Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

NJIT New Jersey's Science & Technology University

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

RICE AND MOUSE QUANTITATIVE PHENOTYPE PREDICTION IN GENOME-WIDE ASSOCIATION STUDIES WITH SUPPORT VECTOR REGRESSION

by

Abdulrhman Fahad Aljouie

Quantitative phenotypes prediction from genotype data is significant for pathogenesis, crop yields, and immunity tests. The scientific community conducted many studies to find unobserved quantitative phenotype high predictive ability models. Early genome-wide association studies (GWAS) focused on genetic variants that are associated with disease or phenotype, however, these variants manly covers small portion of the whole genetic variance, and therefore, the effectiveness of predictions obtained using this information may possibly be circumscribed [1].

Instead, this study shows prediction ability from whole genome single nucleotide polymorphisms (SNPs) data of 1940 genotyped stoke mouse with ~ 12k SNPs, and 413 genotyped rice inbred lines with ~ 40k SNPs. The predictive accuracy measured as the Pearson coefficient correlation between predicted phenotype and actual phenotype values using cross validation (CV), and found a predictive ability for mouse phenotypes MCH, CD8 to be 0.64 and 0.72, respectively.

The study compares whole genome SNPs data prediction methods built using Support Vector Regression (SVR) and Pearson Correlation Coefficient (PCC) to perform SNPs selection and then predict unobserved phenotype using ridge regression and SVR. The investigation shows that ranking SNPs by SVR significantly increases predictive

accuracy than ranking with PCC. In general, Ridge Regression perform slightly better prediction ability than predicting with SVR.

RICE AND MOUSE QUANTITATIVE PHENOTYPE PREDICTION IN GENOME-WIDE ASSOCIATION STUDIES WITH SUPPORT VECTOR REGRESSION

by
Abdulrhman Fahad M Aljouie

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Bioinformatics

Department of Computer Science

January 2015

APPROVAL PAGE

RICE AND MOUSE QUANTITATIVE PHENOTYPE PREDICTION IN GENOME-WIDE ASSOCIATION STUDIES WITH SUPPORT VECTOR REGRESSION

Abdulrhman Fahad Aljouie

Dr. Usman Roshan, Thesis Advisor — Date
Associate Professor of Computer Science, NJIT

Dr. Jason T. L. Wang, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Zhi Wei, Committee Member — Date
Associate Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Abdulrhman Fahad M Aljouie

Degree: Master of Science

Date: January 2015

Undergraduate and Graduate Education:

- Master of Science in Bioinformatics,
  New Jersey Institute of Technology, Newark, NJ, 2015
- Bachelor in Computer,
  King Saud University, Riyadh, Saudi Arabia, 2007

Major: Bioinformatics

Dedicated to my parents Nora and Fahad, my wife Nahlah, and my son Fahad, for their encouragements, infinite love, and support.

ACKNOWLEDGMENT

My work would not have been possible without the help from several people. I would like to express my gratitude to my advisor, Dr. Usman Roshan, who has shown significant guidance, immense knowledge, and extraordinary support. I also would like to thank my thesis committee members, Dr. Jason Wang and Dr. Zhi Wei for their valuable mentorship throughout my graduate study.

Special thanks to my employer, King Abdullah International Research Center (KAIMRC), for financially aiding my graduate study and providing generous scholarship award. I also want to thank Dr. Mohamed Alkelya and Dr. Barrak Alsomie for their encouragements and support.

TABLE OF CONTENTS

Chapter	Page
1 INTRODUCTION ……...	1
2 MATERIALS AND METHODS ……...	3
2.1 Data ……..	3
2.1.1 Mice SNPs Data ...……...……..……....	3
2.1.2 Mice Phenotype Data ….....…….……....	3
2.1.3 Rice SNPs Data ...……..……...……..	3
2.1.4 Rice Phenotype Data …….…….	3
2.2 Genotyping Encoding …….	5
2.3 Imputation Method ……..	5
2.4 Cross Validation ……..……....	5
2.5 Feature Selection ……...……..	6
2.6 Support Vector Regression …….……	7
2.7 Ridge Regression ……..……..	7
2.8 Pearson's Correlation Coefficient ……..	8
2.9 Study Workflow …….	9
3 ANALYSIS AND IMPLEMENTATION ...……..	10
3.1 Regressions Models Implementation …….…..……	10
3.2 Predictive Power Computation ……..……	10
4 RESULTS ……...……..	11
4.1 Mice Phenotypes Prediction Ability Results …….…..……	11

TABLE OF CONTENTS
(Continued)

Chapter **Page**

4.2 Rice Phenotypes Prediction Ability Results …….. 12

5 DISCUSSION ……..... 21

6 CONCLUSION ……...……... 22

REFERENCES ……..……...……... 23

LIST OF TABLES

Table		Page
2.1	Preprocessing Genotype Data Sets …..……...	4
2.2	Phenotype Data Sets ……..…….....	4

LIST OF FIGURES

Figure		Page
2.1	Study Workflow…….	9
4.1	Line graph of prediction accuracy average over 10 splits of mouse MCH data after ranking the SNPs based on SVR absolute value, RR coefficient, and Pearson's correlation absolute value and predicting using SVR and RR models on each ranked split…..……..	11
4.2	Line graph of prediction accuracy average over 10 splits of mouse CD8 data after ranking the SNPs based on SVR absolute value, RR Coefficient and Pearson's correlation absolute value and predicting using SVR and RR models on each ranked split...……..…….....	12
4.3	Line graph of rice days to flower data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset…….	13
4.4	Line graph of rice amylose content data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset…….	13
4.5	Line graph of rice days to blast resistance data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset……..…….	14
4.6	Line graph of rice days to flag leaf length data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset…….…….	15
4.7	Line graph of rice flag leaf width data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset…….	15
4.8	Line graph of rice panicle length data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset…….	16
4.9	Line graph of rice panicle number per plant data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset…….…….	17

