Predicting Panicle Traits from QTLs in *Oryza sativa*

Kavya Puranam[1], Prof. Elhan Ersoz[2]

Department of Crop Sciences, College of ACES, University of Illinois at Urbana-Champaign

INTRODUCTION

Oryza sativa, commonly known as Asian rice, is a staple crop which feeds billions of people globally. As a result, farmers and breeders value the study of the selection of traits to increase yield and grain quality.

Rice panicle architecture is a key target of selection for yield and grain quality. Panicle traits of certain sizes such as compact can maximize grain production and harvest.

Panicle phenotypes are hard to genetically select for as they are confounded during genetic mapping due to correlation with flowering and subpopulation structure and therefore, are hard for breeders to optimize panicle structure.

The original study conducted uses a **GWAS** (Genome Wide Association Study) and phenotypic traits such as flowering as a covariate to highlight **pleiotropy** of shared genetic regions such as **QLTs** or **SNPs** over a number of subpopulations of *Oryza sativa.*

AIM

- To enhance rice breeding by investigating genetic loci associated with key agronomic traits – specifically yield and panicle size – and developing predictive models to forecast these traits based on specific Quantitative Trait Loci (**QTLs**) or Single Nucleotide Polymorphisms (**SNPs**).
- This project aims to enhance our understanding of genetic variation of complex traits in rice and contribute further to rice breeding programs and research.
- Utilizing computational and statistical methods can directly impact food security and the development of high-yielding rice varieties tailored to specific subpopulations.

METHOD

Objectives were created to build an understanding of the investigation and finally create the predictive model.

Objective 1. **Identify loci** associated with traits like yield and panicle size

- **Manhattan plot** across the entire dataset is used to visualize significant SNPs linked to traits.
- According to this we can identify significant traits based on the peaks of the plot where they **exceed** the **genome wide significance** (P < 5x 10^-8) as well as the loci (chromosomes) associated with these traits
- Q-Q Plot is used to assess if observed associations are more significant than expected by chance

Objective 2. Compare **subpopulations** (Indica, Japonica, Aus, Ind, Trj) for SNP or QTL significance in traits

- Subset the dataset into separate subpopulations
- **Bar chart** is used to compute the total number of significant SNPs for each subpopulation based on a – p-value threshold greater than 0.05
- **Box plot** is used to visualize the distribution of SNP p-values within each subpopulations
- **Population-Specific Manhattan plots** are generated to visualize unique patterns of significant loci and highlight SNPs that are unique or more prevalent in certain subpopulations

Objective 3. Can we create a model that will successfully **predict** panicle size and yield **traits** based on specific **QTLs** or **SNPs**?

