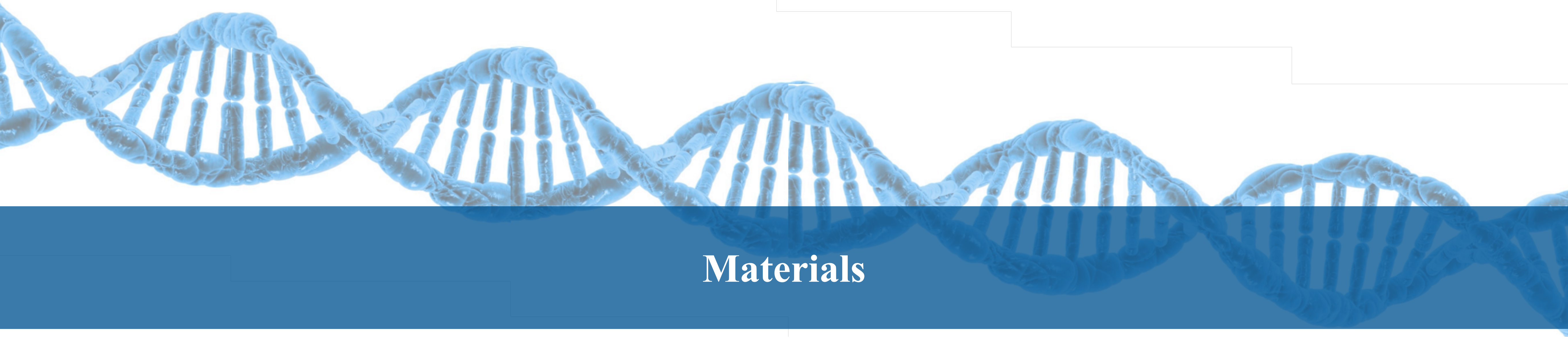
Bioinformatic Analysis of NGS Data

Bioinformatics

Nerea Moreno Ruiz & Jose Serradell

22/07/2022



Materials

https://github.com/nmorenoruiz21/IBECourse2022_SequencingData

nerea.moreno@upf.edu josemiguel.serradell@upf.edu

Install BCFtools

1) Conda

2) Homebrew

3) Download and compile

```
   git clone --recurse-submodules https://github.com/samtools/htslib.git
   git clone https://github.com/samtools/bcftools.git
   cd bcftools
   make
   make install
   export BCFTOOLS_PLUGINS=/path/to/bcftools/plugins
```

Mac users can experience problems with this step. If it happens, try this Problem Mac

Anything you need is in the command guide I provided, also, ASK for help!

Content

0) Overview

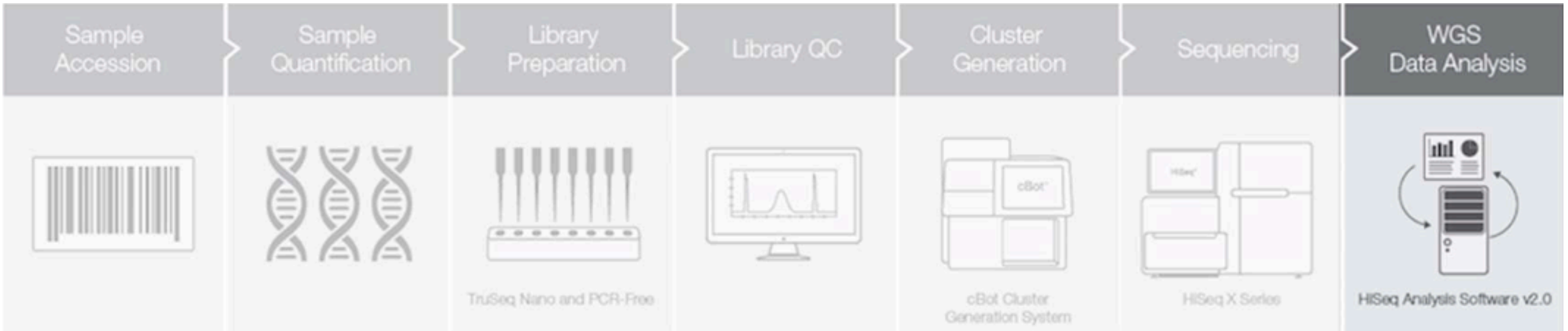
1) Sequencer Output

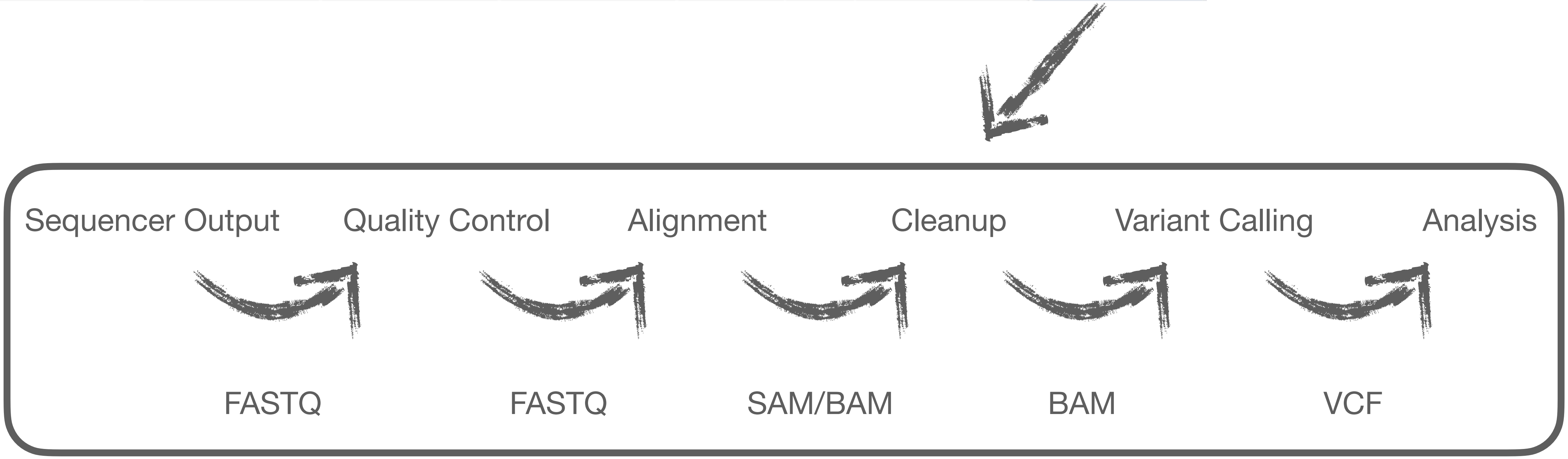
2) Quality Control

3) Alignment

4) Cleanup

5) Variant Calling

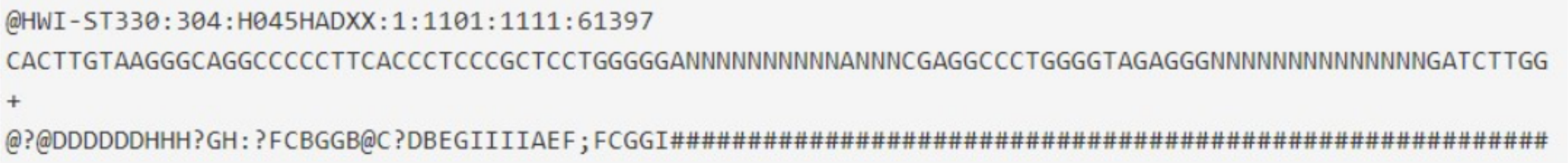




Sequencer Output → FASTQ → Quality Control → FASTQ → Alignment → SAM/BAM → Cleanup → BAM → Variant Calling → VCF → Analysis

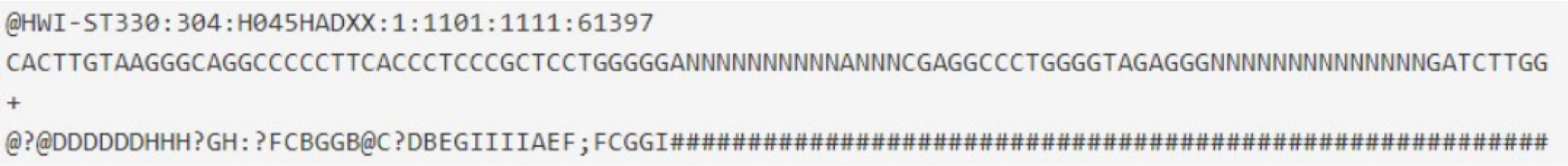
FASTQ format is a text-based format containing nucleotide sequences and their quality scores

Line	Description
1	ALWAYS starts with ‘@’ → info about the read
2	Actual DNA sequence
3	ALWAYS starts with ‘+’ → sometimes info from line 1
4	String of ASCII characters **representing** quality scores; must have same number of characters as line 2



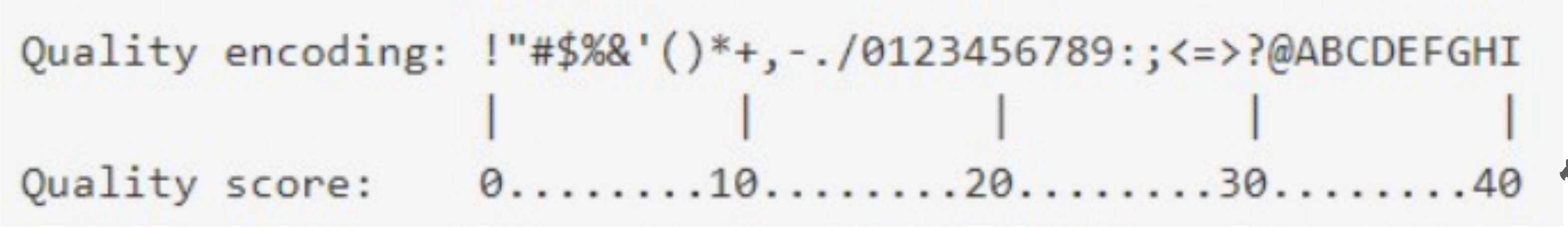
```
@HWI-ST330:304:H045HADXX:1:1101:1111:61397
CACTTGTAAGGGCAGGCCCCTTCACCCTCCCGCTCCTGGGGANNNNNNNNNANNNCGAGGCCCTGGGGTAGAGGGNNNNNNNNNNNNNNGATCTTGG
+
@?@DDDDDHHH?GH:?FCBGGB@C?DBEGIIIIAEF;FCGGI#######################################################
```

FASTQ file



```
@HWI-ST330:304:H045HADXX:1:1101:1111:61397
CACTTGTAAGGGCAGGCCCCCTTCACCCTCCCGCTCCTGGGGGANNNNNNNNNANNNCGAGGCCCTGGGGTAGAGGGNNNNNNNNNNNNNNGATCTTGG
+
@?@DDDDDDHHH?GH:?FCBGGB@C?DBEGIIIIAEF;FCGGI########################################################
```

Quality score codification

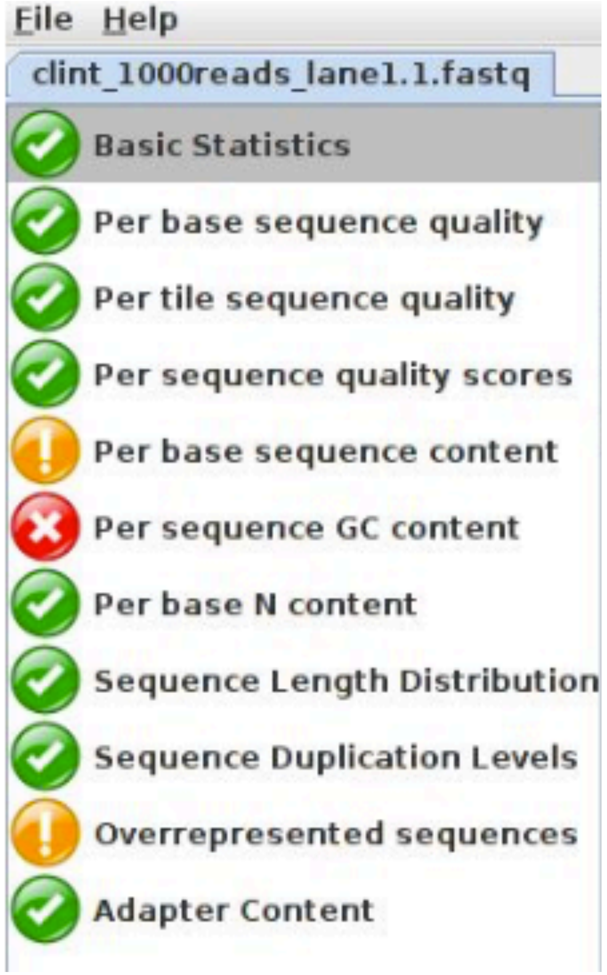


```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
Quality score:    0........10........20........30........40
```

Is the quality of this read good?

