HW #5 due Thursday, April 29
(Note: the HW has 4 pages)

1. For this problem we are going to explore the GWAS Catalog.
Go to: https://www.ebi.ac.uk/gwas/
In the search window type, "type i diabetes mellitus". Click on the first result (it might take a minute to update).

   a. (1 pt) How many associations?
   b. (1 pt) How many studies?

   In Associations, click on the up arrow next to "P-value".

   c. (1 pt) What is the p-value of the variant with the smallest p-value?
   d. (1 pt) What is the risk allele frequency (RAF) of this variant?
   e. (1 pt) What is the odds ratio (OR) of this variant?
   f. (1 pt) Which chromosome is this variant on?

   Click on the "study accession" value for this variant.

   g. (1 pt) How many cases and controls and what population was used for the GWAS study that identified this variant?

   Click on the web browser back arrow. Back in Associations, click on the down arrow next to odds ratio.

   h. (1 pt) What is the p-value of the variant with the largest odds ratio?
   i. (1 pt) What is the risk allele frequency (RAF) of this variant?
   j. (1 pt) What is the odds ratio (OR) of this variant?
   k. (1 pt) Which chromosome is this variant on?

   Click on the "study accession" value for this variant.

   l. (1 pt) How many cases and controls and what population was used for the GWAS study that identified this variant?

   m. (3 pts) Are you surprised that the variant with the smallest p-value is not the variant with the largest odds ratio? Discuss.

2. For the next two problems we are going to use the R code in GWAS lecture 4 (lecture notes on Blackboard, you can just copy and past the commands and then change the numbers for the matrix m and the inflation factor λ). If you do not have R already installed on your computer, use the website mentioned in lecture: https://rdrr.io/snippets/

   For each of the two tables below compute the p-value for the chi-squared test, the p-value for the Cochran-Armitage trend test (with genomic control inflation factor $\lambda = 1$, so no adjustment), the odds ratio, and the 95% confidence interval for the odds ratio.

   a. (5 pts)

	AA	AC	CC
Cases	400	500	100
Controls	540	400	60

   b. (5 pts)

	TT	TG	GG
Cases	95	800	105
Controls	260	490	250

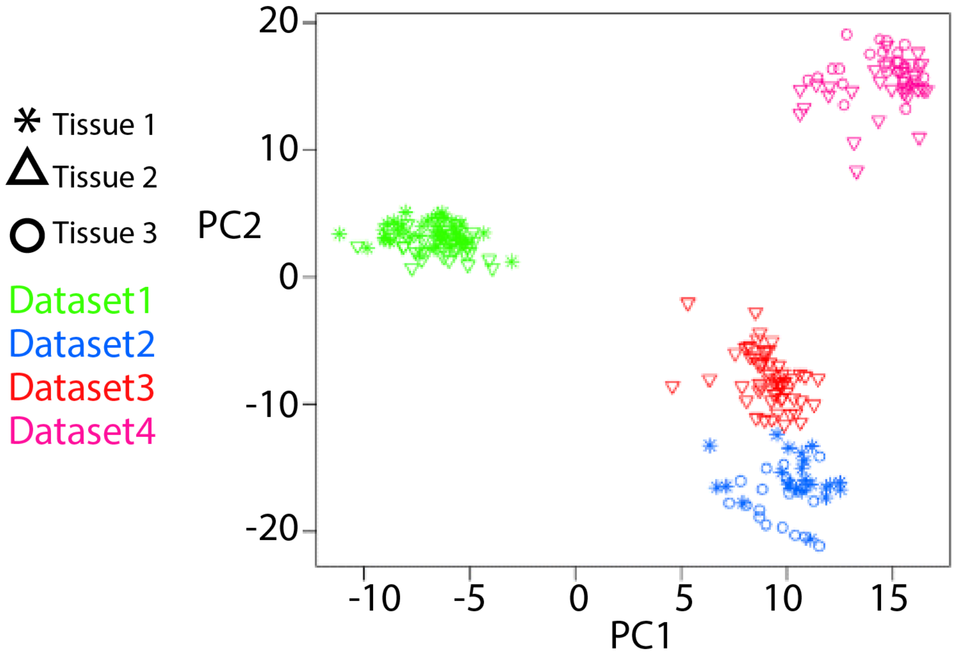
   c. (5 pts) Why do the two p-values roughly agree for (a) but radically disagree for (b)? Discuss.

3. For the data in (2a), compute the p-value for the Cochran-Armitage trend test for the following genomic control inflation factors:

   a. (1 pt) $\lambda = 1$ (no adjustment)
   b. (1 pt) $\lambda = 1.05$
   c. (1 pt) $\lambda = 1.10$
   d. (1 pt) $\lambda = 1.15$
   e. (1 pt) As λ increases, do p-values increase or decrease?

4. (5 pts) In your own words, explain one possible reason for the missing heritability in many GWAS studies.

5. A researcher is interested in studying differential expression (DE) genes in three tissues (Tissue1, Tissue2, and Tissue3) and generated four different RNA-seq datasets (Dataset 1-4). Before carrying out the differential expression analysis, PCA analysis was performed using normalized counts. The plot based on the first two PCs is shown below.



Answer the following questions based on the PCA plot.

a. (1 pt) Which component (PC1/PC2) has better separation of Dataset4 from Dataset2?
b. (2 pts) The differences between different tissues are more pronounced than the differences between different types of datasets (True/False). Explain.
c. (2 pts) If the researcher carries out DE analysis between tissue types using the normalized counts used for generating the PCA plots as it is (without any additional correction), meaningful results cannot be obtained (True/False). Explain.

6. Go to the UCSC genome browser (http://genome.ucsc.edu) and select the Human hg38 genome assembly. In the search box, type the gene name "HBB" and select "HBB" gene. Use the GENCODE gene track to answer the following questions:

    a. (1 pt) Enable dbSNP 153 track under Variation. Using dbSNP153 track, find how many SNPs are reported in the gene body of this gene?
    b. (1 pt) How many SNPs are in exon 1? What are the ids of SNPs present in exon 1?

Enable the "OMIM Alleles" track under "Phenotypes and Literature". OMIM can be used to find information about disease phenotypes associated with this gene. Make sure to refresh after enabling the track. Change OMIM Alleles to pack. Click on OMIM Allelic Variant 141900.0401. It will take you to another page.

c. (1 pt) What is the amino acid substitution for this variant?
d. (1 pt) Is this a synonymous SNP or non-synonymous SNP?
e. (1 pt) What two medical conditions are reported in the literature for the patient with this variant? Find the information in the 141900.0401 link.