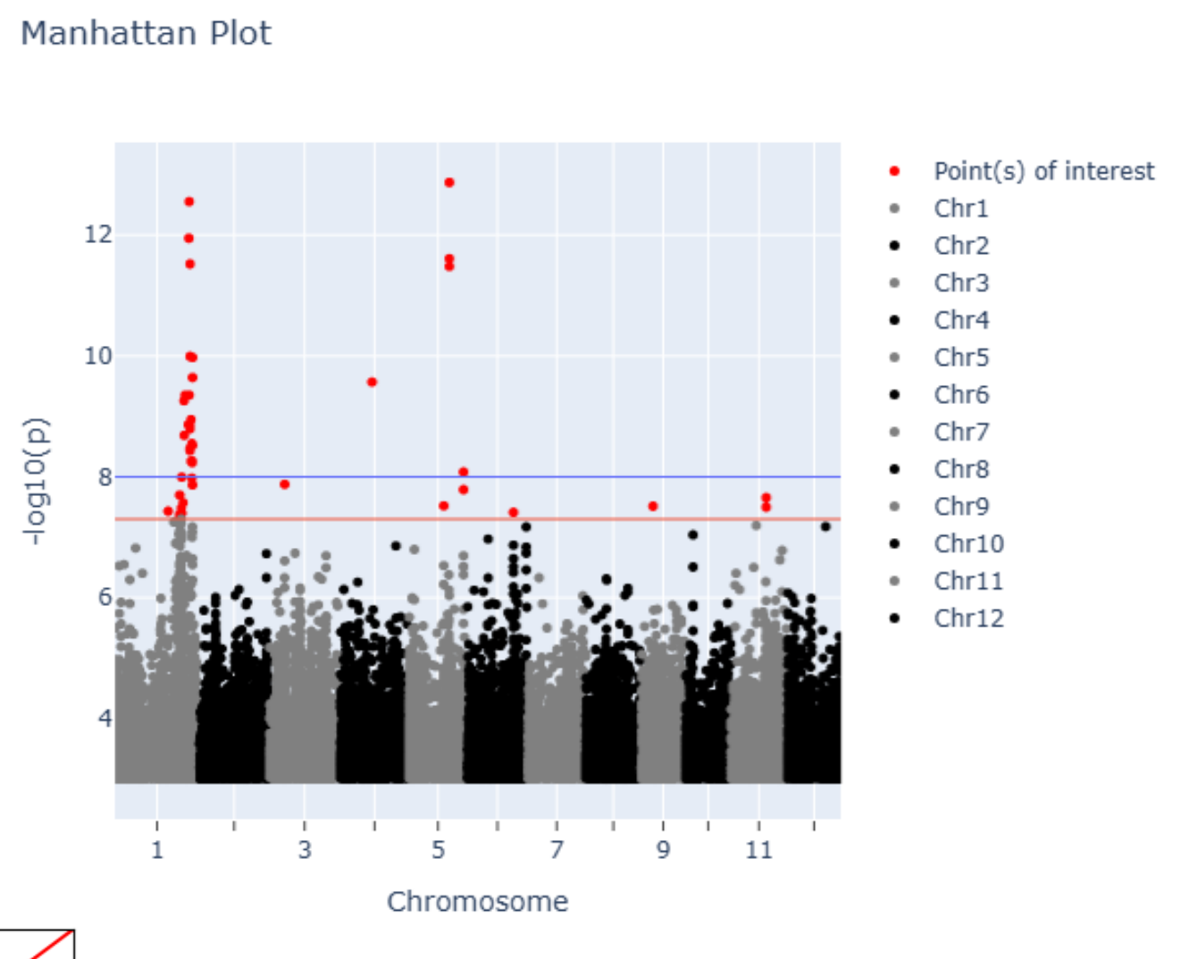
- **Random Forest Classifier** model is developed using significant SNPs as identified in Objective 1 as predictors and preprocessing data using on-hot encoding
- Model trained on a **80%-20% train-test split**
- Perform k-fold cross validation to evaluate model's robustness
- Model performance was assed **using accuracy, precision, and recall**

