LAB 3: manipulating information from SAM files

ASSIGNMENTS

First, take a look at the **LAB3_Tips** file and practice with each section described. We strongly believe that they could be helpful with the following assignments.

Assignment 1: search for SNP and insertions/deletions in a sample

In order to obtain biological information on SNPs, insertions and deletions, manipulate SAM files in the following way:

1. Convert SAM file obtained in LAB2 to BAM
2. Sort BAM to make the following process faster.
3. Use bcftools to obtain a VCF file
4. Write a Python program to parse the VCF file and obtain only Single Nucleotides Polymorphism (SNP) for which the information is complete.

e.g.

```
10          48642          .          C          T,<*>          0          .
DP=2;I16=0,0,1,0,0,0,70,4900,0,0,60,3600,0,0,6,36;QS=0,1,0;SGB=-
0.379885;MQ0F=0     PL     60,3,0,60,3,60
```

5. Write a Python program to parse VCF file and obtain only Insertions or Deletions (INDEL) for which the information is complete

e.g.

```
10     3643016     .     GCTC     G     0     .
INDEL;IDV=2;IMF=1;DP=2;I16=0,0,1,1,0,0,92,5800,0,0,14,98,0,0,50,1250;QS
=0,1;VDB=0.06;SGB=-0.453602;MQSB=1;MQ0F=0     PL     14,6,0
```

Assignment 2: Raw read count for each protein coding gene in chr10 and chr18

With this assignment we want to calculate how many reads of our SAM file have been mapped on the protein coding genes of chromosome 10 and 18. First of all we have to obtain names and genomic positions of each protein coding gene from chr10 and chr18.

- Write a python program that parses the **gtf** file and extract information about chromosome 10 and 18 (means first column equal to 10 or 18). Select only rows for which the feature is equal to “gene” and gene_biotype equal to “protein_coding”. Alternatively, you can perform this step using awk bash command (see awk documentation online https://www.gnu.org/software/gawk/manual/gawk.html#Very-Simple).
- Use `samtools view` to filter unmapped reads and supplementary alignments from the SAM file you obtained in LAB2. Save your results into a new SAM file (e.g. unique_aligned.sam)
- Using information from the reduced SAM and gtf files, write a Python script to calculate the raw read count for each protein coding gene in chr10 and chr18 (since the original files we provided you were truncated for computational feasibility, you will find a lot of genes with read count equal to zero).