Application of genomic tools for Irish pasture improvement

Sai Krishna Arojju

September 2017

Thesis

submitted in partially fulfilment of the requirements for the Doctor of Philosophy at Trinity College Dublin

Supervisors:

Prof. Trevor R. Hodkinson
Head of Department, Department of Botany, School of Natural Sciences
Trinity College Dublin

Dr. Susanne Barth
Senior Research Officer, Crops Research Centre, Oak Park, Carlow
Teagasc

Dr. Dan Milbourne
Senior Research Officer, Crops Research Centre, Oak Park, Carlow
Teagasc

Co-supervisor:

Dr. Stephen L. Byrne
Research Officer, Crops Research Centre, Oak Park, Carlow
Teagasc

Other members:

Dr. Patrick Conaghan, Grassland Science Research Department, Teagasc
Prof. Michael Casler, Department of Agronomy, University of Wisconsin-Madison

This research was conducted under the auspices of School of Natural Sciences Trinity College Dublin

Declaration

I hereby declare that I am the sole author of this dissertation. I also declare that the work presented in this thesis has not been previously submitted in part or whole for any degree to any other university or college. I agree that the library of Trinity College Dublin can lend or copy this thesis on request.

Sai Krishna Arojju

September 2017

Acknowledgements

The work is carried out from October 2013 to September 2017 at Crop Science Department, Teagasc, Oak Park, Carlow. This research project was funded through a DAFM project (RSF 14/S/819) and Teagasc core funding. Part of the scholarship was funded through FEMS postgraduate research scholarship (TCD). This work wouldn't have been possible with the support of my many people, to whom I would like to take a chance to express my appreciation.

First, I would like to thank my supervisors Trevor Hodkinson, Susanne Barth and Dan Milbourne for selecting me for this position and introducing to the world of forage breeding. I really appreciate the support and encouragement throughout the project and it was a pleasure working with you all. I would also like to thank my co-supervisor Stephen Byrne for the discussion and sharing knowledge about the genomic prediction, for his patience and tolerating my knocks at his office door almost everyday in the week. I am very grateful to have him as a mentor for my Ph.D study. Besides my advisors, I would like to thank Patrick Conaghan and Micheal Casler for their valuable comments and suggestions to improve this thesis.

I would like to thank Olivia, Mary, Mick, Jean and Micheal from forage breeding team for there help in phenotyping. I would also like to say thank you to all my colleagues in Crop Science Department for there cooperation and support. A special thank you to John, Dheeraj, Riddhi, Annie and Rose for always supporting me during the difficult times. A big thank you to all my fellow postgraduates and postdocs at Teagasc for the help.

Finally my parents and brother, who worked hard all their life and in doing so provided me a bright future, I will always be grateful for your support.

Summary

Plant breeding is the art of utilizing genetic diversity and selecting the best genotypes using the most efficient methods. The conventional way to improve populations in perennial ryegrass (*Lolium perenne*), the most important forage grass in Ireland, is through recurrent selection. However, despite breeding for nearly a century, the rate of genetic improvement in perennial ryegrass is still in its infancy compared to cereals. With a need to accelerate genetic progress, marker assisted recurrent selection quickly became a promising approach. Initial mapping for quantitative trait loci (QTL) was done in a bi-parental population (linkage mapping), but most of the success was limited to identifying QTLs, with little success in practically using these QTLs in the breeding programme. This was mainly due to inconsistent QTLs, mapping in bi-parental populations, and significant markers failing to explain a large proportion of phenotypic variance. To overcome these limitations genome wide association studies (GWAS) that take account of historical recombination were proposed. This reduced the limitations of mapping in bi-parental populations and enabled QTL to be identified directly in breeding populations; however, GWAS was ineffective for complex traits controlled by many loci with relatively small effect. The next advancement in using markers in breeding came with genomic prediction. In genomic prediction all markers are simultaneously used to estimate allelic effects and generate breeding values. This thesis investigates the use of molecular markers and genomic information to accelerate genetic gains for key traits in the forage perennial ryegrass. Results are described and discussed in three chapters, with chapter two focusing on GWAS and chapter three and four focusing on genomic prediction methodology.

In chapter two a panel consisting of eight full-sib families was used in a GWAS. Full-sib families from a single breeding programme were used to ensure that the alleles were present at the frequency needed to have sufficient statistical power to identify associations. However, GWAS using mixed model failed to identify variants associated with heading and aftermath heading after correcting for multiple testing and population structure. This was put down to low levels of linkage disequilibrium (LD) in the population and the correlation of the trait (heading date) with population structure. However, marker-trait associations were identified within each family using single marker analysis, which takes advantage of linkage present within families. Many of these identified markers were proximal to genes controlling heading date.

Chapter three is a study to predict crown rust resistance using genome-wide markers in a large perennial ryegrass population. In this study the ability of markers to predict crown rust resistance (*Puccinia coronata* f. sp *lolli*) was evaluated using two prediction models and various factors influencing predictive ability were investigated. Predictive ability for crown rust resistance was high, but largely resulted from markers capturing genetic relationships. Using GWAS a small panel of markers were identified that were able to achieve higher predictive ability than the same number of randomly selected marker. Higher predictive ability over random markers indicates they are in LD with QTL for crown rust resistance rather than simply capturing relationship among families. This is relevant because accuracy due to genetic relationships will decay rapidly over generations whereas accuracy due to LD will persist, which is advantageous for practical breeding applications.

In chapter four a restricted population was utilised to develop genomic prediction equations for forage yield in tetraploid perennial ryegrass. Half-sib families were evaluated for yield in both simulated grazing management and conservation management over two years and maternal parents were genotyped using a genotyping-by-sequencing strategy. Genomic predictive abilities for traits ranged from 0.03 for summer yield to 0.30 for spring yield. Genomic prediction for both yield under grazing (calculated as economic value of a plot) and yield under first cut silage was promising. In particular predictive abilities of 0.22 were obtained for both first-cut silage and the economic value of a plot. Based on these results and the fact that we can complete multiple cycles of indirect selection with DNA markers relative to conventional genotypic selection means we can potentially more than double the rate of genetic gain using genomic prediction.

In summary the potential for genomic prediction to reduce the length of time it takes to complete a single cycle of selection from six years to one year makes it particularly attractive for forage breeding. This will lead to increases in the rate of genetic gain for economically important traits. Our results for both crown rust resistance and forage yield were promising and on the back of this genomic prediction is now being implemented in the Teagasc tetraploid forage breeding programme.

Contents

Chapter 1

General introduction

1.1 Perennial ryegrass for agriculture

Perennial ryegrass (*Lolium perenne* L.) is an economically important forage species in northwest Europe and around temperate regions of the world [83]. It has many desirable characteristics for a forage species, such as high water soluble carbohydrate values, crude protein content and dry matter yield, making it a highly valued forage species [37]. Because of its perennial nature, it has long growing seasons often providing more green feed compared to annual forage species, and its nutritive values are higher, making it cheaper and a better feed [164, 181]. It grows well on rich loam and clay soils, but growth rates are slow in acid rich soils and under drought conditions [7, 181]. Perennial ryegrass belongs to the family Poaceae, sub-family Pooideae and naturally occurs as a diploid species (2x) with seven pairs of chromosomes, but tetraploid (4x) varieties have also been created by doubling the chromosomes using colchicine treatment [131].

Perennial ryegrass is an out crossing species with a high degree of self-incompatibility (SI). The gametophytic SI system is found in perennial ryegrass and is controlled by independent multiallelic loci called S and Z loci [40]. This two locus system is found in other major forage species such as *Lolium multiflorum*, *Festuca pratensis* and *Dactylis aschersoniana* and is reviewed by [11]. Because of SI, modern crop improvement techniques such as hybrid breeding are not popular in perennial ryegrass and recurrent selection has remained the method of choice for population improvement. Breeding for diploid species dates back to the 1920s and the use of tetraploids for pasture began in the 1960s [83]. In general, tetraploid varieties tend to have larger leaves and tillers with higher water soluble carbohydrate content compared to diploids. Tetraploids have high yields with upright growth habit, making them suitable for grazing, but they have low sward density, dry matter and persistency compared to diploids. Diploid varieties have a high level of tolerance to biotic and abiotic stress and are well suited for different management regimes. To take advantage of traits from both ploidies, mixtures of diploid and tetraploid ryegrass seeds are often used for pasture establishment [83].

1.2 Traits for improvement

Forage breeding is relatively recent and there is a substantial amount of genetic variation present within forages such as ryegrass, offering great potential for trait improvement [28]. Traits considered to be improved are mainly driven by market needs and often depend on the location varieties are breed for. Traits such as seasonal and total dry matter yield, tolerance to diseases and nutritive value of the species are considered important all over the world [186]. In the beginning of forage breeding, the main focus was given to improving yield and persistency of the species [187], but later on attention was given towards others traits such as heading and aftermath heading, biotic and abiotic stress tolerance and seed yield [83]. However, increasing the combined number of selected traits in breeding programmes generally doesn't improve the genetic gain for each individual trait [34]. Hence, breeders always choose a few uncorrelated traits which are most economically important for breeding. A brief overview of the key traits for perennial ryegrass are discussed below.

Heading date is one of the most economically important traits for perennial ryegrass. Heading is considered as a time point at which the flowering structure (the spike in *Lolium*) emerges from the leaf sheath. When heading occurs, digestibility of forage decreases, as a result of lignin and cellulose deposition in the stems. The leaf to stem ratio is also significantly reduced, which in turn effects nutritive value and persistency of the sward [90]. Based on heading date, perennial ryegrass populations can be classified into early, intermediate and late varieties. Most breeding programmes select for intermediate and late heading varieties to extend vegetative growth and avoid aftermath heading which is highly associated with early heading varieties [192]. Aftermath heading is repeated heading which occurs later in the growing season after primary heading. It can decrease persistency and nutritive value of the swards [192]. Dry matter digestibility, water soluble carbohydrates and crude protein content are the principal measures of nutritive value of forage. Increasing nutritive value increases ruminant digestion and total energy available, which improves animal performance [164]. Disease resistance is another important trait for improvement, because disease infestation can reduce palatability, yield and nutritive value of forage [100, 106]. One of the major diseases in perennial ryegrass is crown rust, caused by *Puccinia coronata* f. sp *lolli* [140]. Severe crown rust infection in susceptible varieties can negatively influence yield and lead to a poor quality forage. Loss in dry matter yield up to 56% has been reported in susceptible varieties [141]. Direct impact assessment of crown rust infection on animal performance is difficult, but reports suggest that low dry matter yield and poor quality forage which is highly associated with crown rust infection as a negative effect on animal performance [26, 100, 164]. Although disease pressure in Ireland is moderate, continuous monitoring and selecting for disease resistance is essential in the breeding programme. If the disease is not monitored continuously there are high chances of losing resistances to one or more pathogens [34, 187].

Dry matter yield per unit of nitrogen input is the most important trait for perennial ryegrass and is measured in every breeding cycle, as well as during variety evaluation trials [34, 83]. Yield measurements are critical and generally undertaken later in the breeding cycle due to low correlation between spaced plants and swards [80]. Yields are always measured in swards and seasonal yields are more important than annual yield, as the economic value of grass will change during the growing season [49]. Fresh matter and dry matter yield are highly correlated, so indirect selection for fresh matter yield can significantly improve dry matter yield in the population [35].

1.3 Traditional forage breeding

In the beginning of forage breeding, natural variation in ecotypes was used to improve varieties. But to improve breeding pools by introducing breeding material from another programmes or another hemisphere, existing accessions were crossed. These varieties made a huge improvement in yield and persistency [83, 187]. Important traits in perennial ryegrass are quantitative in nature and are controlled by many genes with small effect, so the effective way for population improvement is by recurrent selection [16, 34, 187]. Recurrent selection is a breeding method used to increase the frequency of favourable alleles in the population by repeated selection of best performing plants. Each cycle of selection is completed when a new population is formed from crossing best plants from an existing population. Multiple cycles of selection are needed to improve the overall population. The aim of recurrent selection is to complete a cycle of selection as early as possible. This can be achieved by implementing two methods (i) phenotypic recurrent selection (ii) genotypic recurrent selection [37]. The best method is usually determined based on the traits considered to improve.

1.3.1 Phenotypic recurrent selection

In phenotypic recurrent selection superior plants are selected based on the phenotypic value of individual plants. Phenotypic recurrent selection begins with the establishment of populations as spaced plants and evaluation of individual plants for the trait of interest (Figure 1.1). Best performing plants are selected based on the phenotypic values and are recombined to create a new population for the next cycle(s) of selection. Crossing can be done either by (i) open pollination of selected plants and bulking the harvest seeds (producing half-sibs) or (ii) pair-cross selected plants and bulking of the seeds in isolations (producing full-sibs) (Figure 1.1). Using full-sibs as parents for the next cycle of selection doubles the theoretical genetic gain, but also increases time and cost associated with each cycle compared to half-sibs [34]. Phenotypic recurrent selection is the simplest and shortest breeding system and is often useful to improve traits which have high correlation between spaced plants and swards such as heading date, disease resistance and quality. Generally, phenotypic recurrent selection is based on data from unreplicated trials and single environments, so traits with higher heritability can be improved using phenotypic recurrent selection [34]. But dry matter yield and persistency has poor correlation between spaced plants and swards and would require replicated, multi year data to make selection decisions [35]. Hence, phenotypic recurrent selection is not suitable for improving yields, which is by far the most economically important trait.

1.3.2 Genotypic recurrent selection

Genotypic recurrent selection assigns genetic merit for individual plants based on the performance of progeny. Evaluation can be carried out either as half-sib or full-sib families. Based on phenotypic information, best performing plants from the spaced plant nursery are selected and crossed to produce full-sib and half-sib seeds. Full-sib seeds are produced by pair-crossing selected plants and bulking seeds from each pair-cross in isolation. Half-sib seeds are produced by polycrossing selected plants in isolation in multiple replicates. Seeds from matching maternal parents are

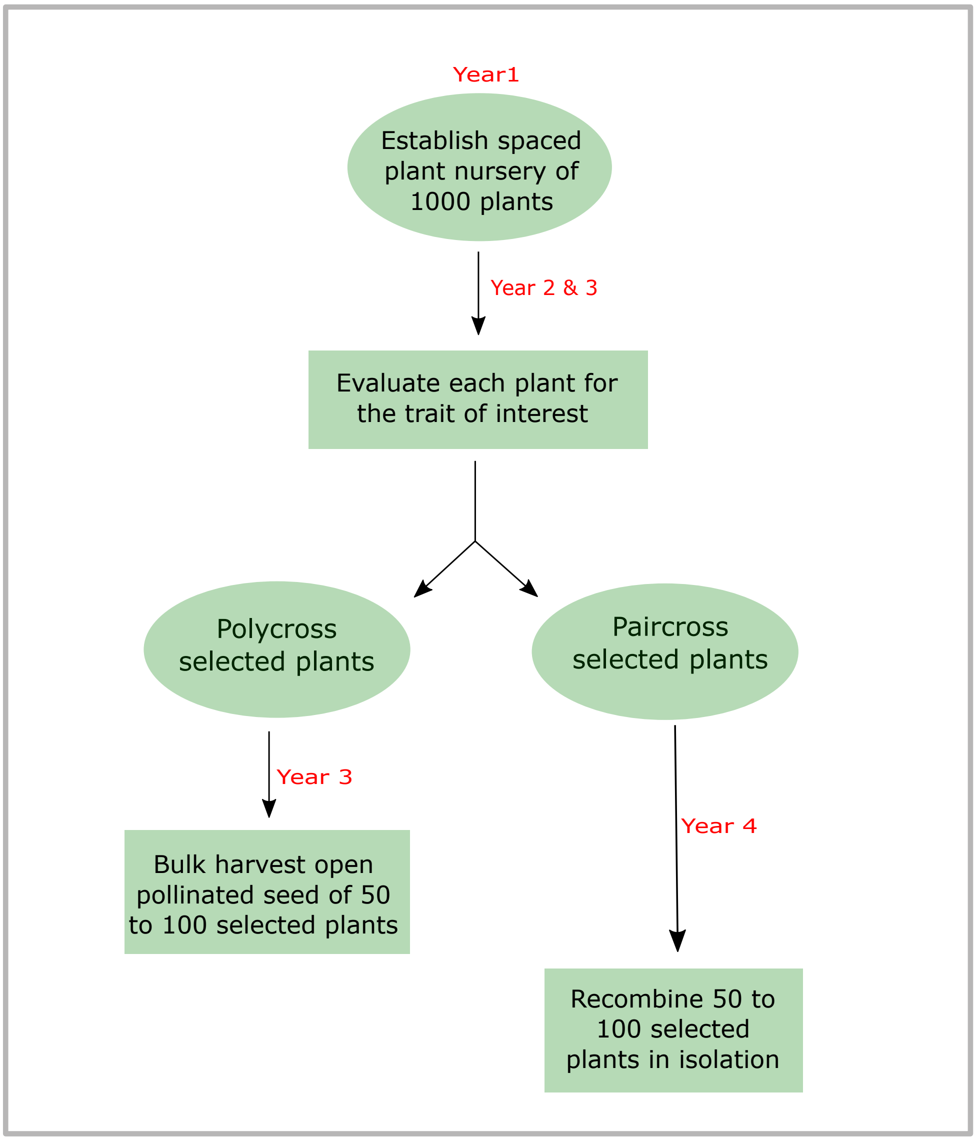
Year1
Establish spaced plant nursery of 1000 plants
Year 2 & 3
Evaluate each plant for the trait of interest
Polycross selected plants
Paircross selected plants
Year 3
Bulk harvest open pollinated seed of 50 to 100 selected plants
Year 4
Recombine 50 to 100 selected plants in isolation

Figure 1.1: Phenotypic recurrent selection using uniparental (polycross) and biparental control (pair-crosses) in perennial ryegrass with time scale needed to complete one cycle of selection. Figure adopted from Conaghan et al. [34].

harvested and combined. Seeds grown from each maternal parent represent each half-sib family. These families are evaluated for yield under sward conditions. Best performing families are crossed in three ways to produce new parents for next cycle of selection or for development of synthetic cultivars. Crossing can be done by using (i) saved parental clones of the best performing families, (ii) plants from randomly grown seeds of original crosses, or (iii) selected plants from full-sib or half-sib families (Figure 1.2 and 1.3) [19, 34]. Synthetic cultivars are further evaluated for yield before being released as commercial varieties. Genotypic recurrent selection offers inclusion of replicated, multi location trials, which are suitable for traits such as yield, which has low heritability and has a high degree of genotype environment interaction. Selection is based on the performance of progeny under sward conditions, which is useful to improve yields.

Genetic gains using recurrent selection remained low for economically important traits such as yield, persistency and quality, despite breeding for nearly 100 years [187]. Although there has been an improvement in seasonal yields (summer and autumn), it was mainly due to improvements in secondary traits such as aftermath heading and crown rust resistance [19, 155]. Some of the reasons for poor genetic gains are due to longer breeding cycles, as each cycle of selection takes up to three to five years and synthetic cultivar evaluation trials takes place every 10 years. Thus it takes almost 15 years to releasing new synthetic cultivars. Both phenotypic and marker information suggests that there is a great amount of genetic variation within and between perennial ryegrass populations, but recurrent selection is long and inefficient and takes many generations to capture a greater amount of genetic variation [28]. But with the help of genomic tools, breeders can now accelerate genetic gains by speeding up the selection cycle.

1.4 Genomic assisted breeding

Genomic assisted breeding is an indirect selection process, where molecular markers linked to the trait of interest is used to predict breeding values, rather than actual phenotype. One approach is marker assisted recurrent selection (MARS), where quantitative trait loci (QTL) are searched in the genome and markers linked to QTLs due to genetic linkage are used to predict phenotypes. MARS was based on mapping QTL in a bi-parental population using a method known as linkage mapping. But after developing high density single nucleotide polymorphisms (SNPs), genome wide association studies (GWAS) became a major tool for QTL detection over linkage mapping and genomic selection replaced MARS. In perennial ryegrass hundreds of QTLs were mapped for morphology, physiology and stress related loci in bi-parental mapping populations and were reviewed by Shinozuka et al. [160]. Despite mapping large number of QTLs, when it comes to practical context of using

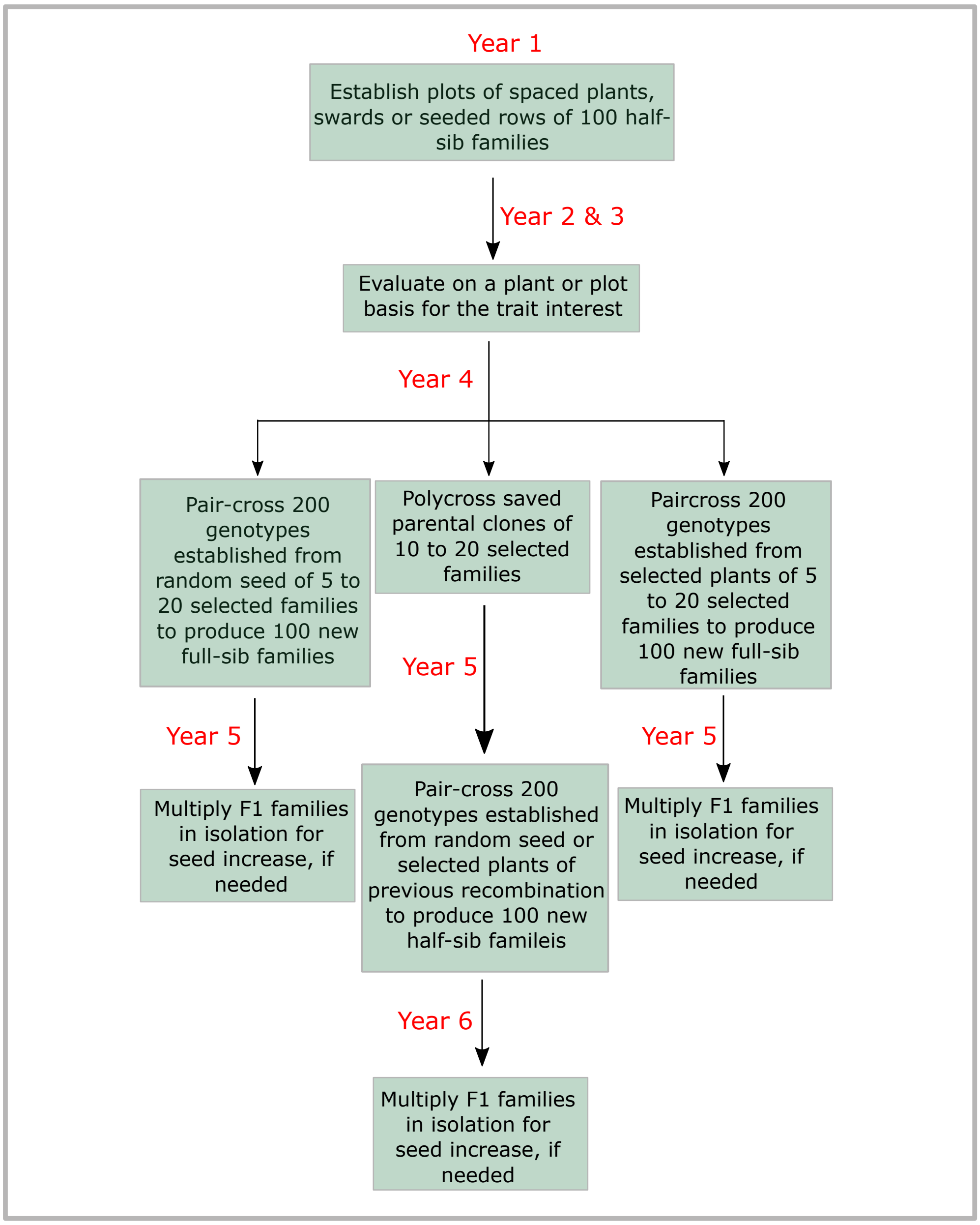
Year 1
Establish plots of spaced plants, swards or seeded rows of 100 half-sib families
Year 2 & 3
Evaluate on a plant or plot basis for the trait interest
Year 4
Pair-cross 200 genotypes established from random seed of 5 to 20 selected families to produce 100 new full-sib families
Polycross saved parental clones of 10 to 20 selected families
Paircross 200 genotypes established from selected plants of 5 to 20 selected families to produce 100 new full-sib families
Year 5
Year 5
Year 5
Multiply F1 families in isolation for seed increase, if needed
Pair-cross 200 genotypes established from random seed or selected plants of previous recombination to produce 100 new half-sib famileis
Multiply F1 families in isolation for seed increase, if needed
Year 6
Multiply F1 families in isolation for seed increase, if needed

Figure 1.2: Genotypic recurrent selection using full-sib families, figure illustrates the time scale needed to complete one cycle of selection. Figure adopted from Conaghan et al. [34].

these QTLs, MARS suffers from huge drawbacks [19]. Reasons for lack of success are due to difficulty in identifying reliable QTL-marker links in a bi-parental mapping population, that can be easily transmitted to breeding material. Another reason is the small population size with limited markers, resulting in large interval QTLs [14, 19, 89]. Dense genotyping opened up the prospect for GWAS, where

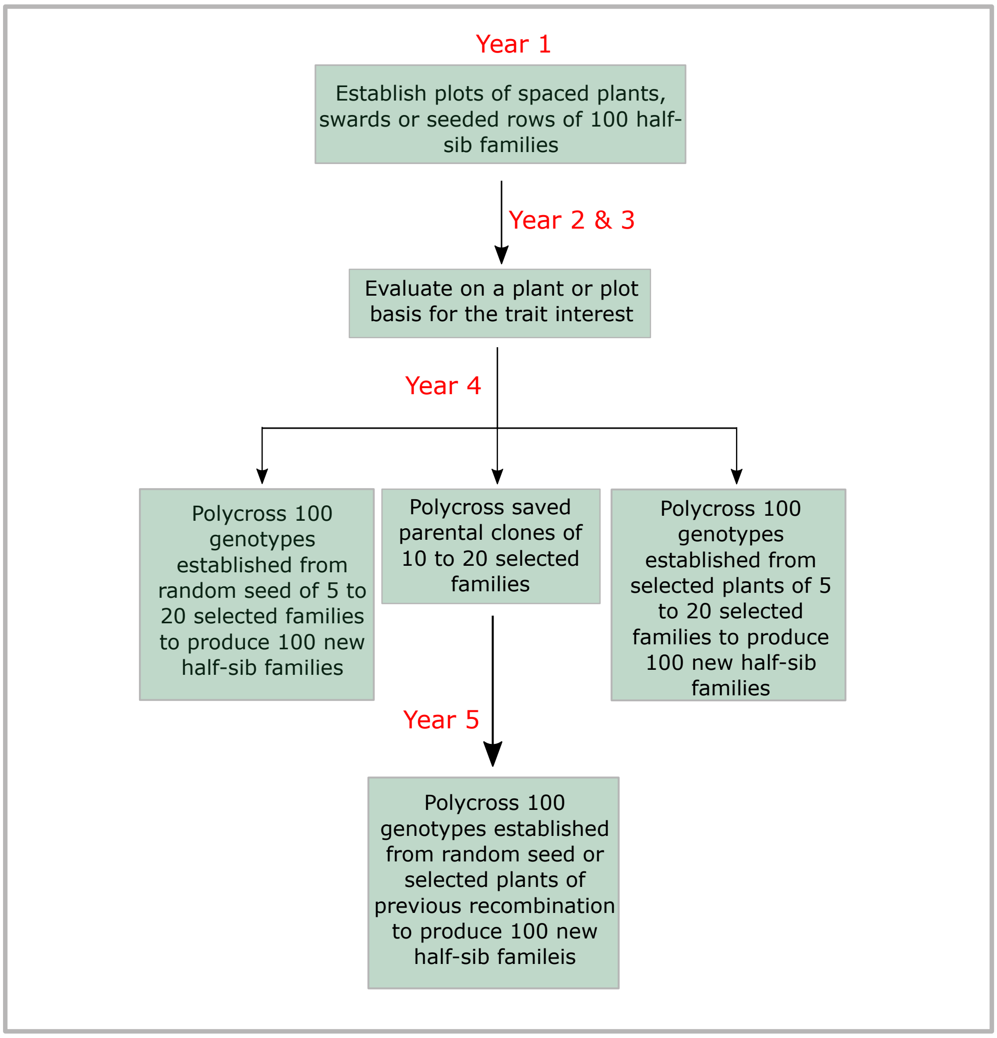
Year 1
Establish plots of spaced plants, swards or seeded rows of 100 half-sib families
Year 2 & 3
Evaluate on a plant or plot basis for the trait interest
Year 4
Polycross 100 genotypes established from random seed of 5 to 20 selected families to produce 100 new half-sib families
Polycross saved parental clones of 10 to 20 selected families
Polycross 100 genotypes established from selected plants of 5 to 20 selected families to produce 100 new half-sib families
Year 5
Polycross 100 genotypes established from random seed or selected plants of previous recombination to produce 100 new half-sib famileis

Figure 1.3: Genotypic recurrent selection using half-sib families, figure illustrates the time scale needed to complete one cycle of selection. Figure adopted from Conaghan et al. [34].

marker-trait association can be identified in same breeding population that is used for selection purpose. In using genome wide markers, linkage disequilibrium (LD) between markers persist, this will enable identifying SNPs which are in LD with causative QTL. GWAS can overcome some of the limitations caused by linkage mapping. But implementing in perennial ryegrass has its own challenges, due to low levels of LD and structure within population. For instance, in perennial ryegrass F_2 families, GWAS for heading date was only able to explain 20% of variation, despite of using nearly one million markers and a large population size [61]. Also MARS is a two step process, only markers passing the significance threshold are used for predicting, this results in accounting for only a fraction of the additive variance for the trait and tends to overestimate marker effects [12]. One way to overcome this limitation, is by using all markers simultaneously in the prediction model without any prior selection.

1.4.1 Genomic prediction

Genomic prediction was proposed to capture total additive variance using genome wide markers without prior selection of markers [123]. Genomic prediction relies on the assumption that using genome wide markers, QTL will be in LD with at least a few markers population wide, so marker effects can be estimated consistently. Genomic prediction combines marker data with phenotypes to predict genomic estimated breeding values (GEBVs). This avoids the need for sub-setting markers and will facilitate capturing small effect QTLs, which are usually missed by MARS [46, 89]. Most of the economically important traits in perennial ryegrass are quantitative in nature [16]. They are either controlled by many small effect QTLs or a mixture of large and small effect QTLs. Genomic prediction is ideal for traits with quantitative inheritance and can accelerate genetic gains in traits which are expensive, inconsistent and time consuming to phenotype [89]. It is well established in dairy cattle breeding where it has been shown to reduce the breeding cycle from 5 to 6 years to 1.5 years, but is still in early phase in plant breeding [89, 144]. To implement genomic prediction, we need a training population which is genotyped with high density markers and phenotyped for the traits of interest. A prediction model is developed on a training population using the phenotypic and genotypic information. The breeding population (population under selection) is only genotyped and GEBVs are estimated using the model trained previously. GEBVs do not provide any information about the underlying QTL, but give criteria for selecting plants [114]. As a result of using high density markers for predictive modeling, a large number of markers (p) need to be estimated compared to number of individuals (n). Classical multiple linear regression cannot handle large the "p", small "n" problem. To counter this, many statistical methods have been proposed for genomic prediction, which are reviewed by Lorenz et al. [114] and de los Campos et al. [43]. Some models are suited for quantitative traits while others perform well for traits that fall between quantitative and qualitative inheritance [114]. But the most commonly used models in plants are known as (1) genomic best linear unbiased prediction (G-BLUP) and (2) ridge regression BLUP (RR-BLUP). G-BLUP is similar to a traditional BLUP model but instead of pedigree relationship, genomic relationship is used as covariance in a mixed model. RR-BLUP is a penalized regression model, which imposes shrinkage on marker effects pushing to zero and reducing the variance of all estimates. It assumes that the variance of markers is equal and can simultaneously estimate many more marker effects. Marker effects are estimated on a training population and those effects are multiplied to marker genotypes from breeding population to estimate GEBVs. Both G-BLUP and RR-BLUP use a mixed model approach and are considered statistically similar [114]. RR-BLUP was initially proposed for MARS [183], but Meuwissen et al. [123] used it for genomic prediction, by fitting all markers as a random effect. Both models are well suited for quantitative traits with small effects [147]. RR-BLUP and G-BLUP models assume that traits are controlled by many loci with small effect (infinitesimal model) and captures only additive genetic variance.

Accuracy of genomic prediction can be assessed based on predictive ability, which is the Pearson's correlation coefficient of true phenotypic value and estimated breeding value. Cross validation is the preferred method to estimate correlation. Factors such as LD within the population, marker density, training population size, relationships between training and test set and genetic architecture of the trait can influence predictive ability. These factors are also interrelated with each other. Extent of LD within a population can determine the marker density and training population required for achieving maximum predictive ability. But useful LD in the population is largely dependent on the past effective population size (N_e). Lower effective population size leads to long range LD in the population due to rapid genetic drift. Based on previously observed LD patterns [8, 63], effective population size for perennial ryegrass is assumed to be very large and appropriately 10,000 according to Hayes et al. [77]. To capture LD between adjacent markers, for perennial ryegrass with large N_e, we need a very large training population and almost a million markers [77]. But empirical studies have shown that as we increase marker density and training population size, predictive ability increases but not linearly. Predictive ability doesn't show any improvement after reaching certain marker density and population size [29, 61, 148]. The level at which marker density and training population reaches its threshold depends on the degree of relationship between training and test set. In a population with low levels of LD, predictive ability can still reach maximum if there is a relationship between the training and test set. Habier et al. [73] demonstrated by simulation, even in the absence of LD in the population predictive ability was non zero, because of markers capturing genetic relationship [73, 112]. Using a lower marker density and smaller training population, predictive ability was high in several plant species [72, 148]. Complexity of genetic architecture is inversely proportional to predictive ability, as the QTL number increases, we see lower predictive ability. For example, traits like heading date can be predicted more accurately compared to crown rust or yield [60, 61, 72]. Factors like genetic architecture and LD within the population cannot be easy controlled. Designing the training population in such a way that the relationship and LD persists longer within the training and breeding population, can impact on marker density and number of individuals needed to phenotype and genotype to achieve maximum predictive ability.

