Instructions for Independent Work

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

    a. **Phenotype 1 - Chromium concentration**
    This is the leaf chromium concentration of plants grown in the greenhouse.
    Phenotype file = ./data/cadmium_concentration.csv
    A GWAS for this dataset is published:
    https://doi.org/10.1371/journal.pgen.1002923
    QUESTION: What does the Manhattan plot look like for a simple trait?

    b. **Phenotype 2 - The entire dataset for flowering time at 16 degrees.**
    This is the entire flowering time dataset for 1100 accessions.
    Phenotype file = ./data/flowering_time_16.csv
    QUESTION: How much longer does this GWAS take to run? Are the Manhattan and QQ plots different using the full dataset?

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 "chromium_concentration" or "flowering_time_16". You shouldn't change any other part of the file names.

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. Do you understand:
    a. What an appropriate phenotype for GWAS looks like?
    b. The input matrices/variables that limix is using?
    c. How a linear mixed model tests for association between genotypes and phenotypes?
    d. How to read and interpret a Manhattan plot (including Bonferroni cutoff)?
    e. What a QQplot looks like if p-values are inflated?
    f. Why we use a minor allele frequency cutoff?

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 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 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?

Some hints about using jupyter notebooks:

1. Shift-enter runs the cell and moves to the next one.
2. Control-enter runs the cell and doesn't move.
3. An asterisk in brackets next to a cell means that it is running.
4. Hitting "esc" puts you in a mode where you can move between cells with your arrow keys. This is called command mode.
5. Hitting "enter" puts you in a mode to edit cells. This is edit mode.
6. **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.
7. There are many keyboard shortcuts for jupyter notebooks! Use a cheatsheet to explore them more:
   https://www.cheatography.com/weidadeyue/cheat-sheets/jupyter-notebook/