Development and Utilization of Nested Association Mapping Population in Rice

A Dissertation

Submitted to the Graduate School of Bioagricultural Sciences of Nagoya University

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

In Agricultural Sciences

by

KITONY Justine Kipruto

Laboratory of Information Sciences in Agricultural Lands

September 2021

Abstract

In the present era of high-throughput sequencing, quantitative trait loci (QTL) mapping is no longer limited by the number of genetic markers but rather by the genetic material being deployed. To uncover QTL that will facilitate crop improvements, a genetic resource for studying the genetic architecture of traits and adaptation to the environments is necessary. Nested association mapping (NAM) is a technique of genetic mapping that integrates QTL linkage mapping and genome-wide association study (GWAS). A NAM population consists of a group of recombinant inbred lines (RILs) derived from crosses between a common parent and multiple diversity donors. NAM approach can support joint-family QTL linkage analysis and GWAS in addition to classical single-family QTL linkage analysis, and a combination of these methods is expected to have higher resolution gene mapping and power for QTL detection. Furthermore, the fixed nature of NAM RILs will facilitate analysis of genotype-environment (GxE) interactions. On the other hand, genomic selection (GS) is a new breeding approach for screening genotypes without field evaluations. For example, a breeding program can utilize GS in screening a large population having only genotype information based on models constructed using a part of the population. The NAM population is suitable for studying methods for constructing the models of genomic prediction. In the present study, a rice NAM population (*aus*-NAM) derived from crosses between T65, a *japonica* variety, as the common parent and 14 *aus* varieties as diversity donors was developed. Firstly, a NAM population consisting of 7 combinations (*aus*-NAM-I) was used to confirm the resolution of QTL mapping for days to heading (DTH), a highly heritable trait. Secondly, the population was expanded to 14 families (*aus*-NAM-II), then the population was used to perform GWAS and genomic predictions.

The *aus*-NAM-I population contained 895 RILs derived from 7 diversity donors. Out of the 7 families, 5 were constructed from *aus* varieties: Kasalath (WNAM02), Kalo Dhan (WNAM29), Shoni (WNAM31), ARC5955 (WNAM35), and Badari Dhan (WNAM39), and 2 families, DV85 (WNAM72) and ARC10313 (WNAM73) were generated at Kyushu University. Genotyping of *aus*-NAM-I was conducted using genotyping by sequencing (GBS), based on double-digestion with *Kpn*I and *Msp*I restriction enzymes. The number of clean single nucleotide polymorphisms (SNPs) in each family ranged from 2868 to 4285. A total of 887 RILs were retained for downstream genetic analysis. Analysis for population structure revealed that WNAM29 (T65 x WRC29) was an isolated group, but generally, *aus*-NAM-I showed weak stratification among the RILs and was therefore considered sufficient for QTL mapping. As a model trait, DTH was recorded in normal cultivation season at Nagoya University`s Togo field in 2015. A total of 1,786 SNP markers from GBS were used to construct a common linkage map. Single-family QTL analysis detected significant QTL on chromosomes 5, 6, 7, and 10. Joint-family QTL linkage analysis on the other hand detected similar QTL to those detected in the single-family analysis except on chromosomes 5, and new QTL on chromosomes 1, 2, and 3. The joint-family QTL peaks on chromosomes 6 and 7 appeared to be integrated peaks from single families in the population. For GWAS, parental DNA variants from whole-genome resequencing were firstly projected onto genotypes of each RIL to obtain 41,561 SNPs. GWAS was thereafter performed using a mixed linear model considering the population structure and the degree of genotype relatedness (MLM (Q + K)), with pedigree information provided as covariates. GWAS results showed significant QTL on chromosomes 6, 7, and 10. The commonly detected QTL on chromosomes 6, 7, and 10 were considered to be *RFT1, Hd3a, Hd1, Ghd7, and Ehd1* and were used to evaluate mapping accuracy in *aus*-NAM-I. The peak harboring

Ehd1 spanned approximately 741kb and included the true position of *Ehd1* locus. The QTL detected on chromosome 6 in WNAM39 using single-family analysis was in good agreement with the actual position of *Hd1*. Multiple sequence alignments confirmed that Badari Dhan (WNAM39) was the only donor parent with a functional allele of *Hd1.* Single-family QTL analysis and joint-family QTL linkage analysis correctly detected *Hd1*. However, in GWAS, a broad peak spanning wider than 6 Mbp including not only *Hd1*, but also *RFT1*/*Hd3a* was detected, and it was concluded that it would be difficult to identify genes underlying these QTL without prior information. For *Ghd7* on chromosome 7, joint-family QTL analysis showed a single peak while the GWAS peak was unclear. These results indicated that single-family QTL analysis and joint-family QTL analysis performed better in gene mapping. GWAS could be used to detect marker-trait association in local regions though. In total, *aus*-NAM-I showed sufficient performance in QTL mapping.

Next, *aus*-NAM-I was modified and expanded to 14 families (*aus*-NAM-II), by removing the 2 families DV85 (WNAM72) and ARC10313 (WNAM73) from *aus*-NAM-I and adding 9 new families. To genotype the increased population size, our group developed a new GBS method called iGBS (Vincent P. Reyes dissertation 2021). *i*GBS combines Illumina's index and GBS barcodes allowing more sample combinations compared to the conventional GBS that utilizes barcodes-only. The total number of RILs in *aus*-NAM-II population was 1,797. Traits used in the genetic analyses were evaluated in 2015, 2018, and 2019. The traits included DTH, culm length (CL), panicle length (PL), panicle rachis length (PRL), panicle number per plant (PN), panicle weight (PW), shoot weight (SW), number of primary branches per panicle (NPB), number of spikelets per panicle (NSPP), seed setting rate (SSR) and plant biomass (BM), BM was calculated as the summation of PW and SW. GWAS for DTH using *aus*-NAM-II detected known

QTL (*Ehd1, Hd1, Hd3a / RFT1, HD9, DTH2, Ghd7, Se14,* etc.) correctly, GWAS performed better than *aus*-NAM-I because of the increased population size and the crossing combinations. For other traits, known and novel QTL were detected as well. To evaluate the predictive ability of traits, 5-fold cross-validation was applied on 4 different genomic prediction models, (i) ridge regression best linear unbiased prediction (rrBLUP), (ii) Bayesian least absolute shrinkage regression (Bayesian LASSO, BL), (iii) Bayesian B (BayesB), and (iv) reproducing kernel Hilbert space regression (RKHS). The correlation coefficient (*r*) between observed and predicted phenotypes was regarded as the prediction accuracy for the model. The highest prediction accuracy for DTH in 2019 was 0.89 by RKHS model. In 2018, the prediction accuracy for phenotypes ranged from 0.89 to 0.29 for DTH and SSR respectively while the range in 2015 was from 0.90 to 0.34 for CL and BM respectively. RKHS was the most robust among the 4 statistical models. Comparison of using different numbers of the markers in modeling indicated that 2,006 GBS markers were sufficient to obtain optimum predictive ability, any additional markers by methods like parental variants projection contributed little to improving predictive ability. Based on these results, the author proposes that DTH would be the target for performing genomic prediction in actual breeding programs.

In conclusion, a rice NAM population was developed and established as a genetic resource in the present study. The NAM population showed sufficient performance in gene mapping and genomic prediction. It is expected that the immortal nature of the materials will facilitate a broad range of genetic analyses including gene discovery, modeling of traits, and analysis of genotype-environment interaction.

Table of Contents

Abstract 1

List of Figures 6

List of Tables 7

Chapter 1 General Introduction 8

Background 8

Mapping Populations 10

Considerations for Constructing NAM Population 13

Applications of NAM Population 17

Studies in the Dissertation 20

Figure 21

Chapter 2 Development of *aus*-NAM-I Population in Rice 22

Introduction 22

Materials and Methods 24

Results 26

Discussion 27

Figures and Tables 28

Chapter 3 QTL Mapping using *aus*-NAM-I Population 34

Introduction 34

Materials and Methods 36

Results 39

Discussion 42

Figures and Tables 44

Chapter 4 Expansion of *aus*-Derived NAM Population in Rice (*aus*-NAM-II) 55

Introduction 55

Materials and Methods 56

Results 58

Discussion 59

Figures and Tables 61
Chapter 5 GWAS and Genomic Predictions using *aus*-NAM-II Population.. 67
Introduction 67
Materials and Methods 70
Results 75
Discussion 78
Figures and Tables 80
Chapter 6 General Conclusions 100
Summary 100
Challenges, Limitations and Opportunities 101
Future Perspectives 102
Acknowledgments 103
Literature Cited 104
List of Publications 119

List of Figures

Figure 1-1. Common mapping populations. 21
Figure 2-1. *aus*-NAM-I population development. 28
Figure 2-2. *aus*-NAM-I population structure (PC 1-3). 29
Figure 2-3. *aus*-NAM-I population structure (PC 2-6). 30
Figure 2-4. Phenotype diversity in an *aus*-NAM population. 31
Figure 3-1. Genes in rice photoperiodic flowering pathways. 44
Figure 3-2. Days to heading frequency distributions. 45
Figure 3-3. Interval mapping for days to heading in *aus*-NAM-I. 46
Figure 3-4. Joint-family QTL and additive effects in *aus*-NAM-I. 47
Figure 3-5. GWAS for DTH using GLM in *aus*-NAM-I. 48
Figure 3-6. GWAS for DTH using MLM (Q+K) in *aus*-NAM-I. 48
Figure 3-7. Local GWAS for days to heading in *aus*-NAM-I. 49

Figure 3-8. Accuracy for *Hd1* mapping. 50
Figure 4-1. *aus*-NAM-II population development. 61
Figure 4-2. Correlations of phenotypes in 2015 and 2018. 62
Figure 4-3. *aus*-NAM-II population structure (PC 1-3). 63
Figure 4-4. *aus*-NAM-II population structure (PC 2-6). 64
Figure 5-1. BGLR package regression prior assumptions. 80
Figure 5-2. Linkage disequilibrium (LD) decay pattern. 81
Figure 5-3. GWAS for days to heading in 2019 using *aus*-NAM-II. 82
Figure 5-4. GWAS for phenotypes in 2018 using *aus*-NAM-II. 85
Figure 5-5. GWAS for phenotypes in 2015 using *aus*-NAM-II. 87
Figure 5-6. Genomic prediction for days to heading in 2019. 88
Figure 5-7. Genomic prediction for eleven traits in 2018. 90
Figure 5-8. Genomic prediction for ten traits in 2015. 92

List of Tables

Table 2-1. List of parents and RILs in *aus*-NAM-I population. 32
Table 2-2. Taxa and SNPs in *aus*-NAM-I. 33
Table 3-1. Joint-family linkage map statistics in *aus*-NAM-I. 51
Table 3-2. QTL in *aus*-NAM-I using single-family QTL analysis. 52
Table 3-3. QTL in *aus*-NAM-I using joint-family QTL analysis. 53
Table 3-4. Additive effects in *aus*-NAM-I using joint-family QTL analysis. 54
Table 4-1. List of RILs in *aus*-NAM-II population. 65
Table 4-2. Taxa and SNPs in *aus*-NAM-II population. 66
Table 5-1. GWAS and candidate genes in 2019 using *aus*-NAM-II. 93
Table 5-2. GWAS and candidate genes in 2018 using *aus*-NAM-II. 94
Table 5-3. GWAS and candidate genes in 2015 using *aus*-NAM-II. 97
Table 5-4. Prediction accuracy for traits in 2019 and 2018. 98
Table 5-5. Prediction accuracy for traits in 2015. 99

Chapter 1 General Introduction

Background

Rice (*Oryza sativa* L.) is a perennial crop of economic importance and staple food for more than half of the world's population (FAO 2006). It is believed that rice was domesticated approximately 8,000 to 10,000 years ago from wild rice (*O. rufipogon* Griff.) in East Asia (Doebley *et al.*, 2006). Even though rice is a facultative short-day plant, through natural mutations and artificial selection, rice has evolved and adapted to a wide range of geographical areas and seasons. The key factor that enabled rice broad adaptations is linked to photoperiod sensitivity and flowering time also known as days to heading (DTH) (Yano et al. 2000; Tsuji et al. 2008; Takahashi et al. 2009; Huang et al. 2012a).

To improve rice productivity and futureproof ourselves against the effects of climate change on food production, it is important to identify the genes underlying quantitative trait loci (QTL), by which most of the agronomic traits are regulated. One way of achieving this is by finding significant DNA markers that tag the QTL. DNA markers tagging a QTL can be used for introgressions or stacking of novel variations into elite adapted lines, mostly from wild/exotic un-adapted species. This process is commonly known as marker-assisted selection (MAS) (Moose and Mumm, 2008). Recently, methods that allow for the selection of candidate breeding materials based on the predicted trait values were proposed. These methods use DNA markers covering the whole genome even if no significant QTL is detected. This approach is commonly known as genomic selection (GS) (Meuwissen *et al.*, 2001).

Modern plant breeding is a predictive science driven by new technologies and knowledge with productivity/quality, time, and cost under consideration (Crossa *et al.*, 2021). Recently, many genomics data have been made public (Kojima *et al.*, 2005; Ebana *et al.*, 2008; The 3000 rice genomes, 2014; Sun *et al.*, 2017). These genetic resources are useful for QTL mapping or breeding. Unlike the traditional phenotype-based selection, the new breeding approaches i.e. genomic selection (GS), utilize such free genomics datasets. GS has been touted as the promising approach for shortening breeding time and mitigating against the increasing cost of phenotyping. Evaluation of a genomic model predictive ability (PA), a key parameter in GS is an important issue.

QTL mapping accuracy mainly depends on (1) recombination fraction between QTL and the available markers, (2) QTL heritability, and (3) the size of the mapping population. Other factors that influence QTL detection power include (1) population structure, (2) phenotypic variations, (3) genotype information quality, and (4) robustness of computer software (Singh and Singh, 2015). The majority of the factors that affect QTL accuracy are directly associated with the genetic material deployed in the study, a controlled or structured mapping population is, therefore, a good starting point in the genetic analysis of traits.

Mapping Populations

Mapping populations in rice are generally a group of lines obtained from controlled crosses between two or more founders. Mapping populations are suitable for QTL linkage mapping using the principles of Mendelian inheritance (Singh and Singh, 2015). Mapping populations are categorized into three main generations (**Figure 1-1**).

The first-generation mapping population consists of bi-parental populations (BP), examples include (a) F_2 (b) F_2-derived F_3 ($F_{2:3}$) (c) backcross populations e.g. (BC_1, BC_2) (d) Near isogenic lines (NILs) (e) recombinant inbred lines (RILs) (f) doubled haploid population (DH) (g) chromosome segment substitution lines (CSSLs). Most major genes/QTL in rice were genetically identified using bi-parental populations (Yano *et al.*, 2000; Ashikari *et al.*, 2005; Yamamoto *et al.*, 2012). The backcross progeny enables precise estimation of allelic effects thus used in QTL fine-mapping. The limited allele richness in BP i.e. only two possible alleles segregating for each locus is an apparent disadvantage. Efforts to address the issue through advanced inter-cross design (AIC) (Lee *et al.*, 2002; Balint-Kurti *et al.*, 2010; Fitz Gerald *et al.*, 2014) were not successful either, similar results could simply be achieved through increasing bi-parental population size.

The second-generation mapping population came about with the advent of next-generation sequencing (NGS) technology, NGS enabled direct detection of genetic loci associated with traits, the so-called genome-wide association study (GWAS) (Ogura and Busch 2015). GWAS utilizes natural germplasm collections i.e. landraces, breeding lines, and varieties that have accumulated recombination events both recent or historic, this attribute gives GWAS a higher gene mapping resolution (Mogga et al. 2018; Yano et al. 2016; Norton et al. 2018). However, unaccounted population structures/stratifications and hidden relatedness make

GWAS prone to false associations. Additionally, the filtering of minor allele frequency (MAF) to thresholds like 0.02, or 0.05 for QTL integrity lowers GWAS detection of true effects that have large/small effects (Cockram and Mackay, 2018). The absence of pedigree information in GWAS diversity panels prohibits classical pedigree-based haplotype mapping, consequently, reduced accuracy to consider rare alleles in analysis (Xiao et al. 2017; Zhu et al. 2008; Korte and Farlow). These factors made the criteria for declaring "statistically significant" too strict and limited the statistical power of GWAS. To date, only major-effect QTL have been identified by plant GWAS studies such as *OsSPL13* (Si *et al.*, 2016), *bZIP73*(Liu *et al.*, 2018) and four novel genes associated with agronomic traits(Yano *et al.*, 2016).

The third-generation mapping populations include Nested Association Mapping (NAM) (Yu et al. 2008), Multi-parent Advanced Generation Inter-Crosses (MAGIC) (Dell'Acqua et al. 2015), and Random-open-parent Association Mapping (ROAM) (Xiao et al. 2016). Multi-parental populations have advantages such as (i) allele richness from the diversity donors (ii) weak population structure due to additional recombination (historic and by artificial crossing), and (iii) flexibility to be used as a pre-breeding tool.

Various QTL mapping approaches have been proposed based on the mapping population category. In convectional QTL linkage analysis, BP with recent genetic recombination is utilized. This method requires few markers to find QTL but due to low allele richness in BP, linkage mapping is synonymous with low mapping resolutions. On the other hand, GWAS takes advantage of historic recombination (high allelic richness) to scan polymorphic DNA sites in linkage disequilibrium (LD) with the variation of trait resulting in high mapping resolution. However, GWAS requires extensive knowledge of population structures before analysis.

To take the advantage of both historic and recent recombination events such as low marker density requirements, high allele richness, high mapping resolution, and high statistical power NAM approach was proposed (McMullen *et al.*, 2009). NAM population consists of independent RIL populations derived from crossing several diverse donor parents with a common parent (Yu *et al.*, 2008). The first NAM population to be constructed was in maize. Maize NAM contains 5000 RILs consisting of 25 families. The most successful commercial inbred line (B73) was used as the common parent (McMullen et al. 2009). The population was successfully used to profile the genetic architecture of agronomic traits, such as flowering time, leaf architecture, stalk strength, and plant height (Buckler et al. 2009; Tian et al. 2011; Peiffer et al. 2014). Following the successes in maize genetic analyses, NAM type populations were developed in other crops like rice (Fragoso et al. 2017), wheat (Bajgain et al. 2016), barley (Maurer et al. 2015), soybean (Song et al. 2017), sorghum (Bouchet et al. 2017), and rapeseed (Hu et al. 2018). Other NAM-based mating designs include near-isogenic lines (NIL-NAM) and double haploid (DH-NAM) (Singh and Singh, 2015; Li *et al.*, 2016; Nice *et al.*, 2016).

Considerations for Constructing NAM Population

Parents

The choice of NAM parents should be selected bearing in mind that the NAM population can serve both as a genetic analysis tool and a resource for breeding. NAM population common parent is often selected from an elite variety that is environmentally adapted. The common parent should support multi-location evaluations and adoption as a new variety in farmers` fields (McMullen *et al.*, 2009). Donor parent/diversity founder, on the other hand, should be selected based on breeder/geneticist purpose i.e. the trait of interest. The diversity donors should contain maximum genetic diversity for the focal trait. They are often selected from adopted cultivars, germplasms, landraces, and wild species. Diversity donors may also be selected based on allelic diversity for biotic/abiotic stress tolerance (Kihupi, 2001; Travis *et al.*, 2015; Norton *et al.*, 2018).

Mating Design

The number of founders/ RILs per family affects recombination events/allele frequencies which in turn influence QTL mapping powers. Several mating designs have been proposed and implemented: (i) Recombinant inbred Line-NAM (RIL-NAM). In this kind of design, the diversity donors are selected from established germplasms, the F_2 generation is advanced using the single seed descent (SSD) method to obtain homozygosity, RIL-NAM constitutes single-family RILs (McMullen *et al.*, 2009; Kitony *et al.*, 2021) and is the common type of NAM population, (ii)Backcross-NAM/Advanced Backcross-NAM (BC-NAM/AB-NAM) populations, in this design, diversity donors are selected from un-adapted germplasms such as wild species or landraces. An example of this design is a backcross of 25 wild barley accessions with an elite variety (Nice *et al.*, 2016). The smaller the recombinant region from the exotic donor parent the easier will be the genetic characterization and phenotyping in this type of design. (iii) Double Haploid-NAM (DH-NAM) populations, this is a special type of design only used in crops with advanced DH protocols (attain homozygosity within single generation development) such as maize (Gireesh *et al.*, 2021).

Phenotyping

Because the size of NAM population is often big, phenotype evaluations are conducted in un-replicated field designs with serpentine plot numbering (Federer and Crossa, 2012). Other field designs include; augmented randomized complete block design (RCBD) which is suitable when the number of genotypes is large and one-way elimination of heterogeneity is required; augmented split-plot design, suitable when testing two different factors with varying importance like genotypes and spacing, etc. With the availability of resources and technical knowledge, high throughput phenotyping techniques such as using information communication technology (barcodes, tablet terminals, etc.) are recommended options for NAM phenotyping (Araus and Cairns, 2014).

Genotyping

The knowledge about organism's ploidy, genome size, heterozygosity, and repetitive sequences are necessary knowledge required to appropriately genotype NAM population (Ray and Satya, 2014). However, the development of next-generation sequencing (NGS) technology has eliminated most of the difficulty. Each RIL in the NAM population is often genotyped using low-cost sequencing techniques like genotyping by sequencing (GBS) or single nucleotide polymorphism (SNP) based arrays. After genotyping RILs, parental lines are genotyped using high-density sequencing techniques like whole-genome shotgun sequencing. Parental genomic sequences are thereafter projected onto recombination blocks in a single family linkage map deriving moderate to high density genotyped NAM population. Since sequencing costs mainly vary according to the technology and the number of lanes used in a flow cell. With the advent of NGS, many options to cut costs for genotyping by pooling multiple samples have been developed (Nakano and Kobayashi, 2020). In a method reported by Poland et al.(Poland *et al.*, 2012), the density of the markers and number of the samples pooled can be optimized by choosing appropriate restriction enzymes. Our group developed a combination of Poland-style GBS and Illumina's genuine index system (iGBS) for obtaining NAM genotypes in the present study (Vincent P. Reyes dissertation 2021).

Statistical Methods

To realize NAM's optimum statistical powers, all families should be analyzed together via joint-family QTL linkage analysis and GWAS (NAM-GWAS). For joint-family QTL linkage mapping, two algorithms are available (i) Joint Inclusive Composite Interval Mapping (JICIM)(Li *et al.*, 2011), which uses stepwise regression model or principal component regression to select co-factors used in the analysis (ii) Joint Composite Interval Mapping (JCIM)(Li *et al.*, 2016), which uses the least absolute shrinkage and selection operator (LASSO) regression to select co-factors used in the analysis. Joint-family QTL linkage analysis can detect more than two alleles per locus compared to single-family linkage analysis (Buckler *et al.*, 2009a; Ogut *et al.*, 2015). In Arabidopsis NAM, JICIM detected more number of QTL and with higher LOD scores compared to single-family QTL analysis (Li *et al.*, 2011). In rapeseed NAM, JCIM displayed higher detection power compared to single marker modeling (Li *et al.*, 2016).

Natural population GWAS can outperform NAM-GWAS in QTL mapping when the size of a population is sufficient (Bouchet *et al.*, 2017). However, increasing the number of founders can make NAM-GWAS very competitive (Gage *et al.*, 2020). NAM-GWAS takes advantage of enriched rare alleles derived from one or few diversity donors to detect QTL. Compared to natural population GWAS, NAM-GWAS can best detect QTL in high/low heritability using a smaller population size due to controlled allelic diversity.

Although NAM can benefit from both joint-family QTL linkage analysis and linkage disequilibrium (LD) (Lu *et al.*, 2010). Contrasting results from an integrated approach have been reported such as the independent mapping of hypersensitive defense response (HR) in maize NAM-GWAS (Olukolu *et al.*, 2014) that showed 36 genes co-localizing within 23 QTL identified by joint-family QTL linkage analysis while leaf architecture analysis (Tian *et al.*, 2011) displayed non-overlapping QTL between the joint-family linkage and LD mapping approaches.

NAM Population Construction Limitations

The likelihood of sterility (Wambugu *et al.*, 2015) is high especially when donors are from genetically distant germplasms, this phenomenon can affect the sizes of NAM families and by extent allele frequencies. In addition, it takes a lot of time to stabilize segregation distortions in developed populations if present.

Another limitation noticed in NAM type design was limited haplotype diversity. With 26 founders in maize NAM for example, a maximum of only 50 recombinant haplotypes could be generated. To increase haplotypes in maize NAM, an additional number of donors (Gage *et al.*, 2020) or reference/common parents was proposed (Guo *et al.*, 2010; Cockram and Mackay, 2018).

Applications of NAM Population

Explore the Genetic Basis of Key Agronomic Traits

NAM is a tailor-made mapping population that can effectively dissect the genetic basis of key agronomic traits, these traits may be controlled by rare and low effects alleles. Plant traits are mostly dependent on interactions between environments (E), genotypes (G), management practices (M), and microbes interactions(M) (hereafter: G × E x M x M) (Holland 2007). The genetically fixed nature of NAM RILs permits analysis of these interactions by repeated and multi-locational trials. Flowering time for example was found to be influenced by additive effects by environment interaction, and several flowering time genes were profiled (Buckler *et al.*, 2009a). Putative QTL for days to heading in rice were also confirmed using NAM population constructed using IR64 *indica* with 10 diverse tropical *japonica* lines (Fragoso *et al.*, 2017). Furthermore, NAM design using exotic donor germplasm has shown an opportunity to detect novel useful alleles for salinity tolerance (Saade *et al.*, 2016). Simply put, NAM is suitable for calculating the sizes of QTL effects and testing scientists' hypotheses for traits.

Discriminate Linkage from Pleiotropy

Co-inheritance of QTL and traits is influenced by either physical proximity of QTL (linkage) or multiple effects of a single QTL/gene (pleiotropy) (Gireesh *et al.*, 2021). Genetic mating design in NAM allows genome shuffling during RIL development (Fragoso *et al.*, 2017). The resultant recombination events in the NAM population enable us to distinguish linkage from pleiotropy effects on traits. A study on maize NAM (McMullen *et al.*, 2009), showed a strong negative correlation between days to anthesis (DTA) and northern leaf blight (NLB) resistance in the founder lines (r = −0.59) compared to RIL-NAM (r = −0.32). The authors concluded that NLB resistance and DTA were confounded by population structure and to some extent genetic linkage within families, rather than the pleiotropic effects of DTA on NLB resistance (Poland *et al.*, 2011).

Detect Segregation Distortions

Genotype information in NAM can be used to identify recombination events and segregation distortions (SD). Due to NAM diversity donors, a high degree of recombination events and SD is expected. In rice, a total of 18.9 recombination events was detected while the recombination rate was close to the expected value of 4.1 cM: Mb (Chen *et al.*, 2002; Fragoso *et al.*, 2017). This advantage is common with usual RIL populations. On the other hand, SD regions should be taken into account when interpreting QTL results, because SD might result in false-positive or false-negative QTL.

Genomic Selection

Recent development in genome sequencing technologies has made it possible to use only high-throughput genetic markers (e.g. SNPs) to select lines with desired traits. This technique/model is known as genomic selection (GS) or genomic prediction (GP) (Meuwissen *et al.*, 2001). Briefly, models are trained using a subset of data with phenotype and genotype information, marker effects are then estimated which are in turn used to model estimated breeding values (GEBVs) for lines. Based on GEBVs best performing lines (good ranking) are selected for onward evaluations.

Generally, genomic prediction methods differ in their assumptions for estimating variances in complex traits (Wang *et al.*, 2018). Examples of GS methods include ridge-regression-based (rrBLUP), Bayesian-based, and machine learning-based and they vary mainly on prior at α (equation 1) (Wang et al. 2018).

Most of these methods are for estimating additive genetic effects. Since the number of markers (k) is often bigger than lines (n), GS methods treat marker effects as random to eliminate the insufficient degree of freedom and multi collinearity issues. The genomic prediction model is generally described by equation (1).

$$y = X\beta + Z\alpha + \varepsilon \qquad (1)$$

where y is a vector of observation; β represents a vector of fixed effects (e.g. PCA) for which the prior distribution is often assumed flat, α is a vector of random effects, and ε is a vector of residuals. X and Z are matrices for fixed effect and random effects respectively. The residuals distribution is assumed normal with a mean value of zero and covariance matrix represented by $R\sigma^2\varepsilon$, where R is an identity matrix, $\sigma^2\varepsilon$ has a scaled inverse chi-square distribution (Wang *et al.*, 2018).

Because prediction accuracy affects genetic gains (Li *et al.*, 2018), several studies have used NAM populations to investigated factors affecting prediction accuracy (Xavier *et al.*, 2016; Zhang *et al.*, 2019). Accurate marker effects estimation is key to both genetic gain and prediction accuracy. The prediction accuracy is influenced by the following factors: the number of markers, statistical models, training population size, relationships between training and breeding population, the genetic architecture of the trait, the heritability of a trait(positively correlated with accuracy), and the interplay between all mention the factors with the environment (Spindel *et al.*, 2015; Fragoso *et al.*, 2017).

