ĐẠI HỌC QUỐC GIA HÀ NỘI
ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN – CƠ – TIN

BÁO CÁO THỰC TẬP

Analyzing high-dimensional GWAS data to infer SNPs associated with rice's phenotypes.

Giảng viên phụ trách thực tập
TS. VŨ TIẾN DŨNG
Giảng viên hướng dẫn
PGS. TS. LÊ ĐỨC HẬU
Học Viên
TẠ VĂN NHÂN

Lớp **Khoa học dữ liệu**
Khóa **1**

LỜI CẢM ƠN

Trước hết, tôi xin được tỏ lòng biết ơn và gửi lời cám ơn chân thành đến **PGS.TS. Lê Đức Hậu**, người trực tiếp hướng dẫn thực tập tốt nghiệp, đã tận tình chỉ bảo và hướng dẫn tôi tìm ra hướng nghiên cứu, tiếp cận thực tế, tìm kiếm tài liệu, xử lý và phân tích số liệu, giải quyết vấn đề. Tôi cũng xin được gửi lời cảm ơn đến **TS. Vũ Tiến Dũng, TS. Nguyễn Thị Bích Thủy** phụ trách công việc thực tập của học viên, và **PGS.TS. Nguyễn Thị Hồng Minh** đã giới thiệu tôi đến thực tập tại phòng Tin sinh học, viện Dữ liệu lớn, tập đoàn Vingroup. Đồng thời tôi muốn gửi lời cảm ơn đến bạn **Trần Thế Hùng**, chuyên viên tin sinh học đã cùng tham gia làm việc với tôi, nhờ đó tôi mới có thể hoàn thành khóa thực tập tốt nghiệp.

Ngoài ra, trong quá trình học tập, nghiên cứu và thực hiện đề tài tôi còn nhận được nhiều sự quan tâm, góp ý, hỗ trợ quý báu của quý thầy cô, đồng nghiệp, bạn bè và người thân.

Tôi xin bày tỏ lòng biết ơn sâu sắc đến:

- Quý thầy cô Khoa Toán-Cơ-Tin và quý thầy cô Khoa Sau đại học trường Đại học Khoa học tự nhiên, Đại học Quốc gia Hà Nội đã tạo điều kiện để tôi thực tập tại nơi phù hợp với chuyên môn của mình.
- Ban lãnh đạo và các đồng nghiệp tại phòng Tin sinh học, viện Dữ liệu lớn thuộc tập đoàn Vingroup luôn động viên, hỗ trợ tôi trong quá trình học tập và nghiên cứu.

Học Viên
Tạ Văn Nhân.

NHẬN XÉT KẾT QUẢ THỰC TẬP

(Cho đơn vị học viên thực tập)

Họ tên học viên thực tập: **Tạ Văn Nhân**
Cán bộ hướng dẫn thực tập: **PGS.TS. Lê Đức Hậu**
Phòng: Tin sinh học, viện Dữ liệu lớn, tập đoàn Vingroup
Sau thời gian học viên *Tạ Văn Nhân* thực tập tại đơn vị, chúng tôi có những nhận xét sau:

Xác nhận của đơn vị

……., ngày……, tháng……, năm 2020

Cán bộ hướng dẫn

NHẬN XÉT KẾT QUẢ THỰC TẬP

(Cho cán bộ phụ trách thực tập)

Họ tên học viên thực tập: **Tạ Văn Nhân**
Cán bộ phụ trách thực tập: **TS. Vũ Tiến Dũng**
Khoa: Toán-Cơ-Tin, trường Đại học Khoa học tự nhiên, Đại học Quốc gia Hà Nội.
Sau thời gian học viên *Tạ Văn Nhân* thực tập, tôi có những nhận xét sau:

……., ngày……, tháng……, năm 2020

Cán bộ phụ trách thực tập

Contents

Task Division

Datasets [Hung]
Data preprocess [Hung]
Association Analysis [Hung]
Regressors [Nhan]
Performance Evaluation [Nhan]
Results

- Individual SNP association analysis [Hung]
- Combinatorial SNP association analysis for individual traits (either GW or T2F) [Hung & Nhan]
- Combinatorial SNP association analysis for both traits (GW and T2F) [Hung & Nhan]
- Conclusions [All]

1. Introduction

Rice is one of the most important staple food in the world and is the main food crop in Asia. Along with the increase in the global population, rice consumption will also increase proportionally [1]. However, the production of rice is limited by the area of soil available. The adverse effect of climate change also contributes to the decrease in land and water resource [2]. Developing rice crop is important for the world population, especially in the face of climate change and a growing population. Many breeding methods are currently in used to create high-yielding rice line which can tolerate the harsh everchanging climate and can yield a substantial amount of grains on a limited area of soil [3]. The conventional method of breeding using trait has been successful in creating cultivars that have a dramatic increase in yield [4].

Nowadays, advancement in genomics technology facilitates the usage of genetic information in the breeding scheme. This can substantially improve the desired characteristics of cultivars [5]. One notable method is a marker-assisted selection (MAS) [6]. Markers used in MAS can be classified into two categories: classical markers and DNA markers. The classical marker was used in breeding since long ago with visible traits such as leaf shape, flower colour, grain weight, fruit shape, etc. With the advancement of technology, more information about chromosome shape, protein in the cultivars can be integrated into the breeding scheme. The DNA marker can be very diverse based on different polymorphism-detecting method: southern blotting, polymerase chain reaction, DNA sequencing, etc. [6]. MAS relies on the tight association between the DNA marker and the major gene responsible for the desired trait. The DNA marker based on single nucleotide polymorphism (SNP) has rapidly gained the centre stage of molecular genetics [7]. SNP is a single nucleotide base difference between 2 individuals' DNA sequences. SNPs occur very commonly in the genome of plants and animals. Typically, there is 1 SNP for every 100-300 base pairs of DNA in plants [8]. Therefore, SNPs are very abundant in the genomes, and new technology also allows high-throughput detection of them.

Because of the ever-increasing amount of genetic information generated by new technologies, the main challenge is in the computational method to identify the informative SNPs which can give benefit to the breeding scheme. A popular approach to identify SNPs that can explain a complex trait is a genome-wide association study (GWAS). The concept behind GWAS is to genotype a large number of SNPs distributed across the genome so that the quantitative trait locus (QTL) would be in linkage disequilibrium (LD) with at least a few SNPs [9]. The goal of this process is to identify the action, interaction and precise location of these QTL. Then it can be applied in the breeding selection scheme.

The GWAS approach is not without limitations [10]. GWAS can reliably identify the loci with large effect size on the trait, but its power is diminished if the genetic architecture of the trait is more complex. For example, if the frequency of the genetic variant is very rare, it requires a very large sample size to ascertain its association with the trait with enough power. And even then, the density of genetic markers needs to be very high to capture the SNP in LD with the causative rare genetic variants [11]. On the other hand, if the trait of interest is determined by many genetic variants having small effect size, the p-value of the association test of these causative variants can be relatively non-distinctive from other confounding signals [12]. Therefore, the application of machine learning methods in prediction of trait holds great promise in ameliorating the crop yield [13, 14]. Therefore, we intend to

investigate some machine learning methods in predicting some trait of the rice plant. We reason that, if the SNPs selected from the prediction model can better predict the trait, it will hold a greater role in defining the trait of interest.

