Instructions for Independent Work

1. In this section, You should work through the three notebooks with two different phenotypes:

A. Phenotype 1 - Cadmium concentration

This is the leaf cadmium concentration of plants grown in the greenhouse.

Phenotype file = ./data/cadmium_concentration.csv

A GWAS for this dataset has been published: https://doi.org/10.1371/journal.pgen.1002923

QUESTION: What does the Manhattan plot look like for a simple trait?

B. Phenotype 2 - Flowering tine in Sweden

This is another subset of the flowering time data.
This time, instead of a random subset, we are considering only Swedish accessions.

Phenotype file = ./data/nsweden_flowering_time_16.csv

QUESTION: How does this analysis differ from the first one we ran together (hint - compare the Manhattan plots). Why do you get different answers when you use different subsets of the same phenotype? (If interested in understanding more, check out the preprint at https://www.biorxiv.org/content/10.1101/2021.02.26.433043v1)

2. Make sure you ***change the names of input and output files in section 1B*** of all three notebooks. To do this, just replace "subset_flowering_time_16" with either "cadmium_concentration" or "sweden_flowering_time_16". Don't change any other part of the file names or change the names of the files with the genotypes or K matrix.

3. Run the three notebooks ***step-by-step***. Focus on what each step of code is doing and why (rather than trying to understand each line individually).

4. Ask yourself:

a. What does an appropriate phenotype for GWAS look like?

b. What input data do you need to run GWAS?

c. How does a linear mixed model test for association between genotypes and phenotypes?

d. **How would you read and interpret a Manhattan plot (including Bonferroni cutoff)?**

e. **What does a QQplot look like if p-values are inflated by population structure?**

(Think about these as you work on your own today and please just ask if you need clarification about any of these points!)

5. What are the differences in GWAS results among the three phenotypes? Which traits are simple and which are complex? Which have more p-value inflation? Which one do you think is more interesting and why?

6. If you are working more quickly than the others, why not try one of the following **optional challenge exercises**?

a. Run another GWAS with a phenotyping dataset whose accessions cover a small geographic area (./data/rosette_color.csv). This is a measure of the color of plants growing in the field, which is often a sign of stress. What's different about GWAS here?

b. Try to run a GWAS with different minor allele frequency cutoffs. You will have to figure out how to change input files and variables accordingly!

c. If you are interested in hdf5 files and how to use them in python, how about trying to understand the code in notebook 2 line by line?

7. Some hints about using jupyter notebooks:

a. Shift-enter runs the cell and moves to the next one.

b. Control-enter runs the cell and doesn't move.

c. An asterisk in brackets next to a cell means that it is running.

d. Hitting "esc" puts you in a mode where you can move between cells with your arrow keys. This is called command mode.

e. Hitting "enter" puts you in a mode to edit cells. This is edit mode.

f. **Help, a cell is acting weird!** (a cell of code won't run **or** a cell of text runs and gives weird errors) In this case, you might be in the wrong mode. A cell can be either markdown mode (M) which is for text, or script mode (Y) which is for writing code. In command mode (hit esc), use arrows to select a cell and then hit either M or Y to toggle between the two.

g. There are many keyboard shortcuts for jupyter notebooks! Use a cheatsheet to explore them more:

   https://www.cheatography.com/weidadeyue/cheat-sheets/jupyter-notebook/