FastQC is the reference software to check quality of raw fasta sequences

- It gives us a global overview of the state of the data

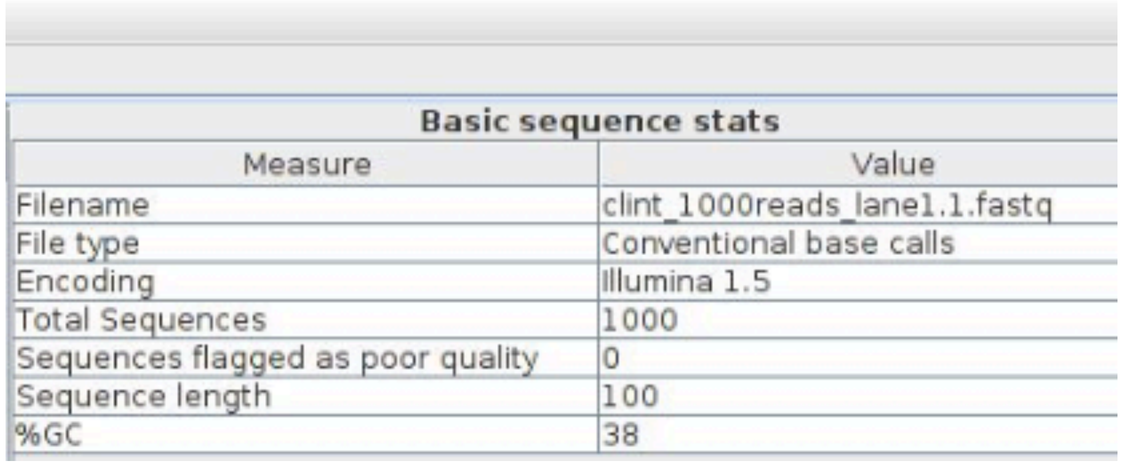


- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content

This statistics list may vary from version to version of FastQC

- Clues on factors that might be affecting the reliability of our reads

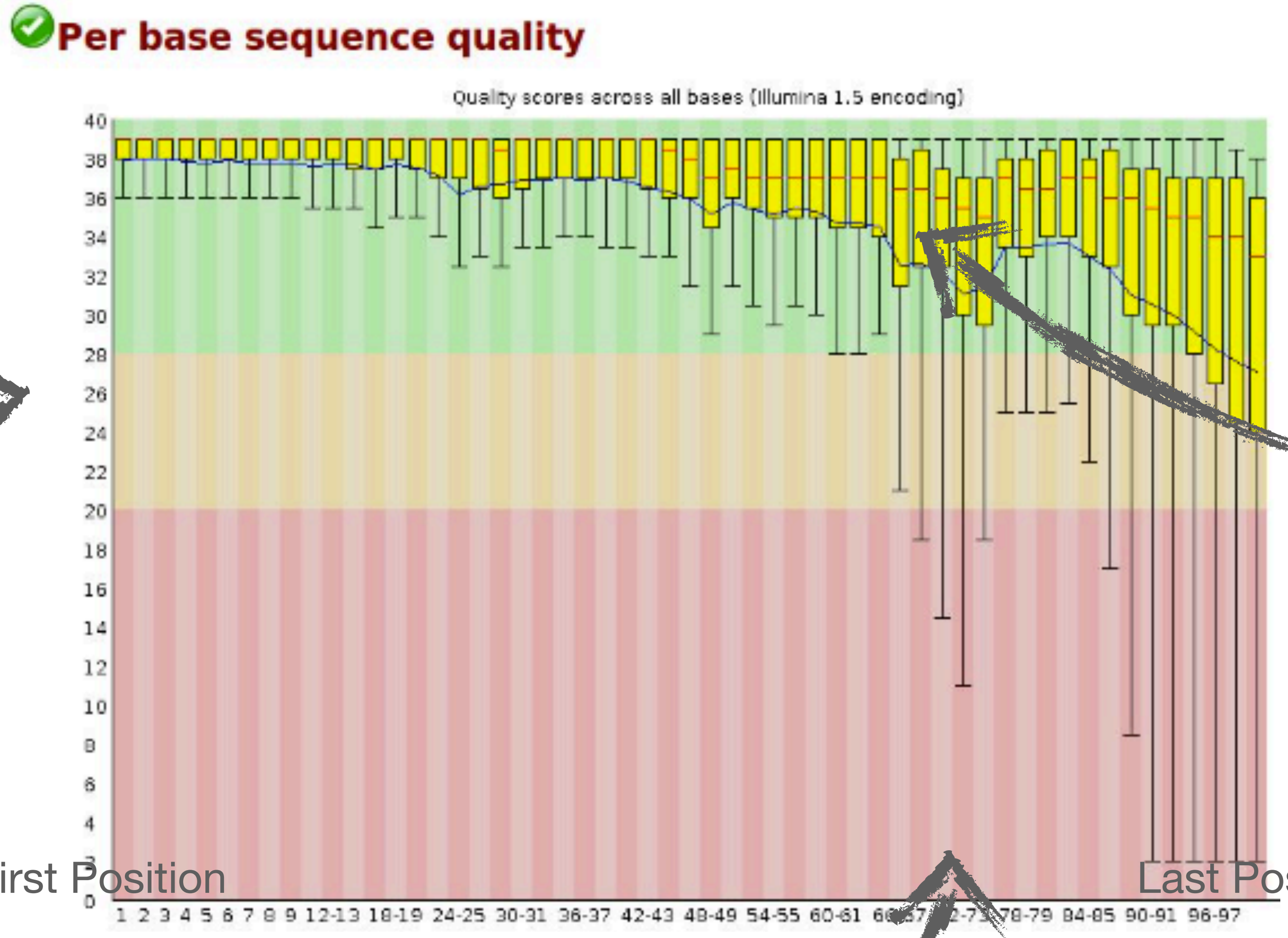
- Important to consider this in the context of different samples, experimental designs, sequencers…



Basic sequence stats

Measure	Value
Filename	clint_1000reads_lane1.1.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	100
%GC	38

- One of the most relevant data that FastQC will provide is the **per base sequence quality**



Read Quality

Box Plot indicates the **distribution** of qualities in each position across all the reads in the **sample**

First Position

Last Position

Nucleotide Position in the Read

Where does each read belong in the reference genome?

Mapping

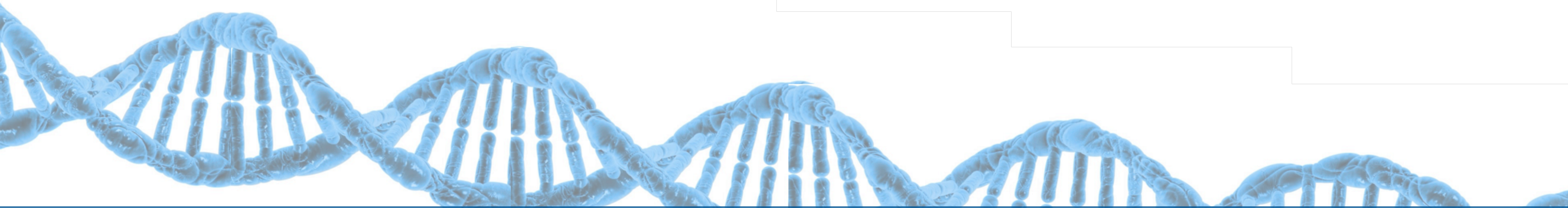
- Region where a read sequence is placed (correct if it overlaps the true region)

```
GTGGTGCATCTGTTCTCCCCGGCGGAAGTA    oqxB_EU370913
```

Alignment

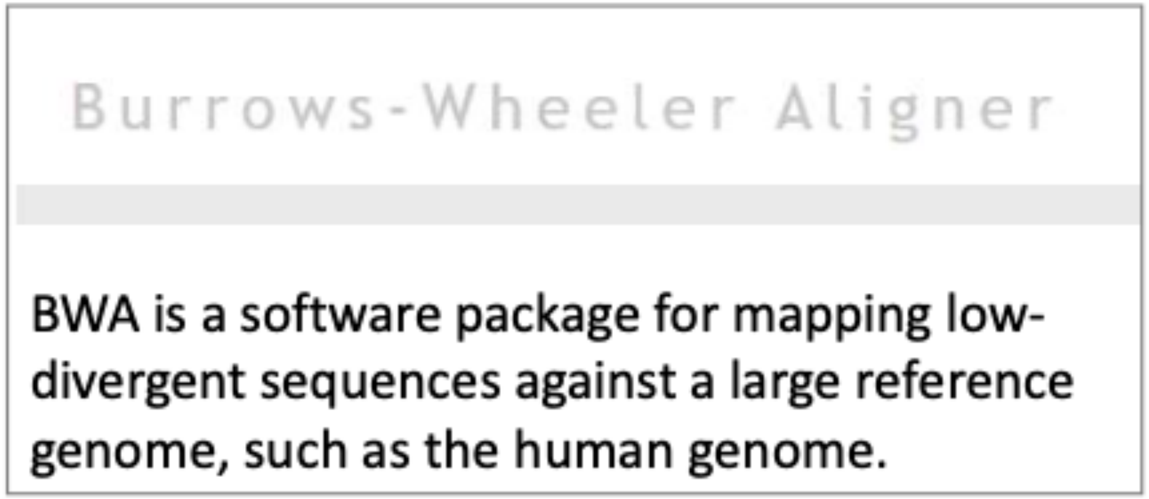
- Detailed placement of each base in a read (correct if each base is placed correctly)

```
GTGGTGCATCTGTTCTCCCCGGCGGAAGTACGACTCGCTGTATATG
||||||||||||||_||_||___|||_|||||||||||||||||||
GTGGTGCATCTGTTTTCGCCAAACGGTAAGTACGACTCGCTGTATATG
```



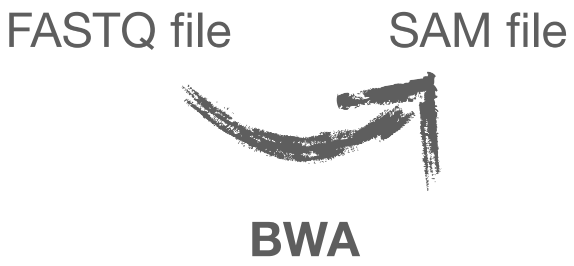
How do we perform the alignment?

