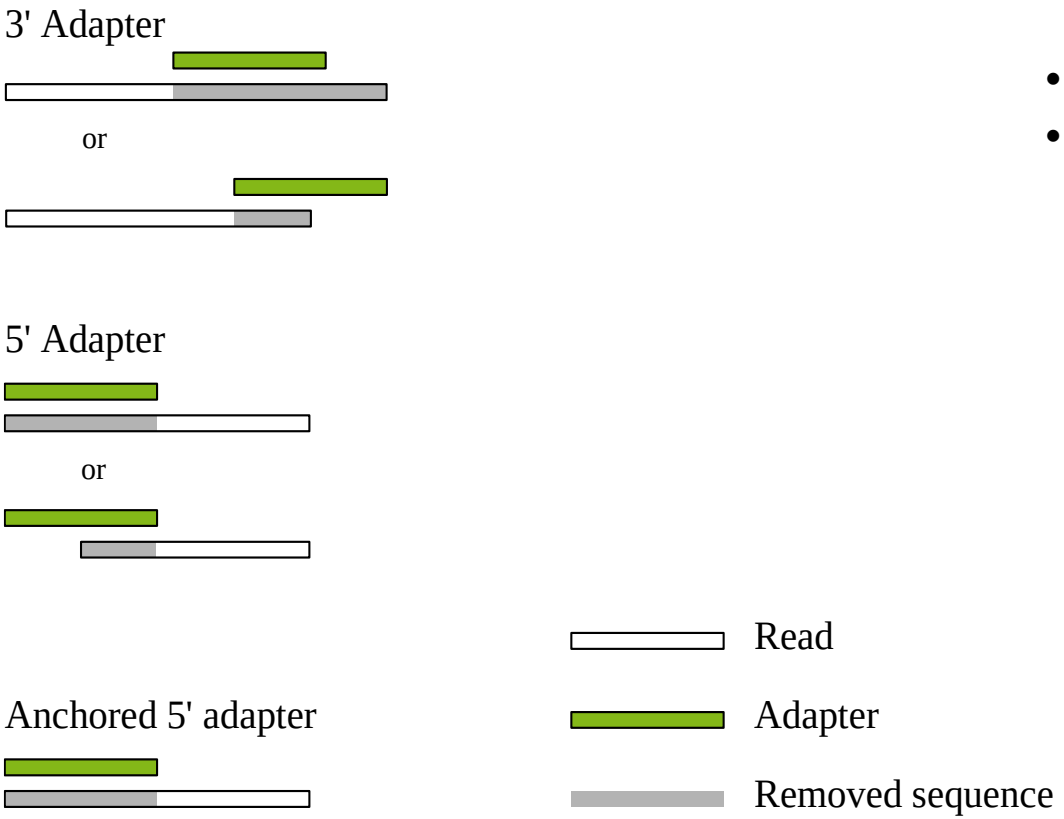


Adapters Absent:



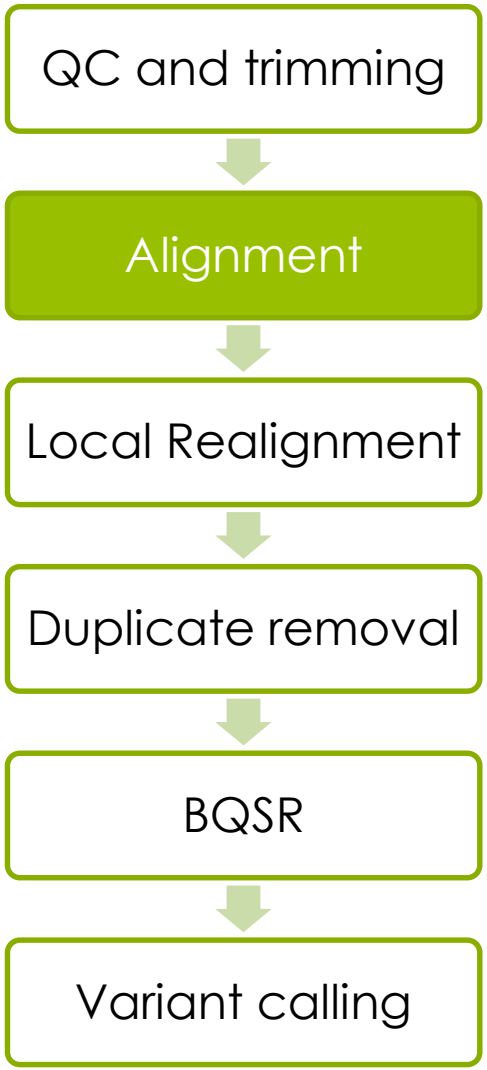
Trimming

module load cutadapt / TrimGalore / trimmomatic

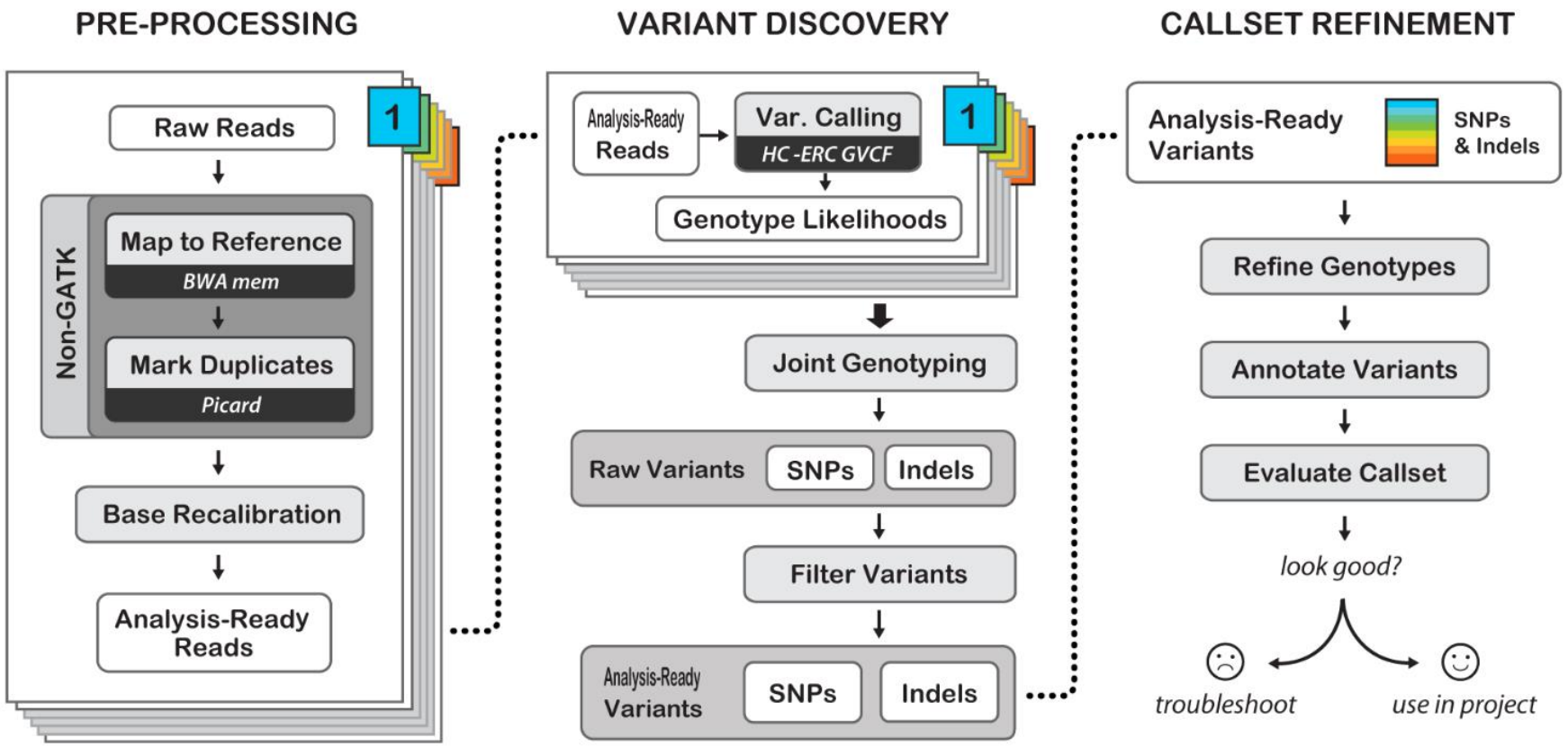


- Remove bad quality reads
- Remove adapters

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

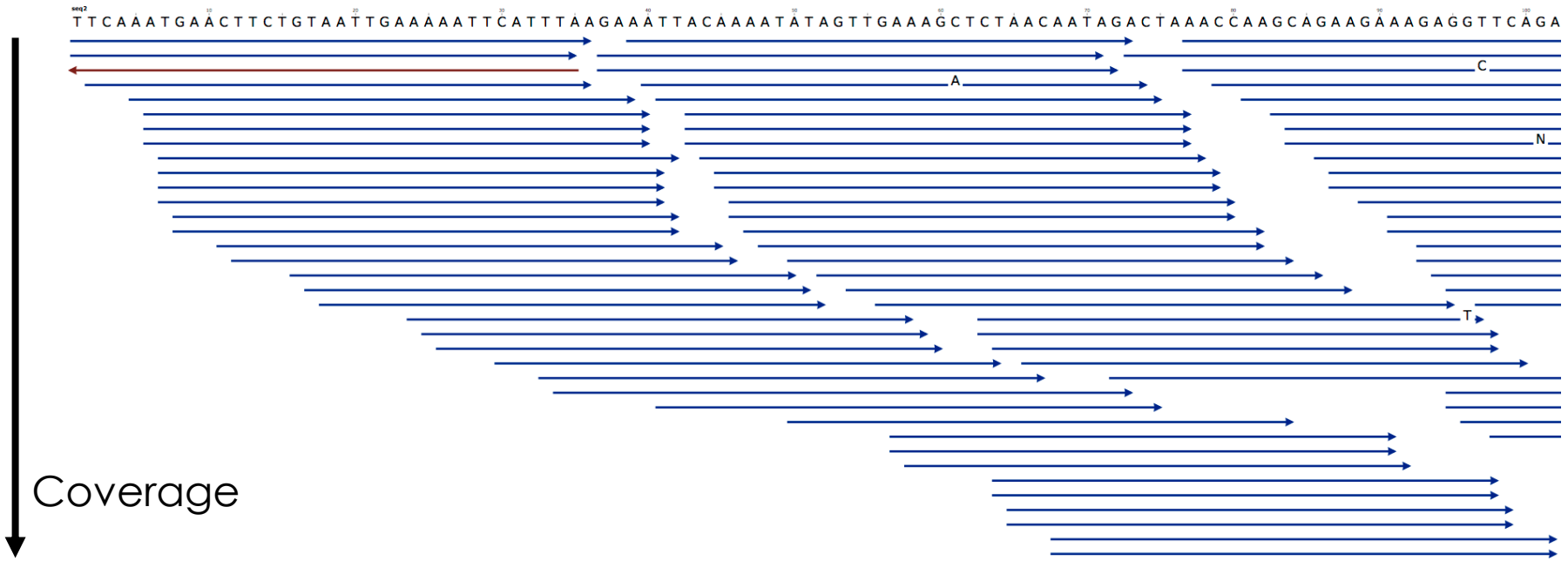
Reference GACCTCATCGATCCCACTG

Alignment

module load bwa

Alignment

module load bwa



Paired-end data & Alignment

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

Sam format

```
Coor      12345678901234   5678901234567890123456789012345
Ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001/1          TTAGATAAAGGATA*CTG
r002          aaaAGATAA*GGATA
r003        gcctaAGCTAA
r004                          ATAGCT..............TCAGC
r003                                ttagctTAGGC
r001/2                                           CAGCGGCAT
```

```
@SQ SN:ref LN:45
r001  99    ref  7   30  8M2I4M1D3M  =  37  39   TTAGATAAAGGATACTG  *
r002  0     ref  9   30  3S6M1P1I4M  *  0   0    AAAAGATAAGGATA    *
r003  0     ref  9   30  5S6M        *  0   0    GCCTAAGCTAA       *
r004  0     ref  16  30  6M14N5M     *  0   0    ATAGCTTCAGC       *
r003  2064  ref  29  17  6H5M        *  0   0    TAGGC             *
r001  147   ref  37  30  9M          =  7   -39  CAGCGGCAT         *  NM:i:1;
```

Read groups

- Link information of *sample id*, *library prep*, *flowcell* and *sequencing runs* to fastq file.
- 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

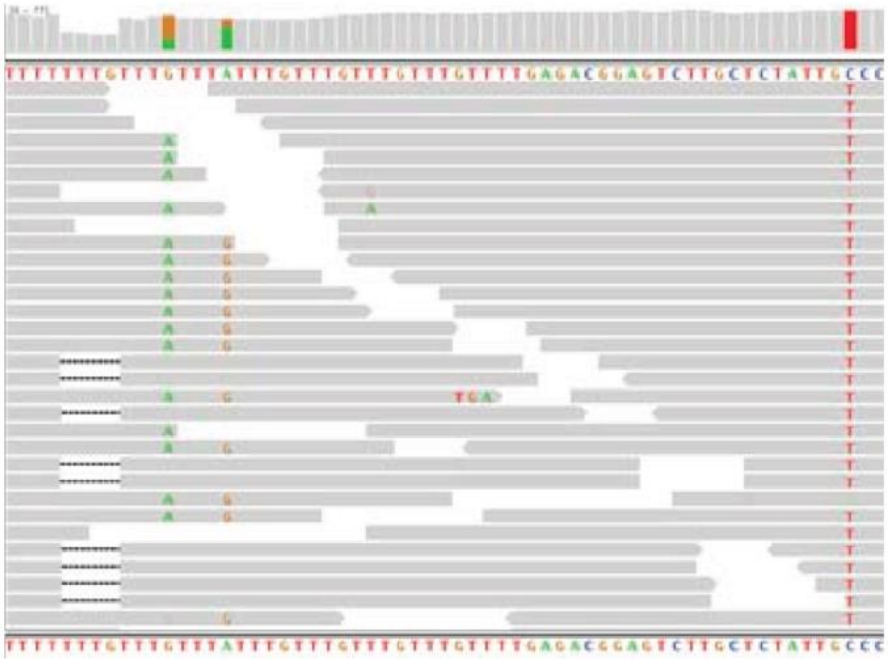
Bam file is a binary representation of the Sam file