LIST OF FIGURES
(Continued)

Figure		Page
4.10	Line graph of rice primary panicle branch number data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset…….…….	17
4.11	Line graph of rice seed number per panicle data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset…….…….	18
4.12	Line graph of rice seed width data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset……..	19
4.13	Line graph of rice plant height data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset…….…….	19

LIST OF SYMBOLS

Chr	Chromosome
≈	Approximately
#	Number Sign
AUC	Accuracy
RR	Ridge Regression
SVR	Support Vector Regression
PCC	Pearson’s Correlation Coefficient

CHAPTER 1

INTRODUCTION

High accuracy prediction of unobserved genetic values for quantitative phenotype are important to understand human diseases, as well as animal and plant breeding [1,2]. Two main approaches has been followed to predict complex traits: (1) genome wide association studies (GWAS), and (2) whole genome prediction (WGP). Both approaches use SNPs and phenotype data to predict genetic values for unobserved traits [2].

GWAS identified genetic markers associated with quantitative phenotype or risk to common disease. Many human and other organisms' quantitative trait loci (QTL) has been detected [2]. However, these variants manly covers small portion of the whole genetic variance, and therefore, the effectiveness of predictions obtained using this information may possibly be circumscribed [1,2]. In WGP, all organism genetic markers are considered for predicting a specific trait, it has been found that predicting with whole genetic variants is encouraging and could increase the prediction accuracy [2,3].

In this study, whole genome single nucleotide polymorphisms (SNPs) data of 1940 genotyped mouse for ≈ 12k SNPs, and 413 genotyped rice inbred lines for ≈ 40k SNPs. The complex traits for mice are Mean Cellular Hemoglobin (MCH), and immunology %CD8 cells. The rice complex traits are days to flower, amylose content, blast resistance, flag leaf length, flag leaf width, panicle length, panicle number per plant, primary panicle per branch, seed number per plant, seed width, and plant height. The predictive accuracy measured as the Pearson's coefficient correlation between predicted

phenotype and actual phenotype values using 10-fold and 5-fold cross validation (CV) for mice and rice data sets, respectively.

To enhance phenotype prediction accuracy genotype has been ranked based on three methods for mice data sets, Support Vector Regression (SVR) weight vector (w) absolute values and Ridge Regression (RR) coefficients as well as Pearson's Correlation Coefficient (PCC) absolute values and then predicting using RR and SVR. In rice data sets, features are ranked by SVR weight vector (*w*) absolute values as a multivariate feature selection method and PCC absolute values as a univariate feature selection.

CHAPTER 2

MATERIALS AND METHODS

2.1 Data

2.1.1 Mice SNPs Data

Mice data sets consists of 12545 SNP from 298 parents and 1940 mouse across 20 chromosomes, and is made publically available by Welcome Trust Centre for Human Genetics, it can be accessed via URL http://mus.well.ox.ac.uk/mouse/HS/.

2.1.2 Mice Phenotypes Data

Two continuous mouse phenotypes data sets has been used in this analysis, Mean Cellular Hemoglobin (MCH), and Immunology CD8 [4].

2.1.3 Rice SNPs Data

Rice data set consist of 36901 SNP from 82 countries and 413 rice plant across 12 chromosomes, and is made publically available by Rice Diversity Panel, it can be accessed via URL http://ricediversity.org/data/sets/44kgwas/ [5].

2.1.4 Rice Phenotypes Data

In this analysis, 11 continuous rice phenotypes has been used in prediction, these phenotypes are days to flower, amylose content, blast resistance, flag leaf length, flag leaf width, panicle length, panicle number per plant, primary panicle per branch, seed number per plant, seed width, and plant height [5].

Table 2.1 Preprocessing Genotype Data Sets

Data set	# of Chromosomes	# of Samples	# of SNPs
Mice	20	1940	12545
Rice	12	413	36901

Table 2.2 Phenotype Data sets

Data set	# of Samples
Mouse %CD8	1521
Mouse MCH	1591
Rice days to flower	374
Rice amylose content	401
Rice blast resistance	385
Rice flag leaf length	377
Rice flag leaf width	377
Rice panicle length	375
Rice panicle number per plant	372
Rice primary panicle per branch	375
Rice seed number per plant	376
Rice seed width	377
Rice plant height	383

2.2 Genotype Encoding

To pass the SNPs matrix into regression models, the genotype data sets were encoded using a Perl script into 0, 1, and 2, these numbers are assigned based on minor allele counts for each subject in a given SNP. Where zero means no minor allele is present for a particular subject in that SNP, 1 means there is one minor allele, and 2 means there are two copies of the minor allele.

2.3 Imputation Method

Missing values are available in all mice and rice data sets. SNPs that contain missing values greater than or equals to 0.01 has been excluded from this analysis. The remaining missing values in SNPs has been imputed by assigning the most occurring encoding in each SNP. The total number of SNPs used in this analysis for mice data sets after imputation is 12145, and the total number of SNPs used for rice data set is 15493.

2.4 Cross Validation

Cross-validation (CV) is a method used in machine learning for model selection [6]. The data set is split into two parts one part is used for building the model and the other part (validation) is used to assess the prediction accuracy [6]. In a 10-fold CV the data set is broken into ten equal parts of size n/10. Then the model is trained on nine parts and tested on the remaining part, this process is repeated ten times, each time using a different part for validation. In a 5-fold CV the data set is broken into five equal parts of size n/5. The model is trained on four parts and tested on the remaining part, this process is repeated five times, each time using a different part for validation [6].

