CAMM535 Assignment 3

Due date: January 5, 2023

Late policy: For each assignment, 30 points deduction will be applied for one day late, and 10 points additional deduction for each extra day.

Remind that the instructor has the right to request a demo of your assignments at any point and can determine the final score based on the demo performance.

Support your findings with screenshots whenever possible.

1. Select a human gene from the given list and search for it using Ensembl. [CREM, BMF, ALB, BAX, CRX]

    a. How many transcripts does this gene have? How many of them are protein-coding? What do the colors indicate regarding the transcript biotypes?

    b. Select one of the golden protein-coding transcripts. How many variant alleles are associated with it? How many of these variants belong to indel class?

    c. How many phenotypes, diseases, and traits are associated with this gene? Select a phenotype source and write a very short information paragraph for the phenotype using the text provided in the source. (E.g. Disease definition / summary / description)

    d. Which GO: Biological processes are associated with the selected gene? Select one of them, write its definition and click Search Biomart link to continue. How many unique genes are present in the Biomart table?

2. Find the genomic region for the human NUMB gene.

    a. What information about NUMB gene are you able to retrieve at UCSC? Write a summary.

    b. Extend the region in the view by adding 10000 bases to each end of the position in the window. Show the new range.

    c. Which transcript variant is highlighted as the canonical transcript of the gene? Select a random exon of the canonical transcript and show its position and size.

d. Determine if structural variation has been indicated in this genomic region by visualizing the copy number variation (CNV) data from the DGV database track. What do the colors tell about the variants? How many variants are there for the extended region in part B ?

e. Export the genome viewer image and add it to your report.

3. SLCO1B1 encodes an organic anion transporter responsible for the disposal of not only methotrexate but many other drugs. There are several SNPs in the SLCO1B1. One of those SNPs is rs11045818. View the rs11045818 SNP in the UCSC browser window.

a. Consider the exon with rs11045818. Select to display the SNP information in dbSNP (all SNPs). How many synonymous and non-synonymous SNPs does the exon have?

b. To which nucleotide and amino acid substitution does rs11045818 correspond?

4. Consider hg19.refGene table at UCSC Table Browser (hg19 represents human).

a. What does the SQL query below search for?

```
SELECT *
FROM hg19.refGene
WHERE exonCount <= 100 AND chrom!=1;
```

b. Use the UCSC Table Browser interface to find the results of this query. How many items did you get from the result statistics?

c. Select one of the transcripts in the resulting table. Visualize it in the Genome viewer at UCSC and add the tracks requested below. Save the view in publication quality. Add the link to visualize the view by other people.

d. Show all SNPs located in the selected transcript.

e. Add UniProt/SwissProt Amino Acid Substitutions. Which amino acid substitutions are present in the selected transcript?

5. Answer the questions below using UCSC table browser (hg38).

a. How many simple repeats are there in the human genome? (check "item count")?

b. How many simple repeats with the sequence CAG or ATG are there in the human genome?

c. How many CG repeats in chromosome 1 and AG repeats in chromosome 2 are there in the human genome?

d. How would you write the SQL queries to ask for these questions above?