- BLAST
- Bowtie2
- **BWA-MEM**
- GraphMap
- MiniMap2
- KMA



Burrows-Wheeler Aligner

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome.

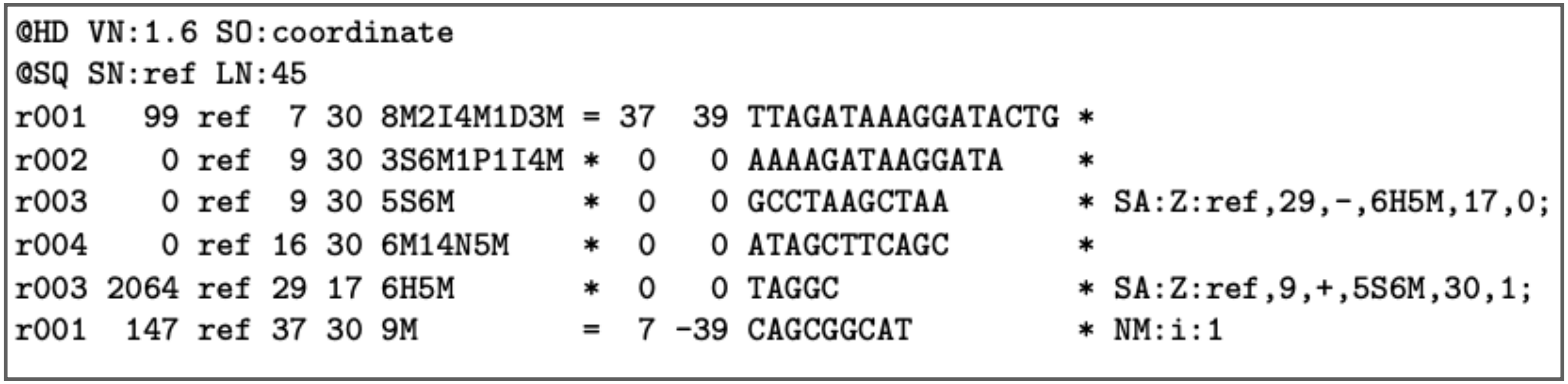


i. Align (FASTQ→SAM)

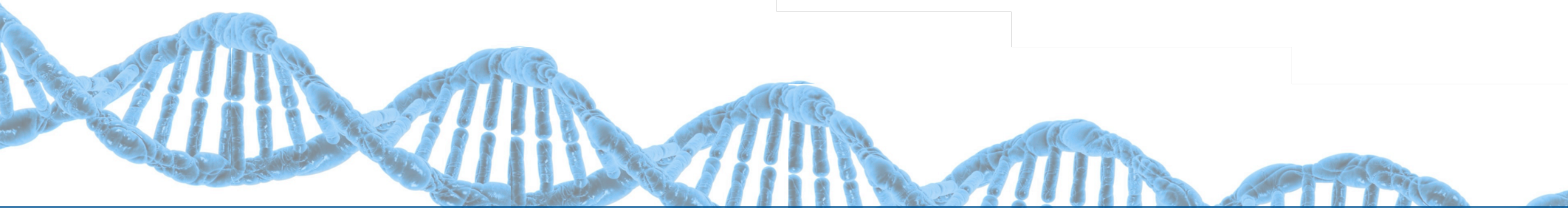
```
>bwa mem -M REFERENCE_GENOME SP1.fastq > SP1.sam
```

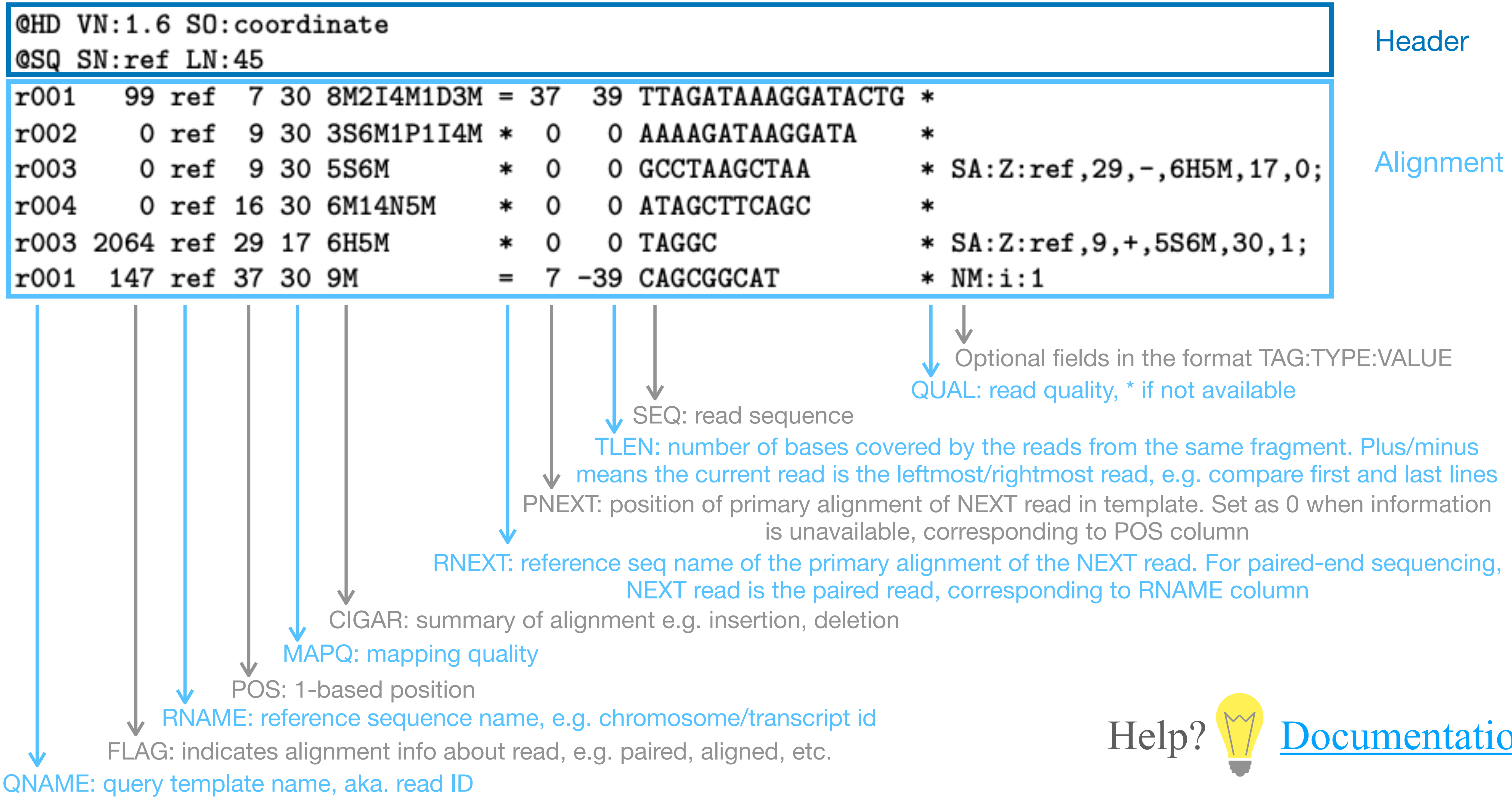
SAM format (Sequence Alignment/Map format)

- TAB-delimited text format with optional header section starting with ‘@’
- Alignment section with 11 mandatory fields for essential alignment information (coordinates, etc) + variable number of optional fields



```
@HD VN:1.6 SO:coordinate
@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       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```





Header

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
```

Alignment

```
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       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Optional fields in the format TAG:TYPE:VALUE

QUAL: read quality, * if not available

SEQ: read sequence

TLEN: number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read, e.g. compare first and last lines

PNEXT: position of primary alignment of NEXT read in template. Set as 0 when information is unavailable, corresponding to POS column

RNEXT: reference seq name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to RNAME column

CIGAR: summary of alignment e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

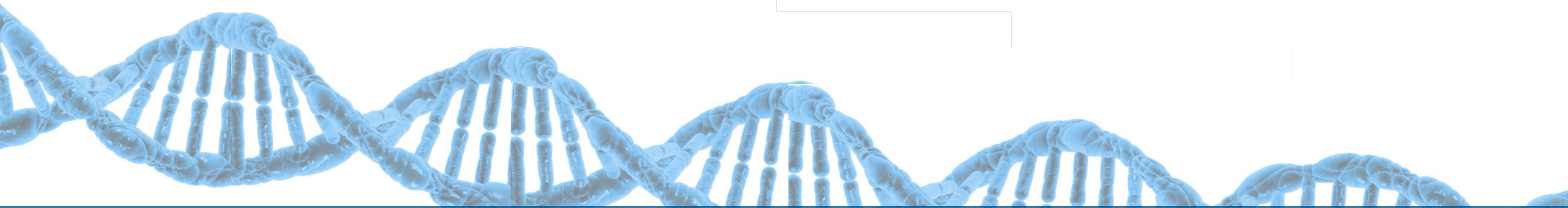
RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment info about read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

Help? [Documentation]

nerea.moreno@upf.edu   josemiguel.serradell@upf.edu



Cleanup

- Post-alignment filtering to ensure quality and discard sequencing artifacts
- First we convert **SAM** into **BAM** files, the **compressed binary** version of SAM

i. Reorder Sam (SAM→BAM)

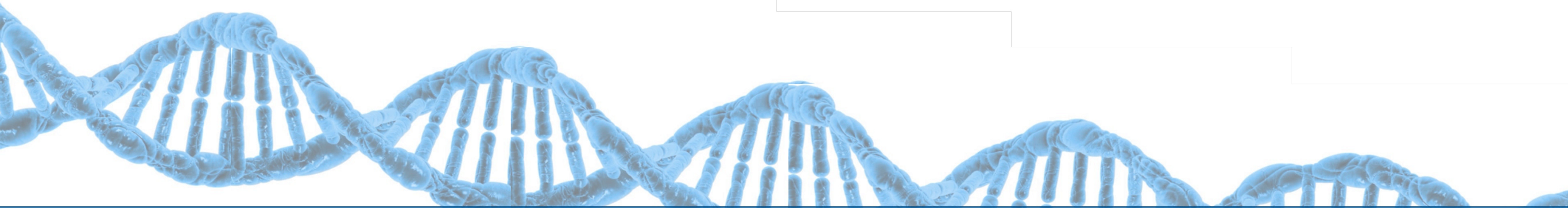
```
>java -Xmx8g -jar ReorderSam INPUT=SP1.sam OUTPUT=SP1_reordered.bam
REFERENCE=REFERENCE_GENOME
```

ii. Sort Sam (BAM→BAM)

```
>java -Xmx8g -jar SortSam SORT_ORDER=coordinate INPUT=SP1_reordered.bam
OUTPUT=SP1_sorted.bam
```

iii. Mark Duplicates (BAM→BAM)

```
>java -Xmx36g -jar MarkDuplicates INPUT=SP1_sorted.bam
OUTPUT=SP1_MD_sorted.bam METRICS_FILE=SP1_metrics.txt
```



iv. Add or Replace Read Groups (BAM→BAM)

