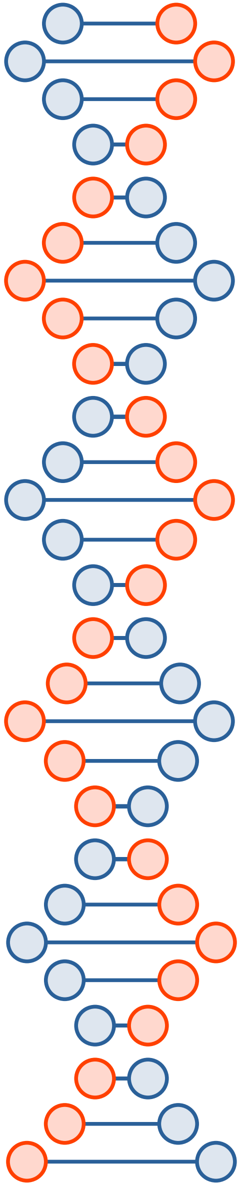
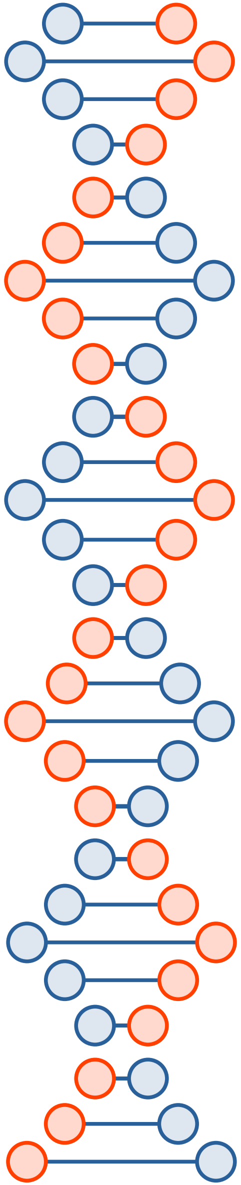


Data formats in NGS data analysis. NGS data analysis workflow.

Dr. Rafiga Masmaliyeva

CARL VON OSSIETZKY universität OLDENBURG

KLINIKUM OLDENBURG Universitätsmedizin Oldenburg

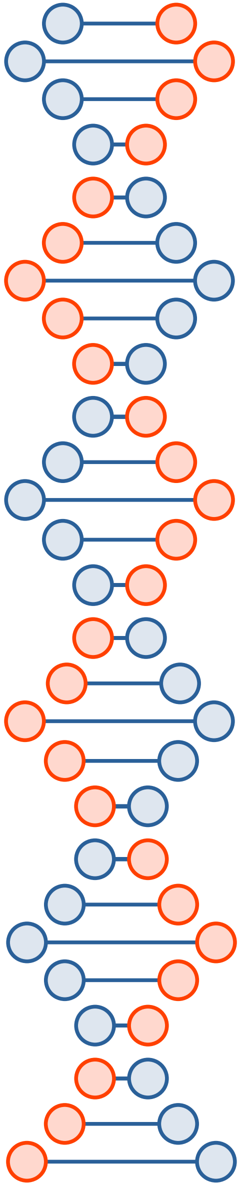


FASTA format

- FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences

An example sequence in FASTA format is:

```
>gi|186681228|ref|YP_001864424.1| phycoerythrobilin:ferredoxin oxidoreductase
MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKGKNQEEVVTTSYAFQTAKLRQIRA
AHVQGGNSLQVLNFVIFPHLNYDLPFFGADLVTLPGGHLIALDMQPLFRDDSAYQAKYTEPILPIFHAHQ
```