In mice data sets a 10-fold CV has been used in this study, and 5-fold CV has been used for rice data sets.

The Ridge shrinking parameter (penalty) and Support Vector Regression cost (penalty) as well as the feature selection has been computed based on the training part of each split.

2.5 Feature Selection

Feature selection plays vital role in eliminating data noise on genomic data sets. Training on a highly correlated data leads to a model that perform poorly on prediction [7]. Therefore, feature selection could enhance the predictive ability. Two methods has been used to rank SNPs data sets namely; multivariate feature selection i.e. Ridge Regression (RR) and Support Vector Regression (SVR), and univariate feature selection using Pearson's Correlation Coefficient (PCC).

Multivariate SNPs selection with regularized linear models using Ridge Regression and Support Vector Regression yielded better predictive ability than univariate SNPs selection using Pearson's Correlation Coefficient on all Mice data sets. In Rice data sets predictive ability using multivariate method by ranking with weight vector of Support Vector Regression was superior to ranking with Pearson's Correlation Coefficient in general.

2.6 Support Vector Regression

Support Vector Regression (SVR) aim to find a function *f(x)* that minimize the deviation ε from the actual dependent variables y_i.

$$\text{Minimize } \frac{1}{n} ||w||^2 + C \sum_{i=1}^{\iota} \xi_i + \xi_i^* \tag{2.1}$$

$$\text{s.t.} \begin{Bmatrix} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{Bmatrix} \tag{2.2}$$

where C is the cost (trade of between tolerating training errors and margin), y_i is the dependent variables, x_i is the independent variables, ξ_i^*,ξ_i is the slack variables.
This study uses the default cost (penalty) $avg.(x.x)^{-1}$.

2.7 Ridge Regression

Multicollinearity (linear relationship between one or more independent variables over 0.90) is a well-known problem in genomic data. Therefore, fitting a model using a multiple linear regression is difficult, since (X'X) is hard to invert.

$$Y = X\beta + \varepsilon \tag{2.3}$$

where *Y* is the target, *X* is the independent variables, and $\beta$is the regression coefficients, and ε is the errors term.

To circumvent this problem, a regularized term is introduced i.e. λI in Ridge Regression (RR). Ridge regression [Hoerl and Kennard (1970)] is a regularized multiple linear regression. RR is a modification of the multiple linear regression.

$$\hat{\beta} = (X'X + \lambda I)^{-1} X'y \tag{2.4}$$

where is "beta hat" is the estimate of β the regression coefficient in the multiple linear regression, λ is the ridge penalty, and I is the identity matrix.

The ridge is penalty used in the study analysis is a semi-automatic following [E.-Culeis- (2012)] method.

$$k_r = \frac{r\,\hat{\sigma}_r^2}{\sum_{j=1}^{r}\hat{\alpha}_j^2} \tag{2.5}$$

where k_r is the shrinking parameter, r is the first principal component, and $\hat{\sigma}_r^2$ is given by

$$\hat{\sigma}_r^2 = \frac{(y-Z_r\hat{\alpha}_r)\prime(y-Z_r\hat{\alpha}_r)}{n-r} \tag{2.6}$$

2.8 Pearson's Correlation Coefficient

Pearson's Correlation Coefficient (PCC) is used to measure the dependence between two variables X and Y, PCC yields a value between -1 and 1, where is 1 means positive correlation, -1 negative correlation, and 0 means no correlation.

$$PCC(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{2.7}$$

where $\bar{X}$ is the X mean, and $\bar{Y}$ is the Y mean.

In this study PCC value has been used to measure the prediction accuracy on validation part, where is Y= true phenotypes, and X = predicted phenotypes. In addition, the absolute value of PCC has been used in feature selection.

2.9 Study Workflow

The Study has been implemented following sequence of steps illustrated in (Figure 2.1).

