

Assignment 2 [29 marks]

Due before 12pm, Tuesday 24th August, 2021

Your answers to all questions should be submitted to myUni as a `.zip` file containing three files: 1) a bash script for Q1 2) a bash script for Q2, and 3) the answers to the statistics questions (Q3 and Q4) in a single Rmarkdown document. Note that the file `my_species_gff_features.txt` is not required as part of your submission for Q1, **only the script which will generate this file!** Similarly, for Q2, only the script is required.

Required scripts [14 marks]

1. Write a script to:
 - Download the gff3 file for your assigned species ([see bottom of page](#)) to your current directory from Ensembl [1 mark]
 - Count how many of each feature type there is, sorted in numerical order [3 marks]
 - Export the results to a file with a name of the form `my_species_gff_features.txt` **where you use your assigned species name instead of `my_species`** [1 mark]. NB: If your actual species is not included in the name, no marks will be given.
 - Include one or more comment lines before the table detailing which build of the genome was used, and the code executed to generate the summary [2 marks]
2. For the file we used in the practicals (`Drosophila_melanogaster.BDGP6.ncrna.fa`), add to the final practical script provided so that:
 - the output contains a meaningful header [1 mark]
 - the output contains column names [2 marks]
 - the output includes: a) gene id; b) chromosome; c) start; d) stop; e) strand and f) gene_biotype [3 marks]
 - Appropriate comments which make the script easier to understand [1 mark]

NB: If identical comments are identified in any submissions, a mark of zero will be given for this question for all suspicious submissions.

Statistics questions [15 marks]

In a single markdown file answer the following questions:

3 Two groups of people have volunteered to take part in a genetic study. Group 1 ($n = 126$) are volunteers with no history of Type I Diabetes in their immediate family, whilst Group 2 ($n = 183$) have all been diagnosed with Type I Diabetes. A genotyping study was undertaken on these volunteers using 25,786 SNPs selected due to their proximity to key immune genes. Researchers are looking to identify any SNP genotypes which may increase the risk of Type I Diabetes. In your answer, consider the reference SNP allele as **A** and the alternate SNP allele as **B**, using the genotypes **AA**, **AB** and **BB**.

- For an individual SNP, what test would be appropriate for this comparison? [1 mark]
- Define H_0 and H_a for the genotype at each individual SNP. [2 marks]
- If there was no true difference in any genotypes between the two groups, how many p-values would you expect to see < 0.05 ? [1 mark]
- Using Bonferroni's method, what would a suitable cutoff value be to consider a SNP as being associated with an increased risk of Type I diabetes, i.e., to reject H_0 [1 mark]
- Given the following genotype table, would you reject or fail to reject H_0 ? Provide your working and a full explanation. [3 marks]

Group	AA	AB	BB
Control	25	60	41
T1D	21	55	103

4 An experiment was repeated multiple times, in which GFP fluorescence was measured in a cell culture as a measurement of gene expression, both *before* and *after* viral transfection. GFP was present on a plasmid as a reporter for activity at a specific promoter. The change in fluorescence values obtained for each repeat are given below as the vector **x**, presented on the log2 scale for your individual subset of experiments.

- Define H_0 and H_a [2 marks]
- Calculate the sample mean and sample variance in **R** [2 marks]
- Calculate the T -statistic using **R**. [1 mark]
- What would the degrees of freedom be for your t -test? [1 mark]
- Calculate the p -value using **R** [1 mark]

Show all working & code.

Species For Question 1

If your student number is not listed, please contact Dave to ensure you are added to the list

You can download your assigned species here: '<http://ftp.ensembl.org/pub/release-100/gff3/>' of course you will have to add the relevant additional information to specify your species and the '.100.gff3.gz' file.