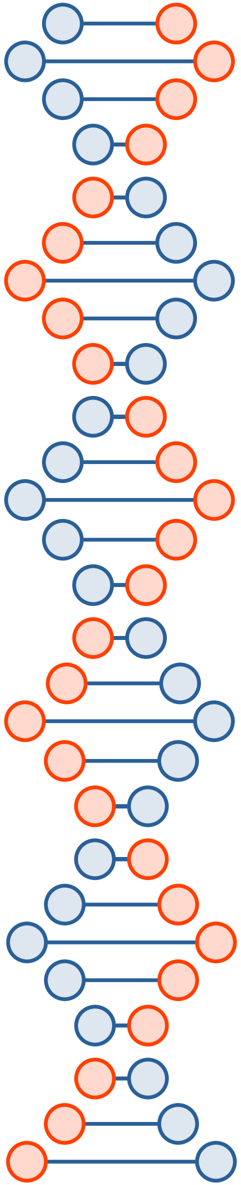


FASTQ format

- fastq format is a text-based format for storing both a biological sequence and its corresponding quality scores.

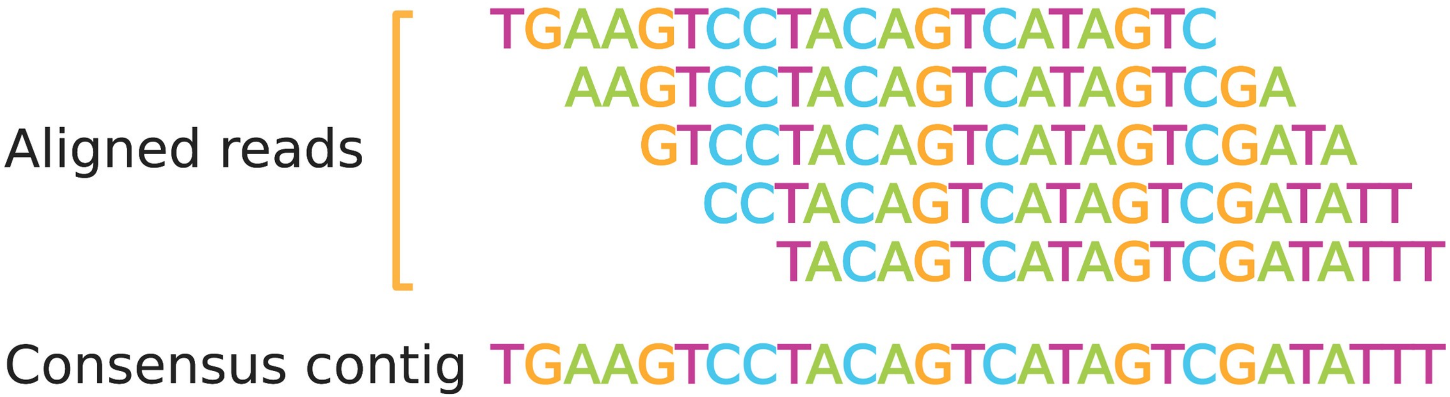
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- Line 1 begins with a ‘@’ character and is followed by a sequence identifier and an optional description.
- Line 2 is the raw sequence letters.
- Line 3 begins with a ‘+’ character and is optionally followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2



SAM & BAM file formats

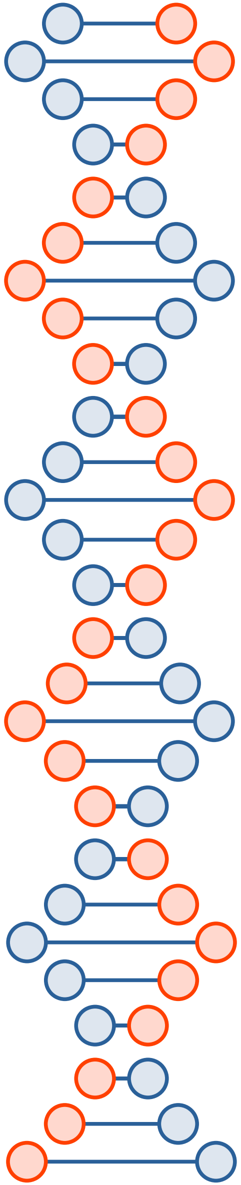
SAM files are text file format that contains the alignment information of various sequences that are mapped against reference sequences.



