Name: Kevin Muirhead

MDSC 679: Machine Learning Project 1

**Predicting Phenotype by Genotypes: A Comparison of Multiple Models**

**Introduction**

*Arabadopsis thaliana* is an important model organism for plant biology research . A. thaliana has a short life cycle, a small genome of 135 megabase pairs (Mbp) and highly curated genomic database TAIR10 (Tayna *et al.* 2015). Having these features allows for a multitude of possible research interests of economic and commercial value for agriculture (The Arabidopsis Genome Initiative. 2000). *A. thaliana* is useful for studying multifactorial traits such as flowering time (FT), which has been extensively studied. The flowering time trait is regulated by the interaction of multiple genes on various chromosomes and stimulus from the environment. The importance of studying the genes involved in flowering time has led to the many advances in plant breeding of various crop varieties such as *Brassica napus, Brassica rapa* and *Brassica oleracea* (Paterson *et al.* 2001). Where flowering time is essential for developing seeds for the food industry or for new crops for the following year. *A. thaliana* has continued to be an important model organism which is still used for plant research.

It is quite possible that the association mapping analysis of flowering time is strongly by population structure given a small sample size. (Salomé et al. 2011). Since the current study involves a smaller sample size of 251 individuals the Genome Wide Association Study (GWAS) may lead to inconclusive results. Most GWAS studies involve at least more than 1000 individuals. The goal of this particular study is to find the relationship of genotype to phenotype for the flowering time (FT) trait using association mapping and three machine learning models; Polygenetic Linear Regression, LASSO Regression (L1 Regularization) and RIDGE Regression (L2 Regularization. The purpose of this study is to predict the phenotype given the genotypes and compare three models. Machine learning is a useful method for many things. Including predicting phenotype given the genotype. This study will compare the power of these three models and will hopefully provide a detailed evaluation of the models.

**Methods**

A panel of 214553 genotyped SNPs from 251 individuals (Atwell *et al.* 2010) were analyzed using a custom pipeline implemented in the python and the R programming languages. The source code with python and R scripts for the workflow are deposited in the github repository <https://github.com/kevmu/MDSC_679> under ML\_Project\_1. Instructions on how to install and configure the pipeline are detailed in the README documentation. The general pipeline workflow is captured in figure 1. In the first step, the quality control workflow was developed in python for filtering the data for further analysis. The quality\_control.py script performs the following tasks Quality filtering, Association mapping, quality control and formatting input files for the machine learning models. The quality control workflow diagram is shown in figure 2.

## Quality Filtering

Flowering time phenotypes are filtered for “NA” values or missing data and were removed. 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 based on the count information. After filtering by phenotypes and biallelic SNPs, the Minor Allele Frequency (MAF) of each variant is calculated.

Association Mapping

Quality filtered SNP variants were retained before association testing using the rMVP Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-Wide Association Study R package version v.1.0.5 (Yin *et al.* 2021). All format conversion, association mapping and adjusted pvalue calculations were performed using the os.system() call in python. The genotype.csv file is parsed and converted into a pedigree formatted file (PED) and a genotypes map formatted file which was generated for input into Plink version 1.90b6.21 (Purcell *et al.* 2007) program. Plink is executed from python with the recode vcf parameter option for converting the PED and genotypes MAP file to Variant Call Format (VCF) file using the following command;

plink --ped plink\_genotype\_ped\_infile --map plink\_genotype\_map\_infile --pca --recode vcf --allow-no-sex --out out\_prefix

Where the plink\_genotype\_ped\_infile is the pedigree format (PED) file, plink\_genotype\_map\_infile is the genotype map format (MAP) file and out\_prefix is the output file path prefix for naming plink output files.

The VCF file was then fixed for input into the rMVP R package because the plink command used the “Family\_ID” as a prefix for the genotype\_id which caused issues when executing the rMVP R package functions data preparation function . The following awk shell command was used to reformat the VCF file into the “fixed” format;

sed 's/Family\_ID\_//g' < mvp\_genotypes\_vcf\_file > mvp\_fixed\_genotypes\_vcf\_file

A phenotypes file is generated using the list of filtered genotype ids against the phenotypes dictionary data structure in python for input into the rMVP program. The rMVP\_marker\_tests.R R script is used to perform the rMVP association tests for SNP variant markers. The following command was used;

Rscript rMVP\_marker\_tests.R -i mvp\_fixed\_genotypes\_vcf\_file -p mvp\_phenotype\_infile -o association\_mapping\_output\_dir.