```
>java -Xmx36g -jar AddOrReplaceReadGroups INPUT=SP1_MD_sorted.bam
SORT_ORDER=coordinate RGLB=algo RGPL=illumina RGPU=7 RGSM=SP1
OUTPUT=SP1_RG.bam
```

v. BaseQualityRecalibrator (BAM→BAM)

```
>gatk --java-options "-Xmx15G" BaseRecalibrator -I SP1_RG.bam -R
REFERENCE_GENOME --known-sites KNOWN_DATABASES -O SP1_RG_data.table
```

vi. Apply BQSR

```
>gatk --java-options "-Xmx15G" ApplyBQSR -R REFERENCE_GENOME -I
SP1_RG.bam --bqsr-recal-file SP1_RG_data.table -O BQSR_SP1.bam
```

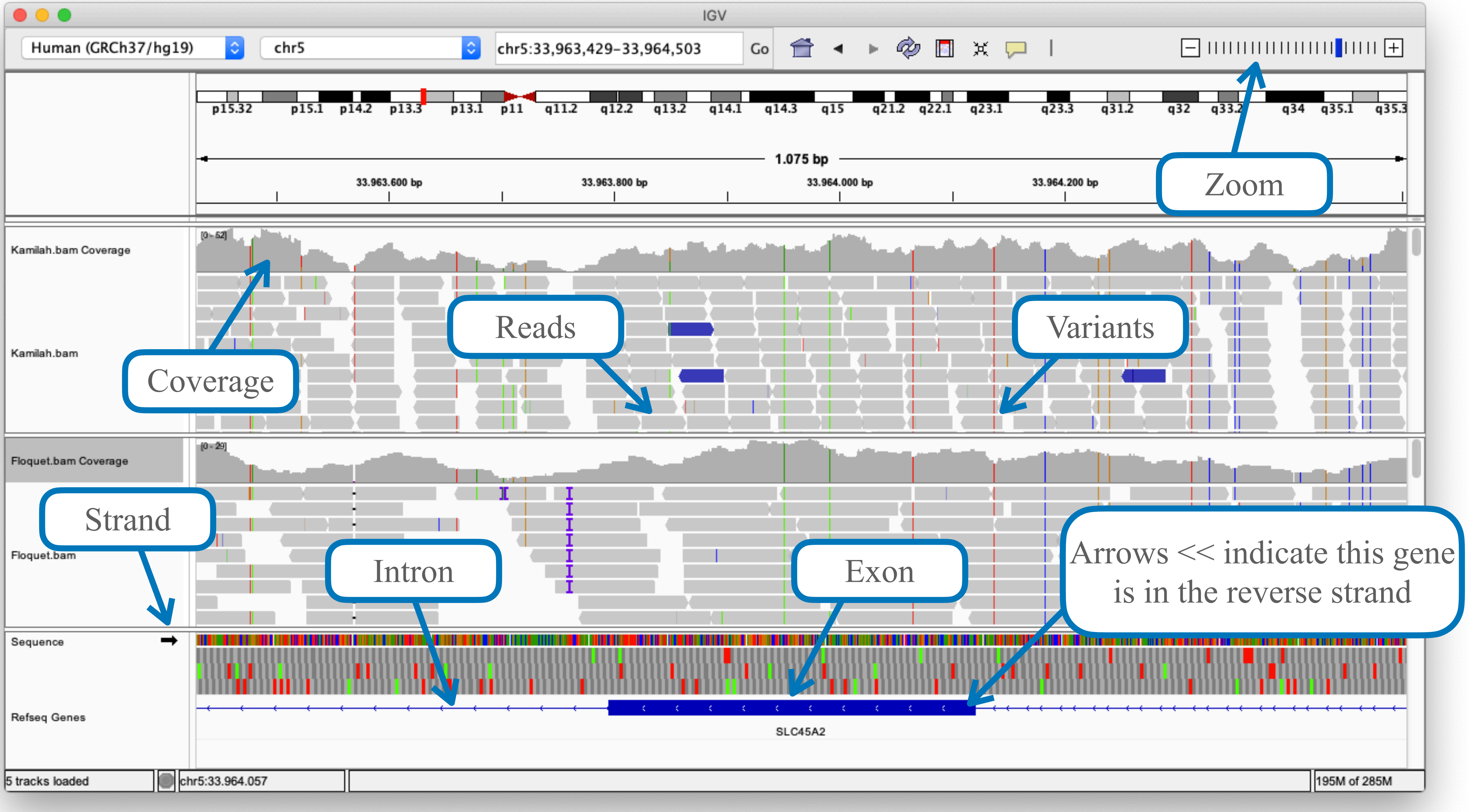
vii. Index (BAM→BAM+BAI)