Aligned reads

TGAAGTCCTACAGTCATAGTC
AAGTCCTACAGTCATAGTCGA
GTCCTACAGTCATAGTCGATA
CCTACAGTCATAGTCGATATT
TACAGTCATAGTCGATATTT

Consensus contig TGAAGTCCTACAGTCATAGTCGATATTT



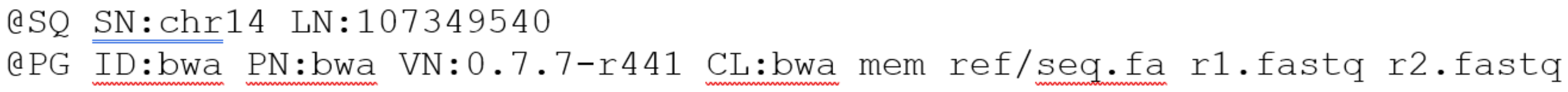
SAM & BAM file formats

BAM files contain the same information as SAM files, except they are in binary file format which is not readable by humans.

The two initial steps taken after the generation of a BAM file are to sort and then index it.

BAM files are often accompanied by a BAM index file also known as a BAI file

SAM & BAM file formats



```
@SQ SN:chr14 LN:107349540
@PG ID:bwa PN:bwa VN:0.7.7-r441 CL:bwa mem ref/seq.fa r1.fastq r2.fastq
```

The first line starts with @SQ, indicating that it is identifying a reference sequence contig.

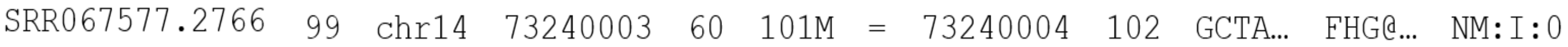
The second line starts with @PG, indicating that it describes the program used to generate the SAM file

SAM & BAM file formats

```
SRR067577.2766  99  chr14  73240003  60  101M  =  73240004  102  GCTA…  FHG@…  NM:I:0
```

