Predictive Analysis of The Genetic Architecture of Panicle Traits in Oryza sativa

Priority Area(s): Genome-wide association studies, Quantitative trait loci, Plant genetics, Data visualization, Statistical modeling, Predictive modeling

Overview

Oryza sativa, commonly known as Asian rice, is the foundational food for many people worldwide, directly impacting access to food for those who consume it. While many crops are produced for feed or fuel, an astounding 95% of rice production is used for human consumption [1]. 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, intermediate, or open can maximize grain production [4]. Panicle phenotypes are hard to genetically select for as they are confounded during genetic mapping due to correlation with flowering and subpopulation structure resulting in difficulty for breeders to optimize panicle structure.

Utilizing the results dataset from a study [2] conducted by a Genome-Wide Association Study (GWAS) using phenotypic traits such as flowering as a covariate, we can highlight pleiotropy of shared genetic regions such as Quantitative Trait Loci (QTLs) or Single Nucleotide Polymorphisms (SNPs). To investigate and demonstrate the relationship between genetic loci and panicle traits we display the results of the GWAS using Manhattan Plots and a predictive model using a Random Forest Classifier with k-fold cross-validation. Manhattan plots are a standard and effective visualization for GWAS to identify genomic regions associated with traits of interest. When trained on predictors, predictive models such as the Random Forest Classifier can help predict desired traits based on genetic loci (QTLs and SNPs). This project is timely given the recent advances in statistical modeling, increased availability of datasets, and growing concerns over food security. Understanding how specific genetic loci result in desired traits in *Oryza sativa* can optimize techniques to handle genetic datasets with fewer resources and lead to more targeted breeding strategies.

Goals and Objectives

Goals:

1. Enhancing 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 QTLs or SNPs.
2. Increasing understanding of genetic variation of complex traits in rice and contribute further to rice breeding programs and research
3. Utilizing computational and statistical methods to directly impact food security and the development of high-yielding rice varieties tailored to specific subpopulations

Objectives:

The overall objective is to analyze GWAS data with statistical methods to predict phenotypic outcomes based on genomic data using specific breeding strategies.

- Manhattan plots are used for standard effective visualization of Genome-Wide Association Studies
  - Evaluating effects of individual and combined SNPs on panicle traits by identifying significant loci where SNPs exceed genome-wide significance threshold
  - Highlights candidate regions that may influence traits like panicle size and yield to enable targeted investigation into their genetic basis
- Random Forest Classifier is used to predict categorical values such as panicle size and yield traits as a function of SNPs and genetic location
  - Understanding trait genetic architecture where we can identify which SNPs and alleles contribute the most to traits like panicle size and yield
  - Accelerating marker-assisted breeding in the future by predicting trait outcomes without needing to conduct extensive trials to focus on loci that are most likely to produce desirable traits

Research Plan (task-based with timeline)

Week 1	*Week 2*	*Week 3*	*Week 4*	*Week 5*	*Outcomes*
Task 1: Data Preprocessing (1 week)					
🟦🟦🟦🟦🟦🟦🟦					Data preprocessing and removing null values from the table
Task 2: Identify loci associated with traits such as yield and panicle size					
	🟦🟦🟦🟦🟦🟦🟦				Generate Manhattan plots related to overall GWAS and isolate significant loci and traits
Task 3: Compare subpopulations (Indica, Japonica, Aus, Ind, Trj) for SNP or QTL significance in traits					
		🟦🟦🟦🟦🟦🟦🟦			Generate Manhattan plots related to subpopulation-specific GWAS and isolate significant traits and loci
Task 4: Creating a model that will successfully predict panicle size and yield traits based on specific QTLs or SNPs					
			🟦🟦🟦🟦🟦🟦🟦	🟦🟦🟦🟦🟦🟦🟦	Build a predictive model based on Loci to predict significant traits assess model and find significant predictors

Task 1: Data Preprocessing (1 week)

This task will be dedicated to understanding the data and setting up the environment to explore the statistics of the data. It is designed to be as extensive as time permitted, because of its emphasis on statistical analysis. The main goal, however, is to be able to create a predictive model implementation to predict traits based on QTLs.

1.1 Collect QTL results Dataset

Data for QTLs and traits are sourced from the original study Genome-wide association and high-resolution phenotyping link Oryza sativa panicle traits to numerous trait-specific QTL clusters from Supplementary Table 5 containing a list of QTL results from a GWAS sourced from The Rice Diversity Project. The dataset explores the pleiotropy of shared genetic regions that affect multiple traits.

1.2 Preprocess Data Points

In the dataset containing 23 columns and 53,280 rows, we can expect there to be invalid or unusable data. Outliers will be removed to have consistency in the analysis I will prepare. The data will be downloaded into a Python notebook and checked for outliers in the dataset by dropping null values. Statistical methods such as finding a Z Score or using the 1.5xIQR rule for identifying outliers will be used as well.