```
>samtools index BQSR_SP1.bam
```

BAI

- Step (vii) creates an index file for the BAM
- This file is created with the same name + the sufix .bai instead of .bam
- This file acts as a table of contents and allows programs to traverse easily through the BAM by jumping directly to specified coordinates, etc
- BAI alone is useless since it doesn’t actually contain any sequence data

Visualization



Zoom

Reads

Variants

Coverage

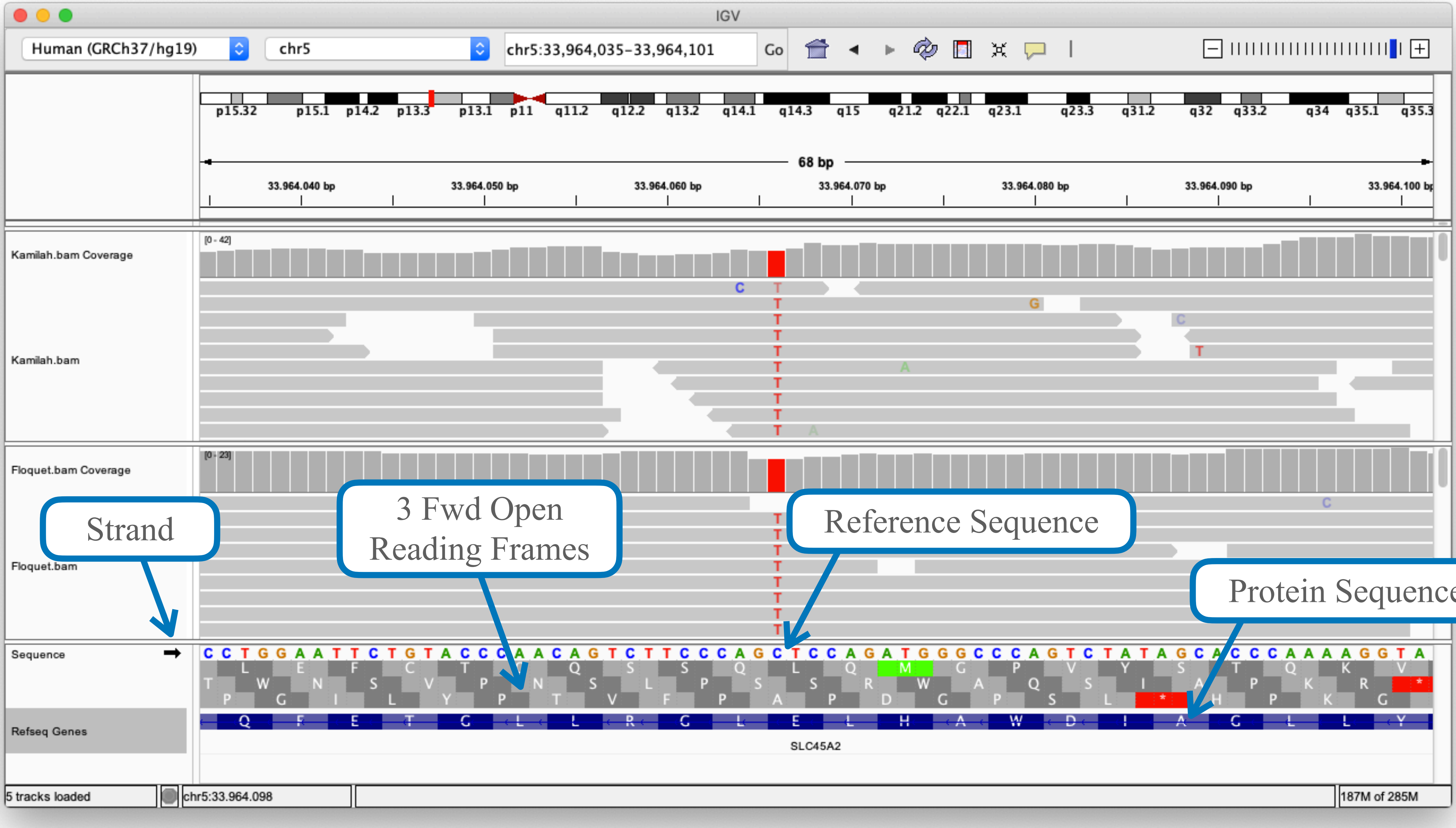
Strand

Intron

Exon

Arrows << indicate this gene is in the reverse strand

Exercise 2. Find the mutation in IGV



Human (GRCh37/hg19)

chr5

chr5:33,964,035-33,964,101

5 tracks loaded

chr5:33.964.098

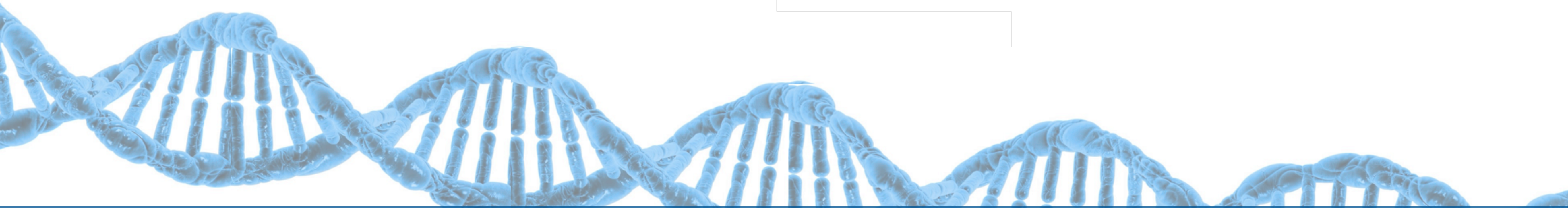
187M of 285M

Strand

3 Fwd Open Reading Frames

Reference Sequence

Protein Sequence



Variant Calling

- As you have seen, visual inspection of BAM files to find variants is hard and time consuming
- The next step of the pipeline is to identify differences between the sample and the reference genome and compile them in a Variant Call Format file (VCF)

In the next session we will learn more about VCF files and how to deal with them

COMING
~~SOON~~
RIGHT NOW

Variant Calling and VCF files

Bioinformatics

Content

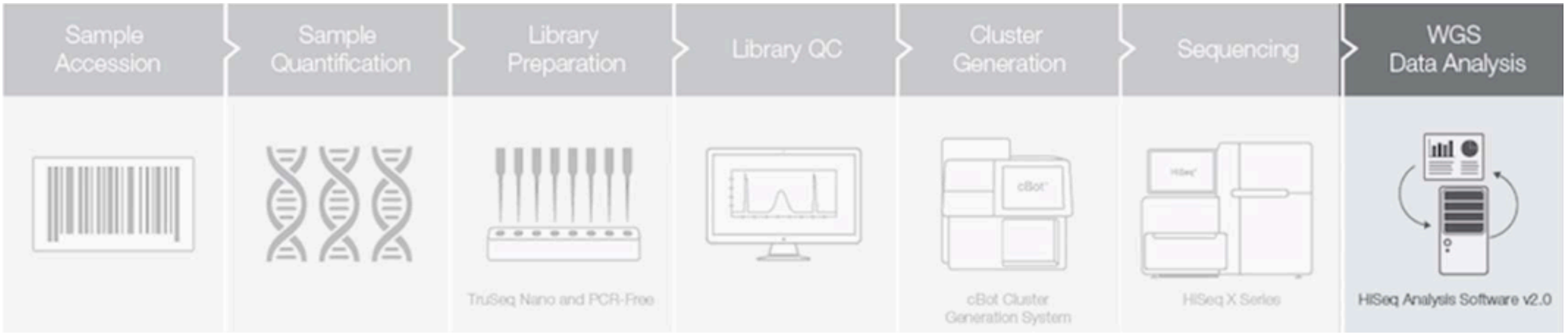
1) Variant Calling

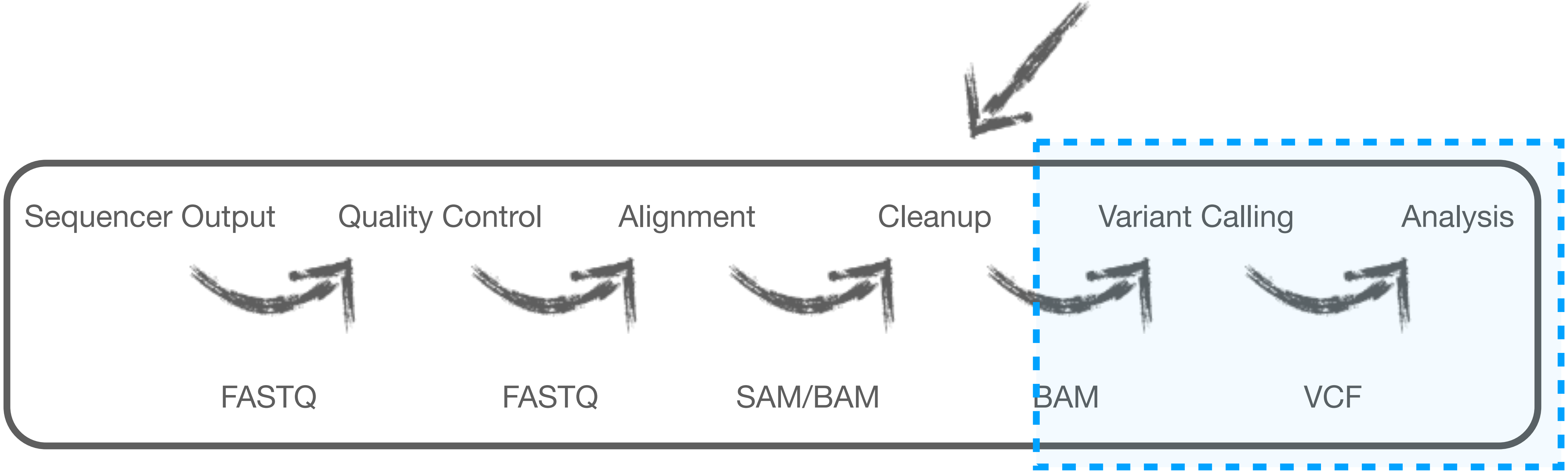
2) Variant Calling Format

3) Possibilities

4) Filtering VCFs

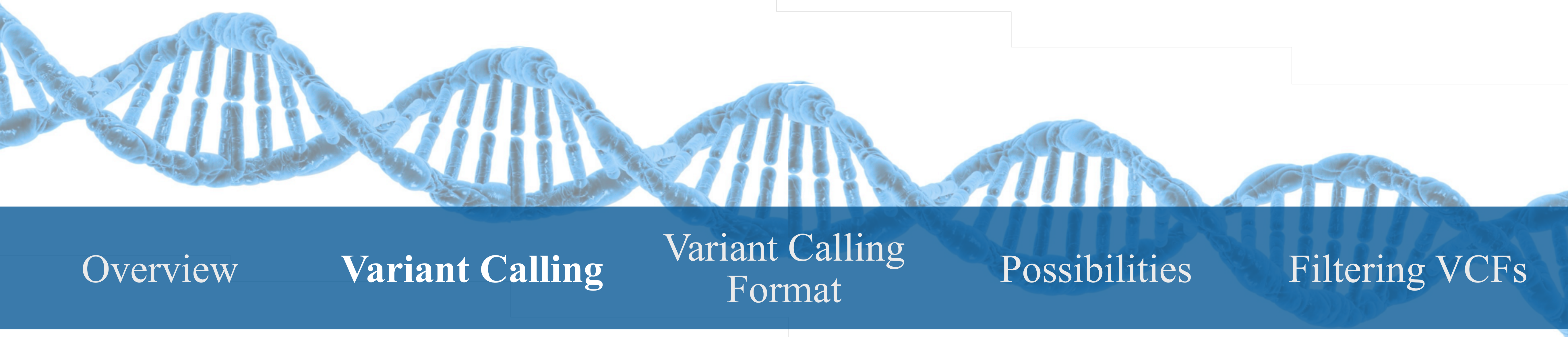
Exercise 1. Use the command line to answer the questions





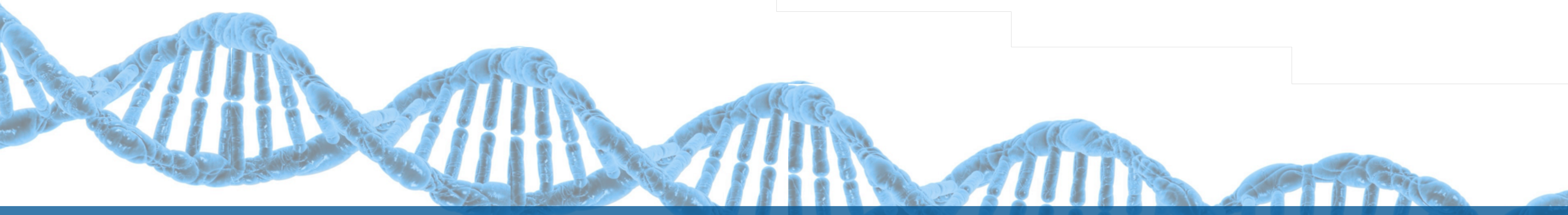
Sequencer Output → Quality Control → Alignment → Cleanup → Variant Calling → Analysis

FASTQ FASTQ SAM/BAM BAM VCF



What is variant calling?

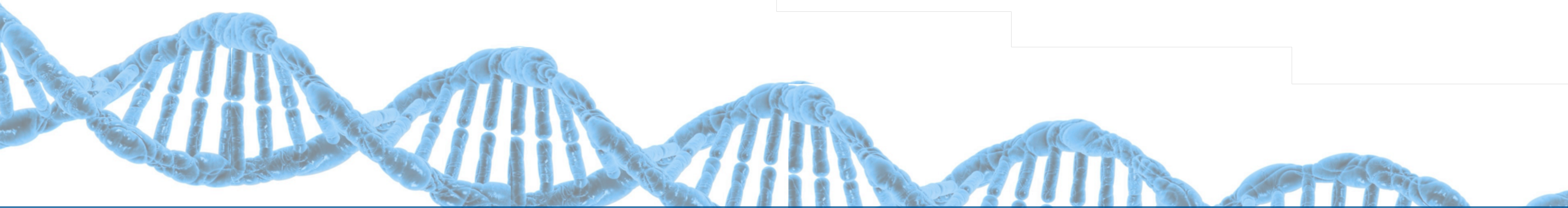
- Identifying where the aligned reads differ from the reference genome and writing the information into a **VCF** file
- The most used tool to call SNVs and indels is **GATK**’s **HaplotypeCaller**
- There are 3 scenarios when calling variants
  - Homozygous for the reference allele
  - Homozygous for an alternative allele
  - Heterozygous



HaplotypeCaller

- Takes a **BAM** file as input
- Calls SNVs and indels simultaneously
- As most modern callers, it uses the Bayes theorem
- Performs local re-assembly to identify haplotypes
- It is more accurate compared to site by site callers, especially for indels
- Returns a **VCF** (Variant Call Format) file as output

Help? 💡 GATK



ix. Haplotype caller

```
>gatk --java-options "-Xmx18G" HaplotypeCaller -R REFERENCE_GENOME -ERC
GVCF -I BQSR_SP1.bam -O BQSR_SP1.g.vcf.gz
```

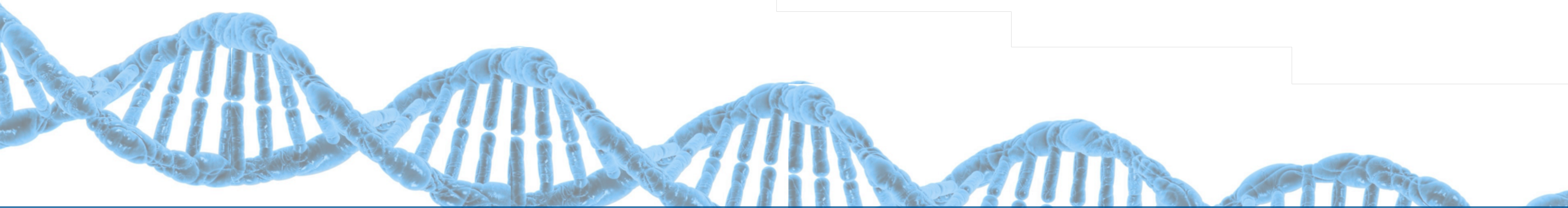
x. Multisample VCF

```
>gatk --java-options "-Xmx130g" GenomicsDBImport -R REFERENCE_GENOME -V
BQSR_SP1.g.vcf.gz --genomicsdb-workspace-path GVCF_DATABASE_CHR# -L CHR#
--batch-size 30 --reader-threads 5 --tmp-dir TMP
>gatk --java-options "-Xmx10g -Xms5g" GenotypeGVCFs -R REFERENCE_GENOME
-V gendb://GVCF_DATABASE_CHR# --create-output-variant-index —output
MULTISP.vcf.gz
```

xi. Merge chromosomes

```
>java -jar PICARD GatherVcfs I=MULTISP_chr1.vcf.gz I=MULTISP_chr2.vcf.gz
... O=MULTISP_merged.vcf.gz
```

xii. Split SNPs and INDELs to faster recalibration with GATK -SplitVcfs or BCFtools



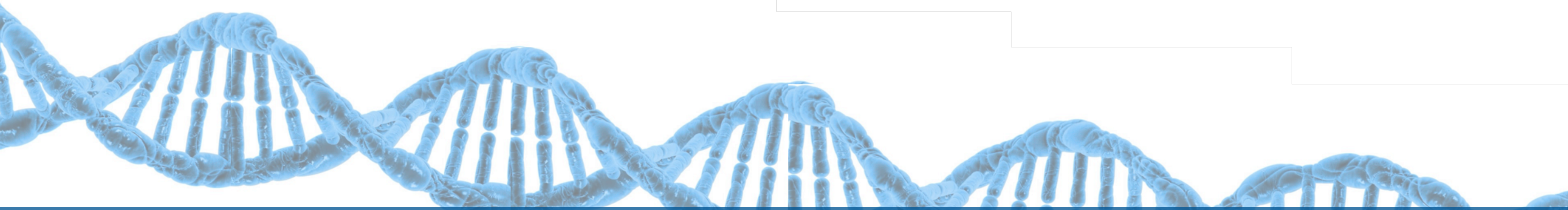
xiii. Variant Quality Recalibration