Inn this study, we used the dataset of rice plant from the 3000 Rice Genomes Project [15] to explore machine learning methods on choosing the genetic variants that are most likely associated with "grain weight" and "time to flowering" traits of Oryza Sativa (i.e. rice plant). Firstly, we found individual SNPs which are significantly associated with the two traits using the traditional GWAS analysis [16]. Then, we investigated several regressors in machine learning to automatically select SNPs and predict the rice's traits. By assuming the non-linear relationship between the genotype data and the traits, we used two non-linear regressors, i.e., support vector regressor (SVR) [17]and random forest regressor (RFR) [18], on significant SNPs. Experimental results showed that the models with SNPs filtered by P-value $\leq 5\times10^{-3}$ for the trait "grain weight" (1,323 SNPs) and for the trait "time to flowering" (3,128 SNPs) gain the best accuracy. For the models that used SNPs satisfied the P-value thresholds in both the traits, the threshold of 5×10^{-2} (929 SNPs) achieve the best accuracy. This indicates that those SNPs should be selected for better prediction of both the traits. In addition, the SVR model gives better prediction performance in terms of both mean square error (MSE) and coefficient of determination (R^2) than the RFR model. Besides, by assuming the linear relationship between the genotype data and the traits, we investigated some advanced linear regressors such as Lasso [19] and Elastic Net [20] for predicting individual traits and multi-task Lasso and multi-task Elastic Net for predicting both the traits. Experimental results showed that with SNPs filtered by pruning to exclude SNPs that are in high linkage with each other (43,527 SNPs) and continue to be automatically filtered by the advanced linear regressors, Elastic Net and multi-task Elastic Net pick more SNPs than Lasso and multi-task Lasso for the trait "time to flowering" and for the mixed trait "grain weight" – "time to flowering". However, the models based on Elastic Net have lower prediction performance than the models based on Lasso. Generally, models based on the advanced linear regressors have worse prediction performance compared to models based on the non-linear regressors.

2. Materials and Methods

2.1. Datasets

The dataset in this study is GWAS data of Oryza Sativa from the 3000 Rice Genome project [15], which can be obtained from the following url: https://snp-seek.irri.org/_download.zul. The dataset is maintained by Rice SNP-Seek Database [21]. It contains the trait and genotype of 3,024 accessions of rice cultivars from 89 countries. Genomic DNA was sequenced in the HiSeq2000 platform, had an average depth of 14x, with an average coverage of 94%. Then, the raw reads are aligned to the Nipponbare reference genome, and variant calling pipeline generated approximately 18.9 million SNPs. Then the datasets went through quality control: removing SNPs with excess heterozygous calls, filtering out SNPs with minor allele frequency <0.01 and missing call rate >0.2. Finally, the dataset was LD pruned to obtain the core SNP set (v0.7) containing 404,388 SNPs. illustrates the density of the SNPs in the dataset.

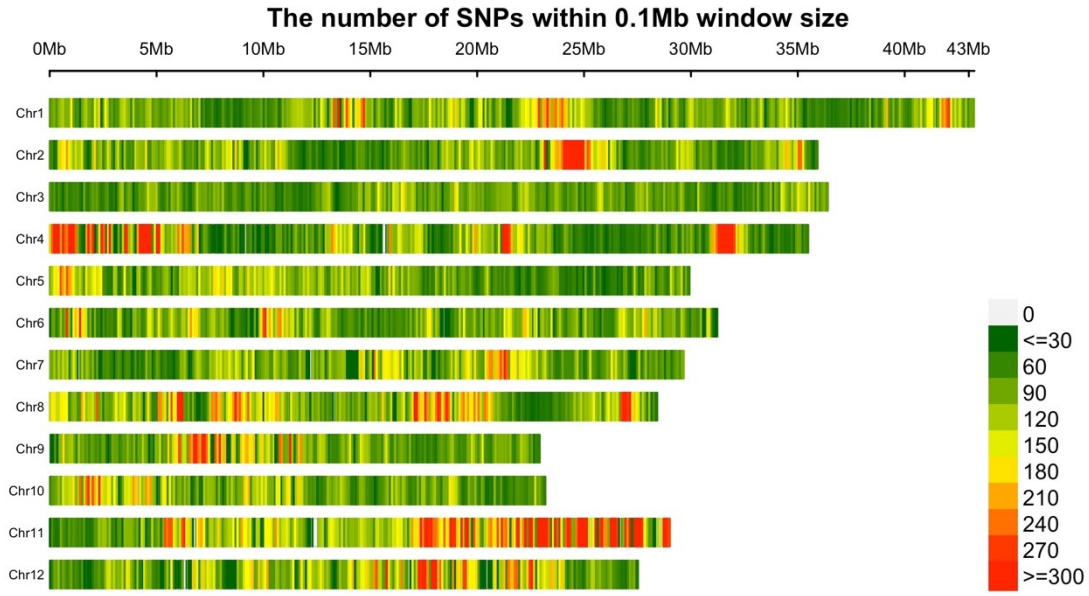

Figure 1: SNPs density. The number of SNPs in every 0.1 million base-pairs distributed on 12 chromosomes of the rice genomes in the dataset.

There are over 50 traits recorded in the dataset with many quantitative traits that are important for agronomy. In this study, we decided to analyze two quantitative traits which have the most impact and have relatively high non-missing trait information in the dataset: "grain weight" and "time to flowering". summarized the characteristics of these two traits.

Table 1: Characteristics of "grain weight" and "time to flowering".

Trait (unit)	**mean ± SD (min-max)**	**Sample size**	**Combination**
Grain weight (gram/100 grains)	2.5 ± 0.52 (1-5)	2260	Union set (2266)
Time to flowering (days)	102.1 ± 24.51 (50-184)	2266	

SD: standard deviation

2.2. Data preprocess

Missing or low-quality data is a major problem for GWAS. Commonly, this is resolved by the imputation of the value of missing data from already available data to improve the power of downstream analysis [22]. In this study, we used LinkImpute [23], a computational tool based on k-nearest neighbour genotype imputation method. This tool could be executed without the physical or genetic map of the species of interest and is suitable for unphased genotype data from heterozygous species.

The distribution of missing genotype per SNPs/individuals were depicted in and . Imputation was carried out by using LinkImpute software on each chromosome of the genotype separately. Then the outputs are joint together using plink 1.90 [24]. Imputation accuracy and corresponding optimized parameters for each chromosome were depicted in .

The parameters reported are k-nearest neighbour based on l SNPs most in LD with the SNP to be imputed.

LD-pruning is a step to select SNPs that are not in strong LD with other SNPs near its. This is conducted by using plink 1.90 with the command --indep-pairwise window_size step_size r2_threshold. In this study, we calculated the LD between the SNPs within the window size of 1000 variants, index the SNP with the highest minor allele frequency and remove any SNPs that have r2 >0.3 with the index SNP. The window will shift forward 10 SNPs after each step. There are 43527 SNPs remained after the LD prune.

We called the allele with the higher frequency the major allele and the one with the lower frequency the minor allele. The genotype value at a specific SNP of an individual consists of two alleles. Each allele can take the value of A, C, G or T. In order to compute with different analysis method, we substitute the genotype with numerical value: “0” for the homozygous genotype of major allele, "2" for the homozygous genotype of minor allele and “1” for heterozygous genotype. This step is performed by the command --recodeA from plink 1.90 [24].

2.3. Association Analysis

The result of GWAS can be confounded due to the genetic background of samples [12]. The fact that related individuals can share both causal and non-causal variants, and that these variants can be in LD with each other, can lead to synthetic association which present itself as false-positive results. A noteworthy method to account for the above artefact has been developed and applied in the field of animal breeding [25]. We used GCTA tool [26] to perform genome-wide complex trait analysis. The fastGWA method [27] integrated into the above tools can control for population stratification by principal components and for relatedness by a sparse genetic relationship matrix.