1.5 Objectives of this thesis

The overall goal of this thesis was to investigate the use of molecular markers to accelerate genetic gain for economically important traits in perennial ryegrass. This thesis focuses on using GWAS to identify markers significantly associated with QTL, and on using genome-wide markers to generate Genomic Estimated Breeding Values (GEBVs) for traits of agronomic importance.

The specific objectives of this thesis were to:

1. Investigate the potential of using GWAS to identify molecular markers for heading date and crown rust resistance that could be utilized in marker assisted recurrent selection strategies (chapter 2 and chapter 3).

2. Investigate the accuracy of using genome-wide markers to generate breeding values and evaluate factors affecting predictive ability (chapter 3).

3. Evaluate genome-wide selection for predicting forage yield under both grazing and conservation managements in a tetraploid perennial ryegrass breeding population (chapter 4).

Chapter 2

Markers associated with heading and aftermath heading in perennial ryegrass full-sib families

Sai krishna Arojju[1,3], Susanne Barth[1], Dan Milbourne[1], Patrick Conaghan[2], Janaki Velmurugan[1], Trevor R. Hodkisnon[3] and Stephen L. Byrne[1*]

[1] Teagasc, Crop Science Department, Oak Park, Carlow, Ireland

[2] Teagasc, Grassland Science Research Department, Animal and Grassland Research and Innovation Centre, Oak Park, Carlow, Ireland

[3] Department of Botany, Trinity College Dublin, Dublin2, Ireland

BMC Plant Biology 16.1 (2016): 160.

Abstract

Heading and aftermath heading are important traits in perennial ryegrass because they impact forage quality. So far, genome-wide association analyses in this major forage species have only identified a small number of genetic variants associated with heading date that overall explained little of the variation. Some possible reasons include rare alleles with large phenotypic effects, allelic heterogeneity, or insufficient marker density. We established a genome-wide association panel with multiple genotypes from multiple full-sib families. This ensured alleles were present at the frequency needed to have sufficient statistical power to identify associations. We genotyped the panel via partial genome sequencing and performed genome-wide association analyses with multi-year phenotype data collected for heading date, and aftermath heading. Genome wide association using a mixed linear model failed to identify any variants significantly associated with heading date or aftermath heading. Our failure to identify associations for these traits is likely due to the extremely low linkage disequilibrium we observed in this population. However, using single marker analysis within each full-sib family we could identify markers and genomic regions associated with heading and aftermath heading. Using the ryegrass genome we identified putative orthologs of key heading genes, some of which were located in regions of marker-trait associations. Given the very low levels of LD, genome wide association studies in perennial ryegrass populations are going to require very high SNP densities. Single marker analysis within full-sibs enabled us to identify significant marker-trait associations. One of these markers anchored proximal to a putative ortholog of TFL1, homologues of which have been shown to play a key role in continuous heading of some members of the rose family, Rosaceae.

Keywords: aftermath heading, flowering, genome wide association, heading, *Lolium perenne*, perennial ryegrass, single marker analysis

2.1 Introduction

Perennial ryegrass (*Lolium perenne* L.) is an important forage species grown in temperate regions of the world where it underpins the dairy and livestock sectors. This is due to a high palatability and digestibility [120]. It also displays relatively rapid establishment and has long growing seasons with relatively high yields in suitable environments [184]. With 38% of global land area available for agriculture, 70% is assigned as pastoral agricultural land [68]. In Europe alone 76 million hectares is used as permanent pasture [64] and in Ireland 80% agricultural land (3.4 million hectares) is used for pasture, hay and silage where perennial ryegrass is the preferred species [134].

Heading date is a trait that can have a large effect to the use of perennial ryegrass as a forage [121]. Heading has an impact on digestibility, biomass production, persistency and nutritional value [85, 164, 192]. The stem and inflorescence formation significantly reduces tiller formation and affects the persistency, digestibility and nutritional value [177]. Perennial ryegrass belongs to the same sub-family (Pooideae) as several other important grain cereals such as barley, oats, rye and wheat [47, 186]. Heading in situations outside of seed production is unwanted as it negatively impacts forage quality by increasing the stem to leaf ratio. Extending the vegetative period would greatly enhance its utility as a forage [27, 66]. Aftermath heading is mainly associated with early heading genotypes, and these tend to show lower persistency and perenniality. There has been limited work done on the genetic control of aftermath heading, and only a single quantitative trait loci (QTL) has been mapped onto linkage group (LG) 6 in an experimental mapping population [192].

In perennial ryegrass, heading is mainly controlled by three main pathways, namely the vernalization pathway, the photoperiod pathway and the circadian clock. To date many QTL mapping studies have been carried out in perennial ryegrass and major loci involved in the floral transition have been identified [2, 5, 6, 9, 22, 92, 160, 162, 163, 167, 192]. QTL for heading date have been detected on all seven LGs of perennial ryegrass, with analogous regions on LG4 and LG7 being linked with large affect QTL across multiple populations [92]. Although genes underlying some of these QTL have been proposed [6, 92] none have been cloned to date.

In addition to within family based QTL analysis, we can also map QTL in populations using genome wide association analysis (GWAS). This offers the benefit of being able to take advantage of historical recombination to more precisely map the QTL region. In the case of a very rapid decay of linkage disequilibrium (LD), the causative quantitative trait nucleotide (QTN) may be elucidated. However, this does necessitate the need for a high marker density. A recent GWAS study of heading date in perennial ryegrass identified markers affecting heading date across 1,000 F2 families [61]. However, the variation explained by the combined marker set was extremely small. LD only extended to very short distances in the study population, and despite using in excess of 0.9 million SNPs the marker density may be insufficient. Alternatively, rare variants affecting the trait may have resulted in low statistical power to identify associations.

Here, we have developed an association mapping population of 360 individuals coming from six full-sib families with contrasting primary heading dates. Multiple individuals from each full-sib family were selected to ensure any allele will be present at a frequency suitable for association analysis. However, the low levels of LD across families restricted GWAS at our marker density, and so we performed single marker analysis within each full-sib family separately. Anchoring markers to the perennial ryegrass GenomeZipper [137, 166] allowed us to identify regions containing clusters of associated markers, some of which co-located with genes having a known involvement in controlling heading and aftermath heading.

2.2 Material and methods

2.2.1 Plant material and phenotypic data

The association population consisted of 360 individual plants from six full-sib F_2 families (60 individuals selected at random from each family) (Table 2.1). Plants were established in the glasshouse and then transplanted into the field in a spaced plant nursery in 2013 at Oak Park, Carlow, Ireland in two replicates. Each replicate consisted of 30 blocks with 2 individuals from each full-sib family within a block. The number of days to heading from April 1^{st} was monitored in 2014 and 2015 for each plant. An individual plant was considered as headed, when three or more heads had emerged from the leaf sheath. In the same population aftermath heading was visually scored in the year 2015 on a scale of 1 (no aftermath heading) to 9 (intense aftermath heading) as described in Fe et al. [62]. Using the R package lme4 [10] variance components were estimated for heading date using genotype, year, and the genotype by year interaction as random effects. Best linear unbiased predictions (BLUPs) were calculated and used for subsequent analysis.

2.2.2 Genotyping full-sib families

We used a genotyping-by-sequencing approach that followed the protocol developed by Elshire et al. [57]. Briefly, genomic DNA was isolated from each individual, digested with ApeKI, samples were grouped into libraries, amplified, and sequenced on an Illumina HiSeq 2000. After sequencing, adaptor contamination was removed with Scythe [20] with a prior contamination rate set to 0.40. Sickle [97] was used to trim reads when the average quality score in a sliding window (of 20 bp) fell below a phred score of 20, and reads shorter than 40bp were discarded. The reads were demultiplexed using sabre [96] and data from each sample was aligned to the perennial ryegrass reference genome[23] using BWA [107]. The Genome Analysis Tool Kit (GATK) [45] was used to identify putative variants across the full-sib families, and also within each full-sib family. Only genotype calls with a phred score of 30 (GQ, Genotype Quality), and only variant sites with a mean mapping quality of 30 were retained. In the case of the SNP set across all full-sib families, we used a minimum minor allele frequency threshold of 5%. When identifying SNP set within each full-sib family we used a minimum minor allele frequency threshold of 10%.

2.2.3 Genome wide association and linkage disequilibrium (LD) analysis

A mixed linear model implemented in the R package GAPIT (Genomic Association and Prediction Integrated Tool) [111] was used to perform an association analysis. The mixed model accounts for population structure and family relatedness using principal component analysis (PCA) and a kinship matrix calculated by GAPIT with available input genotypic data. To account for multiple testing during association analysis, false discovery rate (FDR) [13] with an α level of 0.05 was used as a threshold. To assess the extent of LD across the full-sib populations we identified

SNPs located within a single genomic scaffold, and calculated the inter SNP distance and the squared correlation of the allele counts in Plink 1.9 [30], based on the maximum likelihood solution to the cubic equation [69].

2.2.4 Pipeline for single marker analysis

A SNP panel was developed for each full-sib family, using a 10% of minor allele frequency, and subsequent analyis was performed on each of the six full sib families independently. SNPs segregating in a 1:1 ratio were selected, that is homozygous in parent one and heterozygous in parent two. A X^2 test was used to eliminate SNPs that deviated significantly from a 1:1 segregation. We then performed a Kruskal-Wallis test using R [172] on each marker to check for association with heading date. Using the GenomeZipper [23, 137] we established a putative order for these markers along the genetic map. The median Kruskal-Wallis test statistic was calculated for bins represented by gaps between markers on the genetic linkage map.

2.2.5 Protein datasets and phylogenetic analysis

The query proteins for key heading genes was obtained from *Arabidopsis thaliana*, rice and barley using the uniport database [39]. The complete protein sets from perennial ryegrass [23], *Arabidopsis*[99], *Brachypodium*[179], barley[38], rice[142], *Sorghum* [135] and maize [158] were gathered from PLAZA 3.0 [143] and combined into single file to build a BLASTp database. Using each query we performed a BLASTp with an evalue of 10e-10 and parsed the results for hits with at least 60% coverage and 50% identity. The sequences were aligned using MUSCLE [56], an alignment program implemented in MEGA 6.06 [169]. The phylogenetic analysis was carried out using the Maximum Likelihood method based on the JTT matrix-based model in MEGA 6.06 [95, 169]. Bootstrap values after 100 replicates are shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is mid-point rooted, drawn to scale, with branch lengths measured in the number of substitutions per site.

2.3 Results and Discussion

2.3.1 Phenotypic variation for heading date and aftermath heading

The 360 genotypes were planted in two replicates at Oak Park, Carlow, Ireland, and were scored for days to heading in both 2014 and 2015. This was scored as the number of days from April 1st until three spikes had emerged on a single plant. In all families, days to heading follows a normal statistical distribution. Plants are generally assigned to one of three groups for heading, these are early (head in first half of May), intermediate (head in second half of May), and late (head in first half of June). The full-sib families G15, G16, and G17 were all developed from late heading parents (Table 2.1) and this is evident in the phenotypic distributions for these families (Figure 2.1). Only G11 had an early heading parent, and G12 and G18 involved intermediate heading parents. Scores for heading date were strongly correlated between 2014 and 2015, with a Pearson's product-moment correlation of 0.82 with a 95% confidence interval of (0.79, 0.84). Variance components were calculated using lme4 [10] with genotype, year, and the genotype by year interaction as random effects. From this, we calculated heritability on a line mean basis to be 0.91.

Table 2.1: Full-sib family structure

	Parent1	Parent2	Crosses
G11	Pastour	Genesis	Late X Early
G12	Solomon	Tyrella	Inter X Late
G15	Profit X Hercules	Jumbo X Tyrone	Late X Late X Late X Late
G16	AberAvon	Twystar	Late X Late
G17	Tyrconnell	Majestic	Late X Late
G18	AberSilo	Shandon	Inter X Inter

Aftermath heading was scored only in September 2015 using a visual assessment on a scale of 1 (no aftermath-heading) to 9 (extensive aftermath heading). The Pearson's product-moment correlation between replicates was 0.68 and a 95% confidence interval of (0.61, 0.73). The difference in aftermath heading scores between replicates was not significant($F_{(1,685)} = 3.385$, MSE=21.522, $P = 0.07$) at α = *0.05.*

The population mean, sd, and median scores for aftermath heading were 2.7, 2.7 and 1, respectively.

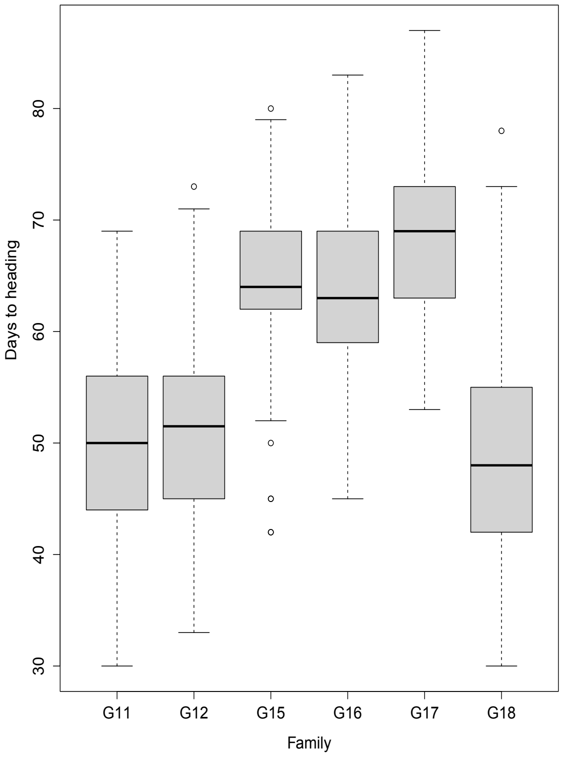

Figure 2.1: Phenotype distribution of heading date in six full-sib families. Box-plots representing heading date in full-sib families with y-axis showing days to heading and families on x-axis.

We only have a single years data for aftermath heading, however, a recent study of 1453 F_2 families of perennial ryegrass determined heritibalities for aftermath heading that were in line with those determined for heading date [62]. The distribution of scores within each full-sib family, we see that one family (G18) has more variation and a higher propensity for aftermath heading. Taking this family in isolation we looked at the association between heading date and aftermath heading. Using aftermath heading as a response variable in linear regression, we can see that earlier heading individuals tend to have higher aftermath heading.

2.3.2 Genome wide association analysis

We used a genotyping by sequencing approach to characterize variation in the association panel. Data were aligned to the reference genome [23] and variants were identified across the entire 360 genotypes (Table 2.2). Only variants present in at least 70% of samples and at a minor allele frequency of 5% were retained. This left 51,846 SNPs for association analysis with the traits heading date and aftermath heading. We corrected for population structure using principal component analysis and the kinship matrix (Figure 2.2). The purpose of including 60 genotypes from each of six full-sib families was to inflate the allele frequencies to ensure we had adequate statistical power for association studies. It is possible that many traits in perennial ryegrass may be controlled by rare alleles with large effects, but in order to detect associations an allele must be present in high enough frequency. Within our association panel the rarest allele would, in theory, be present in 30 individuals (each full-sib family is the result of a single pair cross followed by seed multiplication in isolation plots).

Table 2.2: Markers used at different stages of pipeline

Family	Genotyped markers	Filtered markers	Chi-square	Tagged to zipper
All	51,846	-	-	-
G11	27,934	15,315	7070	2174
G12	59,524	28,606	15,564	4228
G15	77,499	32,805	19,425	4424
G16	62,948	29,263	15,563	3421
G17	63,516	27,007	14,315	3523
G18	17,701	6021	3075	1225

We did not find any markers significantly associated with heading date or aftermath heading after correcting for multiple testing (FDR < 0.05). Heading date is a highly heritable trait [62], and one that can be phenotyped very precisely. It was therefore initially surprising that we did not identify any significant associations. Genome wide association studies can fail for many reasons, including a lack of statistical power due to many rare alleles with large effects or allelic heterogeneity. However, to avoid this problem we have used multiple genotypes (60) from each of six full sib families. Another possible explanation is that heading date (and aftermath heading) is highly correlated with population structure, meaning any correction for population structure will result in false negatives.

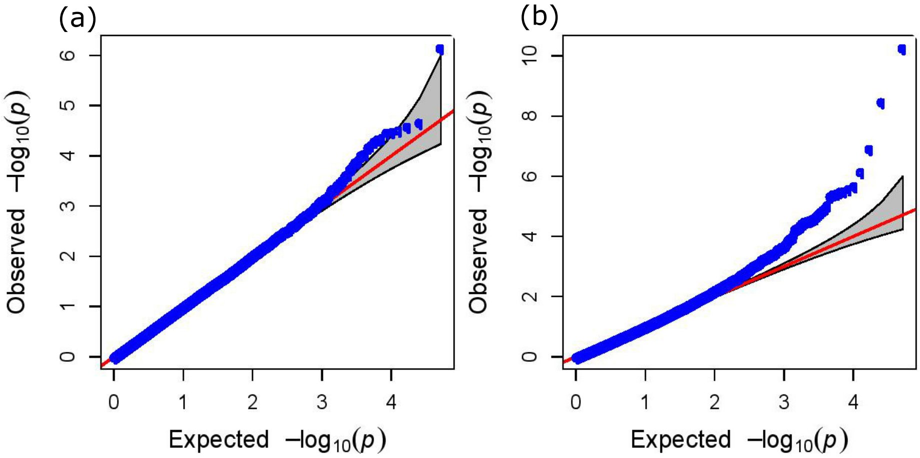

Figure 2.2: QQ-plots for (a) heading and (b) aftermath heading

Designing populations to remove any population structure is complicated for a trait like heading date as synchronization of heading is required for cross pollination. Another possible explanation is that the marker density is insufficient to ensure we always have a marker in LD with a QTL. The six F_2 families were developed from pair-crosses of 12 genotypes taken from a recurrent selection program. When we evaluated the extent of LD in the population, we observed that on average across the genome it decayed very rapidly (Figure 2.3). Based on this, our marker density is not sufficient, even considering that our genotyping approach is focused on the non-repetitive and gene-rich fractions of the genome. In this case it is likely that full re-sequencing of the gene space and regions up and downstream is required to capture alleles associated with a trait. A recent GWAS study in perennial ryegrass using almost 20 times the number of markers ($\sim$ 1 million SNPs), did identify significant SNPs for heading. Some SNPs were in close proximity with key heading genes like *CONSTANS* (*CO*) and *PHYTOCHROME C* (*PHYC*), but the sum of the variances explained by all significant markers was only 20.3% [61].

We have now established that, in general, marker numbers in the region of 50,000 are going to be unsuitable for GWAS in perennial ryegrass. We believe that our inability to find any significant associations with heading date and aftermath heading was due to low marker density and the extremely low LD in the population. Our population is made up of six full-sib families, and within each family a much higher

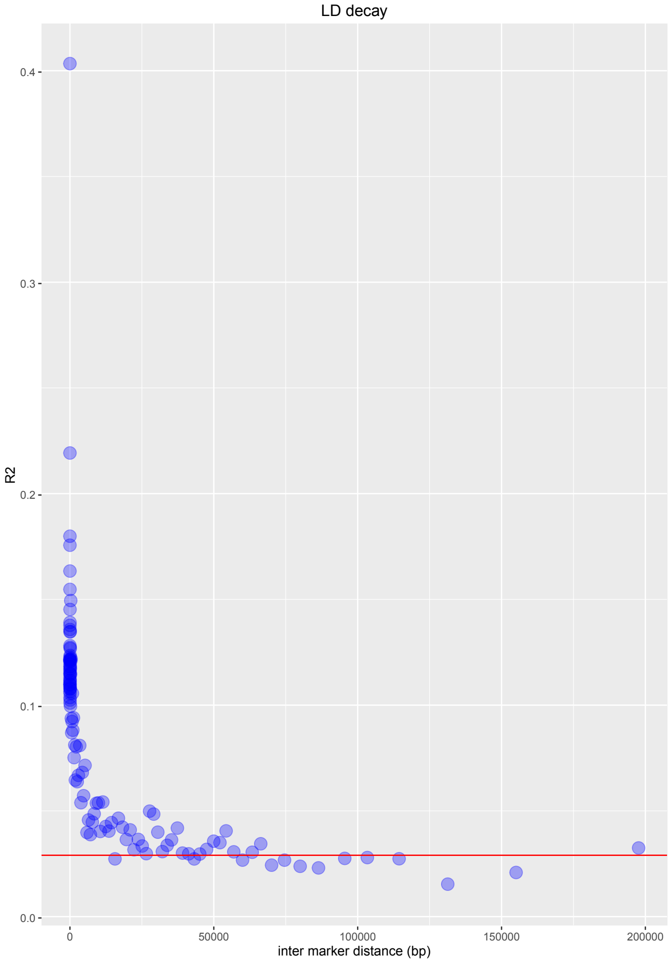

Figure 2.3: Extent of linkage disequilibrium (LD) measured as the squared correlation of allele counts (y-axis), based on the maximum likelihood solution to the cubic equation. The x-axis shows inter marker distance in bp. LD estimates were sorted according to inter-marker distance, and divided into bins of 1000 estimates. Each point on the plot represents the mean R^2 and mean inter-marker distance of 1000 measurements.

LD is expected. An alternative approach would be to perform single marker analysis within each full-sib family. There are only 60 genotypes per full-sib family, however, using this approach there is sufficient SNP density to perform a simple marker-trait association analysis within each family separately. This would not enable us to locate the regions directly affecting a phenotype, but would allow us to identify markers linked to QTL.

2.3.3 Single marker analysis in full-sib families

The original genotypes used in the pair-crosses that generated the six full-sib families were not available for genotyping. We redid the SNP calling on each full-sib family in isolation, and filtered out variants with a minor allele frequency of less than 10%. We only selected SNPs that were segregating in a 1:1 ratio, corresponding to sites that were homozygous in parent one and heterozygous in parent two. This was done because there are only 60 genotypes present in each full-sib family, and so any markers that segregate into more than two marker classes would have a limited number of individuals in each class. A Kruskal-Wallis test was performed on each marker to identify if they were significantly associated with heading date. We then used the perennial ryegrass GenomeZipper [23, 137] to generate a putative order for the markers according to the linkage map upon which the GenomeZipper is based. These data were used to generate heatmaps for each linkage group showing the Kruskal-Wallis test statistic (Figure 2.4-2.6 and Figure S2.1-2.4).

In general the strongest marker-trait associations were identified in the families G11 and G12, particularly on LG4 and LG7 (Figure 2.5 and 2.6). The three families G15, G16, and G17 were all the result of crosses between late heading plants, and these three full-sib families showed the smallest range in days to heading (Figure 2.1). Only G11 was from a cross between an early and late heading populations (Table 2.1). The PCA shows a separation according to the categorization of parental days to heading on the first principal component, which accounts for 7.8% of the variation. The two full-sib families involving crosses between parents falling into different heading categories (G11 and G12) are separated from the others on PC1 (Figure 2.7). We identified many markers associated with days to heading, particularly on LG4 and LG7 (Figure 2.5 and 2.6). This was not too surprising, considering that many studies in experimental cross-populations have identified large effect QTL on the same linkage groups [2, 5, 6, 9, 22, 92, 160, 162, 163, 167, 192].

G18 was the only family that showed a large amount of variation for this trait. Single marker analysis identified markers significantly associated with aftermath heading anchored onto different LGs using the GenomeZipper (Table 2.3). In particular we identified markers in five scaffolds anchored to LG6 in a region covering 35.9 to 56.0 cM (Table 2.3). We also identified markers in two scaffolds on LG2 at 80.4 and 84.2 cM, and markers in two scaffolds anchored to LG1 at 31.5 and 31.2 cM. The recent

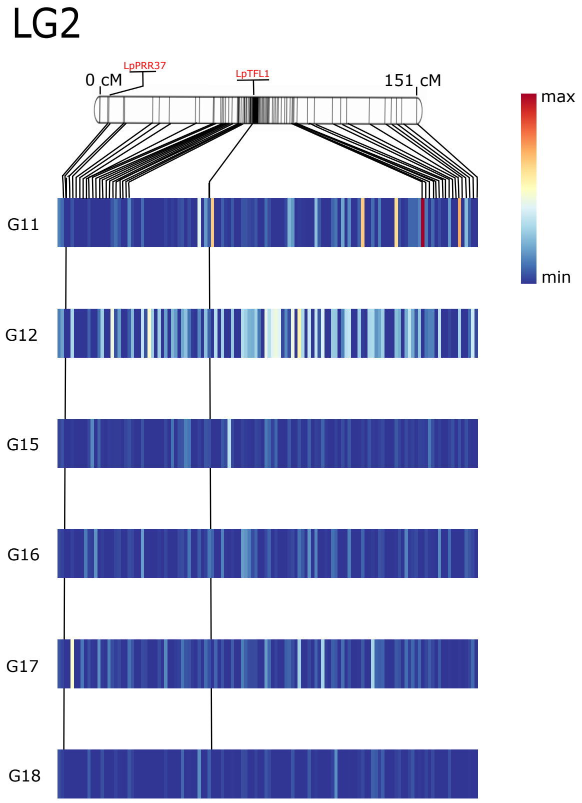

Figure 2.4: Heatmap illustrates regions associated with heading over six full-sib families on perennial ryegrass LG2. A Kruskal-Wallis test was performed on each marker to identify significant regions for heading. Using the perennial ryegrass genome zipper [23, 137] we identified a putative gene order for markers on LG2. These data were used to construct the heatmap for each family. A perennial ryegrass transcriptome-based genetic linkage map upon which GenomeZipper was based was used as reference to construct LG2 [137, 166] and placed above the heatmap. Each bar in the heatmap represents region between two genetic markers from the linkage map. The median Kruskal-Wallis test statistic was calculated for markers binned between markers on the genetic linkage map and used to construct the heatmap. Putative orthologs of LpPRR37 and LpTFL1, were identified in the phylogenetic analysis and placed onto LG2 using genetic positions from genome zipper. The genetic positions of these orthologs were extrapolated over the heatmap. Color of the heatmap illustrates the test-statistic of the Kruskal-wallis analysis from 0 to 21.

release of an annotated draft assembly of the perennial ryegrass genome [23] enables

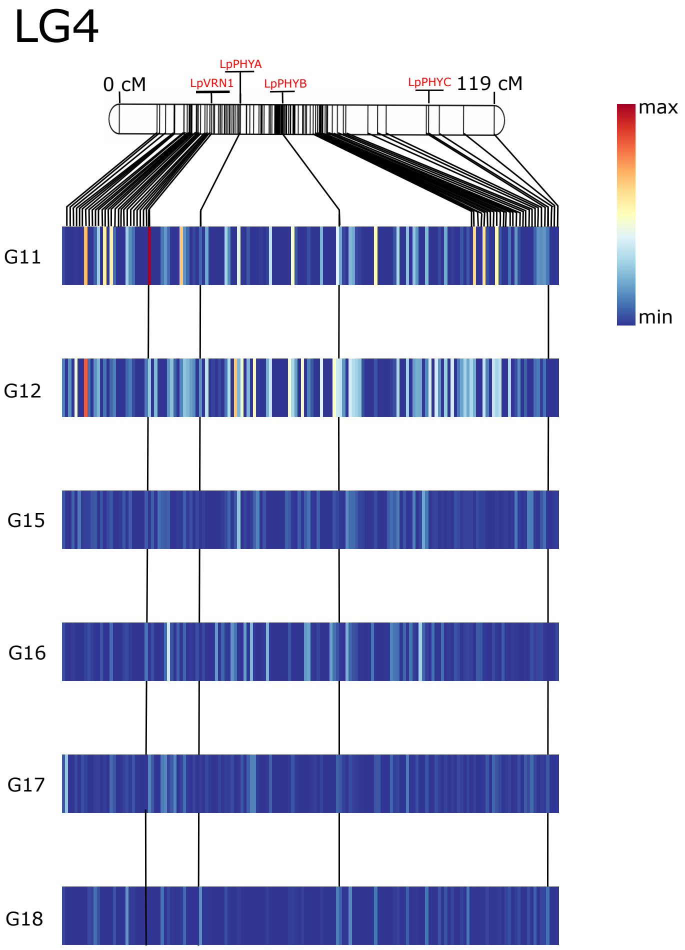

Figure 2.5: Heatmap illustrates regions associated with heading over six full-sib families in perennial ryegrass LG4. Similar methodology was used as described in Figure 2.4. Putative orthologs of LpVRN1, LpPHYA, LpPHYB and LpPHYC, were identified in the phylogenetic analysis and placed onto LG4 using genetic positions from genome zipper. The genetic positions of these orthologs were extrapolated over the heatmap as bars. Color of the heatmap illustrates the test-statistic of the Kruskal-wallis analysis from 0 to 22.

us to identify putative orthologs of key heading genes from model species. Using the GenomeZipper we can locate these on the genetic map and relate them to the marker-trait associations identified above.

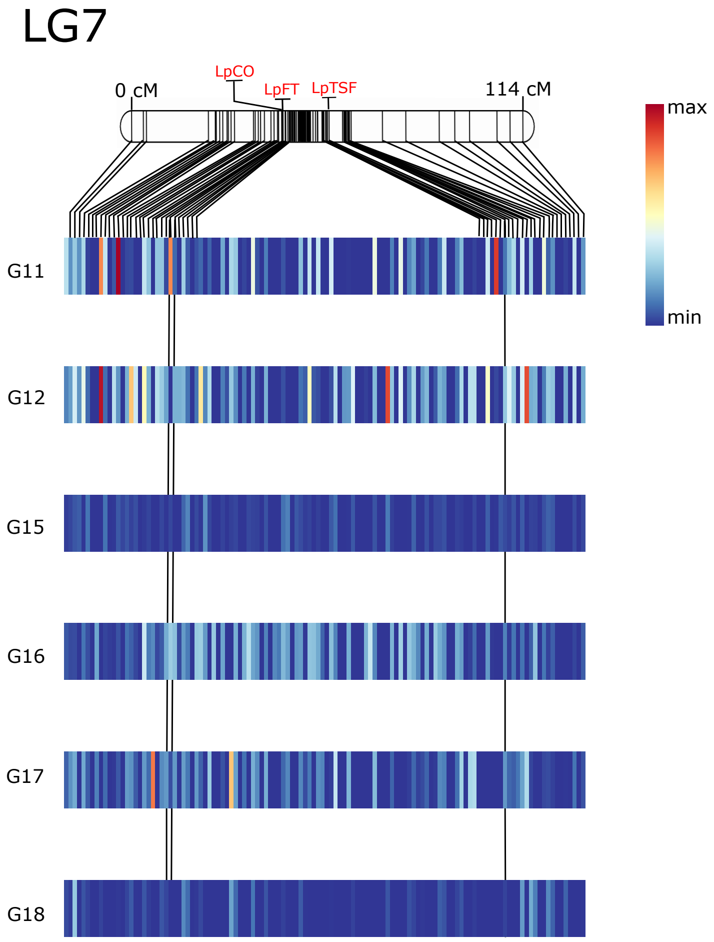

Figure 2.6: Heatmap illustrates regions associated with heading over six full-sib families in perennial ryegrass LG7. Similar methodology was used as described in Figure 2.4. Putative orthologs of LpCO and LpFT, were identified in the phylogenetic analysis and placed onto LG7 using genetic positions from genome zipper. The genetic positions of these orthologs were extrapolated over the heatmap. Color of the heatmap illustrates the test-statistic of the Kruskal-wallis analysis from 0 to 15.

2.3.4 Identifying putative orthologs of key heading genes

Work in model species has identified various genetic pathways controlling heading date (Figure 2.8)[3]. Key genes acting within these pathways have also been characterized. We used the perennial ryegrass genome [23] to identify putative perennial ryegrass orthologs to these regulators of heading date. We used protein sequences from *Arabidopsis*, barley and rice as queries (Table 2.4) to search the perennial ryegrass protein set, and protein sets from *Arabidopsis thaliana* [99], *Brachypodium distachyon* [179], *Hordeum vulgare* [38], *Zea mays* [158], *Sorghum bicolor* [135], and

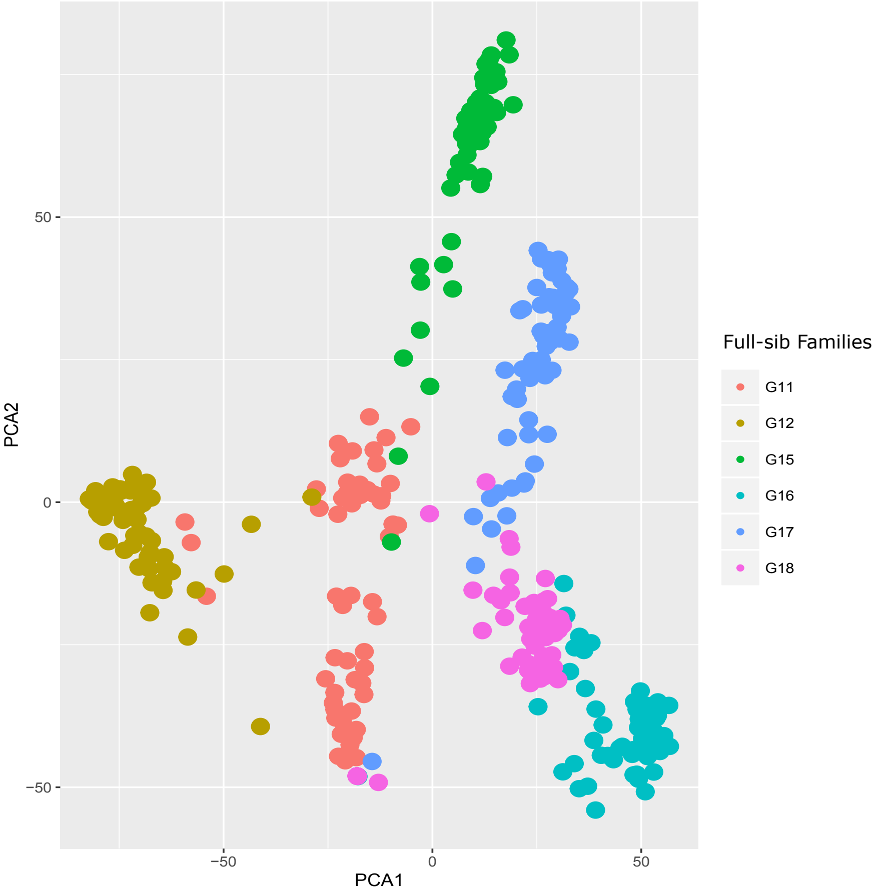

Figure 2.7: Principal component analysis (PCA) of 360 perennial ryegrass individuals, genotyped using 51,846 SNPs. The first two principal components explained 14.8% of total variation. Components are colored according to family (color coding is listed in figure legend).