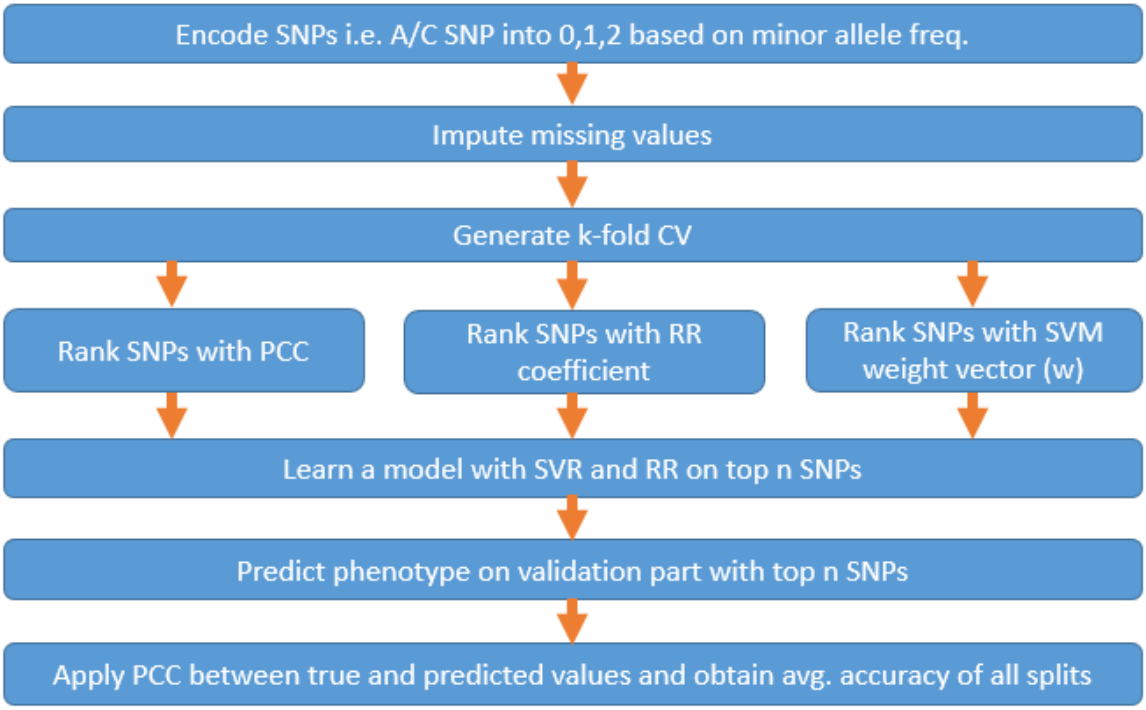

Figure 2.1 Study Workflow.

Ranking SNPs data with Ridge Regression fitted coefficients implemented on mice data sets, however, it was not implemented on rice data sets.

CHAPTER 3

ANALYSIS AND IMPLEMENTATION

3.1 Regressions Models Implementation

In this study R package Ridge is used to conduct Ridge Regression with semi-automatic Ridge parameter (penalty) assignment, which is an implementation of [E. Culeis 2012] method. SVM-Light program, an implementation of Vapnik's Support Vector Machine [Vapnik, 1995], is used to conduct Support Vector Regression with the default cost (penalty) assignment.

3.1 Predictive Power Computation

Predicted phenotypes values are computed using the predict function in R for Ridge Regression and classification module from SVM-light program for Support Vector Regression.

Using 10-fold CV for mice and 5-fold CV for rice, the prediction ability for each phenotype split is obtained by measuring the correlation between true and predicted phenotypes, then all predictive abilities obtained of each split is averaged. The results shown in this analysis for mice and rice phenotypes are averaged accuracies across all splits.

CHAPTER 4

RESULTS

4.1 Mice Phenotypes Prediction Ability Results

MCH prediction ability peaked at top 4k SNPs, while ranking with SVR weight vector (*w*), learning and predicting with Ridge Regression (RR) shows 0.641 averaged prediction accuracy over ten splits and standard deviation 0.04. CD8 prediction ability peaked at top 4k SNPs, while ranking with SVR weight vector (*w*), learning and predicting with Ridge Regression it shows 0.725 averaged prediction accuracy over 10 splits with standard deviation 0.05. With only 100 SNPs ranked by SVR *w* vector, learning and predicting with RR, the average accuracies are 0.48 and 0.61 for MCH and CD8, respectively.

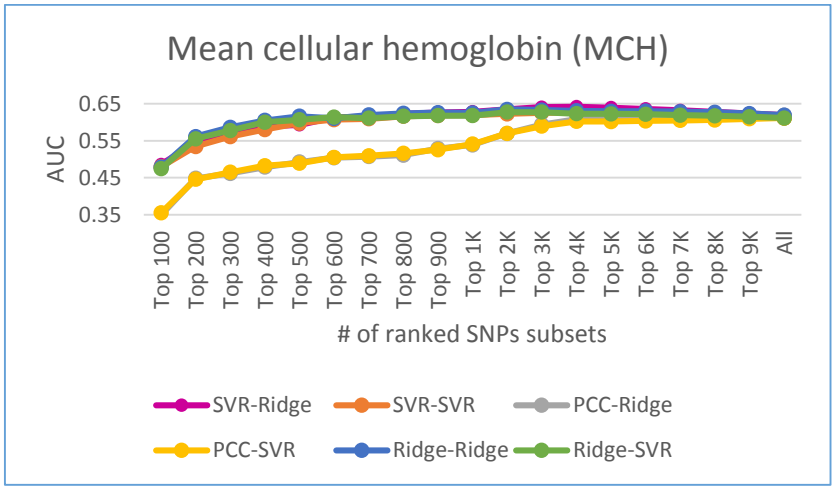

Figure 4.1 Line graph of prediction accuracy average over 10 splits of mouse MCH data after ranking the SNPs based on SVR absolute value, RR coefficient, and Pearson's correlation absolute value and predicting using SVR and RR models on each ranked split.

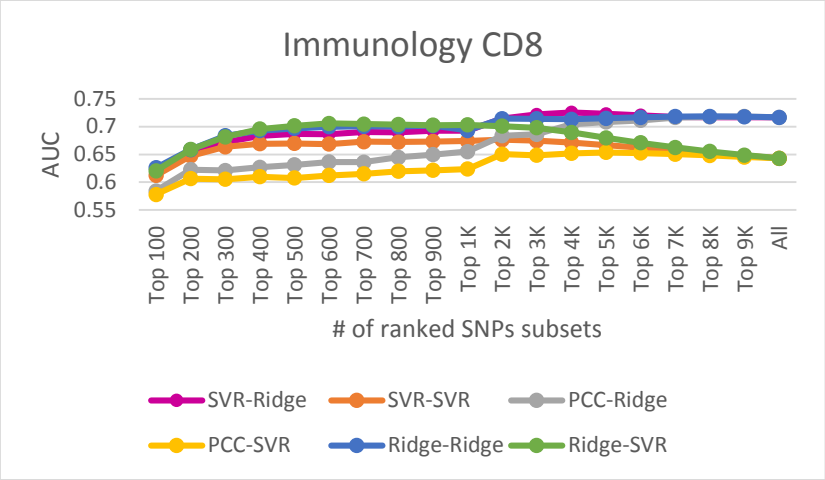

Figure 4.2 Line graph of prediction accuracy average over 10 splits of mouse CD8 data after ranking the SNPs based on SVR absolute value, RR Coefficient and Pearson's correlation absolute value and predicting using SVR and RR models on each ranked split.