GWAS was conducted on the subpanel of the dataset which contains 2266 sample with an available record of "grain weight" and "time to flowering". Principal component analysis (PCA) on the genotypic data was conducted using GCTA, four principal components (PC) were used as covariates in downstream analysis. In order to account for the relatedness among samples, we generated a sparse genetic relationship matrix (GRM) from a full-dense GRM at a cut-off value of 0.05. The GRM was used in restricted maximum likelihood analysis; the estimate of Vg is not statistically significant. This is likely because the number of closely related individuals in the sample is not large enough. In this case, linear regression for the association test adjusted for PCs was used.

Then the GWAS results were presented as Manhattan plot and Quantile-quantile plots. Quantile-quantile plot showed quality control. Manhattan plot showed the p-values for each SNP-trait association. We used a genome-wide significant threshold of 5×10^{-4} [28].

2.4. Regression models

Regression analysis is applied to estimate the relationship between a dependent variable (also called output variable, response variable) and independent variables (also called input variables, predictor variables or features). Here, we consider two types of regression, linear regression such as Lasso and Elastic net, non-linear regression such as support vector regression (SVR) with the kernel functions and random forest regression (RFR). Given an

input data with n samples and p variables corresponds to the n×p matrix X. Considering a single-task problem with an output variable corresponds to the n×1 matrix y, the relationship between X and y is approximately linear and be represented by the equation with an p×1 parameter matrix w, an intercept parameter b and a tolerance ε as:

$$y = Xw + b + \varepsilon. \tag{1}$$

Since the trait of interest is a continuous variable, we used regression models to predict the trait (dependent variables) based on the genotype from 404,388 SNPs (independent variables).

2.4.1 Advanced linear regressors

To find the best parameters, the least-squares regression is known to minimize the sum of the squared errors:

$$min_w \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - (x_i w + b)\right)^2 \tag{2}$$

where y_i gets real value, x_i is a 1×p matrix.

For a multi-task problem with d output variables, we replace the matrix y with the n×d matrix Y, the matrix w by the p×d matrix W.

For high dimensional data, when p is very large compared to n, the least-squares problem may not be solved. To improve the algorithm, l1 norm and l2 norm for regularization are added in different ways in the methods of Lasso, Elastic Net, multi-task Lasso and multi-task Elastic Net.

$$Lasso: min_w \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - (x_i w + b)\right)^2 + \alpha\sum_{j=1}^{p}|w_j|. \tag{3}$$

$$Multitask\ Lasso: min_W \frac{1}{2n}\sum_{i=1}^{n}(Y_i - (x_i W + b))^2 + \alpha\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{p} w_{ij}^2} \tag{4}$$

$$Elastic\ Net: min_w \frac{1}{2n}\sum_{i=1}^{n}(y_i - (x_i w + b))^2 + \alpha\rho\sum_{j=1}^{p}|w_j| + \frac{\alpha(1-\rho)}{2}\sum_{j=1}^{p} w_j^2. \tag{5}$$

$$Multitask\ Elastic\ Net: min_W \frac{1}{2n}\sum_{i=1}^{n}(Y_i - (x_i W + b))^2 + \alpha\rho\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{p} w_{ij}^2} + \frac{\alpha(1-\rho)}{2}\sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}^2. \tag{6}$$

where, α≥0 is a complexity parameter, 0≤ρ≤1 controls the convex combination of l1 and l2 norm for regularization.

Using these methods, we can also select important variables that their coefficients are different from zero.

2.4.2 Non-linear regressors

Support vector regression (SVR) is built with an idea similar to a soft margin algorithm. Slack variables are added to the model to deal with infeasible constraints. Accordingly, the regression tolerance ε can be up to the value of the slack variables ξ and ξ*. For C is a regularization parameter, we have the following optimization problem:

$$min_{w,b,\xi}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*).$$
$$subject\ to \begin{cases} y_i - (x_i \mathrm{w} + b) \le \varepsilon + \xi_i;\ i = 1,2,\dots,n \\ x_i \mathrm{w} + b - y_i \le \varepsilon + \xi_i^*;\ i = 1,2,\dots,n \\ \xi_i \ge 0;\ i = 1,2,\dots,n \\ \xi_i^* \ge 0;\ i = 1,2,\dots,n \end{cases}. \quad (7)$$

Random forest regression is a model that combines multiple regression trees using a random selection of samples and features to avoid overfitting of the individual trees. Random forest regression generates a prediction equal to the average of predictions of the regression trees, thus reduces bias and variance.

2.5. Experimental settings

2.5.1. Performance evaluation

For selection and comparison among the models, we evaluate the performance of the models based on the test set separated from the original data since the set does not join to the training process, and the remaining one participates in the process using grid search with cross-validation (Grid Search CV). When grid search estimates the best hyper-parameters of the model, the k-fold cross-validation helps reduce overfitting on the training set.

In fact, the data is randomly divided into a testing set and training set with ratios of 0.2 and 0.8, respectively. Moreover, one divides the training data into ten folds, and take turns to choose one fold as a test set, the other nine folds as a training set for each model. For SVM and RF, model evaluation metrics of the training set are averaged from MSE of the models. Therefore, the best model is selected in terms of the averaged MSE. For Lasso and Elastic Net with cross-validation, the training process will stop when one reaches the maximum number of iterations or the tolerance for the optimization. Finally, the best models of the different methods are compared to each other by MSE and R-squared on the test set. Figure 2 illustrates the selection and the comparison among the models.

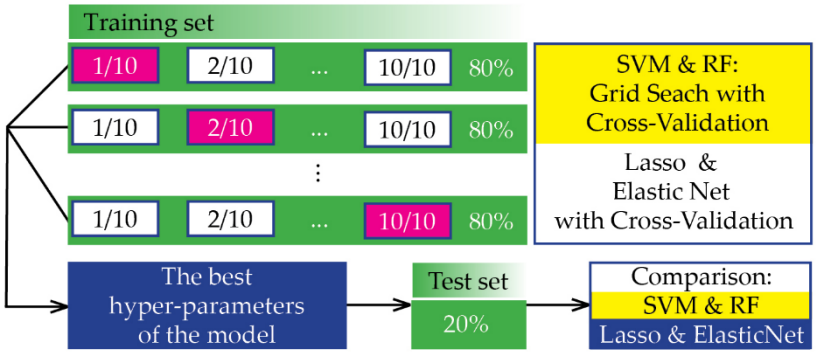

Figure 2: The method of model comparison

2.5.2. Evaluation measures

Mean squared error (MSE) is the average squared difference between the predictor values and observed values. MSE is always greater or equal to zero. The smaller the MSE, the closer the predictor variables are to the observed variables.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{obs}^{(i)} - y_{pred}^{(i)})^2. \quad (8)$$

Coefficient of determination (R^2) represents the proportion of variance for response variables explained by explanatory variables. It is mean that if r-square of the model is 0.7, then 70% observed variation could be explained by input variables.

$$R^2 = 1 - \frac{explained\ variance}{total\ variance} = 1 - \frac{\sum_{i=1}^{m}\left(y^{(i)} - y_{pred}^{(i)}\right)^2}{\sum_{i=1}^{m}\left(y^{(i)} - y_{mean}^{(i)}\right)^2}. \quad (9)$$

R-squared may get a negative value if the fit of the chosen model is worse than the horizontal mean line.