Oryza sativa [142]. Only matches with a minimum query coverage of 60% and a minimum identity of 50% were retained for further analysis. The proteins were aligned with an alignment program MUSCLE and phylogenetic trees were built for each of 18 candidate genes. Using phylogenetic trees is the preferred method to establish orthology relationships [67]. Using this approach we were able to identify putative perennial ryegrass orthologs to eleven of these genes (Table 2.4). We also queried the perennial ryegrass GenomeZipper to identify putative locations for these genes on the genetic map, and relate the locations to markers we anchored as described above.

We identified putative orthologs of the important photo-receptor proteins PHYA (ms_13514|ref0035067-gene-0.0mRNA), PHYB (ms_4484|ref0039062-gene-

Table 2.3: Single marker analysis for aftermath heading

Scaffold	Position	Test statistic	p-value	q-value	LG	cM
7674	27953	16.05	6.15x10^{-05}	0.045	1	57.5
7094	37030	15.46	8.39x10^{-05}	0.045	3	60.4
9166	29048	14.26	0.00015	0.045	3	28.8
4946	43369	13.69	0.00021	0.045	4	0
4946	43375	13.69	0.00021	0.045	4	0
444	103726	13.08	0.00029	0.045	2	84.2
444	103736	12.69	0.00036	0.045	2	84.2
444	103741	12.69	0.00036	0.045	2	84.2
8926	31874	12.55	0.00039	0.045	6	40.8
8309	12264	12.51	0.00040	0.045	1	31.5
8309	12267	12.51	0.00040	0.045	1	31.5
4418	6989	12.43	0.00042	0.045	1	31.2
1397	62074	12.41	0.00042	0.045	4	34.0
4165	42625	12.31	0.00044	0.045	6	55.2
9166	29013	12.23	0.00046	0.045	3	28.8
9166	29061	12.23	0.00045	0.045	3	28.8
4418	6933	12.18	0.00048	0.045	1	31.2
16758	8509	12.03	0.00052	0.045	5	26.4
9159	27424	11.91	0.00055	0.045	6	44.5
500	65053	11.63	0.00064	0.045	6	35.9
444	103718	11.58	0.00066	0.045	2	84.2
444	103742	11.58	0.00066	0.045	2	84.2
19325	5681	11.52	0.00068	0.045	1	13.0
5444	49105	11.51	0.00068	0.045	1	17.2
498	2811	11.51	0.00069	0.045	3	41.8
498	2853	11.51	0.00069	0.045	3	41.8
5379	10675	11.41	0.00072	0.046	3	28.8
4265	21824	11.33	0.00076	0.047	6	56.0
14987	7423	11.15	0.00084	0.048	4	45.3

0.0mRNA), PHYC (ms_2801|ref0025790-gene-0.3mRNA) and CRYTOCHROME 2 (CRY2) (ms_4185|ref0010917-gene-0.1mRNA) (Figure 2.8) (Figure S2.5 and S2.6)

Table 2.4: Phylogenetic relationships of candidate genes involved in heading

Query candidate	Species	Protein near query	Position on Zipper	Genetic position on Zipper	Phylogenetic tree
TFL1	Arabidopsis	ms_821|ref0016245	LG2-79.8cM	LG5-27.5cM	Figure 9
FT	Arabidopsis	ms_13332|ref0029013	LG7-43.6cM	LG7-57.3cM	Figure 9
CRY2	Arabidopsis	ms_4185|ref0010917	LG6-52.5cM	LG6-52.5cM	Figure S2.7
GI	Arabidopsis	ms_1276|ref0038679	LG3-29.6cM	NA	Figure S2.11
PHYA	Arabidopsis	ms_13514|ref0035067	LG4-38.5cM	LG4-38.5cM	Figure S2.6
PHYB	Arabidopsis	ms_4484|ref0039062	LG4-51.0cM	LG4-51.0cM	Figure S2.6
PHYC	Arabidopsis	ms_2801|ref0025790	LG4-98.2cM	LG4-98.2cM	Figure S2.6
PRR37	Rice	ms_13366|ref0021945	LG2-12.4cM	NA	Figure S2.8
SOC1	Arabidopsis	ms_6002|ref0025562	LG6-0cM	NA	Figure S2.9
VRN1	Barley	ms_312|ref0002704	LG4-31.4cM	LG4-31.4Cm	Figure S2.12
CO	Rice	ms_5059|ref001989	LG7-43.5cM	LG7-42.7cM	Figure S2.10

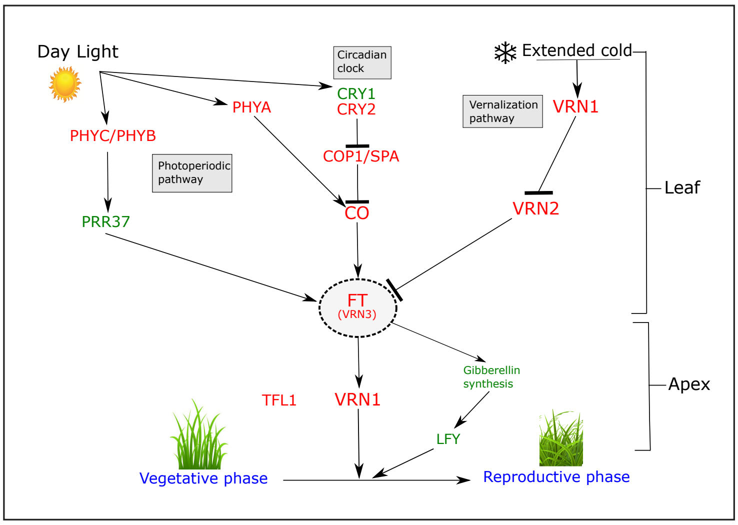

Figure 2.8: Schematic view of genetic pathway controlling heading. Genes promoting heading were shown by arrows and genes acting as repressor shown as lines with bars. External factors like day light and extended cold periods were represented with respective symbols. Pathways were mentioned in grey boxes and genes shown in red were considered as key regulators in heading.

and located these on the genetic map via the GenomeZipper (Figure 2.5 and Figure S2.4). The three Phytochromes, A, B, and C are anchored onto LG4 at different locations. All three are in locations where markers significantly associated with days to heading in one or more full-sib families. CRY2 is located on LG6 at 52.5 cM, in a region where we identified a cluster of markers between 35.9 and 56 cM that were associated with aftermath heading in G18 (Table 2.3). We also identified putative orthologs to PSEUDO RESPONSE REGULATOR PROTEIN 37 (PRR37) (ms_13366|ref0021945-gene-0.0mRNA) and SUPRESSOR OF OVER-EXPRESSION OF CO 1 (SOC1) (ms_6002|ref0025562-gene-0.0mRNA) that play important roles in the central circadian clock (Figure S2.7 and S2.8) (Figure 2.4). Another important photoperiodic pathway gene is CO, which directly regulates the key floral activator FT (Figure 2.8), and we have found a putative ortholog to CO (ms_5059|ref0019898-gene-0.1mRNA) (Figure S2.9) in perennial ryegrass that anchored onto LG7 at 43.5cM (Figure 2.6).

Both the regulation of CO and stability of photoreceptors is controlled by GIGANTEA (GI) that is generally believed to be single copy, with a highly conserved role across the angiosperms [126]. We identified a putative ortholog to GI (ms_1276|ref0038679-gene-0.4mRNA) (Figure S2.10) in our analysis that anchored onto LG3 at 29.3cM [68] (Figure S2.2). A scaffold with markers significantly associated with aftermath heading in G18 was anchored to LG3 at 28.8 cM. In addition to genes from the photoperiodic pathway, we identified a putative ortholog of the barley VERNALIZATION 1 (VRN1) (ms_312|ref002704-gene-0.1mRNA) protein that is involved in the vernalization pathway (Figure S2.11). The VRN1 protein is anchored on LG4 at 31.4cM (Figure 2.5) in a region with markers significantly associated with days to heading. This was most evident in the full-sib family G11 that was generated from crossing early and late heading populations. It has already been shown that a dominant mutation in VRN1 promoter region is responsible for changes in growth habit of winter wheat to spring wheat [115]. The vernalisation and the photoperiod pathway influence heading by acting on the key floral pathway integrator FT (Figure 2.8).

2.3.5 Perennial ryegrass orthologs of FT and TFL1

Floral transition is controlled by FLOWERING LOCUS T (FT) and TERMINAL FLOWERING 1 (TFL1), which are genes that have functionally diverged from a common ancestor MOTHER OF FT AND TFL1 (MFT) [98]. FT promotes heading whereas TFL1 represses heading. In *Arabidopsis* the FT/TFL1 gene family consists of six members: FT, TFL1, MFT, BROTHER OF FT (BFT), CENTRORADIALIS (CEN), and TWIN SISTER OF FT (TSF). They share high sequence similarity but do have different roles in floral transition [102]. Using the *Arabidopsis* FT protein as a query we found perennial ryegrass proteins with sequence similarity to FT/TFL1 family proteins (Figure 2.9). We also identified similar proteins in *Brachypodium* and barley. A phylogenetic analysis using the maximum likelihood method divided the proteins into two distinct groups, one group with the floral inducers FT and TSF and another group with the floral inhibitors TFL1, CEN and BFT (Figure 2.9) [136].

Apart from floral transition, *Arabidopsis* FT also mediates stomatal opening [101]. Likewise, TFL1 is also involved in meristematic development and perennial heading [182]. We identified a putative perennial ryegrass ortholog of FT (ms_13332|ref0029013-gene-0.0mRNA) (Figure 2.9) that was anchored to LG7 at 43.6cM, in a region with markers significantly associated with heading (Figure 2.6). FT was anchored to the same genetic position (43.6cM), as a previously mapped

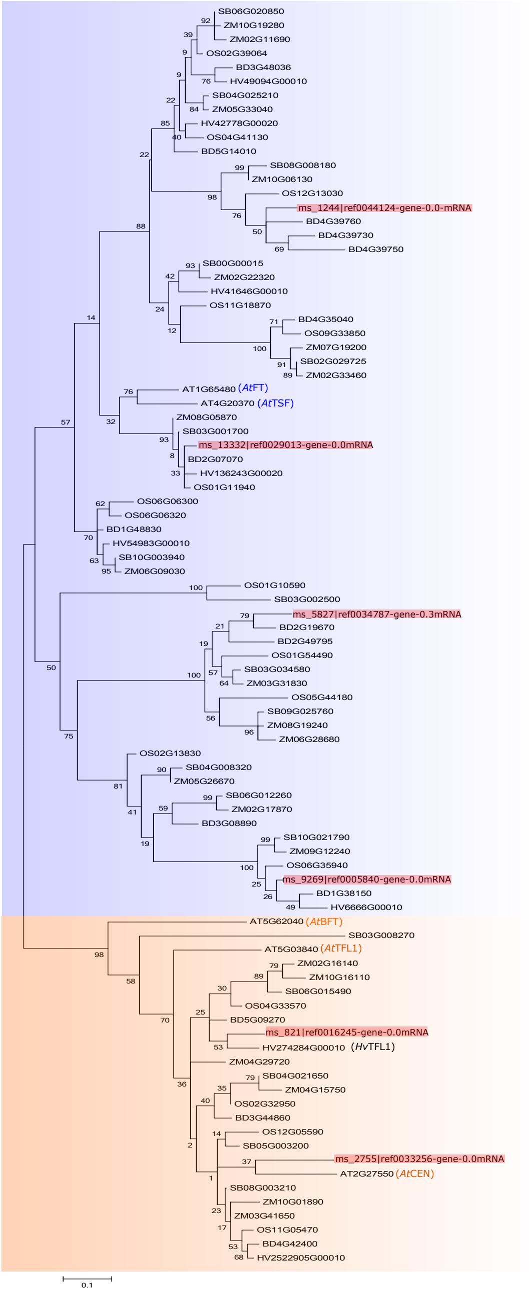

Figure 2.9: Phylogenetic analysis of FT/TFL1 gene family using *Arabidopsis* FT as query. Bootstrap values after 100 replicates were shown next to the branches. The analysis involved 90 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 83 positions in the final dataset. Evolutionary analyses were conducted in MEGA 6.06 [169]. All the associated *Lolium* proteins are in red and *Arabidopsis* proteins were highlighted.

genetic marker, LpVRN3 [137]. LpVRN3 was designed on a sequence that shared 100% identity (alignment length of 80.4%) with the transcript we identified as orthologus to FT. Two genes from the FT/TFL gene family have previously been mapped in perennial ryegrass [166]. Both were mapped in the same experimental population used as the backbone to the GenomeZipper. Using the available genome data we can now better identify the putative perennial ryegrass orthologs to these genes. Based on our phylogenetic anlaysis, the genetic marker previously labeled as LpFT is more likely to be an ortholog of TSF (ms_9269|ref0005840-gene-0.0mRNA) (Figure 2.9). TSF is the closest sequence homologue of FT and they have overlapping roles in promoting heading, however it does have a distinct role to play under short day conditions [193]. TSF was anchored to LG7 at 57.2cM (Figure 2.6) in a region with markers significantly associated with heading date, particularly in family G11.

The *Arabidopsis* floral inhibitors, TFL1, BFT and CEN were grouped in a branch separate to FT. We identified putative perennial ryegrass orthologs of TFL1 (ms_821|ref0016245-gene-0.0mRNA) clustering with barley TFL1 (Figure 2.9) that was anchored on the GenomeZipper to LG2 at 79.8cM (Figure 2.4). Previously a perennial ryegrass gene with sequence homology to TFL1 was anchored to LG5 at 27.5cM using a transcriptome based genetic map [166], however, our phylogentic analysis suggests that this is more likely an ortholog of BFT. In *Arabidopsis* BFT shares highest sequence similarity to TFL1 and functions similar to TFL1 in meristematic development to repress heading [194]. Interestingly, on LG2 markers we identified in single marker analysis for aftermath heading, were located on the GenomeZipper at 80.8cM and 84.2cM. These markers were next to putative perennial ryegrass ortholog of TFL1.

In perennial ryegrass, TFL1 is characterized as a repressor of heading and a regulator of axillary meristem identity [91]. When LpTFL1 was overexpressed in *Arabidopsis*, plants displayed a delayed heading phenotype and extended vegetative growth [91]. In perennial ryegrass expression level of LpTFL1 was observed in leaves, inflorescence, roots, stem and apex. It was found that after a period of cold (primary induction), expression levels of LpTFL1 reduced, allowing plants to prepare for heading. As the day length and temperature increases (secondary induction), LpTFL1 is upregulated in the apex to promote tillering [91]. Unlike annual grasses that flower once in the season and die after seed production, perennial ryegrass continues to grow even after seed production by maintaining at least one tiller in a vegetative phase. It was shown that tillering in ryegrass is mainly controlled by spatiotemporal regulatory mechanism, by activating certain genes to repress heading in vernalized tillers [188]. Interestingly, mutations in homologues of TFL1 in rose (*RoKSN*) and woodland strawberry (*FvKSN*) (both Rosoideae members of the rose family Rosaceae) have been shown to be responsible for continuous heading phenotypes in these species [87]. The putative ortholog of TFL1 identified here, which co-locates with variants for aftermath heading, is an interesting candidate gene for further study of this important forage quality trait.

2.4 Conclusions

In this study we did not detect any SNPs significantly associated with heading and aftermath heading in a genome-wide association analysis, most likely due to the rapid decay of LD we observed in the population and due to the fact that population structure and heading date are confounded. However, using single marker analysis within each full-sib family we did identify linked markers, some in regions containing putative orthologs of key heading genes. Interestingly, in a family segregating for aftermath heading, SNPs were anchored proximal to a putative ortholog of TFL1, homologues of which have recently been shown to play a key role in continuous heading/ aftermath heading of some Rosaceae species [87].

List of abbreviations GWAS: genome wide association study, QTL: quantitative trait loci, LD: linkage disequilibrium, LG: linkage group.

Author's contributions DM, PC, SB, TRH conceived and designed the study. SKA, JV and SLB performed the data analysis. SKA, and SLB drafted the initial paper. SKA, SLB, DM, PC, JV, TRH, and SB contributed to interpretation of data and preparation of the final manuscript. All authors read and approved the final version.

Supplementary files

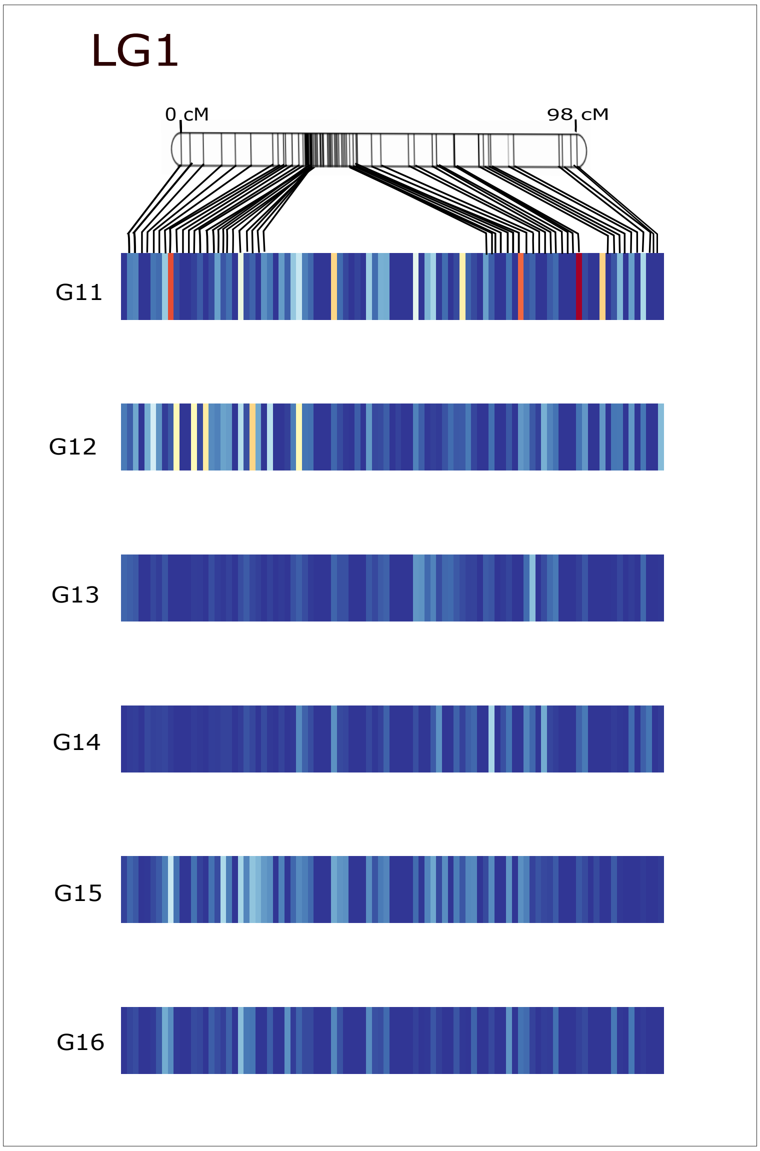

Figure S2.1: Heatmap of perennial ryegrass LG1 over six full-sib families. A Kruskal-Wallis test was performed on each marker to identify significant regions for heading. Using the perennial ryegrass genome zipper [23, 137] we identified a putative gene order for markers on LG1. These data were used to construct the heatmap for each family. A perennial ryegrass transcriptome-based genetic linkage map upon which GenomeZipper was based was used as reference to construct LG1 [137, 166] and placed above the heatmap. Each bar in the heatmap represents region between two genetic markers from the linkage map. The median Kruskal-Wallis test statistic was calculated for markers binned between markers on the genetic linkage map and used to construct the heatmap. Color of the heatmap illustrates the test-statistic of the Kruskal-wallis analysis.

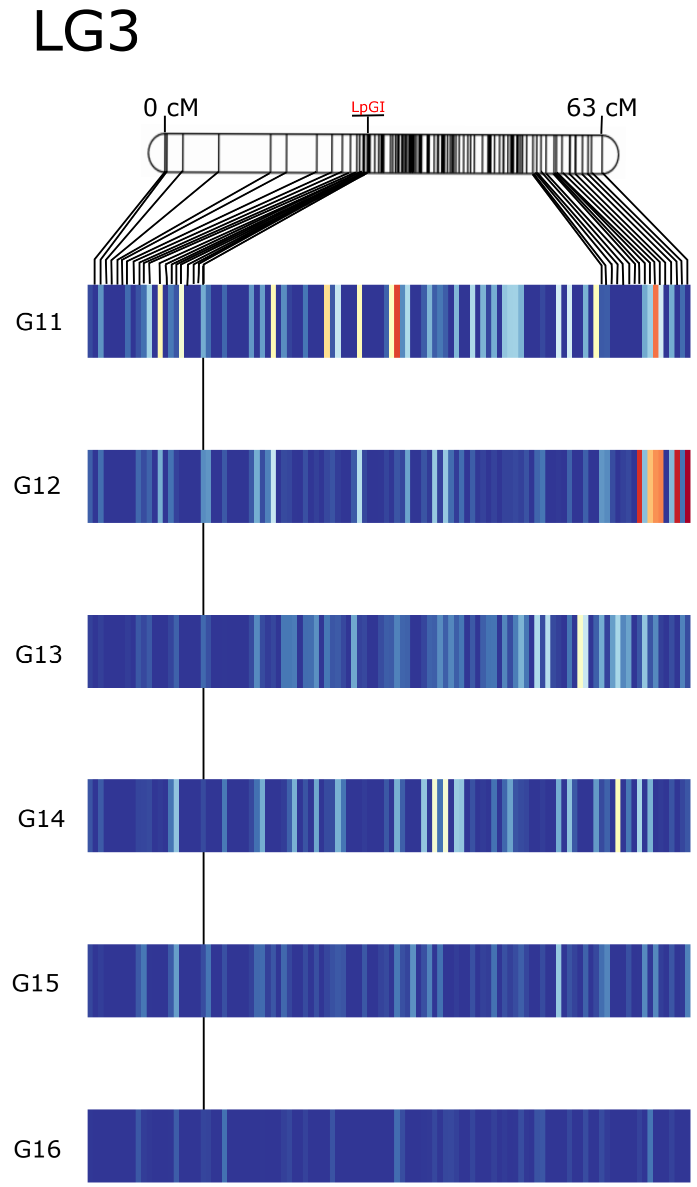

Figure S2.2: Heatmap of perennial ryegrass LG3 over six full-sib families. Similar methodology was used as described in Figure S2.2. Putative ortholog of LpGI, was identified in the phylogenetic analysis and placed onto LG3 using genetic positions from genome zipper. The genetic positions of these orthologs were extrapolated over the heatmap. Color of the heatmap illustrates the test-statistic of the Kruskal-wallis analysis.

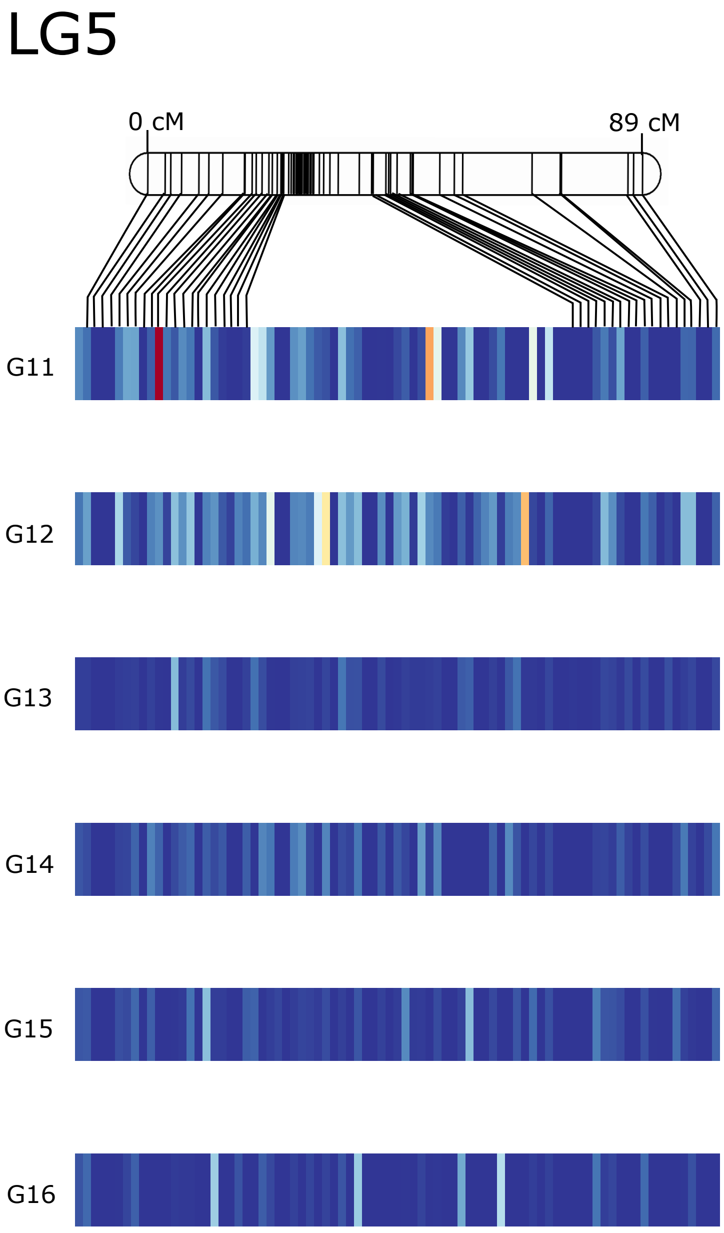

Figure S2.3: Heatmap of perennial ryegrass LG5 over six full-sib families. Similar methodology was used as described in Figure S2.2. Color of the heatmap illustrates the test-statistic of the Kruskal-wallis analysis.

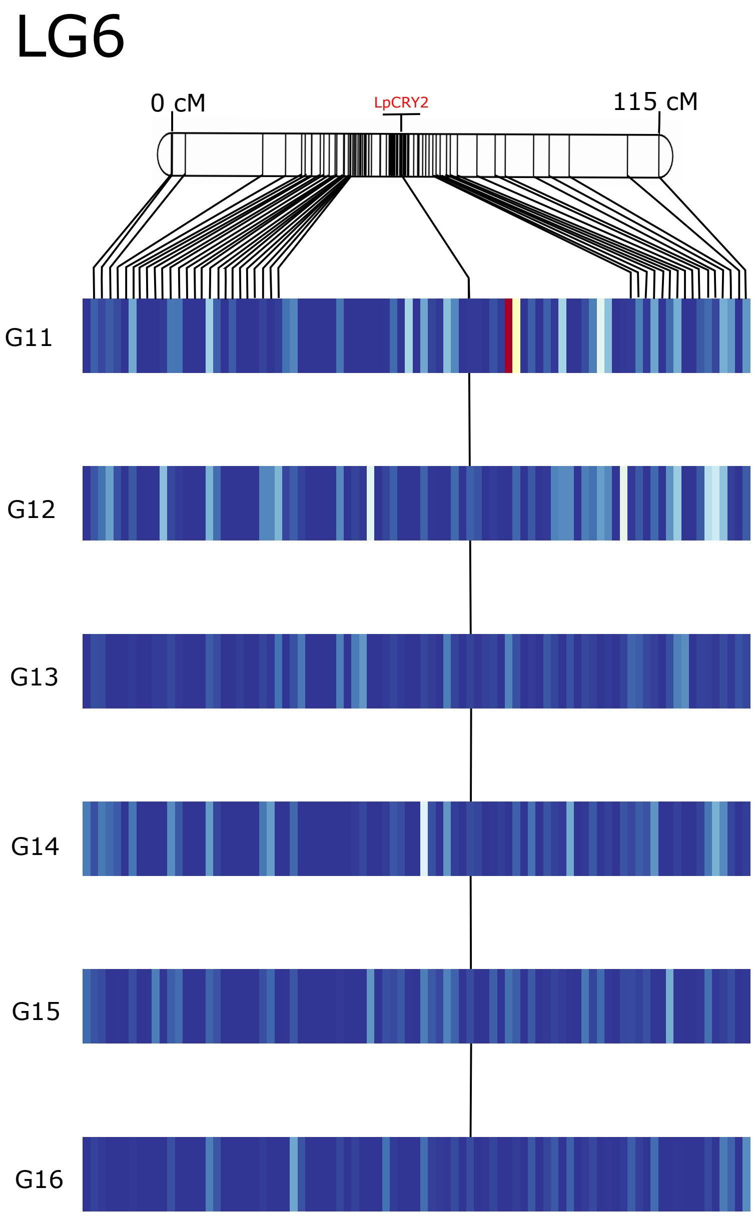

Figure S2.4: Heatmap of perennial ryegrass LG6 over six full-sib families. Similar methodology was used as described in Figure S2.2. Putative ortholog of LpCRY2 was identified in the phylogenetic analysis and placed onto LG6 using genetic positions from genome zipper. The genetic positions of these orthologs were extrapolated over the heatmap. Color of the heatmap illustrates the test-statistic of the Kruskal-wallis analysis.

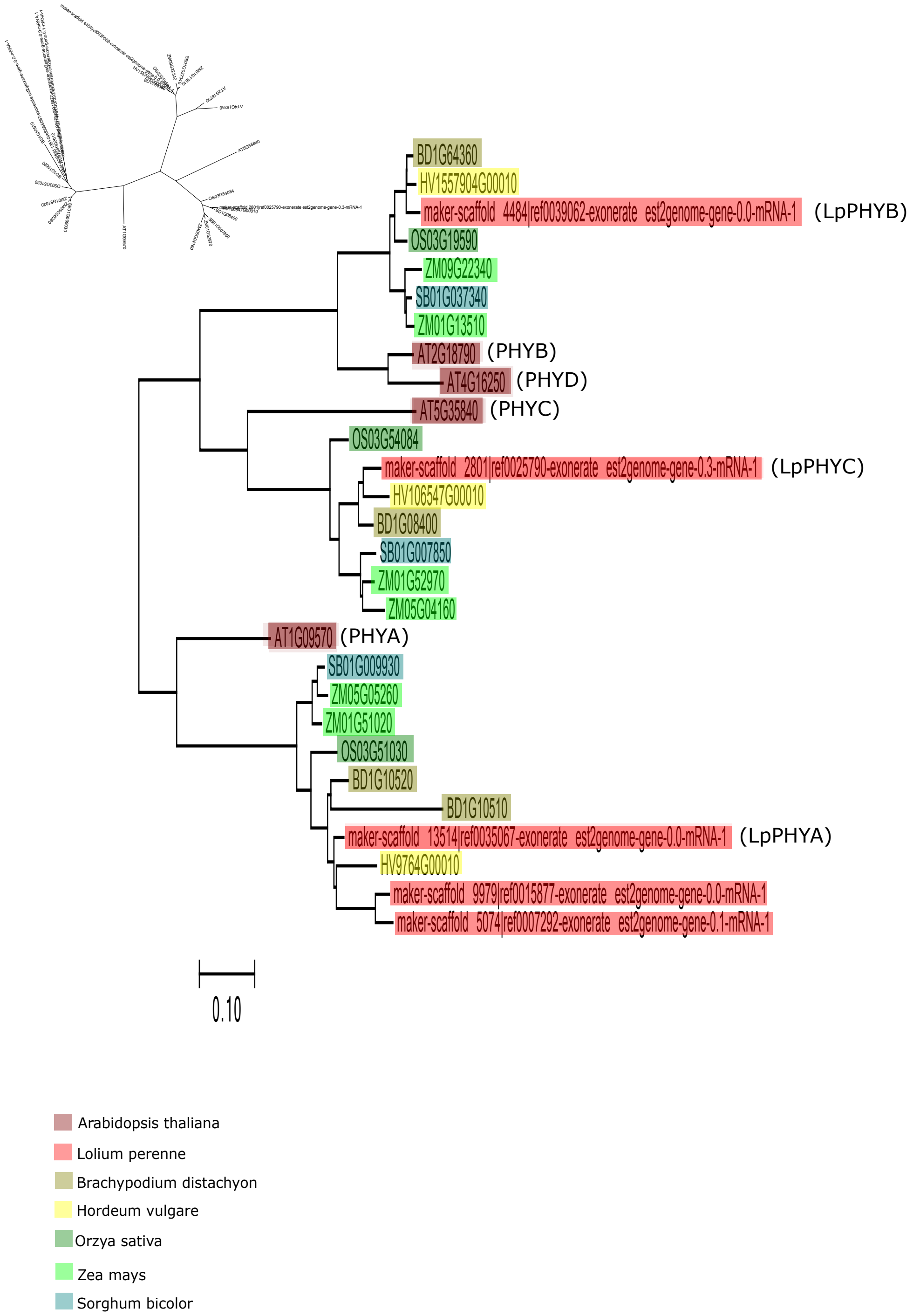

Figure S2.5: Phylogenetic tree of candidate heading genes PHYA, PHYB and PHYC. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [95]. The tree is mid-point rooted, drawn to scale, with branch lengths proportional to the number of substitutions per site. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA 6.06 [169]. All the associated *Lolium* and *Arabidopsis* proteins were highlighted.