Studies in the Dissertation

My research covered the development of an *aus*-derived nested association mapping (*aus*-NAM) population and its utilization in rice genetics. Briefly, in chapter 1, I introduced commonly used mapping populations and things to consider during the construction of the NAM type mapping populations. In chapter 2, I described the construction and genotyping of our first *aus*-derived nested association mapping population(*aus*-NAM-I). I also clarified the population structure of *aus*-NAM-I. In chapter 3, I analyzed phenotypic variations in *aus*-NAM-I RILs. I also characterized the genetic basis of DTH using single-family QTL mapping, joint-family QTL mapping, and the methods based on genome-wide association study (GWAS). In chapter 4, I described the modification and expansion of *aus*-NAM-I to construct *aus*-NAM-II. In chapter 5, I performed GWAS and genomic prediction (GP) using *aus*-NAM-II. The predictive ability discerned by various GP models was profiled. I concluded my dissertation by highlighting the prospects of implementing genomic selection using *aus*-NAM population.

Figure

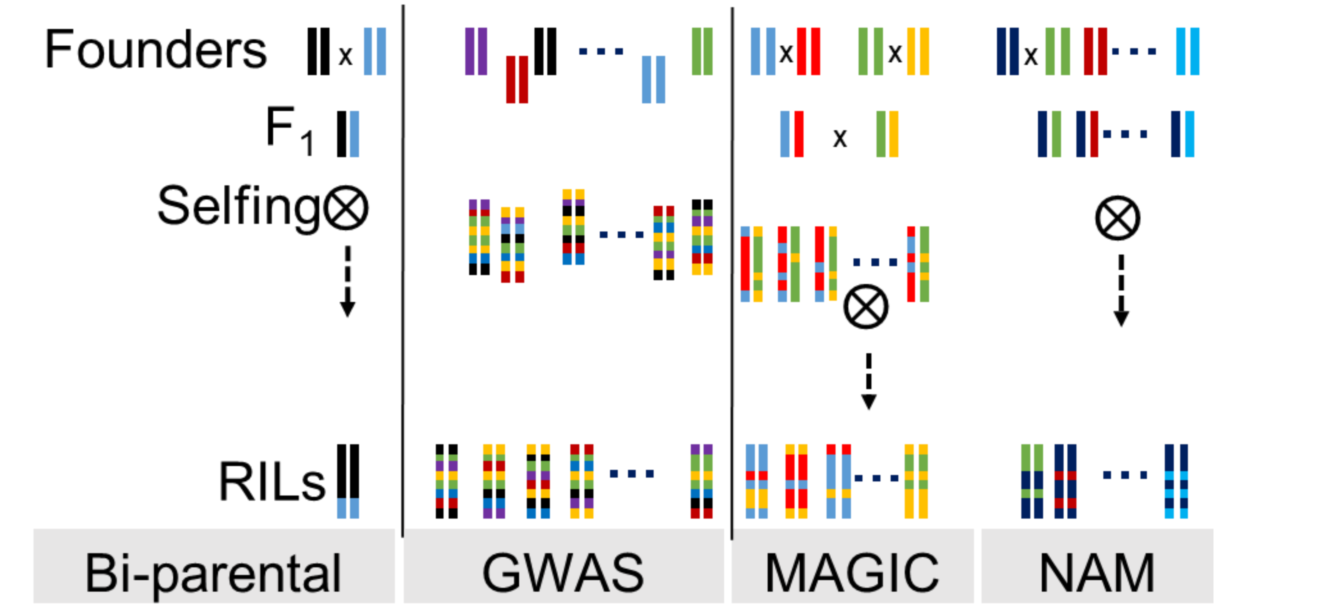


Figure 1-1. Common mapping populations.
Common mapping populations. 1st generation includes bi-parental populations (BP), BP is created from crossing two founders followed by repeated selfing to create a new inbred line whose genome is a mosaic of the parental genomes. Recombination bins are often large, limiting mapping resolution. 2nd generation includes, diversity also called association panels (GWAS), GWAS populations are samples of natural variation from a larger, existing population that has accumulated historical recombination events and mutations. They frequently have greater recombination and allelic richness than 1st generation and 3rd generation populations but are often burdened with inherent population structure that can be difficult to control during analysis. 3rd generation includes; MAGIC populations are often derived from 8 or 16 parental lines (Only four are shown). NAM populations consist of several RIL families that share a common parent (shown in dark blue). This results in improved resolution and a greater number of alleles represented. 3rd generation populations have improved resolution and allelic richness relative to 1st generation populations. Black 'x's indicate crosses between parents and circled '⊗'s indicate self-mating until inbred. Ellipses indicate many other individuals in the population or family(Gage *et al.*, 2020).

Chapter 2 Development of *aus*-NAM-I Population in Rice

Introduction

Improvement of rice (*Oryza sativa* L.) production has been achieved by the development of new varieties and optimization of cultural practices. In rice, genome sequencing enhanced the identification of causal genes related to yield (Ikeda *et al.*, 2013). Most of these genes/QTL were genetically identified using bi-parental populations, combined with the development of backcrossed progeny (Yano and Sasaki, 1997). The backcross progeny enabled precise estimation of allelic effects thus fine-mapping of the target loci. However, limited allele richness is an apparent disadvantage in these types of gene mapping populations.

On the other hand, the development of next-generation sequencing (NGS) technology-enabled direct detection of genetic loci associated with traits, the so-called genome-wide association study (GWAS) (Ogura and Busch, 2015). In plants, GWAS utilizes natural germplasm collections i.e. landraces, breeding lines, and varieties that have accumulated recombination events both recent or historic, this attribute gives GWAS a higher gene mapping resolution (Yano *et al.*, 2016; Mogga *et al.*, 2018; Norton *et al.*, 2018). However, unaccounted population structures or stratifications make GWAS prone to false associations. Additionally, the absence of pedigree information in diversity panels prohibits classical pedigree-based haplotype mapping, consequently, reduced statistical powers (Korte and Farlow; Zhu *et al.*, 2008; Xiao *et al.*, 2017).

To combine the advantage of bi-parental populations and diversity panels, multi-cross mating designs consisting of diverse donors were proposed. Examples of multi-cross designs include Nested Association Mapping (NAM)

(Yu *et al.*, 2008), Multi-parent Advanced Generation Inter-Crosses (MAGIC) (Dell'Acqua *et al.*, 2015), and Random-open-parent Association Mapping (ROAM) (Xiao *et al.*, 2016). Multi-parent populations have advantages such as (1) allele richness coming from the diversity donors (2) weak population structure due to additional recombination (historic and by artificial crossing), and (3) flexibility to be used as a breeding utility.

The first NAM population was reported in maize (McMullen *et al.*, 2009). The population was successfully used to profile the genetic architecture of agronomic traits, such as flowering time, leaf architecture, stalk strength, and plant height (Buckler *et al.*, 2009b; Tian *et al.*, 2011; Peiffer *et al.*, 2014). Following the successes in maize, NAM populations were developed in other crops like rice (Fragoso *et al.*, 2017), wheat (Bajgain *et al.*, 2016), barley (Maurer *et al.*, 2015), soybean (Song *et al.*, 2017), sorghum (Bouchet *et al.*, 2017), and Rapeseed (Hu *et al.*, 2018).

Asian rice (*O. sativa*) is classified into five major varietal groups, namely, temperate *japonica*, tropical *japonica*, *indica*, *aus*, and aromatic (Garris *et al.*, 2005). *aus* rice varieties are considered to have evolved from annual *Oryza nivara* found in Bangladesh, northeast India, Nepal, and northern Myanmar. Most of the *aus* varieties exhibit photoperiod insensitivity, a source of local environment adaptation (Travis *et al.*, 2015). Moreover, *aus* is known to possess genetic properties for enhanced yield traits (Norton *et al.*, 2018); tolerance to rice blast (Takehisa *et al.*, 2009), bacterial blight (Kihupi, 2001), submergence (Xu *et al.*, 2006) and phosphorus (Gamuyao *et al.*, 2012), etc.

In chapter 2, I describe the construction of our first *aus* derived nested association mapping population in rice, hereinafter referred to as ***aus*-NAM-I**.

Materials and Methods

Plant Materials

aus-NAM-I was built using a temperate *japonica* variety, Taichung 65 (T65) as the common female parent. Five *aus* cultivars, Kasalath, Kalo Dhan, Shoni, ARC5955, and Badari Dhan, kindly supplied by the National Agricultural Research Organization (NARO) Genebank, Tsukuba, Japan, were used as diversity donor parents (founders) (Kojima et al. 2005) (**Figure 2-1**). The five *aus* varieties were crossed with T65, and RILs were developed from F2 generation using single seed descent (SSD) method to obtain F5 in 2015. The RILs (F13) derived from T65 x DV85 and T65 x ARC10313 were generated at Kyushu University, Fukuoka, Japan, and kindly provided through National Bioresource Project. The 7 families of RILs were designated as WNAM02 (Kasalath), WNAM29 (Kalo Dhan), WNAM31 (Shoni), WNAM35 (ARC5955), WNAM39 (Badari Dhan), WNAM72 (DV85), and WNAM73 (ARC10313).

Genotyping

For genotyping *aus*-NAM RILs, approximately 5 cm of leaf tissues from each line were sampled into paper envelopes. The samples were dried in an oven at 53 °C overnight and then stored at 6 °C. Total DNA of RILs and founders were extracted using a modified Dellaporta method (Dellaporta *et al.*, 1983). DNA qualities were checked by electrophoresis on a 0.6% agarose gel in 1x Tris/Borate/EDTA (TBE; 40mmol/L Tris, 20 mmol/L acetic acids, and 0.5 mmol/L EDTA-2Na). Quantiflour dsDNA system (Promega, WI, USA) was used for the quantification of the extracted total double-stranded DNA.

GBS libraries were prepared using reported protocols (Poland *et al.*, 2012; Furuta *et al.*, 2017). Briefly, 200 ng (20ng x 10µl) of individual samples of DNA were double-digested with *Kpn*I and *Msp*I enzymes (New England Biolabs, MA, USA), ligated to barcode adaptors, pooled (multiplexed), and purified using QIAquick PCR purification Kit (Qiagen). Flowcell primers were added to the pooled samples and amplified. The library was sequenced using Illumina MiSeq (Illumina, CA, USA).

Raw sequences were processed using the TASSEL-GBS pipeline (Glaubitz *et al.*, 2014) with default parameters, except (1) minimum allele frequency, higher than 0.02 (2) minimum locus coverage set to 0.3 (3) heterozygous sites and taxa that exceeded 0.125 were filtered out. Os-Nipponbare-IRGSP-1.0 (Kawahara *et al.*, 2013) was used as the reference for SNP identification. SNPs were further filtered based on parental polymorphism, sites that were polymorphic between parents but monomorphic in each parent were only included. Additionally, missing data were imputed using the FSFHap algorithm (Swarts *et al.*, 2014).

Population Structures Estimation

Genotype information obtained from GBS was used to estimate population stratifications. Probabilistic PCA (PPCA) algorithm in the Bioconductor package PCA methods (Stacklies *et al.*, 2007) and implemented on R(R Core Team, 2020) was deployed in this analysis.

Results

aus-NAM-I Population

Out of the seven families used in this study, five (WNAM02, WNAM29, WNAM31, WNAM35, and WNAM39) were newly developed. The *aus*-NAM-I population development scheme is shown in **Figure 2-1** and the numbers of plants in F2 and F5 are listed in **Table 2-1**. Because of hybrid sterility and late heading, a substantial number of the plants in F2, F3, and F4 could not be harvested. The residual rate for families at F5 ranged from 46.8% to 74.4%. In total, 895 RILs ranging from 107 to 163 per family were obtained (**Table 2-1**).

Genotyping

The retained number of SNPs for onward analyses after the filtering process ranged from 2,868 to 4,285 in 7 families while 887 lines without excess heterozygosity (>0.125) were retained (**Table 2-2**).

Population Structure

Estimation of population structure using probabilistic principal component analysis (PPCA) showed fairly controlled stratification, the R^2 values were: 0.067, 0.044, 0.041, 0.038, 0.037 and 0.034 for PC1 to PC6 respectively (**Figure 2-2** and **Figure 2-3**). WNAM29 formed an isolated group from other families (**Figure 2-2**).

Discussion

The Development of aus-NAM-I Population

NAM population brings the advantages of joint-family QTL linkage mapping and GWAS in dissecting the genetic basis of complex traits (Yu *et al.*, 2008; McMullen *et al.*, 2009; Tian *et al.*, 2011). In this study, we have developed and characterized *aus* derived NAM population using *aus* varietal group as diversity donors. Because of the hybrid sterility and late heading, a substantial number of lines were lost in the process of SSD. However, the population size allowed detections of known and novel QTL (**Chapter 3**).

Genotyping of aus-NAM-I Population

A GBS method (Poland *et al.*, 2012; Furuta *et al.*, 2017) was used to obtain marker genotypes for *aus*-NAM RILs. DNA was digested with *KpnI* and *MspI* enzymes (New England Biolabs, MA, USA), which are "rare cutter" and "common-cutter" respectively. An approximately 5000 sites were generated after digestion by the two enzymes and GBS. With the advancement of NGS and reducing the cost of sequencing, the use of more "frequent" cutters instead of *KpnI* or even whole-genome sequencing for all of the lines will be possible.

Population Structure

While examining the population structure of *aus*-NAM, PPCA showed a weak stratification, with the half-sib RILs dispersed around the T65 (common female parent) and *aus* (diversity donor parent) (**Figure 2-2** and **Figure 2-3**). This showed that *aus*-NAM retained the genetic diversity but fairly controlled the population structure (Myles *et al.*, 2009). A similar low population structure was reported in oilseed rape NAM (Hu *et al.*, 2018) and sorghum NAM (Bouchet *et al.*, 2017).

Figures and Tables

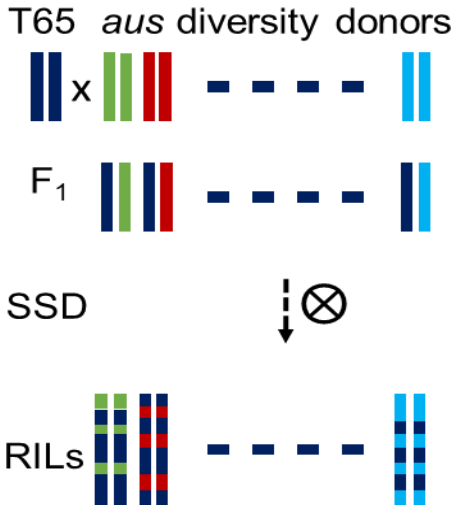


Figure 2-1. *aus*-NAM-I population development.
Depiction of *aus*-NAM-I population construction, temperate *japonica* Taichung 65 (T65) as the common female parent was crossed with 7 *aus* diversity donor parents. RILs were thereafter developed from F2 generation using single seed descent (SSD) method. Some families are not shown in the picture for the sake of visualization. The 7 families RILs were designated as WNAM2 (Kasalath), WNAM29 (Kalo Dhan), WNAM31 (Shoni), WNAM35 (ARC5955), WNAM39 (Badari Dhan), WNAM72 (DV85) and NAM73 (ARC10313). Black 'x's indicate crosses between parents and circled '⊗'s indicate self-mating. Ellipses indicate other individuals in the population.

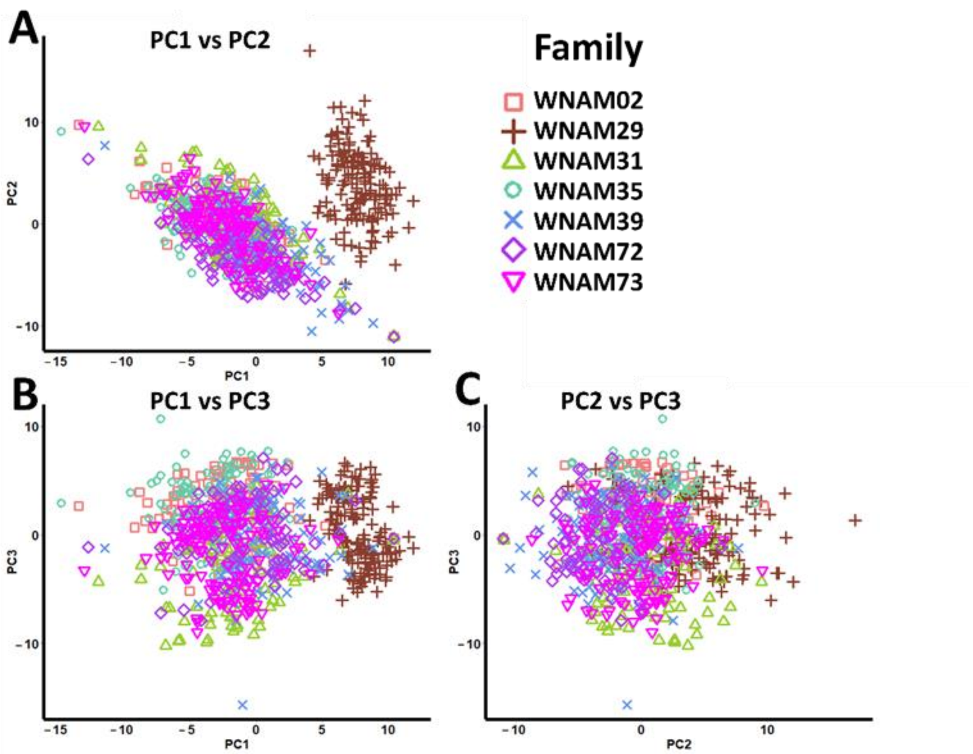


Figure 2-2. *aus*-NAM-I population structure (PC 1-3).

Probabilistic principal component analysis (PPCA) was used to estimate population structure, (A) PC1 vs PC2, (B) PC1 vs PC3, and (C) PC2 vs PC3. Different shapes and colors represent different *aus*-NAM-I families.

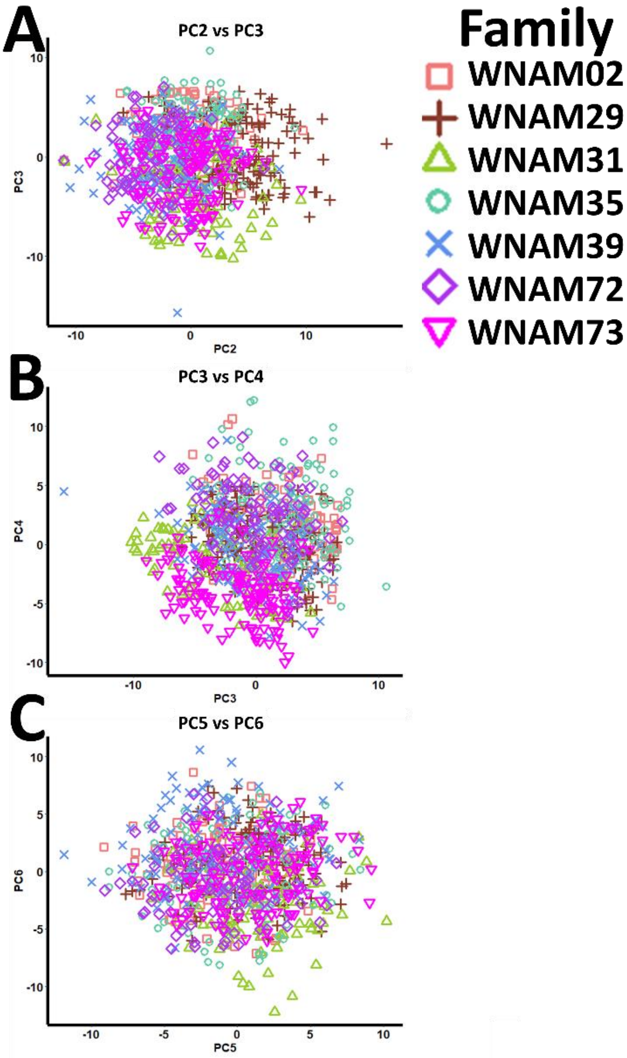


Figure 2-3. *aus*-NAM-I population structure (PC 2-6).

Probabilistic principal component analysis (PPCA) was used to estimate population structure, (A) PC2 vs PC3 (B) PC3 vs PC4 (C) PC5 vs PC6. Different shapes and colors represent different *aus*-NAM-I families.

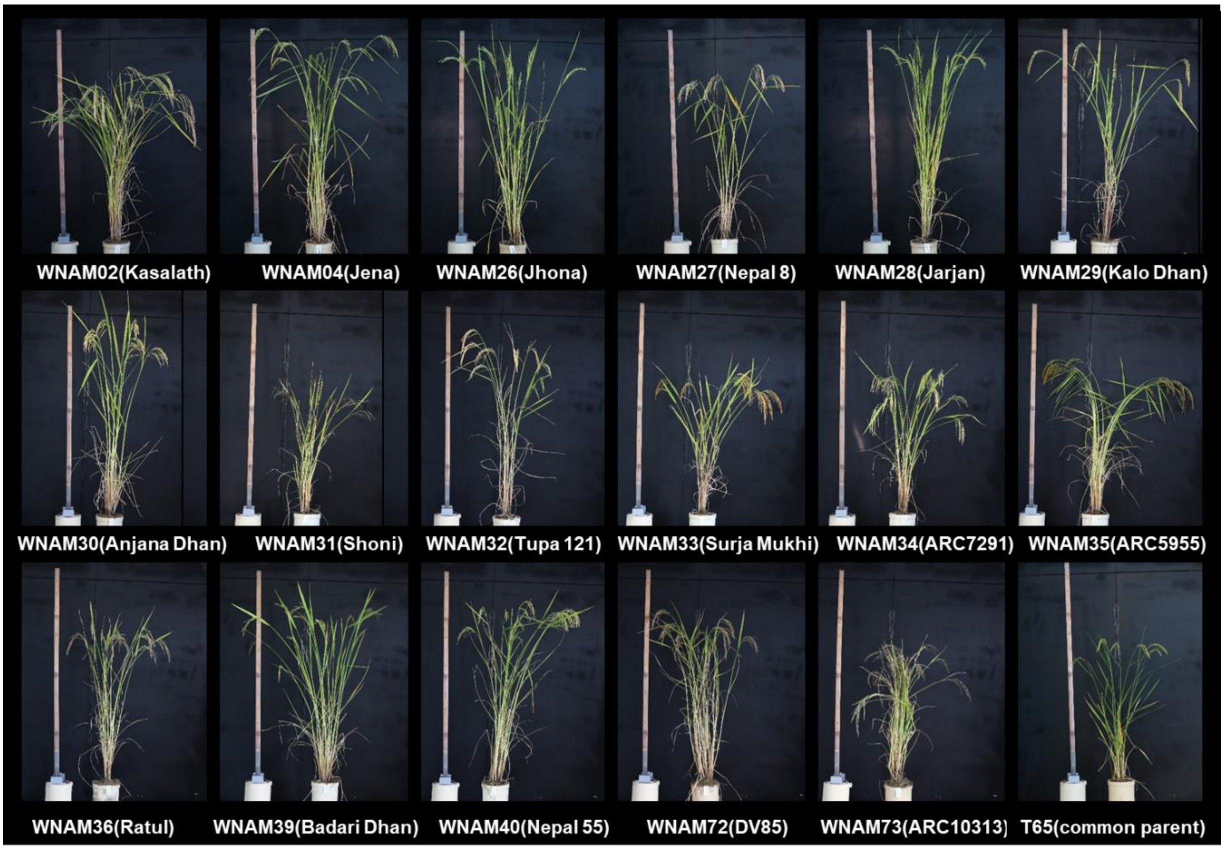


Figure 2-4. Phenotype diversity in an *aus*-NAM population.

The phenotype diversity was demonstrated by their plant height. WNAM02(Kasalath), WNAM04(Jena), WNAM26(Jhona), WNAM27(Nepal 8), WNAM28(Jarjan), WNAM29(Kalo Dhan), WNAM30(Anjana Dhan), WNAM31(Shoni), WNAM32(Tupa 121), WNAM33(Surja Mukhi), WNAM34(ARC7291), WNAM35(ARC5955), WNAM36(Ratul), WNAM39(Badari Dhan), WNAM40(Nepal 55), WNAM72(DV85) and WNAM73(ARC10313)

Table 2-1. List of parents and RILs in *aus*-NAM-I population.
[1]World Rice Core Collection (WRC)(Tanaka *et al.*, 2020)

Family name	Founder's name	WRC No.[1]	F_2	F_5	Residual rate
WNAM02	Kasalath	WRC02	233	109	46.78%
WNAM29	Kalo Dhan	WRC29	219	163	74.43%
WNAM31	Shoni	WRC31	174	121	69.54%
WNAM35	ARC5955	WRC35	229	137	59.83%
WNAM39	Badari Dhan	WRC39	213	126	59.15%
WNAM72	DV85	-	-	107	-
WNAM73	ARC10313	-	-	132	-

Table 2-2. Taxa and SNPs in *aus*-NAM-I.

Family	No. of lines genotyped	No. of lines retained after SNP discovery (TASSEL 5)	No. of raw SNPs	No. of retained SNPs after parental filter	No. of retained SNPs after site filter (min allele freq 0.02)
WNAM02	110	108	19434	4399	4285
WNAM29	164	164	13708	3428	3336
WNAM31	122	120	12756	2980	2852
WNAM35	138	129	13756	3444	3346
WNAM39	132	128	11033	3408	3325
WNAM72	123	106	13522	3616	3510
WNAM73	136	132	10075	2985	2868
aus NAM-I	925	887			

Chapter 3 QTL Mapping using *aus*-NAM-I Population

Introduction

The majority of phenotypic variation of agricultural traits is determined by many loci with small effects. Detection of the genomic regions (QTL) that control these traits is referred to as QTL mapping. In the recent past, the advent of DNA marker technologies has enabled the application of DNA markers in QTL mapping and selection (marker-assisted selection) (Reyes *et al.*, 2021). The rich allele diversity present in multi-parental mapping populations like NAM has an advantage of high mapping power and resolutions compared to the classical bi-parental populations when conducting QTL analysis (Stadlmeier *et al.*, 2018).

Since rice domestication in southeast Asia (Doebley *et al.*, 2006), rice cultivation has expanded to wide geographical regions. The key factor that enabled rice adaptation is selection for photoperiod insensitivity both naturally and/or artificially (Izawa, 2007). Genetic studies of days to heading (DTH) also known as the flowering time in plants have revealed complex gene networks (**Figure 3-1**). In addition to roles in plants adaptations, DTH also influences crop yields, reproductive isolation, growth, and development. DTH is controlled by internal and external signals such as levels of phytohormones, temperature, and photoperiod (Zhu *et al.*, 2017).

Flowering time genes control the transition from vegetative meristem to floral meristem (Gaudinier and Blackman, 2020), The regulatory networks of flowering time genes are well characterized in rice (**Figure 3-1**) (Hayama and Coupland 2004; Tsuji et al. 2008). The first flowering time gene to be cloned in rice was *Hd1* (Yano *et al.*, 2000). Since then, over 12 flowering genes have been isolated and mapped to specific pathways. Analysis of rice flowering genes revealed two main flowering pathways, *Hd1–Hd3a* and *Ghd7–Ehd1–Hd3a/RFT1*(Matsubara and Yano, 2018). Detecting novel genes involved in flowering time is paramount for crop improvements, particularly in temperate environments where seasonal changes will require flowering time plasticity (Gaudinier and Blackman, 2020).

The main objective of this study was to identify quantitative trait loci (QTL) controlling DTH and demonstrate the utility of *aus*-NAM-I for QTL mapping. Single-family QTL linkage mapping, joint-family QTL linkage mapping, and the methods based on genome-wide association study (GWAS) were deployed in the analyses.

Materials and Methods

Plant Materials

Rice nested association mapping population (*aus*-NAM-I) containing 7 *aus* varieties as diversity donors and T65 as the common parent was utilized in this study. Details of the plant materials are presented in **Chapter 2**.

Trait Evaluation and Statistical Analysis

Field trials were conducted in the year 2015 at Togo Field, Nagoya University, Aichi, Japan (35°06'36.5"N, 137°05'06.3"E). Four seedlings per line per row were transplanted with a spacing of 20 cm between the hills and 30 cm between rows. Standard agronomic management was followed during the experiment, except no fertilizer was applied.

Days to heading (DTH) was calculated as the difference between the date of emergence of inflorescence and sowing. Phenotype values distributions across subpopulations were examined. To find trait means that were significantly different among *aus*-NAM families, a one-way analysis of variance (ANOVA) followed by Tukey HSD with a 95% confidence level was performed. All statistical analysis and visualization were performed using R version 4.0.3 (R Core Team, 2020).

Genotyping of aus-NAM-I RILs

GBS libraries were prepared using reported protocols (Poland *et al.*, 2012; Furuta *et al.*, 2017) and sequenced using Illumina MiSeq (Illumina, CA, USA). Raw reads were processed using TASSEL-GBS pipeline (Glaubitz *et al.*, 2014) Genotyping details are presented in **Chapter 2.**

Whole-Genome Resequencing of aus-NAM-I Founders

DNA of the founders was extracted using the cetyltrimethylammonium bromide (CTAB) method then fragmented using Covaris Model S2 (Covaris, MA, USA), and used to construct sequencing library by TruSeq DNA LT kit (Illumina, CA, USA). Sequencing was conducted by using Illumina Miseq with Miseq Reagent Kit v3 (600 cycles). Variants calling was conducted following the standard protocol of Genome Annotation ToolKit (GATK)(DePristo *et al.*, 2011) using Os-Nipponbare-IRGSP-1.0 (Kawahara *et al.*, 2013) as the reference.

Projection of Parental Variants

SNPs from parental read sequences were projected onto each RIL. Projections were performed in two steps (1) employ GBS markers as skeletons (2) check adjacent skeleton markers, if homozygous and have the common allele type as one of the parents; project the parental SNPs onto the intervals, otherwise set the intervals as missing.