Where the mvp\_fixed\_genotypes\_vcf\_file is the fixed VCF file mvp\_phenotype\_infile is the filtered phenotypes file and the association\_mapping\_output\_dir is the association mapping output directory for writing the output files and various plot image files. All variants were subjected to Generalized Linear Model (GLM), Mixed Linear Model (MLM) and FarmCPU associations tests in the rMVP R Package. SNP variants are retained for further filtering of p-values. Using a shell command for determining the number of SNPs for each chromosome was counted for distribution purposes as SNP Density plots can be deceiving for how many SNPs are actually contained in the >2697 category.

for i in {1..5};

do

echo "Chr${i}";

grep "Chr${i}" phenotype.FarmCPU.csv | wc -l;

done

These SNP variants were assessed for tests of association with the flowering time (FT) phenotype. Various plots are generated including QQ-Plots, Manhattan Plots, SNP density plot, and Phenotype distribution histogram

Quality Control

P-values of tested variants are adjusted using adjusted p-value calculations for Bonefferoni correction and q-values using the calculate\_adjusted\_pvalues.R script which uses the pvalue.adjust function and the qvalue library in R. The following command was used;

Rscript calculate\_adjusted\_pvalues.R -i mvp\_phenotype\_association\_file -o adjusted\_pvalues\_outfile.

Where the mvp\_phenotype\_association\_file was the phenotype.MLM.csv file and the adjusted\_pvalues\_outfile is the phenotype.FarmCPU.adjusted.pvalues.tsv file.

It was determined that an alpha value less than or equal to 0.05 of each pvalue using Bonferroni correction was used to retain SNP variants for further processing.

File Formatting For Machine Learning

Genotypes at each SNP variant that passed the filtering procedure were encoded 0 for the major allele and 1 for the minor allele. Genotypes and corresponding phenotypes were written to a file for input into the machine learning models using python.

Machine Learning Models

The second step of the workflow executes the machine learning models, polygenetic linear regression, LASSO regression (L1 Regularization) (Tibshirani *et al.* 1996), and RIDGE regression (L2 Regularization) (Hoerl *et al.* 1970). All machine learning models were implemented in the python scikit-learn package version 0.24.1 (Pedregosa et al. 2011). The encoded genotype files are parsed using the python pandas version 1.2.3 (The pandas development team, 2020) and loaded into a numpy array using the python numpy package version 1.20.2 (Harris *et al.* 2020) for each model. A k-fold cross-validation and randomized search approach was used in order to validate the alpha value hyperparameter for the LASSO and RIDGE regression models. The models were evaluated using 70% for the training set and 30% for the testing set and the following metrics were calculated. The Mean Squared Error (MSE), Root Mean Squared Error (RMSE), the Coefficient of determination R-squared (R2), Mean Absolute Percentage Error (MAPE) and MAPE accuracy. The phenotype predictions based on phenotypes are then performed for each model.

**Results**

In order to study the genetic causation of flowering time (FT), a phenotype to genotype mapping study was conducted to assess the underlying mechanism of the trait using 251 genotyped individuals. A panel of 214553 variant SNPs were genotyped and analyzed using a custom pipeline implemented in the python and the R programming languages shown in Figures 1 and 2.

## Flowering time phenotypes were filtered for “NA” values, which indicated that those particular genotypes had missing data and were removed. 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, the Minor Allele Frequency (MAF) of each variant was calculated. Variants with a MAF of greater than or equal to 0.01 or 1% were retained and considered for further analysis.

A total of 230 genotypes and 214218 variants were retained before association testing of SNP variants using the rMVP Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-Wide Association Study R package version 1.0.5 (Yin *et al.* 2021). All format conversion, association mapping and adjusted pvalue calculations were performed using the os.system() function call in python. The genotype.csv file was parsed and converted into a pedigree formatted file (PED) and a genotypes map formatted file. Plink version 1.90b6.21 (Purcell *et al.* 2007) was executed with the recode vcf parameter option for converting the PED and genotypes MAP file to Variant Call Format (VCF). The VCF file was then fixed for input into the rMVP R package. A phenotypes file was generated using the list of filtered genotype ids in python for input into the rMVP program. The rMVP\_marker\_tests.R R script was used to perform the rMVP association tests for SNP variant markers. All variants were subjected to Generalized Linear Model (GLM) (Price *et al.* 2006), Mixed Linear Model (MLM) (Yu *et al.* 2006) and FarmCPU (Liu *et al.* 2016) associations tests in the rMVP R Package and compared. A total of 212218 SNP variants were retained for further filtering of p-values.

