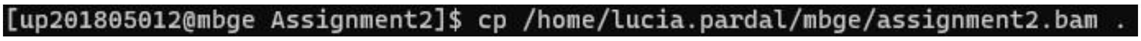
Assignment 2 – Margarida Pinheiro (201805012)

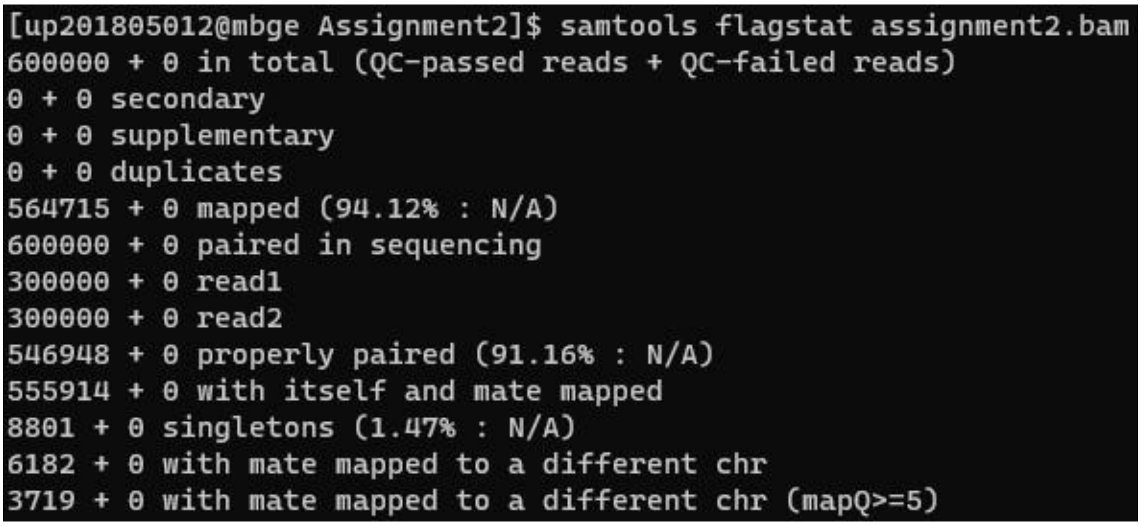
1.

Copy file **assignment2.bam** stored at /home/lucia.pardal/mbge using **cp**:



```
[up201805012@mbge Assignment2]$ cp /home/lucia.pardal/mbge/assignment2.bam .
```

Using the flagstat command from samtools in **assignment2.bam** file and its output:



```
[up201805012@mbge Assignment2]$ samtools flagstat assignment2.bam
600000 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
564715 + 0 mapped (94.12% : N/A)
600000 + 0 paired in sequencing
300000 + 0 read1
300000 + 0 read2
546948 + 0 properly paired (91.16% : N/A)
555914 + 0 with itself and mate mapped
8801 + 0 singletons (1.47% : N/A)
6182 + 0 with mate mapped to a different chr
3719 + 0 with mate mapped to a different chr (mapQ>=5)
```

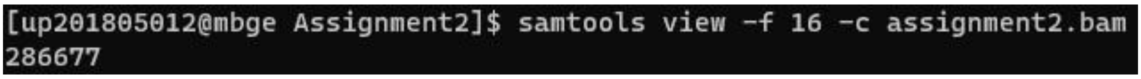
Flagstat provides counts for 13 categories based primarily on bit flags. In the first output line, we have information about the reads that pass the quality control (QC – passed) and the reads that didn't pass the quality control (QC – failed). In this case there are a total number of 600000 reads in the input file and they all passed the quality control. There are no reads marked as secondary, supplementary or duplicates. 94.12% of reads were mapped, corresponding to 564715 reads. The number of paired reads is the same as total primary reads, so we can infer that only paired reads remained after trimming. Because of that, half of the reads (300000) are forward, and the other half are reverse (300000). The percentage of properly paired reads is 91.16%, corresponding to 546948 reads with theirs mates on the same chromosome. There are 555914 reads that are mapped as well as their mate. There are 1.47% of singletons, corresponding to 8801 reads that are mapped but their corresponding reverse / forward read did not map. From this results, 6182 reads have their mate mapped to a different chromosome, but only 3719 reads have their mate mapped to a different chromosome with good quality for the alignment.