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

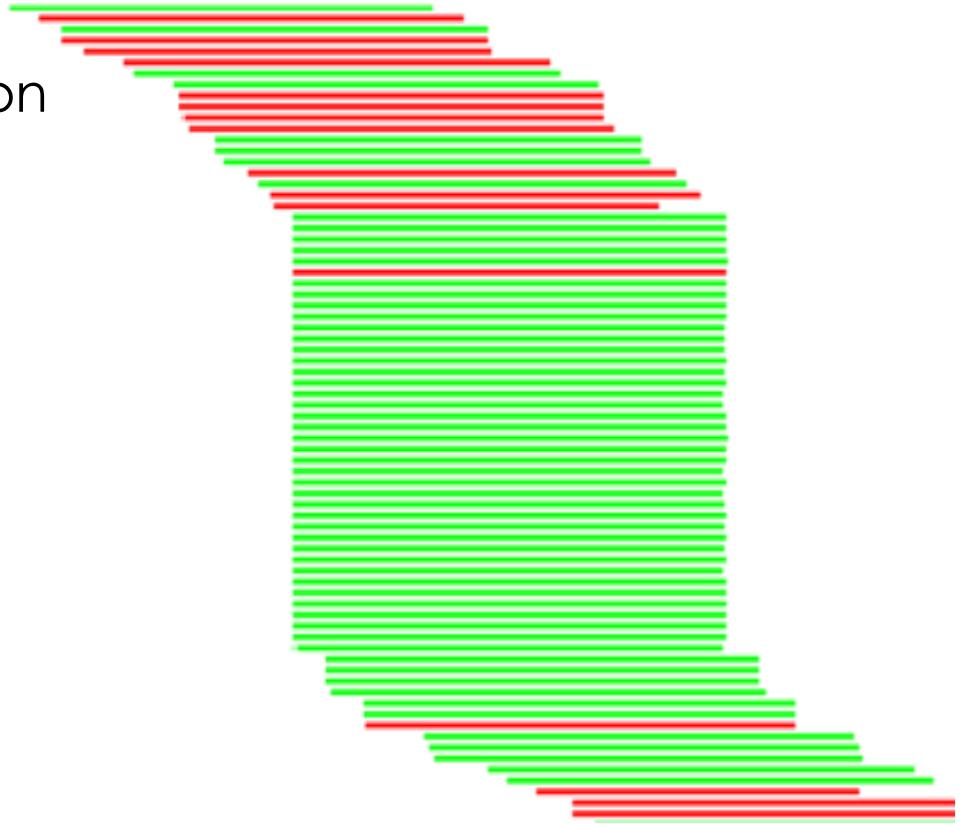
- Genome Analysis ToolKit
  - RealignerTargetCreator
  - IndelRealigner
- Local realignment, still needed?
  - HaplotypeCaller (HC)
  - Mutect2

NGS workflow

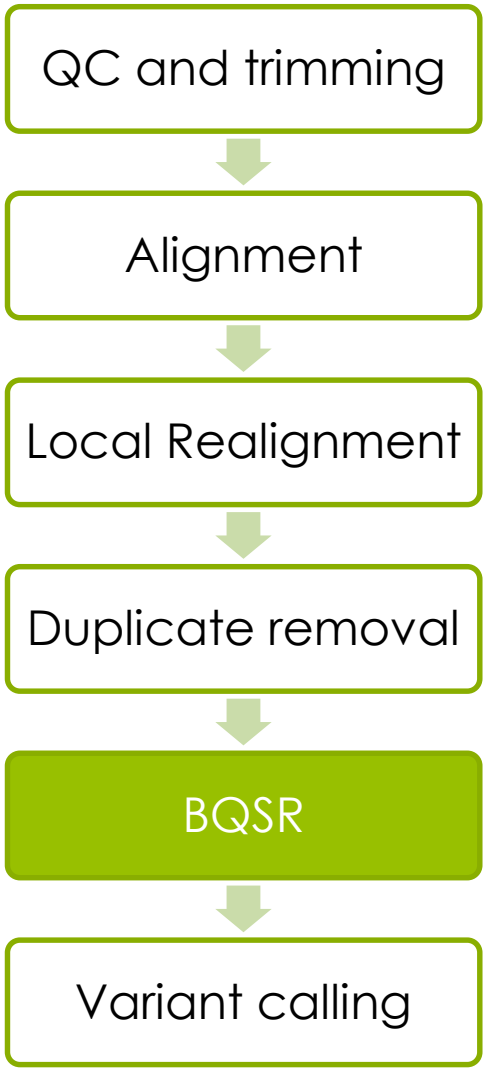
PCR duplicates & removal

module load picard

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



NGS workflow



Base Quality Score Recalibration

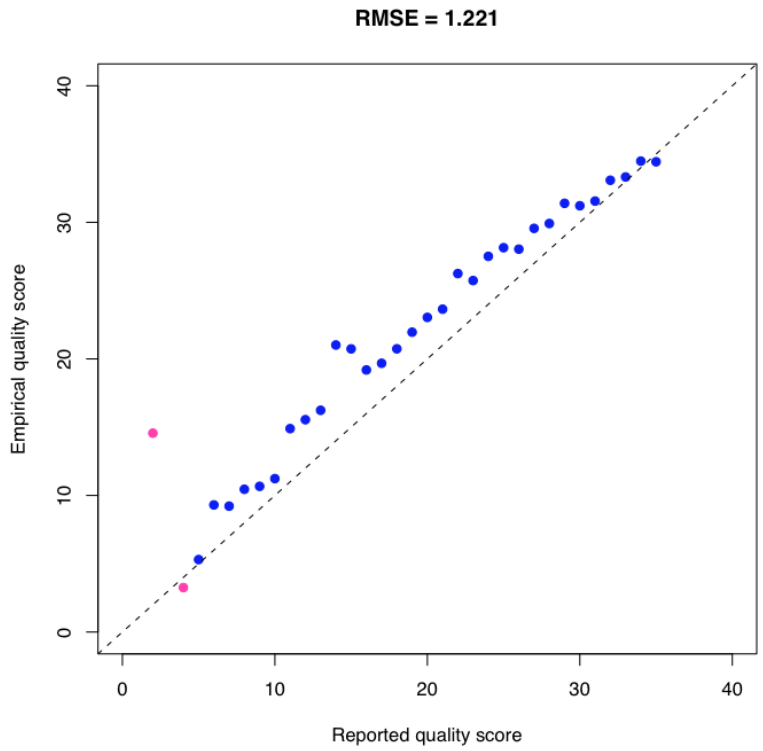
module load GATK

- **Identifies and corrects systematic (non-random) technical errors during sequencing**
- Compares covariation between
  - Reported quality score
  - The position within the read (Machine cycle)
  - The two preceding and current nucleotide (sequencing chemistry effect) observed by the sequencing machine

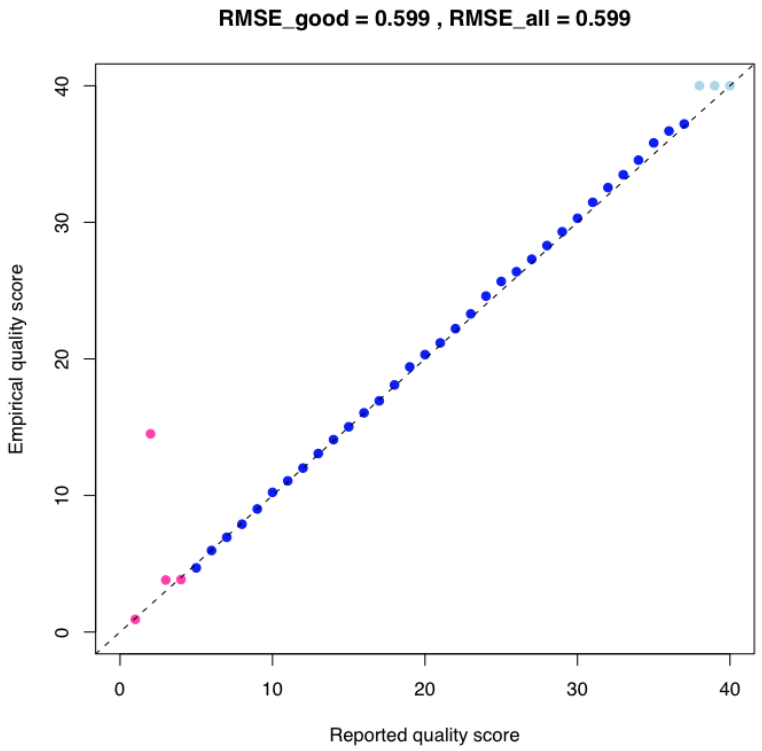
- Over-/Underestimation of quality scores
  - Helps fight False positives
  - Rescues False negatives