The distribution of the flowering time phenotype trait are shown in figure 3. A binomial distribution is shown where the majority of phenotypes are above the 80.0 value. Phenotypes are within the range of 40-120 values marks. With a mean of 86.66 and a standard deviation at 21.11. The density of genotyped SNPs is shown in figure 4 giving a layout of concentration of variants that the three models, GLM, MLM and FarmCPU within a 1 Mb window size. The number of SNPs for each chromosome was counted for distribution purposes as SNP Density plots can be deceiving for how many SNPs are actually contained in the >2697 category. The number of SNPs per Chromosome is shown in Table 1. This plot was generated using the genotype data file and does not reflect the real length of the genome.

|  |  |
| --- | --- |
| Chromosome ID | Number of SNPs |
| Chr1 | 51366 |
| Chr2 | 28148 |
| Chr3 | 43104 |
| Chr4 | 36742 |
| Chr5 | 52858 |

Table 1: The number of genotyped SNPs for each chromosome. Used as an aid for the SNP density plot in figure 4.

These SNP variants were assessed for tests of association with the flowering time (FT) phenotype. The type I error rate was assessed for each of the three models using a Quantile-Quantile Plot (QQ-Plot) shown in figure 5. The QQ-Plot illustrates whether or not the models for association testing control false positives and false negatives by comparing the expected and observed negative-log of association probabilities (Kaler *et al.* 2020). The GLM model depicts a sharp derivation from the 1:1 line first indicating that the SNP are significant. This is in fact not the case as the GLM models is known to generate a false representation of the genotype to phenotype relationship. The QQ-Plot for the FarmCPU model is shown in figure 6. This indicates that the type I error rate for the FarmCPU model is under control and was selected as the association testing model conducted for further analysis. 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. As shown in figure 7 the Manhattan plot of the MLM model shows no significant pvalues for that model. The Manhattan plot for the FarmCPU model is shown in figure 8, the P-values are at a reasonable value. The MLM and FarmCPU models were chosen going forward.

P-values of tested variants were subjected to adjusted p-value calculations for Bonferroni correction using the calculate\_adjusted\_pvalues.R script which uses the pvalue.adjust() function in R. The Bonferroni correction was calculated for FarmCPU model results, alpha’ = 0.05/212218 where the alpha’ = 2.35606782 x 10-7 and pvalues were adjusted. This accounted for the multiple testing problem. It was determined that an alpha value less than or equal to 0.05 of each Bonferroni corrected pvalue was used to retain SNP variants for further processing for the FarmCPU model results. It was determined that the FarmCPU model was the most reliable method. The significant signal line in figure 7 for FarmCPU had nine significant SNP variants. These nine SNPs were used for input into the three machine learning models, polygenetic linear regression, LASSO regression (L1 Regularization) and RIDGE regression (L2 Regularization). The SNPs from the FarmCPU model results were used as a test for the Apriori algorithm in ML\_Project\_2.

Genotypes at each SNP variant that passed the filtering procedure were encoded 0 for the major allele and 1 for the minor allele. Genotypes and corresponding phenotypes were written to a file for input into the machine learning models using python. The three machine learning models, polygenetic linear regression, LASSO regression (L1 Regularization) (Tibshirani *et al.* 1996), and RIDGE regression (L2 Regularization) (Hoerl *et al.* 1970) were compared with the following criteria.

The dataset criteria for input into the three models were the following. Genotypes were filtered using MAF >= 0.01, Bonferroni corrected adjusted p-value of alpha < 0.05 for filtering significant pvalues and the major allele (reference) was encoded as 0 and the minor allele (alternative) was encoded as 1. Phenotypes were filtered by missing values and filtered genotype ids in the quality control step. There were 230 genotyped individuals and 9 SNP variants used as input for the machine learning models. The encoded genotypes format file and filtered phenotypes file is input into the three machine learning models, the Polygenetic Linear Regression model, the LASSO Regression (L1 Regularization) model and the RIDGE Regression (L2 Regularization).

The alpha value hyperparameter for the LASSO and RIDGE regression models was optimized using a k-fold cross-validation using k = 10 and number of repeats = 10 and randomized search approach in order to validate the regularization models. The best alpha value for LASSO was blank and the best alpha value for RIDGE was blank, respectively.

