

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 **Tuesday, February 7**. 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.

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 persistence), 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. How many clusters are there (hard to see, but they're there)? 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: Trait prediction (3 points)

Overview

Exercises

1. **(1 point)** Calculate a polygenic risk score for height for the CEU and TSI populations, using the genome-wide significant SNPs. You might find the script `run_ps3_predict.sh` helpful.
2. **(1 point)** Convert this to a height prediction based on the population-wide mean height distribution described above.
3. **(1 point)** Repeat this process to predict eye color of each samples. Which population is more likely to have blue eyes?
4. **(Bonus)** Repeat the trait prediction on your own genome!