Base Quality Score Recalibration

Reported Quality vs. Empirical Quality



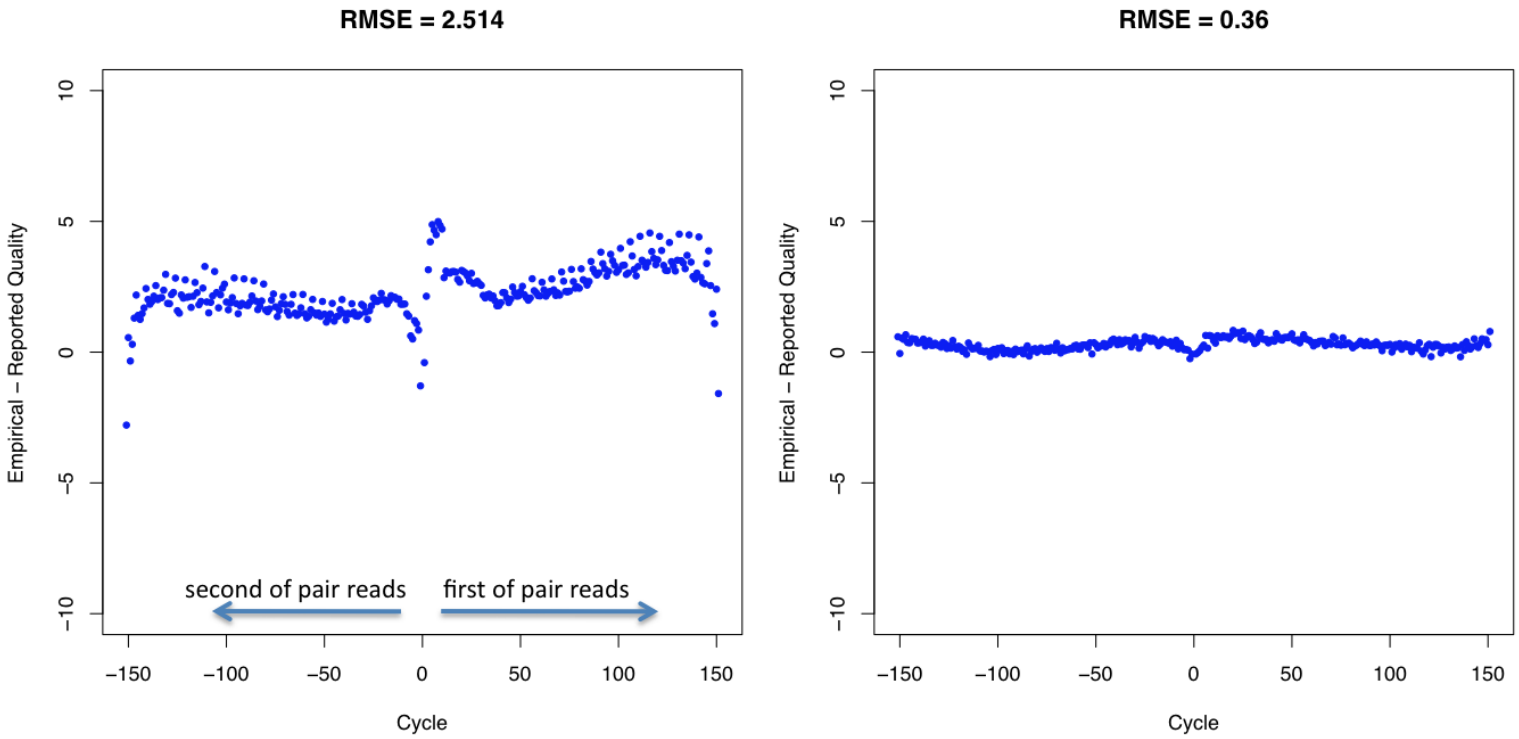
Original Data



After GATK Recalibration

Base Quality Score Recalibration

Residual Error by Machine Cycle

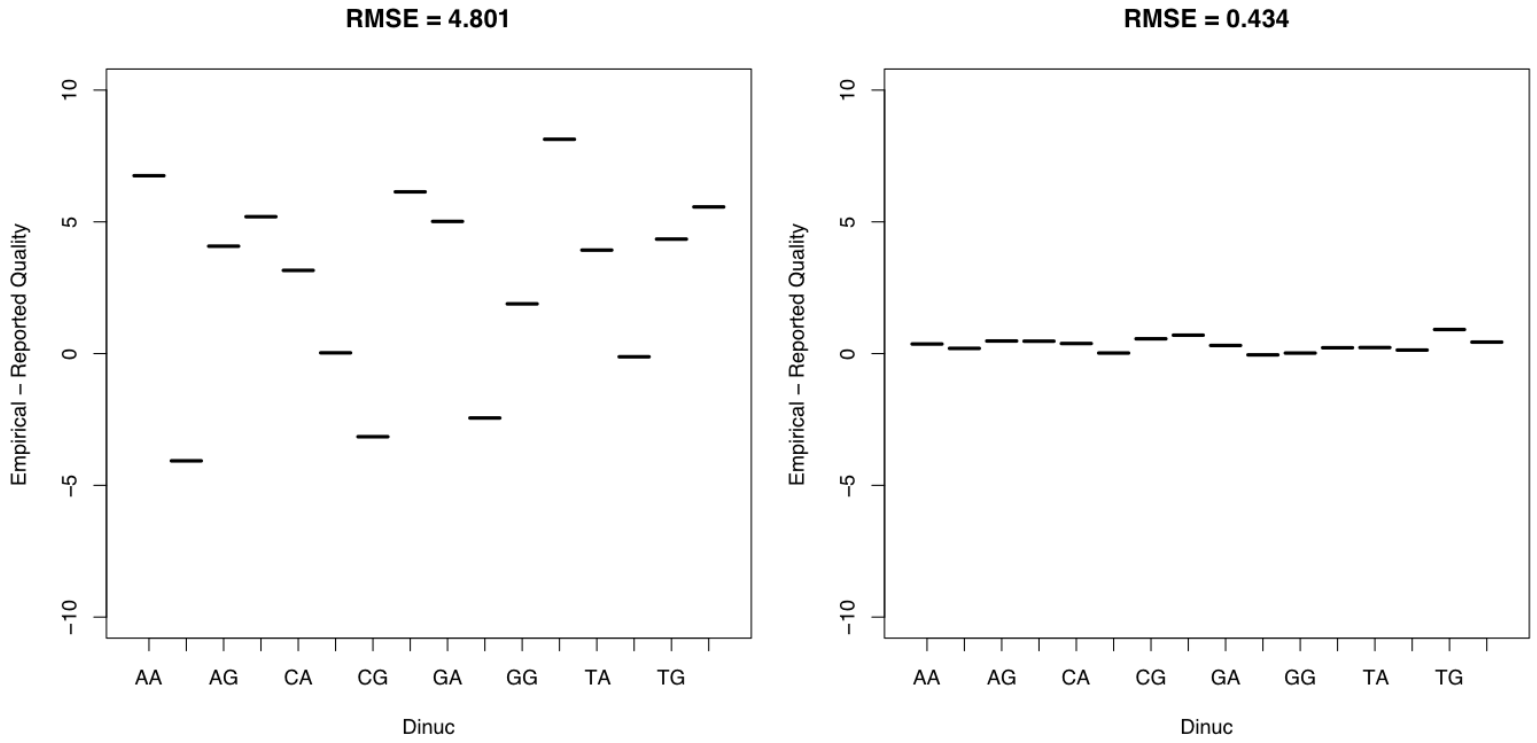


Original Data

After GATK Recalibration

Base Quality Score Recalibration

Residual Error by Dinucleotide

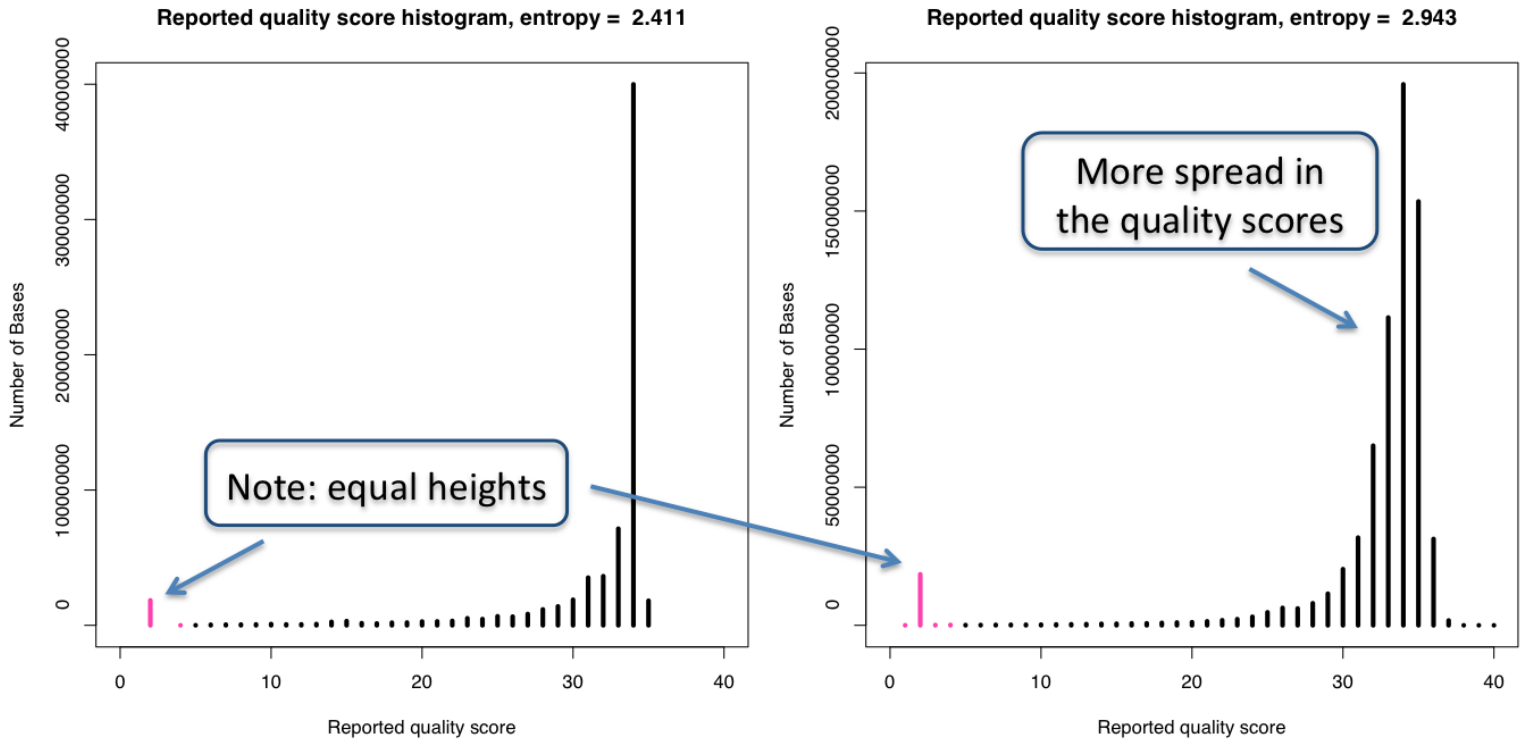