3. Results

3.1 Individual SNP association analysis

The percentage of explained variance for the first ten PCs were illustrated in and the grouping of samples based on PC 1 and 2 was illustrated in . The results of GWAS for "grain weight" and "time to flowering" traits were illustrated in with the log10-transformed of p-values as the y axis and the chromosome number as the x axis. The quantile-quantile plot of these two traits was illustrated in

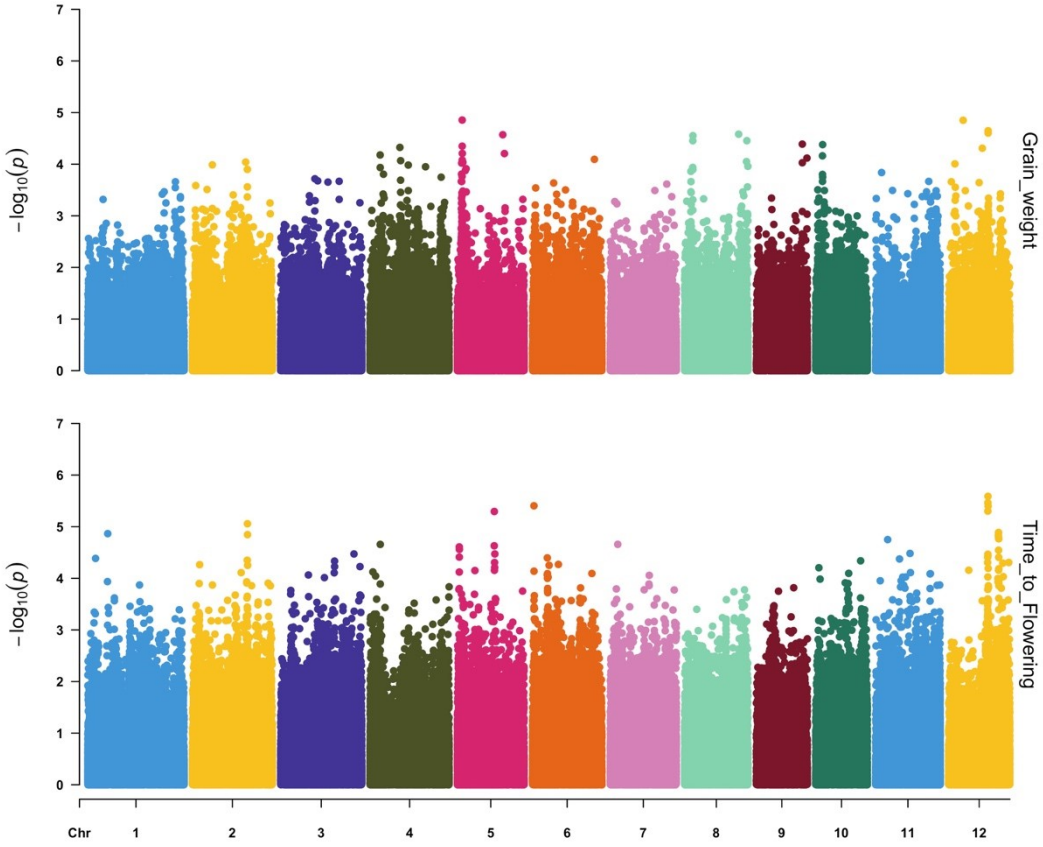

Figure 3: Manhattan plot of GWAS of "grain weight" and "time to flowering" traits.

At a genome-wide significant level of 5×10^{-4} for the genome of Oryza Sativa, there are 120 significant SNPs for "grain weight" trait and 304 significant SNPs for "time to flowering" trait. Samplings of the list of significant SNPs are written in Table A 1 and Table A 2.

3.2. Non-linear regressors on individual traits

The top SNPs are selected based on their P-value of the association test with the corresponding trait. The number of SNPs at each P-value threshold for individual traits are shown in . Because association test is the classical method to identify the SNPs having the most effect on the trait, we assumed that by choosing the SNPs having the lowest P-value of the association test as input for the machine learning models, we could improve its predicting performance. Hence, we select some typical P-value thresholds starting from 5×10^{-2} and decrease incrementally by a factor of 10.

For "grain weight" trait (-A), the RFR has a minimum MSE of 0.234 at P-value threshold of 5×10^{-3} with 1,323 SNPs. At the same threshold, the R^2 is at a maximum of 0.116. Meanwhile, SVR has a minimum MSE of 0.206 at P-value threshold of 5×10^{-3} with 1,323 SNPs also. At the same threshold, the R^2 is at a maximum of 0.224. For "time to flowering" trait (-B), the RFR has a minimum MSE of 471.4 at P-value threshold of 5×10^{-3} with 3,128 SNPs. At the same threshold, the R^2 is at a maximum of 0.083. Meanwhile, SVR has a

minimum MSE of 452.1 at p-value threshold of 5×10^{-3} with 3128 SNPs also. At the same threshold, the R^2 is at a maximum of 0.121.

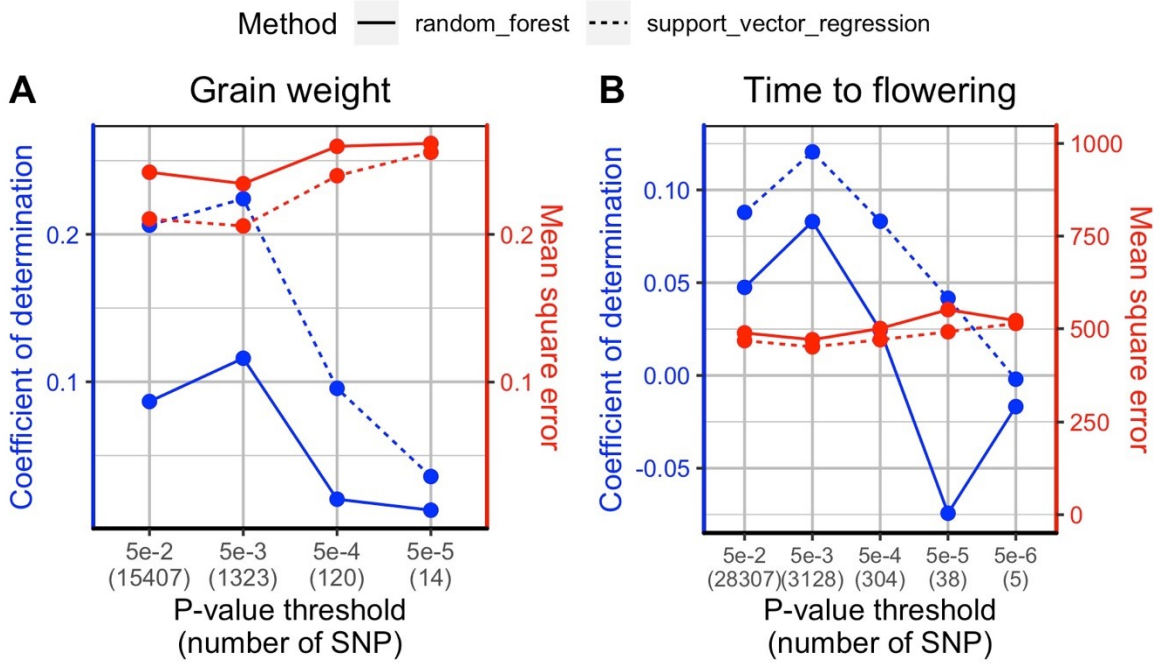

Figure 4: Comparison of random forest and support vector regression models in predicting quantitative trait using P-value thresholds. (A) R^2 and MSE of models predicting "grain weight". (B) R^2 and MSE of models predicting "time to flowering". (5e-2 is the scientific notation for 5 multiply by 10 to the power of -2)

For the non-linear regressors (i.e., RFR and SVR), SVR model consistently provides better performance in terms of both MSE and R^2 than RFR model. Those regressors with different P-value thresholds provide different accuracy of the resulted model. However, we can observe that the best accuracy is attained at 5×10^{-3} for both RFR and SVR models. We hypothesize that at higher P-value thresholds, the models included more noise SNPs that contribute little to the prediction models and at lower P-value thresholds, the models excluded too many SNPs that have a significant effect on the traits.