4.2 Rice Phenotypes Prediction Ability Results

The analysis results of rice phenotypes mostly show that ranking with SVR outperform ranking with PCC. The learning and predicting with RR yield slight better prediction ability than learning and predicting with SVR. The results, of the analysis of all rice phenotypes studied are listed in the next pages.

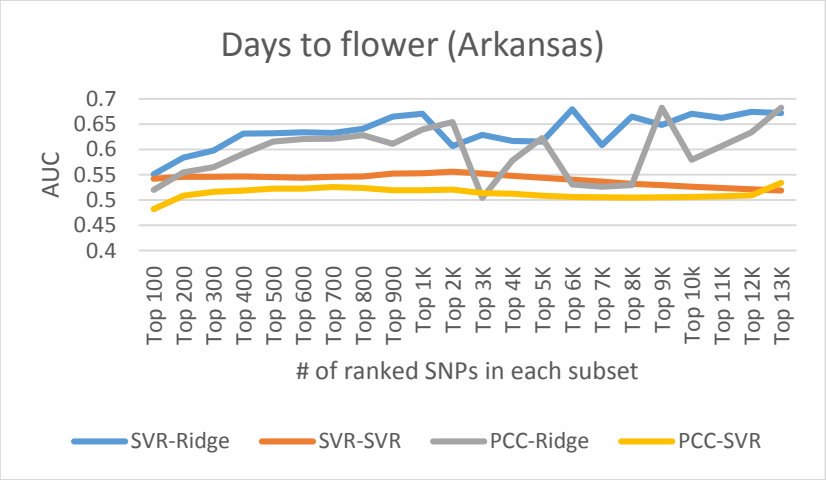

Figure 4.3 Line graph of rice days to flower data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.3 shows that days to flower phenotype prediction ability peaked at top 9k SNPs. The study found that ranking with PCC absolute value, learning and predicting with RR yielded higher accuracy than ranking with SVR weight vector absolute value. It shows 0.68 (0.08) averaged prediction accuracy (standard deviations) over 5 splits.

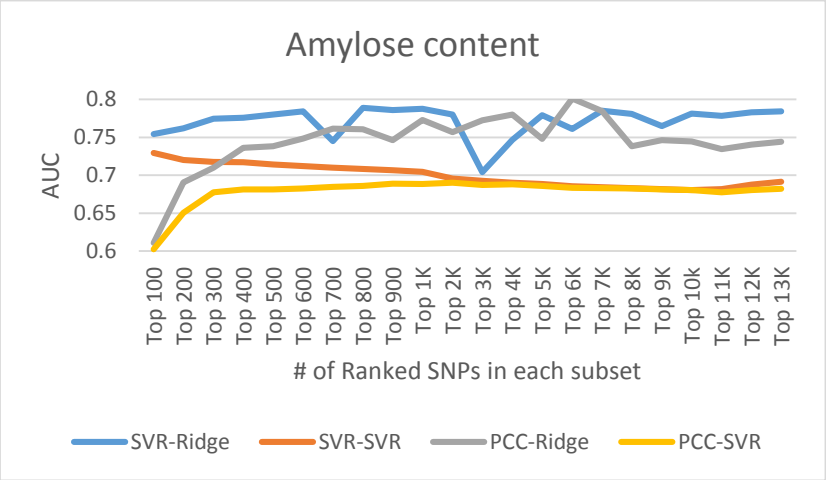

Figure 4.4 Line graph of rice amylose content data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.4 shows that amylose content phenotype prediction ability peaked at top 6k SNPs. The study found that ranking with PCC absolute value, learning and predicting with RR yielded higher accuracy than ranking with SVR weight vector absolute value. It shows 0.80 (0.05) averaged prediction accuracy (standard deviations) over 5 splits.

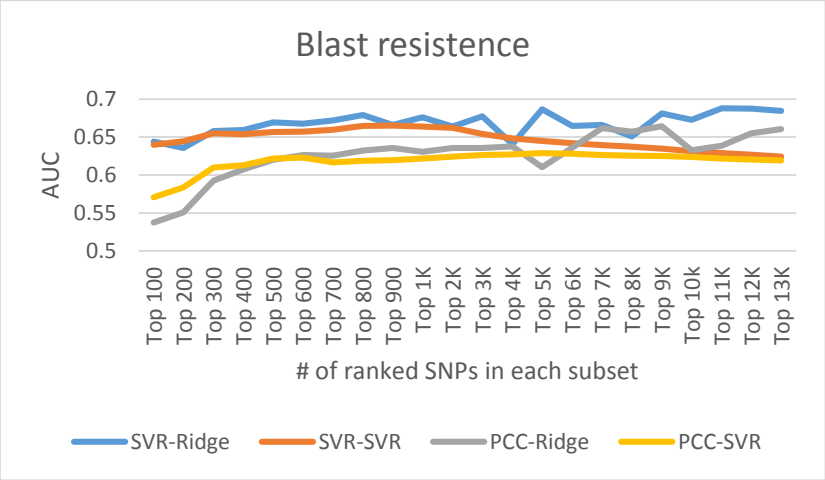

Figure 4.5 Line graph of rice blast resistance data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.5 shows that blast resistance phenotype prediction ability peaked at top 11k SNPs. The study found that ranking with SVR w vector absolute value, learning and predicting with RR yielded higher accuracy than ranking with PCC absolute value. It shows 0.68 (0.04) averaged prediction accuracy (standard deviations) over 5 splits.