Original Data

After GATK Recalibration

Base Quality Score Recalibration

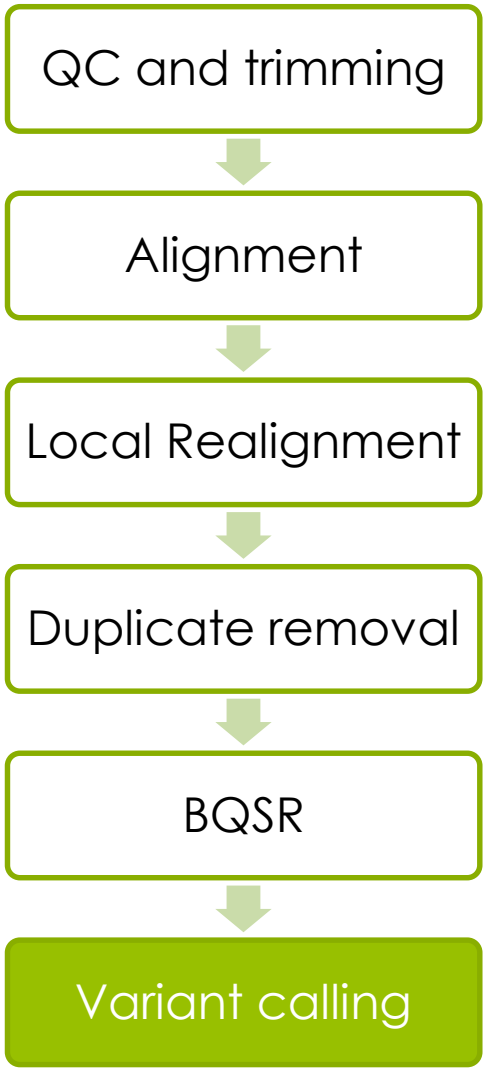
Distribution of Quality Scores



Original Data

After GATK Recalibration

NGS workflow



Variant calling

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...

Sample: ...GTGCGTAGACTGATAGATCGAAGA...

Variant calling

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...