```
>gatk --java-options "-Xmx20g -Xms5g" VariantRecalibrator -R
REFERENCE_GENOME -V MULTISP_snps.vcf.gz -tranche 100.0 -tranche 99.95
-tranche 99.9 -tranche 99.8 -tranche 99.6 -tranche 99.5 -tranche 99.4
-tranche 99.3 -tranche 99.0 -tranche 98.0 -tranche 95.0 -tranche 90.0 –
resource:DB,known=false,training=true,truth=true,prior=15.0 DB … -an QD
-an MQ -an MQRankSum -an ReadPosRankSum -an FS -an DP -mode SNP -O
MULTISP_merged_snps.recal --tranches-file MULTISP_merged_snps.tranches
--rscript-file MULTISP_snp_Recalibration.plots.R
```

xiv. Same with INDELS

xv. Apply Recalibration

```
>gatk --java-options "-Xmx10g -Xms5g" ApplyVQSR -R REFERENCE_GENOME -V
MULTISP_merged_snps.vcf.gz --recal-file MULTISP_merged_snps.recal --
tranches-file MULTISP_merged_snps.tranches --truth-sensitivity-filter-
level 99.9 --create-output-variant-index true -mode SNP -O
MULTISP_merged_snprecal99.9.vcf.gz
```



VCF format is a text-based tab-delimited format that contains variant information

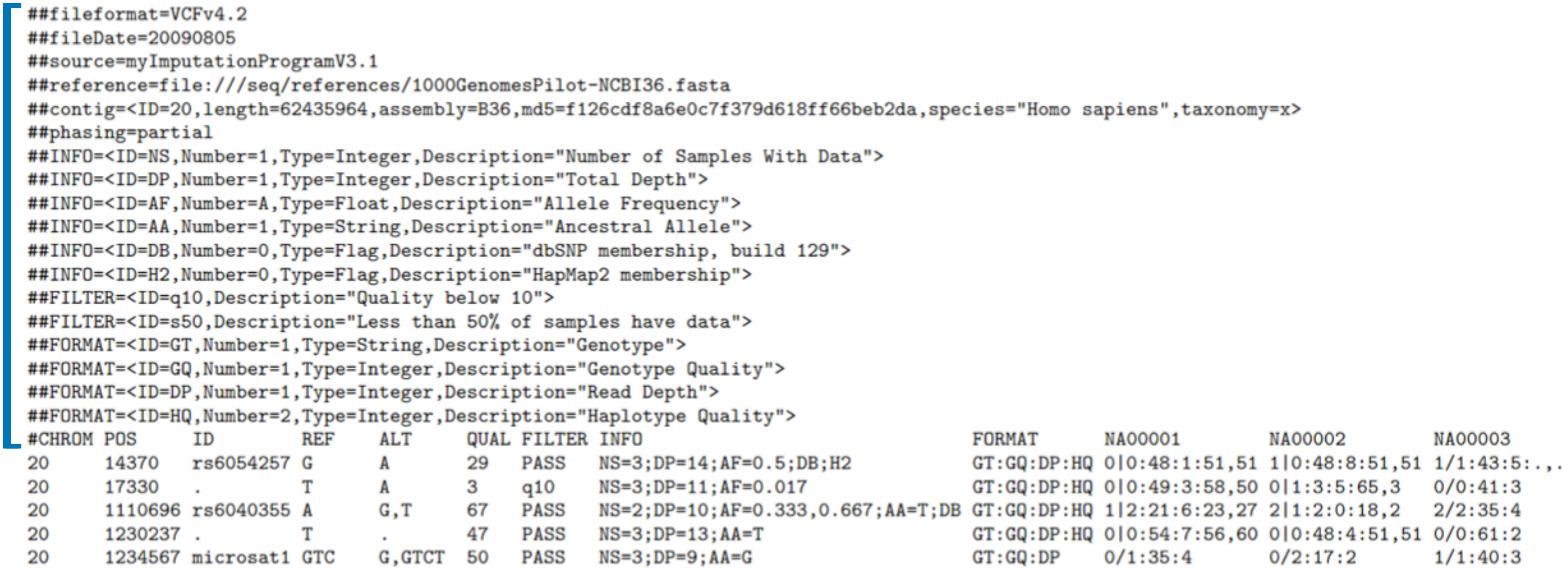
- It always has the same structure, a header (#) and several lines containing the information, each line is a position in the genome

- Usually sorted, compressed and indexed to reduce size and access the information faster

- **BCF** is the binary version of this format and it is also handled using **BCFtools**. This format is used to deal with large amounts of data like whole genome sequencing from several individuals

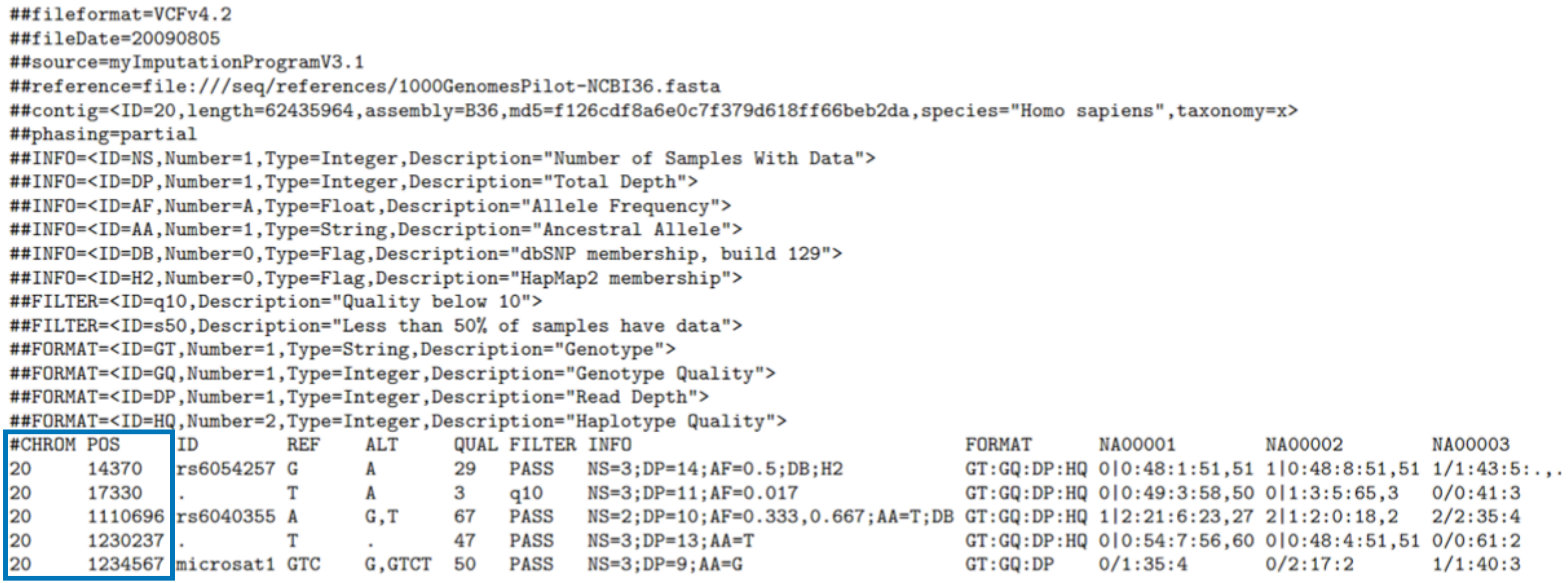
- VCF can contain genotypes for none, one or several individuals

Help? 💡 BCFtools



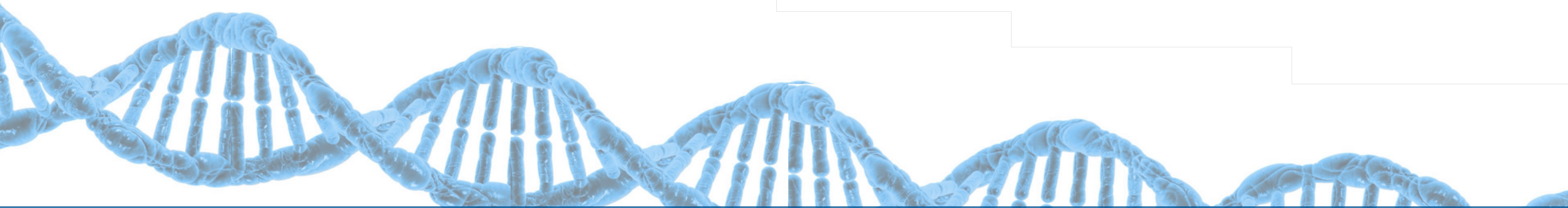
```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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">
##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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 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
```

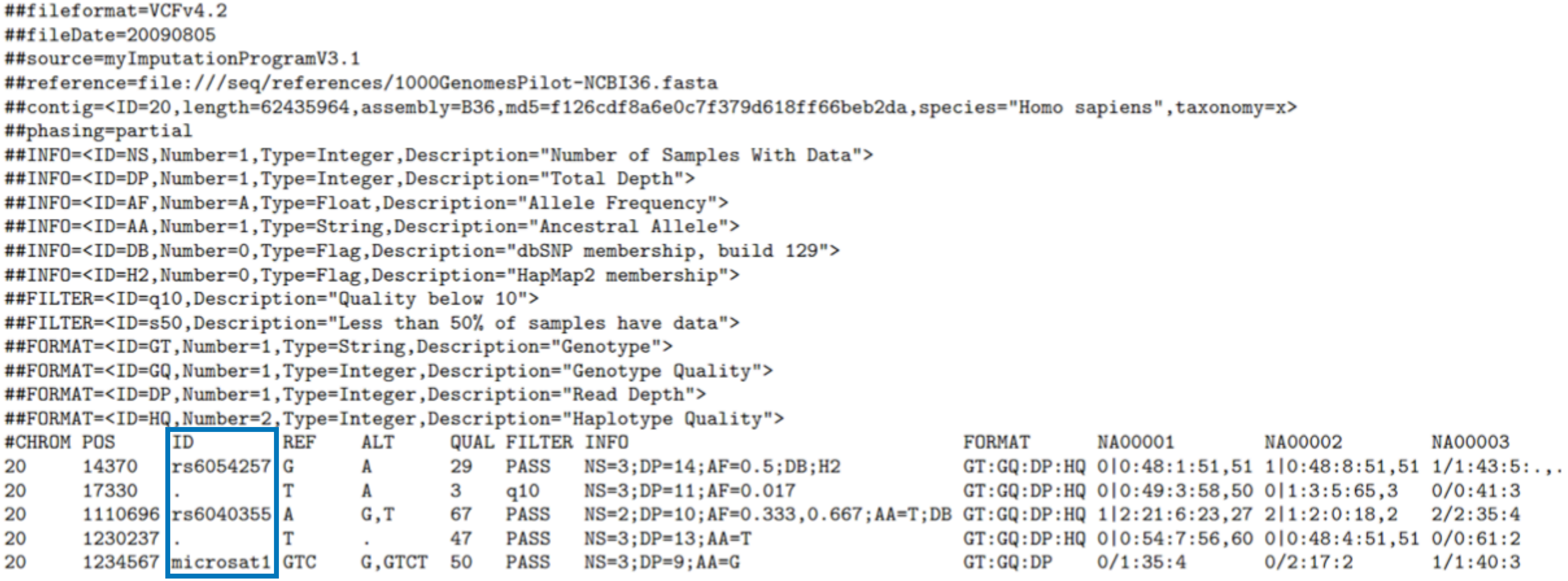
Header lines start with #:
## contains information about the fields and tags that are used throughout the file and record processes applied to the file
# line indicates the name of each column