All three models were evaluated using 70% training set and 30% testing set and the following metrics were calculated. The Mean Squared Error (MSE), Root Mean Squared Error (RMSE), the Coefficient of determination R-squared (R2), Mean Absolute Percentage Error (MAPE) and MAPE accuracy.

The phenotype predictions based on phenotypes are performed for each model.

Need to add table for evaluation metrics

Discussion

Talk about QQ plot and manhattan plots. All figures and machine learning metrics table.

Need to figure out more references

Also need to say stuff about MLM and FarmCPU and how I didn’t know what a QQ-plot really meant and detail it from the email.

Integrate figures into results section.

Talk about mainly in Discussion section.

Discussion

A search for *Arabadopsis thaliana* genes known to be associated with flowering time was conducted using the TAIR (Tayna *et al.* 2015) website (<https://www.arabidopsis.org/>) Accessed on March 11th, 2021. 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 genes were found through association mapping using the FarmCPU model algorithm. The nine significant SNP variants are shown in table 4.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SNP | REF | ALT | Pvalue | TAIR10 Locus ID | Annotation | Location of SNP |
| Chr5\_Pos3188327 | C | T | 3.97E-18 | AT5G10170.1 | myo-inositol-1-phosphate synthase 3 | inside gene |
| Chr2\_Pos9581605 | C | G | 2.42E-10 | AT2G22540 | K-box region and MADS-box transcription factor family protein | inside gene |
| Chr3\_Pos20411685 | C | A | 2.54E-10 | AT3G55080 | SET domain-containing protein | inside gene |
| Chr1\_Pos3877554 | G | A | 1.19E-09 | AT1G11540.2 | Sulfite exporter TauE/SafE family protein | downstream of gene |
| Chr5\_Pos18513605 | T | G | 3.61E-08 | AT5G45650.1 | subtilase family protein | inside gene |
| Chr2\_Pos1022560 | C | A | 6.53E-08 | AT2G03360 | Glycosyltransferase family 61 protein | inside gene |
| Chr1\_Pos25117634 | T | A | 8.84E-08 | AT1G67140.2 | HEAT repeat-containing protein | downstream of gene |  |  |
| Chr1\_Pos19133268 | C | G | 1.49E-07 | AT1G51600 | ZIM-LIKE 2 | inside gene |  |  |
| Chr5\_Pos18591253 | C | T | 1.76E-07 | AT5G45830 | delay of germination 1 | inside gene |  |  |

Table 4: Significant SNP variants from the FarmCPU association tests after Bonferroni correction. Annotations are from the JBrowse TAIR10 instance at jbrowse.arabidopsis.org.

The Arabadopsis thaliana TAIR10 JBrowse instance (<https://jbrowse.arabidopsis.org/>) was used to annotate the nine SNP variants.

There were no annotations related to SNP variants from the nine significant SNPs found in the TAIR 10 search for “Flowering Time” genes. The annotations were then compared to the FLOR-ID (Flowering Interactive) database (Bouché *et al.* 2016) for searching for Arabadopsis thaliana genes associated with the SNPs found via association mapping using FarmCPU. The SNP variant at Chr2\_Pos9581605 is within the AT2G22540: K-box region and MADS-box transcription factor family protein, which encodes a MADS-box transcription factor. This gene is known to contribute to flowering time in the late flowering stage. It is a transcription repressor that inhibits floral transition in the autonomous flowering pathway. It is independent of photoperiod and temperature. (Gregis *et al.* 2006). The other 8 SNPs were found to be associated with expression in the “flowering stage” or “flower” on the TAIR database website. It is possible that they have some relationship directly or indirectly to the flowering time phenotype.

Conclusion

Computation Resources

Program implementation, development and analyses were performed on a MacBook Pro running MacOSX Big Sur with Quad-Core Intel Core i5 2 GHz (4 cores) and 32 GB RAM memory.

Data and Source code Availability.

All python and R scripts (commented), Installation documentation, genotypes and phenotype data are available at the following github repositiory <https://github.com/kevmu/MDSC_679>. Follow the README.md for github repository download and installation instructions for the ML\_Project\_1 pipeline. Can use the conda environments for MacOSX or Linux. Just make sure you install the R library packages as specified. Please read the README.md for more information.

