# Final Assignment

# Next Generation Sequencing

# REPORT

João Cunha nº202103227

Mestrado Bioinformática e Biologia Computacional

**Note:** Since the command lines and scripts used are included in the ".txt" file, they were not added to this report. I will only include the relevant information and comment on the output of each step.

**Execute a SNP calling using the following SRA numbers (SRR2125267, SRR2125268, SRR2125272, SRR2125297)**

First, I had to download some files, with the reference genome, the bed file, and the SRA sequences, 3 files per each different SRA (with the help of "fastq-dump" command).

After a failed attempt to create a script to automatically download the necessary sequences I decided to do this download manually, generating the following 12 new files:

Uma imagem com texto, computador, monitor, interior

Descrição gerada automaticamente

After that, the reference genome was unzipped and indexed (with 710 iterations).

Ending output of the code used:

Uma imagem com texto

Descrição gerada automaticamente

**Create and Preparate the SAM files**

Then, create SAM files from the files provided, one to one again.

After that, I got a SAM file for each of the SRA numbers, meaning 4 new files were created the ".sam" files.

Ending outputs of each code used:

Uma imagem com texto, computador, interior, portátil

Descrição gerada automaticamenteUma imagem com texto, computador, interior, portátil

Descrição gerada automaticamente

Uma imagem com texto, computador, portátil, interior

Descrição gerada automaticamente

Uma imagem com texto, computador, captura de ecrã

Descrição gerada automaticamente

After having the SAM files created, I developed a script for each SRR sequence (file\_67.sh, file\_68.sh, file\_72.sh, file\_97.sh). These scripts transform them into BAM files, also remove duplicated ones.

For each SAM file created:

* I used the bed file to call only the variants included in the file;
* Transformed them into BAM;
* Fill in mate coordinates;
* Obtained the fixmate information;
* Marked the duplicates and removed them;
* Used a thread value of 3, for all files.

Note: After creating each script I had to make them executable before executing each one.

Outputs of each code used:

Uma imagem com texto, computador, portátil, interior

Descrição gerada automaticamenteUma imagem com texto, computador, interior, portátil

Descrição gerada automaticamente

Uma imagem com texto, computador, portátil, interior

Descrição gerada automaticamente

Uma imagem com texto, computador, interior, portátil

Descrição gerada automaticamente

After generating several files and, I will mention the most important ones, 4 files "\*\_final.bam":

* SRR2125267\_final.bam;
* SRR2125268\_final.bam;
* SRR2125272\_final.bam;
* SRR2125297\_final.bam.

From these bam files we can see some information, like:

* Count all the reads before removing duplicates;
* Count all the reads after removing duplicates;
* Count all the reads mapped before removing duplicates;
* Count all the reads mapped after removing duplicates.

**SNP Calling**

After having the bam files, I had all the alignments and base calls ready to start the SNP calling. This step generates a "final.vcf" file.

Output of each codes used:

Uma imagem com texto, computador, captura de ecrã, portátil

Descrição gerada automaticamente

**MAF filter of 0.05**

After the SNP calling, I applied the MAF filter of 0.05. This step only included sites with a Minor Allele Frequency greater than or equal to 0.05 (avoiding alleles that are in low frequency).

Also generated a final file: "analysis\_maf.record.vcf."

Output of each code used:

Uma imagem com texto, captura de ecrã, computador, monitor

Descrição gerada automaticamente

**Remove SAM Files**

Finally I remove all the files that we are not interested in (SAM files).