Single-family QTL Mapping

Genotype files in HapMap format were converted to ABH parent-based format, where A represented T65, B represented *aus* genotype while heterozygotes and missing were represented by H and "-" respectively. Kosambi mapping function in the R/qtl package (Arends *et al.*, 2010) was used to obtain genetic distances in (cM). QTL mapping was performed based on interval mapping using the 'hk' method implemented in R/qtl. The additive effects of a marker were calculated as '((average of *aus*) – (average of T65)) / 2'. Positive and negative additive effects values implied that *aus* and T65 allele increased trait values respectively. The logarithm of odds (LOD) value of 3 was fixed as the significance threshold for QTL although LOD of 3.04 was obtained as the empirical threshold (type I error of 0.05) based on 1000 permutation tests (Churchill and Doerge, 1994).

Joint-family QTL Mapping

For joint-family QTL linkage mapping, genotype information in a common genetic map and DTH were subjected to Joint Inclusive Composite Interval Mapping (JICIM) algorithm (Li *et al.*, 2011). Missing phenotypes were replaced by the mean of the trait, 1 cM step was selected and QTL significance threshold was obtained from 1000 permutation tests with a type I error of 0.05. Genotype file was converted into a numeral format where T65 genotype was represented by 0, *aus* genotype represented by 2, while heterozygous genotypes and missing genotypes were represented by 1 and -1 respectively. Positive and negative additive effects values mean that *aus* and T65 allele increased the trait values respectively.

Genome-Wide Association Analysis

TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) software (Bradbury *et al.*, 2007) was used for QTL association mapping using two GWAS-based methods, General Linear Model (GLM) and Mixed Linear Model(MLM). Principal components (Q) were used to account for population structure and genomic kinship (K) was used to account for hidden relatedness in MLM (Q+K). In addition, pedigree information (family) was given to TASSEL software as covariates in MLM (Q+K). The significance threshold was determined using Bonferroni with the equation: $P \leq 1/N$ (α =0.05) where N is the number of markers (Haynes, 2013).

Results

Phenotypic Characteristics

The mean values of DTH in each family in *aus*-NAM-I varied from 88 days to 105 days with potential transgressive segregation in some families observed i.e. WNAM31 RILs DTH average was extreme compared to both parents **Figure 3-2**. Analysis of variance showed a statistically significant difference among the RILs families with an F value of 73.52 and a P-value < 2 x 10-16.

Linkage Map and Projection of Parental Variants

For joint-family QTL mapping, common 1,786 non-redundant SNP markers, sufficiently covering all the 12 rice chromosomes were discerned and used for QTL analysis. The average distance between markers ranged from 0.41 cM to 0.86 cM (**Table 3-1**). For projections, parental variants obtained from whole-genome resequencing (4,643,123 SNPs) were firstly thinned then projected onto each of the individual family skeleton linkage maps. A total of 41,561 SNPs were obtained and thereafter utilized in GWAS.

QTL Detected by Single Family Analysis, Joint Family Analysis, and GWAS-based Methods

Single-family QTL analysis detected a total of 14 significant additive QTL on chromosomes: 5, 6, 7, and 10 (**Figure 3-3** and **Table 3-2**). The QTL individually explained 12% to 36% of the trait variances. Among the QTL detected, 8 and 6 contained alleles from T65 and *aus* loci increasing days to heading respectively. QTL on chromosome 10 was commonly detected across the seven families, with WNAM73 possessing the highest LOD score (11.81) (**Table 3-2**).

Joint-family QTL linkage mapping identified a total of 19 QTL (**Figure 3-4** and **Table 3-3**). Some of the joint-family QTL overlapped with QTL in single-family, such as QTL on chromosomes 6, 7, and 10. The peak on chromosome 10 was detected as a single peak. The peaks on chromosomes 6 and 7 looked like a combination of all the 7 populations, resulting in 2 separate peaks on each chromosome. In addition, 14 putative QTL spanning relatively wide regions were detected on chromosomes 1, 2, and 3 (**Figure 3-4**). On the other hand, a significant QTL on chromosome 5 in WNAM72 (**Figure 3-3** F) was not detected in joint-family QTL analysis.

GWAS by naive GLM revealed significant QTL signals in all chromosomes (**Figure 3-5**) while MLM(Q+K) identified significant SNPs on chromosomes: 6, 7, and 10 (**Figure 3-6**). The total number of SNPs in MLM(Q+K) that met the negative logarithm P-value (-LogP) of 5.9 (Bonferroni threshold at alpha 0.05) was 188 SNPs.

Evaluation of Mapping Accuracy

The QTL commonly detected by three mapping methods on chromosomes 6, 7, and 10 included the region of *RFT1* (Izawa *et al.*, 2002), *Hd3a* (Kojima *et al.*, 2002), *Hd1 (Yano et al., 2000)*, *Ghd7* (Xue *et al.*, 2008), and *Ehd1* (Doi *et al.*, 2004). Assuming that *Hd1*, *Ghd7,* and *Ehd1* were the genes underlying the detected QTL, these loci were used to evaluate the accuracy of gene mapping in *aus*-NAM.

A major QTL on chromosome 10 was identified in every individual family (**Figure 3-3** and **Table 3-2**). This QTL corresponded to *Ehd1* (Os10g0463400) (Doi *et al.*, 2004) which is located between 17076098 bp to 17081344 bp on chromosome 10. The peak was detected at the marker position from 16626134bp to 17367103bp in single-family QTL mapping (**Figure 3-3**), while joint-family QTL analysis identified marker position 16772764bp as the peak (**Figure 3-4**). MLM (Q+K) detected a peak spanning from 17095439 to 17164368 bp (**Figure 3-7**C). All of the statistical methods successfully detected *Ehd1*.

Closely to *Hd1* on chromosome 6 (9,335,377 bp to 9,337,570 bp), a significant QTL was detected in WNAM39 (Badari Dhan) (**Figure 3-3**E). This peak was flanked by the markers S06_8837126 and S06_9318022 in joint-family QTL linkage analysis with a LOD score of 7.84 (**Figure 3-4** chr.6). In MLM (Q+K), the marker S06_9338330 was contained in association mapping with a -LogP value of 11.6 (**Figure 3-7**A). To better understand the individual family contribution to the joint peak on chromosome 6 (**Figure 3-3**), additive effects at this locus were further analyzed. Multiple alignments of the amino acid sequences deduced using the genomic sequences confirmed loss-of-function in *aus* varieties except for Badari Dhan (**Figure 3-8**). Additionally, Badari Dhan allele had the highest additive effect values (6.72 days) compared to the rest, **Table 3.4.**

Another signal in the vicinity of *RFT1* and/or *Hd3a* on chromosome 6 was also detected by GWAS (**Figure 3-6** and **Figure 3-7**), although the signal did not reach the significance threshold.

Ghd7 is located between 9,184,534 bp to 9,187,187 bp on chromosome 7 (Xue *et al.*, 2008; Yamamoto *et al.*, 2012). Single-family QTL analysis showed that WNAM02 (T65 x Kasalath) and WNAM35 (T65 x ARC5955) had a significant QTL at the vicinity of *Ghd7* (**Figure 3-3**A and **Figure 3-3**D). Therefore, it was hypothesized that Kasalath (WNAM02) and ARC5955 (WNAM35) possess functional (late) alleles for *Ghd7*. The peak corresponding to *Ghd7* in joint-family QTL was from 8.93Mbp to 9.35Mbp on chromosome 7, which contained the *Ghd7* locus (**Figure 3-4**). MLM (Q+K) detected a cluster of markers around *Ghd7* locus with -LogP values greater than 5, surrounded by the markers that showed higher

-LogP values than the significance threshold, and thus the position of *Ghd7* was not clear. In WNAM39, a QTL near *Ghd7* was detected where *aus* allele had a negative additive effect that was opposite from other families, this peak was also detected in joint-family QTL mapping.

Discussion

Accuracy of QTL Mapping using aus-NAM-I Population

To date, over 40 flowering QTL have been identified in rice (Yamamoto et al. 2012). In this study, *Ehd1* (Doi et al. 2004) was detected as the most common major QTL. The mapped positions of the QTL on chromosome 10 corresponded to the actual position of *Ehd1* (**Figure 3-3**, **Figure 3-4**, **Figure 3-5**, and **Figure 3-7**C). Another heading time locus, *Hd1* (Yano et al. 2000) was detected only in WNAM39, and analysis of deduced amino acid sequence confirmed that the founder of WNAM39 (Badari Dhan) was the only variety possessing functional allele of *Hd1* (**Figure 3-8**). The effect of a functional allele of *Hd1* in the environment of this study (long-day) was to delay heading, and it matched the observed result. However, the *Hd1* peak in GWAS (**Figure 3-7**A) was not surrounded by markers with smaller –Log10(P) values like *Ehd1*. This was probably because of the sequence difference between Badari Dhan family and others, signals of linked markers were diluted by other families. Without prior information, *Hd1* would not be mapped to a precise position using MLM (Q+K).

Unlike *Ehd1*, it was not possible to discriminate alleles at *Hd3a*, *RFT1*, and *Ghd7* despite the previous reports (Kojima et al. 2002; Xue et al. 2008; Izawa et al. 2002). However, it should be noted that joint-family QTL mapping precisely mapped the peaks of *Ehd1, Hd1, Ghd7,* and a combined peak of *RFT1/Hd3a* (**Figure 3-4**). The joint-family QTL mapping approach has been reported to amplify small effects signals found on individual family RILs (Fragoso et al. 2017). The results in the present study indicated that joint-family QTL mapping is advantageous in the precision of QTL positioning compared with MLM (Q+K).

A QTL tightly linked to *Ghd7* was detected in WNAM02 and WNAM35. Another QTL near but not tightly linked to *Ghd7* (11.0Mb on chromosome 7, 2.08Mb apart from *Ghd7* (8.93Mb)) was detected in WNAM39, where the functional allele of *Hd1* segregated. This QTL showed an opposite additive effect as that expressed in WNAM02 and WNAM35. *Ghd7* was reported to have the ability to switch its additive effects by epistasis with *Hd1* (Fujino et al. 2019). The underlying gene in the QTL detected in WNAM39 remains unclear. A well-saturated linkage map will facilitate the characterization of genes underpinning this QTL.

Figures and Tables

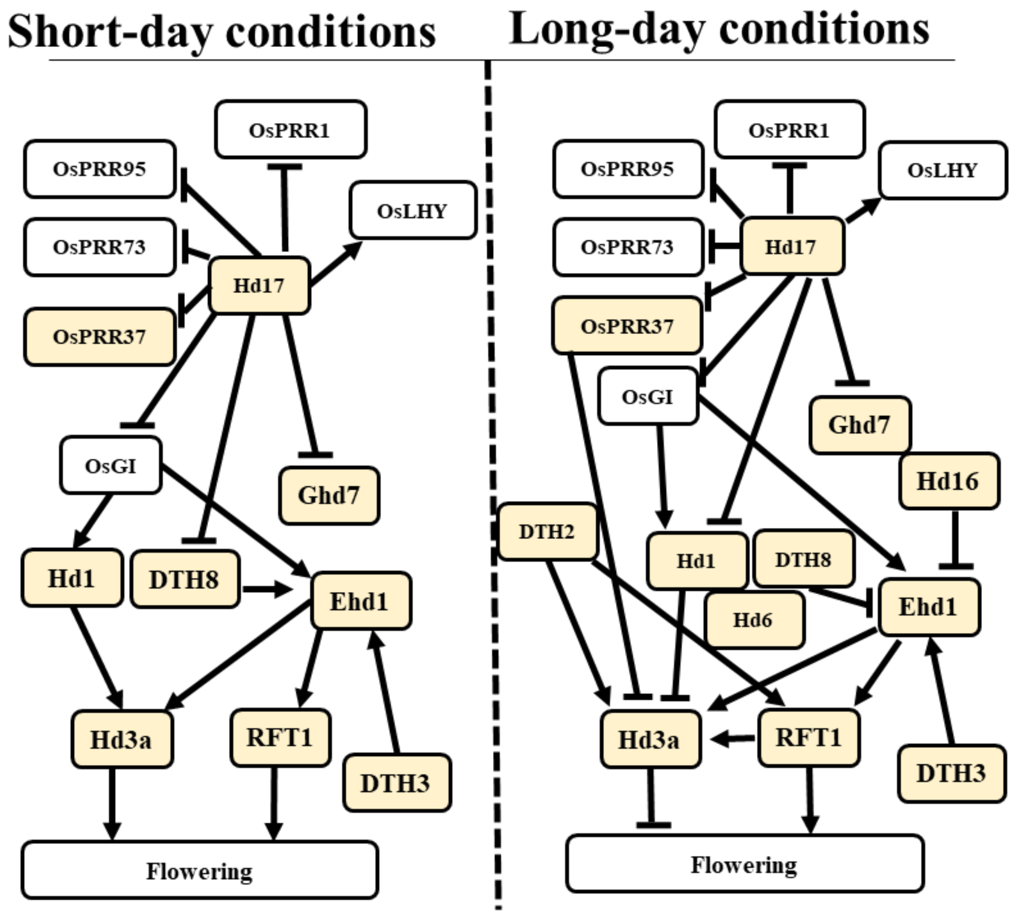


Figure 3-1. Genes in rice photoperiodic flowering pathways.
Arrowheads denote up-regulation; Bars denote down-regulation. Genes that were cloned by QTL analysis of natural variation are shown in the highlighted boxes (Matsubara *et al.*, 2014).

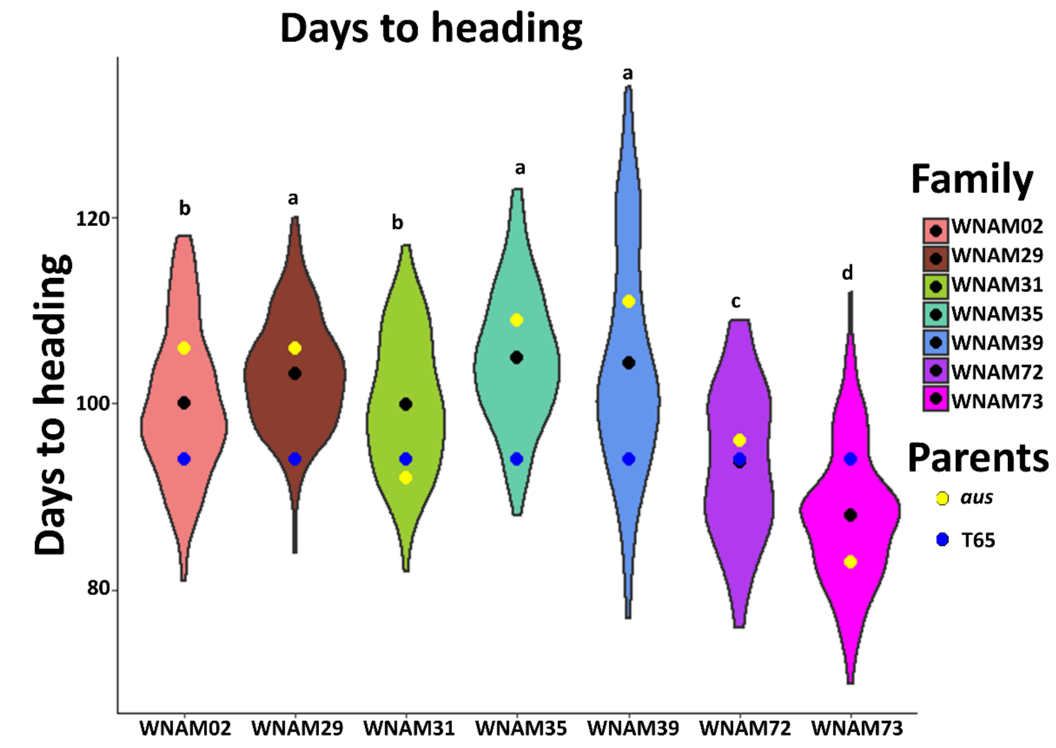


Figure 3-2. Days to heading frequency distributions.

Violin plots showing frequency distributions of DTH. Yellow and blue dots represent *aus* and T65 founders respectively. Black dots show the RILs average in each family. Groups with no significant difference by Tukey HSD with a 95% confidence level are represented by the same letters above the plots.

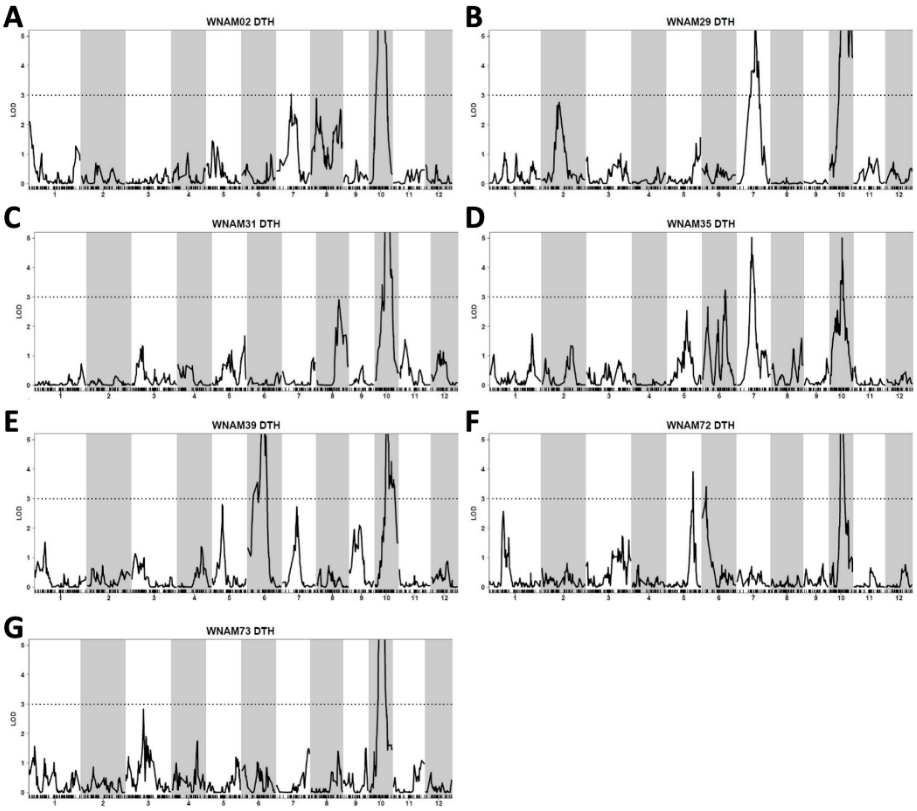


Figure 3-3. Interval mapping for days to heading in *aus*-NAM-I.

Interval mapping using R/QTL scanone function (A) WNAM02, (B) WNAM29, (C) WNAM31, (D) WNAM35, (E) WNAM39, (F) WNAM72, and (G) WNAM73. The black solid lines correspond to the LOD score profile (y-axis) as a function of distance in cM across each chromosome (x-axis). Horizontal dotted lines in all panels indicate the LOD significance threshold value of 3.

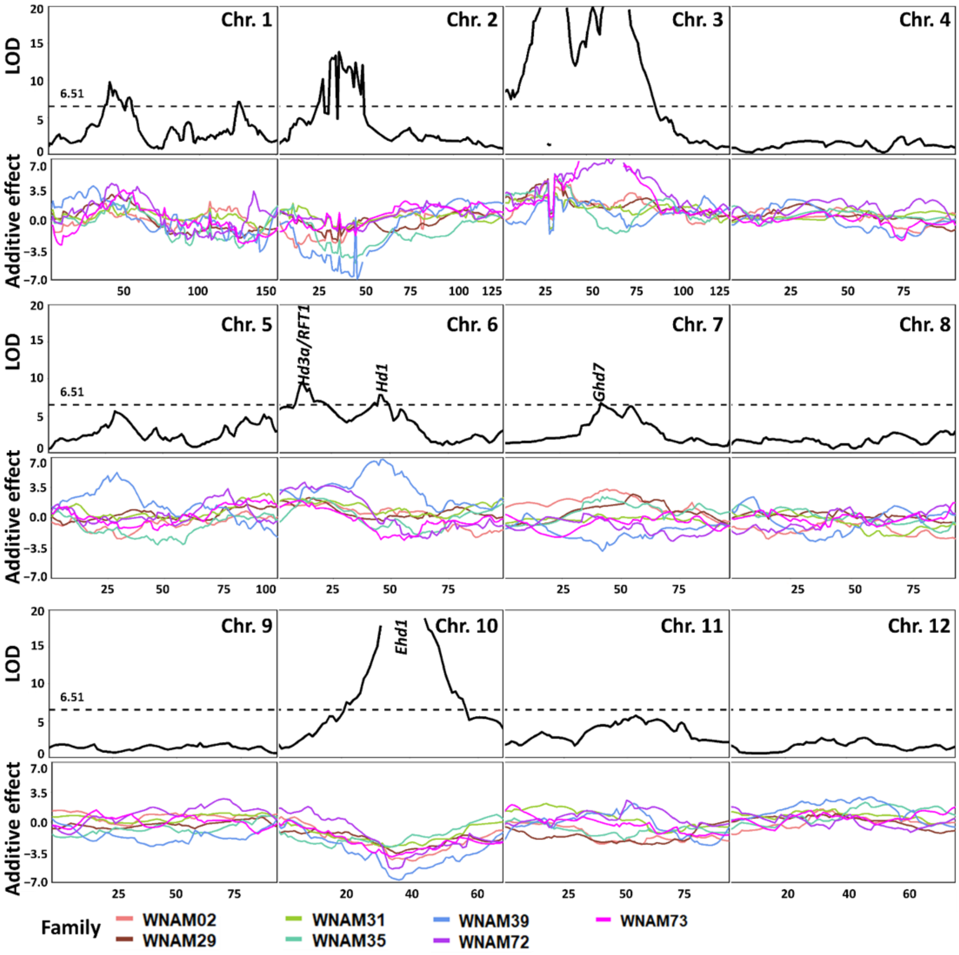


Figure 3-4. Joint-family QTL and additive effects in *aus*-NAM-I.

LOD profiles detected by the JICIM algorithm were plotted on top panels while additive effects are shown on bottom panels. The scanning step was 1 cM and the dotted horizontal line (6.51) represents the significance threshold obtained from 1000 permutation tests with a type I error of 0.05. Different line colors in the additive effects panel represent different families. The positions of known loci (RFT1, *Hd3a*, *Hd1*, *Ehd1*) are shown in the panels of chromosomes 6, 7, and 10.

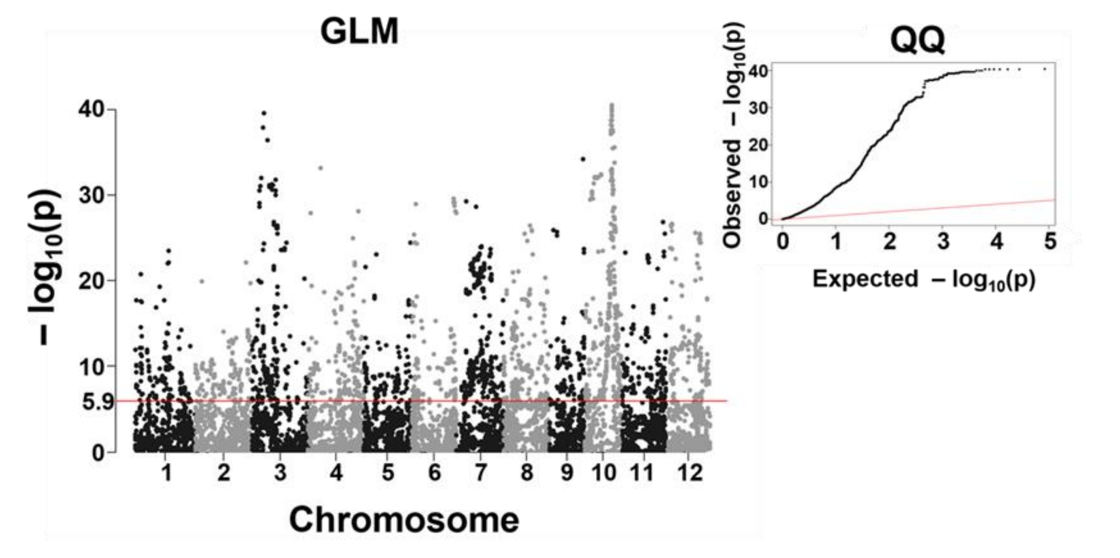


Figure 3-5. GWAS for DTH using GLM in *aus*-NAM-I.
Manhattan plot and quantile-quantile (QQ) plot for days to heading. The red horizontal line marks the threshold (5.9). A QQ plot is shown in the right panel, where the expected P-values vs. the observed P-values are plotted on a -log10 scale.

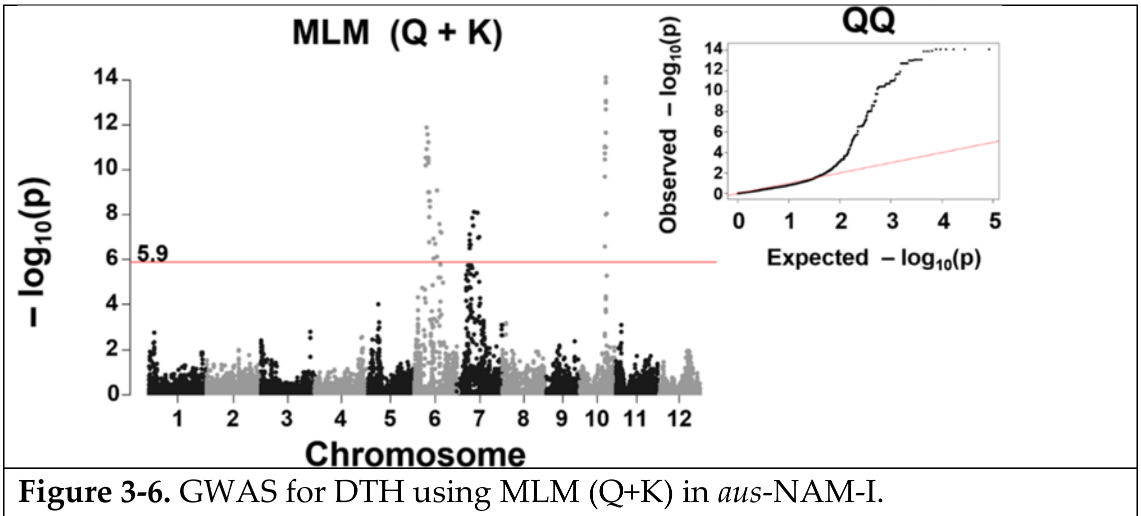


Figure 3-6. GWAS for DTH using MLM (Q+K) in *aus*-NAM-I.
Manhattan plot and quantile-quantile (QQ) plot for days to heading. The red horizontal line marks the threshold (5.9). A QQ plot is shown in the right panel, where the expected P-values vs. the observed P-values are plotted on a -log10 scale.

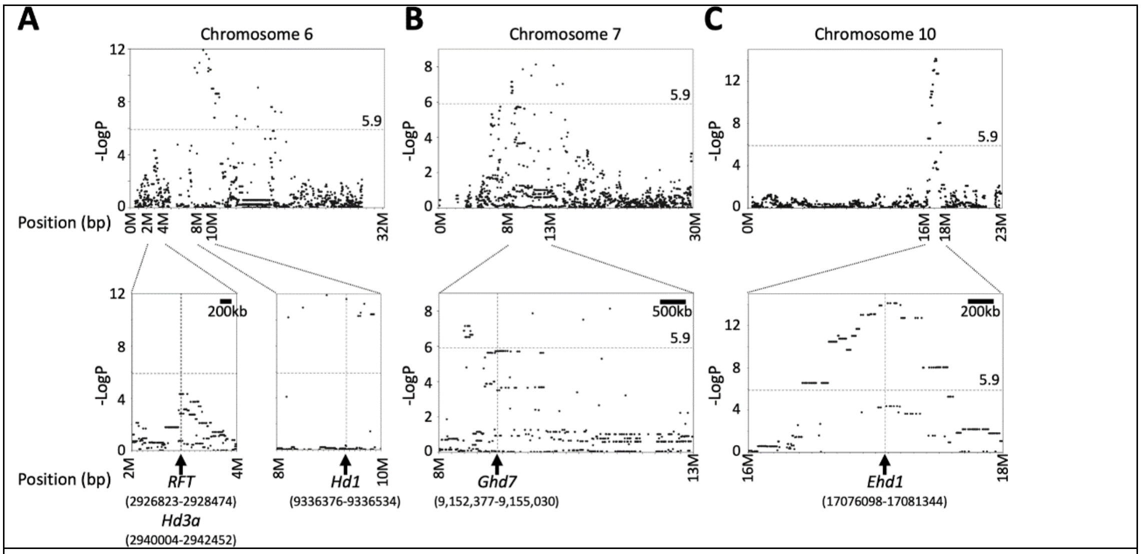
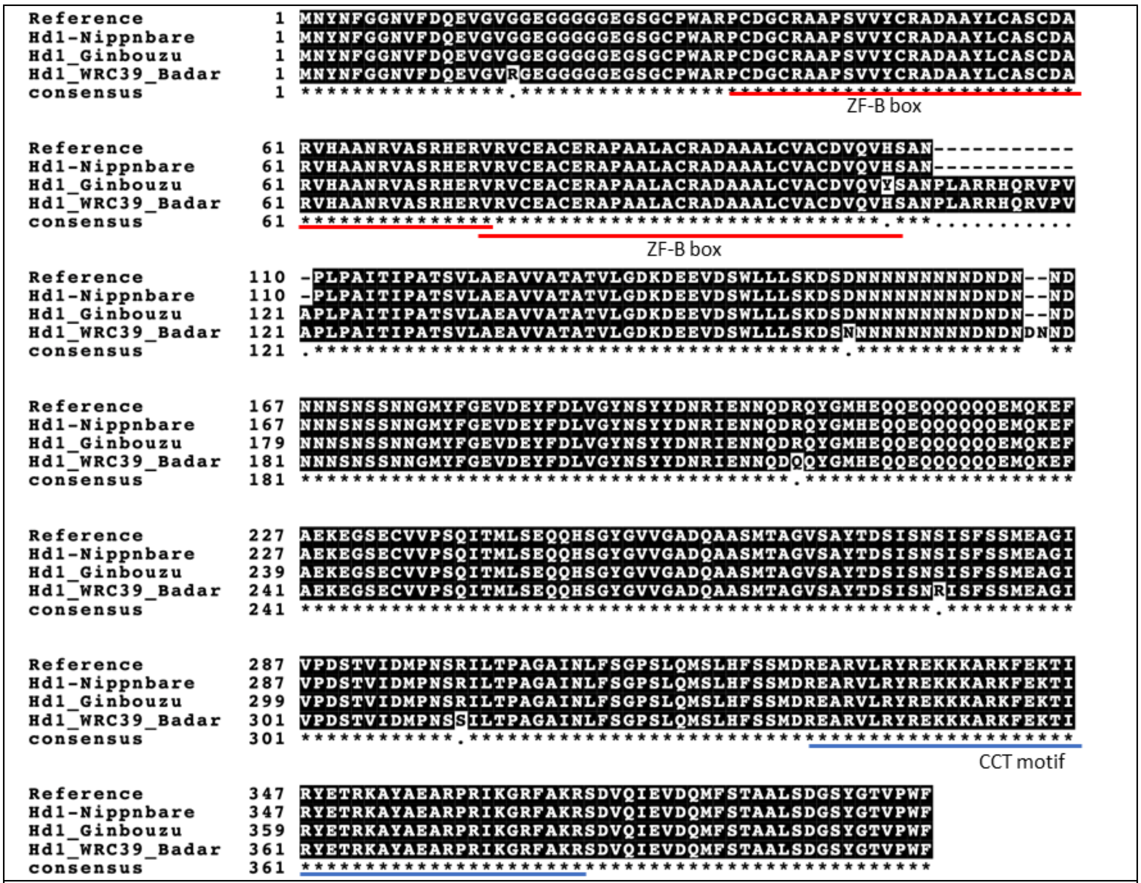


Figure 3-7. Local GWAS for days to heading in *aus*-NAM-I.

