BIOL0050 – Advanced Computational Biology – Assignment 1

Answers due 11:59 pm, Wednesday 5th February 2020

This work is assessed, and this and all other assignments must be attempted to pass the course. Submit your answers via Moodle Turnitin by **11:59 pm** on **Wednesday 5th February**. You should show your working and explain the steps you took to reach your conclusions wherever possible. Please include all code used to answer the questions at the end of the file.

Question 1 ***(60 points)***

The file "snps.RData" contains the genotype information for 53 individuals at selected positions in the genome. The file "mat.gtex.RData" contains the normalized expression values for some genes of interest measured from blood samples from these individuals. Load these two data structures in R using the command ***load***.

(a) Employ these data and simple linear regression models to identify potential eQTLs, i.e. SNPs influencing the expression levels of the given genes (**15 points**). Make sure you perform all major quality control steps that would be employed in an association study and that are possible given the data (**15 points**). Using a cut-off of 10^{-8}, can you identify potential eQTLs for all the genes? Illustrate the identified associations using suitable graphs (**5 points** for using base R; **10 points for using** ggplot2-based graphics).

(b) Create a new covariate representing the tissue where the expression measurement of PIK3CA was performed for each individual. To do so, choose randomly between the following tissues: lung, pancreas, colon, oesophagus, brain, kidney. Cases where PIK3CA expression is lower than 6 should only be found in pancreas, and cases with PIK3CA expression greater than 10 only in brain. Build a linear regression and a mixed effect model taking into account the tissue of origin. How do the two models compare to your previous findings? (**20 points**)

Question 2 ***(20 points)***

Download and read the "gwas.als.RData" file into R using the ***load*** command. This file contains the results of a GWAS study on loci conferring risk for the development of Amyotrophic Lateral Sclerosis (ALS).

(a) How many significant SNPs are there and which chromosomes are they located on? (**1 points**)

(b) Generate Manhattan plots for chromosomes 9 and 17 using base R (**3 points/plot**). Also generate a single Manhattan plot for all the data using the qqman package (**3 points**). What do you observe? Is there anything particularly striking? How can you explain these observations? (**10 points for interpretation**)

Question 3 ***(20 points)***

Using the data from Question 2, select the top most significant SNP from chrosomosome 9 and the top most significant SNP from chromosome 17. Using GTEx (https://gtexportal.org/home/), search for single-tissue eQTLs of these SNPs. Download the results in csv format.

(a) Which genes are affected and in which tissues? Use a cut-off of 10^{-8}. (**4 points**)

(b) How many *cis* and how many *trans* eQTLs do you find in each case (explain how you derived this)? (**6 points**)

(c) Comment on whether these genes could be involved in the development of ALS and how that might look like at a functional level, providing further support from scientific literature or other sources. (**10 points**)