SAM or BAM line

1	2	3	4	5	6	7	8	9	10	11	12
SRR067577.2766	99	chr14	73240003	60	101M	=	73240004	102	GCTA…	FHG@…	NM:I:0



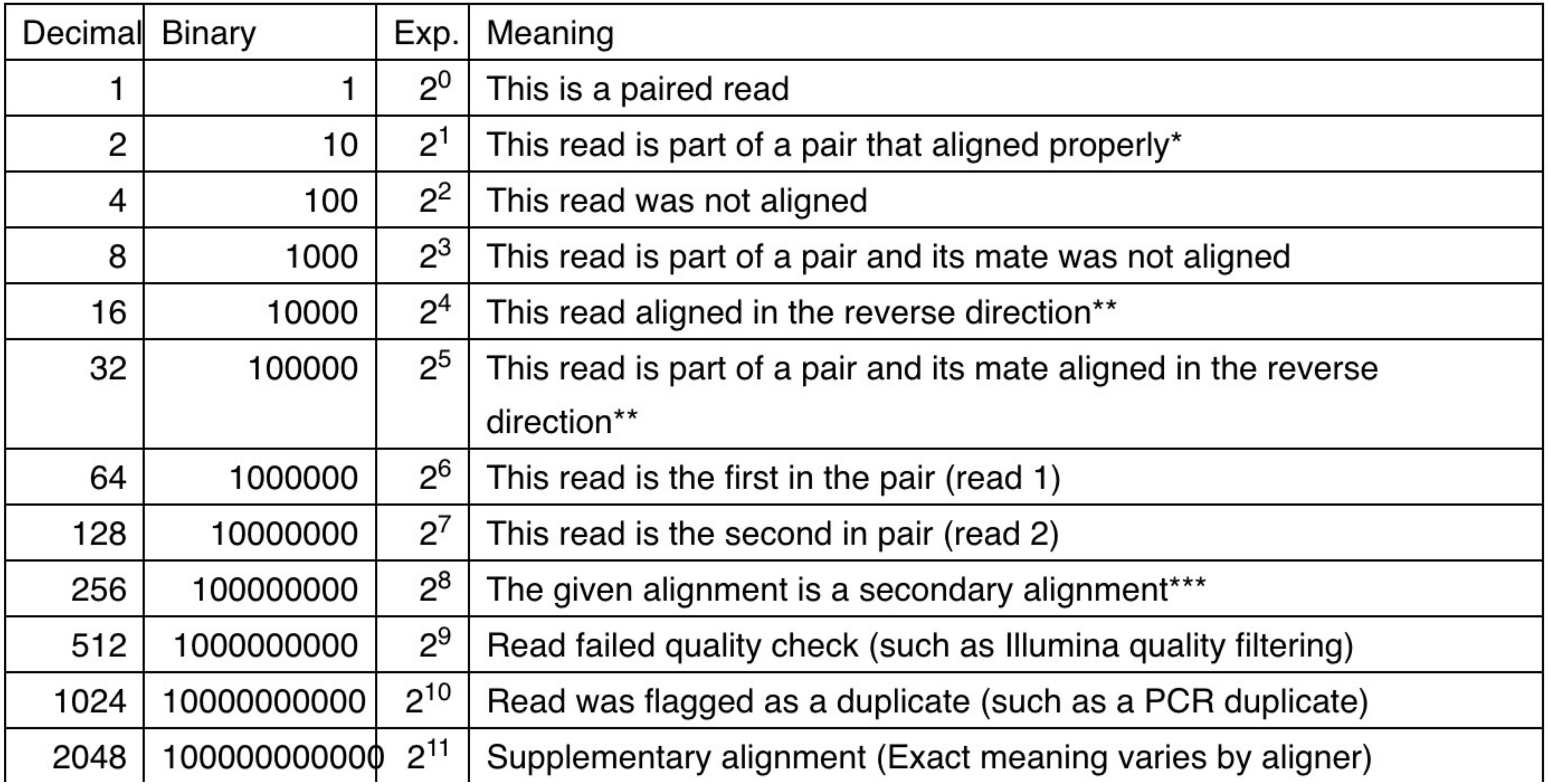
```
SRR067577.2766  99  chr14  73240003  60  101M  =  73240004  102  GCTA…  FHG@…  NM:I:0
```

1	2	3	4	5	6	7	8	9	10	11	12
SRR067577.2766	99	chr14	73240003	60	101M	=	73240004	102	GCTA…	FHG@…	NM:I:0

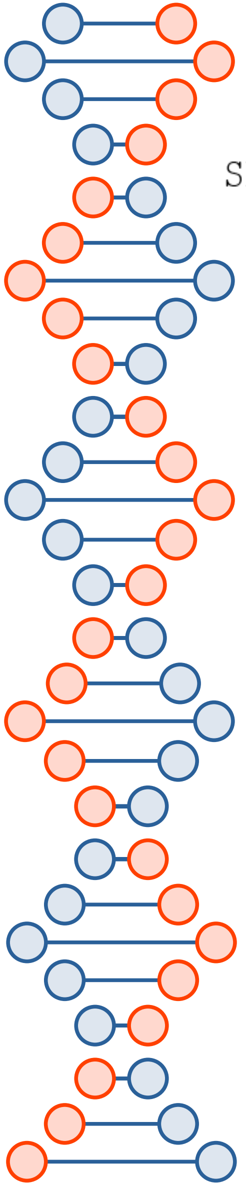
1. Query Name or QNAME is an identifier that is unique to the read within the file and can be used to identify any individual read.