The Manhattan scatter plots using MLM (Q+K) show a local association of days to heading (A) chromosome 6, (B) chromosome 7, and (C) chromosome 10. The panels at the bottom are magnification around RFT/*Hd3a*, *Hd1*, *Ghd7*, and *Ehd1* loci.



```
Reference          1 MNYNFGGNVFDQEVGVGGEGGGGGEGSGCPWARPCDGCRAAPSVVYCRADAAYLCASCDA
Hd1-Nippnbare      1 MNYNFGGNVFDQEVGVGGEGGGGGEGSGCPWARPCDGCRAAPSVVYCRADAAYLCASCDA
Hd1_Ginbouzu       1 MNYNFGGNVFDQEVGVGGEGGGGGEGSGCPWARPCDGCRAAPSVVYCRADAAYLCASCDA
Hd1_WRC39_Badar    1 MNYNFGGNVFDQEVGVRGEGGGGGEGSGCPWARPCDGCRAAPSVVYCRADAAYLCASCDA
consensus          1 ****************.*******************************************

Reference         61 RVHAANRVASRHERVRVCEACERAPAALACRADAAALCVACDVQVHSAN-----------
Hd1-Nippnbare     61 RVHAANRVASRHERVRVCEACERAPAALACRADAAALCVACDVQVHSAN-----------
Hd1_Ginbouzu      61 RVHAANRVASRHERVRVCEACERAPAALACRADAAALCVACDVQVYSANPLARRHQRVPV
Hd1_WRC39_Badar   61 RVHAANRVASRHERVRVCEACERAPAALACRADAAALCVACDVQVHSANPLARRHQRVPV
consensus         61 *********************************************.***...........

Reference        110 -PLPAITIPATSVLAEAVVATATVLGDKDEEVDSWLLLSKDSDNNNNNNNNNDNDN--ND
Hd1-Nippnbare    110 -PLPAITIPATSVLAEAVVATATVLGDKDEEVDSWLLLSKDSDNNNNNNNNNDNDN--ND
Hd1_Ginbouzu     121 APLPAITIPATSVLAEAVVATATVLGDKDEEVDSWLLLSKDSDNNNNNNNNNDNDN--ND
Hd1_WRC39_Badar  121 APLPAITIPATSVLAEAVVATATVLGDKDEEVDSWLLLSKDSNNNNNNNNNNDNDNDNND
consensus        121 .*****************************************.*************  **

Reference        167 NNNSNSSNNGMYFGEVDEYFDLVGYNSYYDNRIENNQDRQYGMHEQQEQQQQQQEMQKEF
Hd1-Nippnbare    167 NNNSNSSNNGMYFGEVDEYFDLVGYNSYYDNRIENNQDRQYGMHEQQEQQQQQQEMQKEF
Hd1_Ginbouzu     179 NNNSNSSNNGMYFGEVDEYFDLVGYNSYYDNRIENNQDRQYGMHEQQEQQQQQQEMQKEF
Hd1_WRC39_Badar  181 NNNSNSSNNGMYFGEVDEYFDLVGYNSYYDNRIENNQDQQYGMHEQQEQQQQQQEMQKEF
consensus        181 **************************************.*********************

Reference        227 AEKEGSECVVPSQITMLSEQQHSGYGVVGADQAASMTAGVSAYTDSISNSISFSSMEAGI
Hd1-Nippnbare    227 AEKEGSECVVPSQITMLSEQQHSGYGVVGADQAASMTAGVSAYTDSISNSISFSSMEAGI
Hd1_Ginbouzu     239 AEKEGSECVVPSQITMLSEQQHSGYGVVGADQAASMTAGVSAYTDSISNSISFSSMEAGI
Hd1_WRC39_Badar  241 AEKEGSECVVPSQITMLSEQQHSGYGVVGADQAASMTAGVSAYTDSISNRISFSSMEAGI
consensus        241 *************************************************.**********

Reference        287 VPDSTVIDMPNSRILTPAGAINLFSGPSLQMSLHFSSMDREARVLRYREKKKARKFEKTI
Hd1-Nippnbare    287 VPDSTVIDMPNSRILTPAGAINLFSGPSLQMSLHFSSMDREARVLRYREKKKARKFEKTI
Hd1_Ginbouzu     299 VPDSTVIDMPNSRILTPAGAINLFSGPSLQMSLHFSSMDREARVLRYREKKKARKFEKTI
Hd1_WRC39_Badar  301 VPDSTVIDMPNSSILTPAGAINLFSGPSLQMSLHFSSMDREARVLRYREKKKARKFEKTI
consensus        301 ************.***********************************************

Reference        347 RYETRKAYAEARPRIKGRFAKRSDVQIEVDQMFSTAALSDGSYGTVPWF
Hd1-Nippnbare    347 RYETRKAYAEARPRIKGRFAKRSDVQIEVDQMFSTAALSDGSYGTVPWF
Hd1_Ginbouzu     359 RYETRKAYAEARPRIKGRFAKRSDVQIEVDQMFSTAALSDGSYGTVPWF
Hd1_WRC39_Badar  361 RYETRKAYAEARPRIKGRFAKRSDVQIEVDQMFSTAALSDGSYGTVPWF
consensus        361 *************************************************
```

Figure 3-8. Accuracy for *Hd1* mapping.

The figure shows amino-acid sequences alignment of *Hd1* in functional alleles of Nipponbare and Ginbouzu and Badari Dhan. Regions of the 2 ZF-B box and CCT motif are indicated. The sequence of Badari Dhan contained 6 non-synonymous mutations, but it was considered that the allele retains function.

Table 3-1. Joint-family linkage map statistics in *aus*-NAM-I.

Chromosome	No. of markers	Genetic length(cM)	Average spacing (cM)	Maximum spacing (cM)
1	230	150.06	0.66	11.41
2	157	127.16	0.82	13.36
3	219	132.04	0.61	11.22
4	136	97.62	0.72	16.1
5	179	100.04	0.56	8.46
6	136	98.67	0.73	12.34
7	113	96.87	0.86	14.37
8	113	92.14	0.82	9.66
9	112	71.26	0.64	7.76
10	166	67.28	0.41	8.08
11	115	90.82	0.8	9.98
12	110	75.98	0.7	11.04
Overall	1786	1199.94	0.68	16.1

Table 3-2. QTL in *aus*-NAM-I using single-family QTL analysis.

Family[1]	chr[2]	position[3]	LOD[4]	a[5]	PVE[6]	SNP marker
WNAM02	10	34.784954	9.54	-9.45	0.35	S10_16626134
	7	42.869937	3.04	5.86	0.13	S07_13768989
WNAM29	10	36.342898	9.57	-6.03	0.25	S10_17367103
	7	54.815808	5.87	4.72	0.16	S07_19401587
WNAM31	10	35.788232	9.38	-7.98	0.31	S10_16808215
WNAM35	10	36.179895	5.00	-6.22	0.18	S10_17171636
	6	67.389475	3.25	-5.06	0.12	S06_23867930
	7	42.276444	5.02	6.42	0.18	S07_9436160
WNAM39	10	36.179895	7.27	-12.10	0.23	S10_17200086
	6	45.828088	7.81	12.52	0.25	S06_9324594
WNAM72	10	34.784954	10.20	-9.55	0.36	S10_16626134
	5	77.858431	3.92	6.34	0.16	S05_24089993
	6	12.908983	3.41	6.15	0.14	S06_3043159
WNAM73	10	35.788232	11.81	-8.77	0.34	S10_16808215

[1] *aus*-NAM-I family.

[2] Chromosome number.

[3] Position along the chromosome in cM.

[4] QTL logarithm of odds

[5] Additive effects of the marker calculated as '((average of *aus*) – (average of T65)) / 2'. Positive values indicate that *aus* parent alleles increased the trait value and vice versa.

[6] Percentage of phenotypic variance explained by QTL.

Table 3-3. QTL in *aus*-NAM-I using joint-family QTL analysis.

						LOD for WNAM:						
Chr	Pos	Left marker	Right marker	LOD[1]	PVE[2]	2	29	31	35	39	72	73
1	41.144	S01_8932051	S01_9243217	9.76	1.21	0.14	3.80	1.15	0.71	0.30	2.58	1.08
1	54.144	S01_12059573	S01_12574351	7.72	1.05	0.07	1.60	0.55	1.68	0.00	1.45	2.37
1	125.14	S01_38216481	S01_38471034	7.10	1.13	0.15	1.99	0.01	2.64	0.23	0.40	1.68
2	27.56	S02_5835680	S02_6788077	10.19	1.83	1.50	2.03	0.18	3.15	3.22	0.05	0.07
2	34.56	S02_8446962	S02_8635149	13.34	2.38	1.63	2.54	0.24	3.68	4.70	0.30	0.25
2	36.56	S02_8785972	S02_9401106	13.81	2.58	1.81	1.52	0.42	4.46	5.17	0.22	0.20
2	45.56	S02_11641918	S02_12015489	12.40	2.98	0.09	0.08	0.03	4.77	7.03	0.17	0.23
2	49.56	S02_16916875	S02_17920669	12.06	4.18	0.07	0.02	0.10	3.18	8.40	0.24	0.04
3	4.6551	S03_1504686	S03_1680047	8.74	1.09	0.01	2.76	2.08	2.12	0.03	1.52	0.23
3	26.655	S03_6481871	S03_6797449	29.86	2.89	5.71	11.44	2.44	4.83	0.57	3.49	1.38
3	30.655	S03_7615673	S03_7920537	31.89	3.55	3.60	11.20	2.87	5.83	0.87	4.87	2.65
3	36.655	S03_9368445	S03_9433259	21.59	2.64	3.48	5.10	1.76	4.41	0.19	4.64	2.01
3	53.655	S03_14087144	S03_14687036	19.86	5.00	0.67	1.07	0.78	0.01	0.56	7.77	9.00
3	66.655	S03_16837159	S03_17302066	27.70	4.74	0.87	0.94	0.70	0.65	0.02	10.42	14.10
3	68.655	S03_20983648	S03_21066372	28.73	4.70	1.15	0.90	1.10	0.39	0.06	10.03	15.11
6	12.603	S06_2279056	S06_2984376	9.54	1.39	0.50	1.65	1.57	0.86	1.57	2.73	0.66
6	45.603	S06_8837126	S06_9318022	7.84	2.95	0.21	0.03	0.12	0.22	5.61	0.08	1.57
7	42.143	S07_8936752	S07_9350535	6.74	1.67	2.45	0.82	0.04	1.52	1.90	0.00	0.00
10	35.383	S10_16626134	S10_16772764	30.69	3.83	4.41	6.35	2.99	1.93	6.11	5.21	3.69

[1] Joint-family QTL logarithm of odds

[2] Percentage of phenotypic variance explained by QTL.

Table 3-4. Additive effects in *aus*-NAM-I using joint-family QTL analysis.

						Additive effects[3] for WNAM:						
Chr	Pos	Left marker	Right marker	LOD[1]	PVE[2]	2	29	31	35	39	72	73
1	41.144	S01_8932051	S01_9243217	9.76	1.21	0.78	2.89	1.71	1.72	1.57	4.13	2.12
1	54.144	S01_12059573	S01_12574351	7.72	1.05	0.53	1.90	1.24	2.60	-0.01	2.99	3.05
1	125.14	S01_38216481	S01_38471034	7.10	1.13	0.76	-2.11	0.19	-3.26	-1.70	-1.98	-2.61
2	27.56	S02_5835680	S02_6788077	10.19	1.83	-2.42	-2.16	-0.71	-3.63	-4.98	-0.55	-0.54
2	34.56	S02_8446962	S02_8635149	13.34	2.38	-2.52	-2.46	-0.80	-3.91	-5.96	-1.36	-1.04
2	36.56	S02_8785972	S02_9401106	13.81	2.58	-2.76	-1.89	-1.05	-4.27	-6.21	-1.16	-0.91
2	45.56	S02_11641918	S02_12015489	12.40	2.98	-0.59	-0.43	-0.27	-4.30	-6.96	-1.01	-0.96
2	49.56	S02_16916875	S02_17920669	12.06	4.18	-0.53	0.19	-0.50	-3.57	-8.72	-1.21	-0.41
3	4.6551	S03_1504686	S03_1680047	8.74	1.09	0.21	2.42	2.32	2.88	-0.49	2.98	0.98
3	26.655	S03_6481871	S03_6797449	29.86	2.89	4.49	4.65	2.69	4.30	2.11	4.49	2.93
3	30.655	S03_7615673	S03_7920537	31.89	3.55	3.72	4.56	3.10	4.67	2.62	5.26	5.00
3	36.655	S03_9368445	S03_9433259	21.59	2.64	3.73	3.28	2.23	4.12	1.21	5.08	4.12
3	53.655	S03_14087144	S03_14687036	19.86	5.00	1.85	1.62	1.45	0.20	2.23	6.50	8.62
3	66.655	S03_16837159	S03_17302066	27.70	4.74	1.91	1.46	1.34	-1.62	0.40	7.31	7.12
3	68.655	S03_20983648	S03_21066372	28.73	4.70	2.17	1.41	1.67	-1.26	0.69	7.11	7.37
6	12.603	S06_2279056	S06_2984376	9.54	1.39	1.44	2.08	2.15	2.04	3.57	4.09	1.73
6	45.603	S06_8837126	S06_9318022	7.84	2.95	0.93	-0.27	0.56	-0.98	6.72	-0.68	-2.51
7	42.143	S07_8936752	S07_9350535	6.74	1.67	3.11	1.34	0.34	2.45	-3.84	-0.14	0.10
10	35.383	S10_16626134	S10_16772764	30.69	3.83	-4.15	-3.60	-2.72	-2.78	-6.59	-5.31	-3.88

[1] Joint-family QTL logarithm of odds

[2] Percentage of phenotypic variance explained by QTL.

[3] Positive and negative additive effects values mean that *aus* and T65 alleles increased trait values respectively.

Chapter 4 Expansion of *aus*-Derived NAM Population in Rice (*aus*-NAM-II)

Introduction

The number of markers is no longer a limitation in QTL mapping but the genetic material employed (Stadlmeier *et al.*, 2018). A large mapping population with more founders has a higher representation of rare alleles, an important factor for crop improvement (Kremling *et al.*, 2018; Valluru *et al.*, 2019; Gage *et al.*, 2020). In the models used for genomic selection in practical breeding programs, population size is more important than marker density (Xu *et al.*, 2018). Further expansion of our *aus*-NAM-I (**Chapter 2**) was conducted in this study. The extended and modified population contained 14 families and herein referred to as ***aus*-NAM-II**. To genotype the increased population size, our group developed a new GBS method called *i*GBS (Vincent P. Reyes dissertation 2021). From the new method, approximately 500M reads could be generated using HiSeqX (Illumina, CA, USA) at a comparable cost to the 25M reads generated by Illumina MiSeq (Illumina, CA, USA).

Materials and Methods

Plant Materials

A temperate *japonica* variety, Taichung 65 (T65), was used as the common female parent. The five *aus* cultivars Kasalath, Kalo Dhan, Shoni, ARC5955, and Badari Dhan (kindly supplied by the National Agricultural Research Organization (NARO) Genebank, Tsukuba, Japan) were used as diversity donor parents (founders) [28]. The five *aus* varieties were crossed to T65, and RILs were derived from the F2 generation using single-seed descent (SSD) to obtain F5 in 2015. The 5 families of RILs were designated as WNAM02 (Kasalath), WNAM29 (Kalo Dhan), WNAM31 (Shoni), WNAM35 (ARC5955), and WNAM39 (Badari Dhan). The nine new additional families were constructed in a similar way as the first five families. They were designated as WNAM04 (Jena), WNAM27 (Nepal 8), WNAM28 (Jarjan), WNAM30 (Anjana Dhan), WNAM32 (Tupa 121), WNAM33 (Surja Mukhi), WNAM34 (ARC7291), WNAM36 (Ratul) and WNAM40 (Nepal 55)

Phenotyping

A total of eleven rice traits were evaluated and measured in normal season field conditions at Nagoya University Togo field. Days to Heading (DTH) was recorded at Nagoya University Togo field, DTH was calculated as the difference between the emergence of inflorescence and sowing date. At the maturity stage, plants above the ground were sampled and dried in a well-ventilated vinyl house for one month. The following additional traits values were recorded: Culm Length (CL), Panicle Length (PL), Panicle Rachis Length (PRL), Panicle Number

Per Plant (PN), Panicle Weight (PW), Shoot Weight (SW), Number of Primary Branches Per Panicle (NPB), Number of Spikelets Per Panicle (NSSP) and Seed setting rate (SSR). Plant biomass (BM) is a summation of PW and SW.

Correlation Analysis

The correlation between traits was estimated using R (R Core Team, 2020) by Pearson method and visualized in corrplot package (v0.84) (Wei and Simko, 2017).

Genotyping

Nine new families in *aus*-NAM-II were sequenced using Illumina HiSeqX (Illumina, CA, USA). GBS library was prepared following protocol in **Chapter 2** except Illumina 'index' was added to the sequences of flowcell primers. Raw sequences were processed using the TASSEL-GBS pipeline (Glaubitz *et al.*, 2014) as explained in **Chapter 2.**

Estimation of Population Structure

The cleaned genotype information obtained from GBS was used to estimate population stratifications. Probabilistic PCA (PPCA) algorithm in the Bioconductor package PCA methods (Stacklies *et al.*, 2007) and implemented on R(R Core Team, 2020) was deployed.

Results

aus-NAM-II Population

The *aus*-NAM-II population development scheme is shown in Figure 4 1 out of the 14 families, 9 WNAM04, WNAM27, WNAM28, WNAM30, WNAM32, WNAM33, WNAM34, WNAM36, and WNAM40 were newly developed. The numbers of plants in F2 and F5 are listed in **Table 4-1**. Because of hybrid sterility and late heading, a substantial part of the plants in F2 through F5 could not be harvested. The residual rate at F5 ranged from 46.8% to 74.4% (**Table 4-1**). In total, 1,797 RILs, ranging from 54 to 169 per family were obtained (**Table 4-1**).

Correlation of Phenotypes

Correlations of phenotypes in 2015 (10 traits) and 2018 (Eleven traits) are shown in **Figure 4-2 .** Positive correlation between SW and DTH was observed in 2015 and 2018 while PN and NPB were negatively correlated. SSR positively correlated with PW and conversely to SW. Because BM was a component of SW and PW, a strong positive correlation was displayed as expected. The correlation trends between traits in 2015 and 2018 were generally almost similar.

Genotyping

The retained number of SNPs for onward analyses in *aus*-NAM-II after filtering for 'MAF = 0.02', 'max-missing = 0.5', and 'thin = 64bp', ranged from 2,522 to 5,019. The number of genotypes without excess heterozygosity (>0.125) was 1,818 (**Table 4-2**).

Population Structure

Estimation of population structure using probabilistic principal component analysis (PPCA) showed substantial controlled stratification, the R^2 values were: 0.096, 0.060,0.057, 0.053, 0.051 and 0.049 for PC1 to PC6 respectively. (**Figure 4-3** and **Figure 4-4**).

Discussion

Correlation of aus-NAM-II Traits

The correlation pattern of traits in 2015 and 2018 appeared similar. This suggested common environmental effects and uniform plots. Moreover, DTH was positively correlated to SW in both years, the delayed heading time likely enabled plants to accumulate photosynthates during the vegetative period thus getting higher SW. Pleiotropy or linkage disequilibrium is thought to be the source of traits correlations(Hill, 2013). The correlations of traits are useful for handling selection responses in plant breeding.

Development of aus-NAM-II Population

Rice *aus*-NAM-II population containing fourteen families and 1790 RILs was constructed in this study. The new population with additional founders was expected to have, a higher number of recombinant haplotypes just like MAGIC populations (Ladejobi *et al.*, 2016) as well as wider coverage of rare alleles, an important property in QTL mapping for crop improvement (Kremling *et al.*, 2018; Valluru *et al.*, 2019). The big size of *aus*-NAM-II population was also expected to mitigate over-estimation of QTL effects "Beavis effect", a common problem in small-sized mapping populations (Utz *et al.*, 2000). According to the retrospectives of maize NAM populations (Gage *et al.*, 2020), the authors recognized the importance of expanding NAM populations, they suggested that two to four times the original NAM size will enable the parental variety to share rapid LD decay which will, in turn, facilitate high-resolution gene mapping.

Genotyping of aus-NAM-II Population

The first five families (WNAM02, WNAM29, WNAM31, WNAM35, and WNAM39) were sequenced using MiSeq (Illumina, CA, USA) while the remaining nine families were sequenced using HiSeqX (Illumina, CA, USA). The potential number of the markers obtained by the restriction enzyme used in this study (*Kpn*I-*Msp*I) was estimated at approximately 5000. This number is sufficient to model the phenotypes based on rrBLUP related methods (Xu *et al.*, 2018), hence, enough for gene mapping. Therefore, the choice of the enzyme was appropriate for *aus*-NAM-II.

Population Structure

PPCA using the expanded population (*aus*-NAM-II) showed that RILs were evenly distributed including the fifteen parents. I presumed that no strong population structure was present in *aus*-NAM-II. A large population (14 families) and the common T65 parent considerably destroyed the population structure. Controlled crossing of varieties was previously reported to destroy population structure, this increases the power to detect QTL (Myles *et al.*, 2009)

Figures and Tables

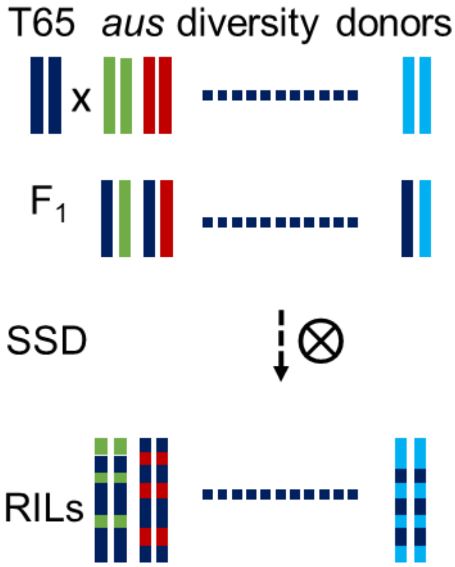


Figure 4-1. *aus*-NAM-II population development.

Depiction of *aus*-NAM-II population construction, temperate *japonica* Taichung 65 (T65) as the common female parent was crossed with 14 *aus* diversity donor parents. RILs were thereafter developed from F2 generation using single seed descent (SSD) method to obtain F5. Some families are not shown in the picture for the sake of visualization. The 14 families RILs were designated as WNAM2(Kasalath), WNAM4(Jena), WNAM27(Nepal 8), WNAM28(Jarjan), WNAM29(Kalo Dhan), WNAM30(Anjana Dhan), WNAM31(Shoni), WNAM32(Tupa 121), WNAM33(Surja Mukhi), WNAM34(ARC7291), WNAM35(ARC5955), WNAM36(Ratul), WNAM39(Badari Dhan), WNAM40(Nepal 55). Black 'x's indicate crosses between parents and circled '⊗'s indicate self-mating. Ellipses indicate many other individuals in the population.

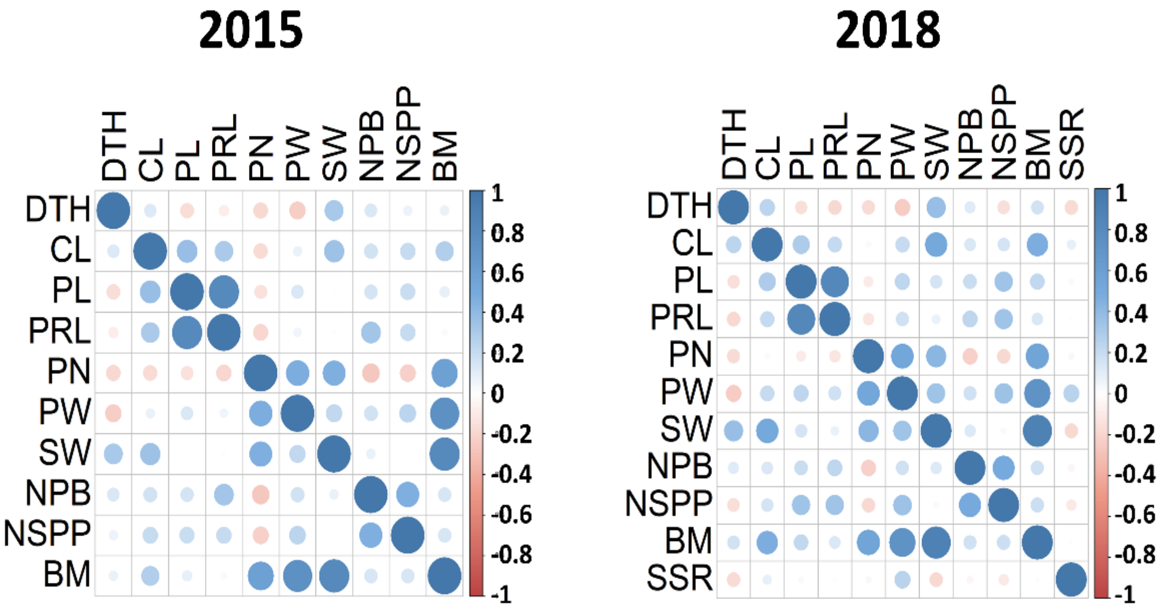


Figure 4-2. Correlations of phenotypes in 2015 and 2018.
The color of the dots indicates the correlation coefficient values corresponding to the bar on the right.

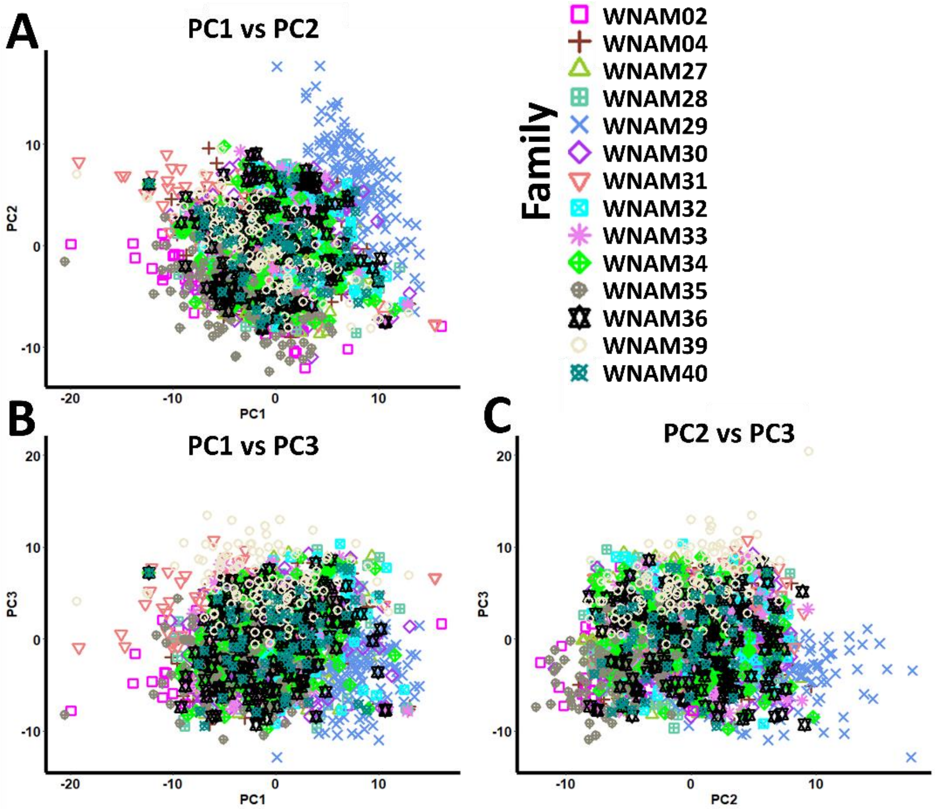


Figure 4-3. *aus*-NAM-II population structure (PC 1-3).