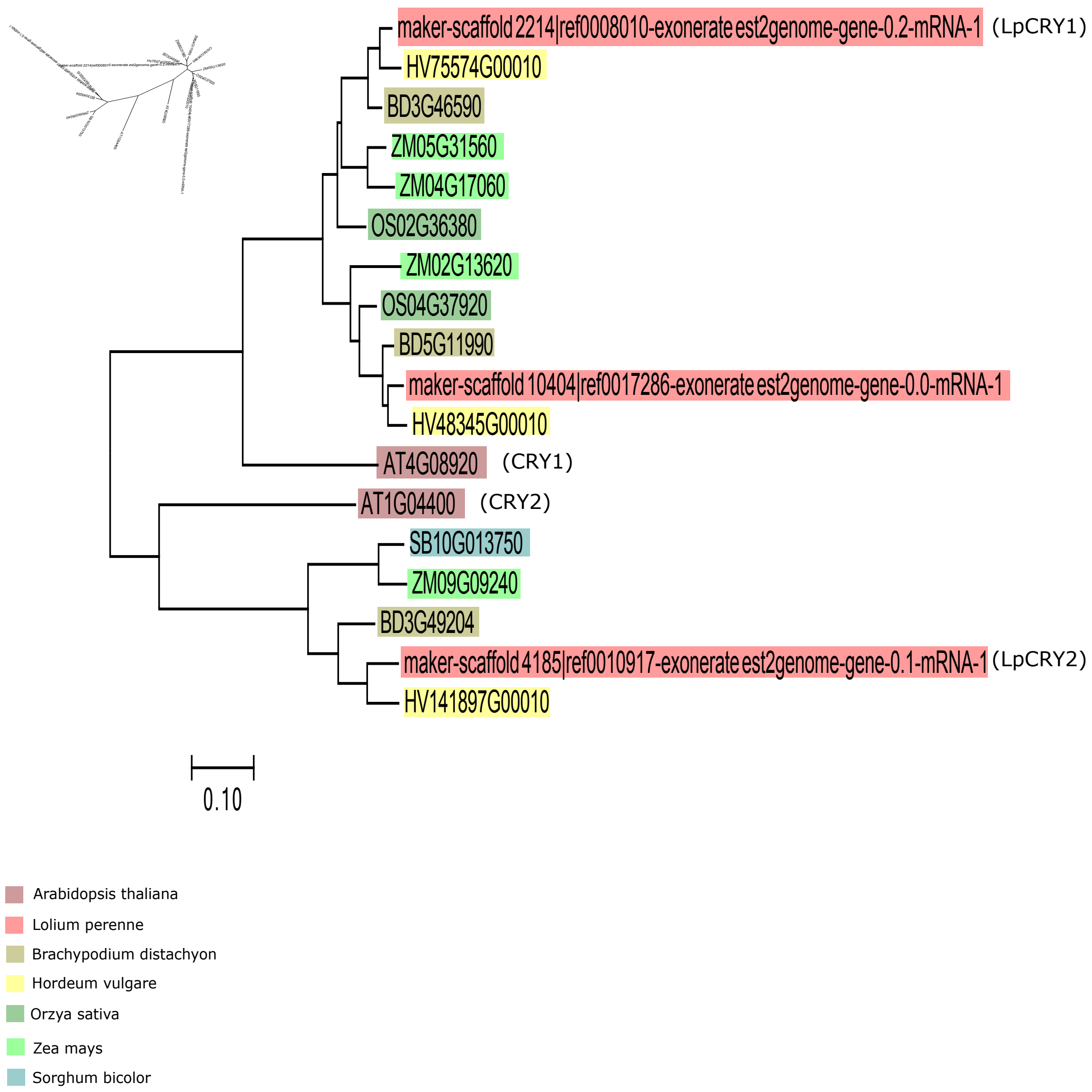

Figure S2.6: Phylogenetic tree of candidate heading gene CRY2. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [95]. The tree is mid-point rooted, drawn to scale, with branch lengths proportional to the number of substitutions per site. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA 6.06 [169]. All the associated *Lolium* and *Arabidopsis* proteins were highlighted.

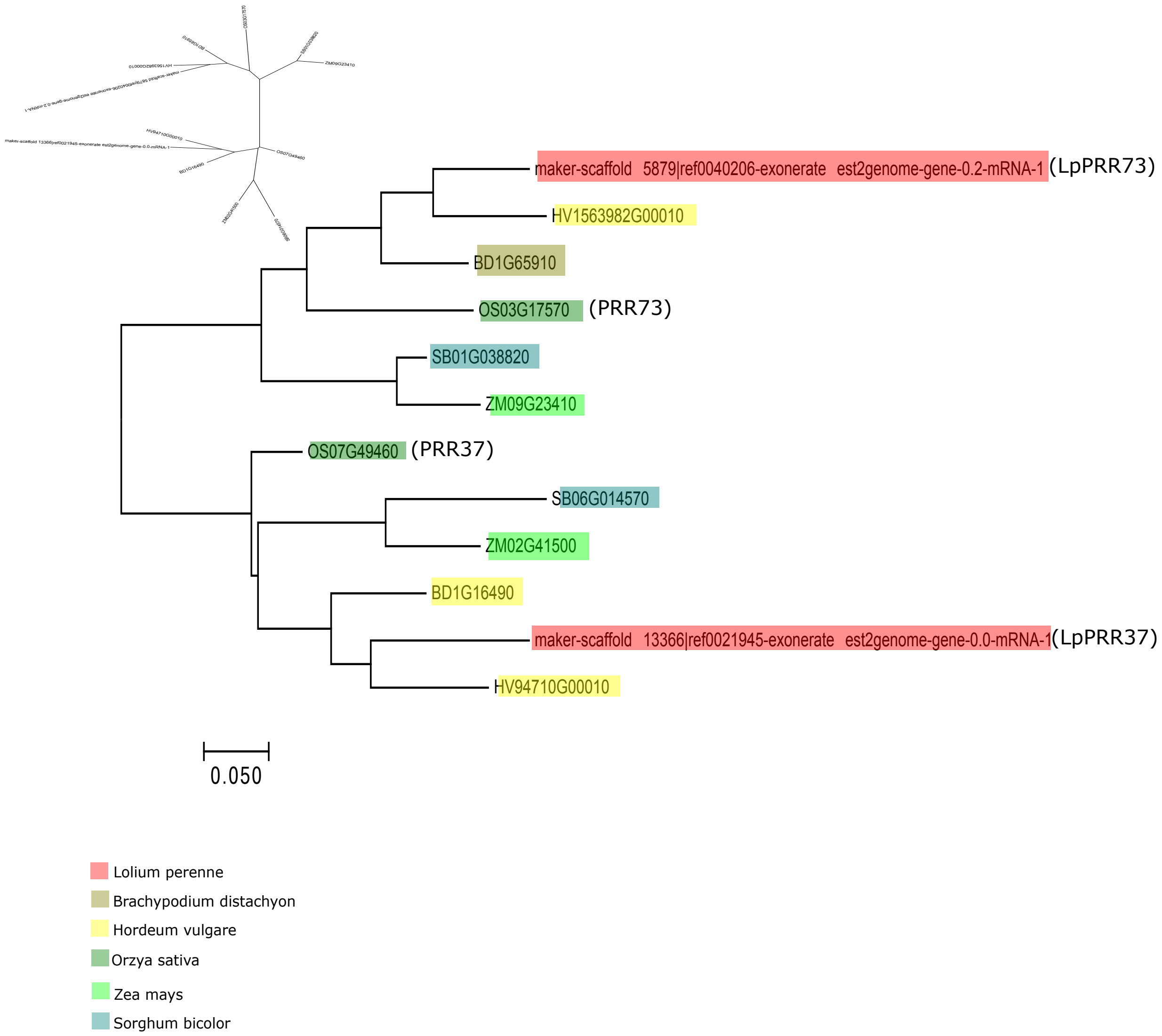

Figure S2.7: Phylogenetic tree of candidate heading gene PRR37. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [95]. The tree is mid-point rooted, drawn to scale, with branch lengths proportional to the number of substitutions per site. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA 6.06 [169]. All the associated *Lolium* and rice proteins were highlighted.

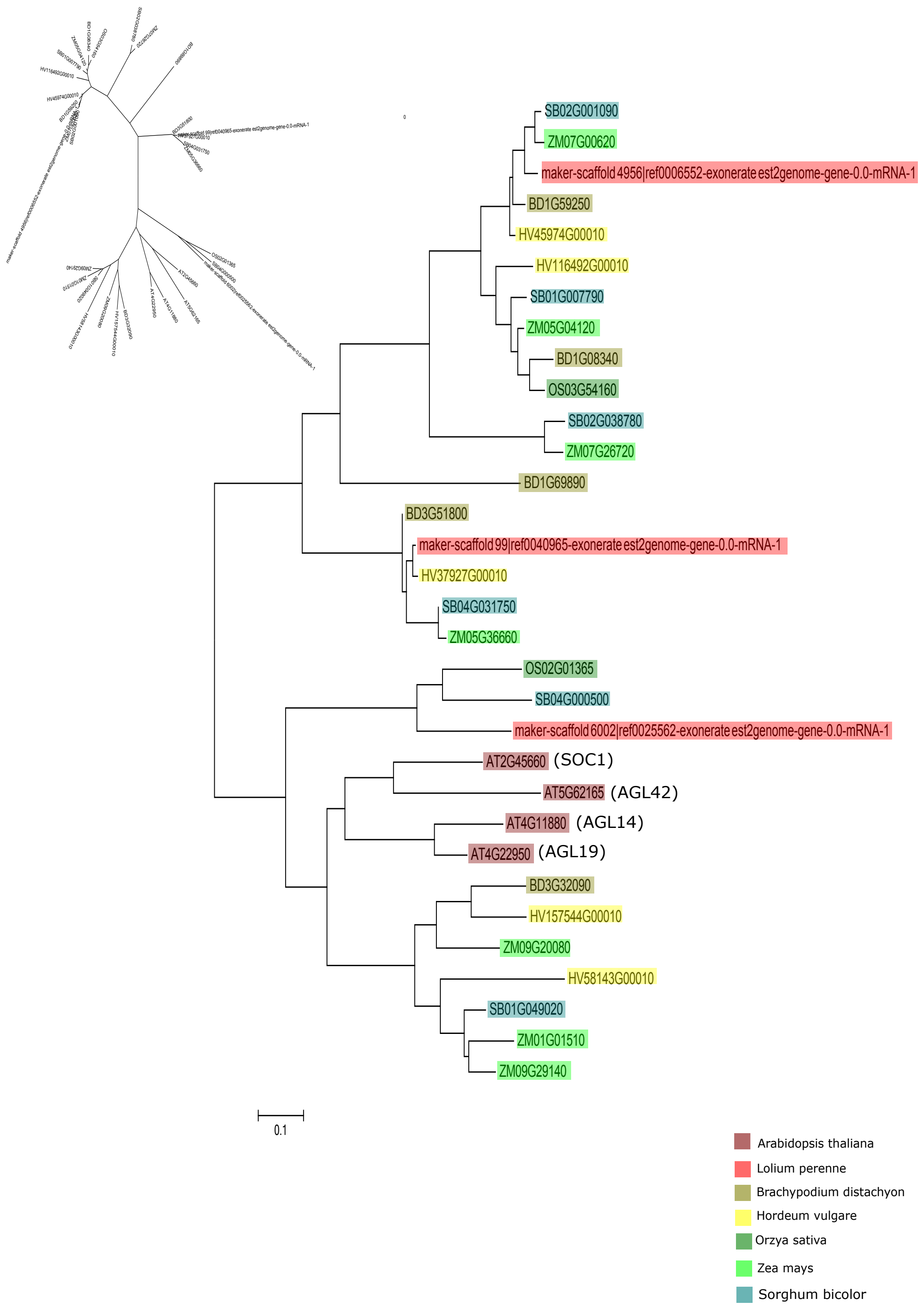

Figure S2.8: Phylogenetic tree of candidate heading gene SOC1. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [95]. The tree is mid-point rooted, drawn to scale, with branch lengths proportional to the number of substitutions per site. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA 6.06 [169]. All the associated *Lolium* and *Arabidopsis* proteins were highlighted.

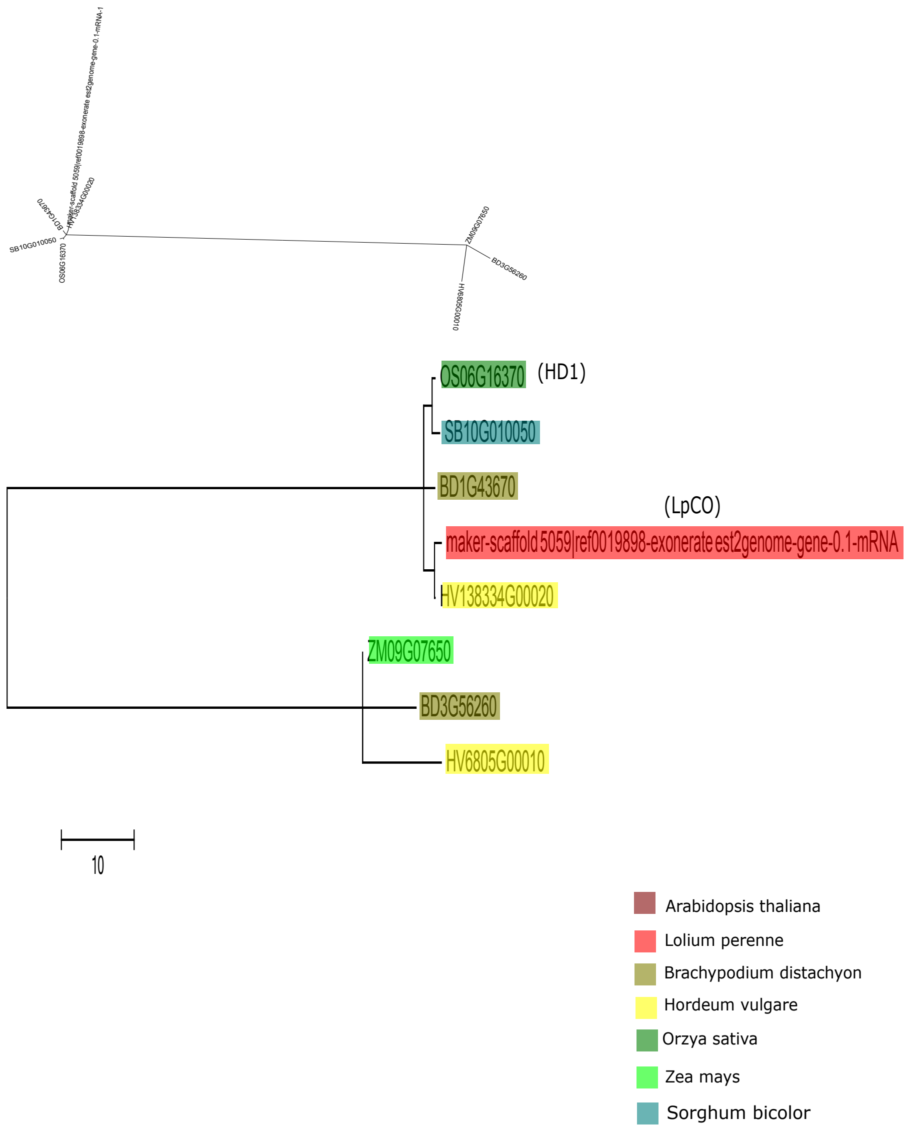

Figure S2.9: Phylogenetic tree of candidate heading gene CO. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [95]. The tree is mid-point rooted, drawn to scale, with branch lengths proportional to the number of substitutions per site. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA 6.06 [169]. Associated *Lolium* and rice proteins were highlighted.

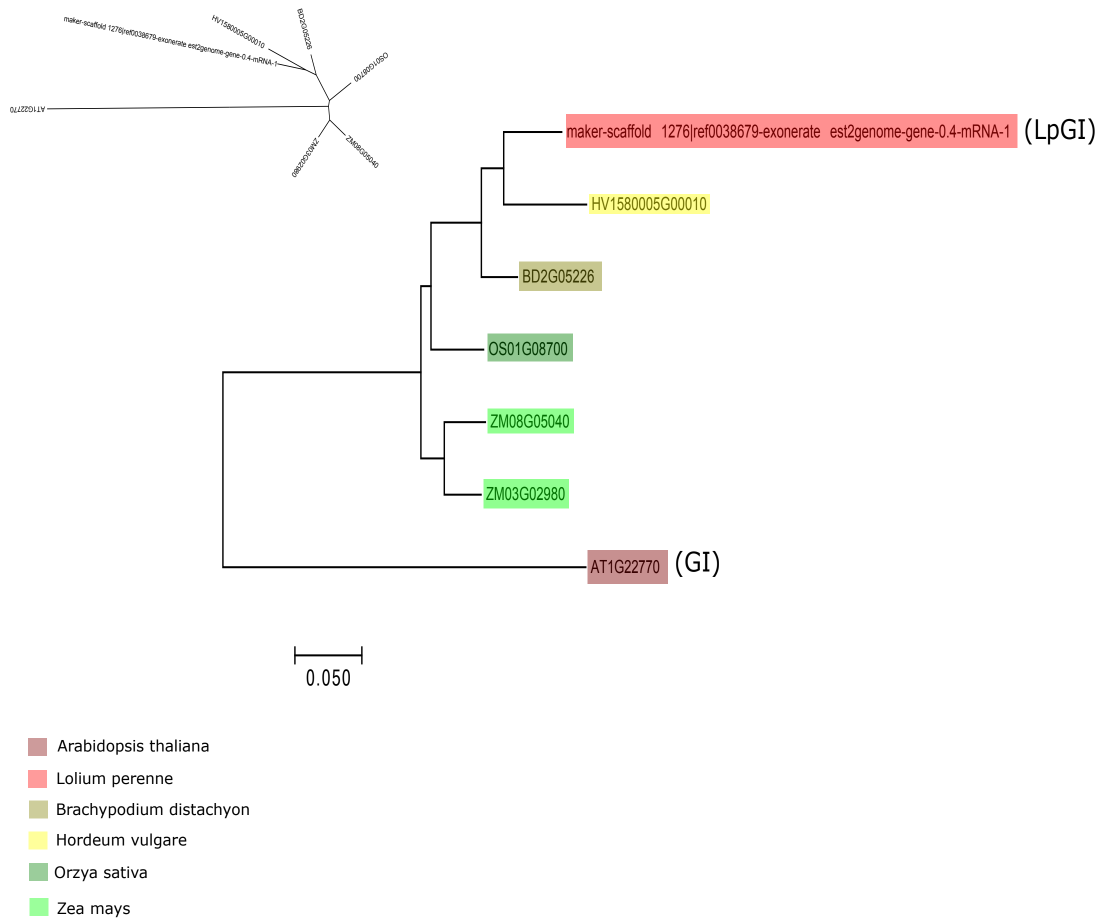

Figure S2.10: Phylogenetic tree of candidate heading gene GI. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [95]. The tree is mid-point rooted, drawn to scale, with branch lengths proportional to the number of substitutions per site. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA 6.06 [169]. Associated *Lolium* and *Arabidopsis* proteins were highlighted.

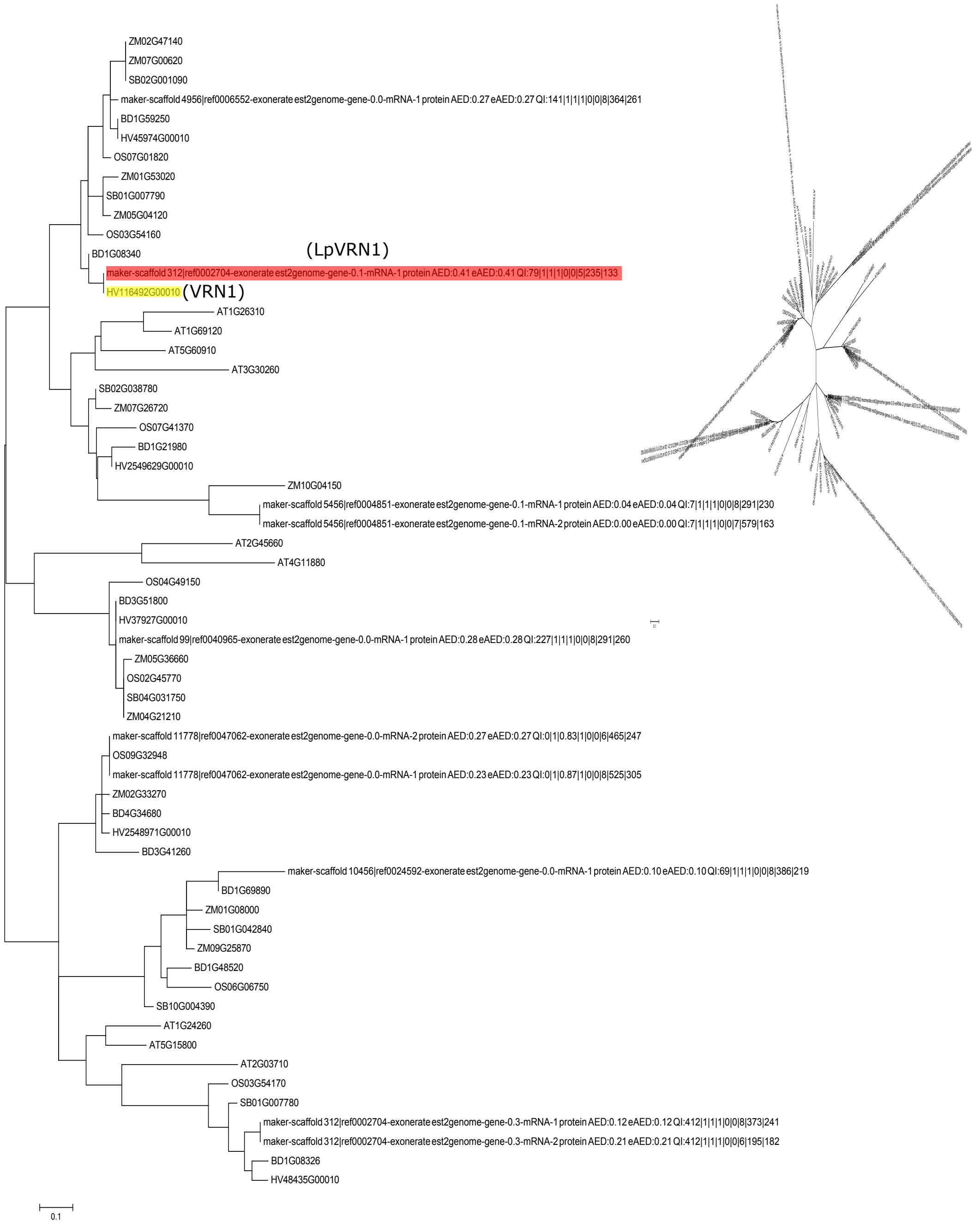
ZM02G47140
ZM07G00620
SB02G001090
maker-scaffold 4956|ref0006552-exonerate est2genome-gene-0.0-mRNA-1 protein AED:0.27 eAED:0.27 QI:141|1|1|1|0|0|8|364|261
BD1G59250
HV45974G00010
OS07G01820
ZM01G53020
SB01G007790
ZM05G04120
OS03G54160
BD1G08340
(LpVRN1)
maker-scaffold 312|ref0002704-exonerate est2genome-gene-0.1-mRNA-1 protein AED:0.41 eAED:0.41 QI:79|1|1|1|0|0|5|235|133
HV116492G00010 (VRN1)
AT1G26310
AT1G69120
AT5G60910
AT3G30260
SB02G038780
ZM07G26720
OS07G41370
BD1G21980
HV2549629G00010
ZM10G04150
maker-scaffold 5456|ref0004851-exonerate est2genome-gene-0.1-mRNA-1 protein AED:0.04 eAED:0.04 QI:7|1|1|1|0|0|8|291|230
maker-scaffold 5456|ref0004851-exonerate est2genome-gene-0.1-mRNA-2 protein AED:0.00 eAED:0.00 QI:7|1|1|1|0|0|7|579|163
AT2G45660
AT4G11880
OS04G49150
BD3G51800
HV37927G00010
maker-scaffold 99|ref0040965-exonerate est2genome-gene-0.0-mRNA-1 protein AED:0.28 eAED:0.28 QI:227|1|1|1|0|0|8|291|260
ZM05G36660
OS02G45770
SB04G031750
ZM04G21210
maker-scaffold 11778|ref0047062-exonerate est2genome-gene-0.0-mRNA-2 protein AED:0.27 eAED:0.27 QI:0|1|0.83|1|0|0|6|465|247
OS09G32948
maker-scaffold 11778|ref0047062-exonerate est2genome-gene-0.0-mRNA-1 protein AED:0.23 eAED:0.23 QI:0|1|0.87|1|0|0|8|525|305
ZM02G33270
BD4G34680
HV2548971G00010
BD3G41260
maker-scaffold 10456|ref0024592-exonerate est2genome-gene-0.0-mRNA-1 protein AED:0.10 eAED:0.10 QI:69|1|1|1|0|0|8|386|219
BD1G69890
ZM01G08000
SB01G042840
ZM09G25870
BD1G48520
OS06G06750
SB10G004390
AT1G24260
AT5G15800
AT2G03710
OS03G54170
SB01G007780
maker-scaffold 312|ref0002704-exonerate est2genome-gene-0.3-mRNA-1 protein AED:0.12 eAED:0.12 QI:412|1|1|1|0|0|8|373|241
maker-scaffold 312|ref0002704-exonerate est2genome-gene-0.3-mRNA-2 protein AED:0.21 eAED:0.21 QI:412|1|1|1|0|0|6|195|182
BD1G08326
HV48435G00010
0.1

Figure S2.11: Phylogenetic tree of candidate heading gene VRN1. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [95]. The tree is mid-point rooted, drawn to scale, with branch lengths proportional to the number of substitutions per site. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA 6.06 [169]. Associated *Lolium* and barley proteins were highlighted.

Chapter 3

Genomic prediction of crown rust resistance in *Lolium perenne*

Sai krishna Arojju[1,5], Patrick Conaghan[2], Susanne Barth[1], Dan Milbourne[1], Michael Casler [3,4], Trevor R Hodkinson[5], Thibauld Michel[1] and Stephen L Byrne[1*]

[1] Teagasc, Crop Science Department, Oak Park, Carlow, Ireland

[2] Teagasc, Grassland Science Research Department, Animal and Grassland Research and Innovation Centre, Oak Park, Carlow, Ireland

[3] Department of Agronomy, University of Wisconsin-Madison, WI53706, Madison, USA

[4] Agricultural Research Service, United State Department of Agriculture, WI53706, Madison, USA

[5] Department of Botany, Trinity College Dublin, Dublin2, Ireland

Submitted to BMC Genetics

Abstract

Genomic selection (GS) can accelerate genetic gains in breeding programmes by reducing the time it takes to complete a cycle of selection. *Puccinia coronata* f. sp *lolli* (crown rust) is one of the most widespread diseases of perennial ryegrass and can lead to reductions in yield, persistency and nutritional value. Here, we have used a large perennial ryegrass population to assess the accuracy of using genome wide markers to predict crown rust resistance and to investigate the factors affecting predictive ability. Using these data, predictive ability for crown rust resistance in the complete population reached a maximum of 0.52. Much of the predictive ability resulted from the ability of markers to capture genetic relationships among families within the training set, and reducing the marker density had little impact on predictive ability. Using permutation based variable importance measure and genome wide association studies (GWAS) to identify and rank markers enabled the identification of a small subset of SNPs that could achieve predictive abilities close to those achieved using the complete marker set. Using a GWAS to identify and rank markers enabled a small panel of markers to be identified that could achieve higher predictive ability than the same number of randomly selected markers, and predictive abilities close to those achieved with the entire marker set. This was particularly evident in a sub-population characterised by having on-average higher genome-wide linkage disequilibirum (LD). Higher predictive abilities with selected markers over random markers indicates they are in LD with QTL for crown rust resistance rather than simply capturing relationships among families. Accuracy due to genetic relationships will decay rapidly over generations whereas accuracy due to LD will persist, which is advantageous for practical breeding applications.

Keywords: genomic selection, crown rust, perennial ryegrass, genetic relationship, GWAS

3.1 Introduction

Perennial ryegrass (*Lolium perenne* L.) is the predominant forage species grown in temperate regions of the world [83]. *Puccinia coronata* f. sp *lolli* (crown rust) is one of the most widespread diseases of perennial ryegrass and can lead to a reduction in forage nutritive value, yield and persistency [140, 141, 177]. Poor quality, rust infected swards can impact animal performance and well-being [106, 138, 164]. Developing resistant cultivars is the most viable option for disease control and it has been shown that resistance to crown rust is conferred by both quantitative and qualitative inheritance [79, 100, 122, 146]. As an obligate out-crossing species, perennial ryegrass germplasm has high variation for disease resistance that can be utilized to develop resistant cultivars [70, 121, 146]. Phenotypic recurrent selection is typically used to develop cultivars with improved resistance and selection is often carried out on spaced plants [70, 100, 146, 185]. There is a high correlation between spaced plants and swards for disease resistance and indirect selection for disease resistance on spaced plants can improve resistance in sward conditions [54]. However, with the advancements in molecular marker development over the last decade, efforts to use marker assisted breeding strategies have been pursued. One such strategy involves identifying quantitative trait loci (QTL) in bi-parental mapping populations and using markers to efficiently backcross the QTL into elite breeding material [32]. Although QTLs explaining significant phenotypic variation for crown rust resistance were mapped onto linkage group (LG) 1-5 and 7 [51–53, 129, 130, 157, 174], it is unclear if any of these QTLs were successfully introduced into breeding material. Genome wide association studies (GWAS) are another approach to identify markers linked to QTL. In this case breeding populations can directly be used to identify marker-trait associations, although identified markers tended to explain a small proportion of the total additive genetic variance, resulting in smaller genetic gains [75, 81, 89].

Genomic selection (GS) was first proposed by Meuwissen et al. [123], as a method to capture complete additive genetic variance using genome wide markers. GS is a form of marker assisted breeding, which accounts for all marker effects across the entire genome to calculate genomic estimated breeding values (GEBVs), which are used to select individual plants for advancement [89]. Use of genome-wide markers will include small effect loci and is ideal for complex traits with low to moderate heritability. In GS, a training population is genotyped with genome wide markers and phenotyped for the trait under selection and models to predict breeding values from marker data are developed. Implementing GS for complex traits like yield and quality is a primary objective of many perennial ryegrass breeding programmes. In contrast to yield and quality traits, the cost (labour and time) of phenotyping for disease resistance is much lower. However, it is important that any GS approaches targeting yield and quality improvements also ensure adequate disease resistance is maintained, particularly where multiple rounds of marker based selections are performed between field evaluations. Opportunities for GS in perennial ryegrass were first reviewed by Hayes et al. [77], and the earliest empirical study was done by Fè et al. [61] for heading date, which confirmed the superiority of GS over marker assisted selection. Later Fè et al. [60], Grinberg et al. [72] and Byrne et al. [21] reported high predictive ability for important agronomical traits in perennial ryegrass. In particular, predictive ability for crown rust reached up to 0.58 [60] when genotypes and phenotypes were evaluated on F_2 families. In this study, we evaluated predictive ability for crown rust resistance on individual plants in a large perennial ryegrass population, and assessed factors contributing to predictive ability, such as training population size and marker density. We also performed a GWAS to identify a small to moderately sized panel of markers with good predictive ability for crown rust resistance.

3.2 Material and methods

3.2.1 Plant material, phenotyping and genotyping

The training population consists of 30 diploid perennial ryegrass families that have been described previously Byrne et al. [21]. Each family consists of 60 genotypes making up a population of 1800 individuals. The complete population consists of ten cultivars, eight full-sib families, eight half-sib families and four ecotypes. Plants were established in a glasshouse and later transplanted to the field in 2013 at Oak Park, Carlow, Ireland (52° $51'$ $34.2"N$ 6° $55'$ $03.0"W$). Plants were grown in two replicates in a partially balanced incomplete block design. Each block consists of 60 test genotypes and 5 check genotypes and was surrounded by a 1.5m sward consisting of a four way mix of crown rust susceptible perennial ryegrass cultivars.

Crown rust was recorded in the years 2014 and 2015 as mean percentage disease score on each plant. Briefly, percentage disease score was obtained by combining scores of percentage of leaves with infection and average percentage of infection on diseased leaves. Scoring was done at multiple time points in September to November without any harvest cuts between scorings (Table 3.1). We are trying to develop genomic models to identify plants with good resistance to crown rust across the season, and we decided to use all time points for constructing a quantitative summary for crown rust resistance. To do this we calculated AUDPC for each spaced plant in both years. Using multiple time point data, we generated area under disease progress curve (AUDPC) as follows:

$$A_k = \sum_{i=1}^{N_i-1} \frac{(y_i + y_{i+1})}{2}(t_{i+1} - t_i) \tag{3.1}$$

where y_i is the extent of infection (percentage disease score) at i^{th} observation and t_i is the time point at i^{th} observation. N_i is the number of individuals in the data set.

Variance components for crown rust were estimated using an R package called lme4 (linear mixed-effects models using 'eigen' and S4) [10]. Broad sense heritability was estimated as follows:

$$H^2 = \frac{\sigma_g^2}{(\sigma_g^2) + (\sigma_{g*yr}^2)/2 + (\sigma_{res}^2)/4} \tag{3.2}$$

where σ_g^2 is the total genetic variance among individuals, σ_{g*yr}^2 is the variance associated with genotype by year interaction and σ_{res}^2 is residual variance. With genotype and block within replicate as random effects and year and checks as fixed effects, conditional modes (BLUPs) were calculated in lme4 and used as input for genomic prediction.

Genotyping was done using genotyping by sequencing (GBS) approach described by Elshire et al. [57]. and detailed in Byrne et al. [21]. Briefly, genomic DNA was extracted from leaf samples and GBS libraries were prepared using the restriction enzyme ApeKI, libraries were amplified and sequenced on an Illumina Hiseq2000. Panels of SNPs were identified in the complete population, as well as in all sub populations separately (half-sibs, full-sibs, ecotypes, cultivars). Individuals with missing marker information and phenotypic data were eliminated from the analysis giving a final population for analysis of 1582 individuals.

3.2.2 Genomic prediction models

We used two statistical models for genomic prediction, ridge regression best linear unbiased prediction (rrBLUP) [123] and random forest [17]. rrBLUP is a mixed model approach, which was initially proposed for GS. We used an R package called rrBLUP [58] for fitting the mixed model as follows

$$y = \mu + Xg + \epsilon \tag{3.3}$$

where μ is the overall mean, X is the marker matrix, g is the matrix of marker effects, ϵ is a vector of residual effects and y is a vector of conditional modes for crown rust. Random forest is a machine-learning tool, in which series of regression trees were grown independently to the largest extent possible using subsets of bootstrap samples. At each split of the tree, a random subset of variables is selected to identify the best split. We implemented random forest using the R package randomForest [110], setting the number of variables at each split to 1/3 of the total variables, and using a terminal node size of five and minimum of 500 trees per forest. We also used random forest to rank variables using the variable importance measure. Its a permutation based measure in which variables are ranked based on the mean decrease in accuracy. The top 100 selected variables are used for the model developed on the training set using rrBLUP and predicted in the test set.