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

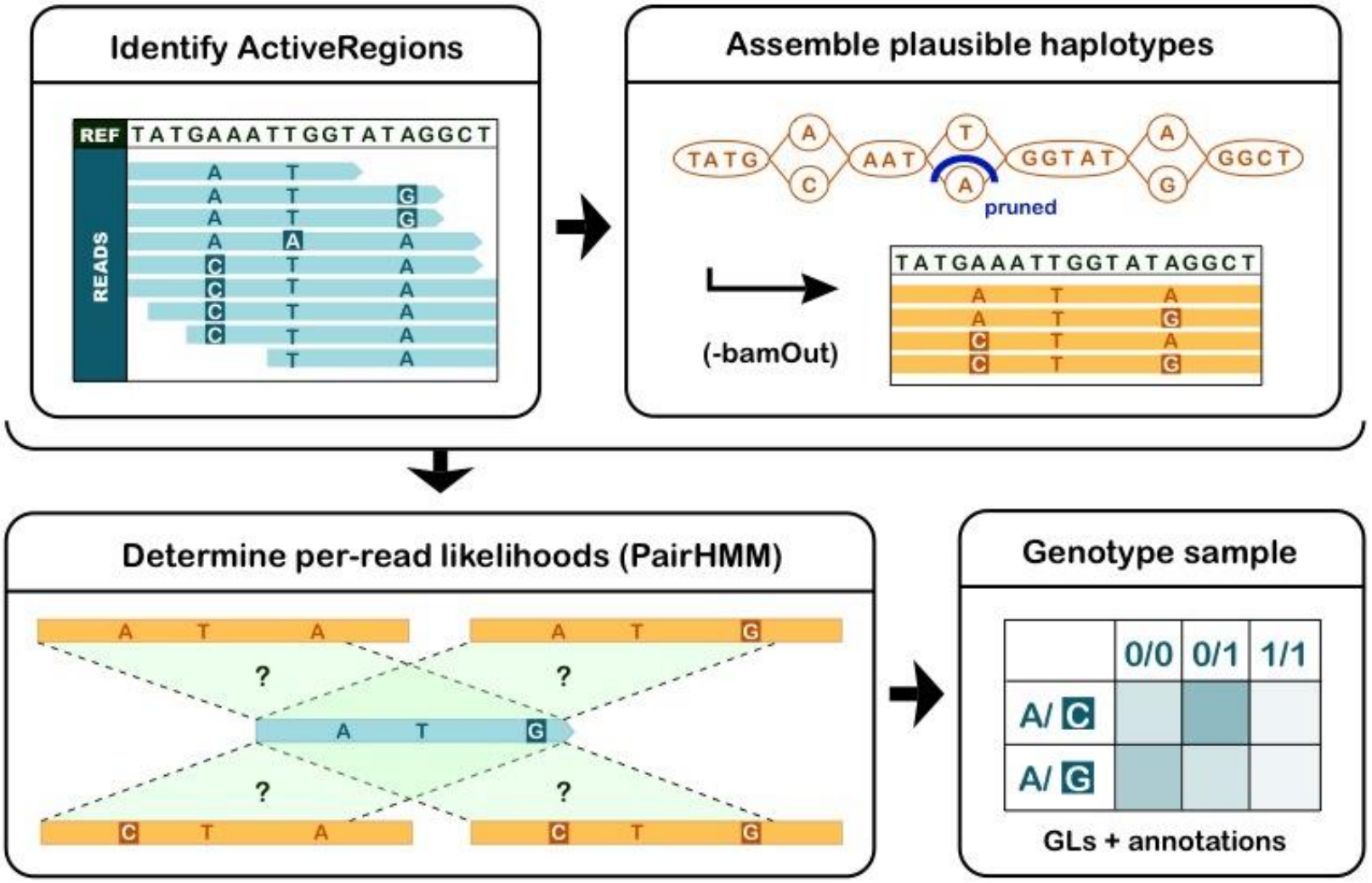
Variant calling

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...

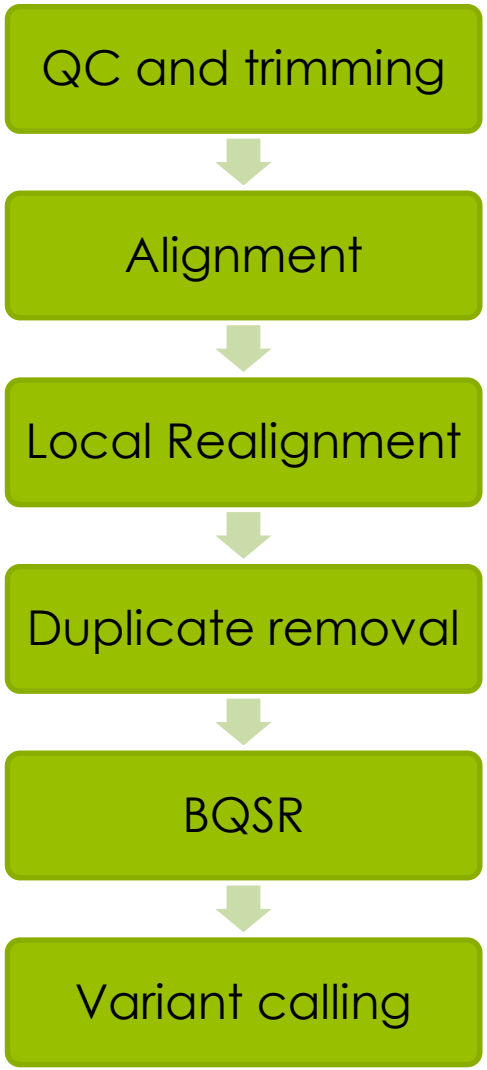
Sample:
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...
...GTGCGTAGACTGCTAGATCGAAGA...
...GTGCGTAGACTGATAGATCGAAGA...

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

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

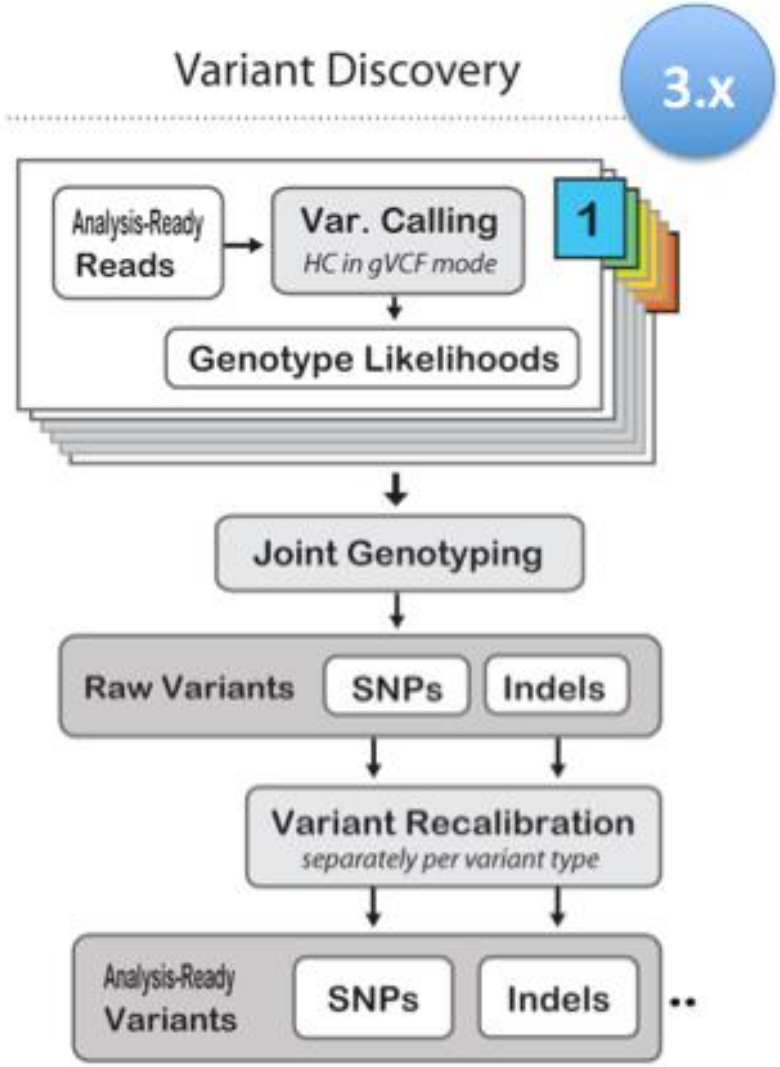
```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele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

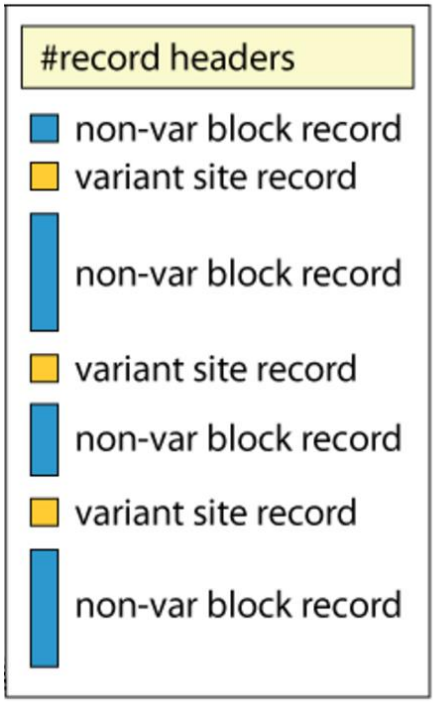
#FORMAT	NA00001	NA00002	NA00003
GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4

Joint genotyping

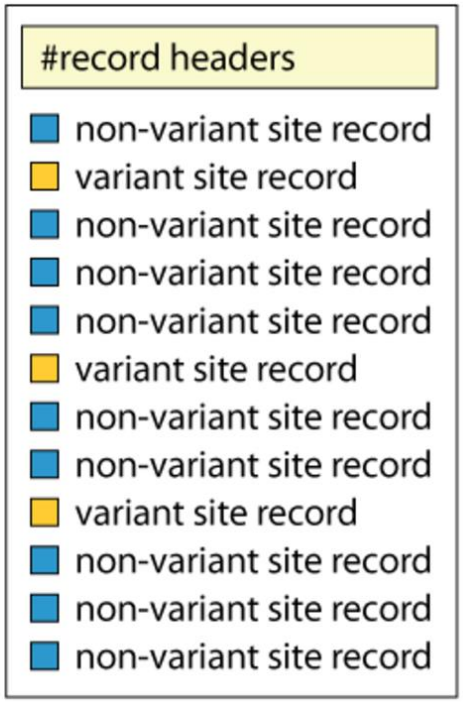


gVCF Files

New gVCF



Old gVCF