```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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">
##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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 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
```

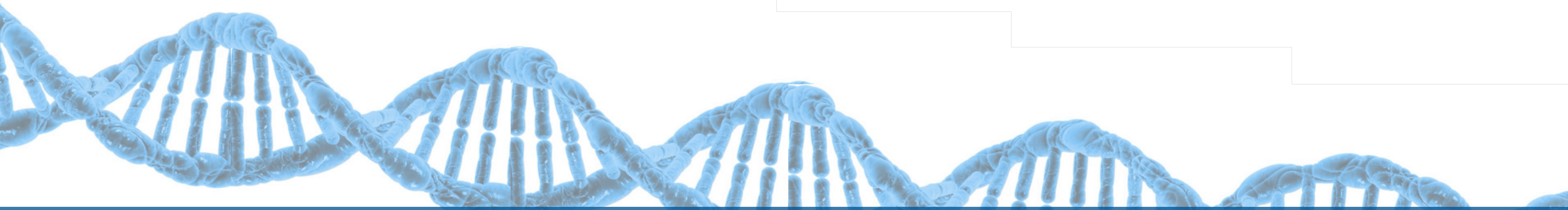
Chromosome and position of the variant in the reference genome

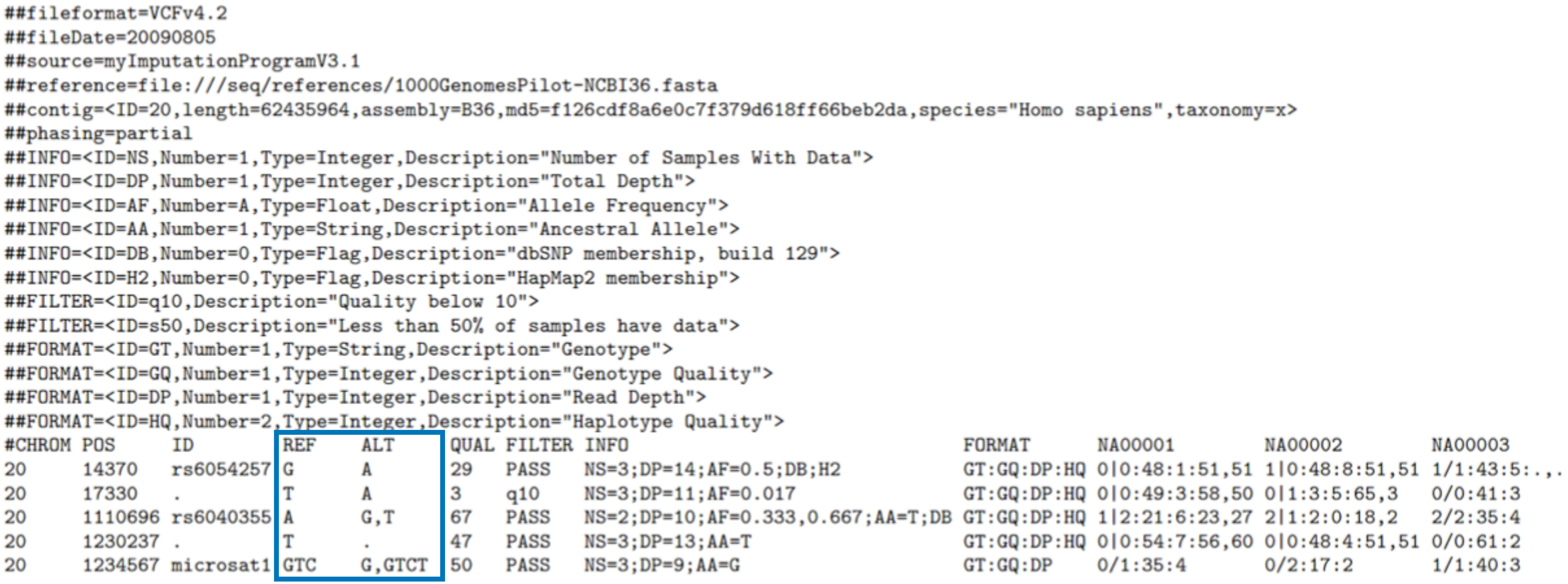




```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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">
##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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 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
```

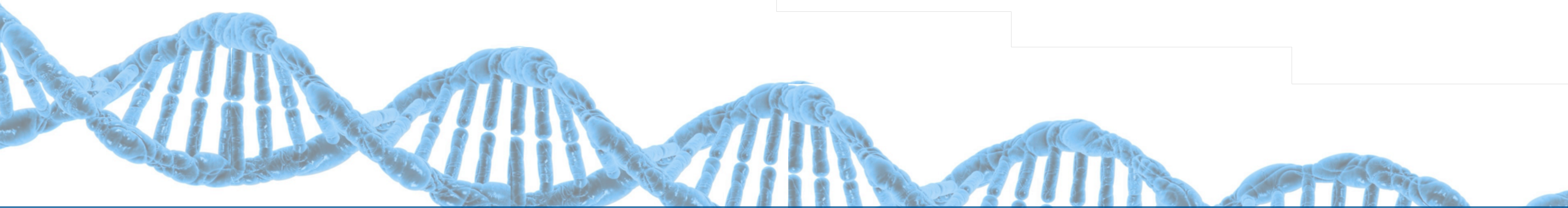
Some information of the variant. In this case we see dbSNP IDs.

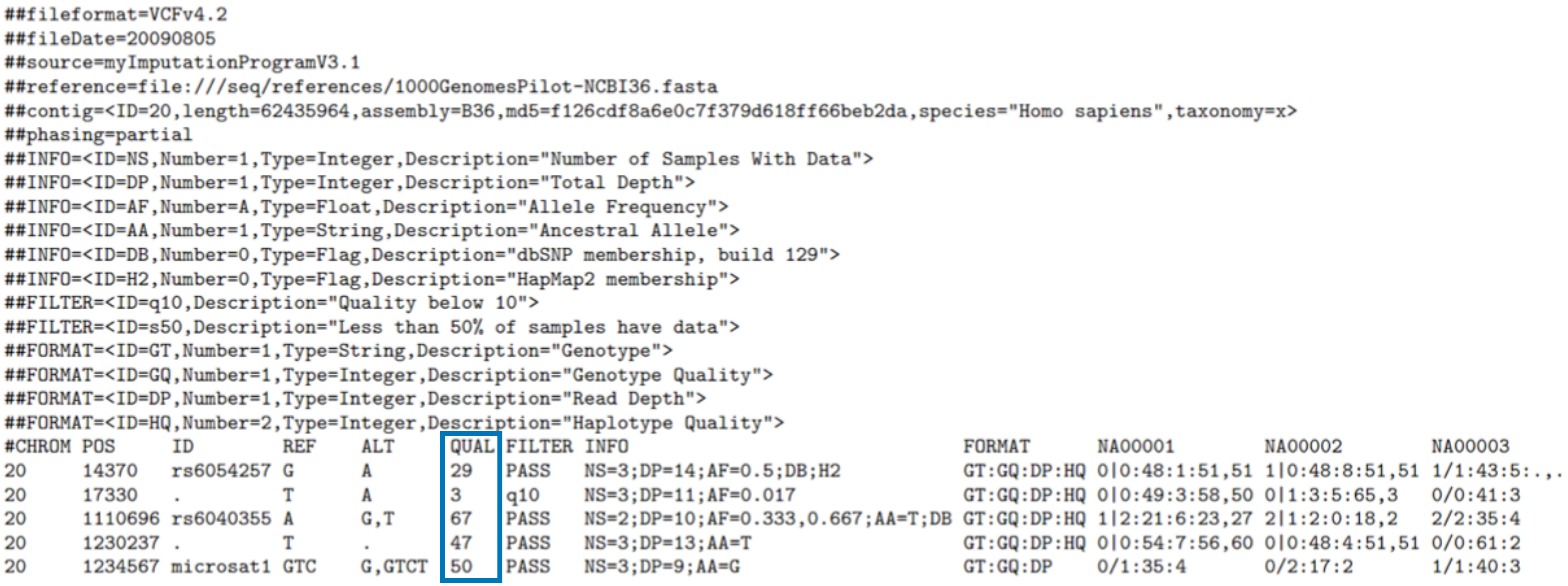




