BIN508 Assignment 3

Due date: 21 April, Monday, 13:30 **Cut-off:** 21 April, Monday, 19:30

Late policy: After the due time, 5 points of deduction will be applied for each extra hour. No submission is allowed after 19:30.

- **Use EU servers instead of USA servers for this assignment.**
- **You need to install IGV on your local machine.**
- **If you plan to use Linux/Unix, you need to download the mouse reference genome, unzip and index it with `bowtie2-build` first.**

  ```
  bowtie2-build filename_of_the_reference.fa index_name_of_your_choice
  ```
- **Use Galaxy for the second part (not obligatory but highly suggested).**
- When going over the tutorials, remember that the "Hands-on" parts are the ones you need to complete, unless specified otherwise.
- **Make your Galaxy history and steps visible, and add a link to your Galaxy history at the top of the assignment. If you use a Linux environment, add your script to the end of the assignment with a mono-space font and include comments.**

Useful links: YouTube channel of IGV

1) Complete this tutorial (excluding JBrowse part). Add a link to your Galaxy History.
   - Summary of the steps: Load the data > Align with Bowtie2 > Inspect with Samtools Stats > Visualise with IGV (Reference genome: Mouse mm10)
   a) Add a screenshot showing the **samtools stats** output and comment on the results.
   b) Add a screenshot of the IGV window with the position given in the tutorial.
   c) Select a read within the BAM file and add a screenshot of it. Find the same read on IGV and click on it. Add a screenshot of both the IGV window and the details window of that read in IGV.
   d) Find the position in the reference genome with the maximum amount of aligned reads (like in the previous assignment). Reveal the details of that position by clicking on the bar (that represents the depth) on top. Share a screenshot of the IGV window showing these details and comment on them.
   e) Show your workflow with a diagram.
2) Complete this tutorial. (Variant Analysis) Add a link to your Galaxy History.
   a) Answer the questions below the "Querying genotypes" title, using screenshots.

- GEMINI is a tool that creates a “database” from the input you provide. As it is a database, SQL queries are used for pulling information.

b) Visualise the mapped reads and the VCF output of FreeBayes on IGV together (Make sure you have selected Human hg19 as your reference genome on the top-left corner of the IGV window first). For this, you need to download the BAM and VCF files and load them into IGV (These files are provided to you with the assignment for ease).
   i) Colour the reads to identify forward and reverse reads. Share a screenshot of your IGV window. Make sure the position on the chromosome is visible.
   ii) Colour the reads by the samples (mother, father, son).
   iii) Select a variant from the VCF and add a screenshot of it (like 1c). Now find it on IGV and zoom into it. What is the nucleotide on the reference genome at that location? Click on the variant to see the details for the mother, the father, and the son. Compared to the reference genome, who carries which variant? Share (a) screenshot(s) of the window containing the information.
   iv) Click on the coverage bar on the variant's location (coloured bar on top of the reads, shown in the BAM part). What is the percentage of each nucleotide amongst the reads on that location? Share a screenshot of the window containing the information.
   v) Click on one of the reads on the variant’s location. Whose sample does this read belong to? What else can you tell about this read looking at the window? Explain and share a screenshot.

c) SnpEff is going to provide a detailed report. Please go over the report by providing screenshots and very short comments.

d) Show your workflow with a diagram.

3) Briefly answer the following questions.
   a) What is the difference between SNV and SNP?
   b) Please define “haplotype”. If there are 2 SNPs close to each other and SNP 1 has two possible alleles, A and T, and SNP 2 has two possible alleles, C and G, what are the possible haplotypes?
   c) What are heterozygous and homozygous variants?
   d) Suppose I have 100 reads aligned to a position on the reference genome. I am investigating whether the individual from whom I obtained the DNA harbours a genetic variation at this specific region, which is known to contain a single nucleotide polymorphism (SNP).
      i) 80 of the reads are reverse reads.

ii) 50 of the reads carry the reference nucleotide on the position of the SNP, while the other 50 carry the alternative nucleotide.

iii) All of the alternative nucleotides are observed in the reverse reads.

Can we say that the individual carries a heterozygous variant in this position? Why?