3.2.3 Cross validation scheme

We evaluated genomic prediction models using Monte-Carlo cross-validation by randomly assigning plants into training (70%) and test (30%) sets. Predictive ability and bias were assessed in the complete population and in each sub-population. Predictive ability (r_p) was determined as the Pearson's correlation coefficient between observed phenotypic value and predicted phenotype over 100 iterations. Bias was evaluated by regressing observed phenotypic value on predictions. We reduced training population size and marker density in order to identify the impact of training population size and marker number on predictive ability. To compare predictive ability for traits with contrasting genetic architecture we compared heading date, a highly heritable trait, with crown rust. Predictive ability for heading date has already been shown to be high (0.81) in this population [21]. We re-analyzed data for heading date according to methods described above and made a comparison with crown rust. To evaluate the impact of leaving related material out of the training set we also performed cross validation by leaving one family out. In this approach one complete family (up to 60 individuals) is left out of the training set and only used for testing. This was repeated so that each family in turn is used as a test set.

3.2.4 Genome wide association

A mixed linear model (MLM) was also used for association mapping, implemented in the R package rrBLUP [58]. Population structure and family relatedness was accounted for in the mixed model using principal component analysis and a kinship matrix calculated by rrBLUP from the input genotypic data. We accounted for multiple testing using false discovery rate (FDR) and markers passing an α level 0.05 threshold were considered statistically significant.

3.3 Results and discussion

3.3.1 Phenotypic analysis for crown rust

The mean percentage disease score for crown rust infection in the population increased over time in both evaluation years as infection levels accumulated (Table 3.1). In both years, evaluations were carried out in the period from September to November during a time when disease pressure tends to be at its greatest [54, 159]. The highest mean percentage disease score was seen in late October 2015 and was more than double the highest mean percentage disease score from 2014 (Table 3.1).

In addition to plant health and level of host resistance, crown rust infection is influenced by various environmental factors, such as temperature, relative humidity, and

Table 3.1: Mean percentage disease score for crown rust resistance at different time points (TP) in Year1 (2014) and Year2 (2015).

Time point/Dates	Mean	SD	min	max
Year 1				
TP1 (13/10/14)	3.1	6.1	0	40
TP2 (20/10/14)	5.2	7.6	0	45
TP3 (29/10/14)	9.6	10.8	0	60
TP4 (10/11/14)	9.8	8.7	0	45
Year 2				
TP1 (21/09/15)	2.0	4.4	0	32
TP2 (05/10/15)	11.2	10.0	0	60
TP3 (19/10/15)	19.9	9.0	0	63

light [149, 150, 156]. The latency period is reduced and spore production increased as temperature increases [149], and it has been shown that when temperatures exceed 25°C, the susceptibility of previously resistant cultivars can be increased [150]. It has already been shown that there is variability within pathogen populations, and different races can be found within and between locations. It is also possible that the composition of a pathogen population can change over short periods of time and plants that are resistant at one point in time will become susceptible as the pathogen population shifts or evolves.

AUDPC values ranged from 0 to 1371 and the Pearson correlation co-efficient between replicates within years was moderate (0.69 in 2014 and 0.59 in 2015). However, the Pearson correlation co-efficient between years was low (0.28), and there was a significant genotype by year interaction ($F_{(1761)} = 3.025$, $MSE = 60676$, $p = 0.0001$). The broad sense heritability for crown rust infection was moderate (0.36), which is in line with previous estimates of heritability calculated in other populations [62, 146]. Overall there is a good phenotypic variation for crown rust infection among and within the 30 families/cultivars/ecotypes making up the entire population (Figure 3.1). Plants were placed into one of four categories (sub-populations) based on mating type or origin, these were (i) full-sib families, (ii) half-sib families, (iii) cultivars, and (iv) ecotypes. In general the ecotypes were more susceptible to crown rust infection than cultivars or breeding material (Figure 3.1), which presents a challenge for the incorporation of ecotypes into breeding programmes. The broad-sense heritability calculated in each sub-population varied between 0.17 in the cultivars to 0.44 in the full-sib families.

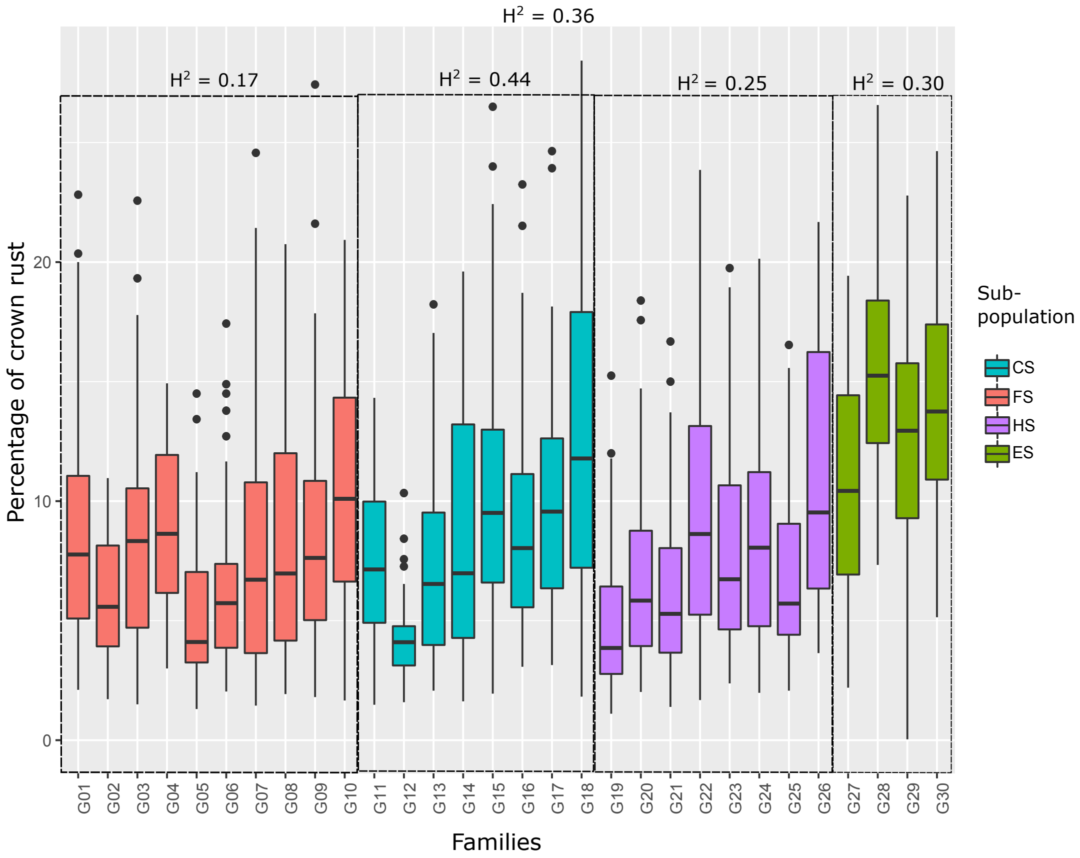

Figure 3.1: Phenotypic variation for crown rust resistance in complete population, grouped according to sub-population types: cultivars (CS), ecotypes (ES), full-sibs (FS) and half-sibs (HS). Broad sense heritability (H^2) in complete population and sub-populations is highlighted over the figure.

Crown rust infection is typically evaluated in breeding programmes by growing spaced plants or potted plants from a population and visually scoring the level of crown rust infection. A mean score is assigned to each family and used to aid selection of the top performing families from which to construct the synthetic cultivars. During construction of synthetics a spaced plant nursery may be established to evaluate heading date and crown rust resistance before selecting individual genotypes from which to construct synthetics (within family selection). In practice, this has a time cost of 2 to 3 years (establishment, evaluation, selection and recombining), and using molecular markers offers an opportunity to reduce this to one year in those selection cycles where GEBVs are predicted. This depends on our ability to accurately predict traits such as crown rust from genomic data.

3.3.2 Predicting crown rust resistance with genomic data

We evaluated two algorithms for prediction of crown rust infection from genomic data, rrBLUP and random forest. The mean predictive ability after cross-validation within the complete population was 0.52 using rrBLUP and 0.49 using random forest (Figure S3.1). rrBLUP was computationally faster and consistently gave higher predictive abilities with lower bias, and therefore results from all further analysis are only reported for rrBLUP. The predictive ability of 0.52 is in line with previous estimates reported in perennial ryegrass where predictions were based on mean genotypes and phenotypes of F_2 families [60]. Using the broad sense heritability of 0.36 as an upper limit on predictive ability, the accuracy of prediction is 0.87. Predictive ability did not differ depending on whether the equations were developed using phenotypes from the last time point scored or the AUDPC values incorporating all time points. This indicates that a single scoring each year would have sufficed. However, the importance of evaluating crown rust in more than one year was emphasised by the low correlation between scores in 2014 and 2015.

When we calculate the predictive ability within each of the sub-populations (cultivars, half-sib families, full-sib families, and ecotypes), the highest predictive ability for crown rust was obtained using plants from full-sib families (0.54) and the lowest predictive ability for crown rust was obtained with the plants from the ecotypes (0.24) (Figure 3.2). Generally, traits with higher heritability achieve higher predictive abilities [128, 189], and we see that here where crown rust measurements taken in the full-sib families had the highest broad-sense heritability and the highest predictive ability. In general, there was a good correlation between predictive ability and both phenotypic variance and heritability. This relationship between phenotypic variance and predictive ability has been observed previously [86, 128].

We also evaluated the predictive ability using a leave-one-family-out cross validation scheme. The complete population is comprised of 30 families/cultivars/ecotypes, each with up to 60 individual genotypes. The predictive ability was assessed in the complete population by selectively leaving one family out of the training set and using it for testing. In addition to crown rust we also evaluated predictive ability for heading date phenotypes previously reported [21]. The predictive ability for both

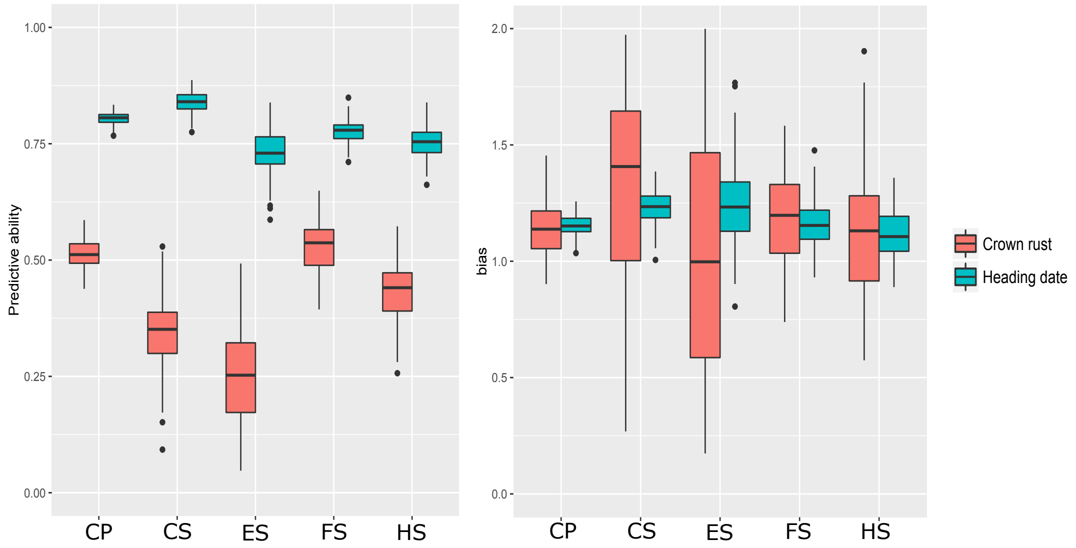

Figure 3.2: Predictive ability in different population types. Complete population (CP), cultivars (CS), ecotypes (ES), full-sibs (FS) and half-sibs (HS) are listed on x-axis, predictive ability (left) and bias (right) on y-axis. Crown rust is in red and heading date in blue.

crown rust (r_p=0.02, min=-0.36, max=0.36) and heading date (r_p=0.29, min=-0.14, max=0.65) varied greatly depending on which family was left out, and having related material in the training set (shared parentage) greatly improved predictive ability.

3.3.3 Effect of training population size and marker density on predictive ability

As we reduced the number of individuals in the training population we saw a decrease in predictive ability and an increasingly upward bias in the variance of predictions for both crown rust resistance and heading date (Figure 3.3). The drop in predictive ability was more pronounced as we reduced the training population size for crown rust resistance than it was for heading date.

The predictive ability for crown rust resistance when using 90% of the population as a training set was 0.52 and the predictive ability was 0.38 when using just 10% of the population. Irrespective of the trait, as the training population size increased there was an increase in predictive ability which is consistent with similar correlations between training population size and predictive ability reported previously

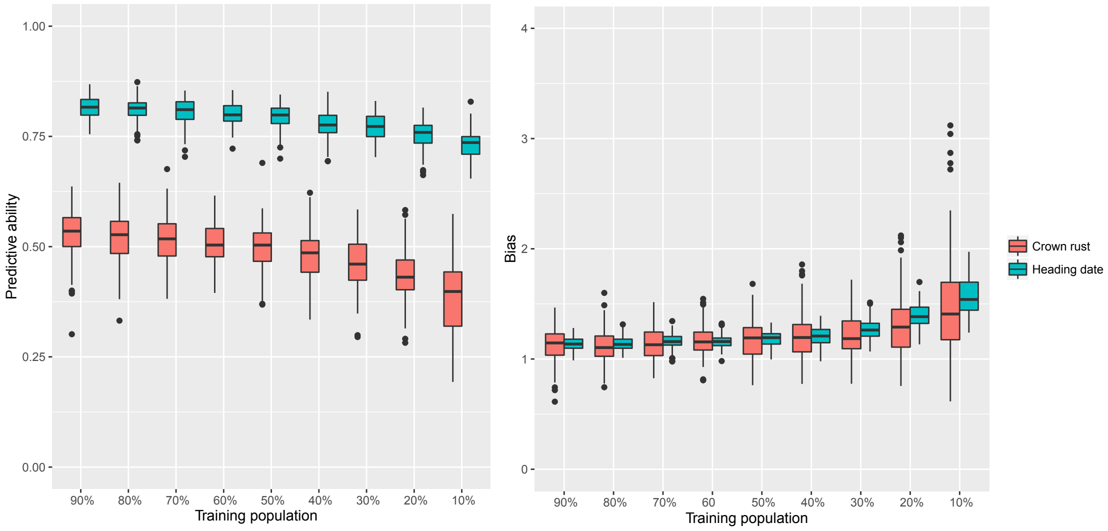

Figure 3.3: Effect of training population size on predictive ability. Training population is varied from 90% (1423 individuals) to 10% (158 individuals) on x-axis and predictive ability (left), bias (right) on y-axis. Crown rust is in red and heading date in blue.

for perennial ryegrass [60, 61] and other crops [86, 113, 148, 171]. Useful linkage disequilibrium (LD) only extends over short distances in perennial ryegrass and it has been suggested that this is the result of a very large past effective population size, which is likely larger than that of humans [77]. This impacts both the size of the reference population and marker density required to achieve high accuracies when predicting traits from genomic data. The fact that we are able to achieve high predictive abilities with relatively small training populations is likely a result of strong genetic structure and differentiation in our diverse population and the use of the marker data to capture genetic relationships [73].

The limited LD also affects the number of markers required to obtain high predictive accuracies, and given the extent of LD in the broader perennial ryegrass population, marker numbers in excess of one million have been suggested for achieving high accuracies [77]. When we reduced marker number in the complete population and the various sub-populations we observed very little impact on the predictive ability for either trait (Table 3.2). But due to limited number of individuals and markers in ecotypes, the effect of marker density on predictive ability was not assessed in this sub-population. Reducing the marker set to 5% of the total available had virtually no impact on predictive ability in all cases. This would support our observation that much of the predictive ability can be derived from markers capturing genetic relationships. When marker number dropped below 5% (10878) predictive ability for both traits in the complete population began to drop. However, even with 0.05% (109) of markers the mean predictive ability was 0.30 for crown rust resistance and 0.52 for heading date. Knowing the contribution of genetic relationships to predictive ability is important because it will change over generations. In contrast, predictive ability due to LD has greater persistence over generations and is therefore preferential [73]. Schemes for implementing genomic selection in perennial ryegrass that pursue a reduction in effective population size from the outset have been proposed. Such schemes would lead to an increase in the extent of LD and ensure that predictive ability due to LD can be captured using a reasonable number of markers and a reference population size that is feasible in breeding programmes.

3.3.4 Identifying SNPs associated with crown rust resistance

The cost of genotyping impacts the number of selection candidates that can be evaluated and therefore impacts the selection intensity. Different approaches to low density SNP genotyping for genomic selection have been proposed. These include variable selection methods to identify a small subset of markers in strong LD with the trait [195] or using a small random subset of markers to impute from low-to-high density [74]. Until a chromosome scale assembly of the perennial ryegrass genome becomes available the latter remains a challenge. We used both permutation based variable importance measures and GWAS analysis to identify a subset of markers capable of predicting crown rust resistance. Using permutation based variable importance measures we were able to rank markers by mean decrease in accuracy and select the top ranked markers for use in genomic prediction. In the case of GWAS we ranked SNPs based on significance and again selected the top ranked markers for use in genomic prediction. All variable importance measures and GWAS were identified and ranked in the training set and used to predict phenotypes in the test set via cross-validation. When we used the top 100 ranked markers from the permutation based variable importance measures, the mean predictive ability of 100 iterations was 0.42 (ranging from 0.36 to 0.48). When we used the top 100 ranked markers from the GWAS analysis, the mean predictive ability of 100 iterations was 0.36 (ranging from 0.25 to 0.44). In both cases the mean predictive ability with selected markers is higher than the predictive ability with random markers, which was 0.28 (ranging from 0.18 to 0.39). The lower predictive ability using GWAS marker

Table 3.2: Predictive ability (r_p) and bias for crown rust (CR) and heading date (HD) by selecting random markers of 100% to 0.05%, in complete population (CP), cultivars (CS), full-sibs (FS) and half-sibs (HS).

Pop	100%		60%		20%		5%		1%		0.5%		0.1%		0.05%	
	r_p	*bias*	r_p	*bias*	r_p	*bias*	r_p	*bias*	r_p	*bias*	r_p	*bias*	r_p	*bias*	r_p	*bias*
CR																
CP	0.52	1.22	0.52	1.22	0.52	1.21	0.51	1.18	0.46	1.10	0.43	1.07	0.36	1.04	0.30	1.48
CS	0.29	1.28	0.28	1.26	0.28	1.24	0.27	1.18	0.22	0.97	0.17	0.80	0.14	0.95	0.10	1.13
FS	0.54	1.13	0.54	1.13	0.54	1.13	0.54	1.14	0.52	1.07	0.50	1.03	0.45	1.00	0.40	0.99
HS	0.49	1.24	0.49	1.24	0.49	1.24	0.49	1.24	0.48	1.23	0.46	1.21	0.42	1.23	0.36	1.22
HD																
CP	0.81	1.16	0.81	1.16	0.81	1.16	0.80	1.14	0.75	1.07	0.72	1.05	0.62	1.01	0.52	1.00
CS	0.84	1.25	0.81	1.19	0.81	1.20	0.81	1.18	0.78	1.11	0.77	1.12	0.66	1.03	0.56	1.02
FS	0.76	1.00	0.75	1.16	0.75	1.16	0.75	1.16	0.74	1.14	0.68	1.27	0.64	1.26	0.62	0.87
HS	0.74	1.18	0.74	1.09	0.74	1.10	0.74	1.09	0.73	1.08	0.72	1.15	0.67	1.10	0.62	1.09

selection is not surprising considering that we corrected for population structure using a kinship matrix, and we are more reliant on identifying markers in LD with the trait. As discussed above, the predictive ability of these markers is expected to be more persistent over subsequent generations. Using GWAS selected markers it is clear to see that they are superior to randomly selected markers up to the point, beyond which adding more markers does not improve predictive ability in either case (Figure 3.4). The ability of a GWAS within each sub-population to identify and select a small set of SNPs with excellent predictive ability varied, and in some cases was little better than random SNP selection (Figure 3.5). The GWAS on plants originating from IBERs bred cultivars identified a small set of twenty SNPs with 77% of the predictive ability achieved with 20,000 SNPs. The power of a GWAS to identify markers with high predictive ability was much greater within the population made up of IBERs plants than within cultivars, and full-sib families where twenty SNPs could only achieve 46% and 48% of the predictive ability with 20,000 SNPs, respectively. On average LD is higher within the sub-population with IBERs plants, which may explain the greater ability to identify markers associated with crown-rust resistance.

In order to characterise the markers associated with crown rust resistance we redid the GWAS analysis without division of genotypes into training and testing sets. We carried out GWAS using the complete population and found 55 markers significantly associated with crown rust resistance after correction for multiple testing (Table S3.1). Using the perennial ryegrass genome [23] as a reference, we located 29 significant markers within 22 genomic scaffolds that contained 50 predicted genes. Using the Genome Zippper [23, 137], we anchored ten scaffolds onto LG2, 3, 4, 5 and 7 (Table S3.3). Similarly, we did GWAS on IBERS material and found 24 markers associated with crown rust resistance (Table S3.2). All markers were located within 16 genomic scaffolds containing 44 predicted genes. Out of 16 scaffolds we were able to place seven scaffolds onto LG3, 5 and 7 (Table S3.3,S3.4). We found five common scaffolds between the complete population and the IBERS and only two of these scaffolds were mapped, onto LG3. On LG3 five markers were anchored within 60.4-61.21 cM. Genes present on these scaffolds were coding for domains including Mon1, Aquaporin, DUF1635, Nucleoredoxin, Beta-glucan export ATP-binding/permease protein, BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1, Alpha N-terminal protein methyltransferase 1. The gene func-

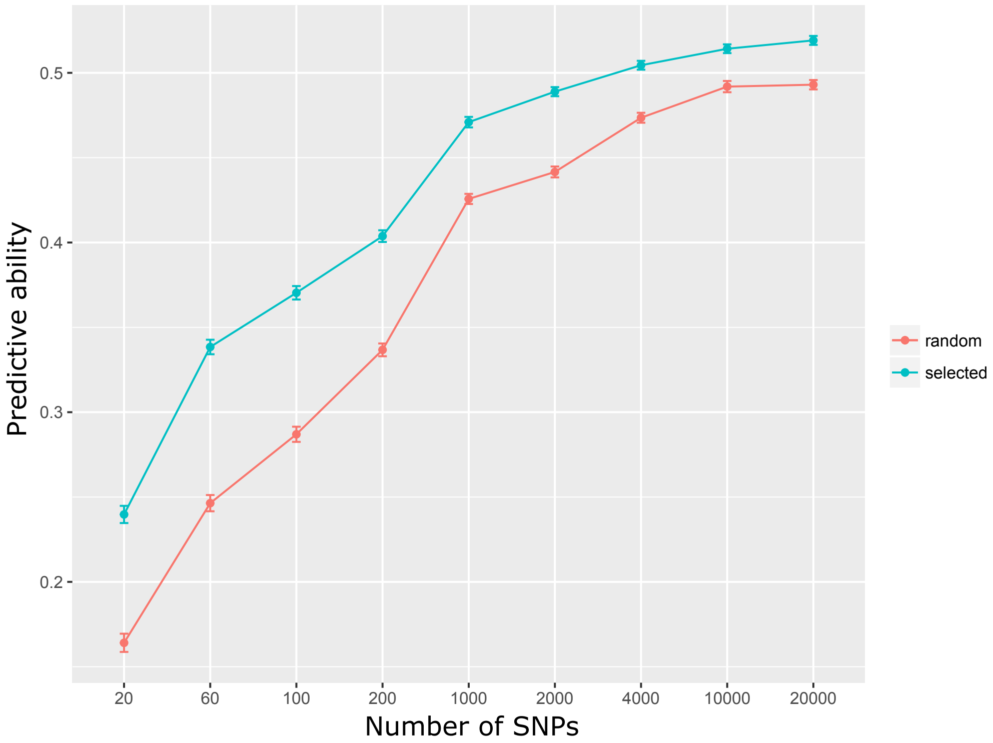

Figure 3.4: Predictive ability of selected markers versus random markers in complete population. Markers were selected based on the ranking from genome wide association studies and compared with random markers of similar size.

tion of these domains plays a key role in ATP-binding, membrane proteins, enzyme catalysis and pathogen-associated molecular pattern (PAMP)-triggered immunity (PTI) (Table S3.5, S3.6, S3.7, S3.8) [139].

Using small subsets of trait associated markers may be an effective strategy for within-family prediction of traits such as heading date, crown rust resistance and some quality traits. Predicting heading date from markers would enable plants to be matched in heading date to ensure sufficient cross-pollination when constructing synthetic cultivars [21]. Combining these with markers to predict crown rust resistance would also avoid the inclusion of plants with high levels of susceptibility, and furthermore prediction models can be based on multi-year evaluations. It is clear from the phenotypic data presented here that there is substantial within family variation for crown rust resistance. Opportunities already exist to genotype small to moderate sized marker panels in 1000s of samples at low cost [25]. Using these approaches small fragments (200-300bp) are amplified and sequenced at hundreds of loci. These amplicons can be used as short haplotypes in marker aided selection

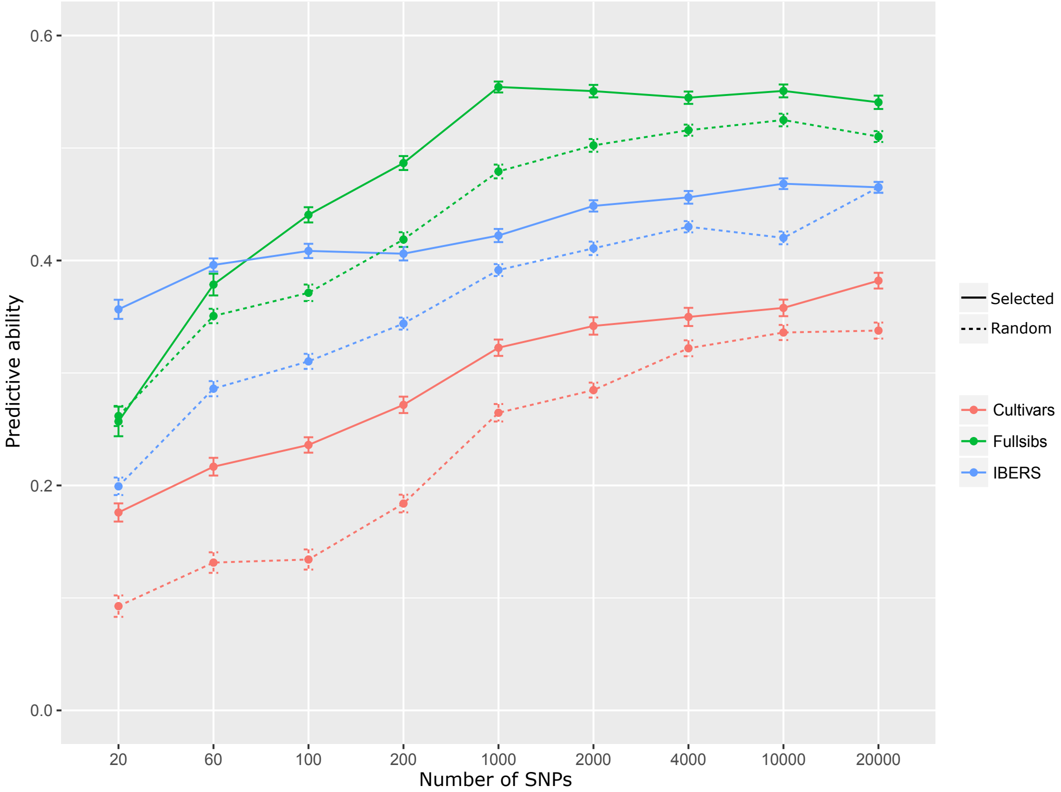

Figure 3.5: Comparing predictive ability of selected versus random markers. Markers were selected based on the ranking from genome wide association studies in Cultivars, Full-sibs and IBERS material and compared with random markers of similar size.

strategies. Initially, we will develop the assay to target loci in linkage with QTL for heading date [21] and crown rust resistance, and a suite of loci with a good distribution throughout the genome. Once high yielding families are identified in field trials, within family selection can be performed with this molecular marker assay to select plants for synthetics with synchronized flowering time and acceptable crown rust resistance. We plan to expand the assay to include loci linked to QTL associated with quality traits, allowing a breeding scheme where among-family selection is based on yield evaluations in the field and within-family selection for quality, heading date and crown rust resistance is based on markers.

3.4 Conclusions

Our findings show that predicting crown rust resistance in perennial ryegrass can be achieved with high accuracy using AUDPC scores on spaced plants. However, there was no difference in predictive ability when equations were developed using phenotypes from the last time-point scored or the AUDPC values. This means that scoring at a single time point was adequate to evaluate the crown rust susceptibility of the spaced plants and calculating AUDPC was unnecessary. Much of the predictive ability comes from markers capturing genetic relationships among the families, highlighted by the observation that there was no drop in predictive ability when going from the entire marker set down to only 5% (10,878) of the marker set. Accuracy due to genetic relationships will decay rapidly over generations whereas accuracy due to LD will persist. Using a GWAS we attempted to identify and rank markers in LD with QTL. This enabled a small panel of markers to be identified that had higher predictive ability than the same number of randomly selected markers, and had predictive abilities close to those achieved with the entire marker set.

List of abbreviations GS: genomic selection; GWAS: genome wide association study; LD: linkage disequilibrium; AUDPC: area under disease progress curve; QTL: quantitative trait locus; GEBV: genomic estimated breeding value

Author's contributions DM, PC, SB, TRH, MC conceived and designed the study. SKA and SLB performed the data analysis. SKA and SLB drafted the initial paper. SKA, SLB, DM, PC, TRH, MC, TM and SB contributed to interpretation of data and preparation of the final manuscript. All authors read and approved the final version.