Task 2: Identify loci associated with traits such as yield and panicle size (1 week)

2.1 Build visualizations to understand and investigate data

Create Manhattan plots across the entire dataset 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

2.2 Evaluate accuracy with Statistical analysis

Q-Q Plot is used to assess if observed associations are more significant than expected by chance

Task 3: Compare subpopulations (Indica, Japonica, Aus, Ind, Trj) for SNP or QTL significance in traits (1 week)

3.1 Subset the dataset into separate subpopulations

Each subset will be used to create visualizations of QTLs to understand and investigate if certain SNPs are significant in their subpopulations.

3.2 Statistical Models for Data Discovery

A 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. A box plot is used to visualize the distribution of SNP p-values within each subpopulation.

3.2 Build Visualizations for each subpopulation to compare the significance of QTLs for traits

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

Task 4: Creating a model that will successfully predict panicle size and yield traits based on specific QTLs or SNPs (2 weeks)

4.1 Developing model

A Random Forest Classifier model is developed using significant SNPs as identified in the second task as predictors and preprocessing data using on-hot encoding. The model uses an 80%-20% train-test split.

4.2 Performing K-fold Cross Validation

Perform k-fold cross-validation to evaluate the model's robustness

4.3 Assessing Accuracy of Model

Model performance will be assessed using accuracy, precision, and recall.

Evaluation Metrics

Milestone 1: Data Preprocessing Completion (Week 1)

- QTL dataset successfully imported with 23 columns and 53,280 rows verified
- Data cleaning was completed with outlier removal using Z-Score and 1.5xIQR methods
- Null values are systematically identified and handled

Milestone 2: Trait-Associated Loci Identification (Week 2)

- Manhattan plots were generated successfully, highlighting significant SNPs
- Genome-wide significant loci identified (P < 5x 10^-8)
- Q-Q plot created to validate the statistical significance of associations
- Comprehensive documentation of trait-associated chromosomal regions

Milestone 3: Subpopulation Analysis (Week 3)

- Dataset successfully segmented into Indica, Japonica, Aus, Ind, and Trj subpopulations
- Bar chart completed showing significant SNPs per subpopulation
- Box plots generated displaying p-value distributions across subpopulations
- Population-specific Manhattan plots were created, revealing unique genetic patterns

Milestone 4: Predictive Modeling Completion (Weeks 4-5)

- Random Forest Classifier developed with 80-20 train-test split
- K-fold cross-validation performed to assess model robustness
- Model performance metrics calculated:
    - Accuracy score
    - Precision for trait prediction
    - Comprehensive model interpretation and feature importance analysis completed

Anticipated Outcomes & Impacts

1. **Technical Contributions to Predictive Modeling in Agriculture**

    - **Experience:** Explore and implement predictive modeling techniques for forecasting rice yield potential based on genetic markers and environmental factors. Develop models to optimize strategies for handling complex genetic datasets with limited resources.

    - **Learning outcome:** Acquire skills in applying predictive analytics to address challenges in agriculture, contributing to the development of high-yield rice varieties and breeding strategies tailored to specific subpopulations.

2. **Advanced Agricultural Sciences for Food Security**

    - **Experience:** Engage in research prioritizing crops critical to human consumption, emphasizing global food security and the need for targeted breeding efforts.

- **Learning outcome:** Develop an understanding of the intersection between agricultural research, global food systems, and the importance of prioritizing human-consumption crops in scientific innovation.

