# Analysis journal

Jay Gillenwater

8/25/2021

## Overview

This document will be a day-by-day record of what I do in the development of this workflow. I'm making this a few days into development though so I've grouped the first few days into one section.

#### Before 2021-08-25

Set up the workflow with tflow and targets. Got some functions working to read in and clean up the phenotype data from the excel spreadsheet. Also worked on some functions for some simple exploratory visualizations (boxplots, histograms, and QQ plots). I used these plots to find one sample that was likely an error (LA PLOT: 255) and removed it from the data. I also added a function that takes the clean data with the likely bad sample removed and calculates average values for each phenotype for each genotype. The function also calculates lsmeans but the average may be more appropriate to use since many of the genotypes were not balanced across locations. I still want to go back and compare the outputs from this function/look into what version would be best to use as the phenotypes for mapping. Maybe change it to also give by location averages for each phenotype in addition to the average across location.

Yesterday I worked on cleaning up the genotype data into a format that works with r/qtl. I settled on the "csvsr" format since the export from genomestudio was already close to this format and the phenotype data is already separate. The main challenge was converting the genotype data into the "ABH" format expected by r/qtl. The NC-Raleigh parent (code 2104) was genotyped, but the soja parent was not so I was not able to directly convert the observed progeny genotypes to the ABH format based on the observed founder genotypes. One approach I tried was to use the genotypes provided when the USDA germplasm collection was genotyped with the 50k SNP chip. The 6K SNP chip that was used to genotype this population uses a subset of the markers from the 50K chip and as such, the historical genotypes could also be subset to the markers used to genotype the progeny. I tried this approach but the alleles present in the historical sample did not match up with those observed to be segregating in the current population. As an example, the Raleigh parent had a genotype of AA at a SNP, and the historical Soja genotype had a genotype of GG, but the observed genotypes across samples in the current data were only AA, AC, and CC. Because of this mismatch, I instead opted to "derive" the soja parents genotype at each marker by observing what alleles were present at each site. Using the same example from above, I would assume that the founder parent had a genotype of CC at that particular SNP. Each SNP in the data only had two alleles, so this approach worked although I wasn't able to so some quality control steps like removing SNPs that were missing or heterozygous in the soja parent.

I want to revisit using the historical data though. I expected more of the alleles to fit with the observed segregation patterns just trough chance alone than I saw when I joined the historical soja genotypes to the new data. I was worried that there could be a problem with how I joined the data with the SNP names, but I haven't been able to find an error in that part (although I'm still looking). Maybe something in the conversion of the dbSNP names in the historical data to the longer names that are used in the 6K chip? Differences in the way the genotypes are presented in the two sources may also be causing the mismatch, I have to look into the technical details of the two chips to be sure though.

Added functions to export the cleaned phenotype and genotype data to external files that are ready to be input into r/qtl and another function to actually read them into r/qtl.

**TODO** Figure out why the alleles in the historical SNP data for the soja parent don't match those in the current data.

**2021-08-25**

Explored some basic quality checks for linkage mapping with the cross data by using ASMap. Really just exploring the data and seeing how many markers pass different thresholds for missingness and segregation distortion.

The data does seem to have more heterozygosity than I expected, about 8.5% before any filtering. I'll have to check if this is due to a few problematic SNPs or is more general. I think I should go back to the starting genotype matrix as it was exported from genomestudio and get some summary stats for each marker and the data as a whole to get a better sense of the magnitude and distribution of the heterozygosity.

I ran the mapping function (mstmap.cross) on the data without much filtering too, just to see what I'd get out really. In general, the marker orders math well with their physical orders so that's good. The genetic lengths of the linkage groups are much larger than what I'd expect from a final map (going off past maps using the same SNP chip at least). I haven't done much filtering of the SNPs or genotypes yet so that wasn't unexpected. Next I want to work on cleaning up the map. Right now, I think the best way to do this is still to use the ASMap functions interactively, I'll have to be careful as I go along though to keep everything organized because ultimately, I want to repeat the steps to get the final map in a function that I can use in the overall workflow.