Supplementary files

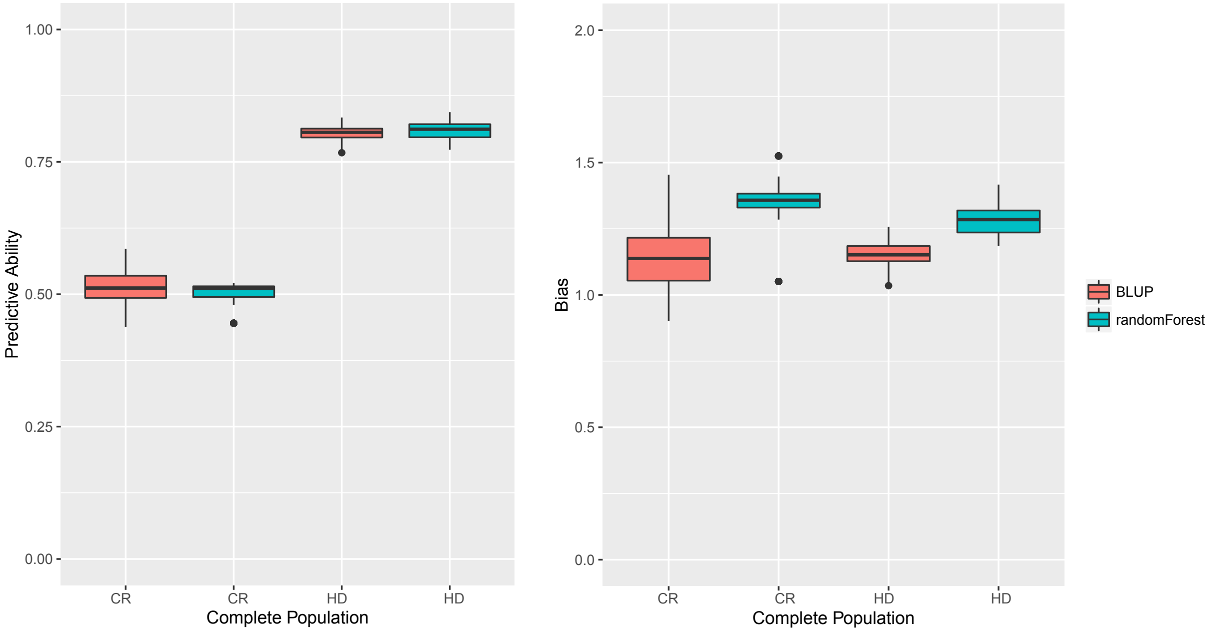

Figure S3.1: Predictive ability and bias for crown rust (CR) and heading date (HD) using rrBLUP (in red) and random forest model (in blue)

Table S3.1: List of markers associated with crown rust resistance based on genome wide association studies in complete population

marker	chrom	pos	(-log10p)	p-value
scaffold4630ref0024910	NA	20804	7.941769608	1.00E-09
scaffold2543ref0034320	NA	48546	7.311879805	2.29E-09
scaffold322ref0033181	NA	103294	7.128888413	2.95E-09
scaffold2543ref0034320	NA	48293	7.097824221	3.08E-09
scaffold19449ref0033434	NA	799	7.056263718	3.27E-09
scaffold745ref0048287	NA	126827	6.968761236	3.70E-09
scaffold4630ref0024910	NA	20880	6.841750463	4.45E-09
scaffold4566ref0015013	NA	56827	6.738381996	5.18E-09
scaffold4630ref0024910	NA	20797	6.544374232	6.94E-09
scaffold10537ref0014202	NA	14951	6.534674878	7.04E-09
scaffold3061ref0020001	NA	19424	6.417930836	8.43E-09
scaffold1959ref0042697	NA	51699	6.277526822	1.05E-08
scaffold10537ref0014202	NA	14952	6.274204153	1.06E-08
scaffold165ref0012847	NA	21652	6.241120127	1.12E-08
scaffold745ref0048287	NA	126873	6.161191063	1.27E-08
scaffold4576ref0045556	NA	2201	6.152848062	1.29E-08
scaffold3049ref0007382	NA	30526	6.075206319	1.46E-08
scaffold16787ref0028195	NA	6376	6.005295262	1.64E-08
scaffold4448ref0017406	NA	51686	5.918343404	1.90E-08
scaffold1959ref0042697	NA	51682	5.913685214	1.91E-08
scaffold41003ref0040763	NA	731	5.89101557	1.99E-08
scaffold176ref0043339	NA	61013	5.863611873	2.08E-08
scaffold322ref0033181	NA	103268	5.829832889	2.21E-08
scaffold15267ref0023130	NA	6908	5.809996767	2.28E-08
scaffold5403ref0046155	NA	46603	5.809241394	2.28E-08
scaffold12079ref0011165	NA	15633	5.765243739	2.47E-08
scaffold6898ref0036069	NA	47774	5.714738327	2.69E-08
scaffold10240ref0008205	NA	25006	5.689459688	2.81E-08
scaffold3232ref0030065	NA	66830	5.658335868	2.97E-08

Table S3.2: List of markers associated with crown rust resistance based on genome wide association studies in IBERS

marker	chrom	pos	(-log10p)	p-value
scaffold165ref0012847	NA	32653	7.154076371	2.8475E-09
scaffold13324ref0046917	NA	97792	6.985229661	3.61571E-09
scaffold237ref0040978	NA	109547	6.869525112	4.273E-09
scaffold13324ref0046917	NA	97791	6.834872487	4.49465E-09
scaffold1052ref0009181	NA	30902	6.716013897	5.35654E-09
scaffold1326ref0023230	NA	175102	6.555063165	6.82702E-09
scaffold2467ref0011548	NA	167776	6.545575638	6.92662E-09
scaffold2467ref0011548	NA	167780	6.425233	8.33895E-09
scaffold15267ref0023130	NA	102811	6.378657866	8.96823E-09
scaffold2543ref0034320	NA	172296	6.328177961	9.70986E-09
scaffold5403ref0046155	NA	61950	6.203242357	1.18526E-08
scaffold13324ref0046917	NA	97793	6.0815543	1.44495E-08
scaffold176ref0043339	NA	181764	5.864719152	2.07744E-08
scaffold4870ref0016801	NA	6337	5.863728591	2.08095E-08
scaffold48168ref0018173	NA	127401	5.832090144	2.19663E-08
scaffold165ref0012847	NA	32668	5.646030106	3.03786E-08
scaffold322ref0033181	NA	140884	5.588431859	3.36589E-08
scaffold2543ref0034320	NA	172290	5.452009907	4.30957E-08
scaffold4398ref0019381	NA	13496	5.138549	7.79103E-08
scaffold2712ref0010036	NA	172297	5.070353988	8.90466E-08
scaffold4398ref0019381	NA	13493	5.021706637	9.80588E-08
scaffold6131ref0028460	NA	112091	4.962099377	1.10496E-07
scaffold2543ref0034320	NA	172291	4.945148952	1.14342E-07
scaffold4398ref0019381	NA	13495	4.733071063	1.77243E-07

Table S3.3: List of genomic scaffolds where all the significant markers from genome wide association studies were located in the complete population. Scaffolds were placed onto linkage groups with the aid of the Genome Zipper

marker	LG	cM	no of markers
scaffold 6898 ref0036069	2	99.0 - 103.6	1
scaffold 3232 ref0030065	2	80.3 - 80.4	1
scaffold 2543 ref0034320	3	60.4 - 61.2	2
scaffold 5403 ref0046155	3	60.4 - 61.2	1
scaffold 4630 ref0024910	4	64.6 - 65.3	3
scaffold 745 ref0048287	4	50.7 - 51.8	2
scaffold 4576 ref0045556	4	55.1 - 55.4	1
scaffold 1959 ref0042697	5	9.1 - 11.0	2
scaffold 4566 ref0015013	5	31.7 - 32.8	1
scaffold 4448 ref0017406	7	44.7 - 45.3	1
scaffold 322 ref0033181	NA	NA	2
scaffold 10537 ref0014202	NA	NA	2
scaffold 19449 ref0033434	NA	NA	1
scaffold 3061 ref0020001	NA	NA	1
scaffold 165 ref0012847	NA	NA	1
scaffold 3049 ref0007382	NA	NA	1
scaffold 16787 ref0028195	NA	NA	1
scaffold 41003 ref0040763	NA	NA	1
scaffold 176 ref0043339	NA	NA	1
scaffold 15267 ref0023130	NA	NA	1
scaffold 12079 ref0011165	NA	NA	1
scaffold 10240 ref0008205	NA	NA	1

Table S3.4: List of genomic scaffolds where all the significant markers from genome wide association studies were located in the complete population. Scaffolds were placed onto linkage groups with the aid of the Genome Zipper

marker	LG	cM	no of markers
scaffold 1052 ref0009181	3	39	1
scaffold 4398 ref0019381	3	43.3	3
scaffold 6131 ref0028460	3	60.4	1
scaffold 2543 ref0034320	3	60.4 - 61.2	3
scaffold 5403 ref0046155	3	60.4 - 61.2	1
scaffold 4870 ref0016801	5	31.7 - 32.8	1
scaffold 237 ref0040978	7	44.7 - 45.3	1
scaffold 165 ref0012847	NA	NA	2
scaffold 13324 ref0046917	NA	NA	3
scaffold 1326 ref0023230	NA	NA	1
scaffold 2467 ref0011548	NA	NA	2
scaffold 15267 ref0023130	NA	NA	1
scaffold 176 ref0043339	NA	NA	1
scaffold 48168 ref0018173	NA	NA	1
scaffold 322 ref0033181	NA	NA	1
scaffold 2712 ref0010036	NA	NA	1

Table S3.5: List of predicted proteins on the genomic scaffolds. Markers located on these scaffolds were associated with crown rust resistance in complete population. BLAST was done on the predicted protein sequences using PLAZA to obtain the gene function and only predicted proteins with gene function were reported

protein	accession	comment
ms4576ref0045556–gene-0.0-mRNA-1	HV1562535G00010	purple acid phosphatase
ms4576ref0045556–gene-0.2-mRNA-1	PP00096G00250	Pectinesterase/pectinesterase inhibitor
ms4576ref0045556–gene-0.3-mRNA-1	BD1G68220	LRR-repeat protein
ms5403ref0046155–gene-0.0-mRNA-1	BD3G35480	Probable nucleoredoxin 1
ms5403ref0046155–gene-0.1-mRNA-1	BD2G62540	Beta-(1–2)glucan export ATP-binding/permease protein NdvA
ms176ref0043339–gene-0.1-mRNA-1	HV109398G00010	Protein phosphatase 2C
ms176ref0043339–gene-0.3-mRNA-1	HV39472G00010	BRASSINOSTEROID INSENSITIVE 1-associated kinase 1
ms176ref0043339–gene-0.4-mRNA-1	BD2G62680	threonine-protein kinase ATM
ms176ref0043339–gene-1.0-mRNA-1	HV324067G00010	metalloprotease ZmpB
ms176ref0043339–gene-1.2-mRNA-1	BD2G62730	protein DnaJ
ms176ref0043339–gene-1.3-mRNA-1	BD2G62697	SUMO-protein ligase SIZ1
ms2543ref0034320–gene-0.1-mRNA-1	BD2G62532	fusion protein MON1 homolog
ms2543ref0034320–gene-0.2-mRNA-1	HV43071G00010	Aquaporin
ms2543ref0034320–gene-0.3-mRNA-1	HV137222G00010	Protein of unknown function (DUF1635)
ms1959ref0042697–gene-0.0-mRNA-1	BD4G00740	BEL1-like homeodomain protein
ms3061ref0020001–gene-0.0-mRNA-1	BD4G38757	RNA polymerase gld-2 homolog
ms3049ref0007382–gene-0.0-mRNA-1	PP00156G00780	RNA polymerase GLD2
ms4448ref0017406–gene-0.0-mRNA-1	HV37339G00010	MADS-box transcription factor 55

Table S3.6: List of predicted proteins on the genomic scaffolds. Markers located on these scaffolds were associated with crown rust resistance in complete population. BLAST was done on the predicted protein sequences using PLAZA to obtain the gene function and only predicted proteins with gene function were reported

protein	accession	comment
ms12079ref0011165–gene-0.1-mRNA-1	SB02G010650	Alpha-galactosidase
ms4566ref0015013–gene-0.0-mRNA-1	HV64311G00020	Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B
ms4566ref0015013–gene-0.1-mRNA-1	HV223763G00020	RPP8-like protein 4
ms4566ref0015013–gene-0.2-mRNA-1	BD4G33450	Putative BPI/LBP family protein
ms4566ref0015013–gene-0.3-mRNA-1	HV52308G00010	Nuclear ribonuclease Z
ms4630ref0024910–gene-0.0-mRNA-1	HV1567116G00010	40S ribosomal protein SA
ms4630ref0024910–gene-0.1-mRNA-1	BD1G72470	Encodes initiation factor 3k
ms745ref0048287–gene-0.0-mRNA-1	BD1G63960	Membrane steroid-binding protein 2
ms745ref0048287–gene-0.1-mRNA-1	HV48620G00010	Guanylate kinase
ms745ref0048287–gene-0.2-mRNA-1	HV1577599G00010	Leucine-rich receptor-like protein kinase family protein
ms3232ref0030065–gene-0.0-mRNA-1	BD5G07790	sorting-associated protein 11 homolog
ms165ref0012847–gene-0.0-mRNA-1	BD2G61950	factor 4E-1
ms322ref0033181–gene-0.0-mRNA-1	OS05G02130	E3 ubiquitin-protein ligase XB3
ms322ref0033181–gene-0.1-mRNA-1	BD2G62410	Pectinacetylesterase family protein
ms322ref0033181–gene-0.2-mRNA-1	BD5G01462	Receptor like protein kinase
ms6898ref0036069–gene-0.0-mRNA-1	BD2G44160	11-oxo-beta-amyrin 30-oxidase
ms6898ref0036069–gene-0.1-mRNA-1	HV53759G00010	Lactosylceramide 4-alpha-galactosyltransferase

Table S3.7: List of predicted proteins on the genomic scaffolds. Markers located on these scaffolds were associated with crown rust resistance in IBERS population. BLAST was done on the predicted protein sequences using PLAZA to obtain the gene function and only predicted proteins with gene function were reported

protein	accession	comment
ms165ref0012847-est2-gene-0.0-mRNA-1	SI005G47950	Eukaryotic translation initiation factor 4E-1
ms237ref0040978-est2-gene-1.0-mRNA-1	BD1G45567	DIMBOA UDP-glucosyltransferase BX9
ms237ref0040978-est2-gene-1.1-mRNA-1	BD1G74770	hydroxynicotinate 3-monooxygenase
ms1052ref0009181-est2-gene-0.1-mRNA-1	BD2G21960	Protein BREAST CANCER SUSCEPTIBILITY 1 homolog
ms1052ref0009181-est2-gene-0.2-mRNA-1	SI002G36320	Mevalonate kinase
ms1052ref0009181-est2-gene-0.3-mRNA-1	BD2G13320	Light-mediated development protein DET1
ms2467ref0011548-est2-gene-0.1-mRNA-1	HV1570054G00010	Growth inhibition and differentiation-related protein 88
ms2467ref0011548-est2-gene-0.0-mRNA-1	BD2G61740	36 and H4 lysine-20 specific
ms2543ref0034320-est2-gene-0.1-mRNA-1	BD2G62532	fusion protein MON1 homolog
ms2543ref0034320-est2-gene-0.3-mRNA-1	HV43071G00010	Aquaporin
ms2543ref0034320-est2-gene-0.2-mRNA-1	HV137222G00010	DUF1635
ms5403ref0046155-est2-gene-0.0-mRNA-1	BD3G35480	Probable nucleoredoxin 1
ms5403ref0046155-est2-gene-0.1-mRNA-1	BD2G62540	Betaglucan ATP-binding/permease protein NdvA

Table S3.8: List of predicted proteins on the genomic scaffolds. Markers located on these scaffolds were associated with crown rust resistance in IBERS population. BLAST was done on the predicted protein sequences using PLAZA to obtain the gene function and only predicted proteins with gene function were reported

protein	accession	comment
ms176ref0043339-est2-gene-0.1-mRNA-1	HV109398G00010	Protein phosphatase 2C
ms176ref0043339-est2-gene-0.3-mRNA-1	HV39472G00010	BRASSINOSTEROID INSENSITIVE 1-associated kinase 1
ms176ref0043339-est2-gene-0.4-mRNA-1	BD2G62680	threonine-protein kinase ATM
ms176ref0043339-est2-gene-1.2-mRNA-1	BD2G62730	protein DnaJ
ms176ref0043339-est2-gene-1.3-mRNA-1	BD2G62697	SUMO-protein ligase SIZ1
ms4870ref0016801-est2-gene-0.2-mRNA-1	HV100406G00010	gamete expressed protein 1
ms4870ref0016801-est2-gene-0.3-mRNA-1	HV244641G00010	Cathepsin L-like cysteine proteinase
ms322ref0033181-est2-gene-0.0-mRNA-1	SI003G06460	E3 ubiquitin-protein ligase XB3
ms322ref0033181-est2-gene-0.1-mRNA-1	BD2G62410	Pectinacetylesterase family protein
ms322ref0033181-est2-gene-0.2-mRNA-1	BD5G01462	Receptor like protein kinase
ms4398ref0019381-est2-gene-0.0-mRNA-1	HV130622G00010	GTP-binding protein TypA/BipA
ms4398ref0019381-est2-gene-0.2-mRNA-1	BD2G50060	28 kDa heat- and acid-stable phosphoprotein
ms4398ref0019381-est2-gene-0.3-mRNA-1	HV68202G00010	Vascular plant one zinc finger protein

Chapter 4

Genomic prediction for pasture yield in tetraploid perennial ryegrass

Sai Krishna Arojju[1,3], Patrick Conaghan[2], Dan Milbourne[1], Susanne Barth[1], Trevor R Hodkinson[3], and Stephen L Byrne[1*]

[1] Teagasc, Crop Science Department, Oak Park, Carlow, Ireland

[2] Teagasc, Grassland Science Research Department, Animal and Grassland Research and Innovation Centre, Oak Park, Carlow, Ireland

[3] Department of Botany, School of Natural Sciences, Trinity College Dublin, Dublin2, Ireland

Manuscript in preparation will be submitted to journal before thesis defense

Abstract

Forage yield is the most important trait in perennial ryegrass breeding. In production systems where animals are at pasture for up to 300 days per year it is important that forage yield meets feed demand throughout that time. The value of forage at different times during the year can be captured and used as an index to aid selection. In this study we have evaluated genomic prediction as a means of accelerating the rate of genetic gain for forage yield to meet the feed demand. Tetraploid half-sib families were evaluated for forage yield in both simulated grazing and conservation management regimes over two years, and their maternal parents were genotyped. Linkage disequilibrium in the population above background levels extended over large cM distances. Marker-based heritabilities for traits varied from 0.07 to 0.27, and genomic predictive abilities for traits ranged from 0.03 to 0.30. Predictive abilities of 0.22 were achieved for both first cut silage under conservation management and economic value under simulated grazing management. Our results indicate that genomic prediction for both yield under grazing (calculated as economic value of a plot) and yield under first cut silage in tetraploid perennial ryegrass is promising, and that the ability to complete multiple cycles of indirect selection with DNA markers relative to conventional genotypic selection will result in increased genetic gains.

Keywords: Genomic selection, linkage disequilibrium, Lolium perenne, tetraploid, yield

4.1 Introduction

Perennial ryegrass (*Lolium perenne*) arguably is the most important grass species in Ireland and around temperate regions of the world [83]. Yield is an important trait for perennial ryegrass, to provide a natural low-cost feed for ruminants [84, 187]. High yielding cultivars produce more herbage, making it a relatively cheap and high quality feed for animals. However, grass growth is uneven throughout the year, with excess of grass growth during the summer and a growth deficit in spring and autumn in a typical grazing system. To compensate the feed demand in spring and autumn, surplus yields from summer are often harvested and stored as silage. Silage production is an expensive process and decreases the overall profitability of the farm [133]. Developing cultivars with increased yield in spring and autumn will increase the overall grazing period and thus decrease the feed cost [127, 187].

McEvoy et al. [117, 119] introduced the pasture profit index (PPI), to identify and rank cultivars that will be profitable at farm level. The PPI assigns economic values for key traits such as dry matter yields (spring, summer and autumn yields), first and second silage cut, dry matter digestibility and persistency. The PPI was based on simulating a spring-calving dairy farm model over a period of 12 months. Net profits per hectare were calculated for the model farm and by simulating a change in each trait separately, the effect of change on the model farm was calculated. The difference between the change in net margin per hectare (before and after simulating change) divided by change in trait is considered as the economic value of the trait [117, 119]. Currently ryegrass cultivars produce around 17 t/ha in Europe and there is a potential to increase forage yield up to 25 t/ha [83]. However, genetic gain for annual dry matter yield is about 0.3 to 0.5% per year [116, 187] and these rates of genetic improvement are significantly lower compared to cereals, which are 1.0 to 1.5% per year [132]. Some of the reasons for these poor genetic gains compared to cereals were due to (i) longer breeding cycles in forage crops, with each selection cycle taking up to three to five years, (ii) inability to exploit heterosis, as in hybrid crops, and (iii) selecting for multiple traits, which are not correlated or negatively correlated with forage yield [28, 82].

With the development of molecular markers, marker assisted selection (MAS) looked like a promising approach to speed up the selection cycle and thus increase genetic gain for economically important traits in perennial ryegrass [32]. Studies have mapped quantitative trait loci (QTL) linked to yield in bi-parental mapping populations on all linkage groups [59, 160, 176]. However, multiple inconsistent QTLs were detected due to high G x E interactions and MAS would be ineffective for improving complex traits [14, 32]. With the development of high density genotyping panels, genome wide association studies (GWAS) became a popular choice for QTL identification in breeding material, but its use has been limited due to the requirement of additional steps in marker selection and validation, potential overestimation of marker effects and the ability to explain only a small proportion of variance [12, 32]. Meuwissen et al [123] proposed an approach to integrate all markers simultaneously in the model to estimate breeding values. This approach is known as genomic prediction or genome wide selection. To implement genomic prediction, initially a training population is established which is genotyped with high density markers and phenotyped for the trait under evaluation. Using both genotypic and phenotypic information, marker effects are estimated to predict genomic estimated breeding values (GEBVs) in the selection population [81, 89]. Because of the high density of markers used for predictive modeling, it is possible to capture marker effects consistently across the population, making predictions more reliable [44]. Genomic prediction revolutionized animal breeding and continues to be successfully applied [124], but it is still in an early phase in plant breeding.

Previous studies have demonstrated the potential of genomic prediction in diploid perennial ryegrass populations [21, 60, 61, 72]. A prediction accuracy of 0.22 was reported for total yield in diploid perennial ryegrass families [72]. While, these results are encouraging with high accuracy for most of the traits, no study has yet explored the potential of genomic prediction in tetraploid perennial ryegrass families. In this study, we aim to evaluate genomic prediction for forage yield under both conservation and simulated grazing management in a tetraploid perennial ryegrass population, and to develop genomic prediction models for the economic value.

4.2 Material and methods

4.2.1 Plant material and experimental design

The tetraploid breeding material used in this study were developed from a commercial cultivar, which has been on the Irish recommended list since 2012. The cultivar was developed by intercrossing 75 plants from four full-sib families. A set of 120 plants from the cultivar was planted out in a polycross nursery and allowed to cross pollinate. This was replicated seven times. Seeds from matching maternal parents were harvested and bulked, producing half-sib families. Out of 120 half-sib families, 109 families produced enough seeds for a replicated field trial and two managements were planted out in a partially balanced incomplete block design with seven blocks within each replicate for measuring yield in 2015 and 2016 at Oak Park, Carlow, Ireland

Yield was measured as fresh weight under two management schemes, (i) simulated grazing management (SGM) with seven harvest cuts per year (each cut every four weeks from March to October) (Table 4.1) and (ii) conservation management (CM) with four harvest cuts per year (Table 4.1). For CM only first cut (May) was used for model development, due to its economical importance. The experiment was carried out with two controls, Kintyre and Abergain in each block. Each plot size was 6 x 1.5 m and harvested approximately at 4 cm above the ground using a plot harvester.

4.2.2 Phenotyping and data analysis

In SGM, yield data was collected from seven harvest cuts in two years. Total yield in each year was the sum of all seven harvest cuts. We estimated economic value of the plot (EV) using weightings from the PPI [117, 119]. Cuts in SGM were divided into spring yield (cuts 1 and 2), summer yield (cuts 3, 4 and 5) and autumn yield (cuts 6 and 7) (Table 4.1). Yields in spring, summer and autumn were multiplied by €0.16, 0.04 and 0.11 respectively [117]. The EV was calculated by summing spring, summer and autumn values. Under CM, yield data of the first harvest cut was used for the analysis. In total five traits from SGM (total yield, EV, spring, summer and autumn yields) and one trait from CM (first silage cut) were used for further analysis.

Table 4.1: Forage yield measured under simulated grazing management (SGM) and conservation management (CM) was collected during the dates mentioned below in the table.

Management	Year 1 (2015)	Year 2 (2016)
SGM		
Cut 1	09/04 to 10/04	14/03 to 15/03
Cut 1	29/04 to 01/05	18/04 to 19/04
Cut 3	21/05 to 22/05	12/05 to 13/05
Cut 4	10/06 to 19/06	07/06 to 08/06
Cut 5	15/07 to 17/06	04/07 to 05/07
Cut 6	10/08 to 12/08	04/08 to 05/08
Cut 7	07/09 to 08/09	14/09 to 21/09
CM		
Cut 1	21/05 to 22/05	12/05 to 13/05

Phenotypic analysis was carried out in two stages. In stage one, grand mean and mean of controls in each block was calculated and the difference was considered as the adjusted control. The difference between the adjusted control and the mean of each plot value is the adjusted mean. In stage two, adjusted means were used to fit a mixed model with genotype as a random effect, and year and replicate as fixed effects, to obtain conditional modes (also called best linear unbiased predictions, BLUPs) for all six traits. The repeatability (broad sense heritability) for each trait was calculated as follows:

$$v^2 = \frac{\sigma_g^2}{(\sigma_g^2 + \sigma_{res}^2/r)} \tag{4.1}$$

The variance components for genotype and residuals were estimated based on analysis of variance (ANOVA) and r is the replicates per genotype. Marker-based narrow sense heritability was estimated based on marker data, genotypic and residual variance were computed using a mixed model, based on restricted maximum likelihood estimates. The repeatability and marker-based heritability were calculated using the R package heritability [103].

4.2.3 Genotyping and variant calling

DNA was extracted from leaf samples of maternal genotypes using standard cetyl trimethyl ammonium bromide (CTAB) method [50]. A genotyping by sequencing (GBS) approach was used for library preparation and was carried out as a service by LGC Genomics, Berlin, Germany. Sequence reads were quality filtered, demultiplexed and aligned to a reference perennial ryegrass genome [23]. After demultiplexing, genotype calling was done according to Li et al [109] where all three heterozygous states were called as heterozygous. Briefly, for each SNP a minimum of 11 reads were required to call a homozygote (i.e AAAA). If fewer than 11 reads were present, the genotype was considered as missing to avoid misclassifying a triplex heterozygote (i.e AAAT) as homozygous. To call a SNP heterozygous, two reads per allele was set as the threshold and a minor allele frequency (MAF) for the given SNP set to be greater than 0.10, otherwise it was considered as a missing genotype. Any SNPs with less than 5% MAF and missing data points greater than 50% were eliminated. These filtering resulted in 45,569 genome wide markers.

4.2.4 Linkage disequilibrium and population structure

To determine linkage disequilibrium (LD) we obtained genetic positions for 26,333 markers using GenomeZipper [23, 137]. All heterozygous markers were marked as missing values and marker loci containing more than 50% missing genotype values were removed. MAF was calculated on these markers and markers with less than 5% MAF were removed, reducing the dataset to 1,029 markers. The extent of LD in the population was estimated on these 1,029 markers using PLINK 1.9 [30] and there are various approaches to set a threshold to determine the extent of LD in the population. The most common approach is to use an r^2 threshold value of 0.1 or 0.2, but recently published literature used a different approach to set a threshold value based on background LD [1, 18, 178, 180]. LD was assessed for linked loci (located on the same linkage group) and unlinked loci (located on different linkage groups), as an r^2 value between pairs of markers. Background LD, which is the level of disequilibrium between unlinked loci was determined based on the distribution of r^2 values. Unlinked r^2 values were log transformed to approximate normally distributed variables, and then a parametric 95th percentile of the distribution was considered as the background LD [18]. Intersection of background LD value with the loess fit curve is considered as the extent of LD in each linkage group [18, 175, 180].

Structure in the population was assessed using fastStructure [145], based on markers with genetic positions (26,333 markers). The program fastStructure determines the optimum number of subpopulations using a model based Bayesian algorithm. Models were run with a varying number of clusters (K) from K = 1 to 10. The best fit model was selected based on marginal likelihood values using Python chooseK.py script implemented in fastStructure. In addition, principal component analysis (PCA) was performed using prcomp function in R [173] and the first two components were visualized using an R package ggfortify [170].

4.2.5 Genomic prediction models and cross validations

We evaluated four statistical models for genomic prediction: Genomic best linear unbiased prediction (GBLUP), ridge regression BLUP (rrBLUP), random forest regression and support vector regression. GBLUP is a widely used genomic prediction model for estimating marker effects [58]. GBLUP is a mixed linear model, in which covariance among individuals is the realized genetic relationship estimated using genome wide markers. The mixed model can be expressed as follows:

$$y = \mu + Xu + \epsilon \tag{4.2}$$

where y is the vector of input phenotypic values, μ is the overall mean, X is the marker matrix, $u \sim N(0, I\sigma_e^2)$ represents the realized genetic relationship matrix calculated from genome wide markers, ϵ is a vector of residual effects. rrBLUP is also a mixed linear model, but covariance among markers is considered to be zero. In the mixed model equation, $u \sim N(0, I\sigma_e^2)$ is the vector of markers effects. Both GBLUP and rrBLUP rely on the assumption that the trait is controlled by many genes with small effects (infinitesimal model). We used an R package rrBLUP to implement GBLUP and rrBLUP [58].

Random forest regression is an ensemble learning algorithm, in which a series of regression trees are grown to the largest extent possible with a subset of bootstrapped samples. At each split of the tree, a random subset of variables is selected to identify the best split. This is repeated for each of the bootstrap samples and finally trees are averaged. We used an R package randomForest [110], setting the number of variables at each split to 1/3 of the total variables, and using a terminal node size of five and minimum of 500 trees per forest.

Support vector regression is a supervised learning method, in which input variables are initially mapped into multi-dimensional feature space using non-linear mapping and then a linear model is developed within this feature space [41]. We used an R package e1071 [125] for implementing support vector regression, using linear regression model with cost of constraint violation (C) as 1 and epsilon (ϵ) insensitive regression as 0.1.

Cross-validation was performed by assigning 80% of the population as training set and assigning 20% of the population as test set. Predictive ability was measured as the Pearson correlation coefficient of true phenotypic value and predicted phenotype (GEBVs) of the test set averaged over 1,000 iterations. We assessed predictive ability for EV, total yield and first silage cut using conditional modes. In addition, predictive ability was estimated for spring, summer and autumn yields.

4.3 Results and discussion

4.3.1 Phenotypic analysis

The production environment defines the breeding goals and in Ireland perennial ryegrass is utilised in production systems that are predominantly based on grazing for up to 300 days in a year. While forage yield is one of the important trait selected for in breeding programmes; the value of any additional yield differs depending on the time of year. Overall, there is a large amount of variation in forage yield across the different cuts under SGM, and also between years (Figure 4.1) (Table 4.2). In SGM, yield measured in the second year was slightly lower than the first year and a moderate correlation was observed between years. Under CM, yields were similar in both the years, but correlation between years was very low (Table 4.2). Previous studies have highlighted the significant G x E interaction for forage yield, making the trait complicated to measure [33, 187]. Weather has a large impact on the year to year yield differences and is the primary contributor for G x E interactions [33]. Yield differences between years may also be caused by cultivar persistence [24, 33]. The change in the persistence, tolerance to grazing, and susceptibility to disease can all lower the yields in second or subsequent years [33, 187].

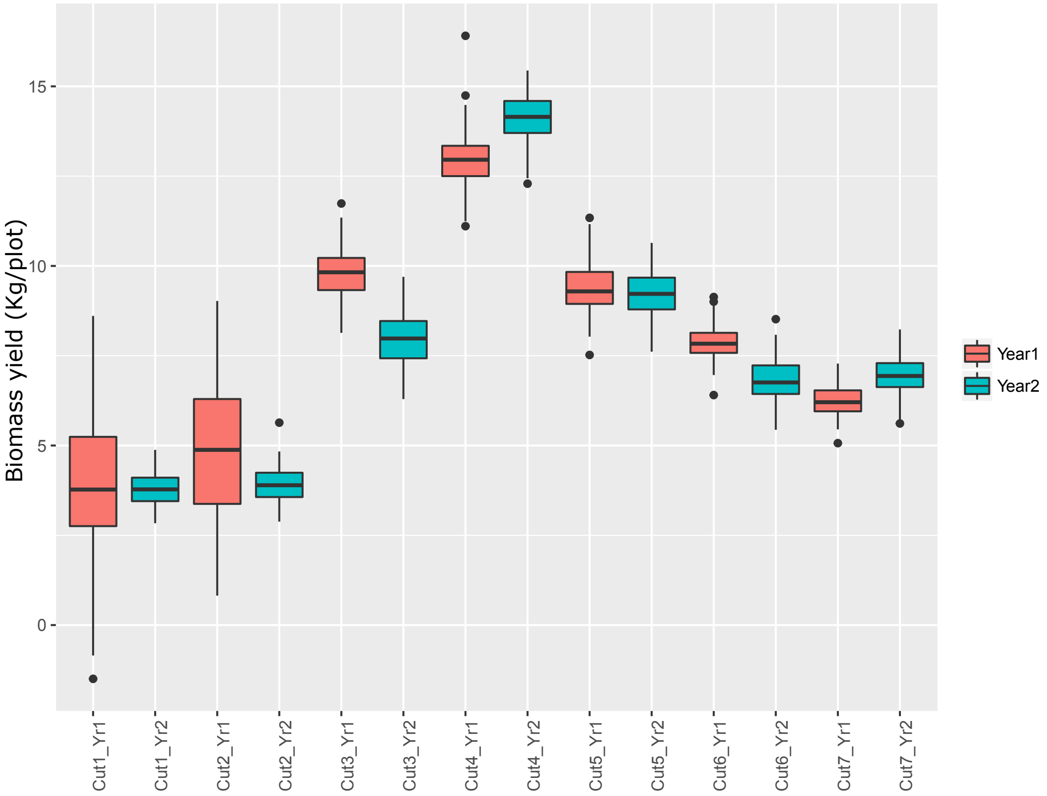

Figure 4.1: Phenotypic description of seven harvest cuts evaluated for two years under grazing management. Boxplot on the X-axis represents each cut and is matched with the closest cut in year 2. Scale on the Y-axis represents kilograms of yield produced per plot.

Table 4.2: Summary statistics of yields produced in each year under grazing and conservation management. In grazing management, total yield in each year was the sum of seven harvest cuts and in conservation management, the first silage cut is the value of a single harvest cut. Correlation is between two years for each trait

Trait	Year	Average	SD	Min	Max	Correlation
Total yield	Year 1	55.0	2.3	48.7	63.8	0.54
	Year 2	52.7	2.3	46.7	58.4	
First silage cut	Year 1	31.7	1.9	24.7	35.6	0.14
	Year 2	31.5	2.8	17.7	38.7	

Perennial ryegrass cultivars, perform differently under different managements [71, 116, 118]. This is due to a high degree of genetic independence between yield of reproductive growth, which primarily contributes to the silage yield and the yield of vegetative growth which primarily contributes to grazing yield [187]. In this study we have managed the population under two managements (SGM and CM), and we observed that the ranking of families differed between the two managements. When we ranked the families in each management and compared them there were only three families (K20, K65 and K142) in common across the top 20 of SGM and CM.

Forage yield is highly variable throughout the year [48, 187], with yields that cannot meet the annual demand in spring and autumn. The relative value of yield at different times of the year has been captured in the PPI [117, 119], which was developed in Ireland to place an economic value on cultivars. The PPI is primarily aimed at helping producers select new cultivars when reseeding pastures but also serves as a selection index for breeders developing new cultivars for Irish production systems. We have used the weightings for spring, summer, and autumn growth to assign an economic value to a family on a plot basis. These values are being used to develop genomic prediction models for EV.

4.3.2 Genotyping, linkage disequilibrium and population structure

In total, 45,569 markers were identified in this population and we were able to place 26,333 markers onto the seven linkage groups of perennial ryegrass using the GenomeZipper [23, 137]. Markers were evenly distributed across all seven linkage groups with the least number of markers on linkage group five and highest number of markers on linkage group four. We conclude from this that the markers are well distributed and are suitable for genome wide selection. The placement of markers on the linkage map also enabled us to estimate long range LD in this population.

We estimated background LD by determining pairwise LD between markers on different linkage groups, and based on the 95th percentile of pairwise r^2 values we determined background LD to be 0.064. LD between the linked loci (markers on the same linkage group) was estimated for all the seven linkage groups. The extent of LD (above background) in linkage group 1 to 7 varied from 2.5cM to 10cM (Figure 4.2, Figure S4.1-S4.6). Extent of LD is population specific and is considered to be extremely low in broad collections of perennial ryegrass, due to large past effective population size [77]. Because we have a restricted population, we see much higher levels of LD. In linkage group 1, the median r^2 value was 0.017 with 23% of pairs of markers above the background LD and 16% of pairs of markers above the r^2 value 0.1. Considering 0.064 as the base value, the LD in linkage group 1 extended to 10cM (Figure 4.2). Similar results were obtained for other linkage groups (Figure S4.1 - S4.6). It has already been demonstrated in many plant studies that predictive ability can be largely dependent on markers capturing genetic relationships between individuals [148, 175]. Similar observations have been made in cattle and sheep, where decreases in prediction accuracy were observed after correction for population structure [42, 152]. This is relevant because accuracy due to genetic relationships will decay rapidly over generations whereas accuracy due to LD will persist longer.