Probabilistic principal component analysis (PPCA) was used to estimate population structure, (A) PC1 vs PC2, (B) PC1 vs PC3, and (C) PC2 vs PC3. Different shapes and colors represent different *aus*-NAM-II families.

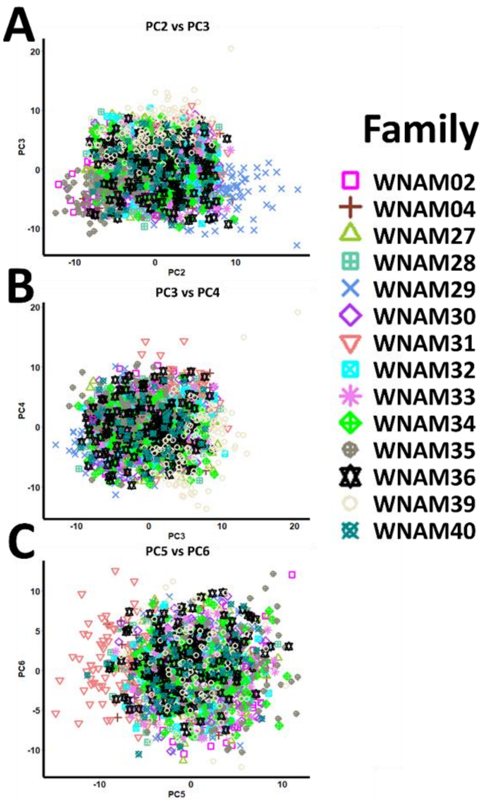


Figure 4-4. *aus*-NAM-II population structure (PC 2-6).

Probabilistic principal component analysis (PPCA) was used to estimate population structure, (A) PC2 vs PC3 (B) PC3 vs PC4 (C) PC5 vs PC6. Different shapes and colors represent different *aus*-NAM-II families.

Table 4-1. List of RILs in *aus*-NAM-II population.

[1]World Rice Core Collection (WRC)(Tanaka *et al.*, 2020)

Family name	Founder's name	WRC No.[1]	F_2	F_5	Residual rate
WNAM02	Kasalath	WRC02	233	109	46.78%
WNAM29	Kalo Dhan	WRC29	219	163	74.43%
WNAM31	Shoni	WRC31	174	121	69.54%
WNAM35	ARC5955	WRC35	229	137	59.83%
WNAM39	Badari Dhan	WRC39	213	126	59.15%
WNAM04	Jena035	WRC04	-	129	-
WNAM27	Nepal 8	WRC27	-	145	-
WNAM28	Jarjan	WRC28	-	151	-
WNAM30	Anjana Dhan	WRC30	-	169	-
WNAM32	Tupa 121	WRC32	-	88	-
WNAM33	Surja Mukhi	WRC33	-	86	-
WNAM34	ARC7291	WRC34	-	151	-
WNAM36	Ratul	WRC36	-	168	-
WNAM40	Nepal 555	WRC40	-	54	-

Table 4-2. Taxa and SNPs in *aus*-NAM-II population.

Family	No. of lines genotyped	No. of lines after basic filtering	No. of raw SNPs	No. of lines after MAF 0.02 ,max-missing 0.5 and thin 64bp	No. of SNPs after MAF 0.02 ,max-missing 0.5 and thin 64bp
WNAM02	114	110	19434	110	4399
WNAM04	135	131	9871	131	2799
WNAM27	151	147	11265	147	3211
WNAM28	157	153	11546	153	2522
WNAM29	170	166	13768	166	3732
WNAM30	175	171	13635	171	3919
WNAM31	126	122	12756	122	2980
WNAM32	94	90	12293	90	3700
WNAM33	92	88	10975	88	3359
WNAM34	157	153	18098	153	5019
WNAM35	135	131	13756	131	3444
WNAM36	174	170	11345	170	3276
WNAM39	134	130	11033	130	3408
WNAM40	60	56	11333	56	3555

Chapter 5 GWAS and Genomic Predictions using *aus*-NAM-II Population

Introduction

In plants genetics, genome-wide association study (GWAS) is a method to uncover QTL associated with traits variations, usually by statistically examining the relationship between whole-genome sequence variants and traits (Nordborg and Weigel, 2008). GWAS can be advantageous to plants because of the possibility to develop fixed populations thus genotype once and phenotype many times for different traits (Xiao *et al.*, 2017). In the recent past, many GWAS studies have been conducted in plants using natural populations, for example, rice (Huang *et al.*, 2010) and maize (Li *et al.*, 2013). However, GWAS power to detected QTL has been dismal due to uncounted population structures (Zhou and Huang, 2019). To circumvent the problem, a nested association mapping (NAM) population was proposed (McMullen *et al.*, 2009) and since then several GWAS studies have been conducted using NAM populations (Tian *et al.*, 2011).

Modern plant breeding is a predictive science driven by new technologies and knowledge (Crossa *et al.*, 2021). Predictive ability (PA) is one of the key factors determining the application of genomics-based breeding, and prediction models are designed to improve PA. Traditionally, organism selection was based on phenotypes, with the advent of quantitative genetics and statistics, the best linear unbiased prediction (BLUP) method that utilizes phenotypic and pedigree information was proposed (Henderson, 1985; Bernardo, 1996). Because this method was time-consuming and costly, marker-assisted selection (MAS) was championed as an alternative (Fujino *et al.*, 2019). MAS success relied on QTL with large effects (Bernardo and Yu, 2007). However, quantitative traits are complex and influenced by many genes with small effects that are not significant in QTL analysis, and interacts with environments (Yonemaru *et al.*, 2010). As a new genotypes selection alternative, methods using molecular markers without the need of computing QTL statistical significance were proposed (Meuwissen *et al.*, 2001; Nakaya and Isobe, 2012) This approach is commonly referred to as genomic selection (GS).

The hypothesis behind GS is that with high-density markers, each QTL should be associated with at least one marker. GS ranks individuals for selection based on their estimated breeding values (GEBVs). GS can be applied at any stage of the breeding cycle providing much-needed flexibility to breeders. In addition, GS is a promising technology for traits that are difficult or expensive to measure like milling quality (Monteverde *et al.*, 2018).

GS models use two types of population (i) Training population (TP) which has genotype and phenotype data (ii) Breeding population or validation population which has genotype information only. Since genetic data are mostly high-dimensional i.e. more marker number (p) than individuals (n), the so called large-p-with-small-n problem (Bellman, 1961). Various prediction methods employ different dimension reduction and selection on the many parameters(p). The various statistical models implemented use the TP, where marker effects are firstly estimated (Jonas and de Koning, 2013). The marker effects are then used to calculate GEBVs of individuals in the breeding populations. The response to selection can be evaluated in plant breeding by the genetic gain achieved in GS. Genetic gain can be calculated using equation (2):

$$\mathbf{R_t} = \frac{\mathbf{ir\sigma_A}}{\mathbf{y}} \qquad (2)$$

Where R_t denotes a genetic gain over time, i represent selection intensity, r corresponds to selection accuracy measured as the correlation between actual breeding values and estimated breeding values. In the case of repeated measurements, accuracy is adjusted by dividing it by the square root of the narrow-sense heritability (h). σ_A denotes the genetic variance brought about by diversity in the population and y is the number of years in the breeding cycle (Li *et al.*, 2018).

To achieve optimum genetic gain, selection intensity, accuracy and genetic variance should be improved i.e. through increasing the population size and accounting for environmental artifacts correctly (Xu *et al.*, 2020). For GEBVs obtained from markers, optimum accuracy depends on the choice of models which rely on the genetic architectures of traits, sample size and linkage disequilibrium(Campos *et al.*, 2013).

As the development of computing and quantitative genetics improves and the cost of sequencing reduces, attempts for the implementation of GS in crops continue to rise. Genomic predictions have been applied to many rice traits such as heading dates, grain yield, plant height, milling quality, arsenic concentration, and biomass (Onogi *et al.*, 2016; Monteverde *et al.*, 2018; Frouin *et al.*, 2019; Toda *et al.*, 2020). Recently, genetic researches have elucidated that incorporating trait-related data i.e. growth-related (Toda *et al.*, 2020), environments data (Monteverde *et al.*, 2018), and phenological data i.e. heading date (Onogi *et al.*, 2016) improves genomic predictions. Despite several studies of genomic predictions in rice, this study, to the best of my knowledge, is the first to apply genomic prediction in *aus* derived NAM population.

In this chapter, I evaluated the accuracy of four genomic prediction methods. The performance of several methods and markers size impacts on genomic prediction was examined.

Materials and Methods

Genotype and Phenotype Data

Genotype and phenotype information used here are described in **Chapter 4** of this dissertation.

GWAS

For GWAS, the GWAS function implemented in the rrBLUP package (Endelman, 2011) was utilized. This function performs GWAS based on a mixed model (Yu *et al.*, 2006) as shown in equation (3).

$$y = x\beta + S\tau + Zg + \varepsilon \qquad (3)$$

Where β vector of fixed effects (e.g. PCA), τ models additive SNP effect as a fixed effect; g models genetic background of each line as a random effect with variance explained by [g] = Kσ2. ε is residuals with variance explained by [ε] = Iσ2. Two sets of marker genotype data were prepared: (i) Fourteen families merged GBS SNPs (ii) Fourteen families merged GBS SNPs plus the parental variants projected onto it. GWAS with Q (population structure) with K (kinship relatedness) was executed in this study. Minor allele frequency was set to 0.05 and the number of principal components (n. PC) was set to 4. To visualize the results, the qqman package was used (Turner, 2018). Since GWAS performs hypothesis testing for each of the large number of SNPs, the significance level was calculated using the p. adjust function with the false discovery rate (FDR) set to 0.05 (5%). Based on FDR <0.05, SNPs that were considered significant were highlighted in green color on the Manhattan plots. The Linkage disequilibrium (LD) values between pairs of SNPs were determined from the squared correlation coefficients (r^2) values using plink (Purcell *et al.*, 2007) with following settings (--ld-window-kb 43000 --ld-window 100000), LD decay pattern was rendered according to the Hill and Weir function (Weir and Hill, 1980).

Methods for Genomic Prediction

In this study, methods in rrBLUP (Endelman, 2011) and BGLR (Pérez and Campos, 2014) packages were utilized. The results were visualized using ggplot2 package. Of the four methods used, three were parametric; (rrBLUP, BayesB, and Bayesian LASSO) and one non-parametric method (RKHS). The general equation for the parametric model is described by equation (4).

$$y = X\beta + \sum_{k=1}^{m} Z_k \gamma_k + \varepsilon \qquad (4)$$

Where y is a vector for n observed phenotypes, X is a matrix of fixed effects of size n× q, β is a q × 1 vector of fixed effects, m is the number of markers (SNPs), Z_k corresponds to vector for genotype indicator variable, γ_k is the additive genetic effect of marker k, and residual errors is represented by ε, n × 1 vector with an assumed N (0, $I\sigma^2$) distribution. The genotypic indicators of marker k for individual j (where j = 1, 2.. n) are defined as − 1, 0, 1 i.e. homozygote of the minor allele, heterozygote, and the homozygote of the major allele, respectively (Xu *et al.*, 2018).

Genetic Relationship Matrices

Kinship coefficients between RILs were estimated from genomic data. A realized genomic relationship matrix (G) was calculated as described by Van Raden (VanRaden, 2008) and shown in equation (5).

$$\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})'}{2\sum p_i(1-p_i)} \qquad (5)$$

where M and P are two matrices of dimension n (number of individuals) × p (number of markers). In matrix M, homozygote, the heterozygote, and the other homozygote are represented by −1, 0, and 1 respectively. P denotes the matrix containing the allele frequencies in this form: $2(p_i - 0.5)$, where p_i is the observed allele frequency at the marker *i* for all individuals. The use of minor allele frequency scales G to the expected additive genetic relationship (Bartholome *et al.*, 2016).

Bayesian Least Absolute Shrinkage and Selection Operator (Bayesian LASSO)

The BGLR package implements many Bayesian-based regression methods, the algorithms are based on Gibbs sampler with scalar updates(Pérez and Campos, 2014). The BGLR Bayesian LASSO (BL) was selected for analysis in this study because of its computing speed and the few prior assumptions assigned to the model. The default BGLR Bayesian LASSO parameters were applied to run the model. The marker effects in BGLR Bayesian LASSO are assigned double exponential distribution (**Figure 5-1**) (Heslot *et al.*, 2012; Pérez and Campos, 2014).

BayesB

BGLR BayesB model uses a mixed prior distribution with mass at zero and a slab that has scaled t distribution (**Figure 5-1**)(Meuwissen *et al.*, 2001). The prior used in BGLR BayesB has the potential to be used for variable selection due to assigning non-null prior for marker effects to be zero (Pérez and Campos, 2014). BGLR BayesB variance can summarized as $\gamma_k \sim N\ (0; \sigma^2\ \gamma_k)$, where $\sigma^2\ \gamma_k = 0$ with prob = π and $\sigma^2\ \gamma_k \sim \chi{-2}(v, S)$ with prob = $1 - \pi$. The package`s default parameters were used(Pérez and Campos, 2014).

Reproducing Kernel Hilbert Space (RKHS)

BGLR RKHS utilizes the *Gaus*s kernel function to fit the model, the model is described by equation (6).

$$y = \mu + K_h\alpha + \varepsilon \qquad (6)$$

where μ is mean of population and α corresponds to covariance matrix $K_h\sigma^2_\alpha$; ε is residuals with distribution $\varepsilon \sim N\ (0, I_n\sigma^2)$; K_h is a kernel function that represents the correlation between individuals and is represented by the equation (7).

$$\mathbf{K}_h(x_i, x_j) = \exp(-hd_{ij}) \qquad (7)$$

where d_{ij} is the squared Euclidean distance between individuals i and j calculated based on their genotypes, and the smoothing parameter h is defined as h = 2/d* and d* is the mean of d_{ij}, h value was fixed to 0.5 in this study. BGLR RKHS can handle epistasis and is solved using a Gibbs sampler in a Bayesian framework, or using a mixed linear model (Wang *et al.*, 1994).

Ridge Regression with rrBLUP

Marker effects and GEBVs for traits were predicted using rrBLUP 'mixed. solve' function(Endelman, 2011). The basic model in rrBLUP is described by equation (8)

$$y = WGu + \varepsilon \qquad (8)$$

Where y is observation, u is a vector of marker effects with distribution u~N (0, Iσ^2u), G is genotype matrix and W design matrix relating to observation(y). BLUP solution for marker effects can be described by $\hat{u}$=Z'(ZZ' + λI) $- ^1$ y; where Z = WG and the ridge parameter $\lambda = \sigma^2 e/\sigma^2 u$ is the ratio of residual and marker effects variances(Searle *et al.*, 2006). 'mixed. solve' function calculates maximum likelihood solutions (ML/REML) for mixed models where a single variance component apart from residual error has a relationship with ridge regression.

Cross-Validation

For each dataset in the study, RILs were randomly sampled into five groups. Each model was trained on 4/5 of the data subsets and accuracy tested on the remaining 1/5 of the subset. The cross-validation was replicated 10 times before the average prediction accuracy for each trait was calculated. Prediction accuracy (r values) was defined as the linear correlation between true phenotypic records and the predicted individual's breeding value (Pearson, 1895).

Results

Linkage Disequilibrium (LD) Decay

After filtering SNPs and performing imputations, the merged 56,042 SNPs were analyzed for the distance at which LD was half of its maximum value. The squared correlation coefficient (r^2) values of the pairwise LD were plotted using a nonlinear regression curve against physical distance (kb) to estimate the LD decay pattern. The regression curve pattern showed that LD decayed to half ($r2 < 0.26$) within 2.86 Mb (**Figure 5-2**).

GWAS for DTH in 2019

Days to heading (DTH) was the only trait evaluated trait in 2019. GWAS for DTH using 2006 GBS SNPs (marker set 1) and 78,154 projected markers (marker set 2) were profiled. The number of SNPs that were considered significant (false discovery rate (FDR) set to 0.05) were: 63 and 395 for marker set 1 and marker set 2 respectively. For marker set 1, significant SNPs were detected on chromosomes 4 (S04_253061326), 6 (S06_3043159), 7 (S07_15686122) and 10 (S10_17200086). The candidate genes were annotated based on QTARO database (Yamamoto *et al.*, 2012). QTL detected on chromosomes 6 and 10 were considered *RFT1* and *Ehd1* respectively (**Table 5-1**), and QTL linked to peaks on chromosomes 4 and 7 were probably novel. Marker set 2 showed two additional candidate genes *DTH2* and *Hd9* for S02_29724553 and S03_1825245 respectively (**Table 5-1** and in **Figure 5-3**).

GWAS for Phenotypes in 2018

Manhattan and QQ plots for 2018 are shown in **Figure 5-4**. There was an outstanding QTL approximately 16Mbp to 17Mbp on chromosome 10 for DTH using both marker sets, herein referred to as qDTH10. Other chromosomes with significant SNPs associated with traits were: Marker set 1, DTH (chromosomes 1, 4, 6 and 7), CL (1, 5, 6 and 10), PL (2, 3, 4, 6, and 11), PRL (4 and 11), PN (4), SW (5, 6 and 10), NSPP (3, 4 and 7) and BM (10). For marker set 2: DTH (2, 3, 6 and 7), CL (2, 3, 5, 8 and 9), PL (2, 4 and 6), PRL (2, 4, 5 and 6), PN (1, 2, 3, 4, 5, 7 and 12), PW (6), SW (1, 5, 6, 7, 9 and 10), NSPP (3, 4, 7, 8 and 11) and BM (10) (**Table 5-2**).

GWAS for Phenotypes in 2015

Manhattan and QQ plots for 2015 are shown in **Figure 5-5**. The chromosomes with significant SNPs associated with traits variations were as follows: Marker set 1, DTH (chromosomes 6, 7 and 10), CL (1 and 5), PL (4), and PRL (8) while the chromosomes of significant SNPs in the marker set 2, DTH (3, 6, 7 and 10), CL (1, 6 and 12), PL (4 and 6), PRL (4 and 8), PN (2, 4 and 12), PW (8), SW (10) and NPB (2) (**Table 5-3**).

GP for DTH in 2019

Four statistical models (i) Bayesian B (BayesB), (ii) Bayesian least absolute shrinkage and selection operator (BL), (iii) reproducing kernel Hilbert space regression (RKHS), and (iv) ridge regression best linear unbiased prediction (rrBLUP) were explored. The prediction accuracies for DTH from the models tested are shown in **Figure 5-6**. The highest average accuracy 'r' was 0.894 by RKHS model while the least prediction accuracy was 0.757 by BL model (**Table 5-4**). When only significant GWAS markers (63 SNPs) in marker set 1 (**Table 5-1**) were utilized as the explanatory variables in the model, prediction accuracy 'r' dropped to 0.61 (**Figure 5-6** C).

GP for Phenotypes in 2018

Prediction accuracies for traits evaluated in 2018 using BayesB, BL, RKHS, rrBLUP models are shown in **Figure 5-7**. The highest and lowest accuracies were DTH and SSR with correlation coefficients (r) of 0.894 and 0.292 respectively. RKHS model yielded the highest prediction accuracy while rrBLUP was the least performing model (**Table 5-4**).

GP for Phenotypes in 2015

Prediction accuracies for traits evaluated in 2015 using BayesB, BL, RKHS, rrBLUP models are shown in **Figure 5-8**. The highest and lowest accuracies were CL and BM with correlation coefficients (r) of 0.903 and 0.345 respectively. The model with the highest accuracy was RKHS while rrBLUP was the least performing model (**Table 5-5**).

Discussion

Improved Statistical Power of GWAS Using aus-NAM-II

To test GWAS power for discovering QTL, marker-trait association analysis was performed for DTH in 2019, eleven traits in 2018, and ten traits in 2015. Several QTL were detected confirming NAM population design improved GWAS power (McMullen *et al.*, 2009; Zhou and Huang, 2019), some of the QTL co-localized with reported QTL. There was one stable QTL for DTH on chromosome 10 (qDTH10), qDTH10 was detected across the 3 years. Besides qDTH10, another stable QTL for PL was detected across 2 years (2015 and 2018), the QTL was located at around 24Mbp on chromosomes 4 (**Table 5-2** and **Table 5-3**).

Predictive Ability for Traits

To explore the potentials of genomic prediction using *aus*-NAM-II population, four models that vary in hypotheses for marker effects sizes and parameters dimension reduction were evaluated. The prediction results suggested that the RKHS model which is based on kernel matrix thus can handle epistasis in addition to additive effects, was the best performing in most traits. Similar findings were reported in the *japonica* diversity panel and advanced breeding lines with *japonica* genetic backgrounds (Frouin *et al.*, 2019).

The level of predictive ability for DTH and CL using five-fold cross-validations was similar to other rice studies (Isidro *et al.*, 2015; Spindel *et al.*, 2015; Onogi *et al.*, 2016) and other crops too (Guo *et al.*, 2014; Cantelmo *et al.*, 2017). The prediction accuracy for BM was low as expected i.e. a trait with low heritability and the fact that BM accumulated experimental noises from the summation of

PW and SW, similar results were observed in SSR. Other traits examined in this study showed mid to high-level prediction accuracy.

Within the cross-validation folds, predictive ability was slightly different, this phenomenon has been reported as well and it was attributed to differences in linkage disequilibrium and allele frequencies between the datasets. I also surveyed prediction accuracies using only significant markers that were detected in GWAS, this phenomenon mimics MAS in away i.e. significant QTL only(Reyes *et al.*, 2021). The reduced DTH prediction accuracies when significant QTL were only utilized elucidates that marker number is not the only important factor in genomic prediction, other factors such as the genetic architecture of the traits play a major role.

In summary, traits that were detected to have significant QTL in GWAS showed higher predictive abilities generally, this suggests that future research aimed at improving genomic predictions should include the available QTL information as was previously proposed (VanRaden, 2008; Zhang *et al.*, 2014) or incorporate other omics data such as crop models (Onogi *et al.*, 2016).

Figures and Tables

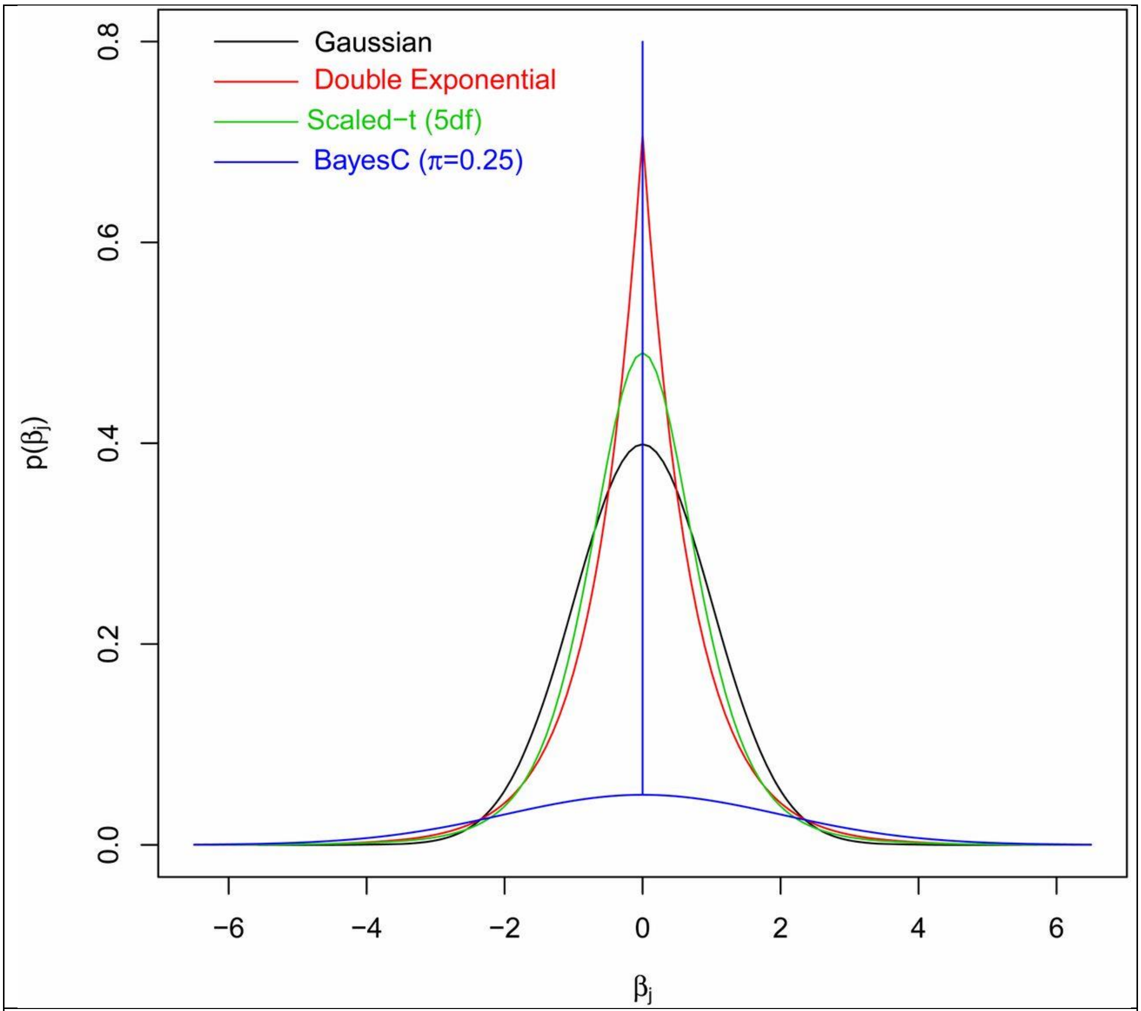


Figure 5-1. BGLR package regression prior assumptions.
Regression coefficients for priors to determine the size of shrinkage in estimating marker effects. Gaussian prior induces shrinkage like that of ridge regression, double exponential prior induces shrinkage that depends on the size of the effect and used in BL, BayesB utilizes a mixture of priors which can induce variable selections.