```
##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

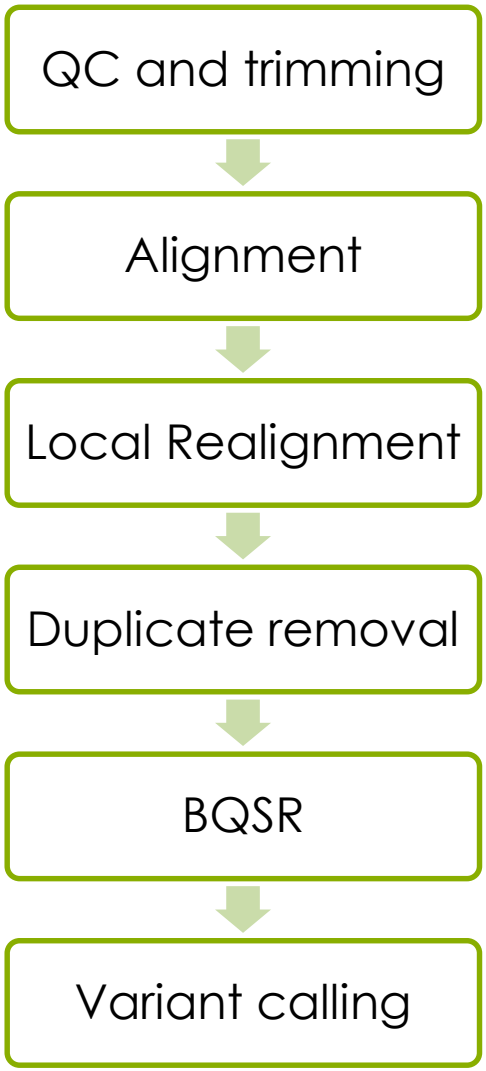
module load annovar /snpEff / vep

```
#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

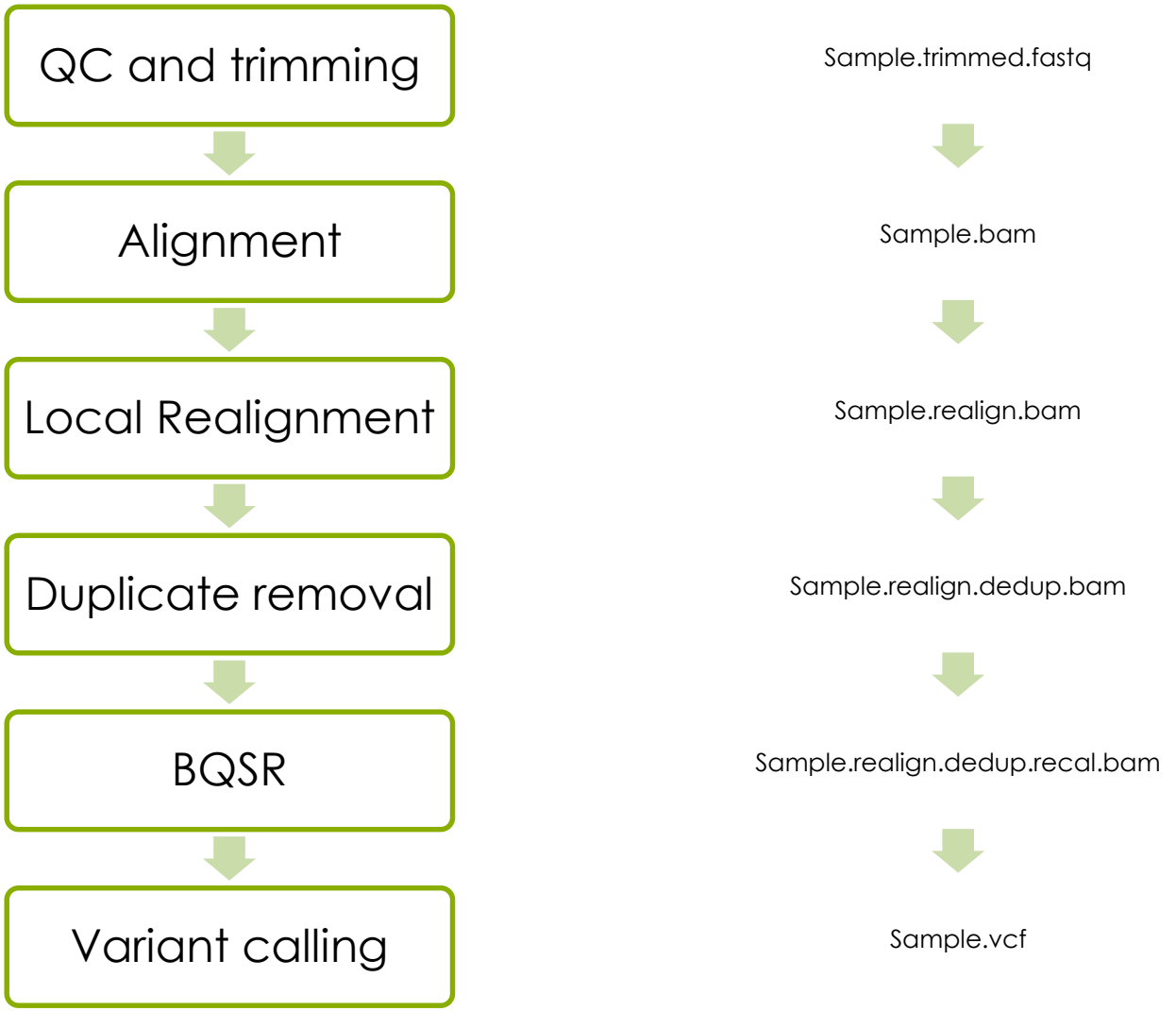
USE THE SAME REFERENCE!

File naming conventions



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

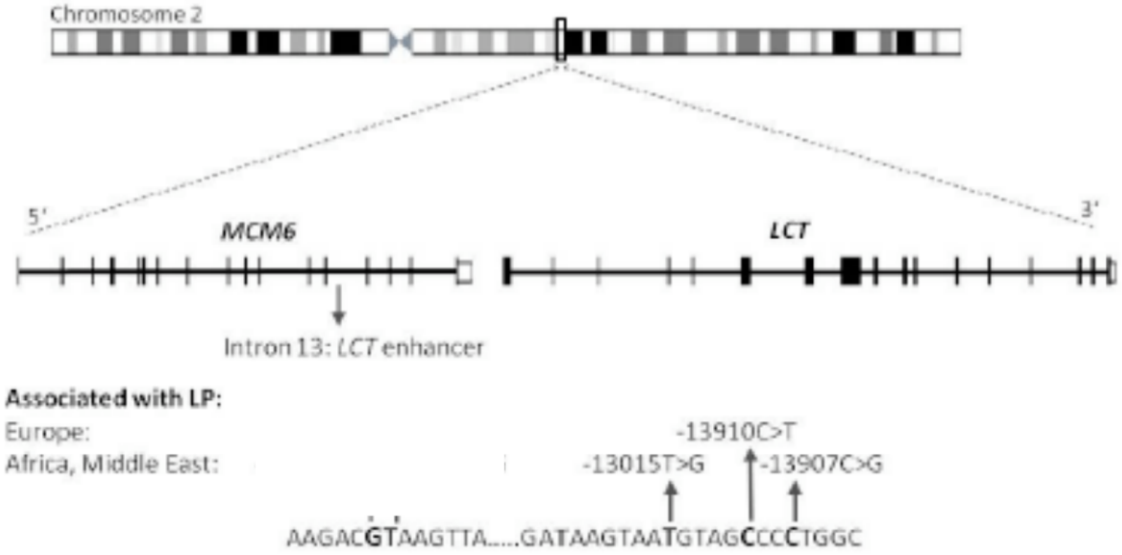
File naming conventions



Variant relating lactase persistence