References:

1. Atwell, S., Huang, Y., Vilhjálmsson, B. *et al.* “Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines.” *Nature* **465,**627–631 (2010). <https://doi.org/10.1038/nature08800>
2. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,**796–815 (2000). https://doi.org/10.1038/35048692
3. 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
4. Paterson, A.H., Lan, Th., Amasino, R. *et al.* *Brassica* genomics: a complement to, and early beneficiary of, the *Arabidopsis*sequence. *Genome Biol* **2,**reviews1011.1 (2001). <https://doi.org/10.1186/gb-2001-2-3-reviews1011>
5. 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.
6. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC. “PLINK: a toolset for whole-genome association and population-based linkage analysis.” American Journal of Human Genetics, 81. (2007).
7. Lilin Yin, Haohao Zhang, Zhenshuang Tang, Jingya Xu, Dong Yin, Zhiwu Zhang, Xiaohui Yuan, Mengjin Zhu, Shuhong Zhao, Xinyun Li, Xiaolei Liu, “rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study”, Genomics, Proteomics & Bioinformatics, ISSN 1672-0229, (2021). <https://doi.org/10.1016/j.gpb.2020.10.007>.
8. Price, A., Patterson, N., Plenge, R. *et al.* “Principal components analysis corrects for stratification in genome-wide association studies.” *Nat Genet* **38,**904–909 (2006). <https://doi.org/10.1038/ng1847>
9. Yu, J., Pressoir, G., Briggs, W. *et al.* “A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.” *Nat Genet* **38,**203–208 (2006). <https://doi.org/10.1038/ng1702>
10. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. “Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies.” PLOS Genetics 12(2): e1005767. (2016) https://doi.org/10.1371/journal.pgen.1005767
11. Kaler AS, Gillman JD, Beissinger T, Purcell LC. Comparing Different Statistical Models and Multiple Testing Corrections for Association Mapping in Soybean and Maize. *Front Plant Sci*. 2020;10:1794. (2020). doi:10.3389/fpls.2019.01794
12. Bonferroni, Carlo E.*, "Teoria statistica delle classi e calcolo delle probabilità", Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze (in Italian),****8****: 1–62, [Zbl](https://en.wikipedia.org/wiki/Zbl_(identifier)" \o "Zbl (identifier)" \t "_blank)*[*0016.41103*](https://zbmath.org/?format=complete&q=an:0016.41103) *(1936)*
13. Benjamini, Y., Hochberg, Y., “Controlling the false discovery rate, A practical and powerful approach to multiple testing.” *J. R. Stat. Soc. Series. B. Stat. Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x. (1995).
14. Benjamini, Y., and Yekutieli, D. “The control of the false discovery rate in multiple testing under dependency.” *Annals of Statistics*, **29**, 1165–1188. doi: [10.1214/aos/1013699998](https://doi.org/10.1214/aos/1013699998). (2001).
15. Pedregosa et al., “Scikit-learn: Machine Learning in Python”, JMLR 12, pp. 2825-2830 (2011).
16. The pandas development team, “Pandas”, Zenodo, version, doi: 10.5281/zenodo.3509134 (2020).
17. Harris, C.R., Millman, K.J., van der Walt, S.J. et al. “Array programming with NumPy.” Nature 585, 357–362 (2020). DOI: [0.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2)
18. Tibshirani, R. “Regression shrinkage and selection via the lasso.” J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288 (1996).
19. Arthur E. Hoerl & Robert W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems”, Technometrics, 12:1, 55-67, DOI: 10.1080/00401706.1970.10488634. (1970).
20. Mark A. Taylor, Amity M. Wilczek, Judith L. Roe, Stephen M. Welch, Daniel E. Runcie, Martha D. Cooper, Johanna Schmitt. “Large-effect flowering time mutations reveal conditionally adaptive paths through fitness landscapes in *Arabidopsis thaliana”.* Proceedings of the National Academy of Sciences Sep 2019, 116 (36) 17890-17899; DOI: 10.1073/pnas.1902731116
21. 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).
22. Veronica Gregis, Alice Sessa, Lucia Colombo, Martin M. Kater, “AGL24, SHORT VEGETATIVE PHASE, and APETALA1 Redundantly Control AGAMOUS during Early Stages of Flower Development in Arabidopsis”, The Plant Cell, Volume 18, Issue 6, June 2006, Pages 1373–1382, <https://doi.org/10.1105/tpc.106.041798>

Appendix A: Figures

Diagram

Description automatically generated

Diagram

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart

Description automatically generated with medium confidence

Timeline

Description automatically generated