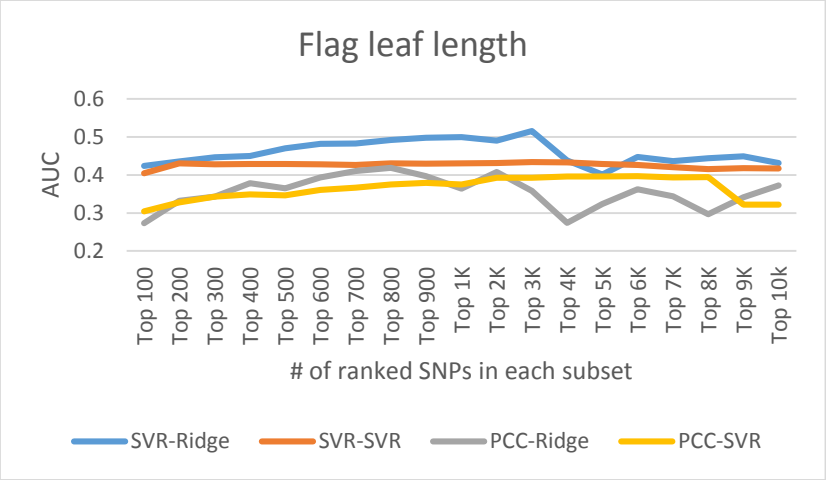

Figure 4.6 Line graph of rice flag leaf length data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.6 shows that flag leaf length phenotype prediction ability peaked at top 3k SNPs. The study found that ranking with SVR w vector absolute value, learning and predicting with RR yielded higher accuracy than ranking with PCC absolute value. It shows 0.51 (0.06) averaged prediction accuracy (standard deviations) over 5 splits.

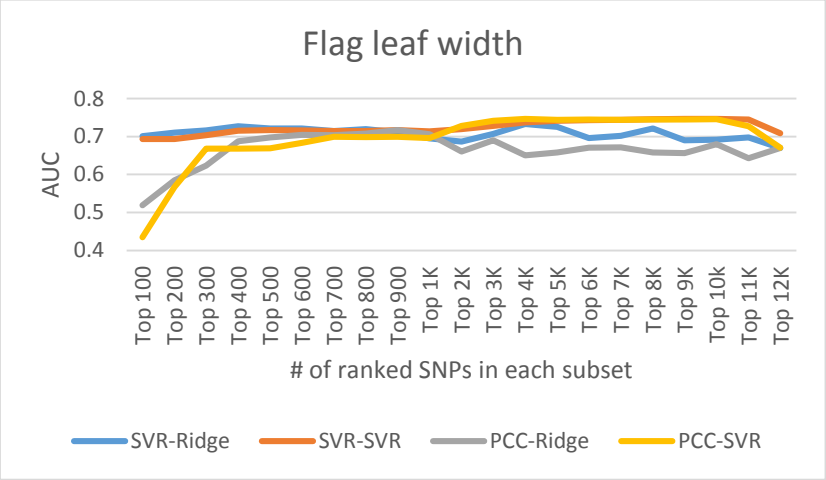

Figure 4.7 Line graph of rice flag leaf width data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.7 shows that flag leaf width phenotype prediction ability peaked at top 8k SNPs. The study found that ranking with SVR w vector absolute value, learning and predicting with SVR yielded higher accuracy than ranking with PCC absolute value. It shows 0.74 (0.09) averaged prediction accuracy (standard deviations) over 5 splits.

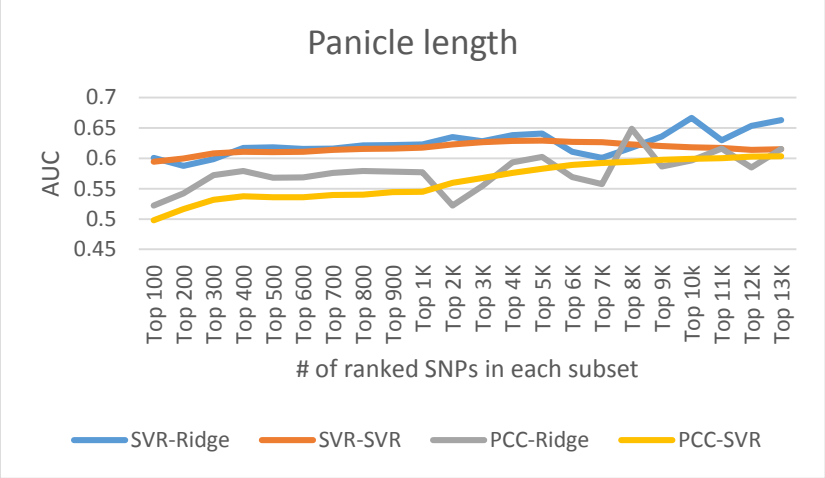

Figure 4.8 Line graph of rice panicle length data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.8 shows that panicle length phenotype prediction ability peaked at top 10k SNPs. The study found that ranking with SVR w vector absolute value, learning and predicting with RR yielded higher accuracy than ranking with PCC absolute value. It shows 0.66 (0.08) averaged prediction accuracy (standard deviations) over 5 splits.