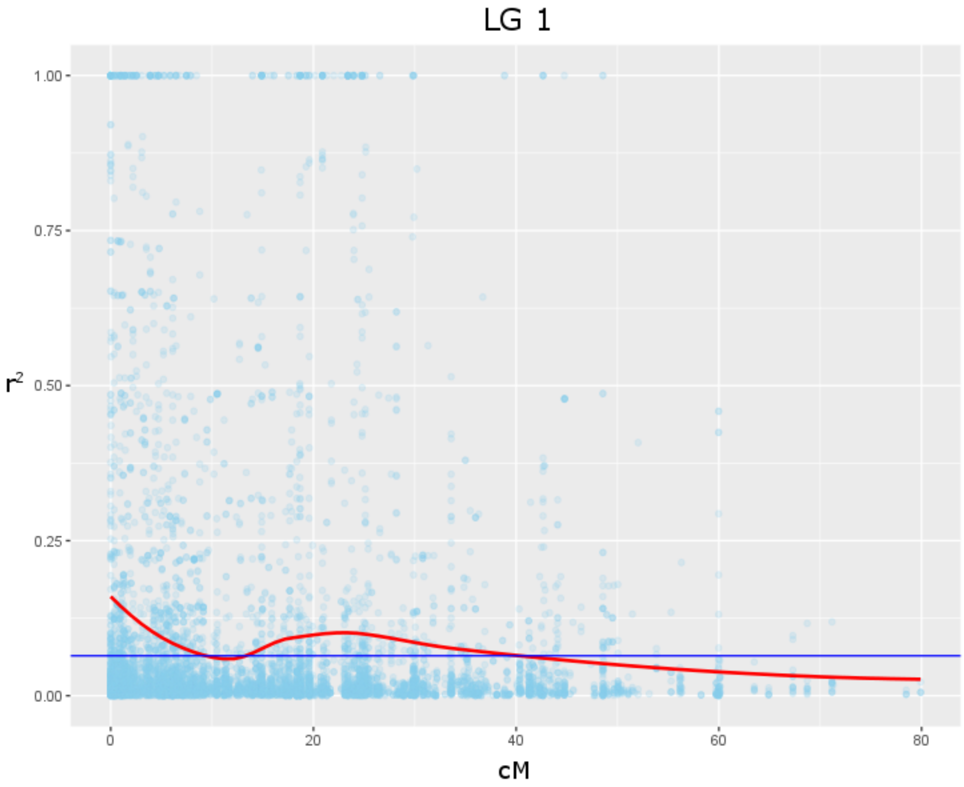

Figure 4.2: Extent of linkage disequilibrium (LD) on linkage group 1, estimated as r^2 value over the distance (cM). Red line represents the smoothing curve fitting using loess curve. Blue line shows the 95th percentile of unlinked loci (background LD)

Absence of population structure simplifies methodologies for genomic prediction. It was envisaged that the design of this population would mean it was free of population structure and this was confirmed. The optimum number of clusters in the population were determined based on marginal likelihood values. When K varied from 1 to 10, marginal likelihood values ranged from -0.8249 to -0.8387, with K=1 generating the highest value. This suggests a lack of population structure that was supported by the PCA plot, which also lacked evidence of clear groupings (Figure 4.3). One of the main challenges with implementing genomic prediction in perennial ryegrass is that useful LD only extends over short distances resulting from a very large past effective population size. This impacts the size of the reference population and number of markers needed to achieve high predictive accuracies. Seeking out inflated LD has already been suggested as an approach to implementing genomic prediction in outbred forage species such as perennial ryegrass [77].

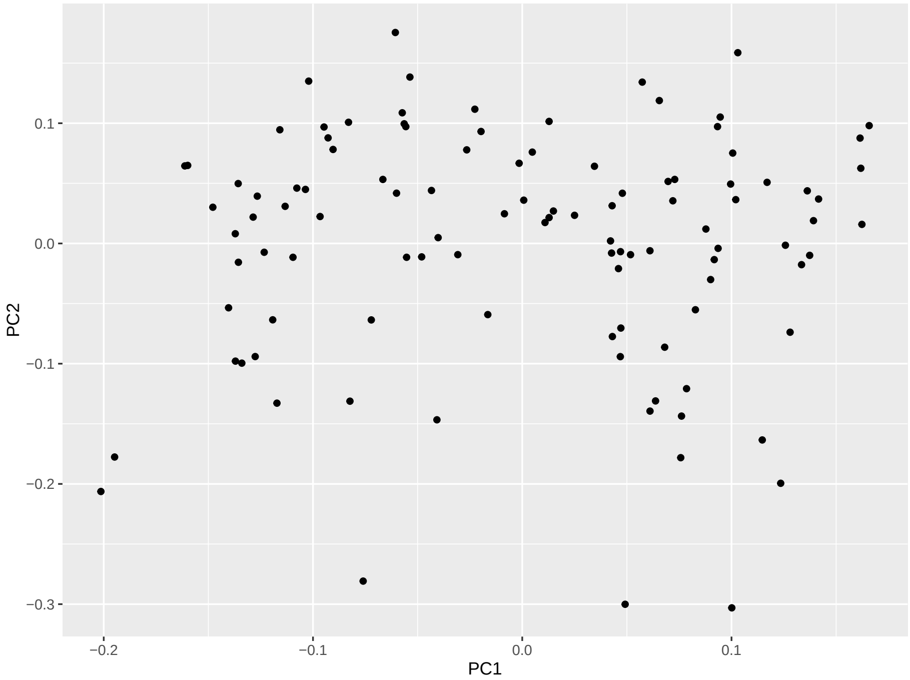

Figure 4.3: Principal component analysis for complete population based on 45569 markers. A lack of population structure was evident because of absence in groupings

4.3.3 Estimates of repeatability and marker-based heritabilities

We calculated repeatability and marker-based heritability for all the traits (Table 4.3). Marker-based heritability is an estimation of narrow sense heritability and calculates the proportion of the variance explained only by additive effects, whereas repeatability (broad-sense heritability) accounts for additive and non-additive (dominance and epistatic) effects [103, 190]. Marker-based heritability is calculated using a mixed model to explain the phenotypic variance by accounting for genetic relatedness and within genotype variability using replicated plot data [103]. Marker-based heritability for all the traits ranged from 0.07 to 0.27 (Table 4.3). Generally, heritability estimates are population specific and can depend on the evaluation environments. This makes it difficult to compare estimates of heritability across studies. Most of the studies in perennial ryegrass estimated yield related traits to be of low to moderate (0.20 to 0.50) heritability, and these statements agree with our results [61, 65, 72, 88].

Table 4.3: Repeatability (v^2) and marker-based heritability (h^2) for first silage cut (harvest cut under conservation management), EV (economic value for each plot estimated based on pasture profit index), total yield (sum of seven cuts measured under simulated grazing management), spring yield (sum of cut 1 and 2 measured under simulated grazing management), summer yield (sum of cut 3, 4 and 5 measured under simulated grazing management) and autumn yield (sum of cut 6 and 7 measured under simulated grazing management). Repeatability is based on phenotypic values and Heritability is based on markers with confidence interval (CI) of 95%.

Trait	Repeatability (v^2)	Heritability (h^2)	CI (h^2)
First silage cut	0.36	0.12	0.06 - 0.25
EV	0.43	0.27	0.18 - 0.39
Total yield	0.43	0.25	0.16 - 0.38
Spring yield	0.22	0.16	0.08 - 0.28
Summer yield	0.20	0.07	0.02 - 0.22
Autumn yield	0.42	0.19	0.10 - 0.31

4.3.4 Genomic prediction for yield

In this study we evaluated accuracy of genomic prediction using genotypes of maternal plants and phenotypes collected on half-sib progeny. This is an ideal evaluation of the potential of genomic prediction for forage yield as it focuses on the additive genetic variation which is relevant for predicting parental breeding values during synthetic cultivar development. We evaluated genomic prediction for forage yield traits measured under two different managements, SGM and CM. Under CM first cut silage is the trait of greatest importance and we obtained a maximum mean predictive ability of 0.22. The marker-based estimate of narrow-sense heritability for this trait was 0.12, giving a prediction accuracy for 1st cut silage of 0.63. Under SGM we determined predictive abilities for total yield, spring yield, summer yield, autumn yield and EV based on weightings in the PPI. Predictive ability for these traits ranged from 0.03 (summer yield) to 0.30 (spring yield) (Table 4.4). The low predictive ability for summer yield corresponded to the low estimate of marker-based heritability of 0.07. The relationship between prediction accuracy and heritability has already been shown in other studies, with low heritable traits having low predictive ability in many crops species [104, 105, 153]. We evaluated four statistical models (rrBLUP, GBLUP, random forest and support vector regression) to develop the genomic prediction models and for most traits rrBLUP/GBLUP gave the highest values (Table 4.4). GBLUP and rrBLUP both rely on the assumption that genetic control of the trait follows an infinitesimal model and are considered as statistically similar [78, 123]. In only one case (spring yield) did random forest outperform other models and one case (summer yield) where support vector regression outperformed other models (Table 4.4). This is similar to findings from other studies where in general GBLUP or rrBLUP outperformed other models [4, 72].

Table 4.4: Comparing mean predictive ability using four genomic prediction models for first silage cut (first harvest cut under conservation management), *EV* (economic value for each plot estimated based on pasture profit index), total yield (sum of seven cuts measured under simulated grazing management), spring yield (sum of cut 1 and 2 measured under simulated grazing management), summer yield (sum of cut 3, 4 and 5 measured under simulated grazing management) and autumn yield (sum of cut 6 and 7 measured under simulated grazing management). Values in the bracket represent the median predictive ability for each trait

Trait	rrBLUP	GBLUP	RF	SVR
First silage cut	0.22 (0.23)	0.22 (0.23)	0.21 (0.22)	0.18 (0.19)
EV	0.22 (0.24)	0.22 (0.23)	0.16 (0.17)	0.18 (0.18)
Total yield	0.16 (0.16)	0.16 (0.17)	0.11 (0.11)	0.12 (0.13)
Spring yield	0.28 (0.30)	0.28 (0.30)	0.30 (0.32)	0.29 (0.32)
Summer yield	0.03 (0.03)	0.03 (0.04)	0.05 (0.05)	0.07 (0.08)
Autumn yield	0.16 (0.16)	0.16 (0.16)	0.12 (0.13)	0.09 (0.10)

There has been one other study in perennial ryegrass (although with diploid types) that investigated predictive ability for forage yield using half-sib families [72]. In that study, prediction accuracies were generally low for the conservation cut (second cut). However, they obtained prediction accuracies of up to 0.275 for total yield in year one, which compares favorably to the predictive ability of 0.16 obtained here for total yield based on conditional modes calculated from two years data. Another very similar study to the one presented here was carried out recently on Alfalfa and included a similarly sized training population of 125 half-sib families [4]. In that case predictive abilities of 0.32 were achieved for forage yield. Of particular interest in this study was the development of prediction models based on forage yield weighted according to economic value across the season. The marker-based heritability for this trait was 0.27 and we achieved a predictive accuracy of 0.22. When calculating EV, forage yield in spring (cuts 1 and 2) is awarded the highest value and predictive accuracy for spring yield was 0.30.

Our results on genomic prediction for forage yield and weighted forage yield (presented as EV) in a tetraploid perennial ryegrass breeding population are encouraging. These were achieved with a modestly sized training population that had been developed to (i) be free from population structure and (ii) to have inflated linkage disequilibrium. The fact that we are able to achieve good predictive abilities with relatively small training populations is likely a result of the inflated LD we observe in this population and that the testing material is closely related to the training material. The predictive ability was moderate (0.22) for the EV calculated from the PPI values for spring, summer, and autumn yield. Considering the marker-based heritability for this trait as 0.27 the relative selection efficiency of indirect selection with markers vs. conventional genotypic selection is 0.42. This is assuming identical selection intensities, however, it would be expected that higher selection intensities could be achieved with genomic prediction. Considering that we can complete five cycles of genomic prediction in the same time it takes to do a single cycle of genotypic selection, there is a 2.1 fold greater efficiency for genomic prediction. This assumes no degradation in the predictive accuracy over generations. This is in general agreement with a very similar study carried out in alfalfa [108]. That study also had a small training population made up of half-sib families, and without population structure. Furthermore, as pointed out by Li et al. [108], we should not expect large decreases in accuracy over generations because of the complex nature of forage yield that is likely to be controlled by thousands of loci.

4.4 Conclusions

This is the first report on genomic prediction for forage yield in tetraploid perennial ryegrass families. Our results indicate that indirect selection with genome wide markers for both yield under grazing (calculated as EV) and first cut silage is promising, and that the ability to complete multiple cycles of selection with genomic prediction relative to conventional genotypic selection will result in increased genetic gains.

List of abbreviations LD: linkage disequilibrium; GEBV: genomic estimated breeding value; PPI pasture profit index; EV economic value of the plot; SGM: simulated grazing management; CM: conservation management

Supplementary files

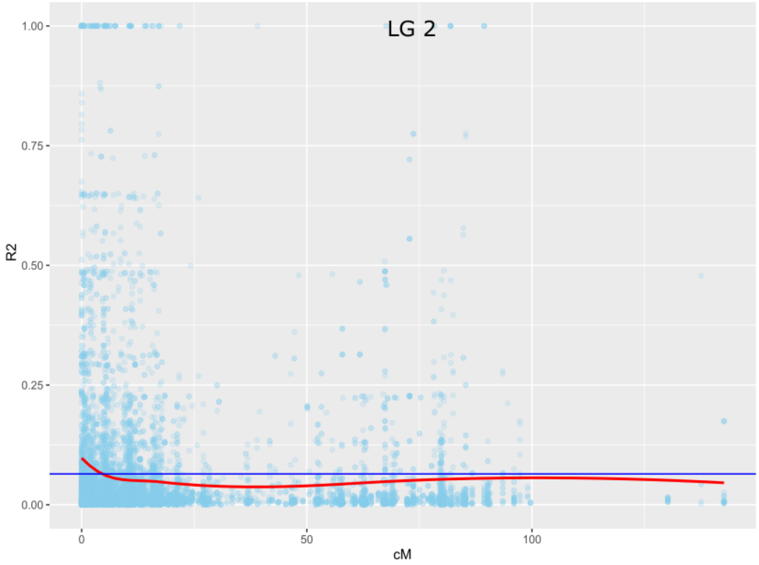

Figure S4.1: Extent of linkage disequilibrium (LD) on linkage group 2, estimated as r^2 value over the distance (cM). Red line represents the smoothing curve fitting using loess curve. Blue line shows the 95th percentile of unlinked loci (background LD)

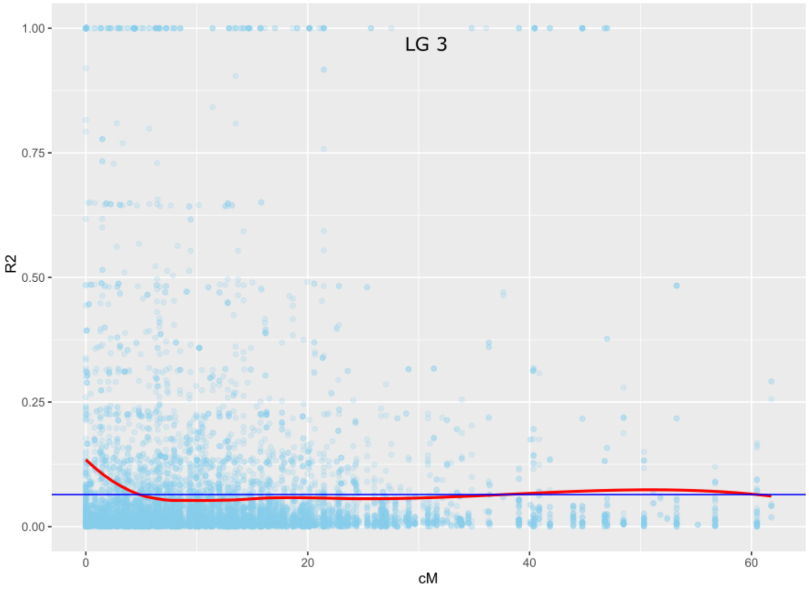

Figure S4.2: Extent of linkage disequilibrium (LD) on linkage group 3, estimated as r^2 value over the distance (cM). Red line represents the smoothing curve fitting using loess curve. Blue line shows the 95th percentile of unlinked loci (background LD)

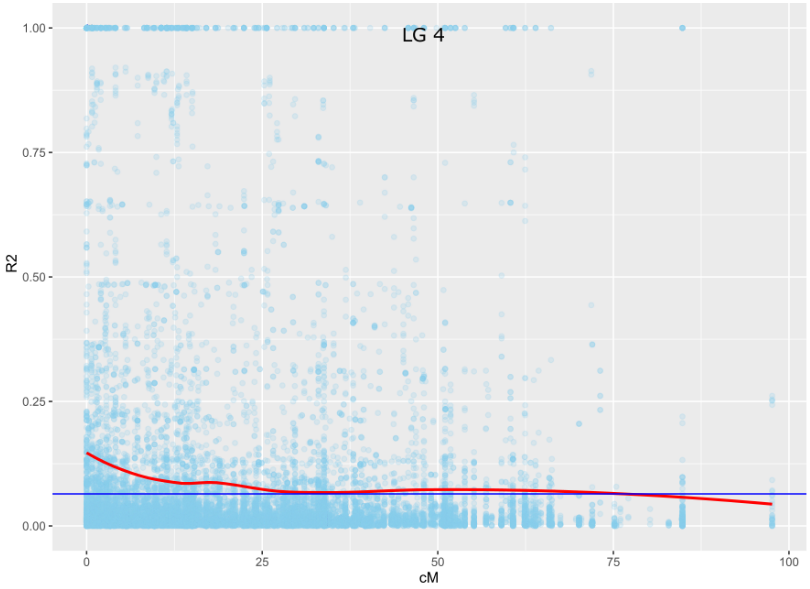

Figure S4.3: Extent of linkage disequilibrium (LD) on linkage group 4, estimated as r^2 value over the distance (cM). Red line represents the smoothing curve fitting using loess curve. Blue line shows the 95th percentile of unlinked loci (background LD)

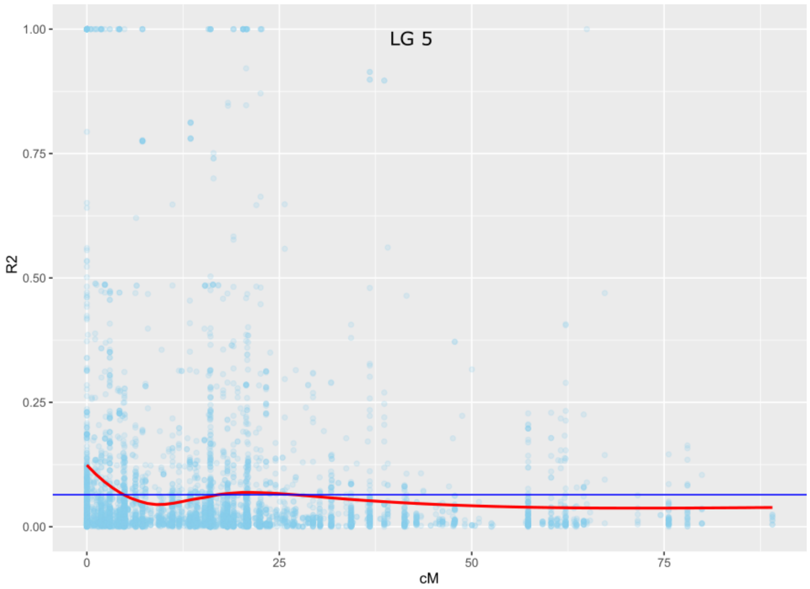

Figure S4.4: Extent of linkage disequilibrium (LD) on linkage group 5, estimated as r^2 value over the distance (cM). Red line represents the smoothing curve fitting using loess curve. Blue line shows the 95th percentile of unlinked loci (background LD)

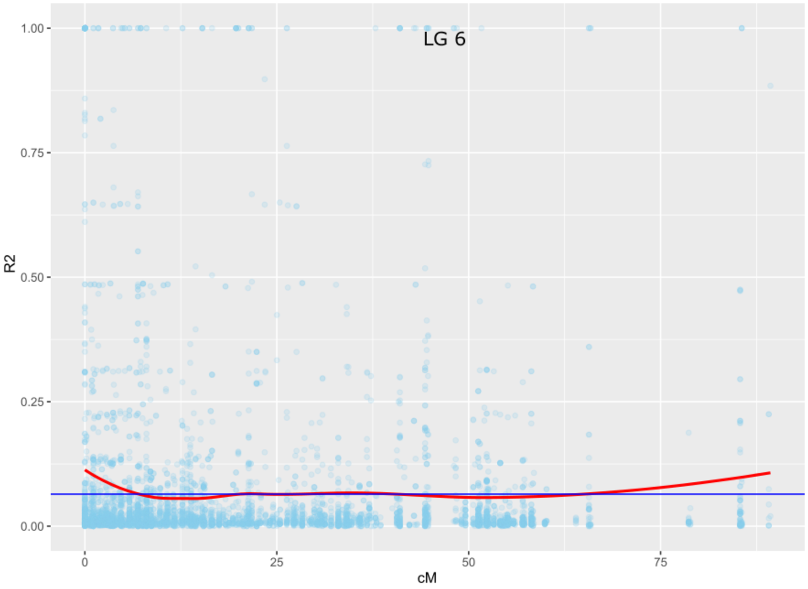

Figure S4.5: Extent of linkage disequilibrium (LD) on linkage group 6, estimated as r^2 value over the distance (cM). Red line represents the smoothing curve fitting using loess curve. Blue line shows the 95th percentile of unlinked loci (background LD)

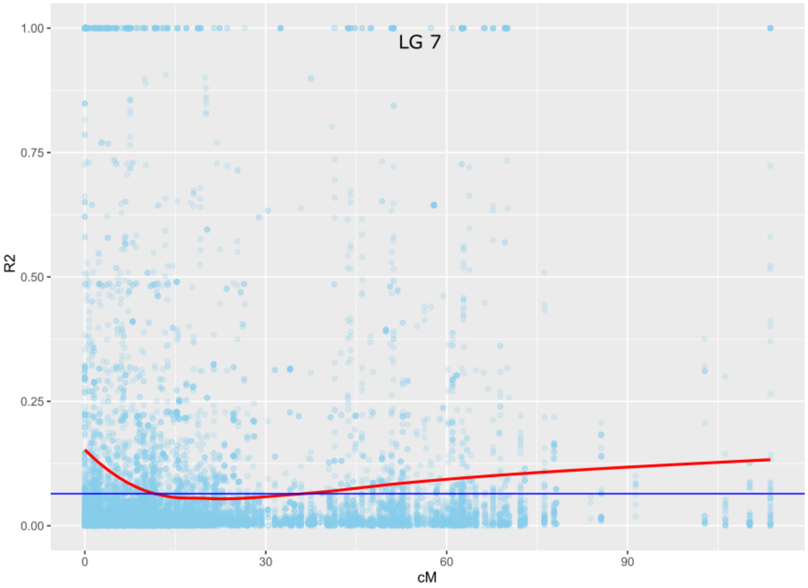

Figure S4.6: Extent of linkage disequilibrium (LD) on linkage group 7, estimated as r^2 value over the distance (cM). Red line represents the smoothing curve fitting using loess curve. Blue line shows the 95th percentile of unlinked loci (background LD)

Chapter 5

General discussion

5.1 General discussion

Conventional plant breeding is largely the art of utilizing natural genetic variation and selecting the best genotypes using the most efficient methods. The typical approach to improve populations in perennial ryegrass is by recurrent selection. Currently, there are two recurrent selection methods used for population improvement in forage grasses, phenotypic and genotypic recurrent selection. Genotypic recurrent selection is the standard breeding system for improving forage yield where breeding values for yield are determined by evaluating either half-sib or full-sib progeny in swards. Phenotypic recurrent selection can be used to improve traits, where there is good agreement between phenotypes scored on spaced plants and phenotypes scored in swards [34]. Genetic gains for economically important traits such as forage yield, dry matter digestibility and plant/genotype persistency have generally been low. Annual genetic gains of 0.3 to 0.5% for forage yield have been reported [55, 82, 116, 187] and no improvements in digestibility and persistency [116]. This is significantly lower than the rates of genetic gain achieved in cereals, where gains of 1.0 to 1.5% per year have been achieved [132]. Reasons include; (i) a longer breeding cycles in forage crops, with each selection cycle taking five to six years, (ii) an inability to exploit heterosis, (iii) selection for multiple traits, which can be poorly correlated or even negatively correlated and (iv) due to the out-breeding nature of perennial ryegrass [28, 82]. In this thesis I have looked at various approaches to use molecular markers and genomic information to increase the rate of genetic gain for traits especially crown rust resistance and forage yield.

Indirect selection based on molecular markers (marker assisted recurrent selection or MARS) was identified as a promising approach to improve the rate of genetic gain per unit time and cost [14]. The goal of MARS is to improve the overall population performance by improving the frequency of favorable alleles [34]. The success of MARS depends on how proximal the marker is to the QTL [14, 31]. The linked markers are used to select seedlings that have the desired allelic combination, thereby reducing the time it takes to complete a cycle of selection. In perennial ryegrass, marker-trait associations have typically been identified in bi-parental mapping populations. However, there are few reports of these markers being used in practical breeding. MARS has had some success in other crops such as rice, maize, soybean and wheat [154, 191]. For example, cyst nematode resistance in soybean was initially screened in glasshouse trials, which took up to 30 days. Using MARS reduced this to 1-2 days, and at a greatly reduced cost [36]. MARS can be successful, especially for traits controlled by few QTLs, each with large effects [14]. In perennial ryegrass economically important traits, especially forage yield are quantitative in nature, and controlled by many genes with relatively small effects. Therefore there is little opportunity to improve such traits with MARS. As mentioned by Conaghan and Casler [34], population size required to introduce one QTL into breeding material is four, but for 10 QTLs the population size required is 1,048,576, making it impractical. The identified QTLs may not even segregate in the breeding material, because the QTL has been identified in a bi-parental population and may already be at a high frequency in breeding material. Furthermore, the number of QTLs identified for the trait doesn't represent all the major QTLs present in the breeding material. Because of this the focus has shifted from identifying QTL in bi-parental populations to identifying QTL directly in breeding material. This results in marker-trait associations that are much more useful in practical breeding.

GWAS overcomes many limitations of bi-parental mapping by identifying QTL segregating in the breeding material. This takes the advantage of historical recombinations to identify marker-trait associations. The success of GWAS depends on the marker being in LD with a QTL and the QTL allele must be at a high enough frequency in the population to be detected and its effect estimated. GWAS for discovery of markers in perennial ryegrass (for MARS) is challenging because the extent of LD is low and allelic diversity is very high. This indicates a very large past effective population size (N_e) [77], requiring higher marker density and large population sizes to detect marker-trait associations. In Chapter 2 we attempted to overcome the challenges of rare alleles by using a number of full-sib families in an association mapping analysis. We identified no significant marker trait associations for heading date, which we put down to challenges identifying significant associations when correcting for population structure, and also low levels of LD across the full-sib families. However, we were able to identify marker-trait associations when focusing on single marker analysis within each family where we observe long range LD. Many markers were proximal to genes controlling heading date (Chapter 2). Interestingly, another study I contributed to during the course of my PhD used variable importance measures to identify markers capable of predicting heading date in the complete training population and succeeded in identifying markers that were within or proximal to genes with known involvement in heading date [21]. In this case no correction for population structure was performed and markers were simply ranked based on their predictive ability. It was clear from the results that this approach did identify markers in LD with QTL for heading date. The initial experiment described above were carried out on full-sib families, which are a subset of larger population of spaced plants consisting full and half sib families, cultivars and ecotypes. When data for the complete population was available we did association analysis for crown rust resistance. As we reported above, heading date correlates quite well with family structure in this population, therefore GWAS for heading date was avoided. A marker density of 200,000 and a genome wide association panel of 1582 individuals was used for GWAS and we were able to identify 29 markers linked to crown rust resistance (Chapter 3). However, the total phenotypic variance accounted by these significant markers was very low (7%) (Chapter 3). Another study in perennial ryegrass reported GWAS for heading date using a similar size of population but the marker density was five times higher. Even with much higher marker density the overall phenotypic variance explained by the significant markers was very low (20%) [61]. The reasons for this so called "missing heritability" in plants was explained by Brachi et al. [15]. The most likely explanation is stringent thresholds imposed by multiple testing and the control of traits by rare alleles. An allele has to be present in high enough frequency for it effect to be accurately estimated. Strategies to overcome many of the problems associated with GWAS have been implemented in other species, however, they required the development of specific experimental populations. These include Multi-parent Advanced Generation InterCrosses (MAGIC) and nested association mapping (NAM) populations. Producing such populations in self-incompatible outbreeding species such as perennial ryegrass is extremely challenging.

Using GWAS to identify marker-trait associations when traits are controlled by thousands of loci with small effect, such as forage yield, is unlikely to be successful. The next innovation in using markers in breeding was suggested by Meuwissen et al. [123] and is referred to as genomic prediction. Here all markers are used simultaneously to predict breeding values. The major difference between genomic prediction and MARS is that markers do not have to cross significant thresholds to be used as predictors of breeding value . In recent years, genomic prediction has been successfully implemented in dairy cattle, which has reduced progeny testing from 6 years to 1.5 years [94, 124]. Its potential has already been demonstrated in annual crops [29, 165, 175], perennial trees [105] and in forage grasses [4, 21, 60, 61, 72, 108].

In chapter 3, we developed genomic prediction models for crown rust resistance in a large spaced plant population. Results were encouraging with high predictive ability for crown rust resistance (0.52). Genomic prediction for crown rust resistance was previously investigated in perennial ryegrass families. In that case predictions were based on mean genotypes and phenotypes of F_2 families, and similar predictive abilities were observed [60]. Genomic prediction relies on using high density markers to ensure all QTL (even small effect QTL) are in LD with at least one marker and all markers are used to predict accurate breeding values [123]. To achieve this in perennial ryegrass we need very large marker densities given the low levels of LD we observed in this population. However, with only 10,000 markers we were still able to achieve a high predictive ability for crown rust resistance (Chapter 3). This was mainly due to markers capturing genetic relationships among families. While predictive ability due to LD will persist over many generations, predictive ability due to capturing genetic relationships is expected to decay rapidly. Habier et al. [73] demonstrated using simulations that markers capturing genetic relationship significantly contributes to predictive ability. However predictive ability decreased after a few generations due to the decay in genetic relationships. In animal breeding, when marker effects were estimated from one breed (Jersey) of animals, and used to estimate effects in another breed (Holstein) lower accuracies were reported [76]. This was mainly due to lack of genetic relationship between the breeds. Similar observations were reported in our study, when predicting allelic effects in a family of unrelated individuals to the training set, predictive ability was mostly zero or negative depending upon the relationship with the training set (Chapter 3). Other studies in perennial ryegrass also assessed the influence of genetic relationship on predictive ability, by predicting allelic effects between material from different breeding programmes which lead to a loss in predictive ability [21, 61]. In order to predict with high accuracy over several generations (required in practice), predictive ability due to capturing markers in LD with QTL is more important than markers capturing genetic relationships. Given the low levels of LD in perennial ryegrass, high marker densities are required, which increases the overall cost of genotyping and limits our ability to increase selection intensity with genomic prediction. An alternative may be a two step approach similar to MARS where we identify a subset of markers predictive of a trait and use these to generate GEBVs. Genotyping opportunities exist for targeted sequencing of 100s of loci at low cost. Genotyping in Thousands by sequencing (GT-seq) is one method, where Campbell et al. [25] reported that 2068 individuals can be genotyped at 192 targeted loci at a cost of $ 3.98 per sample including DNA isolation and PCR. These approaches will be most beneficial if the markers are based on GWAS ranking, as the predictive ability of these markers will be mainly due to LD with QTL. When we ranked markers based on GWAS significant values (although ignoring significance thresholds) and used these to develop genomic prediction models we observed good predictive ability (Chapter 3). The selected top markers based on the GWAS ranking achieved higher predictive ability compared to a similar number of random markers. While this approach may be possible for traits such as crown rust resistance or quality traits, it is unlikely to be successful for more complex traits such as forage yield.

Genomic prediction will be most beneficial in forage breeding for phenotypes where the correlation between measurements taken on single plants and measurements taken in swards is low. This makes within family selection difficult [161]. Forage yield has poor agreement between spaced plants and swards, and is generally measured as the mean of full- or half-sib progeny grown in swards. Due to high G x E interactions selection decisions are always made after multi-year replicated field trials. Hence, forage yield is the perfect target for implementing genomic prediction. In chapter 4, we evaluated the potential of genomic prediction for forage yield traits in tetraploid perennial ryegrass families. In this study we evaluated accuracy of genomic prediction using genotypes of maternal plants and phenotypes collected on half-sib progeny. This is an ideal evaluation of the potential of genomic prediction for forage yield as it focuses on the additive genetic variation which is relevant for predicting parental breeding values during synthetic cultivar development. Using maternal parents have a genotyping advantages over pooled plants. It is easier to genotype a single plant than a pooled plants, as it requires higher sequencing depth to represent true allele frequency of pooled plants. As discussed previously N_e presents challenges for genomic prediction in that we require large reference populations and very high marker densities. Collecting forage yield data (in swards) from large reference populations is challenging. For example, the Teagasc breeding programme has the capacity to evaluate approximately 250 full-sib or half-sib families in two managements with two replications in each. Expanding this to 1000s of families is not feasible. The approach proposed by Hayes et al. [77] was to seek a reduction in N_e from the outset and operating the breeding programme as a closed system. This would lead to increase in LD and higher accuracies using genomic prediction. In chapter 4 we took advantage of a population that had many of these characteristics. The tetraploid half-sib families we used were developed from a commercial cultivar. The cultivar was developed by intercrossing 75 plants from four full-sib families and a set of 120 plants from the cultivar were polycrossed to produce half-sib families. Forage yield was evaluated on these half-sib families under both simulated grazing management and conservation management over two years. Forage yield is highly variable throughout the year and the relative value of the yield at different times of the year has been captured in pasture profit index (PPI) [117, 119]. The PPI was developed in Ireland to place an economic value on cultivars and assist breeders as a selection index for developing new cultivars for Irish production system. We estimated economic value of the plot using weightings from the PPI. The aim was to evaluate genomic prediction for forage yield traits. Predictive ability for traits ranged between 0.03 and 0.30. It is likely that the restricted population and higher LD we observed have led to our reasonable predictive ability despite such a small reference population. However, predictive ability for summer yield was low and this correlated with low marker-based heritability for this trait. Grinberg et al. [72] also evaluated forage yield in (diploids) perennial ryegrass, using a limited number of families (254) as the reference population. Overall, the predictive abilities were also high reaching up to 0.31. A study in alfalfa used a training population size of 100 [108] and another study with two training populations of 124 and 154 families [4] to predict forage yield. Both studies reported good predictive ability for forage yield. Although, predictive ability for forage yield is not as high as crown rust resistance or heading date these results are still promising. Selection based on field evaluations takes up to six years, whereas indirect selection with molecular markers takes 1 year. Because of this, there is huge potential to complete multiple cycles of selection with genomic prediction in the same time it takes to complete a single cycle using conventional selection. Therefore even with predictive abilities of 0.22 for economic value of a plot the relative selection efficiency of genomic prediction over conventional selection is 2.1 (Chapter 4).