3.3. Non-linear regressors on combinatorial traits

Top N selected SNPs must satisfy a P-value threshold on both "grain weight" and "time to flowering" traits (). If we used the same P-value threshold as the above method, there are only two thresholds (5×10^{-2} and 5×10^{-3}) where there are non-zero SNPs selected. Therefore, we selected additional thresholds to compare the performance of the models at different number of SNPs included. Tree more P-value thresholds are chosen: 2.5×10^{-2}, 1×10^{-2}, 7.5×10^{-3}.

In predicting the trait "grain weight", the best model is observed at the P-value threshold of 5×10^{-2} with 929 SNPs (). SVR model gives a better prediction than the RFR. SVR model has MSE of 0.2323 corresponding to an R^2 of 0.1237 at P-value threshold of 5×10^{-2} (-A)

In predicting the trait "time to flowering", the best model is observed at the P-value threshold of 5×10^{-2} with 929 SNPs. SVR model gives a better prediction than the RFR. SVR model have MSE of 480.5 corresponding to an R^2 of 0.0653 at p-value threshold of 5×10^{-2} (-B)

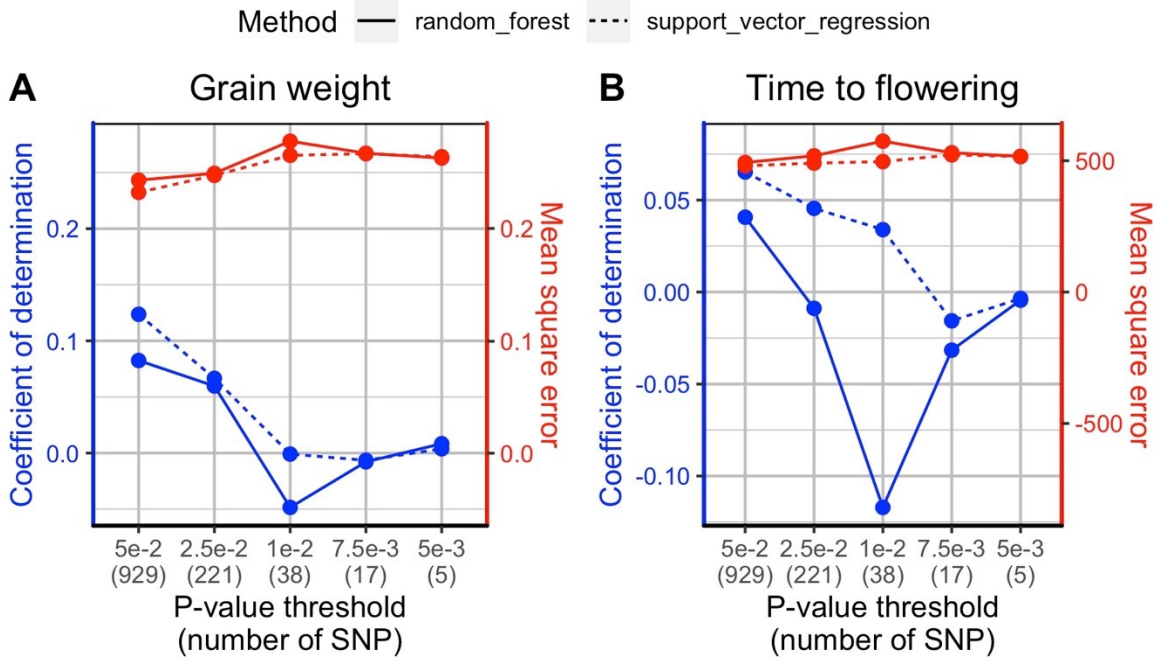

Figure 5: Comparison of random forest and support vector regression model in predicting quantitative trait using combinatorial P-value thresholds. (A) R^2 and MSE of models predicting "grain weight". (B) R^2 and MSE of models predicting "time to flowering". (5e-2 is the scientific notation for 5 multiply by 10 to the power of -2)

We observe that the best accuracy is attained at 5×10^{-2} for both RFR and SVR models. However, for this experiment (i.e., combinatorial traits), the R^2 is significantly lower than that of models in the previous experiment (i.e., individual traits). We speculate that is because the SNPs that are significantly associated with "grain weight" have little effect on "time to flowering" and vice versus. Therefore, the combination of "grain weight" and "time to flowering" resulted in fewer SNPs included in the prediction models. Moreover, these SNPs have little effect on either trait, so the model has lower accuracy. Interestingly, whether testing for individual traits or combinatorial traits, the R^2 of SVR model is usually higher than that of RF model. Accordingly, the MSE obtained by SVR model is lower than that of RF model.

3.4. Advanced linear regressors

The Lasso method will perform variable selection and regularization before producing the statistical model. However, Lasso tends to select one variable from a group and ignore others. The variables in our data are SNPs are not independent. Some SNPs have high LD with the adjacent SNPs on the chromosomes. Therefore, we also used Elastic Net, which overcomes the limitation of Lasso by adding a penalty which, when used alone is Ridge regression, to the model. Table 2 shows a comparison of prediction performance in terms of MSE and R^2 among advanced linear regressors.

For models based on individual traits, Lasso and Elastic Net get the same result with "grain weight" (1 SNPs picked, MSE is 0.2652, R-squared is -0.00052). Elastic Net picks 26 SNPs with "time to flowering" when Lasso only picks 4 SNPs, but the performance of Lasso is better than the performance of Elastic Net a little. For models based on combinatorial traits, multi-task Elastic Net picks 26 SNPs while multi-task Lasso only picks 4 SNPs, but the performance of multi-task Lasso is better than the performance of multi-task Elastic Net a little.

Table 2: Comparison of Lasso and Elastic Net

	Grain weight			Time to flowering		
Advanced linear regressors	#SNP	MSE	R^2	#SNP	MSE	R^2
Lasso	1	0.26524	-0.00052	4	514.4029	-0.00064
Multi-task Lasso	4	**0.26484**	**0.00100**	4	**514.4028**	**-0.00064**
Elastic Net	1	0.26524	-0.00052	26	514.7855	-0.00138
Multi-task Elastic Net	26	0.26501	0.00034	26	514.7853	-0.00138

The list of SNPs chosen in each method is written in and

Advanced linear regression methods generally have worse performance than the non-linear regression ones. The R^2 of single-task Lasso and single task Elastic Net all have a negative value. Only the multi-task Lasso and multi-task Elastic Net model predicting "grain weight" have a positive value. Moreover, all R^2 are very close to zero.

One interesting phenomenon that we observed is, the SNPs chosen for multi-task Lasso model is exactly the same as the SNPs chosen for single-task Lasso model predicting "time to flowering" (Table A 3). Similarly, the SNPs chosen for the multi-task Elastic Net model is exactly the same as the ones chosen for the single-task Elastic Net model predicting "time to flowering" (Table A 4). The unit of "time to flowering" trait is "day", and its standard deviation is 24.51. The unit of "grain weight" is "gram/100 grains" and its standard deviation is 0.52. When we computed the multi-task models with the mixed trait, it is apparent that the variance of "time to flowering" would represent most of the variance of the mixed trait of "grain weight & time to flowering". If the model only takes into account the variance of "time to flowering" trait because it is many times greater than the variance of "grain weight", it is explainable that the variables selected in the multi-task Lasso model are the same as the variables selected in Lasso model for "time to flowering" trait.

4. Conclusion

The genotype of the rice plant from the 3000 Rice Genome project is used in this study. In addition to performing traditional GWAS analysis to identify SNPs significantly associated with the two traits "grain weight" and "time to flowering", we investigated the prediction performance of some machine learning methods on two traits. The association testing between genotype and trait was done with a linear mixed model; meanwhile, the machine learning models were non-linear regressors (RFR and SVR) and advanced linear regressors (Lasso and Elastic Net).