ID	Species	Taxonomy ID	Common Name
a1705481	Larimichthys crocea	215358	Large Yellow Croaker
a1707609	Fukomys damarensis	885580	Damara Mole-Rat
a1727718	Lepidothrix coronata	321398	Blue-Crowned Manakin
a1734633	mus musculus akrj	10090	
a1743091	panthera tigris altaica	9694	Amur Tiger
a1747876	Geospiza fortis	48883	Medium Ground-Finch
a1767956	Pan paniscus	9597	Pygmy Chimpanzee
a1770858	Pogona vitticeps	103695	Central Bearded Dragon
a1773581	Clupea harengus	7950	Atlantic Herring
a1773594	Mola mola	94237	Ocean Sunfish
a1777472	Oryzias javanicus	123683	Javanese Ricefish
a1778718	Otolemur garnettii	30611	Small-Eared Galago
a1780328	sus scrofa usmarc	9823	
a1828691	Fundulus heteroclitus	8078	Mummichog
a1828993	Serinus canaria	9135	Common Canary
a1835622	Poecilia mexicana	48701	
a1837876	chrysemys picta bellii	8479	Western Painted Turtle
a1841011	Macaca fascicularis	9541	Crab-Eating Macaque
a1843320	terrapene carolina triunguis	158814	Three-Toed Box Turtle

Values For Question 4

If your student number is not listed, please contact Dave to ensure you are added to the list

The results you are analysing for Q4 are as follows. You can simply paste these values into

your RMarkdown document as the object `x` and perform all of your analysis on these values.

ID	Values
a1705481	<code>x <- c(2.2889, 2.1635, 1.3856, -3.6268, 2.3854, 1.599, 0.1233, -0.8989, -0.4923, 0.0417)</code>
a1707609	<code>x <- c(0.4166, 2.1107, -0.3169, 0.5991, 0.3768, -1.5983, -0.8858)</code>
a1727718	<code>x <- c(2.239, -0.4818, -1.7278, 1.3568, -1.387, 4.3296, 1.1261)</code>
a1734633	<code>x <- c(1.1951, 1.9995, 1.6065, 1.2997, 3.4331)</code>
a1743091	<code>x <- c(2.3628, 1.568, 2.6526, 1.1039, 0.7289, 2.0428, -0.4728, -3.5131)</code>
a1747876	<code>x <- c(-2.3702, -0.8255, 2.4861, -1.0819, -0.739, -0.17)</code>
a1767956	<code>x <- c(-1.303, 0.1688, 4.0797, 2.5159, 0.3931, -0.87)</code>
a1770858	<code>x <- c(0.3317, 0.3423, 0.7688, 1.1354, 2.4904)</code>
a1773581	<code>x <- c(-1.6214, -1.4619, 0.6893, -0.5449, 0.6709)</code>
a1773594	<code>x <- c(0.2608, -1.4151, 2.8979, 1.4756, 2.2516)</code>
a1777472	<code>x <- c(-0.5246, 0.3854, -1.0757, 3.222, 3.5642)</code>
a1778718	<code>x <- c(0.7971, -0.822, -3.8884, -1.4897, 0.3142, 3.4211, 0.3878, 2.1248)</code>
a1780328	<code>x <- c(-1.2165, 0.1543, -0.6167, 2.5594, -0.4585, 2.5459, 0.7718)</code>
a1828691	<code>x <- c(2.0259, -0.57, 0.4786, -0.146, 1.4909, 1.3815, -0.9265, 0.9515)</code>
a1828993	<code>x <- c(1.9851, -3.2159, -0.0196, 0.2871, 0.3485)</code>
a1835622	<code>x <- c(1.9688, -1.0354, -0.6331, 1.1205, 1.3101, 0.7891, 0.401, 1.3373, -1.4588)</code>
a1837876	<code>x <- c(-0.4005, -1.0709, 0.9511, 1.0804)</code>
a1841011	<code>x <- c(-0.2152, 1.4955, -1.1043, 1.1306, -1.8557, 0.6466, 1.1056, 1.0517)</code>
a1843320	<code>x <- c(2.0329, 0.8256, 0.8852, -1.3689, -1.9406)</code>