2. Flag. As an example, a common FLAG value is 99, which is made up of:

$$64 + 32 + 2 + 1$$



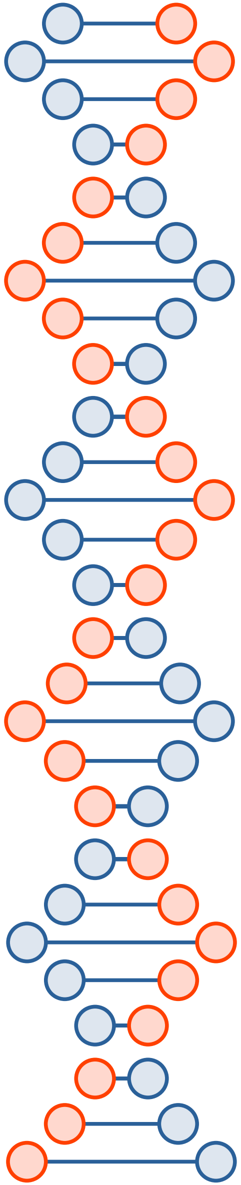
Decimal	Binary	Exp.	Meaning
1	1	2^0	This is a paired read
2	10	2^1	This read is part of a pair that aligned properly*
4	100	2^2	This read was not aligned
8	1000	2^3	This read is part of a pair and its mate was not aligned
16	10000	2^4	This read aligned in the reverse direction**
32	100000	2^5	This read is part of a pair and its mate aligned in the reverse direction**
64	1000000	2^6	This read is the first in the pair (read 1)
128	10000000	2^7	This read is the second in pair (read 2)
256	100000000	2^8	The given alignment is a secondary alignment***
512	1000000000	2^9	Read failed quality check (such as Illumina quality filtering)
1024	10000000000	2^{10}	Read was flagged as a duplicate (such as a PCR duplicate)
2048	100000000000	2^{11}	Supplementary alignment (Exact meaning varies by aligner)



```
SRR067577.2766  99  chr14  73240003  60  101M  =  73240004  102  GCTA…  FHG@…  NM:I:0
```

1	2	3	4	5	6	7	8	9	10	11	12
SRR067577.2766	99	chr14	73240003	60	101M	=	73240004	102	GCTA…	FHG@…	NM:I:0

3. Reference name
4. Position
5. Mapping quality
6. CIGAR - concise idiosyncratic gapped alignment report string



Operator	Meaning
M	Match (base may not be identical to reference, but exists in both)
D	Deletion (base exists in reference, but not read)
I	Insertion (base exists in read, but not reference)

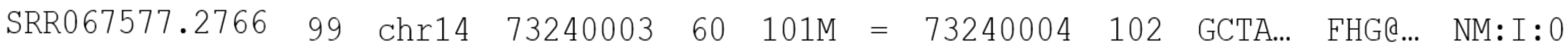
```
SRR067577.2766  99  chr14  73240003  60  101M  =  73240004  102  GCTA…  FHG@…  NM:I:0
```

1	2	3	4	5	6	7	8	9	10	11	12
SRR067577 .2766	99	chr14	73240003	60	101M	=	73240004	102	GCTA …	FHG@ …	NM:I:0

7. Reference Name for Mate. “=” if it is identical to the Reference Name value.

8. Position of Mate

9. Template length. A read with multiple insertions may have a smaller template length than the read length, while a read with multiple deletions may have a template length longer than the read length.



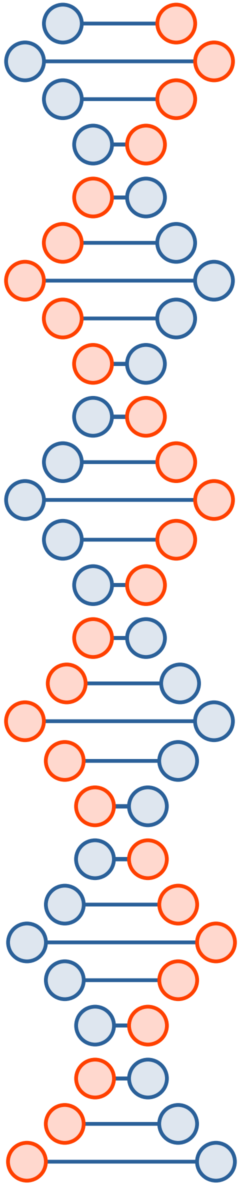
```
SRR067577.2766  99  chr14  73240003  60  101M  =  73240004  102  GCTA…  FHG@…  NM:I:0
```