In the non-linear methods, we adjust the input variables of the models by choosing different number of SNPs based on P-value threshold of the association test (SNPs having P-value lower than the threshold is chosen as the input variables). For individual traits, we observed that both RFR and SVR models performed the best at the P-value threshold of 5×10^{-3}. For the combinatorial traits, the models generally performed worse than themselves for the individual traits. Among the tested methods, the methods based on SVR with input SNPs chosen at P-value threshold of 5×10^{-3} for individual traits give the best performance in terms of MSE and R^2.

In advanced linear regressors, we observed that its models had worse performance than the ones based on the non-linear methods. We investigated four methods: Lasso, multi-task Lasso, Elastic Net and multi-task Elastic Net. Among these methods, only multi-task Lasso and multi-task Elastic Net produce models having positive R^2, the other methods all have a negative value of R^2. The multi-task Lasso and multi-task Elastic Net models predicting "time to flowering" have similar performance to Lasso and Elastic Net ones, consecutively. We argued that multi-task regressors was not suitable to analyze continuous traits with different scale. In addition, we speculate that the advanced linear regression methods performed worse than the non-linear ones because the variables (SNPs) of the genotype data is not independent with each other and the relationship between the genotype data and the traits are non-linear. Hence, the non-linear regression methods are better suited to capture the non-linearity of the GWAS data. The best model in terms of both MSE and R^2 in our experimental results is the support vector regression models. This confirms the result of a study by Grinberg et al.[29] who also argued that SVR method produce the best result compared to other machine learning methods.

Appendix A

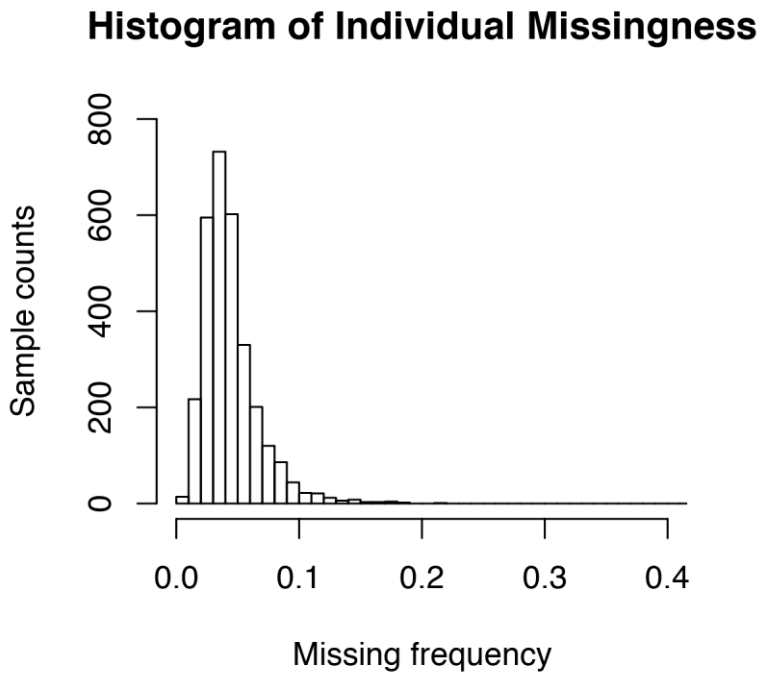

Figure A 1: Histogram of individual missingness. SNPs' missing rate distribution of the sample

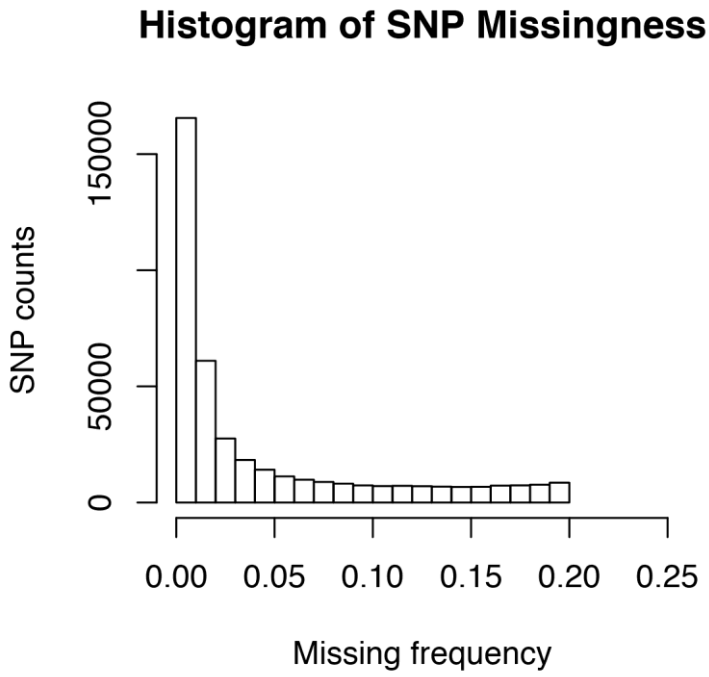

Figure A 2: Histogram of SNPs missingness. Samples' missing rate distribution per SNPs.

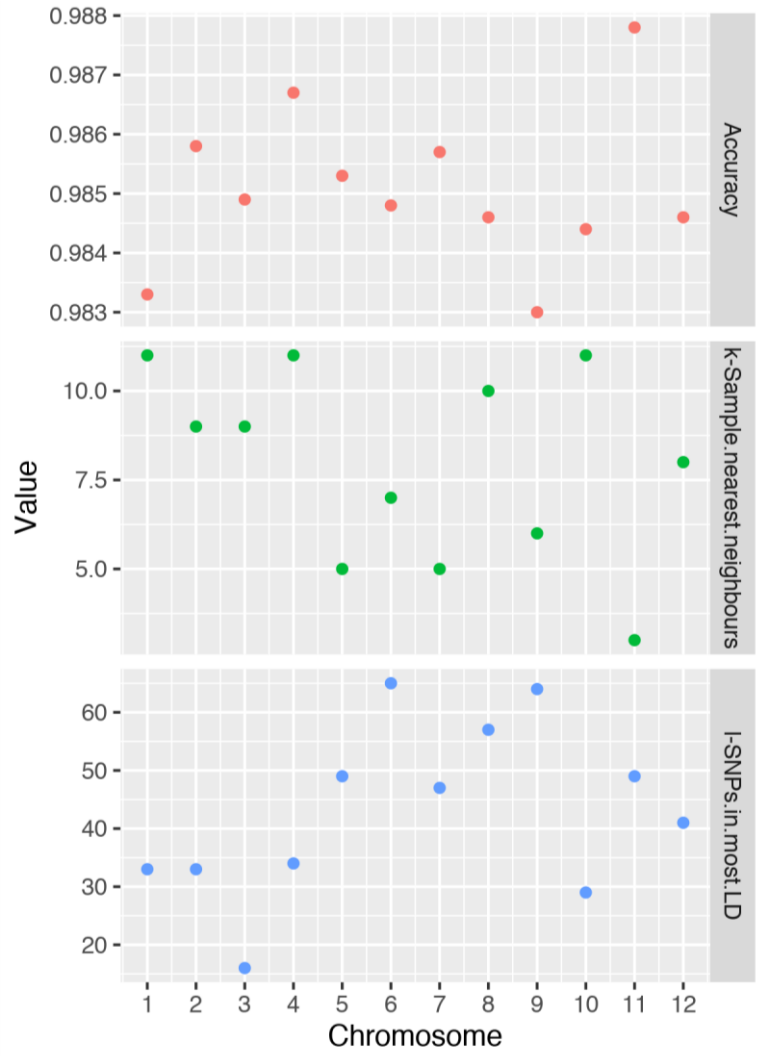

Figure A 3: Imputation accuracy and optimized parameter report for each chromosome. This is the summary of the output result when using LinkImpute to infer the missing value from every chromosome separately. The horizontal axis is the chromosome of the rice plant. The top plot is a scatter plot of the accuracy of the imputation step. The middle and bottom plots show the optimized parameter of the imputation corresponding to every chromosome. l: number of SNPs in high linkage disequilibrium with the imputed SNP. k: number of samples nearest neighbours based on l SNPs.