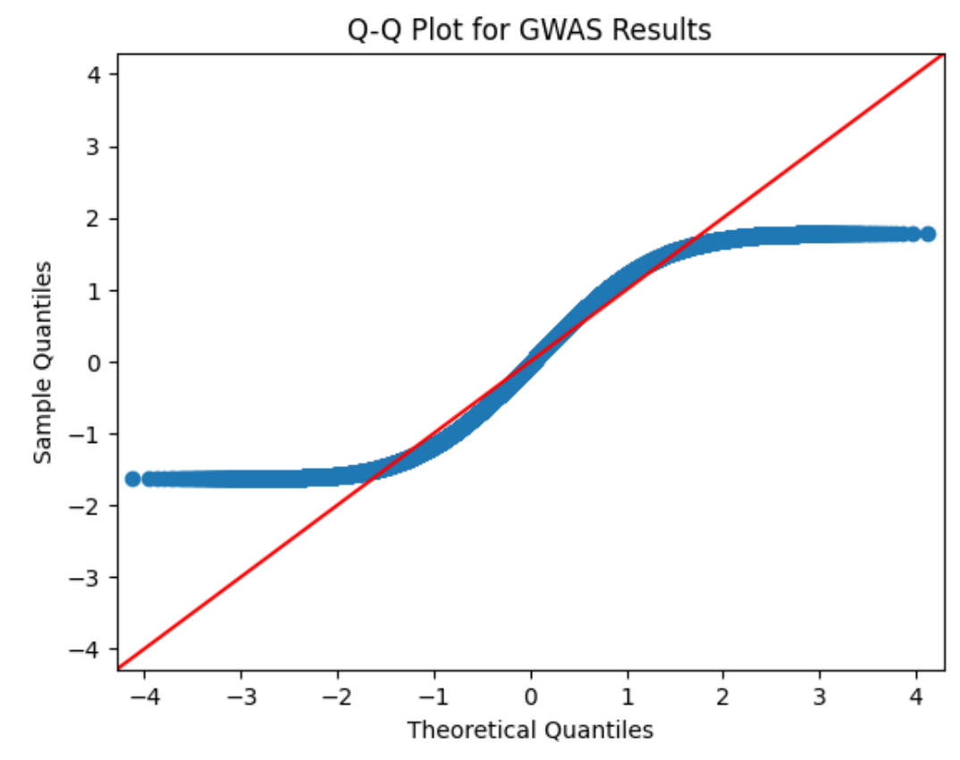
VISUALIZATIONS

Objective 1 Plots

Dataset wide Manhattan Plot

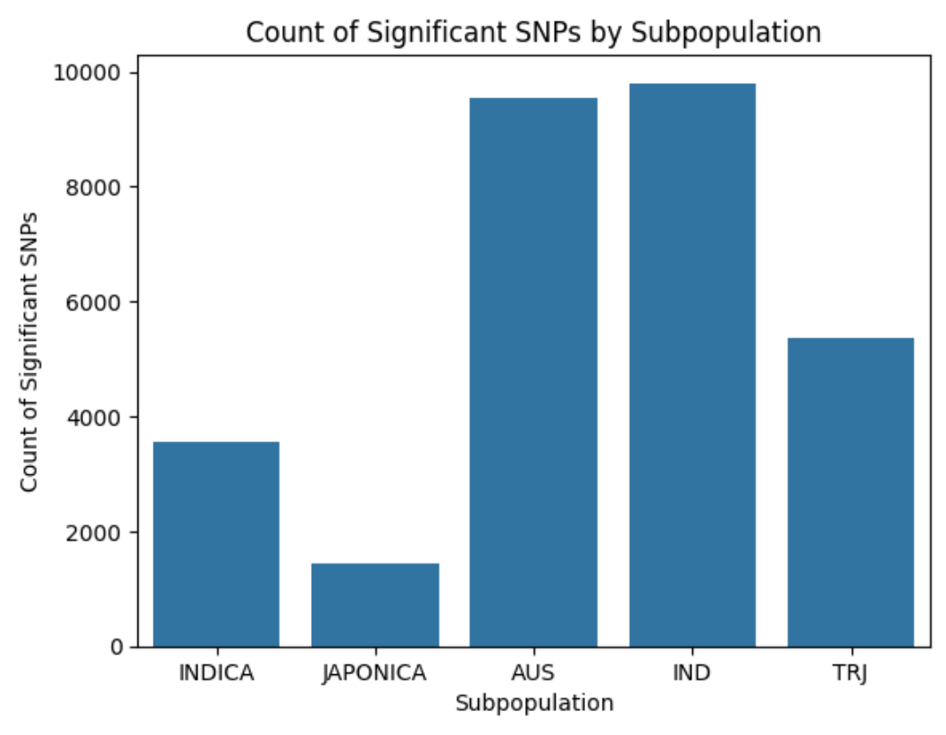
- *Points of interest traits include:*
- *xPBN (panicle branch number)*
- *PH (Plant Height)*
- *BpN (Branches per Node)*
- *BT (Basal Tillers)*
- *PBLin (Primary Branch Length)*
- *totPBL (total Primary Branch Length)*
- *nNL (number of nodes with leaves)*
- *NL (number of leaves)*