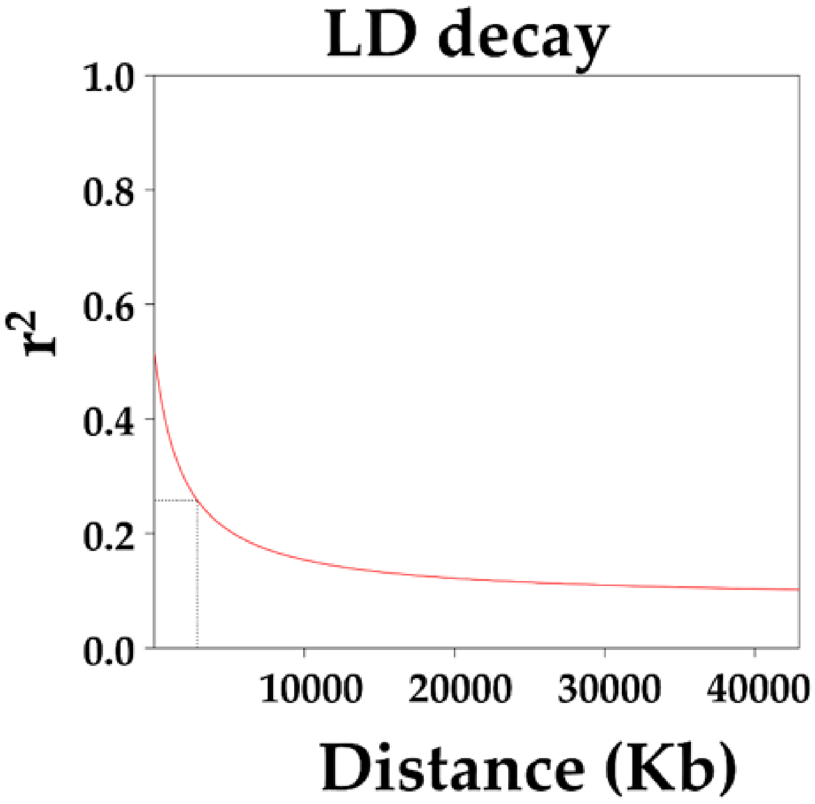
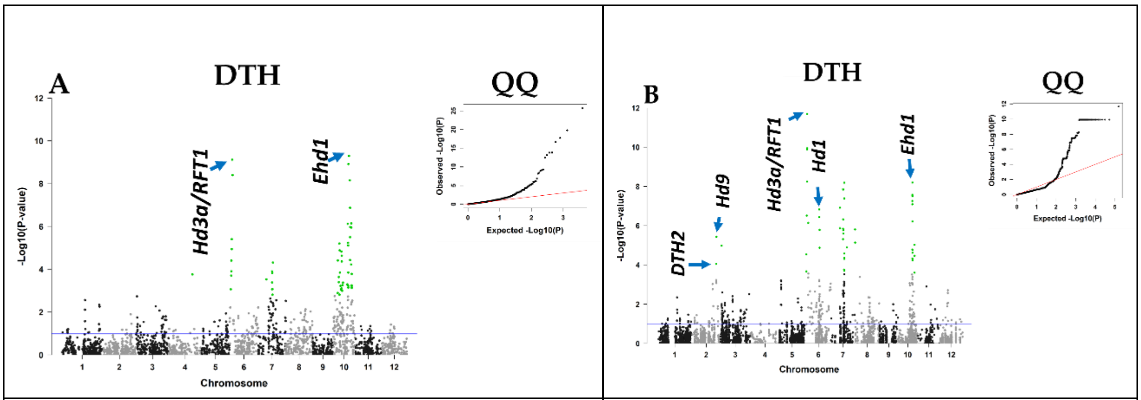


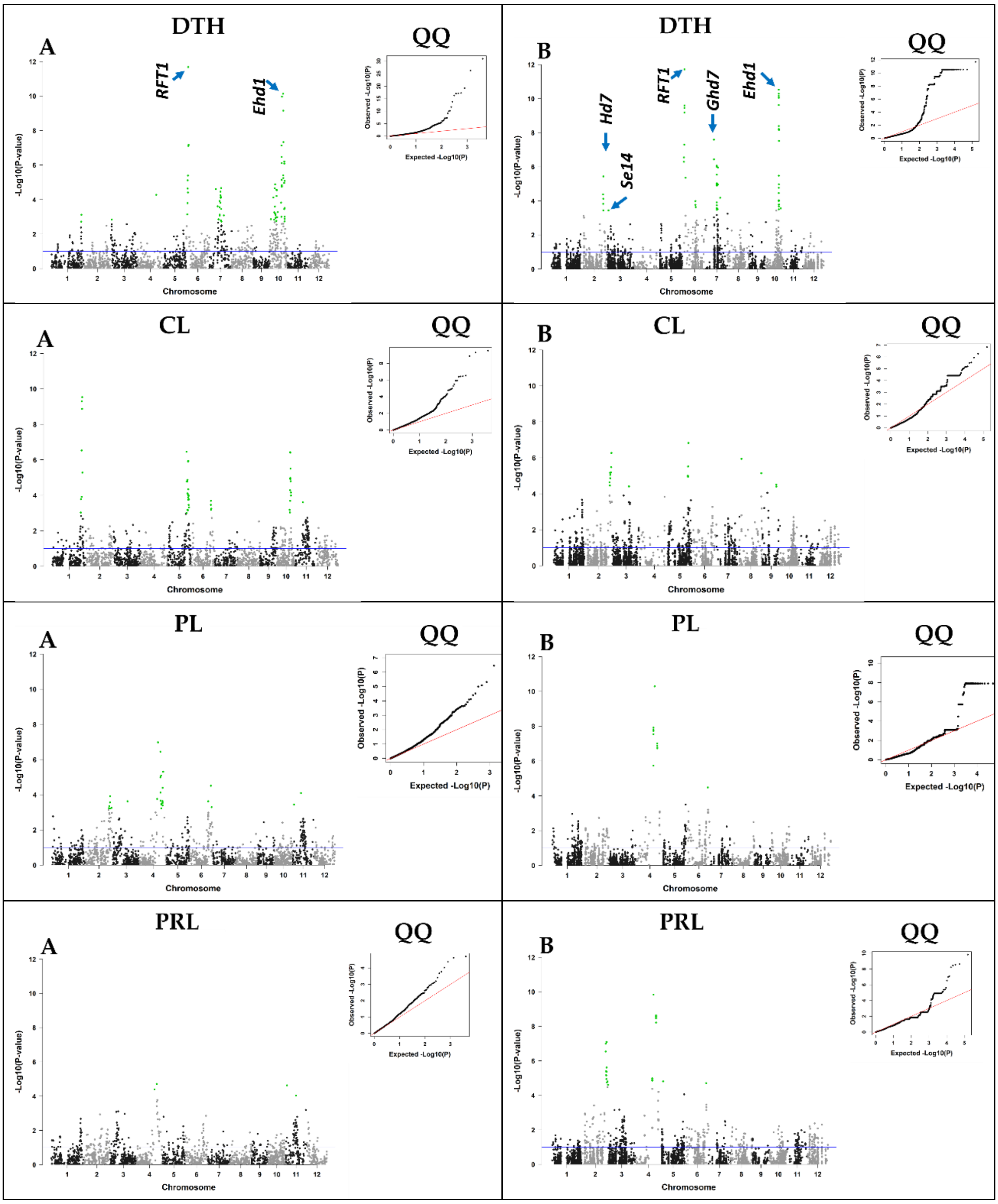
Figure 5-2. Linkage disequilibrium (LD) decay pattern.
LD pattern was calculated based on Hill and Weir function with Kb values on the x-axis and r2 values on the y-axis. Estimates of LD over genetic distance were conducted for all chromosomes in 1,709 RILs with 56,042 SNPs. The red curve indicates the LD decay pattern that was estimated by fitting a trend line based on a nonlinear regression of r2 on physical distance. Vertical dotted lines corresponded to LD halving distance (2.86Mb).

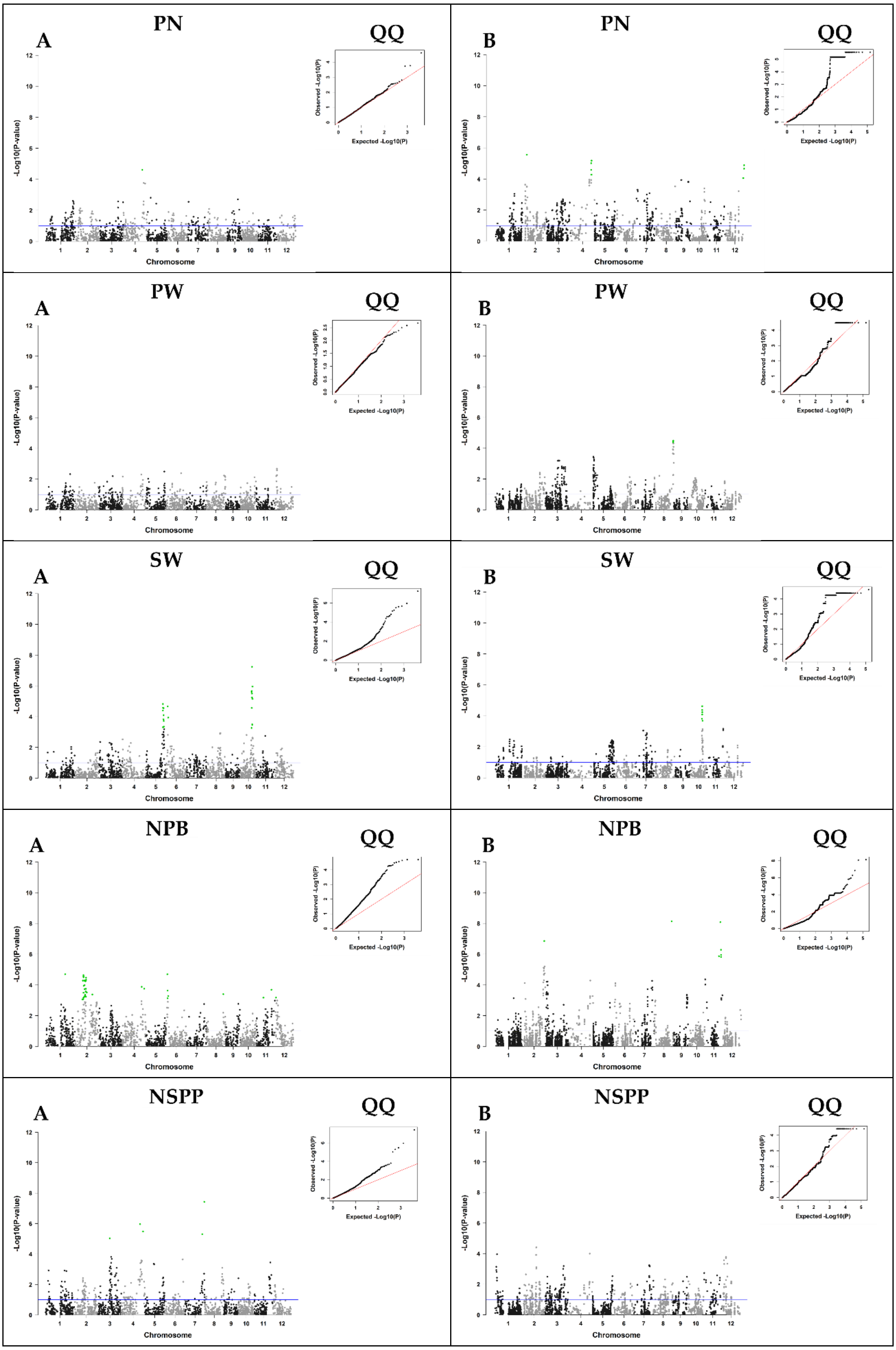


GWAS (Marker set 1 (2006 SNPs))	GWAS (Marker set 2 (78154 SNPs))

Figure 5-3. GWAS for days to heading in 2019 using *aus*-NAM-II.

Manhattan plots for (A) marker set 1; 2006 GBS SNPs (B) marker set 2; 78,154 projected SNPs. The GWAS results were obtained using a mixed linear model with principal components (Q) and genomic kinship (K) as covariates (MLM (Q+K)). The blue horizontal line marks the threshold for genome-wide significance on a -log10 scale. SNPs were considered significant using false discovery rate (FDR) set to 0.05 (5%) and highlighted in green color. A quantile-quantile (QQ) plot is shown in the right panel, where the observed P-values (Y-axis) against the expected P-values (X-axis) under the null hypothesis of no association are plotted on a -log10 scale. Each black dot denotes SNP.


A
DTH
QQ
RFT1
Ehd1
B
DTH
QQ
Hd7
Se14
RFT1
Ghd7
Ehd1
A
CL
QQ
B
CL
QQ
A
PL
QQ
B
PL
QQ
A
PRL
QQ
B
PRL
QQ
-Log10(P-value)
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12
Observed -Log10(P)
Expected -Log10(P)


A
PN
QQ
B
PN
QQ
A
PW
QQ
B
PW
QQ
A
SW
QQ
B
SW
QQ
A
NPB
QQ
B
NPB
QQ
A
NSPP
QQ
B
NSPP
QQ
-Log10(P-value)
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12
0 2 4 6 8 10 12
Observed -Log10(P)
Expected -Log10(P)

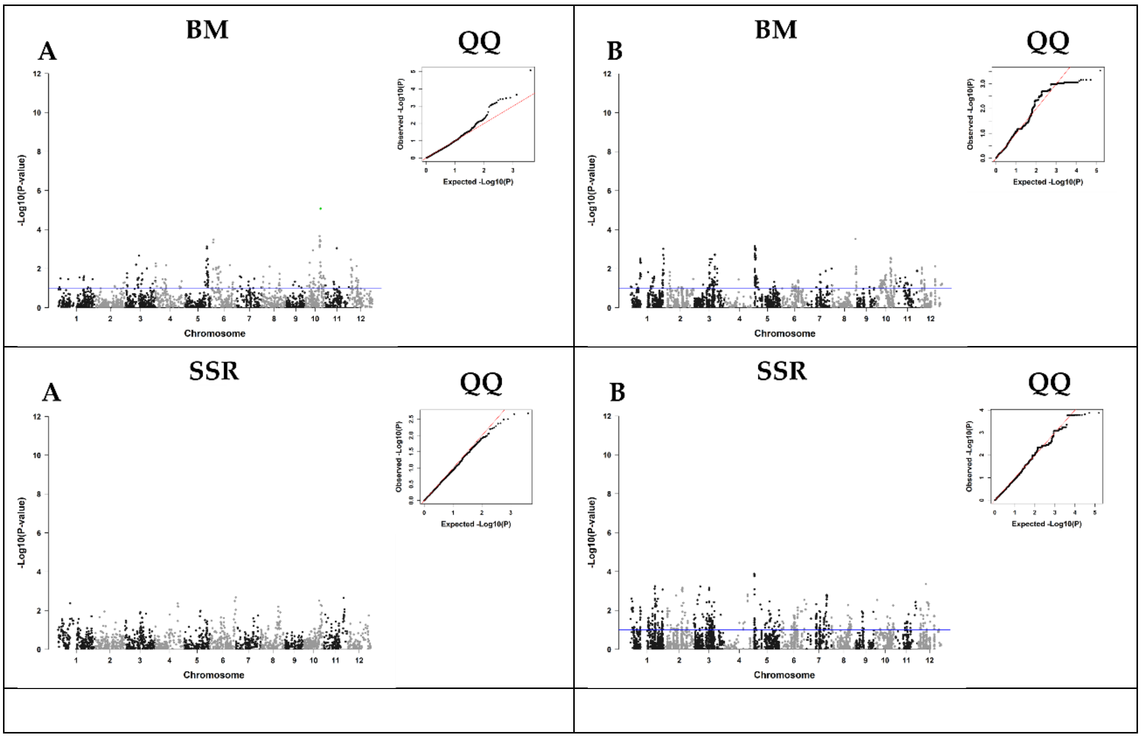
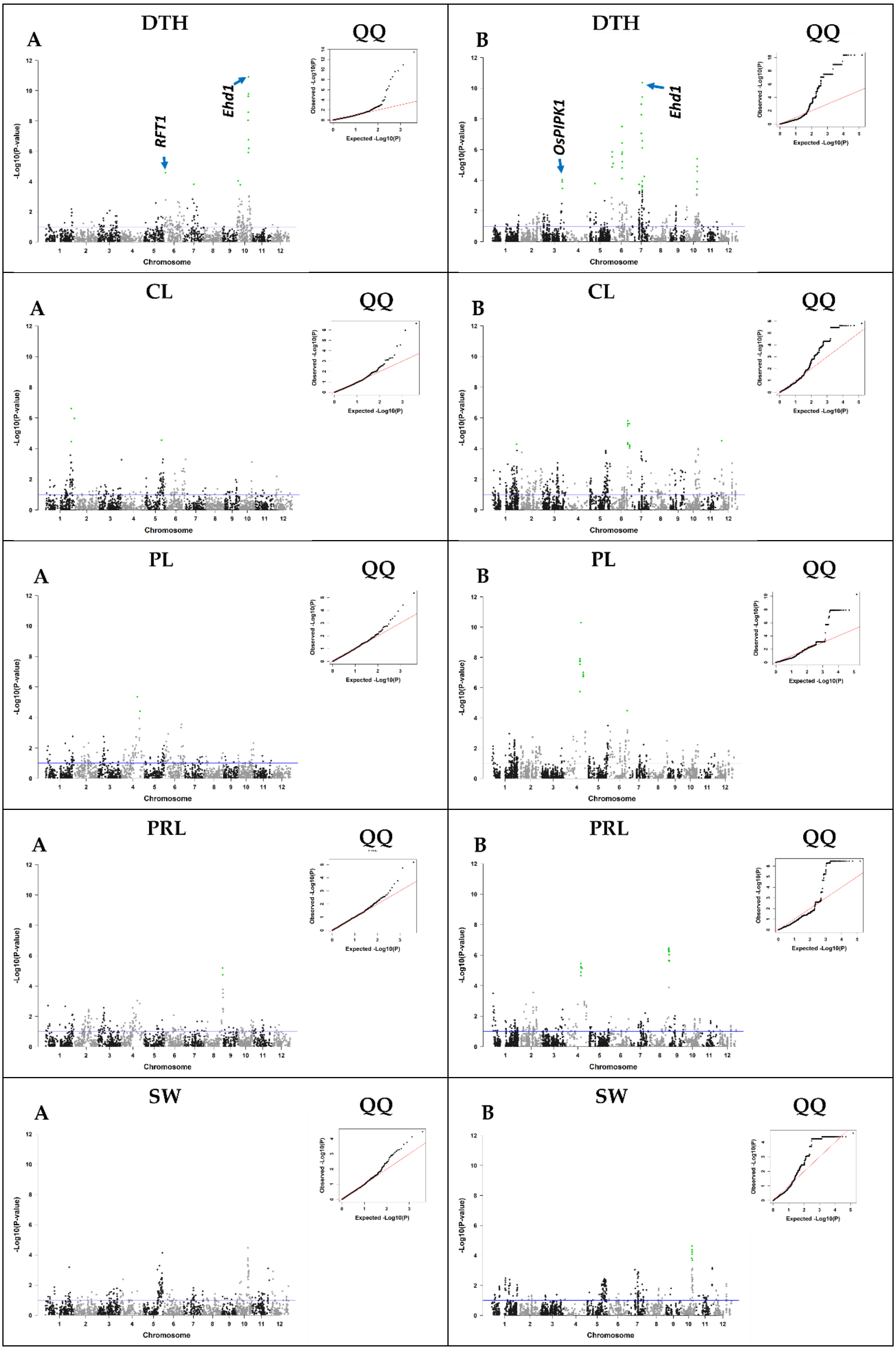


Figure 5-4. GWAS for phenotypes in 2018 using *aus*-NAM-II.

Manhattan plots for (A) marker set 1; 2006 GBS SNPs (B) marker set 2; 78,154 projected SNPs. The GWAS results were obtained using a mixed linear model with principal components (Q) and genomic kinship (K) as covariates (MLM (Q+K)). The blue horizontal line marks the threshold for genome-wide significance on a -log10 scale. SNPs were considered significant using false discovery rate (FDR) set to 0.05 (5%) and highlighted in green color. A quantile-quantile (QQ) plot is shown in the right panel, where the observed P-values (Y-axis) against the expected P-values (X-axis) under the null hypothesis of no association are plotted on a -log10 scale. Each black dot denotes SNP.


DTH
A
QQ
RFT1
Ehd1
-Log10(P-value)
Chromosome
Observed -Log10(P)
Expected -Log10(P)
DTH
B
QQ
OsPIPK1
Ehd1
CL
A
QQ
CL
B
QQ
PL
A
QQ
PL
B
QQ
PRL
A
QQ
PRL
B
QQ
SW
A
QQ
SW
B
QQ
1
2
3
4
5
6
7
8
9
10
11
12

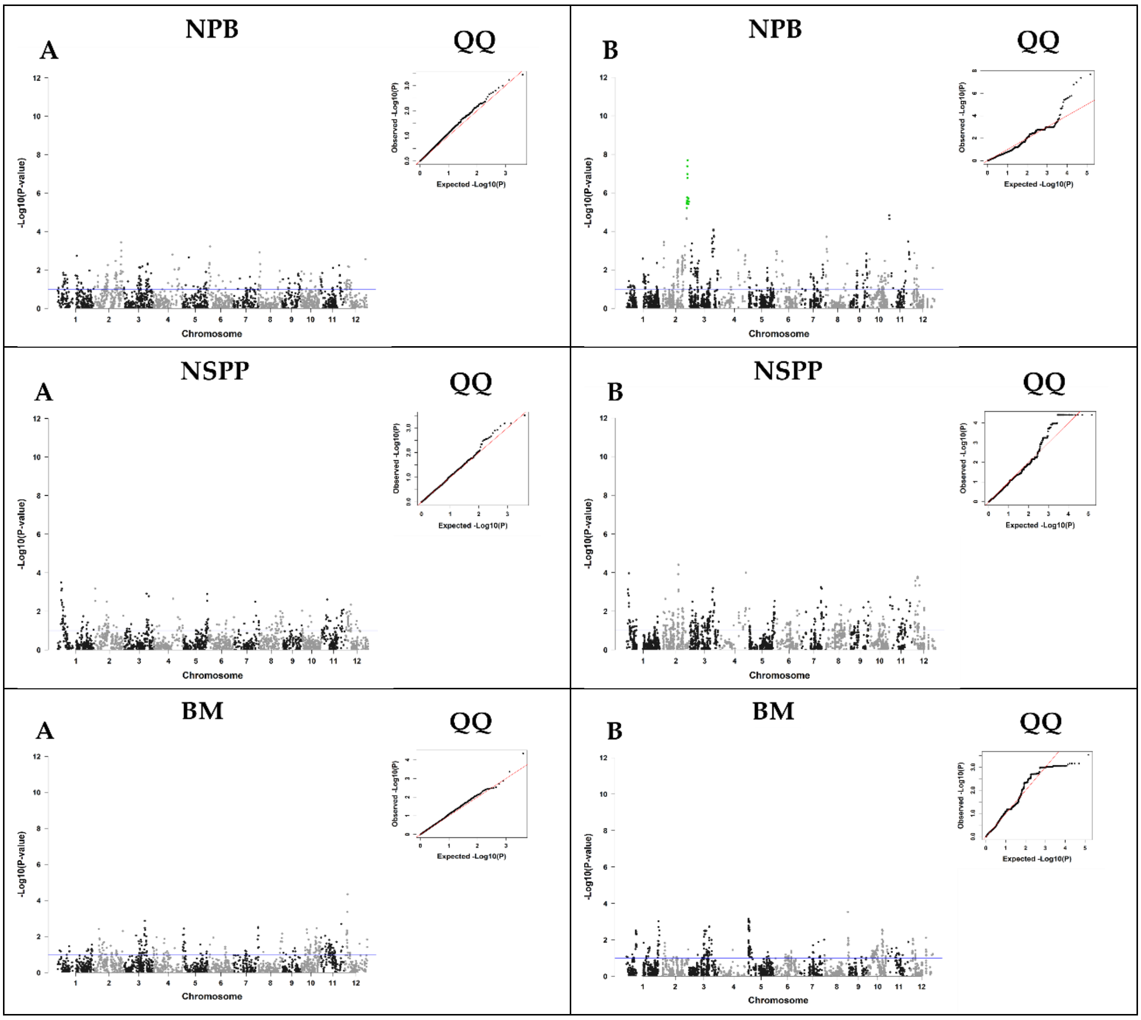


Figure 5-5. GWAS for phenotypes in 2015 using *aus*-NAM-II.

Manhattan plots for (A) marker set 1; 2006 GBS SNPs (B) marker set 2; 78,154 projected SNPs. The GWAS results were obtained using a mixed linear model with principal components (Q) and genomic kinship (K) as covariates (MLM (Q+K)). The blue horizontal line marks the threshold for genome-wide significance on a -log10 scale. SNPs were considered significant using false discovery rate (FDR) set to 0.05 (5%) and highlighted in green color. A quantile-quantile (QQ) plot is shown in the right panel, where the observed P-values (Y-axis) against the expected P-values (X-axis) under the null hypothesis of no association are plotted on a -log10 scale. Each black dot denotes SNP.

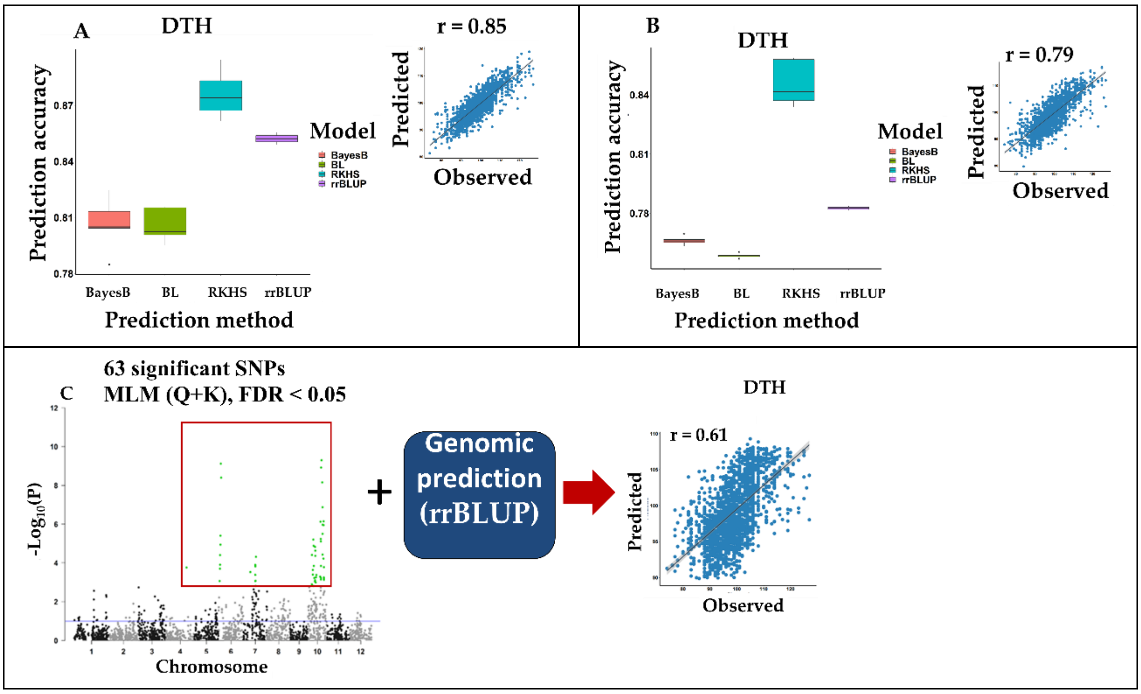
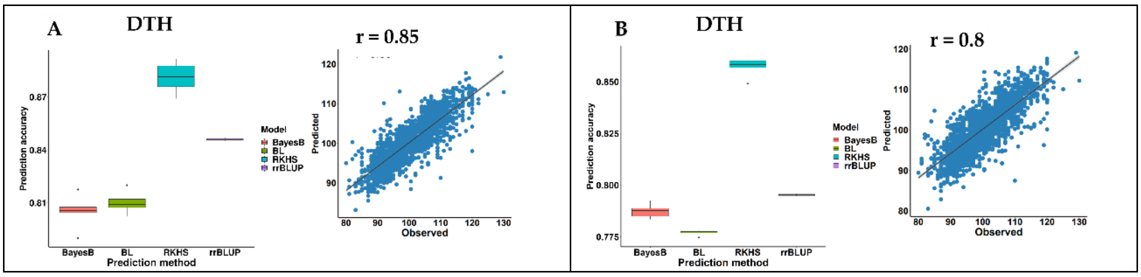
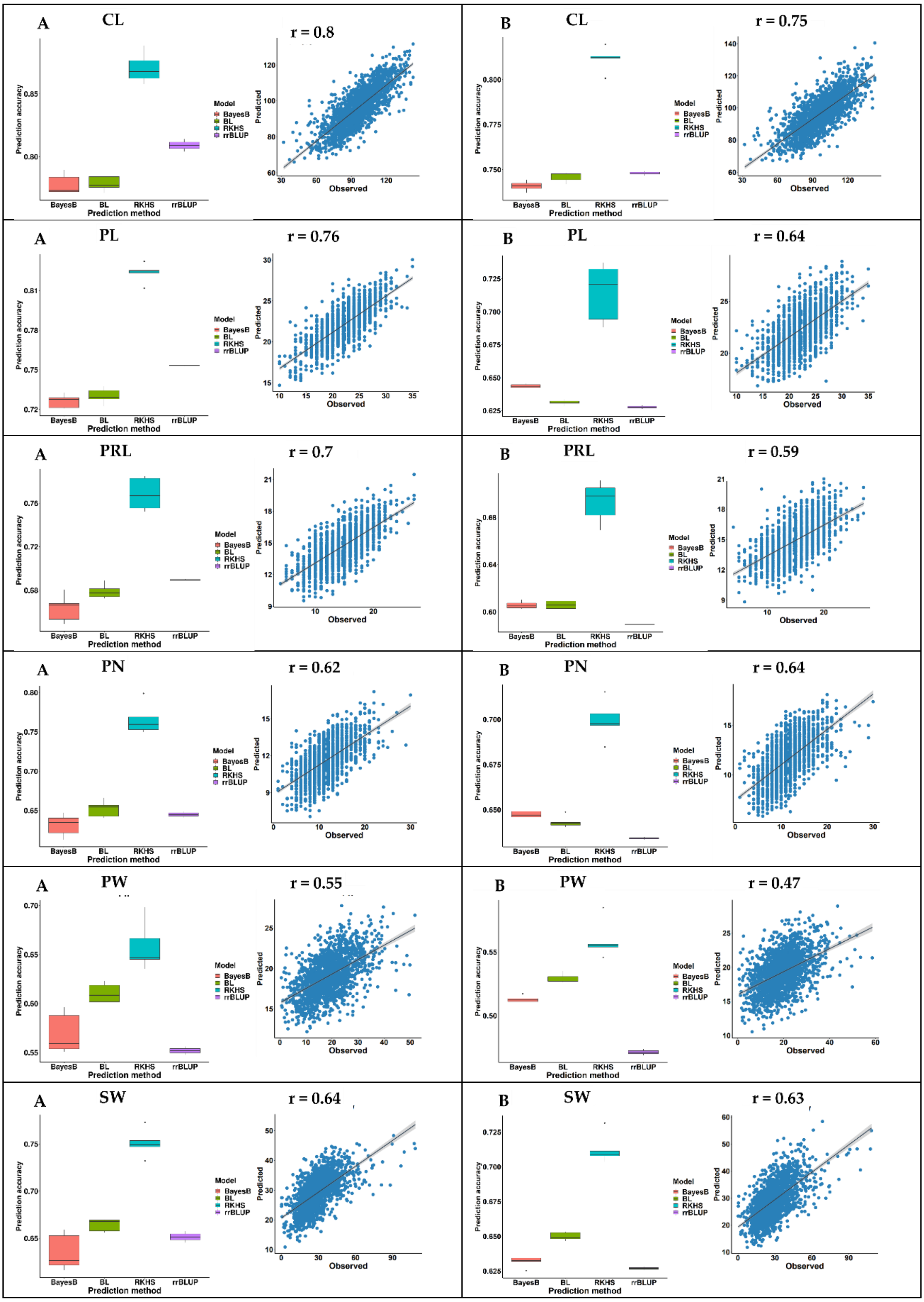


Figure 5-6. Genomic prediction for days to heading in 2019.