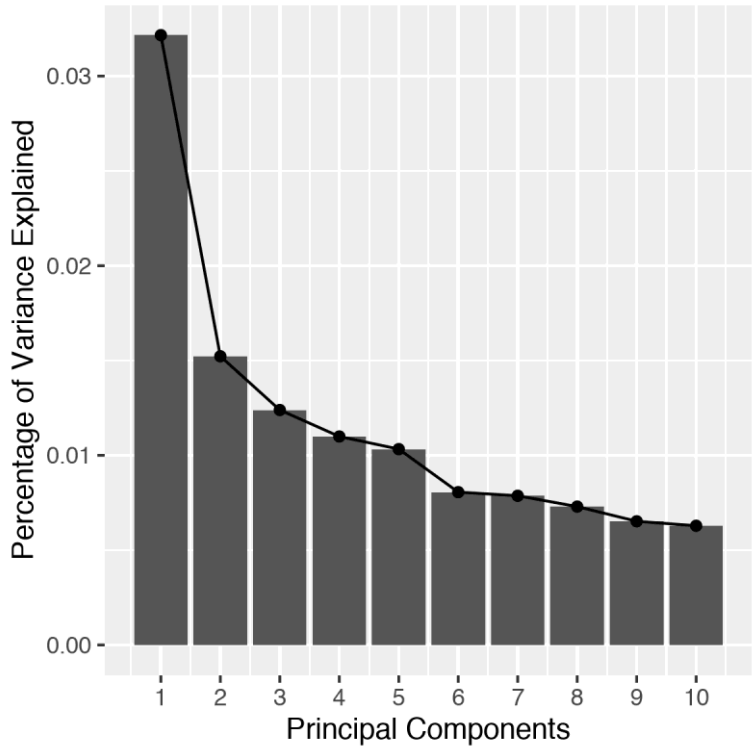

Figure A 4: Percentage of Variance Explained of the first ten principal components.

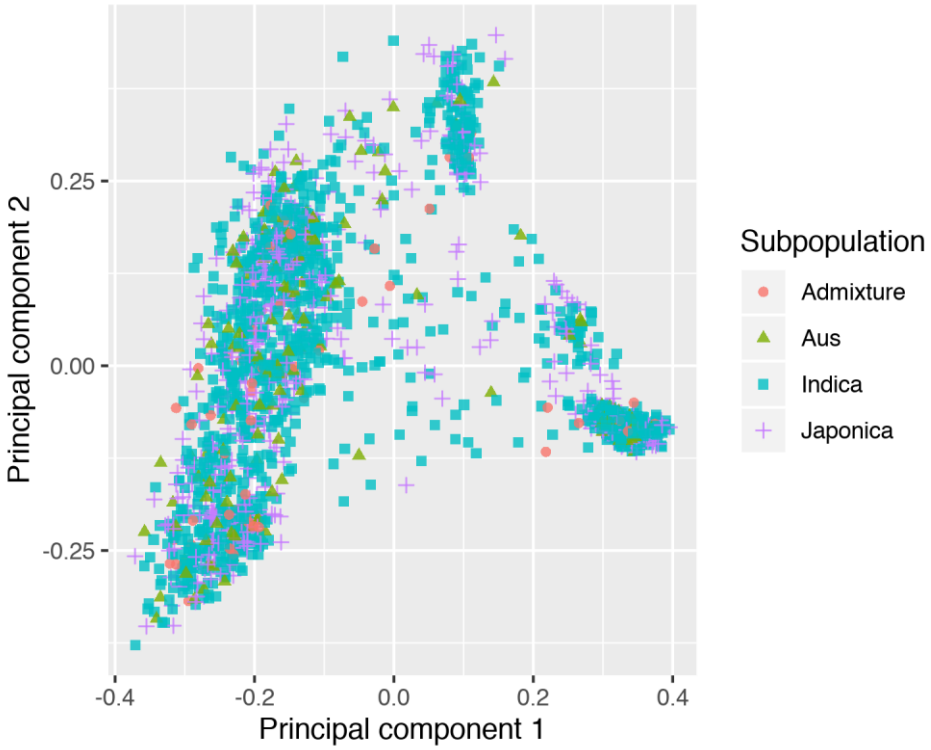

Figure A 5: Scatter plot of every sample based on Principal component 1 and 2.

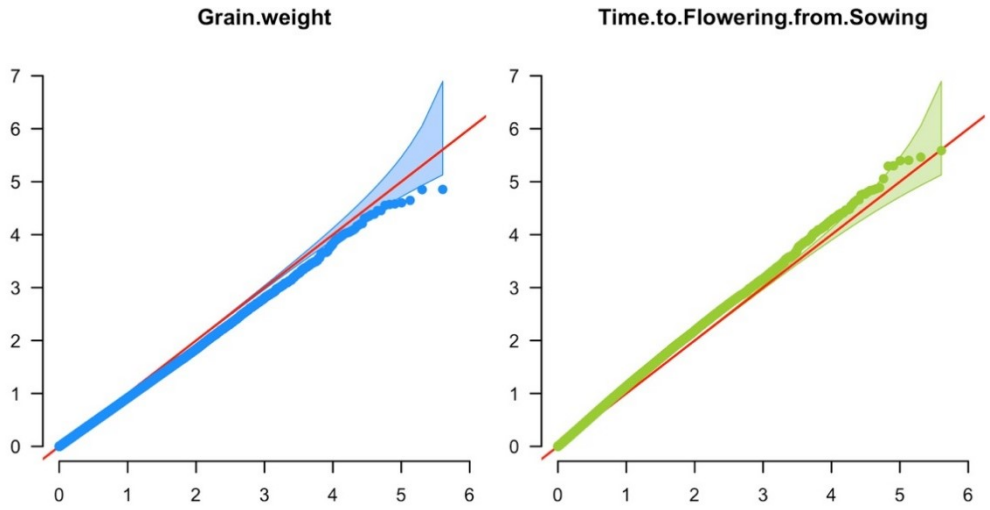

Figure A 6: Q-Q plot of the GWAS results of "grain weight" and "time to flowering". The horizontal axis is -log10(expected P-value) and the vertical axis is the -log10(observed P-value)

Table A 1: List of top 10 significant SNPs for "grain weight" trait at the threshold of 5×10^{-4}

Chromosome	SNP	BP	A1	A2	N	BETA	SE	P-value
12	363422307	17708644	G	T	2266	3.12	0.663	2.58×10^{-6}
12	363379344	17665681	T	C	2266	3.38	0.729	3.44×10^{-6}
6	181775087	691967	A	G	2266	2.65	0.574	3.94×10^{-6}
12	363499529	17785866	G	C	2266	3.8	0.825	4.00×10^{-6}
12	363358521	17644858	A	T	2266	2.55	0.56	4.99×10^{-6}
5	167690301	16565615	T	C	2266	8.91	1.95	5.08×10^{-6}
2	67995743	24724820	T	C	2266	-8.19	1.84	8.77×10^{-6}
12	368167535	22453872	A	G	2266	-5.46	1.25	1.30×10^{-5}
1	8954521	8954521	C	T	2266	-7.13	1.64	1.36×10^{-5}
2	68023241	24752318	A	G	2266	-5.9	1.36	1.43×10^{-5}
With 110 more rows (total 120 SNPs). BP: Base pair position of the SNP. A1: minor allele. A2: major allele. N: number of non-missing samples. BETA: beta coefficient of the linear regression. SE: Standard error								

Table A 2: List of top 10 significant SNPs for "time to flowering" trait at the threshold of 5×10^{-4}.