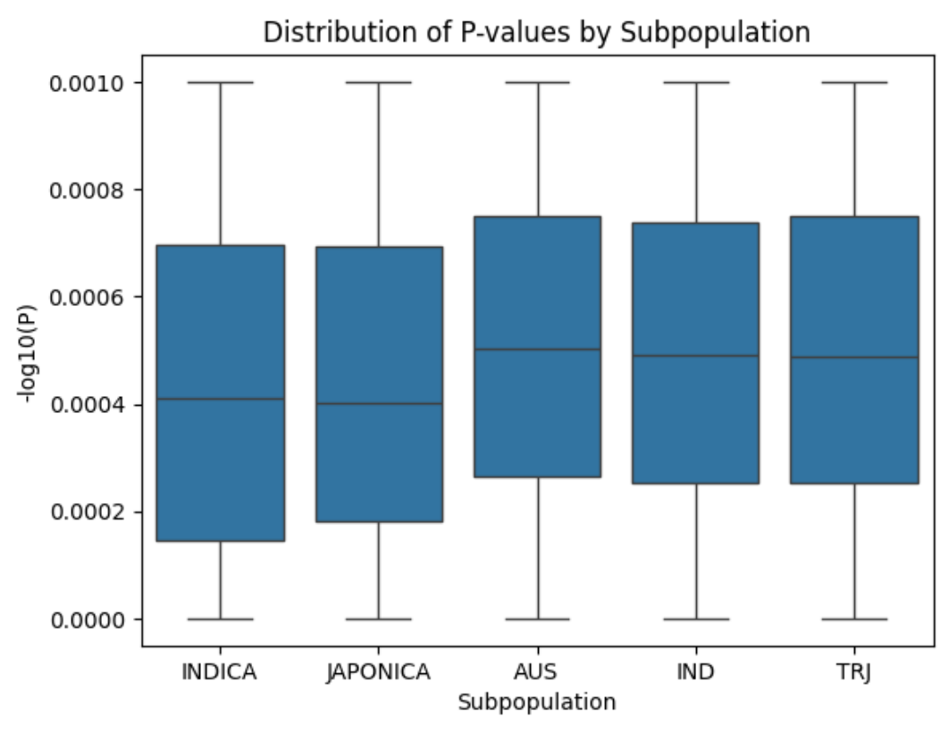


Corresponding Q-Q Plot

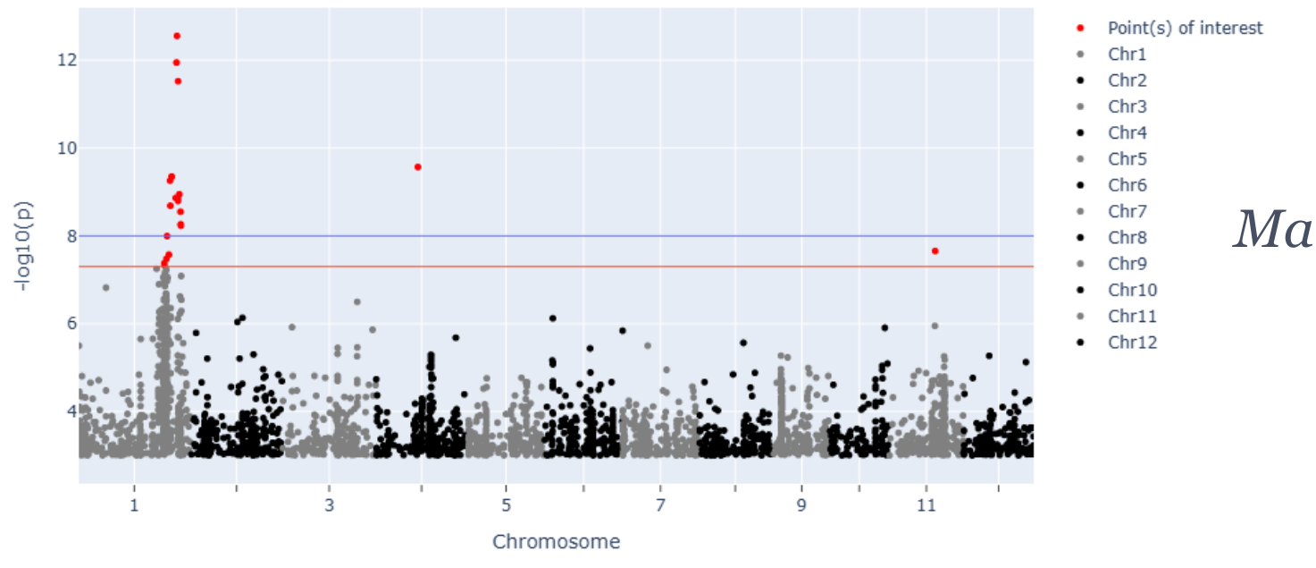
Objective 2 Plots



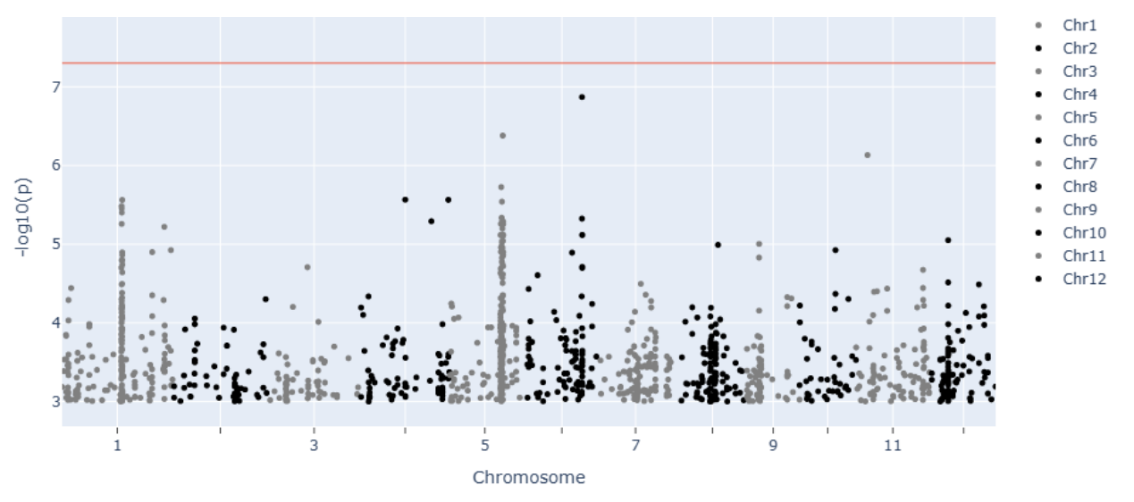
