Problem Set 3 - Complex traits

This homework should be sent to mgymrek@ucsd.edu with subject line [CSE291 PS3:LASTNAME] by the beginning of class on **Thursday, February 9**. The assignment is worth 10 points total.

As in problem set 2, template code is provided for some problems. Using the template code is optional, it is simply there to guide you.

Objectives

- Perform a basic genome-wide association study (GWAS).
- Explore the effect of confounding factors like population structure.
- Predict a person's phenotype for a complex trait (eye color).

A description of data files for this problem set and several setup steps can be found at PS3 resources.

Part 1: A basic GWAS (4 points)

Overview

In this exercise we will perform a GWAS of a quantitative trait. We have a dataset of 206 European individuals, and have recorded the height for each sample.

In the ps3 data directory, you'll find genotype and phenotype data in plink format:

```
ps3_gwas.ped
ps3_gwas.fam
ps3_gwas.map
ps3_gwas.bed
ps3_gwas.bim
ps3_gwas.phen
```

See the plink website http://pngu.mgh.harvard.edu/ purcell/plink/data.shtml for a description of these formats. These are similar to in problem set 2, but now we have added a phenotype file ".phen".

Note, the height data has been normalized to have mean 0 and variance 1, so you'll see some negative numbers. Also of note, we will restrict analysis to a single chromosome (chr2). Finally, this data has been simulated using random SNPs and effect sizes, so don't try to compare to existing height GWAS data.

In the ps3 templates directory, you'll find:

- A bash script for running the GWAS using plink:

  `run_ps3_gwas_plink.sh`

- A script for generating Manhattan plots from plink results:

  `ps3_manhattan.py`

Exercises

1. **(1 point)** Perform a GWAS using plink, which should output a file `ps3_gwas.assoc.linear.tab`.

   Some SNPs could not be tested, and are reported as "NA". What happened to those SNPs?

   Plot the results in the form of a Manhattan plot. You'll notice positions that with many strong signals seemingly forming a vertical line. Describe what is driving that pattern.

2. **(1 point)** For the purposes of illustration, we will choose a significance threshold of $p < 10^{-4}$ to determine genome-wide significance. (Note, the canonical GWAS threshold is 5×10^{-8}). How many SNPs pass our threshold? (Approximately) How many independent signals does this represent? What if you ignore the region spanning from chr2:135397569-137621910, which appears to contain multiple signals? Note, you might find the plink `--clump` option helpful.

3. **(1 point)** Generate a QQ plot of the p-values for each SNP compared to a uniform p-value distribution. Are the p-values well calibrated? Is there any evidence that the study has confounding variables we didn't account for?

4. **(1 point)** The strongest signal should be in the region of chr2:135397569-137621910. Which genes fall in this region (hint, you can look at this region on the UCSC Genome Browser). Do any genes sound familiar (hint: remember our positive selection discussion). Based on the previous question and knowledge of this gene, do you think this represents a true signal? (more hints below).

Part 2: Confounding by population structure (3 points)

Overview

It turns out that our dataset from the previous example consists of a mixture of southern Europeans (TSI, from Sardinia) and northern Europeans (CEU).

Notably, northern Europeans are on average several inches taller than southern Europeans. Thus, mutations that are simply correlated with northern vs. southern European ancestry (e.g., the mutation for lactase persistance), and not necessarily related to height, might still show quite strong signals in our association test.

Below we'll repeat the analysis, this time accounting for population structure, and see how this changes the results.

Exercises

1. **(1 point)** Calculate the top 10 genotype principal components. Plot PC1 vs. PC2 for all samples. You should see two main groups (the PCA won't separate them that well here)? Hint: while we previously wrote code to perform PCA in problem set 2, for this problem set you might find plink's `--pca` option will save you some time.

2. **(1 point)** Repeat the GWAS analysis, now using the top 10 PCs as covariates in the analysis (hint, see plink's `--covar` option). Regenerate the Manhattan plot and QQ plot.

3. **(1 point)** Describe and explain the change in results. What happened to the signal at `LCT`?. What pros and cons can you think of for including PCs as covariates in the analysis?

Part 3: Predicting eye color (3 points)

Overview

In this exercise we'll predict eye color from SNP data. We will be following a model from the IrisPlex Paper and it could be helpful to take a look at that paper.

We will treat eye color as a case/control phenotype, except with the slight complication that rather than two categories (e.g. disease, not disease), we'll predict three: blue eyes, brown eyes, or other colored eyes. Multinomial regression is a good model for this task.

Walsh 2011 and others have shown that just 6 SNPs can do a pretty good job predicting eye color. They trained the following two models (see their Supplementary Table 3):

$$\ln(p_{blue}/p_{brown}) = \alpha_1 + \sum_k \beta_{1,k} X_k \tag{1}$$

$$\ln(p_{other}/p_{brown}) = \alpha_2 + \sum_k \beta_{2,k} X_k \tag{2}$$

where α_i is an intercept term, $\beta_{i,k}$ is the effect size for model i at the kth SNP, and X_k is the number of minor alleles at an individuals genotype for the kth SNP (0, 1, or 2). Rearranging these allows us to predict the probability of each class of eye color:

$$p_{blue} = \frac{e^{\alpha_1 + \sum_k \beta_{1,k} X_k}}{1 + e^{\alpha_1 + \sum_k \beta_{1,k} X_k} + e^{\alpha_2 + \sum_k \beta_{2,k} X_k}} \tag{3}$$

$$p_{other} = \frac{e^{\alpha_2 + \sum_k \beta_{2,k} X_k}}{1 + e^{\alpha_1 + \sum_k \beta_{1,k} X_k} + e^{\alpha_2 + \sum_k \beta_{2,k} X_k}} \tag{4}$$

$$p_{brown} = 1 - p_{blue} - p_{other} \tag{5}$$

Below is a reproduced table of the model parameters from the 6 predictive SNPs, converted to hg19 coordinates:

Chromosome	Position	rsid	Minor allele	β_1	β_2
15	28365618	rs12913832	A	-4.81	-1.79
15	28230318	rs1800407	T	1.40	0.87
14	92773663	rs12896399	G	-0.58	-0.03
5	33951693	rs16891982	C	-1,30	-0,50
11	89011046	rs1393350	A	0.47	0,27
6	396321	rs12203592	T	0.70	0.73
$\alpha_1 = 3.94$	$\alpha_2 = 0.65$				

In the ps3 data directory, you'll find:

- The table of parameters above:

  `eyecolor_snps_irisplex.bed`

- Genotypes for CEU and TSI individuals in VCF format at these 6 SNPs:

```
ps3_pred_eyecolor.vcf.gz
```

If you use Python, you might find the VCF parsing library helpful for the exercises below. You can install it by doing:

```
pip install --user pyvcf
```

Exercises

1. **(1 point)** Using the IrisPlex SNPs, calculate probability that each individual in the dataset has blue, brown, or other colored eyes. What do you predict for sample NA12249? Sample NA20509? Sample NA12750?

2. **(1 point)** Calculate the mean probability of blue, brown, or other colored eyes for each population, CEU and TSI. Which group is more likely to have blue eyes? Does this match with what is known about eye color frequencies in those populations?

3. **(0.5 points)** What is the interpretation of the effect sizes? In particular, how can we interpret $\beta_1 = -4.81$ for rs12913832?

4. **(0.5 points)** Look up the top SNP, rs12913832 in dbSNP or on the UCSD Genome Browser. Does this SNP fall in any predicted coding region? Hypothesize how it might be affecting eye color.