Chromosome	SNP	BP	A1	A2	N	BETA	SE	P-value
5	153259075	2134389	A	C	2260	0.0571	0.0131	1.40×10^{-5}
12	352270867	6557204	T	C	2260	0.233	0.0537	1.41×10^{-5}
12	363489728	17776065	A	T	2260	-0.145	0.0343	2.25×10^{-5}
12	363495388	17781725	C	T	2260	-0.141	0.0335	2.50×10^{-5}
8	266284759	24255231	T	C	2260	0.168	0.0399	2.62×10^{-5}
5	171493034	20368348	A	G	2260	0.239	0.0568	2.69×10^{-5}
8	245730796	3701268	A	G	2260	0.201	0.048	2.80×10^{-5}
8	269958731	27929203	A	G	2260	0.219	0.0528	3.51×10^{-5}
8	245689265	3659737	A	G	2260	0.198	0.0478	3.51×10^{-5}
9	291140088	20667538	C	T	2260	0.0836	0.0204	4.09×10^{-5}
With 294 more rows (total 304 SNPs) BP: Base pair position of the SNP. A1: minor allele. A2: major allele. N: non-missing sample. BETA: beta coefficient of the linear regression. SE: Standard error								

Table A 3: Number of SNPs in several thresholds of P-value

P-value threshold	Grain weight	Time to flowering
5×10^{-2}	15407	28307
5×10^{-3}	1323	3128
5×10^{-4}	120	304
5×10^{-5}	14	38
5×10^{-6}	0	5

Table A 4: Top SNPs selected from combined P-value of grain weight and time to flowering

P-value threshold	Grain weight and Time to flowering
5×10-2	929
2.5×10-2	221
1×10-2	38
7.5×10-3	17
5×10-3	5

Table A 5: SNPs included in the single task and the multi-task lasso models

#SNP	Single task lasso on grain weight	Single task lasso on time to flowering	Multi-task lasso
1	172235150	120600601	120600601
2		180393908	180393908
3		208081820	208081820
4		373102125	373102125

Table A 6: SNPs included in the single task and the multi-task elastic net models

#SNP	Single task elastic net on grain weight	Single task elastic net on time to flowering	Multi-task elastic net
1	172235150	2389069	2389069
2		11760405	11760405
3		34489925	34489925
4		41750977	41750977
5		50628584	50628584
6		61375990	61375990
7		84928121	84928121
8		120600601	120600601
9		125832526	125832526
10		137230497	137230497
11		147879747	147879747
12		180393908	180393908
13		180644828	180644828
14		208081820	208081820
15		227565875	227565875
16		230068383	230068383
17		235911589	235911589
18		239306881	239306881
19		239531623	239531623
20		244281569	244281569
21		274806614	274806614
22		285157810	285157810
23		341280196	341280196
24		345796823	345796823
25		363361795	363361795
26		373102125	373102125

References

1. Seck, P.A., et al., *Crops that feed the world 7: Rice.* Food Security, 2012. **4**(1): p. 7-24.
2. Richard, M.A., et al., *Effects of global climate change on agriculture: an interpretative review.* Climate Research, 1998. **11**(1): p. 19-30.
3. Wassmann, R., et al., *Chapter 2 Climate Change Affecting Rice Production: The Physiological and Agronomic Basis for Possible Adaptation Strategies*, in *Advances in Agronomy*, D.L. Sparks, Editor. 2009, Academic Press. p. 59-122.
4. Tester, M. and P. Langridge, *Breeding technologies to increase crop production in a changing world.* Science, 2010. **327**(5967): p. 818-22.
5. Pérez-de-Castro, A.M., et al., *Application of genomic tools in plant breeding.* Current genomics, 2012. **13**(3): p. 179-195.
6. Guo-Liang, J., *Molecular Markers and Marker-Assisted Breeding in Plants*, in *Plant Breeding from Laboratories to Fields*, S.B. Andersen, Editor. 2013, IntechOpen.
7. Mammadov, J., et al., *SNP Markers and Their Impact on Plant Breeding.* International journal of plant genomics, 2012. **2012**: p. 728398.
8. Xu, Y., *Molecular Breeding Tools: Genetic Markers*, in *Molecular plant breeding*. 2010, Centre for Agriculture and Bioscience International.
9. Hall, D., C. Tegstrom, and P.K. Ingvarsson, *Using association mapping to dissect the genetic basis of complex traits in plants.* Brief Funct Genomics, 2010. **9**(2): p. 157-65.
10. Korte, A. and A. Farlow, *The advantages and limitations of trait analysis with GWAS: a review.* Plant Methods, 2013. **9**(1): p. 29.
11. Gibson, G., *Rare and common variants: twenty arguments.* Nat Rev Genet, 2012. **13**(2): p. 135-45.
12. Vilhjalmsson, B.J. and M. Nordborg, *The nature of confounding in genome-wide association studies.* Nat Rev Genet, 2013. **14**(1): p. 1-2.
13. Libbrecht, M.W. and W.S. Noble, *Machine learning applications in genetics and genomics.* Nature Reviews Genetics, 2015. **16**(6): p. 321-332.
14. Grinberg, N.F., O.I. Orhobor, and R.D. King, *An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat.* Machine Learning, 2019.
15. The, r.g.p., *The 3,000 rice genomes project.* GigaScience, 2014. **3**(1): p. 7.
16. Bush, W.S. and J.H. Moore, *Chapter 11: Genome-Wide Association Studies.* PLOS Computational Biology, 2012. **8**(12): p. e1002822.
17. Hearst, M.A., *Support Vector Machines.* IEEE Intelligent Systems, 1998. **13**(4): p. 18–28.
18. Breiman, L., *Random Forests.* Mach. Learn., 2001. **45**(1): p. 5–32.
19. Tibshirani, R., *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society. Series B (Methodological), 1996. **58**(1): p. 267-288.
20. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. **67**(2): p. 301-320.

21. Mansueto, L., et al., *Rice SNP-seek database update: new SNPs, indels, and queries.* Nucleic Acids Research, 2016. **45**(D1): p. D1075-D1081.
22. Jiao, S., et al., *The use of imputed values in the meta-analysis of genome-wide association studies.* Genet Epidemiol, 2011. **35**(7): p. 597-605.
23. Money, D., et al., *LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms.* G3: Genes|Genomes|Genetics, 2015. **5**(11): p. 2383-2390.
24. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
25. Zhang, Z., et al., *Mixed linear model approach adapted for genome-wide association studies.* Nature Genetics, 2010. **42**(4): p. 355-360.
26. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
27. Jiang, L., et al., *A resource-efficient tool for mixed model association analysis of large-scale data.* Nature Genetics, 2019. **51**(12): p. 1749-1755.
28. Hoang, G.T., et al., *Genome-wide Association Study of a Panel of Vietnamese Rice Landraces Reveals New QTLs for Tolerance to Water Deficit During the Vegetative Phase.* Rice, 2019. **12**(1): p. 4.
29. Grinberg, N.F., O.I. Orhobor, and R.D. King, *An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat.* Machine Learning, 2020. **109**(2): p. 251-277.