```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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">
##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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 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
```

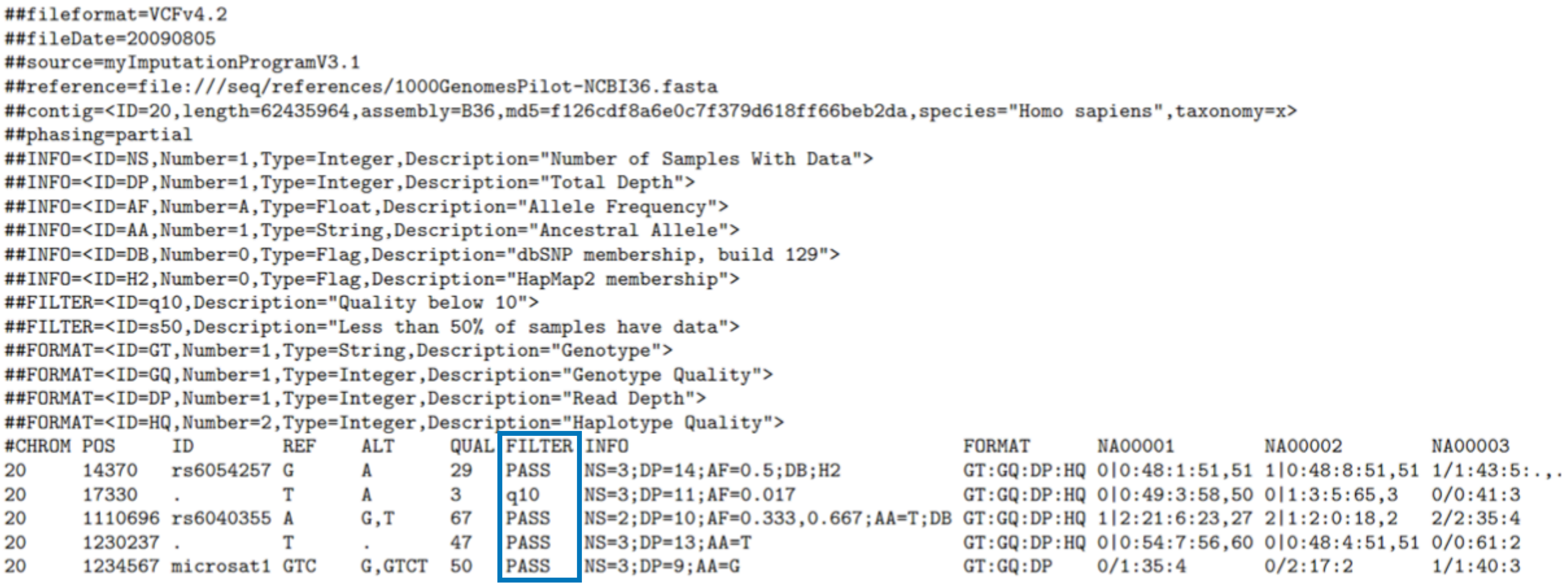
Allele in the reference genome (REF) and in the sample(s) (ALT)





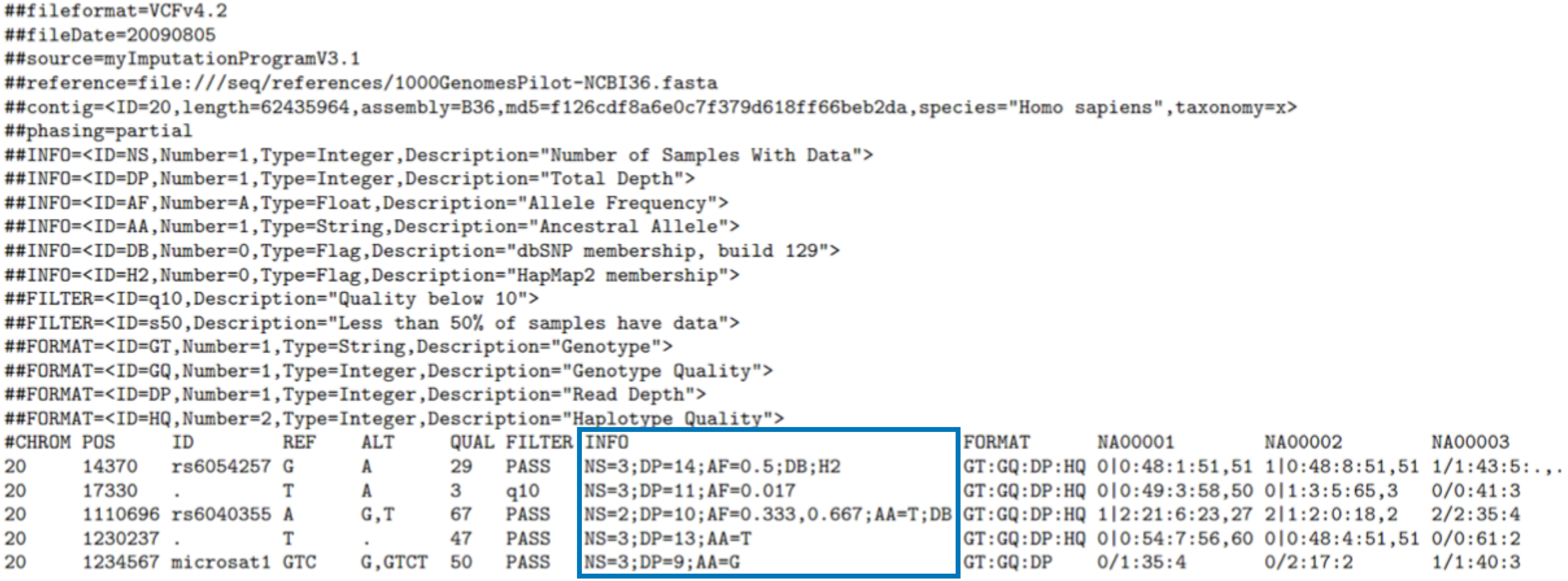
```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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">
##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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 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
```

Phred quality score of the call



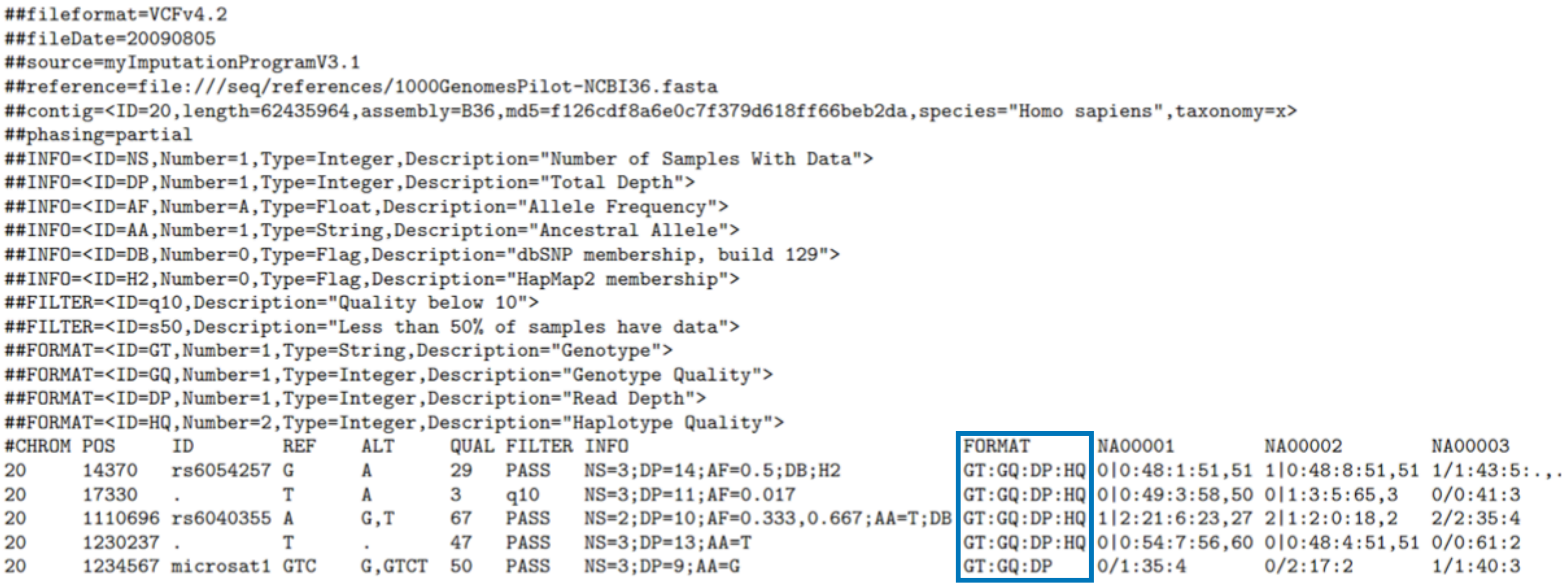
```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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">
##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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 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
```

If the call in this position passes or not the filters applied



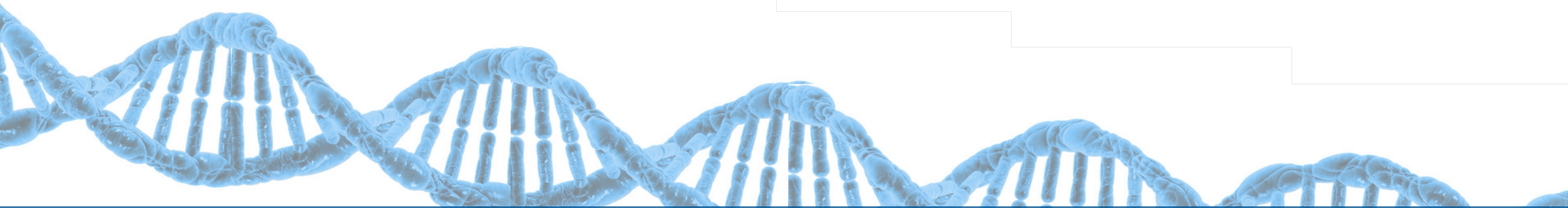
```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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">
##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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 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
```

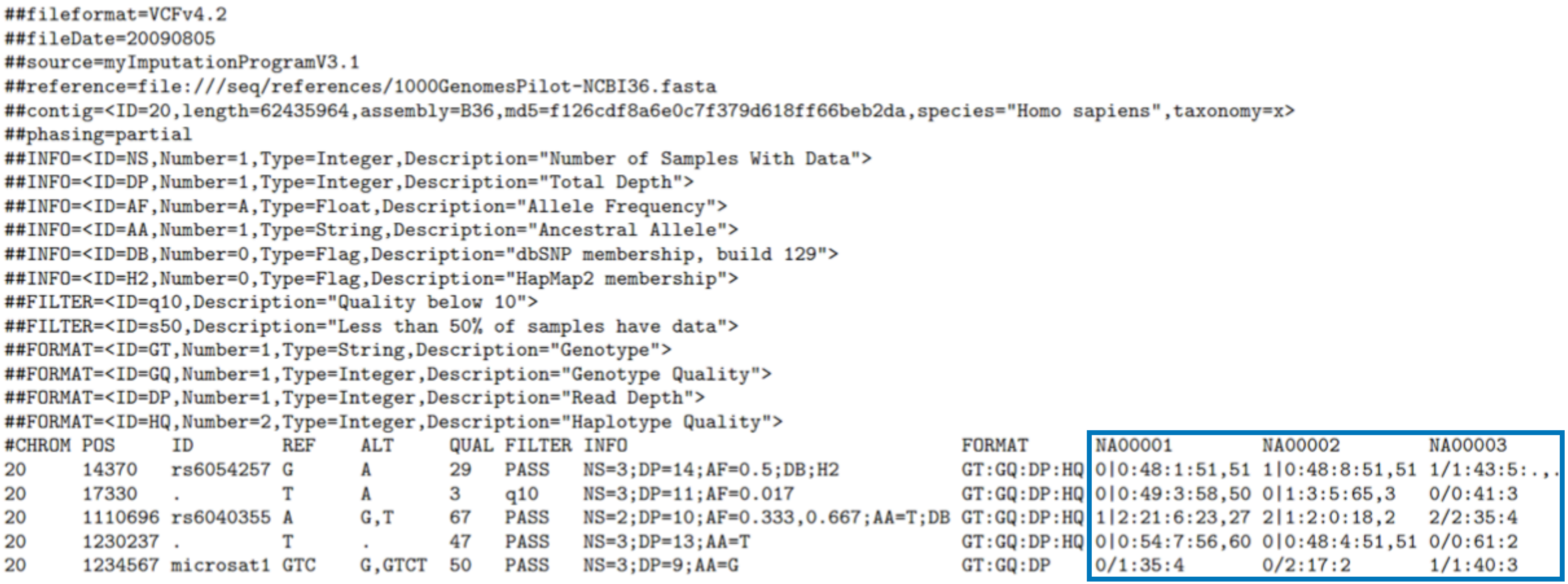
Information of the variants, described in the header



```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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">
##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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 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
```

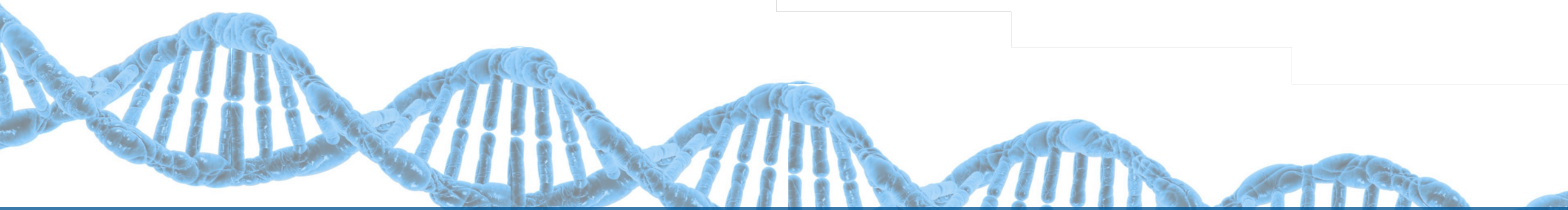
Genotype parameters that we will find for each variant in the subsequent sample columns





```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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">
##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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 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
```

Genotype for each variant in a given sample. Sample name is the header (#) of the column.
In this case we have genotypes for 3 samples



Different things we can do with a VCF file:

- View certain positions of a VCF file
- Access different information from the INFO fields
- Retrieve information from a given individual
- Build a PCA with PLINK
- Annotate functional information of the variants with ANNOVAR, SnpEff, VEP…
- **Filter variants combining the use of BCFtools and AWK**

Additional Tools

Help? [Biowulf GATK Tutorial]

[Data Carpentry Genomics Tutorial]

[GATK Best Practices Workflows]

[Nextflow Variant Calling Tutorial]