1	2	3	4	5	6	7	8	9	10	11	12
SRR067577.2766	99	chr14	73240003	60	101M	=	73240004	102	GCTA …	FHG@ …	NM:I:0

10. Sequence.

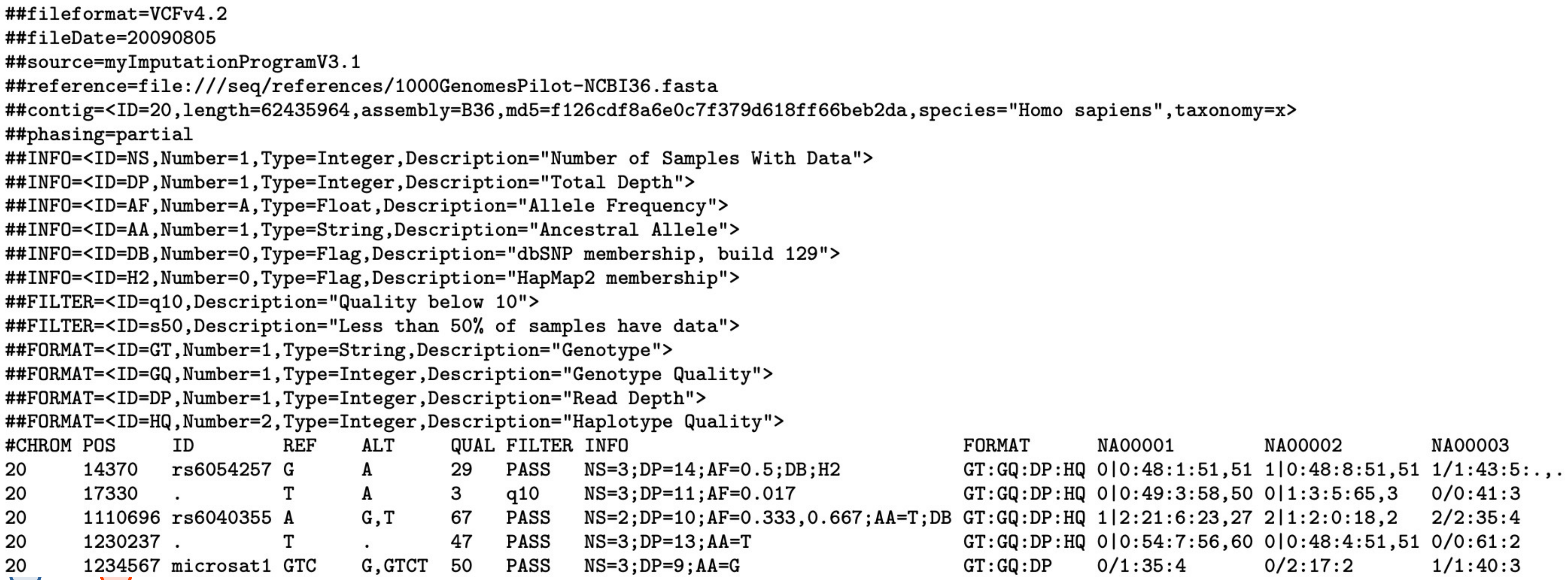
11. Quality string.

12. Reference tags. Gives additional information on the alignment or read



VCF file format

- VCF is the standard file format for storing variation data.



```
##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
```

VCF file format

- VCF is the standard file format for storing variation data.