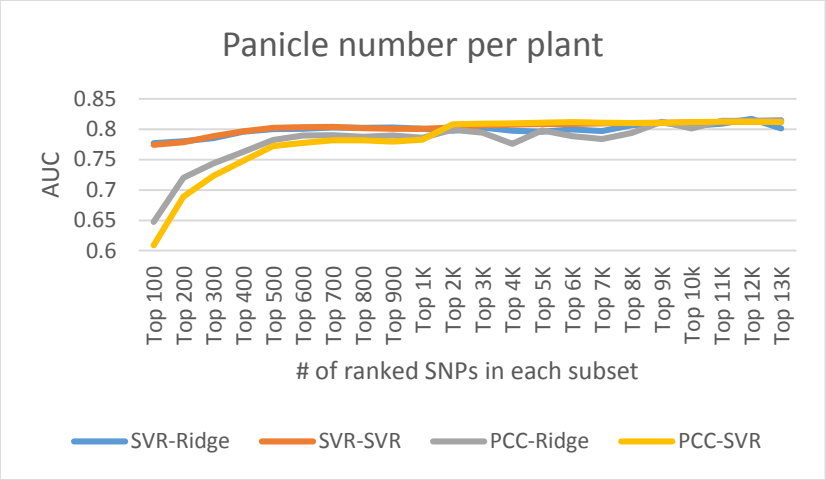

Figure 4.9 Line graph of rice panicle number per plant data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.9 shows that panicle number per plant phenotype prediction ability peaked at top 12k SNPs. The study found that ranking with SVR w vector absolute value, learning and predicting with RR yielded higher accuracy than ranking with PCC absolute value. It shows 0.81 (0.03) averaged prediction accuracy (standard deviations) over 5 splits.

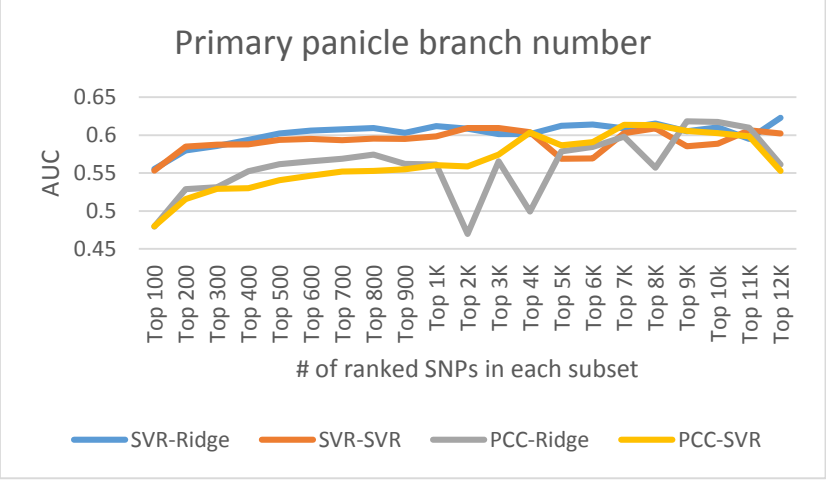

Figure 4.10 Line graph of rice primary panicle branch number data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.10 shows that primary panicle branch number phenotype prediction ability peaked at top 12k SNPs. The study found that ranking with SVR w vector absolute value, learning and predicting with RR yielded higher accuracy than ranking with PCC absolute value. It shows 0.62 (0.05) averaged prediction accuracy (standard deviations) over 5 splits.

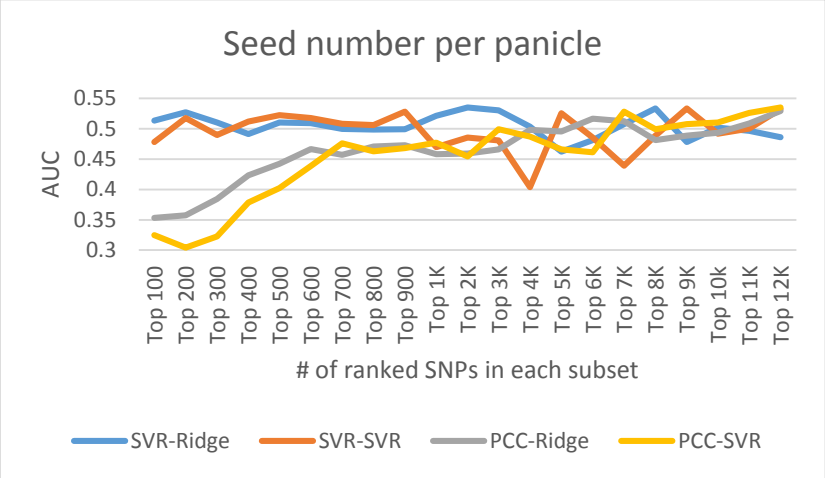

Figure 4.11 Line graph of rice seed number per panicle data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.11 shows that seed number per panicle phenotype prediction ability peaked at top 2k SNPs. The study found that ranking with SVR w vector absolute value, learning and predicting with RR yielded higher accuracy than ranking with PCC absolute value. It shows 0.53 (0.11) averaged prediction accuracy (standard deviations) over 5 splits.