Bar chart depicting counts of significant SNPs by Subpopulation



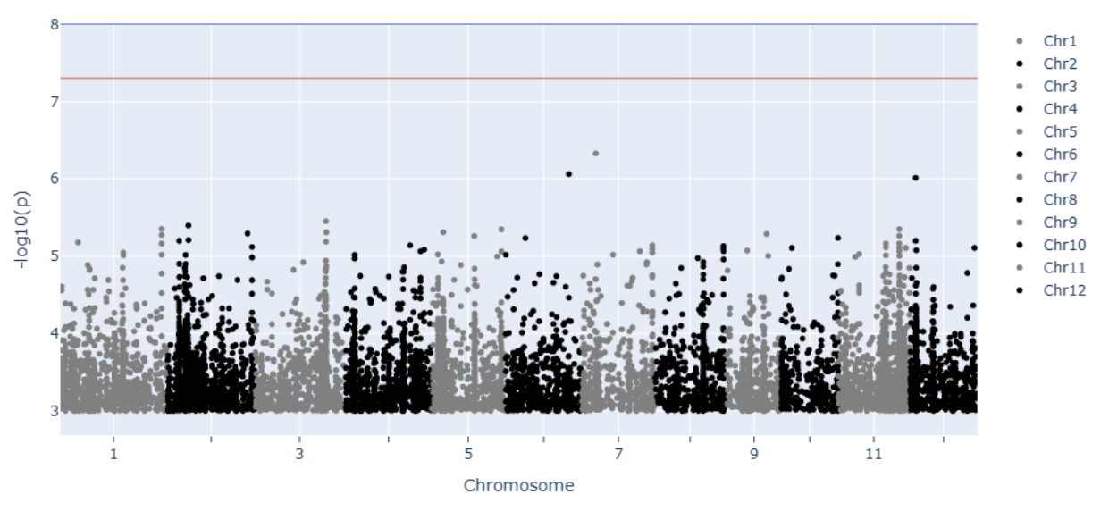
Box plot depicting distribution of p-values by Subpopulation