2.

a)

With **`-f 16`** we are able to select (**`-f`**) the samflag 16 to “read reverse strand”.
`-c` counts the number of samflag 16 in the file.

There are 286677 reads that align to the reverse strand in **assignment2.bam** file.

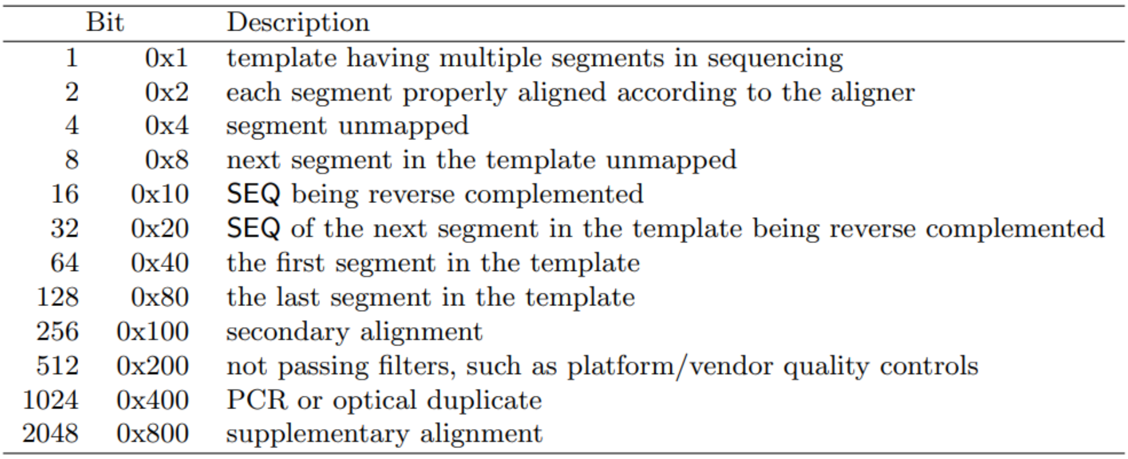


```
[up201805012@mbge Assignment2]$ samtools view -f 16 -c assignment2.bam
286677
```

b)

`-f 25` selects (**`-f`**) the reads with the properties associated with flag 25.

To understand the properties associated with flag 25, we first convert 25 into a binary number and we obtain the number 11001. Next we correspond each number of the binary format to each elementary flag retrieved from the description table from the samtools manual. In conclusion, flag 25 corresponds to the combination of flag 1, flag 8 and flag 16.

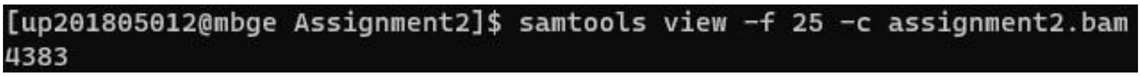


Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Figure 1. Meaning of each bit in the SAM alignment records flags field.

To conclude, **`-f 25`** selects paired reads that are unmapped and reads that are mapped on the reverse strand.

In the assignment2.bam file, there are 4383 reads with these properties.



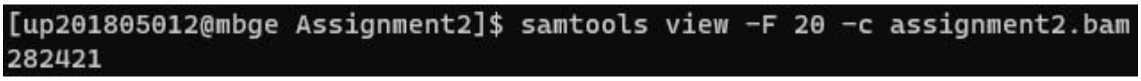
```
[up201805012@mbge Assignment2]$ samtools view -f 25 -c assignment2.bam
4383
```

c)

`-F 20` excludes (**`-F`**) the reads with the properties associated with flag 20. Flag 20 (Binary format: 10100) is the combination of the flag 4 and flag 16.

Therefore, `-F 20` excludes reads that are unmapped and reads that are mapped on the reverse strand.

In the assignment2.bam file, 282421 reads are excluded.



```
[up201805012@mbge Assignment2]$ samtools view -F 20 -c assignment2.bam
282421
```

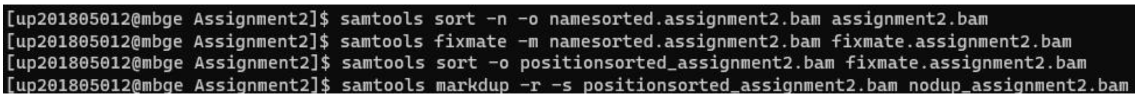
d)

The flag 1105 (Binary format: 10001010001) is the combination of the elementary flags 1, 16, 64 and 1024.

