Name: Kevin Muirhead

MDSC 679: Machine Learning Project 1

Abstract

Introduction

*Arabadopsis thaliana* is an important model organism with several traits making it a useful model for understanding the biology of flowering plants. *A. thaliana* has relatively short life cycle, small genome of size 135 megabase pairs (Mbp) and

It is useful for studying multifactorial traits such as flowering time (FT) that are potentially regulated by the interaction of multiple genes on various chromosomes and stimulus from the environment.

The goal of this particular study is to find the relationship of genotype to phenotype.

In this study only homozygous genotypes were observed.

Methods

In order to study the genetic causation of FT, a phenotype to genotype mapping study was conducted to assess the underlying mechanism of the trait. A genome wide association study (GWAS) was conducted using 251 individuals. A panel of 214553 variant SNPs were genotyped and analyzed using a custom pipeline implemented in the python and R programming languages. The pipeline workflow is captured in the following figure x. The quality control step workflow was developed in python for filtering the data for further analysis. Flowering time phenotypes were filtered for “NA” values indicating missing data. A total of 238 phenotypes were retained for further processing of individual genotypes. Genotypes were first filtered for biallelic SNPs. Genotypes with flowering time phenotype data were assessed for minor and major allele frequency. Alleles of each variant were observed and counted, and minor and major alleles were determined. After filtering by phenotypes and biallelic SNPs, a total of 230 genotypes and a total of 214219 variants were retained before association testing SNP sets. All variants were subjected to SNP-HWE tests (Wigginton et al. 2005) to assess the population for genotyping quality, for derivations from Hardy-Weinberg equilibrium (HWE) and as test of association of the variants. Variants with a P-value of less than or equal to α = 0.001 were retained as this indicated that they were not derived from genotyping error. The Minor Allele Frequency (MAF) of each variant was calculated and variants with a MAF of greater than or equal to 0.05 were retained and considered for further analysis. A total of xxx individual genotypes and XXX variants were retained after the quality control filtering step.

The type I error rate was assessed using a Quantile-Quantile Plot (QQ-Plot) shown in figure x. The plot indicates that the type I error rate is under control and further analysis can be performed. A Manhattan plot (−log10(P) genome-wide association plot)

was generated to visualize significant SNP variants across the entire genome in the locations that were genotyped.

A search for *Arabadopsis thaliana* genes known to be associated with flowering time was conducted using the TAIR website (<https://www.arabidopsis.org/>). A total of 79 genes were found using the search term “flowering time” and the accession numbers were downloaded. The 79 genes found in the TAIR database were compared to the genes with variants to assess whether or not these “flowering time” associated

#### Talk about type I error rate

###More specifically, in order to control the overall type I error rate, the level at which each test is conducted must be adjusted.

genotypes.csv

Annotations (gene\_model.gff)

References:

1. Reference Dr. Long’s paper.
2. Hamazaki, K. and Iwata, H. (2020) RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. PLOS Computational Biology, 16(2): e1007663.
3. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*. 2005;76(5):887-893. doi:10.1086/429864
4. Turner, S.D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. biorXiv DOI: 10.1101/005165 (2014).
5. Tanya Z. Berardini, Leonore Reiser, Donghui Li, Yarik Mezheritsky, Robert Muller, Emily Strait and Eva Huala. The Arabidopsis Information Resource: Making and mining the "gold standard" annotated reference plant genome. genesis 2015 doi: 10.1002/dvg.22877

Salomé PA, Bomblies K, Laitinen RA, Yant L, Mott R, Weigel D. Genetic architecture of flowering-time variation in Arabidopsis thaliana. Genetics. 2011 Jun;188(2):421-33. doi: 10.1534/genetics.111.126607. Epub 2011 Mar 15. PMID: 21406681; PMCID: PMC3122318.

F. Bouché, G. Lobet, P. Tocquin, C. Périlleux, FLOR-ID: An interactive database of flowering-time gene networks in Arabidopsis thaliana. Nucleic Acids Res. **44**, D1167–D1171 (2016).

“””Candidates were selected that contained ‘flowering’ or ‘vernalization’ in their name and/or description.

“”

Look at references in here https://www.pnas.org/content/116/36/17890

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2894516/