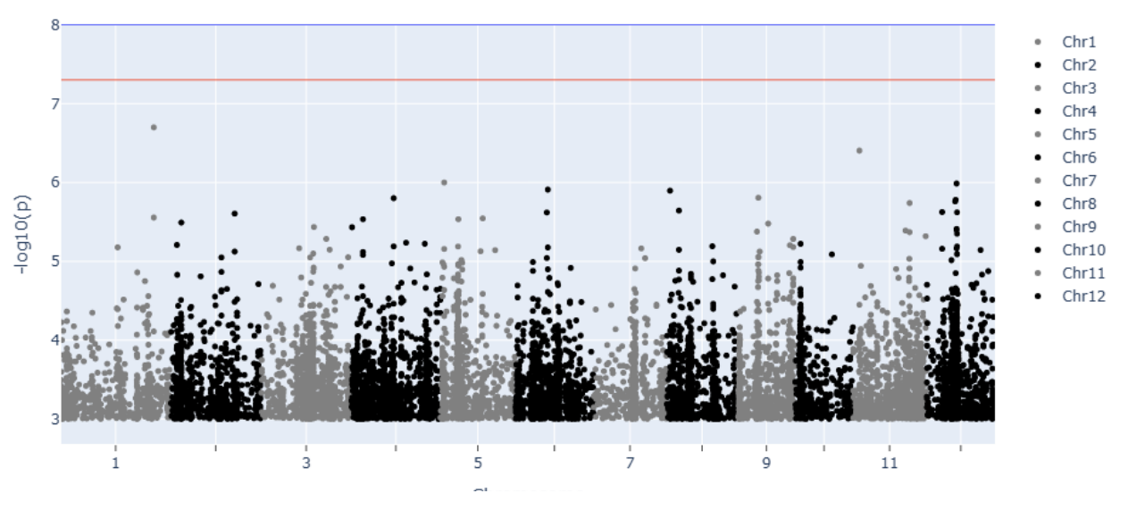
Manhattan plot for Indica



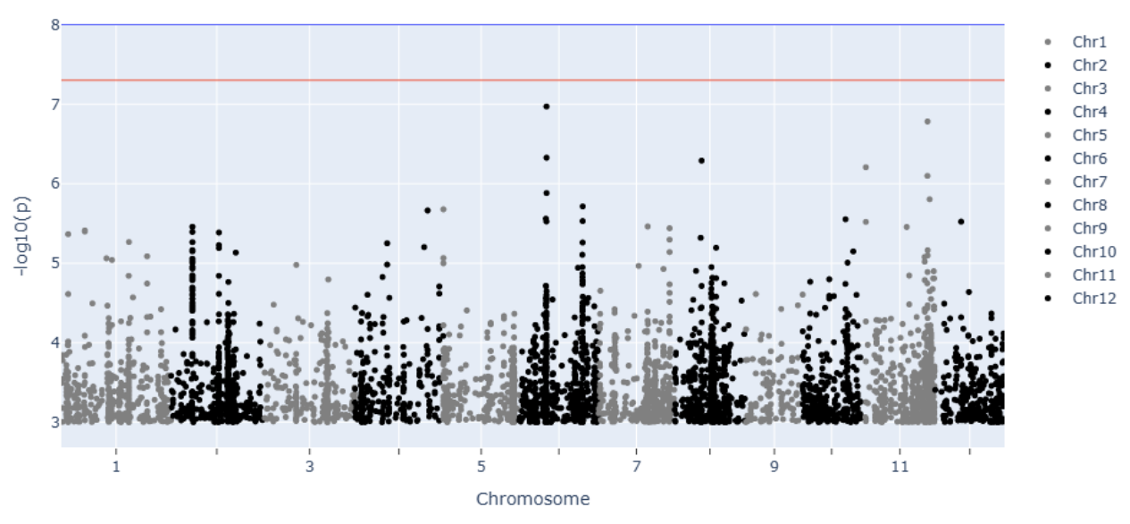
Manhattan plot for Japonica



Manhattan plot for Aus

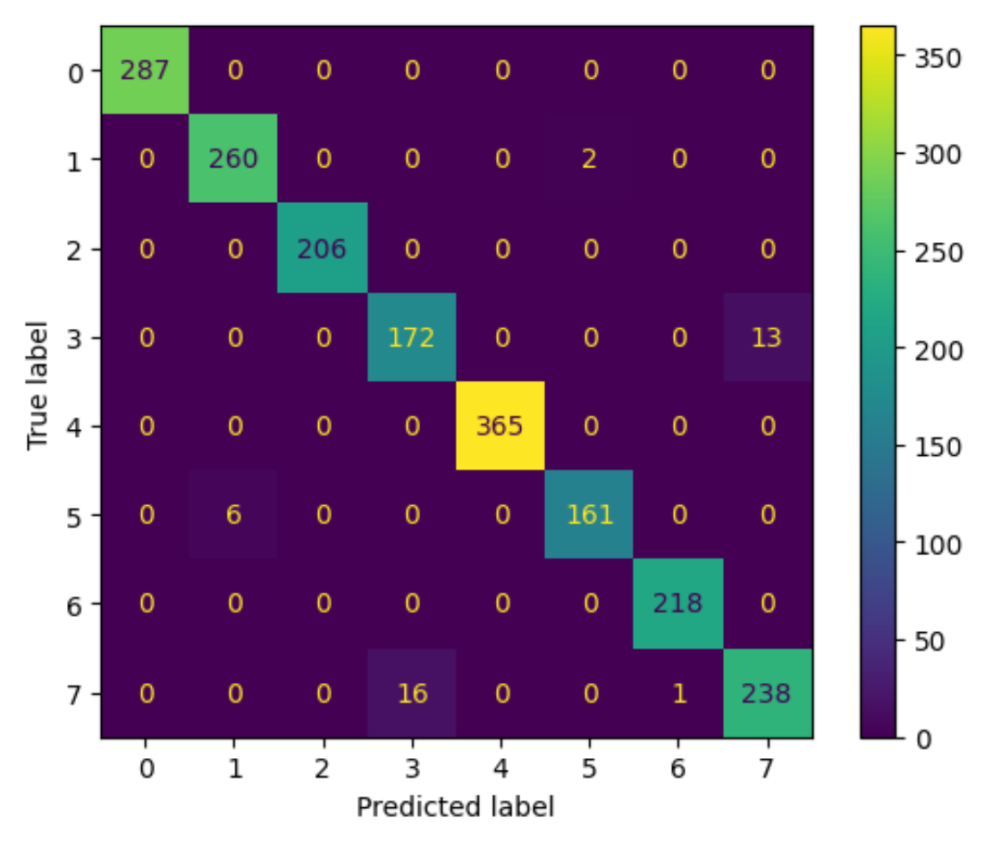


Manhattan plot for Ind



Manhattan plot for Trj

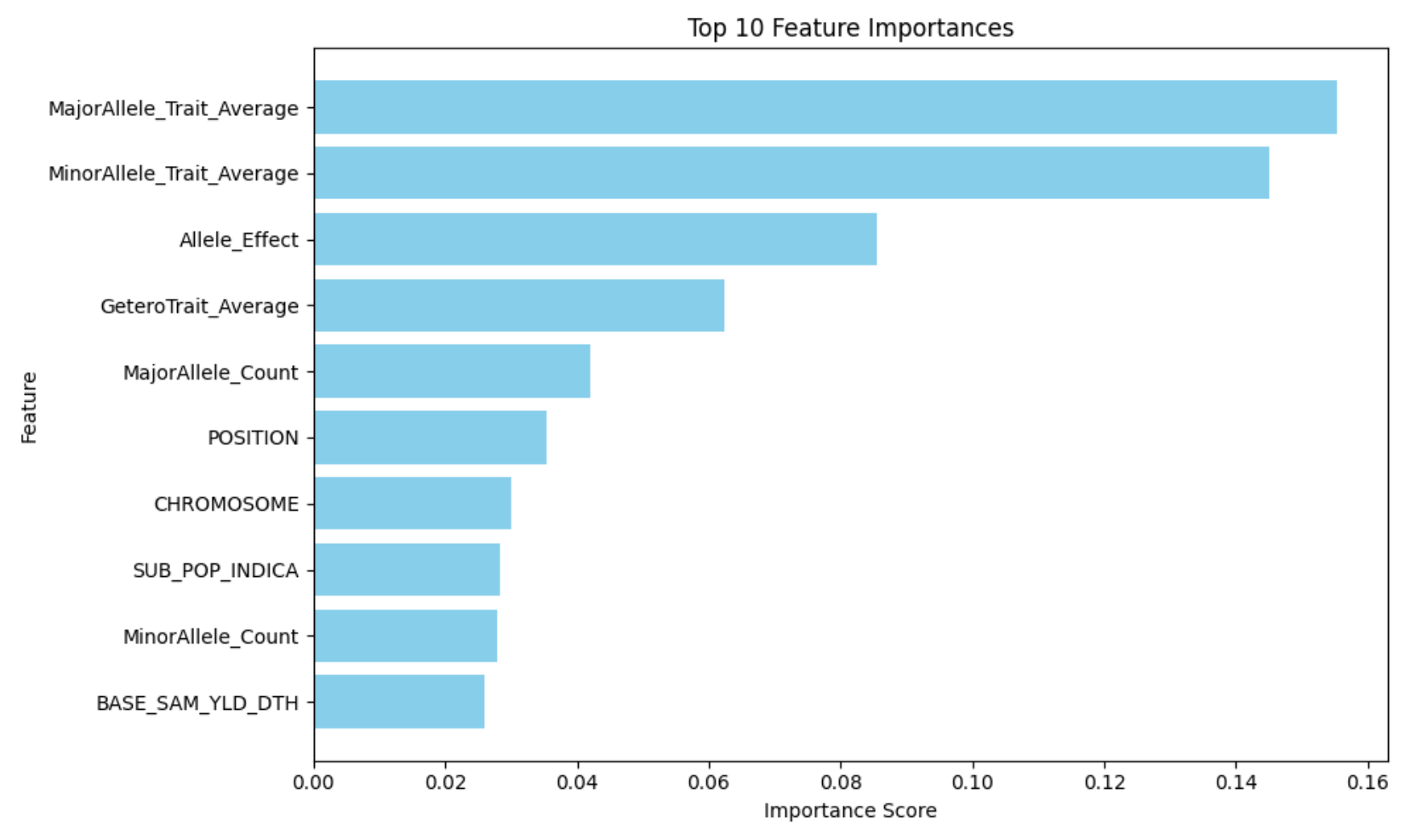
Objective 3 Plots



True label \ Predicted label	0	1	2	3	4	5	6	7
0	287	0	0	0	0	0	0	0
1	0	260	0	0	0	2	0	0
2	0	0	206	0	0	0	0	0
3	0	0	0	172	0	0	0	13
4	0	0	0	0	365	0	0	0
5	0	6	0	0	0	161	0	0
6	0	0	0	0	0	0	218	0
7	0	0	0	16	0	0	1	238

Confusion matrix of Random Forest Classifier after performing k-fold cross validation

Feature importance bar chart based on calculated importance scores



RESULTS

- Manhattan plots of the entire dataset of SNPs show that the specific traits listed that are related to plant yield and panicle size are significant in chromosomes 1, 5, 6, 7, 9 and that these loci are significant as well.
- Corresponding Q-Q plot reveals points deviating from diagonal have smaller p-values than expected as theoretical p-values increases, which correspond to SNPs with strong associations with the trait which supports the Manhattan plot.