Flag 1105 specifies reads that are paired (Flag 1 (0x1)), align to the reverse strand (flag 16 (0x10)), are the first mate in the pair of reads (flag 64 (0x40)) and are PCR or optical duplicates (flag 1024 (0x400)).

3.

a)



```
[up201805012@mbge Assignment2]$ samtools sort -n -o namesorted.assignment2.bam assignment2.bam
[up201805012@mbge Assignment2]$ samtools fixmate -m namesorted.assignment2.bam fixmate.assignment2.bam
[up201805012@mbge Assignment2]$ samtools sort -o positionsorted_assignment2.bam fixmate.assignment2.bam
[up201805012@mbge Assignment2]$ samtools markdup -r -s positionsorted_assignment2.bam nodup_assignment2.bam
```

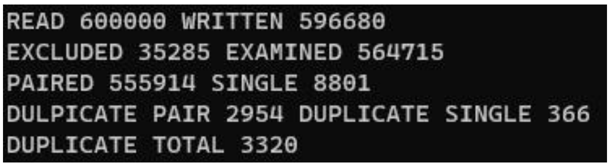
To remove duplicates from the assignment2.bam file we have to name sort before using samtools fixmate. For that we use samtools sort to order the file based on the names of the reads (`-n`) and write the final sorted output to the specified file (-`o`).

Next we use the samtools fixmate to add mate score tags (`-m`). These results will be used by markdup to select the best reads to keep.

Then we need to sort the reads based on chromosome number and coordinates (`samtools sort`) for samtools markdup.

Finally, we use samtools markdup to mark the duplicates, remove them (`-r`) and obtain a report stats (`-s`).

Output:

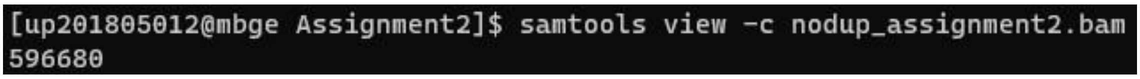


```
READ 600000 WRITTEN 596680
EXCLUDED 35285 EXAMINED 564715
PAIRED 555914 SINGLE 8801
DULPICATE PAIR 2954 DUPLICATE SINGLE 366
DUPLICATE TOTAL 3320
```

b)

The reads of the new file after the removal of the duplicates is obtain in the output of the previous exercise. This means that there are 596680 reads in the file after removing duplicates (WRITTEN 596680). The same happens to find how many reads were removed. In this case, 3320 duplicate reads were removed (DUPLICATE TOTAL 3320).

We can also use **`samtools view -c`** to count all the reads of the new file after the remove duplicates:



```
[up201805012@mbge Assignment2]$ samtools view -c nodup_assignment2.bam
596680
```

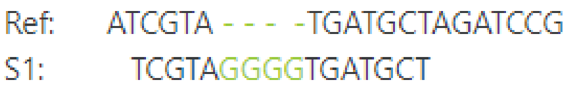
Knowing the total number of reads (600000) and the number of reads written out we find how many reads were removed: 600000-596680=3320

4.

The CIGAR string is a sequence of base lengths and the associated operation to represent alignments in SAM/BAM formats. CIGAR strings have several operators, each preceded by the number of aligned nucleotides.

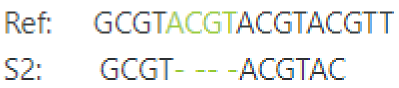
The following table presents the meaning of the operators (single character) used for this exercise:

Operation	Description
M	Match (alignment column containing two letters). This could contain different letters (mismatch) or identical letters.
D	Deletion (gap in the target sequence).
I	Insertion (gap in the reference sequence).
=	Alignment column containing identical letters.
X	Alignment column containing a mismatch.



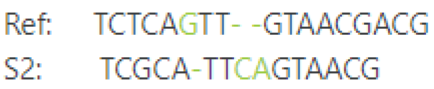
```
Ref:    ATCGTA - - -  -TGATGCTAGATCCG
S1:       TCGTAGGGGTGATGCT
```

CIGAR string S1: 5M4I7M



```
Ref:   GCGTACGTACGTACGTT
S2:     GCGT- -- -ACGTAC
```

CIGAR string S2: 4M4D6M



```
Ref:   TCTCAGTT- -GTAACGACG
S2:     TCGCA-TTCAGTAACG
```

CIGAR string S3: 5M1D2M2I6M or 2=1X2=1D2=2I6=