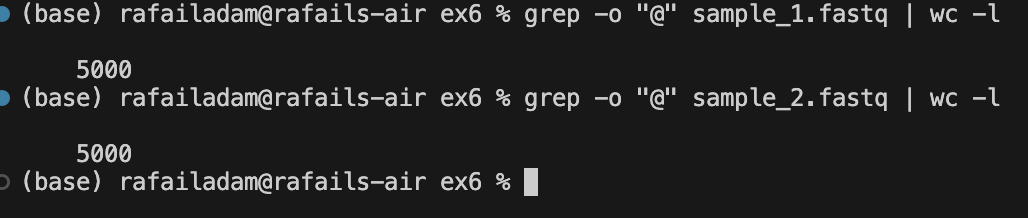
Introduction to Bioinformatics

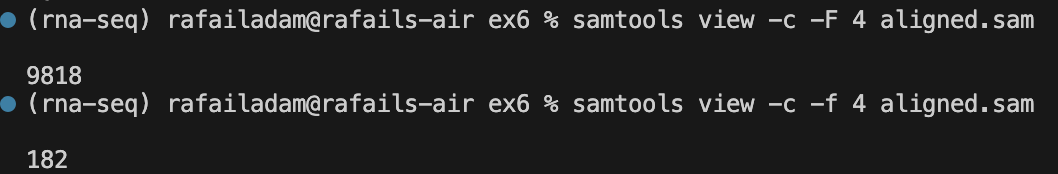
Exercise 6

Rafail Adam

Question 1



Question 2



Question 3,4

***samtools view -S -b aligned.sam > aligned.bam***

***samtools sort -o aligned.sorted.bam aligned.bam***

The sorted bam file is slightly smaller than the original bam file. The reason for that might be that due to sorting, redundant information such as the chromosome that each read is mapped to, is discarded as reads mapped to the same chromosome are grouped together leading to better compression. In the same way other information for each read might also be discarded as they are grouped together.

Question 5

***samtools index aligned.sorted.bam***

It is not possible to index an unsorted bam file because the reads are in random order and this probably creates a problem during indexing as the index file is created so as to not load all the bam file at once.

Question 6

***bedtools genomecov -d -ibam sorted.bam -g human\_g1k\_v37\_chr20 > per\_base\_coverage.txt***

Question 7

***bedtools genomecov -bg -ibam aligned.sorted.bam > coverage.bedgraph***

Question 8

***bedtools bamtofastq -i aligned.qsort.bam -fq forward.fastq -fq2 reverse.fastq***

Question 9

***bedtools slop -i TargetRegion.bed -g human\_g1k\_v37.genome -l 0 -r 100 > TargetRegion.100bp.bed***

Question 10

***bedtools sort -i TargetRegion.100bp.bed | bedtools merge > TargetRegion.100bp.merged.bed***

bedtools sort was needed first as the merge command threw an error for a specific coordinate out of order.

Question 11

***samtools view -b -L TargetRegion.100bp.merged.bed aligned.sorted.bam > aligned.filtered.bam***

***samtools view aligned.filtered.bam | wc -l***

There are only 105 reads after filtering for regions in the merged bed file

Question 12

Unique for A: 565 A.vcf.gz (43.9%)

Unique for B: 551 B.vcf.gz (43.2%)

Common for A and B: 723 A.vcf.gz (56.1%) B.vcf.gz (56.8%)