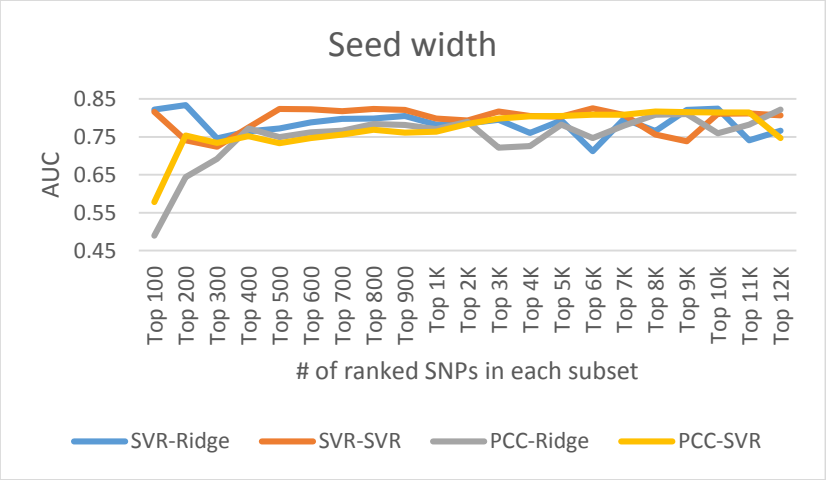

Figure 4.12 Line graph of rice seed width data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.12 shows that seed width phenotype prediction ability peaked at top 200 SNPs. The study found that ranking with SVR w vector absolute value, learning and predicting with RR yielded higher accuracy than ranking with PCC absolute value. It shows 0.83 (0.04) averaged prediction accuracy (standard deviations) over 5 splits.

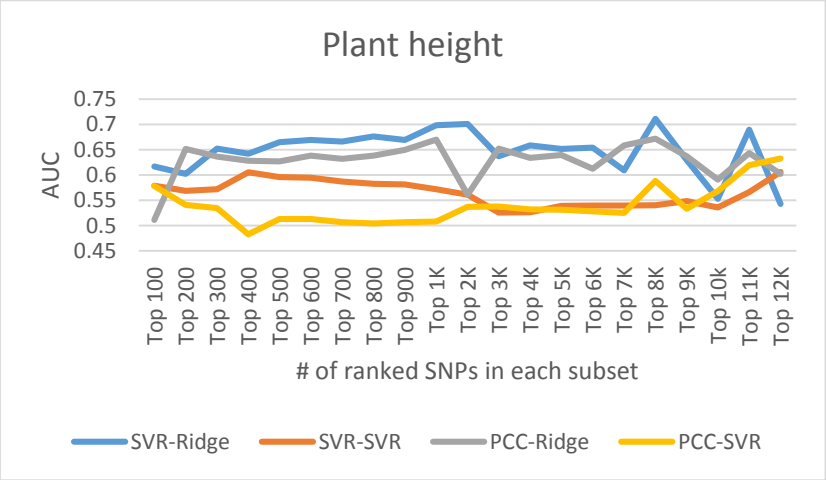

Figure 4.13 Line graph of rice plant height data prediction ability. Ranking SNPs based on SVR and PCC absolute values and predicting with SVR and RR on each subset.

Figure 4.13 shows that plant height phenotype prediction ability peaked at top 8k SNPs. The study found that ranking with SVR w vector absolute value, learning and predicting with RR yielded higher accuracy than ranking with PCC absolute value. It shows 0.71 (0.10) averaged prediction accuracy (standard deviations) over 5 splits.

CHAPTER 5

DISCUSSION

Ranking SNPs with penalized multivariate regression namely; Support Vector Regression and Ridge Regression shows significant improvement in prediction accuracy over ranking with univariate ranking in mice data sets as well as attaining slightly higher accuracy than predicting with all SNPs by 0.04 in MCH and 0.01 in CD8. In fact, prediction accuracy peaks while using 30% of the SNPs, when selecting significant SNPs by SVR w vector absolute value. This shows the potential of feature selection in eliminating data noise in genomic data. Feature selection using SVR and RR shows similar prediction accuracy, however, SVR ranking shows slightly higher accuracy. Ranking with PCC performed poorly when SNPs selected are less than 1K.

In Rice data sets, feature selection with SVR vector w absolute value generally outperformed PCC in most phenotypes except days to flower and amylose content. The averaged prediction accuracies in all rankings was consistent with less than 0.02 standard deviation in most phenotypes, this implies that ranking with only top 100 SNPs yielded high accuracy. In all rice phenotypes prediction with RR shows slightly higher accuracies than prediction with SVR except flag leaf width phenotype.

CHAPTER 6

CONCLUSION

Predicting continuous phenotypes with SNPs data only show promising high prediction accuracy. In this analysis ranking SNPs with SVR weight vector (w) yields slightly better accuracy than predicting with whole SNPs. The study also found that learning a model and predicting with RR slightly outperformed SVR. Overall SNPs ranking with multiple SNPs regression improved the prediction ability compared to ranking with PCC.

REFERENCES

1. Ober, U., Ayroles, J., Stone, E., Richards, S., Zhu, D., Gibbs, R., ...Sticker, C. (2012). Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in Drosophila melanogaster. *PLoS Genetics*, 8(5), e 1002685. doi:10.1371/journal.pgen.1002685

2. Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., ...Cai, X. (2014). Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. *PLoS ONE*, 9(3), e93017. doi:10.1371/journal.pone.0093017

3. Riedelsheimer, C., Technow, F., & Melchinger, A. (2012). Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics*, 13, 452. doi:10.1186/1471-2164-13-452

4. Solberg, L., Valdar, W., Gauguier, D., Cookson, W., Rawlins, J, Mott, R., ...Flint, J. (2006). A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. *Mamm Genome*. 17(2):129–46. PMID:16465593

5. Zhao, K., Tung, C., Eizenga, G., Wright, M., Ali, M., Price, A., ... Mccouch, S. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nature Communications*, 2, 467. doi:10.1038/ncomms1467

6. Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 88(442), 486. doi:10.1080/01621459.1993.10476299

7. Cruz, J. & Wishart, D. (2007) Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Inform*. 2, 59–77. PMCID: PMC2675494