(A) Marker set 1; 2006 GBS SNPs (B) Marker set 2; 78,154 projected SNPs (C) Significant markers from GWAS (63 SNPs). Prediction accuracy was obtained by Pearson correlation of the observed phenotypes against the predicted phenotypes using four models: Bayesian B (BayesB), Bayesian least absolute shrinkage and selection operator (BL), reproducing kernel Hilbert space regression (RKHS), and ridge regression best linear unbiased prediction (rrBLUP). The scatter plot on the right of each plot shows the distribution of observed phenotypes (X-axis) against predicted phenotypes(Y-axis) using rrBLUP with its correlation coefficient (r) shown on the top.




A
CL
r = 0.8
B
CL
r = 0.75
A
PL
r = 0.76
B
PL
r = 0.64
A
PRL
r = 0.7
B
PRL
r = 0.59
A
PN
r = 0.62
B
PN
r = 0.64
A
PW
r = 0.55
B
PW
r = 0.47
A
SW
r = 0.64
B
SW
r = 0.63
Prediction accuracy
Prediction method
BayesB
BL
RKHS
rrBLUP
Model
Predicted
Observed

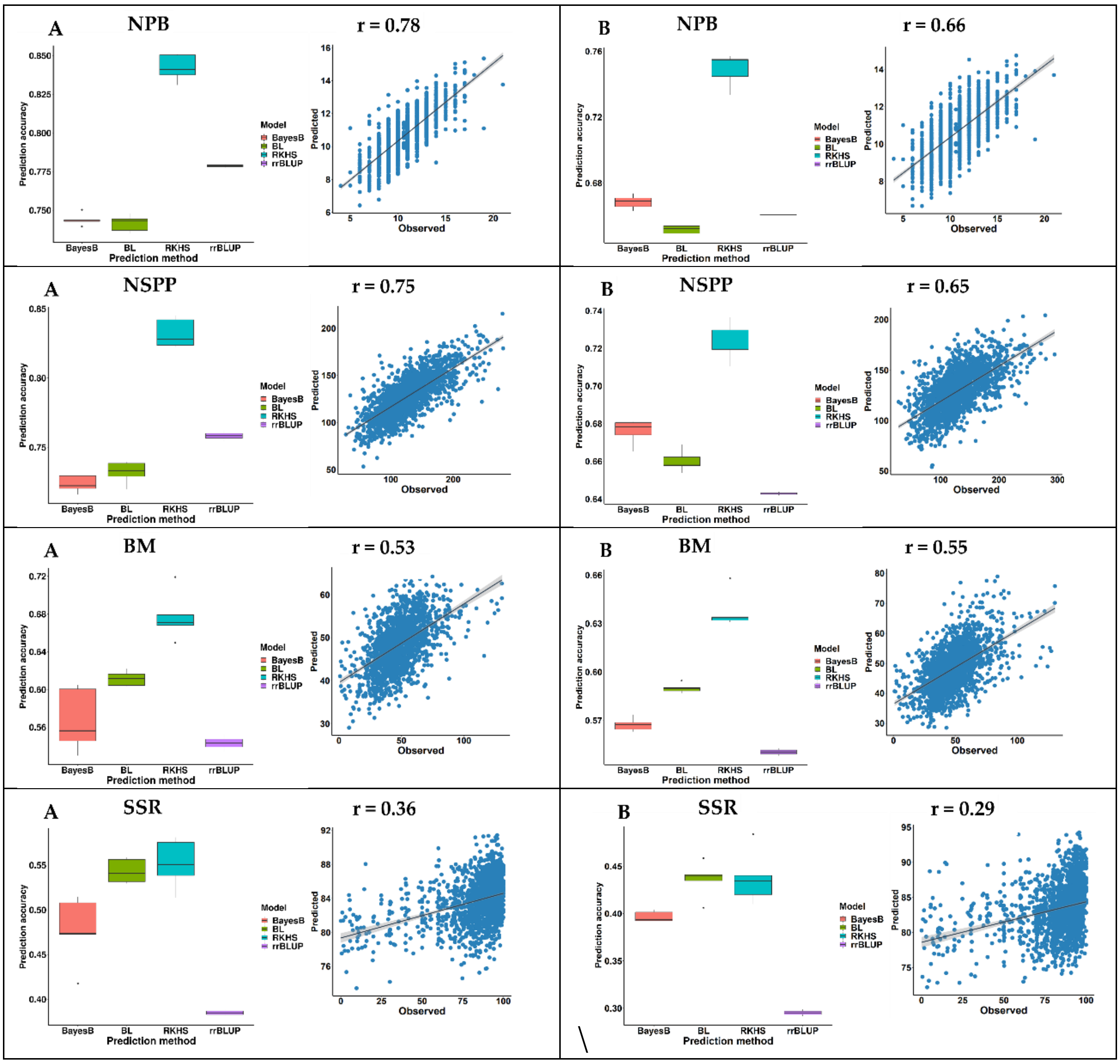
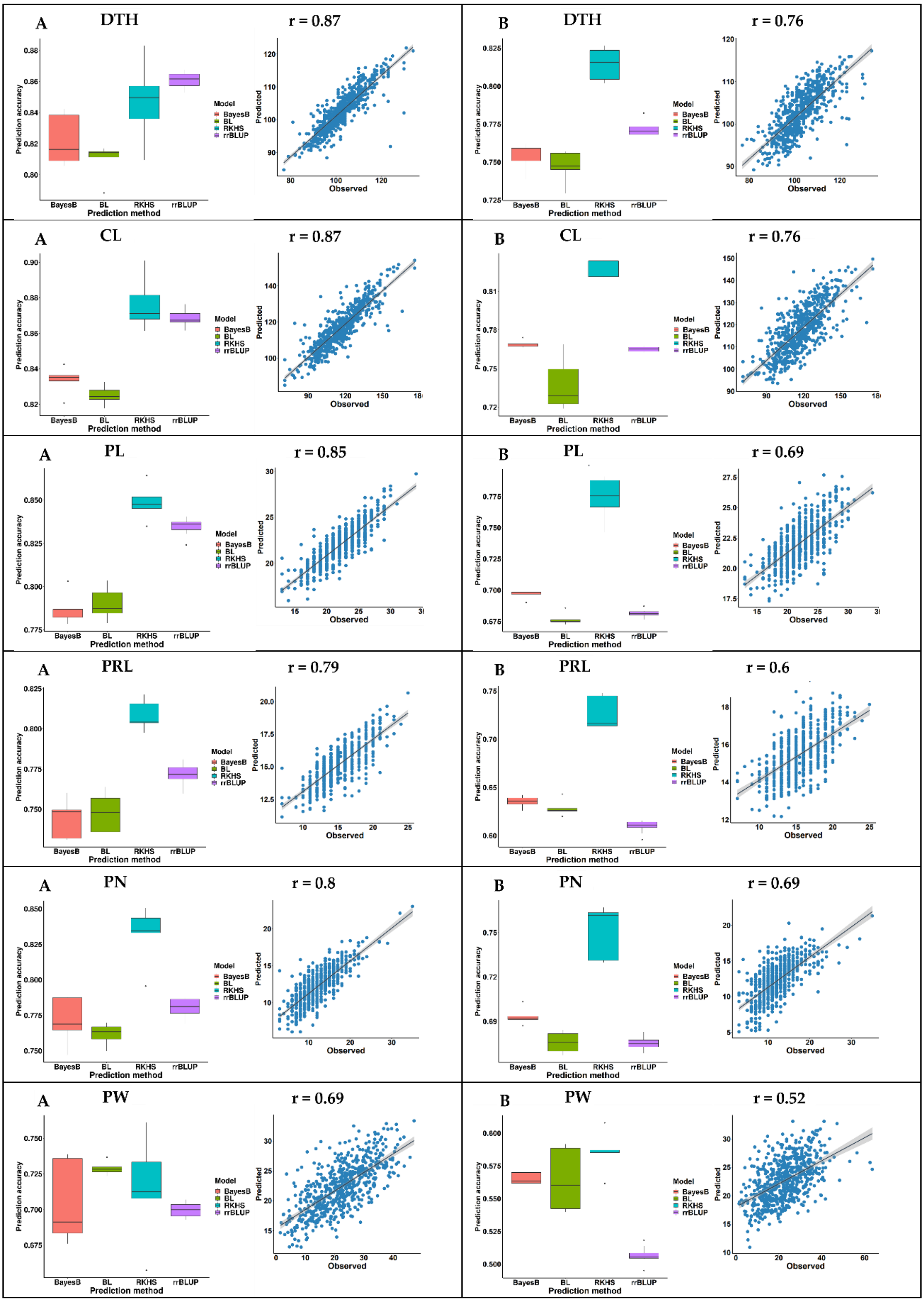


Figure 5-7. Genomic prediction for eleven traits in 2018.

(A) Marker set 1; 2006 GBS SNPs (B) Marker set 2; 78,154 projected SNPs. Prediction accuracy was obtained by Pearson correlation of the observed phenotypes against the predicted phenotypes using four models: Bayesian B (BayesB), Bayesian least absolute shrinkage and selection operator (BL), reproducing kernel Hilbert space regression (RKHS), and ridge regression best linear unbiased prediction (rrBLUP). The scatter plot on the right of each plot shows the distribution of observed phenotypes (X-axis) against predicted phenotypes(Y-axis) using rrBLUP with its correlation coefficient (r) shown on the top.


A
DTH
r = 0.87
B
DTH
r = 0.76
A
CL
r = 0.87
B
CL
r = 0.76
A
PL
r = 0.85
B
PL
r = 0.69
A
PRL
r = 0.79
B
PRL
r = 0.6
A
PN
r = 0.8
B
PN
r = 0.69
A
PW
r = 0.69
B
PW
r = 0.52
Prediction accuracy
Prediction method
BayesB
BL
RKHS
rrBLUP
Model
Predicted
Observed

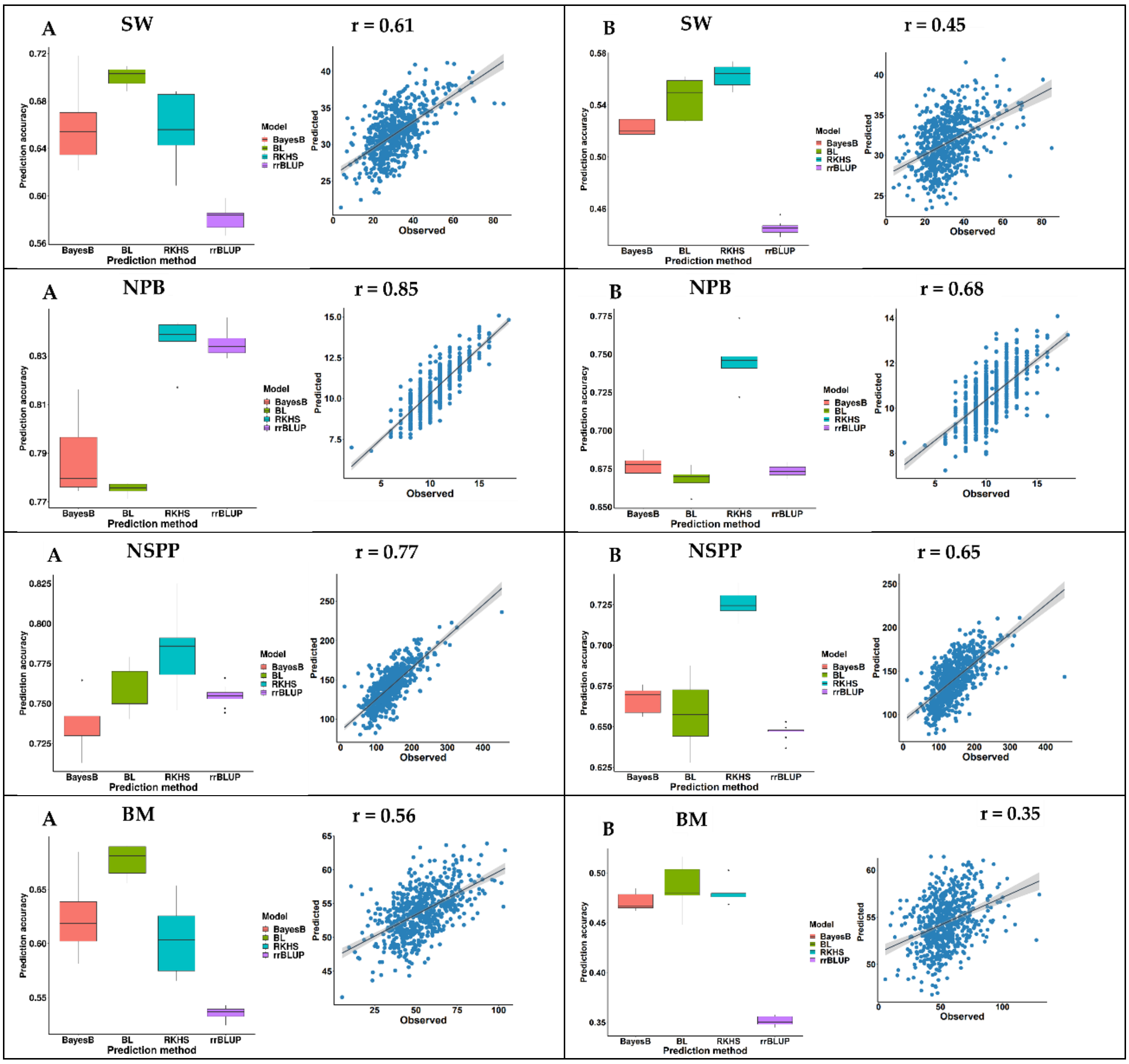


Figure 5-8. Genomic prediction for ten traits in 2015.

(A) Marker set 1; 2006 GBS SNPs (B) Marker set 2; 78,154 projected SNPs. Prediction accuracy was obtained by Pearson correlation of the observed phenotypes against the predicted phenotypes using four models: Bayesian B (BayesB), Bayesian least absolute shrinkage and selection operator (BL), reproducing kernel Hilbert space regression (RKHS), and ridge regression best linear unbiased prediction (rrBLUP). The scatter plot on the right of each plot shows the distribution of observed phenotypes (X-axis) against predicted phenotypes(Y-axis) using rrBLUP with its correlation coefficient (r) shown on the top.

Table 5-1. GWAS and candidate genes in 2019 using *aus*-NAM-II.
Only one top SNP per chromosome is shown. SNPs were considered significant by false discovery rate (FDR) set to 0.05 (5%).

Year	Marker set	Trait	SNP	CHR	bp	P-value	Candidate QTL/gene
2019	Set 1	DTH	S10_17200086	10	17,200,086	1.82E-26	*Ehd1*
			S06_3043159	6	3,043,159	1.60E-14	*RFT1*
			S07_15686122	7	15,686,122	4.83E-05	
			S04_25306132	4	25,306,132	0.00017	
	Set2		S06_3043159	6	3,043,159	2.05E-12	*RFT1*
			S10_17173439	10	17,173,439	6.34E-09	*Ehd1*
			S07_16134353	7	16,134,353	6.56E-09	
			S03_1825245	3	1,825,245	1.04E-05	*Hd9*
			S02_29724553	2	29,724,553	8.81E-05	*DTH2*

Table 5-2. GWAS and candidate genes in 2018 using *aus*-NAM-II. Only one top SNP per chromosome is shown. SNPs were considered significant by false discovery rate (FDR) set to 0.05 (5%).

Year	Marker set	Trait	SNP	CHR	bp	P-value	Candidate QTL/gene
2018	Set 1	DTH	S10_17200086	10	17,200,086	1.02E-31	*Ehd1*
			S06_3043159	6	3,043,159	2.11E-12	*RFT1*
			S07_15686122	7	15,686,122	2.14E-05	
			S04_25306132	4	25,306,132	5.26E-05	
			S01_40094852	1	40,094,852	0.00076	*OsMADS51*
	Set 2		S06_3043159	6	3,043,159	1.92E-12	*RFT1*
			S10_17179241	10	17,179,241	2.94E-11	*Ehd1*
			S07_10940592	7	10,940,592	2.56E-08	*Ghd7*
			S02_29724553	2	29,724,553	4.11E-05	*Hd7*
			S03_2313919	3	2,313,919	0.00036	*Se14*
	Set 1	CL	S01_38613725	1	38,613,725	2.92E-10	
			S05_24022772	5	24,022,772	3.50E-07	
			S10_16772764	10	16,772,764	3.70E-07	
			S06_25796053	6	25,796,053	0.0002	
	Set 2	CL	S05_25485871	5	25,485,871	1.48E-07	
			S02_33844696	2	33,844,696	5.39E-07	
			S08_2514430	8	2,514,430	1.13E-06	
			S09_17869824	9	17,869,824	3.10E-05	
			S03_21448627	3	21,448,627	3.85E-05	
	Set 1	PL	S04_24036340	4	24036340	1.04E-07	
			S06_28256890	6	28256890	3.04E-05	
			S11_11239016	11	11239016	7.98E-05	
			S02_32639480	2	32639480	0.000121	
			S03_19841279	3	19841279	0.000232	

Set 2	PL	S02_31613748	2	31613748	5.72E-14
		S04_24036340	4	24036340	1.33E-13
		S06_28558740	6	28558740	2.12E-07
Set 1	PRL	S04_24036340	4	24036340	4.06E-05
		S11_2186339	11	2186339	2.32E-05
Set 2	PRL	S04_24036340	4	24036340	1.45E-10
		S02_32660615	2	32660615	8.06E-08
		S05_1478272	5	1478272	1.54E-05
		S06_28256890	6	28256890	1.95E-05
Set 1	PN	S04_28029034	4	28029034	2.42E-05
Set 2	PN	S03_27742608	3	27742608	2.68E-08
		S04_29988729	4	29988729	3.93E-08
		S02_4533105	2	4533105	1.39E-06
		S12_25924704	12	25924704	1.41E-05
		S07_345435	7	345435	1.69E-05
		S01_25149006	1	25149006	3.86E-05
		S05_2763539	5	2763539	7.41E-06
Set 2	PW	S06_20720431	6	20720431	1.73E-08
Set 1	SW	S10_17200086	10	17200086	5.76E-08
		S05_25224751	5	25224751	1.51E-05
		S06_3043159	6	3043159	2.16E-05
Set 2	SW	S07_10940592	7	10940592	5.10E-06
		S09_6634648	9	6634648	9.29E-06
		S01_39971552	1	39971552	1.32E-05
		S06_3082602	6	3082602	1.82E-05
		S05_25485871	5	25485871	4.69E-05
		S10_16237006	10	16237006	6.01E-05
Set 1	NPB	S06_2279056	6	2279056	1.99E-05
		S01_28910188	1	28910188	2.00E-05

		S02_12973181	2	12973181	2.34E-05
		S04_28661829	4	28661829	0.000131
		S11_22953724	11	22953724	0.000205
		S06_2611918	6	2611918	0.000228
		S08_25135358	8	25135358	0.000397
		S12_1376568	12	1376568	0.00066
Set 2	NPB	S08_25425641	8	25425641	7.17E-09
		S11_23196408	11	23196408	8.22E-09
		S02_33164470	2	33164470	1.39E-07
Set 1	NSPP	S07_28878093	7	28878093	3.70E-08
		S04_28029034	4	28029034	1.09E-06
		S03_17911107	3	17911107	9.53E-06
Set 2	NSPP	S07_28878093	7	28878093	4.72E-23
		S11_23196408	11	23196408	3.30E-12
		S04_34909036	4	34909036	2.63E-07
		S08_25425641	8	25425641	2.14E-06
		S03_27784127	3	27784127	8.27E-06
Set 1	BM	S10_17973467	10	17973467	8.39E-06
Set 2	BM	S06_20720431	6	20720431	8.41E-08

Table 5-3. GWAS and candidate genes in 2015 using *aus*-NAM-II.
Only one top SNP per chromosome is shown. SNPs were considered significant by false discovery rate (FDR) set to 0.05 (5%).

Year	Marker set	Trait	SNP	CHR	bp	P-value	Candidate QTL/gene
2015	Set 1	DTH	S10_17367103	10	17,367,103	3.38E-14	*Ehd1*
			S06_3043159	6	3,043,159	2.61E-05	*RFT1*
			S07_14908092	7	14,908,092	0.00015	
	Set 2	DTH	S07_16134353	7	16,134,353	4.33E-11	
			S06_17269895	6	17,269,895	3.14E-08	
			S10_17173439	10	17,173,439	3.94E-06	*Ehd1*
			S03_29567496	3	29,567,496	0.00034	OsPIPK1
	Set 1	CL	S01_38216481	1	38216481	2.41E-07	
			S05_24022772	5	24022772	2.80E-05	
	Set 2	CL	S06_26004369	6	26004369	1.52E-06	
			S12_4149852	12	4149852	3.09E-05	
			S01_38216481	1	38216481	5.14E-05	
	Set 1	PL	S04_24036340	4	24036340	4.43E-06	
	Set 2	PL	S04_24036340	4	24036340	5.11E-11	
			S06_28256890	6	28256890	3.28E-05	
	Set 1	PRL	S08_27350299	8	27350299	6.42E-06	
	Set 2	PRL	S08_27802432	8	27802432	3.43E-07	
			S04_22392948	4	22392948	3.55E-06	
	Set 2	PN	S02_6620524	2	6620524	2.71E-06	
			S04_30514855	4	30514855	6.50E-06	
			S12_27269291	12	27269291	1.29E-05	
	Set 2	PW	S08_27802432	8	27802432	3.21E-05	
	Set 2	SW	S10_17175989	10	17175989	2.35E-05	
	Set 2	NPB	S02_33164470	2	33164470	2.02E-08	

Table 5-4. Prediction accuracy for traits in 2019 and 2018.

Prediction accuracy was deduced by Pearson correlation coefficients between predicted and observed traits. The coefficients were averaged from 5-fold cross validation in marker set 1 (2006 GBS SNPs) and marker set 2 (78,154 projected SNPs).

Year	Marker set	Trait	BayesB	BL	RKHS	rrBLUP
2019	Set 1	DTH	0.8064	0.8058	0.8758	0.8525
	set 2		0.7662	0.7584	0.846	0.7825
2018	Set 1	DTH	0.8052	0.8102	0.883	0.846
		CL	0.778	0.7784	0.87	0.809
		PL	0.7264	0.7304	0.8172	0.7535
		PRL	0.6634	0.6788	0.7654	0.6895
		PN	0.6314	0.652	0.7638	0.645
		PW	0.5696	0.611	0.6578	0.552
		SW	0.6352	0.6642	0.75	0.6515
		NPB	0.7438	0.7414	0.8448	0.779
		NSPP	0.7234	0.7318	0.8334	0.7585
		BM	0.5672	0.6116	0.6762	0.543
		SSR	0.4776	0.5432	0.5516	0.385
	set 2	DTH	0.7876	0.7774	0.8574	0.7955
		CL	0.7406	0.746	0.8114	0.748
		PL	0.6438	0.6312	0.7144	0.6275
		PRL	0.6056	0.606	0.693	0.5895
		PN	0.6468	0.643	0.6996	0.6345
		PW	0.5124	0.5292	0.5592	0.4715
		SW	0.6324	0.65	0.7128	0.627
		NPB	0.6686	0.6522	0.749	0.6605
		NSPP	0.6756	0.6602	0.7232	0.643
		BM	0.5672	0.59	0.6378	0.5505
		SSR	0.3966	0.4358	0.4378	0.2955

Table 5-5. Prediction accuracy for traits in 2015.

Prediction accuracy was deduced by the Pearson correlation coefficients between predicted and observed traits. The coefficients were averaged from 5-fold cross validation in marker set 1 (2006 GBS SNPs) and marker set 2 (78,154 projected SNPs).

Year	Marker set	Trait	BayesB	BL	RKHS	rrBLUP
2015	set 1	DTH	0.8222	0.809	0.8424	0.8608
		CL	0.8334	0.8252	0.876	0.8681
		PL	0.7874	0.7904	0.848	0.8349
		PRL	0.7442	0.7482	0.8054	0.7717
		PN	0.7712	0.7616	0.8294	0.7805
		PW	0.7052	0.729	0.7136	0.6996
		SW	0.6598	0.7004	0.6514	0.5806
		NPB	0.7888	0.7752	0.8372	0.835
		NSPP	0.6456	0.7168	0.6828	0.5951
		BM	0.6252	0.6764	0.6018	0.5354
	set 2	DTH	0.754	0.747	0.8146	0.7708
		CL	0.759	0.748	0.7966	0.7649
		PL	0.7568	0.752	0.8066	0.6813
		PRL	0.7572	0.7488	0.8146	0.6095
		PN	0.7614	0.7466	0.8116	0.6751
		PW	0.7486	0.7478	0.8062	0.5067
		SW	0.758	0.7544	0.7846	0.4449
		NPB	0.7594	0.7456	0.8164	0.6736
		NSPP	0.7584	0.7436	0.8042	0.5212
		BM	0.7572	0.7454	0.805	0.3513

Chapter 6 General Conclusions

Summary

An *aus*-NAM population was developed and reported in this dissertation. The population structure analysis of *aus*-NAM-II showed a weak stratification, with half-sib RILs dispersed around the T65 and *aus* diversity donors. This confirmed that the *aus*-NAM-II population successfully expanded genetic diversity and destroyed population structure, thus suitable for analyzing the genetic architecture of complex traits. QTL mapping using *aus*-NAM-I population discerned several known QTL and novel candidate QTL (Kitony *et al.*, 2021), this elucidated that our genetic mapping approaches and genetic material employed were useful. Moreover, the GWAS confirmed that genotype projections from dense parental variants publicly available such as WRC (Tanaka *et al.*, 2020) onto RILs genotyped with a low-cost GBS was a feasible alternative for increasing marker size.

In terms of prediction accuracy, this study verified that sample size, marker type/density, prediction model and the relationship between training population and testing population were important. Reproducing kernel Hilbert space regression (RKHS) method which accommodates markers with large and null effects at the same time capture epistasis was the most robust model. The genomic prediction results were encouraging for the implementation of genomic selection in a practical rice breeding program.

Unlike association panels (Huang *et al.*, 2010), *aus*-NAM combines useful traits derived from multiple donor lines (*aus*) into an elite rice variety (T65). Overall, the immortal nature of *aus*-NAM RILs allows evaluations in various environments. Thus, could be used to reveal QTL functions in detail.

Challenges, Limitations and Opportunities

The biggest challenge in this work like in any other large-scale genetic mapping population is the high-throughput genotyping and phenotyping required. However, with the advent of NGS, automation of laboratory work by robots, and computers are expected to mitigate the limitations thus genotyping will no longer be a problem. As a caveat, the large datasets generated by high-throughput sequencing can sometimes lead to computing issues, mostly in terms of computer memory/storage, and visualization, the use of intuitive desktop applications is practically not possible. Moreover, the varying data formats and tuning parameters required by different tools can be a challenge, however, basic knowledge of Linux scripts and R scripts can come in handy here.

On the other hand, digital tools (barcodes and tablet terminals) were utilized for phenotyping in this study. The devices successfully saved time and labor (unpublished). It should be noted though that time/labor for phenotyping is not expected to reduce anytime soon unless cutting-edge technologies are innovated such as computer vision and robotics. More efforts to fill the "genotype-phenotype gap"(Santini *et al.*, 2021) are necessary.

Future Perspectives

In most of the actual breeding programs, several breeding targets are determined before the commission of the project. DTH is usually included in the targets because it affects the adaptation of new varieties to environments and cropping systems. As presented in this study, DTH is one of the most predictable traits by genomic prediction (GP) models. Therefore, the author proposes that DTH would be the target for performing GP in actual breeding programs. For example, when using a scheme of population breeding (bulk method), thousands of F4 or F5 plants as the materials for starting selection can be generated. The application of GBS to these thousands of plants is realistic because of the advancement of GBS. If parts of these generations could be used for genomic predictions, plants with suitable DTH can be selected based on GP genomic estimated breeding values (GEBVs). This would eliminate most of the unnecessary plants with non-suitable heading time before planting in fields.

The immortal nature of the NAM RILs brings advantages for repeated and multi-locational testing. This will show whether the genetic analysis using a broad range of phenotypes can reveal the desired or undesired results from breeders` perspectives as suggested by QTL analysis/genomic predictions. Furthermore, multi-environment trials can be used to improve the robustness of the prediction models. Incorporating crop growth models and other omics such as *e*RD-GWAS (Lin et al., 2017) can as well be explored using NAM population in the future.

Acknowledgments

I would like to take this opportunity to express my sincere gratitude for the help I received during my research. First, to my advisor Dr. Kazuyuki Doi, for his mentorship. Similar thanks go to my committee members: Dr. Jun Murase, Dr. Motoyuki Ashikari, Dr. Yoshiaki Inukai, and Dr. Shunsaku Nishiuchi for their guidance and direct support throughout my study. I can’t thank enough Dr. Daigo Makihara for his support, he continuously checked on my progress and offered a word of encouragement.

I am also deeply indebted to the present members of our laboratory (Inoue, Kobayashi, Maria, Nakata, Amano, Kato, Abebaw, Kuroki, Takata, Matsuyama, Demizu, Shibata, Soe, and Vincent) and those who have graduated for their help and different learning experiences during my study. I cannot forget those who were involved in developing the *aus*-NAM genetic materials, thank you Dr. Sunohara and Yamasaki san.

For my health and well-being, I would like to appreciate Dr. Kikuta Mayumi Takenaga san and Dr. Kazuyuki Doi for their support during my relocation and stay in Nagoya. To my neighbor, Tomoe san, thank you very much for the concern and help especially during the cycling injury. To Maesaka san and all Togo farm staff members, thank you all for your kindness and support.