3. **Promoting Technological Education in Agriculture**

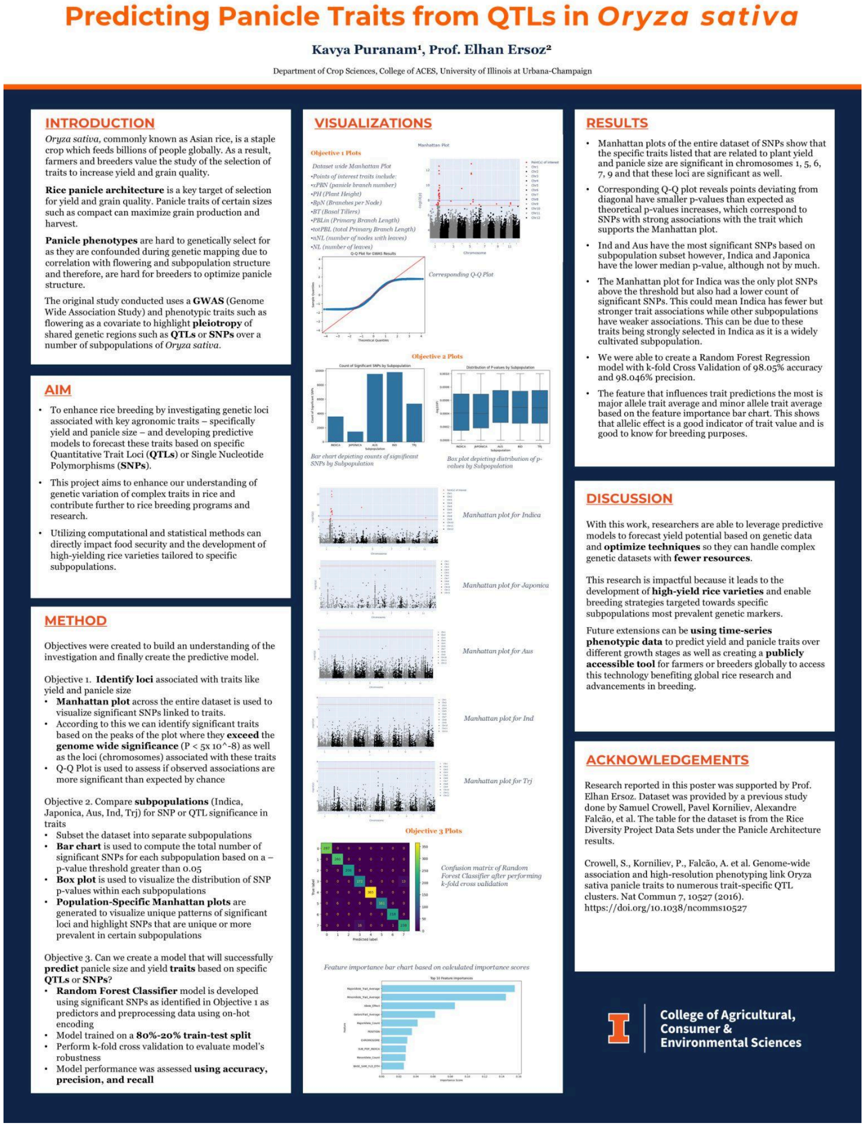
- **Experience:** Share insights on how technological advancements can improve crop outcomes and food security. Present on the importance of data science and statistical modeling in agriculture and genetics.

- **Learning outcome:** Cultivate the ability to communicate complex scientific ideas to non-specialists, fostering a broader appreciation for the role of technology in solving global challenges.

Follow-up Funding/Follow-up Research/Growth Plan

This project is the foundation of a larger, self-sustaining research initiative focused on using predictive modeling and data-driven decision-making for crop breeding. To enhance this research, I aim to study **genotype-by-environment (G×E) interactions** to better understand how environmental factors influence the relationship between SNPs and key traits. This would involve integrating environmental variables such as temperature, rainfall, soil quality, sunlight, latitude, and longitude into predictive models. By developing G×E interaction models, I seek to identify SNPs that are either **robust across diverse environments** or **responsive to specific conditions**. Additionally, I plan to explore the use of **Geographically Weighted Regression (GWR)** to analyze the spatial relationships between SNPs and traits, detect patterns in the data, and forecast SNP effects at different geographic locations.

I plan to expand this research into a graduate school thesis, where I will have the resources and mentorship to further develop these ideas. Over the next six months, I will focus on building a succinct thesis proposal and identifying avenues for data collection. Next fall, I will pursue funding opportunities such as the **NSF Plant Genome Research Program (Grant Number 1026555)** and the **NSF Graduate Research Fellowship Program**. In the long term, I envision streamlining this process to apply the research framework not only to rice but also to other crops, ensuring its accessibility to farmers. This would make predictive modeling a practical tool for addressing global agricultural challenges, including crop yield improvement and climate resilience.



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 **QTLs** 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 assessed **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

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

Google Colab for Analysis and Notes

REFERENCES

1.

Rice | Description, History, Cultivation, & Uses | Britannica. 3 Dec 2024 [cited 16 Dec 2024]. Available: https://www.britannica.com/plant/rice

2.

Crowell S, Korniliev P, Falcão A, Ismail A, Gregorio G, Mezey J, et al. Genome-wide association and high-resolution phenotyping link Oryza sativa panicle traits to numerous trait-specific QTL clusters. Nat Commun. 2016;7: 10527. doi:10.1038/ncomms10527

3.

Importance of Rice. [cited 16 Dec 2024]. Available: http://www.knowledgebank.irri.org/ericeproduction/Importance_of_Rice.htm

4.

Rost TL. Rice Anatomy Stems Panicle Formation. In: UC Davis [Internet]. 1997 [cited 16 Dec 2024]. Available: https://labs.plb.ucdavis.edu/rost/rice/Stems/panicle.html

5.

McCouch SR, Wright MH, Tung C-W, Maron LG, McNally KL, Fitzgerald M, et al. Open access resources for genome-wide association mapping in rice. Nat Commun. 2016;7: 10532. doi:10.1038/ncomms10532