BI-GY 7453: Prof. Manpreet S Katari
April 11, 2022

ASSIGNMENT 4 : Due Date : 20-April-2022

Question 1. Next-generation sequencing technologies are making a large impact in Biology today. The ability to sequence large genomes and RNA molecules from cells in different conditions has allowed biologists to propose many different experiments. At the root of all the analysis, the sequences have to be first aligned to the genome so you know where the sequence comes from.

For this assignment, you will parse the results of an alignment (called a SAM file) and compare it to a file that contains gene annotations (GTF) . The documentation for the SAM file can be found here. (https://samtools.github.io/hts-specs/SAMv1.pdf). The name of the sam file is sample.sam. The two columns of interest for our purpose are column 3, the name of the chromosome, and column 4, the start position of the alignment. For simplicity of the homework, we are not worried about the strand and the length of the match.
The File will look something like this after pre processing :

index	chrom	pos
26	Chr3	14202145
27	Chr2	8023
28	Chr3	14201994
29	Chr2	5881
30	Chr3	14199852
31	chloroplast	133584

First, split all the alignments based on chromosomes. Simply store the start positions for every chromosome. For example, if you are using python, you can create a dictionary where the chromosome is the key and the value is a list of start positions. Note that some lines do not have a value for chromosomes which means that sequence did not align to the genome, so we can ignore them.

1) **Implement without sorting the sam file** : The GTF file (Arabidopsis.gtf) is a tab-separated file already sorted. It contains the chromosome in the first column and the start and end positions of an exon in the 4th and 5th column, and the gene name in the last column. To determine the number of reads that match a gene, count the number of start positions from the SAM file that fall between the start and stop coordinates of an exon that belongs to a gene.
2) **Implement after sorting the sam file** : For each chromosome sort the coordinates. You can use the quicksort function we worked on in class. Now the GTF file (Arabidopsis.gtf) is a tab-separated file already sorted. It contains the chromosome in the first column and the start and end positions of an exon in the 4th and 5th column, and the gene name in the last column. To determine the number of reads that match a gene, count the number of start positions from the SAM file that fall between the start and stop coordinates of an exon that belongs to a gene. ***Note** - when you are counting the number of sorted coordinates located inside a gene boundary, you don't have to continue looking if the coordinate matches past your gene end coordinate. This should speed up your search significantly.*

In both the above cases calculate the total time taken with sorting and without sorting the coordinates from the sam file and clearly explain why sorting reduce/increase the total time for execution?
You can use the below mentioned code to calculate the total time taken to execute a function.

```
# in Python
import time
start = time.process_time()
# your code here
print(time.process_time() - start)
```

```
# in R
start_time <- Sys.time()
# your code here
print(Sys.time() - start)
```