This dissertation and my dream to earn a Ph.D. would not have been possible if it were not for the scholarship I received from JICA. Thanks to JICA Agriculture Studies Networks for Food Security (Agri Net) program for funding my doctoral course. I am also grateful to KALRO Mwea scientists for their encouragement. Last but not least, I want to thank my family and dedicate this work to my parents, wife, and children who encouraged me throughout my studies. Your support and love enabled me to reach this level of academic achievement. To those who supported me in any aspect, I will always remember you in my heart, thank you.

Literature Cited

Araus, J.L., Cairns, J.E., 2014. Field high-throughput phenotyping: the new crop breeding frontier. Trends in plant science 19, 52-61.

Arends, D., Prins, P., Jansen, R.C., Broman, K.W., 2010. R/qtl: high-throughput multiple QTL mapping. Bioinformatics 26, 2990-2992.

Ashikari, M., Sakakibara, H., Lin, S., Yamamoto, T., Takashi, T., Nishimura, A., Angeles, E.R., Qian, Q., Kitano, H., Matsuoka, M., 2005. Cytokinin oxidase regulates rice grain production. Science 309.

Bajgain, P., Rouse, M.N., Tsilo, T.J., Macharia, G.K., Bhavani, S., Jin, Y., Anderson, J.A., 2016. Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. PloS one 11, e0155760.

Balint-Kurti, P.J., Yang, J., Van Esbroeck, G., Jung, J., Smith, M.E., 2010. Use of a maize advanced intercross line for mapping of QTL for northern leaf blight resistance and multiple disease resistance. Crop science 50, 458-466.

Bartholome, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., Bouffier, L., 2016. Performance of genomic prediction within and across generations in maritime pine. BMC genomics 17, 604.

Bernardo, R., 1996. Best linear unbiased prediction of maize single-cross performance. Crop science 36, 50-56.

Bernardo, R., Yu, J., 2007. Prospects for genomewide selection for quantitative traits in maize. Crop science 47, 1082-1090.

Bouchet, S., Olatoye, M.O., Marla, S.R., Perumal, R., Tesso, T., Yu, J., Tuinstra, M., Morris, G.P., 2017. Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. Genetics 206, 573–585.

Bradbury, P.J., Zhang Z Fau - Kroon, D.E., Kroon De Fau - Casstevens, T.M., Casstevens Tm Fau - Ramdoss, Y., Ramdoss Y Fau - Buckler, E.S., Buckler, E.S., 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics,2633-5.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., 2009a. The genetic architecture of maize flowering time. Science 325, 950–953.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., Goodman, M.M., Harjes, C., Guill, K., Kroon, D.E., Larsson, S., Lepak, N.K., Li, H., Mitchell, S.E., Pressoir, G., Peiffer, J.A., Oropeza, R.M., Rocheford, T.R., Cinta, R.M., Romero, S., Salvo, S., Villeda, H.S., Silva, H.S.d., Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., McMullen1, M.D., 2009b. The genetic architecture of maize flowering time. Science 325.

Campos, G.d.l., M. Hickey, J., Pong-Wong, R., D. Daetwyler, H., P. L. Calus, M., 2013. Whole genome regression and prediction methods applied to plant and animal breeding. Genetics 193, 327-345.

Cantelmo, N.F., Von Pinho, R.G., Balestre, M., 2017. Genome-wide prediction for maize single-cross hybrids using the GBLUP model and validation in different crop seasons. Molecular breeding, 37.

Chen, M., Presting, G., Barbazuk, W.B., Goicoechea, J.L., Blackmon, B., Fang, G., Kim, H., Frisch, D., Yu, Y., Sun, S., Higingbottom, S., Phimphilai, J., Phimphilai, D., Thurmond, S., Gaudette, B., Li, P., Liu, J., Hatfield, J., Main, D., Farrar, K., Henderson, C., Barnett, L., Costa, R., Williams, B., Walser, S., Atkins, M., Hall, C., Budiman, M.A., Tomkins, J.P., Luo, M., Bancroft, I., Salse, J., Regad, F., Mohapatra, T., Singh, N.K., Tyagi, A.K., Soderlund, C., Dean, R.A., Wing, R.A., 2002. An integrated physical and genetic map of the rice genome. Plant cell 14, 537-545.

Churchill, G.A., Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics, 963-971.

Cockram, J., Mackay, I., 2018. Genetic mapping populations for conducting High-resolution trait mapping in plants. Plant genetics and molecular biology, pp. 109-138.

Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O.A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A., Bentley, A.R., 2021. The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. Frontiers in plant science 12, 651480.

Dell'Acqua, M., Gatti, D.M., Pea, G., Cattonaro, F., Coppens, F., Magris, G., Hlaing, A.L., Aung, H.H., Nelissen, H., Baute, J., Frascaroli, E., Churchill, G.A., Inze, D., Morgante, M., Pe, M.E., 2015. Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. Genome biology 16, 167.

Dellaporta, S.L., Wood, J., Hicks, J.B., 1983. A plant DNA minipreparation: version II. Plant molecular biology reporter 1, 19-21.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43, 491-498.

Doebley, J.F., Gaut, B.S., Smith, B.D., 2006. The molecular genetics of crop domestication. Cell 127, 1309-1321.

Doi, K., Izawa, T., Fuse, T., Yamanouchi, U., Kubo, T., Shimatani, Z., Yano, M., Yoshimura, A., 2004. *Ehd1*, a B-type response regulator in rice, confers short-day promotion of flowering and controls *FT-like* gene expression independently of *Hd1*. Genes and development 18, 926-936.

Ebana, K., Kojima, Y., Fukuoka, S., Nagamine, T., Kawase, M., 2008. Development of mini core collection of japanese rice landrace. Breeding science 58, 281–291.

Endelman, J.B., 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. The plant genome 4, 250-255.

Federer, W.T., Crossa, J., 2012. I.4 Screening experimental designs for quantitative trait loci, association mapping, genotype-by environment interaction, and other investigations. Frontiers in physiology 3, 156.

Fitz Gerald, J.N., Carlson, A.L., Smith, E., Maloof, J.N., Weigel, D., Chory, J., Borevitz, J.O., Swanson, R.J., 2014. New arabidopsis advanced intercross recombinant inbred lines reveal female control of nonrandom mating. Plant physiology 165, 175-185.

Fragoso, C.A., Moreno, M., Wang, Z., Heffelfinger, C., Arbelaez, L.J., Aguirre, J.A., Franco, N., Romero, L.E., Labadie, K., Zhao, H., Dellaporta, S.L., Lorieux, M., 2017. Genetic architecture of a rice nested association mapping population. G3, 7.

Frouin, J., Labeyrie, A., Boisnard, A., Sacchi, G.A., Ahmadi, N., 2019. Genomic prediction offers the most effective marker assisted breeding approach for ability to prevent arsenic accumulation in rice grains. PloS one 14, e0217516.

Fujino, K., Hirayama, Y., Kaji, R., 2019. Marker-assisted selection in rice breeding programs in Hokkaido. Breed sci 69, 383-392.

Furuta, T., Ashikari, M., Jena, K.K., Doi, K., Reuscher, S., 2017. Adapting genotyping-by-sequencing for rice F2 populations. G3, 7, 881-893.

Gage, J.L., Monier, B., Giri, A., Buckler, E.S., 2020. Ten years of the maize nested association mapping population: impact, limitations, and future directions. Plant cell 32, 2083-2093.

Gamuyao, R., Chin, J.H., Pariasca-Tanaka, J., Pesaresi, P., Catausan, S., Dalid, C., Slamet-Loedin, I., Tecson-Mendoza, E.M., Wissuwa, M., Heuer, S., 2012. The protein kinase pstol1 from traditional rice confers tolerance of phosphorus deficiency. Nature 488, 535-539.

Garris, A.J., Tai, T.H., Coburn, J., Kresovich, S., McCouch, S., 2005. Genetic structure and diversity in Oryza sativa L. Genetics 169, 1631.

Gaudinier, A., Blackman, B.K., 2020. Evolutionary processes from the perspective of flowering time diversity. The new phytologist 225, 1883-1898.

Gireesh, C., Sundaram, R.M., Anantha, S.M., Pandey, M.K., Madhav, M.S., Rathod, S., Yathish, K.R., Senguttuvel, P., Kalyani, B.M., Ranjith, E., Venkata Subbarao, L., Kumar Mondal, T., Swamy, M., Rakshit, S., 2021. Nested association mapping (NAM) populations: present status and future prospects in the genomics era. Critical reviews in plant sciences 40, 49-67.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., Buckler, E.S., 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PloS one, 9.

Guo, B., Sleper, D.A., Beavis, W.D., 2010. Nested association mapping for identification of functional markers. Genetics 186, 373-383.

Guo, Z., Tucker, D.M., Basten, C.J., Gandhi, H., Ersoz, E., Guo, B., Xu, Z., Wang, D., Gay, G., 2014. The impact of population structure on genomic prediction in stratified populations.Theoretical and applied genetics,127, 749-762.

Haynes, W., 2013. Bonferroni correction. Encyclopedia of systems biology, 154-154.

Henderson, C.R., 1985. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. Journal of animal science 60, 111-117.

Heslot, N., Yang, H.-P., Sorrells, M.E., Jannink, J.-L., 2012. Genomic selection in plant breeding: A comparison of models. Crop science 52, 146-160.

Hill, W.G., 2013. Genetic correlation. Brenner's encyclopedia of genetics ,237-239.

Hu, J., Guo, C., Wang, B., Ye, J., Liu, M., Wu, Z., Xiao, Y., Zhang, Q., Li, H., King, G.J., Liu, K., 2018. Genetic properties of a nested association mapping population constructed with semi-winter and spring oilseed rapes. Frontiers in plant science 9, 1740.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., Li, M., Fan, D., Guo, Y., Wang, A., Wang, L., Deng, L., Li, W., Lu, Y., Weng, Q., Liu, K., Huang, T., Zhou, T., Jing, Y., Li, W., Lin, Z., Buckler, E.S., Qian, Q., Zhang, Q.F., Li, J., Han, B., 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nature genetics 42, 961-967.

Ikeda, M., Miura, K., Aya, K., Kitano, H., Matsuoka, M., 2013. Genes offering the potential for designing yield-related traits in rice. Current opinion in plant biology 16, 213-220.

Isidro, J., Jannink, J.L., Akdemir, D., Poland, J., Heslot, N., Sorrells, M.E., 2015. Training set optimization under population structure in genomic selection. Theoretical and applied genetics, 128, 145-158.

Izawa, T., 2007. Adaptation of flowering-time by natural and artificial selection in Arabidopsis and rice. Journal of experimental botany 58, 3091–3097.

Izawa, T., Oikawa, T., Sugiyama, N., Tanisaka, T., Yano, M., Shimamoto, K., 2002. Phytochrome mediates the external light signal to repress *FT* orthologs in photoperiodic flowering of rice. Genes development 16, 2006-2020.

Jonas, E., de Koning, D.J., 2013 Does genomic selection have a future in plant breeding? Trends biotechnology, 497-504.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., Childs, K.L., Davidson, R.M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S.S., Kim, J., Numa, H., Itoh, T., Buell, C.R., Matsumoto, T., 2013. Improvement of the Oryza sativa nipponbare reference genome using next generation sequence and optical map data. Rice 6, 4.

Kihupi, A., Angeles, E. & Khush, G. , 2001. Genetic analysis of resistance to bacterial blight, Xanthomonas oryzae pv. oryzae, in rice, Oryza sativa L. . Euphytica 117, 39–46.

Kitony, J.K., Sunohara, H., Tasaki, M., Mori, J.-I., Shimazu, A., Reyes, V.P., Yasui, H., Yamagata, Y., Yoshimura, A., Yamasaki, M., Nishiuchi, S., Doi, K., 2021. Development of an aus-derived nested association mapping (aus-NAM) population in rice. Plants 10,1255.

Kojima, S., Takahashi, Y., Kobayashi, Y., Monna, L., Sasaki, T., Araki, T., Yano, M., 2002. Hd3a, a rice ortholog of the Arabidopsis FT gene, promotes transition to flowering downstream of *Hd1* under short-day conditions. Plant cell physiol 43, 1096-1105.

Kojima, Y., Ebana, K., Fukuoka, S., Nagamine, T., Kawase, M., 2005. Development of an RFLP-based rice diversity research set of germplasm. Breeding science 55, 431-440.

Korte, A., Farlow, A.,2013. The advantages and limitations of trait analysis with GWAS: a review. Plant methods,9,29.

Kremling, K.A.G., Chen, S.Y., Su, M.H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F., Lorant, A., Bradbury, P.J., Buckler, E.S., 2018. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature 555, 520-523.

Ladejobi, O., Elderfield, J., Gardner, K.A., Gaynor, R.C., Hickey, J., Hibberd, J.M., Mackay, I.J., Bentley, A.R., 2016. Maximizing the potential of multi-parental crop populations. Applied and translational genomics 11, 9-17.

Lee, M., Sharopova, N., Beavis, W.D., Grant, D., Katt, M., Blair, D., Hallauer, A., 2002. Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. Plant molecular biology 48, 453–461.

Li, H., Bradbury, P., Ersoz, E., Buckler, E.S., Wang, J., 2011. Joint QTL linkage mapping for multiple-cross mating design sharing one common parent. PloS one 6, e17573.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., Liu, J., Warburton, M.L., Cheng, Y., Hao, X., Zhang, P., Zhao, J., Liu, Y., Wang, G., Li, J., Yan, J., 2013. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nature genetics 45, 43-50.

Li, H., Rasheed, A., Hickey, L.T., He, Z., 2018. Fast-forwarding genetic gain. Trends in plant science 23, 184-186.

Li, J., Bus, A., Spamer, V., Stich, B., 2016. Comparison of statistical models for nested association mapping in rapeseed (Brassica napus L.) through computer simulations. BMC plant biology 16, 26.

Lin, H.Y., Liu, Q., Li, X., Yang, J., Liu, S., Huang, Y., Scanlon, M.J., Nettleton, D., Schnable, P.S., 2017. Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. Genome biology 18, 192.

Liu, C., Ou, S., Mao, B., Tang, J., Wang, W., Wang, H., Cao, S., Schlappi, M.R., Zhao, B., Xiao, G., Wang, X., Chu, C., 2018. Early selection of bZIP73 facilitated adaptation of japonica rice to cold climates. Nature communications 9, 3302.

Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., Farkhari, M., Ribaut, J.M., Cao, M., Rong, T., Xu, Y., 2010. Joint linkage-linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. Proceedings of the national academy of sciences of the united states of America 107, 19585-19590.

Matsubara, K., Hori, K., Ogiso-Tanaka, E., Yano, M., 2014. Cloning of quantitative trait genes from rice reveals conservation and divergence of photoperiod flowering pathways in arabidopsis and rice. Frontiers in plant science 5, 193.

Matsubara, K., Yano, M., 2018. Genetic and molecular dissection of flowering time control in rice. In: Sasaki, T., Ashikari, M. (Eds.), Rice genomics, genetics and breeding, 177-190.

Maurer, A., Draba, V., Jiang, Y., Schnaithmann, F., Sharma, R., Schumann, E., Kilian, B., Reif, J.C., Pillen, K., 2015. Modelling the genetic architecture of flowering time control in barley through nested association mapping. BMC genomics 16, 290.

McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., 2009. Supporting online material for: genetic properties of the maize nested association mapping population. Science 325, 737–741.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Mogga, M., Sibiya, J., Shimelis, H., Lamo, J., Yao, N., 2018. Diversity analysis and genome-wide association studies of grain shape and eating quality traits in rice (Oryza sativa L.) using DArT markers. PloS one 13, e0198012.

Monteverde, E., Rosas, J.E., Blanco, P., Pérez de Vida, F., Bonnecarrère, V., Quero, G., Gutierrez, L., McCouch, S., 2018. Multienvironment models increase prediction accuracy of complex traits in advanced breeding lines of rice. Crop science 58, 1519-1530.

Moose, S.P., Mumm, R.H., 2008. Molecular plant breeding as the foundation for 21st century crop improvement. Plant physiology 147, 969-977.

Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z., Costich, D.E., Buckler, E.S., 2009. Association mapping: Critical considerations shift from genotyping to experimental design. The Plant cell 21, 2194–2202.

Nakano, Y., Kobayashi, Y., 2020. Genome-wide association studies of agronomic traits consisting of field and molecular-based phenotypes. Reviews in agricultural science 8, 28-45.

Nakaya, A., Isobe, S.N., 2012. Will genomic selection be a practical method for plant breeding? Annals of botany 110, 1303-1316.

Nice, L.M., Steffenson, B.J., Brown-Guedira, G.L., Akhunov, E.D., Liu, C., Kono, T.J., Morrell, P.L., Blake, T.K., Horsley, R.D., Smith, K.P., Muehlbauer, G.J., 2016. Development and genetic characterization of an advanced backcross-nested association mapping (AB-NAM) population of wild x cultivated barley. Genetics 203, 1453-1467.

Nordborg, M., Weigel, D., 2008. Next-generation genetics in plants. Nature 456, 720-723.

Norton, G.J., Travis, A.J., Douglas, A., Fairley, S., Alves, E.P., Ruang-Areerate, P., Naredo, M.E.B., McNally, K.L., Hossain, M., Islam, M.R., Price, A.H., 2018. Genome wide association mapping of grain and straw biomass traits in the rice Bengal and Assam Aus Panel (BAAP) grown under alternate wetting and drying and permanently flooded irrigation. Frontiers in plant science 9, 1223.

Ogura, T., Busch, W., 2015. From phenotypes to causal sequences: using genome wide association studies to dissect the sequence basis for variation of plant development. Current opinion in plant biology 23, 98-108.

Ogut, F., Bian, Y., Bradbury, P.J., Holland, J.B., 2015. Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. Heredity 114, 552-563.

Olukolu, B.A., Wang, G.F., Vontimitta, V., Venkata, B.P., Marla, S., Ji, J., Gachomo, E., Chu, K., Negeri, A., Benson, J., Nelson, R., Bradbury, P., Nielsen, D., Holland, J.B., Balint-Kurti, P.J., Johal, G., 2014. A genome-wide association study of the maize hypersensitive defense response identifies genes that cluster in related pathways. PLoS genetics 10, e1004562.

Onogi, A., Watanabe, M., Mochizuki, T., Hayashi, T., Nakagawa, H., Hasegawa, T., Iwata, H., 2016. Toward integration of genomic selection with crop modelling: the development of an integrated approach to predicting rice heading dates. Theoretical and applied genetics, 129, 805-817.

Pearson, K., 1895. Notes on regression and inheritance in the case of two parents. Proceedings of the royal society of London, 240–242.

Peiffer, J.A., Romay, M.C., Gore, M.A., Flint-Garcia, S.A., Zhang, Z., Millard, M.J., 2014. The genetic architecture of maize height. Genetics 196, 1337–1356.

Pérez, P., Campos, G.d.l., 2014. Genome-Wide Regression and Prediction with the BGLR statistical package. Genetics 198, 483–495.

Poland, J.A., Bradbury, P.J., Buckler, E.S., Nelson, R.J., 2011. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proceedings of the national academy of sciences of the united states of America 108, 6893-6898.

Poland, J.A., Brown, P.J., Sorrells, M.E., Jannink, J.-L., 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by- sequencing approach. PloS one, 7.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575.

R Core Team, 2020. R Core Team (2020). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. URL https://www.R-project.org/.

Ray, S., Satya, P., 2014. Next generation sequencing technologies for next generation plant breeding. Frontiers in plant science 5, 367.

Reyes, V.P., Angeles-Shim, R.B., Mendioro, M.S., Manuel, M.C.C., Lapis, R.S., Shim, J., Sunohara, H., Nishiuchi, S., Kikuta, M., Makihara, D., Jena, K.K., Ashikari, M., Doi, K., 2021. Marker-assisted introgression and stacking of major QTLs controlling grain number (Gn1a) and number of primary branching (WFP) to NERICA cultivars. Plants 10, 844.

Saade, S., Maurer, A., Shahid, M., Oakey, H., Schmockel, S.M., Negrao, S., Pillen, K., Tester, M., 2016. Yield-related salinity tolerance traits identified in a nested association mapping (NAM) population of wild barley. Scientific reports 6, 32586.

Santini, F., Kefauver, S.C., Araus, J.L., Resco de Dios, V., Martin Garcia, S., Grivet, D., Voltas, J., 2021. Bridging the genotype-phenotype gap for a Mediterranean pine by semi-automatic crown identification and multispectral imagery. The new phytologist 229, 245-258.

Searle, S.R., Casella, G., McCulloch, C.E., 2006. Variance components. John Wiley & sons, NJ.

Si, L., Chen, J., Huang, X., Gong, H., Luo, J., Hou, Q., Zhou, T., Lu, T., Zhu, J., Shangguan, Y., Chen, E., Gong, C., Zhao, Q., Jing, Y., Zhao, Y., Li, Y., Cui, L., Fan, D., Lu, Y., Weng, Q., Wang, Y., Zhan, Q., Liu, K., Wei, X., An, K., An, G., Han, B., 2016. OsSPL13 controls grain size in cultivated rice. Nature genetics 48, 447-456.

Singh, B.D., Singh, A.K., 2015. Mapping Populations. Springer, New Delhi.

Song, Q., Yan, L., Quigley, C., Jordan, B.D., Fickus, E., Schroeder, S., Song, B.H., Charles An, Y.Q., Hyten, D., Nelson, R., Rainey, K., Beavis, W.D., Specht, J., Diers, B., Cregan, P., 2017. Genetic characterization of the soybean nested association mapping population. The plant genome 10.

Spindel, J., Begum, H., Deniz , A., Parminder , V., Bertrand, C., Edilberto , R., Gary , A., Jean-Luc, J., McCouch, S.R., 2015. Genomic selection and association mapping in rice (Oryza Sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS genetics,11.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., Selbig, J., 2007. PCAmethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics 23, 1164-1167.

Stadlmeier, M., Hartl, L., Mohler, V., 2018. Usefulness of a multiparent advanced generation intercross population with a greatly reduced mating design for genetic studies in winter wheat. Frontiers in plant science 9, 1825.

Sun, C., Hu, Z., Zheng, T., Lu, K., Zhao, Y., Wang, W., Shi, J., Wang, C., Lu, J., Zhang, D., Li, Z., Wei, C., 2017. RPAN: rice pan-genome browser for ~3000 rice genomes. Nucleic acids res 45, 597-605.

Swarts, K., Li, H., Romero Navarro, J.A., An, D., Romay, M.C., Hearne, S., Acharya, C., Glaubitz, J.C., Mitchell, S., Elshire, R.J., Buckler, E.S., Bradbury, P.J., 2014. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. The plant genome, 7.

Takehisa, H., Yasuda, M., Fukuta, Y., Kobayashi, N., Hayashi, N., Nakashita, H., Abe, T., Sato, T., 2009. Genetic analysis of resistance genes in an indica-type rice (Oryza sativa L.), Kasalath, using DNA markers. Breeding science 59, 253–260.

Tanaka, N., Shenton, M., Kawahara, Y., Kumagai, M., Sakai, H., Kanamori, H., Yonemaru, J., Fukuoka, S., Sugimoto, K., Ishimoto, M., Wu, J., Ebana, K., 2020. Whole-genome sequencing of the NARO world rice core collection (WRC) as the basis for diversity and association studies. Plant cell physiol 61, 922-932.

The 3000 rice genomes, p., 2014. The 3,000 rice genomes project. Gigascience, 3:7.

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B., Buckler, E.S., 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature genetics 43, 159-162.

Toda, Y., Wakatsuki, H., Aoike, T., Kajiya-Kanegae, H., Yamasaki, M., Yoshioka, T., Ebana, K., Hayashi, T., Nakagawa, H., Hasegawa, T., Iwata, H., 2020. Predicting biomass of rice with intermediate traits: Modeling method combining crop growth models and genomic prediction models. PloS one 15, e0233951.

Travis, A.J., Norton, G.J., Datta, S., Sarma, R., Dasgupta, T., Savio, F.L., Macaulay, M., Hedley, P.E., McNally, K.L., Sumon, M.H., Islam, M.R., Price, A.H., 2015. Assessing the genetic diversity of rice originating from Bangladesh, Assam and west Bengal. Rice ,8 , 35.

Turner, S.D., 2018. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Journal of open source software 3.

Utz, H.F., Melchinger, A.E., Schon, C.C., 2000. Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. Genetics 154, 1839–1849.

Valluru, R., Gazave, E.E., Fernandes, S.B., Ferguson, J.N., Lozano, R., Hirannaiah, P., Zuo, T., Brown, P.J., Leakey, A.D.B., Gore, M.A., Buckler, E.S., Bandillo, N., 2019. Deleterious mutation burden and its association with complex traits in sorghum (sorghum bicolor). Genetics 211, 1075.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91, 4414-4423.

Wambugu, P.W., Brozynska, M., Furtado, A., Waters, D.L., Henry, R.J., 2015. Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. Scientific reports 5, 13957.

Wang, C., Rutledge, J., Gianola, D., 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genetics selection evolution, 91-115.

Wang, X., Xu, Y., Hu, Z., Xu, C., 2018. Genomic selection methods for crop improvement: Current status and prospects. The crop journal 6, 330-340.

Wei, T., Simko, V., 2017. R package "corrplot": Visualization of a correlation matrix. https://github.com/taiyun/corrplot.

Weir, B.S., Hill, W.G., 1980. Effect of mating structure on variation in linkage disequilibrium. Genetics 95, 477.

Xavier, A., Muir, W.M., Rainey, K.M., 2016. Assessing predictive properties of genome-wide selection in soybeans. G3, 6.

Xiao, Y., Liu, H., Wu, L., Warburton, M., Yan, J., 2017. Genome-wide association studies in maize: praise and stargaze. Molecular plant 10, 359-374.

Xiao, Y., Tong, H., Yang, X., Xu, S., Pan, Q., Qiao, F., Raihan, M.S., Luo, Y., Liu, H., Zhang, X., Yang, N., Wang, X., Deng, M., Jin, M., Zhao, L., Luo, X., Zhou, Y., Li, X., Liu, J., Zhan, W., Liu, N., Wang, H., Chen, G., Cai, Y., Xu, G., Wang, W., Zheng, D., Yan, J., 2016. Genome-wide dissection of the maize ear genetic architecture using multiple populations. The New phytologist 210, 1095-1106.

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., Ismail, A.M., Bailey-Serres, J., Ronald, P.C., Mackill, D.J., 2006. Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442, 705-708.

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., Prasanna, B.M., Olsen, M.S., Wang, G., Zhang, A., 2020. Enhancing genetic gain through genomic selection: from livestock to plants. Plant communications 1, 100005.

Xu, Y., Wang, X., Ding, X., Zheng, X., Yang, Z., Xu, C., Hu, Z., 2018. Genomic selection of agronomic traits in hybrid rice using an NCII population. Rice 11, 32.

Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L., Zhou, H., Yu, S., Xu, C., Li, X., Zhang, Q., 2008. Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. Nature genetics 40, 761-767.

Yamamoto, E., Yonemaru, J.-i., Yamamoto, T., Yano, M., 2012. OGRO: the overview of functionally characterized genes in rice online database. Rice 5, 26.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P.C., Hu, L., Yamasaki, M., Yoshida, S., Kitano, H., Hirano, K., Matsuoka, M., 2016. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nature genetics 48, 927-934.

Yano, M., Katayose, Y., Ashikari, M., Yamanouchi, U., Monna, L., Fuse, T., Baba, T., Yamamoto, K., Umehara, Y., Nagamura, Y., Sasaki, T., 2000. *Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the arabidopsis flowering time gene CONSTANS. The Plant cell 12, 2473-2483.

Yano, M., Sasaki, T., 1997. Genetic and molecular dissection of quantitative traits in rice. Plant molecular biology 35, 145–153.

Yonemaru, J.-i., Yamamoto, T., Fukuoka, S., Uga, Y., Hori, K., Yano, M., 2010. Q-TARO: QTL annotation rice online database. Rice 3, 194-203.

Yu, J., Holland, J.B., McMullen, M.D., Buckler, E.S., 2008. Genetic design and statistical power of nested association mapping in maize. Genetics 178, 539-551.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S., 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38, 203-208.

Zhang, H., Yin, L., Wang, M., Yuan, X., Liu, X., 2019. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Frontiers in genetics 10, 189.

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., Li, J., Simianer, H., 2014. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PloS one, 9, e93017.

Zhou, X., Huang, X., 2019. Genome-wide association studies in rice: How to solve the low power problems? molecular plant 12, 10-12.

Zhu, C., Gore, M., Buckler, E.S., Yu, J., 2008. Status and prospects of association mapping in plants. The plant genome 1, 5-20.

Zhu, Y.J., Fan, Y.Y., Wang, K., Huang, D.R., Liu, W.Z., Ying, J.Z., Zhuang, J.Y., 2017. Rice flowering locus T 1 plays an important role in heading date influencing yield traits in rice. Scientific reports 7, 4918.

List of Publications

1. Kitony, J.K., H. Sunohara, M. Tasaki, J.-I. Mori, A. Shimazu, V.P. Reyes, H. Yasui, Y. Yamagata, A. Yoshimura, M. Yamasaki, et al. (2021) Development of an *Aus*-Derived Nested Association Mapping (*Aus*-NAM) Population in Rice. Plants 2021, 10(6), 1255; https://doi.org/10.3390/plants10061255.
2. Kitony JK, Reyes VP, Sunohara H, Tasaki M, Yamasaki M, Nishiuchi S and Doi K. (2021). Genome-wide association study and genomic prediction using *aus*-NAM population in rice. Breed. Sci. (in preparation)