- Ind and Aus have the most significant SNPs based on subpopulation subset however, Indica and Japonica have the lower median p-value, although not by much.
- The Manhattan plot for Indica was the only plot SNPs above the threshold but also had a lower count of significant SNPs. This could mean Indica has fewer but stronger trait associations while other subpopulations have weaker associations. This can be due to these traits being strongly selected in Indica as it is a widely cultivated subpopulation.
- We were able to create a Random Forest Regression model with k-fold Cross Validation of 98.05% accuracy and 98.046% precision.
- The feature that influences trait predictions the most is major allele trait average and minor allele trait average based on the feature importance bar chart. This shows that allelic effect is a good indicator of trait value and is good to know for breeding purposes.

DISCUSSION

With this work, researchers are able to leverage predictive models to forecast yield potential based on genetic data and **optimize techniques** so they can handle complex genetic datasets with **fewer resources**.

This research is impactful because it leads to the development of **high-yield rice varieties** and enable breeding strategies targeted towards specific subpopulations most prevalent genetic markers.

Future extensions can be **using time-series phenotypic data** to predict yield and panicle traits over different growth stages as well as creating a **publicly accessible tool** for farmers or breeders globally to access this technology benefiting global rice research and advancements in breeding.

ACKNOWLEDGEMENTS

Research reported in this poster was supported by Prof. Elhan Ersoz. Dataset was provided by a previous study done by Samuel Crowell, Pavel Korniliev, Alexandre Falcão, et al. The table for the dataset is from the Rice Diversity Project Data Sets under the Panicle Architecture results.

Crowell, S., Korniliev, P., Falcão, A. et al. Genome-wide association and high-resolution phenotyping link Oryza sativa panicle traits to numerous trait-specific QTL clusters. Nat Commun 7, 10527 (2016). https://doi.org/10.1038/ncomms10527

College of Agricultural, Consumer & Environmental Sciences