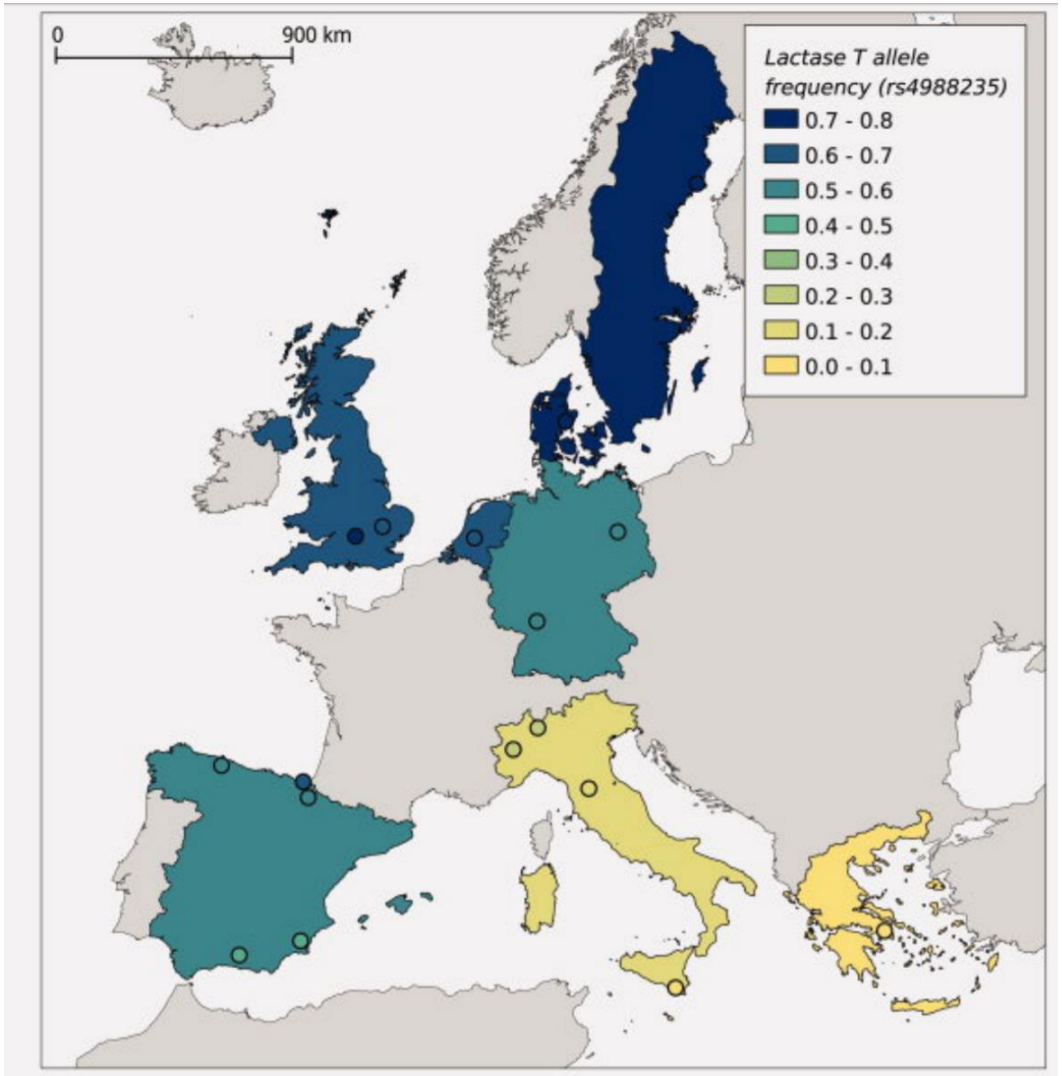
One single variant in the enhancer to the LCT gene is associated with the ability to digest lactase as adults, e.g. lactase Persistence

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

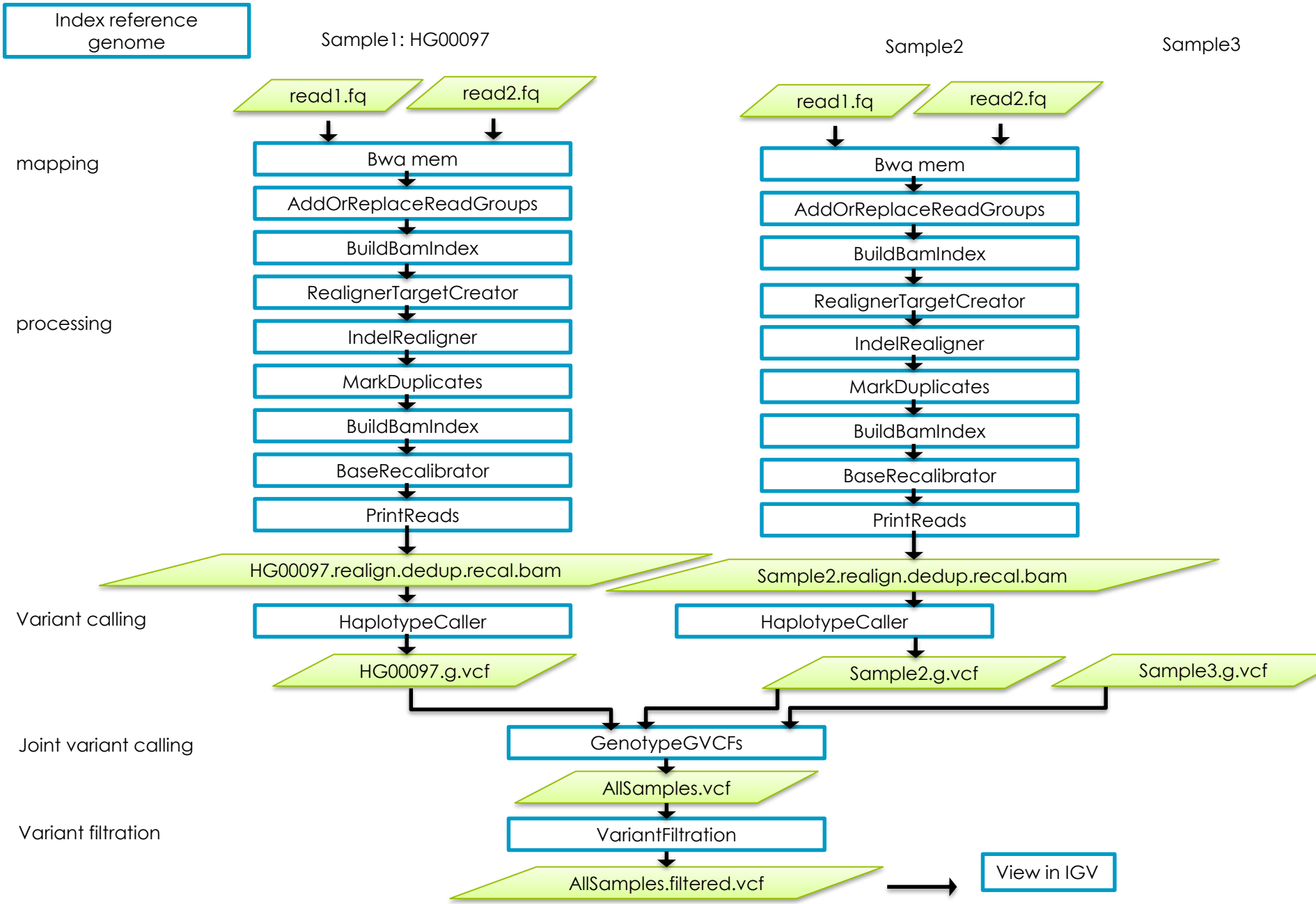


Varying allele frequencies

The allele frequencies varies between different countries, for example 74% has the alternate allele in Sweden, compared 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!