HW2: eQTL, Polygenic Risk Score, RNAseq, and CRISPR

Due: Wednesday, October 7th, 2020 at 5pm EST.

Submission Guidelines

Please read the instructions carefully! Name all files using the following convention:

- lastName_firstName_shortDescriptionOfFile.extension.

You must submit **one** pdf file (with all solutions to questions) and **one** file containing your code. If using R this can either mean including:

- one pdf file and one .R file **or**
- one pdf file and one .Rmd (R markdown) file.

1 eQTL

Download the data set **RMA_Dataset.txt**, which contains Affymetrix expression data for 8 European and 8 African-American individuals. The header contains the ProbeSet ID (each probeset corresponds to a different gene) and sample IDs. For each individual, replicate arrays were performed. Thus, CEU_1_1 and CEU_1_2 correspond to array 1 and array 2 for the first CEU individual, CEU_2_1 and CEU_2_2 correspond to array 1 and array 2 for the second CEU individual, etc. These data have already been quantile normalized. There is also a file labeled **Genotypes.txt** that contains genotype data for a single SNP for each individual. The header contains the SampleID (CEU1, CEU2, ... YRI8) and the genotype of each individual. Genotypes are coded as 0, 1, and 2, which corresponds to CC, CT, and TT genotypes, respectively (i.e., the number of T alleles an individual has).

1. Read the dataset into R and average the two replicates for each individual.
2. Perform ANOVA using the genotype data from all 16 individuals for each probeset (i.e., expression data for each gene measured). What is the most significant probeset and what is its p-value?
3. This analysis is problematic for two important reasons. What are they and how would you address them to make the inference more robust?

2 A Polygenic Risk Score for Eye Color

Walsh et al, 2011 constructed a polygenic risk score (PRS) for eye color from the genotypes at 6 SNPs. The authors constructed their PRS using 3804 Dutch individuals. A substantial fraction of Dutch individuals are blue-eyed. They report a 90% accuracy in predicting blue and brown eye color based on genotype. We will use their PRS to predict eye color from genotypes.

Notation and predictor

For a given individual, let X_k represent the minor allele count at loci $k = 1, \ldots, 6$. Each individual is classified as either brown-eyed, blue-eyed, or other. Eye color predictions can be formed using two log-odds ratios,

$$\ln\left(p_{blue}/p_{brown}\right) = \hat{\alpha}_1 + \sum_k \hat{\beta}_{1,k} X_k \tag{1}$$

$$\ln\left(p_{other}/p_{brown}\right) = \hat{\alpha}_2 + \sum_k \hat{\beta}_{2,k} X_k \tag{2}$$

Here values of p correspond to the probability of an individual in the sample population having a particular eye-color. Walsh et al. formed the estimates $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_{1,k}$, and $\hat{\beta}_{2,k}$. For this question, you will use the following estimates to predict liability for brown eyes in a simulated target population: $\hat{\alpha}_1 = 3.94$, $\hat{\alpha}_2 = 0.65$,

k	Chromosome	Position	rsid	Minor allele	$\hat{\beta}_1$	$\hat{\beta}_2$
1	15	28365618	rs12913832	A	-4.81	-1.78
2	15	28230318	rs1800407	T	1.40	0.87
3	14	92773663	rs12896399	G	-0.58	-0.03
4	5	33951693	rs16891982	C	-1.30	-0.50
5	11	89011046	rs1393350	A	0.47	0.27
6	6	396321	rs12203592	T	0.70	0.73

1. Solving for p_{brown} yields,

$$p_{brown} = \frac{1}{1 + e^{\hat{\alpha}_1 + \sum_k \hat{\beta}_{1,k} X_k} + e^{\hat{\alpha}_2 + \sum_k \hat{\beta}_{2,k} X_k}}$$

   Compute p_{brown} for an individual with genotype $0 = X_1 = X_2 = X_3 = X_4 = X_5 = X_6$.

2. Under this model, which genotype has the highest value of

$$\ln\left(\frac{p_{brown}}{1 - p_{brown}}\right) \tag{3}$$

   Justify your answer.

3. (a) Load the dataset 'geno_prs.txt'
   (b) Compute (3) for each individual in the dataset.
   (c) Plot a histogram of (3) for individuals with each eye color. You should end up with three histograms.

reference Walsh, S., Liu, F., et. al. 2011. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Science International: Genetics, 5(3), pp.170-180.

3 RNA-Seq Analysis

For this problem, we will be using the result of RNA-Seq data. Remember that this type of data gives the number of reads mapped to a specific gene, which is an indication of expression level. The dataset found in the tab-delimited file **rnaseq.txt** contains data from two biological replicates of yeast grown in two conditions: 'E' was grown in ethanol, while 'G' was grown in glucose, and '1' and '2' represent the two biological replicates within those conditions. You will now see whether there is differential expression in any of those genes between those two conditions.

1. Read this downloaded file into R using read.table. Use the edgeR package to test whether these genes are differentially expressed between the ethanol (E) and glucose (G) conditions. Instructions for using the edgeR package are in Section 1.4 in the edgeR Quick Start Guide. Print the p-values obtained for each gene.

   Note: You can get your answer by following the steps exactly as in Section 1.4 of the edgeR Quick Start Guide. Compute your p-values using only **quasi-likelihood F-tests**. If you are interested, you can read in the edgeR guide what each function does. This guide is very complete, but also quite lengthy, so do this at your leisure.

2. edgeR calculates a p-value for each gene, but when you have multiple p-values you need to adjust for multiple hypothesis testing.

   (a) In 1-2 sentences, why is this necessary?

   (b) Correct those p-values using the Bonferroni correction. Show a histogram of your corrected p-values. After Bonferroni correction, how many are significant at a threshold of .05?

3. edgeR also calculates the log fold change (logFC) between G and E for each gene. Show a histogram of the distribution of log fold changes. What is the most overexpressed gene in E relative to G? What is the most overexpressed gene in G relative to E?

4. Use the R graphing package ggplot2 with option 'geom text' to create a plot with the logFC on the x-axis, the log of the Bonferroni-corrected p-value on the y-axis, and the names (not just dots) of each of the 18 genes on the graph. What does this plot suggest is the relationship between effect size (log FC) and significance? Documentation about ggplot can be found here. An advantage of R and ggplot2 is that graphing operations can be written concisely.

4 CRISPR Technologies

1. What is Cas9 and what does it do?
2. How is Cas9 modified to achieve different effects in CRISPRko, CRISPRi, and CRISPRa systems?
3. What is a PAM? Why does it matter?
4. CRISPick (https://portals.broadinstitute.org/gppx/crispick/public) is a webtool that helps biologists design sgRNA sequences for different CRISPR systems. Designing sgRNA sequences is a critical component of CRISPR systems, as these determine where Cas9 is directed. CRISPick helps make this task significantly easier by automatically a list of candidate sequences matching user-input specifications. Use CRISPick to design 5 sgRNA sequences for the gene BRCA1 using reference genome GRCh38, the CRISPRi system, and SpyoCas9. What is the sgRNA sequence for the third result? What does the "On-Target Efficacy Score" column describe and why is it important?