Having markers in LD is beneficial for long term use of genomic prediction. For medium heritability traits not conforming to the infinitesimal model, it is possible to capture LD relationships using approaches such as variable selection and GWAS, potentially identifying smaller, more manageable marker sets for practical application. However, for lower heritability traits, and those with characteristics closer to the infinitesimal model, LD can be artificially inflated by using a restricted population. This will also work for less complex traits.

5.2 Implementing genomic prediction in perennial ryegrass breeding

Implementation of genomic prediction in forage breeding schemes such as those outlined in Chapter 1 require modifications to the programme to really take advantage of genomic prediction. Genomic prediction accuracy can be increased when N_e is small and seeking a reduction in N_e has already been proposed as an approach to implement genomic prediction in forage breeding [77]. Using a restricted population with limited number of founder lines is the potential solution for increasing accuracy. We were able to achieve good predictive ability with modest size of training population when we used such an approach in tetraploid perennial ryegrass (Chapter 4). Therefore, the emphasis should be on re-designing breeding programmes that are pre-disposed to having higher LD from the outset. Predictive accuracies can be improved (and updated) over time by generating new families and carrying out field evaluations. Based on the results of this thesis, I propose a strategy for a genetic improvement scheme for perennial ryegrass, based on implementing genomic prediction in a restricted population (Figure 5.1).

In this scheme, genomic prediction is used to select plants to recombine to produce an improved population. The scheme begins with establishing a spaced plant nursery of 1000 plants and 200 plants are selected for polycross using the breeders visual preference. The maternal plants are genotyped using a GBS approach. After recombining selected plants, the seed of half-sib families is used to establish replicated field trial and evaluated for traits such as yield, disease resistance and persistency. Genomic prediction model can be trained using genotypic information from maternal parents and phenotypes collected on the half-sib progenies. Initially phenotypic records are used for selection of the 20 best performing families (among family selection). Seeds (e.g. 1,500) from these families are germinated and genotyped to generate GEBVs. Selections are made based on GEBVs and 150 plants (10%) are recombined to produce the next generation. This process is repeated until the population exceeds the performance required. New phenotypes can be generated while this is ongoing and as data is available it can be incorporated to update models and improve accuracy of genomic prediction.

5.3 Next generation phenotyping in forage breeding

Apart from genomic prediction there are other innovations which may lead to increased genetic gain in breeding programmes. Application of new phenotyping techniques can significantly improve estimation of breeding values. One example is measuring normalized difference vegetative index (NDVI) on swards repeatedly across growing season using an unmanned aerial vehicle. NDVI has been used for a long time as a predictor of green biomass and seems to be correlated well with grain yield in cereals. I propose that, measured NDVI can be used as secondary trait for forage yield and included in a multivariate genomic prediction model for predicting yields. It was already demonstrated that a multivariate genomic prediction model is better and gives high accuracy compared to an univariate models [93]. A similar approach was performed in wheat, where Sun et al. [168] and Rutkoski et al. [151] included NDVI as a secondary trait in a multivariate model and the predictive ability was improved by 70% for grain yield compared to the univariate model. This approach can also be extended to persistency by phenotyping swards using an image sensing camera. The only challenge to implement this method is that the secondary traits should be highly heritable and correlate well with the target trait. Although, no study has yet reported the heritability estimates of NDVI and image sensing in perennial ryegrass, considering the results from cereals I am very optimistic that this approach will be beneficial for low heritable traits in perennial ryegrass.

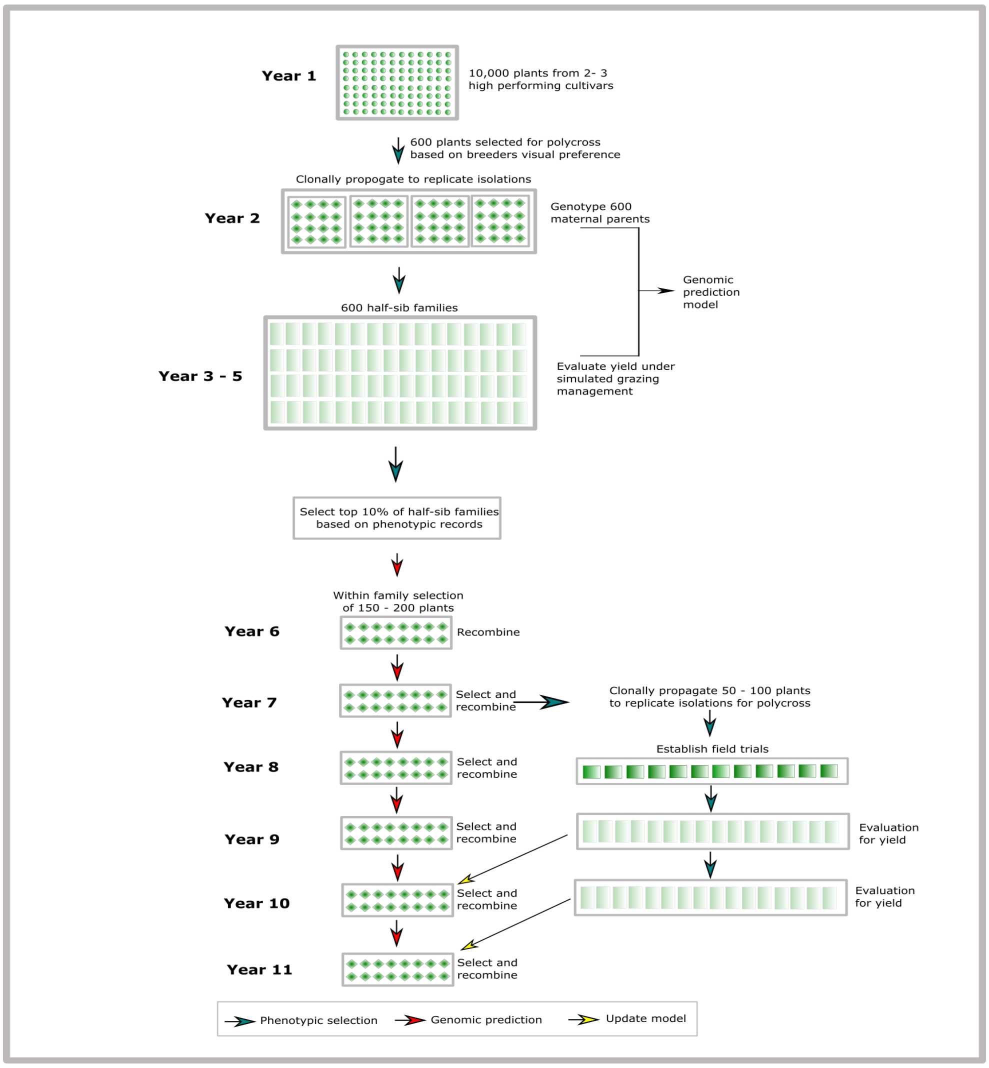
Year 1
10,000 plants from 2- 3 high performing cultivars
600 plants selected for polycross based on breeders visual preference
Clonally propogate to replicate isolations
Year 2
Genotype 600 maternal parents
Genomic prediction model
600 half-sib families
Year 3 - 5
Evaluate yield under simulated grazing management
Select top 10% of half-sib families based on phenotypic records
Within family selection of 150 - 200 plants
Year 6
Recombine
Year 7
Select and recombine
Clonally propagate 50 - 100 plants to replicate isolations for polycross
Establish field trials
Year 8
Select and recombine
Year 9
Select and recombine
Evaluation for yield
Year 10
Select and recombine
Evaluation for yield
Year 11
Select and recombine
Phenotypic selection
Genomic prediction
Update model

Figure 5.1: A proposed genomic prediction scheme for perennial ryegrass. (i) Selection cycle begins with the establishment of a spaced plant nursery of 10,000 plants. (ii) About 600 plants are selected for a polycross based on survivals to produce half-sib families. (iii) These 600 maternal parents are genotyped for GBS markers. (iv) Half-sib families are established in replicated plots and evaluated for forage yields, disease resistance, quality and persistency. (v) Data from the phenotypic information and genotyped maternal parents is used for genomic prediction model development. (vi) Initial selections are based on the phenotypic information to select best families and then random seeds from these selected families are germinated and genotyped, to select best plants within each family. (vii) Based on the ranking of GEBVs, about 150-200 plants are selected and recombined and in the next cycle recombined seeds are germinated and genotyped to select best plants. This process can be repeated for multiple cycles. New phenotype data can be added to update models and improve predictive accuracy

5.4 Conclusions

Genomic prediction can reduce the length of time it takes to complete a single cycle of selection from six years to one year, and therefore increase the rate of genetic gain for economically important traits. Our results for crown rust resistance and forage yield were encouraging. I have reported on the first use of genomic prediction for economic value of a plot in a tetraploid perennial ryegrass population. Results were encouraging and on the back of these genomic prediction is now being implemented in the tetraploid forage breeding programme.

Publication list

Peer-reviewed journal publications

Barth, S., McGrath, S.K., **Arojju, S.K.,** and Hodkinson, T.R., 2015. An Irish perennial ryegrass genetic resource collection clearly divides into two major gene pools. Plant Genetic Resources, 15(3), pp.1-10.

Arojju, S.K., Barth, S., Milbourne, D., Conaghan, P., Velmurugan, J., Hodkinson, T.R. and Byrne, S.L., 2016. Markers associated with heading and aftermath heading in perennial ryegrass full-sib families. BMC Plant Biology, 16(1), p.16.

Byrne, S.L., Conaghan, P., Barth, S., **Arojju, S.K.,** Casler, M, Micheal T, Velmurugan, J. and Milbourne, D. 2017. Using variable importance measures to identify a small set of SNPs to predict heading date in perennial ryegrass. Scientific Reports, 7(1):3566.

Manuscripts

Arojju, S.K., Conaghan, P., Barth, S., Milbourne, D., Hodkinson, T.R., Micheal Casler. and Byrne, S.L. Genomic prediction of crown rust resistance in *Lolium perenne*. Submitted to BMC Genetics.

Arojju, S.K., Conaghan, P., Barth, S., Milbourne, D., Hodkinson, T.R., and Byrne, S.L. Using genome wide markers to predict yield in tetraploid perennial ryegrass families. In preparation, aiming for BMC Genomics

Peer-reviewed book chapter

Arojju, S.K., Milbourne, D., Conaghan, P., Hodkinson, T.R. and Barth, S., 2016. Extent of Crown Rust Infection in a Perennial Ryegrass (Lolium perenne L.) Association Mapping Population. In Breeding in a World of Scarcity (pp. 47-52). Springer International Publishing.

Invited speaker

Plant and animal genome (PAG XXV) conference, San Diego, USA (2017). Title: Genomic prediction for heading date and crown rust in perennial ryegrass (Oral)

Attended conferences

Walsh fellow seminar, Dublin, Ireland (2016). Title: Using genomic selection to increase genetic gain in perennial ryegrass breeding program (Poster)

Eucarpia symposium section fodder crops and amenity grasses, Brussels, Belgium (2015). Title: Extent of crown rust infection in perennial ryegrass association mapping population (Oral)

Irish plant scientist's association meeting, Maynooth, Ireland (2015). Title: Extent of crown rust infection in perennial ryegrass association mapping population (Poster).

References

[1] Adetunji, I., Willems, G., Tschoep, H., Bürkholz, A., Barnes, S., Boer, M., Malosetti, M., Horemans, S., and van Eeuwijk, F. (2014). Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. *Theoretical and applied genetics*, 127(3):559–571.

[2] Andersen, J. R., Jensen, L. B., Asp, T., and Lübberstedt, T. (2006). Vernalization response in perennial ryegrass (*Lolium perenne* L.) involves orthologues of diploid wheat (*Triticum monococcum*) VRN1 and rice (*Oryza sativa*) Hd1. *Plant Molecular Biology*, 60(4):481–494.

[3] Andrés, F. and Coupland, G. (2012). The genetic basis of flowering responses to seasonal cues. *Nature Reviews Genetics*, 13(9):627–639.

[4] Annicchiarico, P., Nazzicari, N., Li, X., Wei, Y., Pecetti, L., and Brummer, E. C. (2015). Accuracy of genomic selection for alfalfa biomass yield in different reference populations. *BMC Genomics*, 16(1):1020.

[5] Armstead, I., Turner, L., Marshall, A., Humphreys, M., King, I., and Thorogood, D. (2008). Identifying genetic components controlling fertility in the outcrossing grass species perennial ryegrass (*Lolium perenne*) by quantitative trait loci analysis and comparative genetics. *New Phytologist*, 178(3):559–571.

[6] Armstead, I. P., Turner, L. B., Farrell, M., Skøt, L., Gomez, P., Montoya, T., Donnison, I. S., King, I. P., and Humphreys, M. O. (2004). Synteny between a major heading-date QTL in perennial ryegrass (*Lolium perenne* L.) and the Hd3 heading-date locus in rice. *Theoretical and Applied Genetics*, 108(5):822–828.

[7] Armstrong, S. F. (1921). *British grasses and their employment in agriculture.* CUP Archive.

[8] Auzanneau, J., Huyghe, C., Julier, B., and Barre, P. (2007). Linkage disequilib-

rium in synthetic varieties of perennial ryegrass. *Theoretical and Applied Genetics*, 115(6):837–847.

[9] Barre, P., Moreau, L., Mi, F., Turner, L., Gastal, F., Julier, B., and Ghesquière, M. (2009). Quantitative trait loci for leaf length in perennial ryegrass (*Lolium perenne* L.). *Grass and Forage Science*, 64(3):310–321.

[10] Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1(7).

[11] Baumann, U., Juttner, J., Bian, X., and Langridge, P. (2000). Self-incompatibility in the grasses. *Annals of Botany*, 85(suppl_1):203–209.

[12] Beavis, W. D. (1998). QTL analyses: power, precision, and accuracy. *Molecular dissection of complex traits*, 1998:145–162.

[13] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

[14] Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Science*, 48(5):1649–1664.

[15] Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology*, 12(10):232.

[16] Breese, E. and Hayward, M. (1972). The genetic basis of present breeding methods in forage crops. *Euphytica*, 21(2):324–336.

[17] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[18] Breseghello, F. and Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (*Triticum aestivum L.*) cultivars. *Genetics*, 172(2):1165–1177.

[19] Brummer, E. C. and Casler, M. D. (2009). Improving selection in forage, turf, and biomass crops using molecular markers. In *Molecular breeding of forage and turf*, pages 193–210. Springer.

[20] Buffalo (2011 (accessed November 7, 2015)). Scythe - a bayesian adapter trimmer version 0.994 beta. `https://github.com/vsbuffalo/scythe`.

[21] Byrne, S., Conaghan, P., Barth, S., Arojju, S. K., Casler, M., Thibauld, M.,

Velmurugan, J., and Milbourne, D. (2017). Using variable importance measures to identify a small set of SNPs to predict heading date in perennial ryegrass. *Scientific Reports*, 7.

[22] Byrne, S., Guiney, E., Barth, S., Donnison, I., Mur, L. A., and Milbourne, D. (2009). Identification of coincident QTL for days to heading, spike length and spikelets per spike in *Lolium perenne* L. *Euphytica*, 166(1):61–70.

[23] Byrne, S. L., Nagy, I., Pfeifer, M., Armstead, I., Swain, S., Studer, B., Mayer, K., Campbell, J. D., Czaban, A., Hentrup, S., et al. (2015). A synteny-based draft genome sequence of the forage grass *Lolium perenne*. *The Plant Journal*, 84(4):816–826.

[24] Camlin, M. and Stewart, R. (1975). Reaction of Italian ryegrass cultivars under grazing as compared with cutting. *Grass and Forage Science*, 30(2):121–130.

[25] Campbell, N. R., Harmon, S. A., and Narum, S. R. (2015). Genotyping-in-thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, 15(4):855–867.

[26] Carr, A. (1975). Diseases of herbage crops-some problems and progress. *Annals of Applied Biology*, 81(2):235–239.

[27] Casler, M. and Vogel, K. (1999). Accomplishments and impact from breeding for increased forage nutritional value. *Crop Science*, 39(1):12–20.

[28] Casler, M. D. and Brummer, E. C. (2008). Theoretical expected genetic gains for among-and-within-family selection methods in perennial forage crops. *Crop Science*, 48(3):890–902.

[29] Cericola, F., Jahoor, A., Orabi, J., Andersen, J. R., Janss, L. L., and Jensen, J. (2017). Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. a case of study in advanced wheat breeding lines. *PLOS ONE*, 12(1):e0169606.

[30] Chang, C. C., Chow, C. C., Tellier, L., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(7).

[31] Collard, B., Jahufer, M., Brouwer, J., and Pang, E. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for

crop improvement: the basic concepts. *Euphytica*, 142(1-2):169–196.

[32] Collard, B. C. and Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1491):557–572.

[33] Conaghan, P., Casler, M., McGilloway, D., O'Kiely, P., and Dowley, L. (2008a). Genotype× environment interactions for herbage yield of perennial ryegrass sward plots in Ireland. *Grass and Forage Science*, 63(1):107–120.

[34] Conaghan, P. and Casler, M. D. (2011). A theoretical and practical analysis of the optimum breeding system for perennial ryegrass. *Irish Journal of Agricultural and Food Research*, pages 47–63.

[35] Conaghan, P., Casler, M. D., O'Kiely, P., and Dowley, L. J. (2008b). Efficiency of indirect selection for dry matter yield based on fresh matter yield in perennial ryegrass sward plots. *Crop Science*, 48(1):127–133.

[36] Concibido, V. C., Diers, B. W., and Arelli, P. R. (2004). A decade of QTL mapping for cyst nematode resistance in soybean. *Crop Science*, 44(4):1121–1131.

[37] Connolly, V. (2001). *Breeding improved varieties of perennial ryegrass.* Teagasc, Crops Research Centre.

[38] Consortium, I. B. G. S. et al. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 491(7426):711–716.

[39] Consortium, U. et al. (2014). Uniprot: a hub for protein information. *Nucleic Acids Research*, page gku989.

[40] Cornish, M., Hayward, M., and Lawrence, M. (1980). Self-incompatibility in ryegrass. *Heredity*, 44(1):55–62.

[41] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[42] Daetwyler, H., Kemper, K., Van der Werf, J., and Hayes, B. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population. *Journal of animal science*, 90(10):3375–3384.

[43] de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327–345.

[44] De Roos, A., Hayes, B., and Goddard, M. (2009). Reliability of genomic predictions across multiple populations. *Genetics*, 183(4):1545–1553.

[45] 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., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498.

[46] Desta, Z. A. and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, 19(9):592–601.

[47] Devos, K. M. and Gale, M. D. (1997). Comparative genetics in the grasses. *Plant Molecular Biology*, 35(1-2):3–15.

[48] Dillon, P. (2007). Achieving high dry-matter intake from pasture with grazing dairy cows. *Frontis*, 18:1–26.

[49] Doyle, C. and Elliott, J. (1983). Putting an economic value on increases in grass production. *Grass and Forage Science*, 38(3):169–177.

[50] Doyle, J. (1991). DNA protocols for plants. In *Molecular techniques in taxonomy*, pages 283–293. Springer, Berlin Heidelberg.

[51] Dracatos, P., Cogan, N., Dobrowolski, M. P., Sawbridge, T., Spangenberg, G., Smith, K. F., and Forster, J. (2008). Discovery and genetic mapping of single nucleotide polymorphisms in candidate genes for pathogen defence response in perennial ryegrass (Lolium perenne L.). *Theoretical and Applied Genetics*, 117(2):203–219.

[52] Dracatos, P. M., Cogan, N. O., Sawbridge, T. I., Gendall, A. R., Smith, K. F., Spangenberg, G. C., and Forster, J. W. (2009). Molecular characterisation and genetic mapping of candidate genes for qualitative disease resistance in perennial ryegrass (Lolium perenne L.). *BMC Plant Biology*, 9(1):1.

[53] Dumsday, J., Smith, K., Forster, J., and Jones, E. (2003). SSR-based genetic linkage analysis of resistance to crown rust (*Puccinia coronata* f. sp. *lolii*) in perennial ryegrass (*Lolium perenne*). *Plant Pathology*, 52(5):628–637.

[54] Easton, H., Cooper, B., Frasers, T., and Widdup, K. (1989). Crown rust on perennial ryegrass in field trials. In *Proceedings of the New Zealand Grassland Association*, volume 50, pages 253–254.

[55] Easton, S., Amyes, J., Cameron, N., Green, R., Kerr, G., Norriss, M., and

Stewart, A. (2002). Pasture plant breeding in New Zealand: where to from here? In *PROCEEDINGS OF THE CONFERENCE-NEW ZEALAND GRASSLAND ASSOCIATION*, pages 173–180.

[56] Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

[57] Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS ONE*, 6(5):e19379.

[58] Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, 4(3):250–255.

[59] Faville, M., Jahufer, M., Hume, D., Cooper, B., Pennell, C., Ryan, D., and Easton, H. (2012). Quantitative trait locus mapping of genomic regions controlling herbage yield in perennial ryegrass. *New Zealand Journal of Agricultural Research*, 55(3):263–281.

[60] Fè, D., Ashraf, B. H., Pedersen, M. G., Janss, L., Byrne, S., Roulund, N., Lenk, I., Didion, T., Asp, T., Jensen, C. S., et al. (2016). Accuracy of genomic prediction in a commercial perennial ryegrass breeding program. *The Plant Genome.*

[61] Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B. H., Pedersen, M. G., Roulund, N., Asp, T., Janss, L., Jensen, C. S., et al. (2015a). Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics*, 16(1):921.

[62] Fè, D., Pedersen, M. G., Jensen, C. S., and Jensen, J. (2015b). Genetic and environmental variation in a commercial breeding program of perennial ryegrass. *Crop Science*, 55(2):631–640.

[63] Fiil, A., Lenk, I., Petersen, K., Jensen, C. S., Nielsen, K. K., Schejbel, B., Andersen, J. R., and Lübberstedt, T. (2011). Nucleotide diversity and linkage disequilibrium of nine genes with putative effects on flowering time in perennial ryegrass (*Lolium perenne* L.). *Plant Science*, 180(2):228–237.

[64] Fischer, G., Prieler, S., van Velthuizen, H., Berndes, G., Faaij, A., Londo, M., and de Wit, M. (2010). Biofuel production potentials in europe: Sustainable use of cultivated land and pastures, part ii: Land use scenarios. *Biomass and Bioenergy*, 34(2):173–187.

[65] Frandsen, K. (1986). Variability and inheritance of digestibility in Perennial Ryegrass (*Lolium perenne*), Meadow Fescue (*Festuca pratensis*) and Cocksfoot (*Dactylis glomerata*) ii. f1 and f2 progeny. *Acta Agriculturae Scandinavica*, 36(3):241–263.

[66] Frank, A. and Hofmann, L. (1994). Light quality and stem numbers in cool-season forage grasses. *Crop Science*, 34(2):468–473.

[67] Gabaldón, T. (2008). Large-scale assignment of orthology: back to phylogenetics. *Genome Biology*, 9(10):235.

[68] Gagic, M., Faville, M., Kardailsky, I., and Putterill, J. (2015). Comparative genomics and functional characterisation of the gigantea gene from the temperate forage perennial ryegrass *Lolium perenne*. *Plant Molecular Biology Reporter*, 33(4):1098–1106.

[69] Gaunt, T. R., Rodríguez, S., and Day, I. N. (2007). Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool'cubex'. *BMC Bioinformatics*, 8(1):428.

[70] Gibbs, J. (1966). Field resistance in *Lolium* sp. to leaf rust (*Puccinia coronata*). *Nature*, 209(5021):420–420.

[71] Gilliland, T. and Mann, R. (2000). Effect of sward cutting management on the relative performance of perennial ryegrass varieties. *The Journal of Agricultural Science*, 135(2):113–122.

[72] Grinberg, N. F., Lovatt, A., Hegarty, M., Lovatt, A., Skøt, K. P., Kelly, R., Blackmore, T., Thorogood, D., King, R. D., Armstead, I., et al. (2016). Implementation of genomic prediction in *Lolium perenne* (L.) breeding populations. *Frontiers in Plant Science*, 7:133.

[73] Habier, D., Fernando, R., and Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397.

[74] Habier, D., Fernando, R. L., and Dekkers, J. C. (2009). Genomic selection using low-density marker panels. *Genetics*, 182(1):343–353.

[75] Hayes, B. and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding this article is one of a selection of papers from the conference "exploiting genome-wide association in oilseed brassicas: a model for

genetic improvement of major OECD crops for sustainable farming". *Genome*, 53(11):876–883.

[76] Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., and Goddard, M. E. (2009a). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution*, 41(1):51.

[77] Hayes, B. J., Cogan, N. O., Pembleton, L. W., Goddard, M. E., Wang, J., Spangenberg, G. C., and Forster, J. W. (2013). Prospects for genomic selection in forage plant species. *Plant Breeding*, 132(2):133–143.

[78] Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(1):47–60.

[79] Hayward, M. (1977). Genetic control of resistance to crown rust (*Puccinia coronato* Corda) in *Lolium perenne* L. and its implications in breeding. *Theoretical and Applied Genetics*, 51(2):49–53.

[80] Hayward, M. and Vivero, J. (1984). Selection for yield in *Lolium perenne*. ii. performance of spaced plant selections under competitive conditions. *Euphytica*, 33(3):787–800.

[81] Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science*, 49(1):1–12.

[82] Humphreys, M. (1997). The contribution of conventional plant breeding to forage crop improvement. In *Proceedings 18th International Grassland Congress'.(Association Management Centre: Calgary, Canada)*.

[83] Humphreys, M., Feuerstein, U., Vandewalle, M., and Baert, J. (2010). Ryegrasses. In *Fodder crops and amenity grasses*, pages 211–260. Springer.

[84] Humphreys, M. O. (2005). Genetic improvement of forage crops–past, present and future. *The Journal of Agricultural Science*, 143(06):441–448.

[85] Humphreys, M. W., Yadav, R., Cairns, A. J., Turner, L., Humphreys, J., and Skøt, L. (2006). A changing climate for grassland research. *New Phytologist*, 169(1):9–26.

[86] Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics*, 128(1):145–158.

[87] Iwata, H., Gaston, A., Remay, A., Thouroude, T., Jeauffre, J., Kawamura, K., Oyant, L. H.-S., Araki, T., Denoyes, B., and Foucher, F. (2012). The TFL1 homologue KSN is a regulator of continuous flowering in rose and strawberry. *The Plant Journal*, 69(1):116–125.

[88] Jafari, A., Connolly, V., and Walsh, E. (2003). Genetic analysis of yield and quality in full-sib families of perennial ryegrass (*Lolium perenne* L.) under two cutting managements. *Irish Journal of Agricultural and Food Research*, pages 275–292.

[89] Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics*, 9(2):166–177.

[90] Jensen, C. S., Salchert, K., Andersen, C., Didion, T., and Nielsen, K. K. (2004). Floral inhibition in red fescue (*Festuca rubra* L.) through expression of a heterologous flowering repressor from *Lolium*. *Molecular Breeding*, 13(1):37–48.

[91] Jensen, C. S., Salchert, K., and Nielsen, K. K. (2001). A TERMINAL FLOWER1-like gene from perennial ryegrass involved in floral transition and axillary meristem identity. *Plant Physiology*, 125(3):1517–1528.

[92] Jensen, L. B., Andersen, J. R., Frei, U., Xing, Y., Taylor, C., Holm, P. B., and Lübberstedt, T. (2005). QTL mapping of vernalization response in perennial ryegrass (*Lolium perenne* L.) reveals co-location with an orthologue of wheat VRN1. *Theoretical and Applied Genetics*, 110(3):527–536.

[93] Jia, Y. and Jannink, J.-L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, 192(4):1513–1522.

[94] Jonas, E. and Koning, D.-J. d. (2015). Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. *Frontiers in Genetics*, 6:49.

[95] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences: CABIOS*, 8(3):275–282.

[96] Joshi NA, F. J. (2011 (accessed November 7, 2015)a). Sabre - a barcode demultiplexing and trimming tool for fastq files. `https://github.com/najoshi/sabre`.

[97] Joshi NA, F. J. (2011 (accessed November 7, 2015)b). Sickle

- a windowed adaptive trimming tool for fastq files using quality. https://github.com/ucdavis-bioinformatics/sickle.

[98] Karlgren, A., Gyllenstrand, N., Källman, T., Sundström, J. F., Moore, D., Lascoux, M., and Lagercrantz, U. (2011). Evolution of the PEBP gene family in plants: functional diversification in seed plant evolution. *Plant Physiology*, 156(4):1967–1977.

[99] Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M.-I., Lin, X., et al. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815.

[100] Kimbeng, C. (1999). Genetic basis of crown rust resistance in perennial ryegrass, breeding strategies, and genetic variation among pathogen populations: a review. *Animal Production Science*, 39(3):361–378.

[101] Kinoshita, T., Ono, N., Hayashi, Y., Morimoto, S., Nakamura, S., Soda, M., Kato, Y., Ohnishi, M., Nakano, T., Inoue, S.-i., et al. (2011). FLOWERING LOCUS T regulates stomatal opening. *Current Biology*, 21(14):1232–1238.

[102] Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M., and Araki, T. (1999). A pair of related genes with antagonistic roles in mediating flowering signals. *Science*, 286(5446):1960–1962.

[103] Kruijer, W., Boer, M. P., Malosetti, M., Flood, P. J., Engel, B., Kooke, R., Keurentjes, J. J., and van Eeuwijk, F. A. (2015). Marker-based estimation of heritability in immortal populations. *Genetics*, 199(2):379–398.

[104] Kumar, S., Bink, M. C., Volz, R. K., Bus, V. G., and Chagné, D. (2012). Towards genomic selection in apple (Malus× domestica Borkh.) breeding programmes: prospects, challenges and strategies. *Tree Genetics & Genomes*, 8(1):1–14.

[105] Kwong, Q. B., Ong, A. L., Teh, C. K., Chew, F. T., Tammi, M., Mayes, S., Kulaveerasingam, H., Yeoh, S. H., Harikrishna, J. A., and Appleton, D. R. (2017). Genomic selection in commercial perennial crops: applicability and improvement in oil palm (*Elaeis guineensis* Jacq.). *Scientific Reports*, 7.

[106] Lancashire, J. and Latch, G. (1970). The effect of crown rust (*Puccinia coronata* Corda) on the yield and botanical composition of two ryegrass/white clover

pastures. *New Zealand Journal of Agricultural Research*, 13(2):279–286.

[107] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[108] Li, X., Wei, Y., Acharya, A., Hansen, J. L., Crawford, J. L., Viands, D. R., Michaud, R., Claessens, A., and Brummer, E. C. (2015). Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. *The Plant Genome*, 8(2).

[109] Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., and Brummer, E. C. (2014). A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-by-sequencing is highly syntenous with the Medicago truncatula genome. *G3: Genes, Genomes, Genetics*, 4(10):1971–1979.

[110] Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3):18–22.

[111] Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., and Zhang, Z. (2012). Gapit: genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397–2399.

[112] Liu, H., Zhou, H., Wu, Y., Li, X., Zhao, J., Zuo, T., Zhang, X., Zhang, Y., Liu, S., Shen, Y., et al. (2015). The impact of genetic relationship and linkage disequilibrium on genomic selection. *PLOS ONE*, 10(7):e0132379.

[113] Lorenz, A., Smith, K., and Jannink, J.-L. (2012). Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. *Crop Science*, 52(4):1609–1621.

[114] Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E., and Jannink, J.-L. (2011). Genomic selection in plant breeding: knowledge and prospects. *Advances in Agronomy*, 110:77.

[115] Loukoianov, A., Yan, L., Blechl, A., Sanchez, A., and Dubcovsky, J. (2005). Regulation of VRN-1 vernalization genes in normal and transgenic polyploid wheat. *Plant Physiology*, 138(4):2364–2373.

[116] McDonagh, J., O'Donovan, M., McEvoy, M., and Gilliland, T. (2016). Genetic gain in perennial ryegrass (*Lolium perenne*) varieties 1973 to 2013. *Euphytica*, 212(2):187–199.

[117] McEvoy, M., McHugh, N., O'Donovan, M., Grogan, D., Shalloo, L., et al.

(2014). Pasture profit index: updated economic values and inclusion of persistency. In *EGF at 50: The future of European grasslands. Proceedings of the 25th General Meeting of the European Grassland Federation, Aberystwyth, Wales, 7-11 September 2014*, pages 843–845. IBERS, Aberystwyth University.

[118] McEvoy, M., O'Donovan, M., and Shalloo, L. (2010). Evaluating the economic performance of grass varieties. *Advances in Animal Biosciences*, 1(1):328.

[119] McEvoy, M., O'Donovan, M., and Shalloo, L. (2011). Development and application of an economic ranking index for perennial ryegrass cultivars. *Journal of Dairy Science*, 94(3):1627–1639.

[120] McGrath, S., Hodkinson, T., and Barth, S. (2007). Extremely high cytoplasmic diversity in natural and breeding populations of *Lolium* (poaceae). *Heredity*, 99(5):531–544.

[121] McGrath, S., Hodkinson, T., Charles, T., Zen, D., and Barth, S. (2010). Variation in inflorescence characters and inflorescence development in ecotypes and cultivars of *Lolium perenne* L. *Grass and Forage Science*, 65(4):398–409.

[122] McVeigh, K. (1975). Breeding for resistance to crown rust (*Puccinia coronata Corda var. lolii Brown*) in turf-type perennial ryegrass (*Lolium perenne*). *New Brunswick, NJ, USA: Rutgers University, PhD thesis.*

[123] Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

[124] Meuwissen, T., Hayes, B., and Goddard, M. (2016). Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers*, 6(1):6–14.

[125] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C., and Lin, C. (2015). Misc functions of the department of statistics. *Probability Theory Group (Formerly: E1071), TU Wien.*

[126] Mishra, P. and Panigrahi, K. C. (2015). Gigantea–an emerging story. *Frontiers in Plant Science*, 6.

[127] Moseley, G., Owen, J., and Barriball, P. (2001). Practice into profit': The results of adopting good grassland practice on contrasting dairy farms. In *The Right Mix: Blending Science and Practice in Dairying. Proceedings of the British Grassland Society Winter Meeting*, pages 29–34.

[128] Muranty, H., Troggio, M., Sadok, I. B., Al Rifaï, M., Auwerkerken, A., Banchi,

E., Velasco, R., Stevanato, P., Van De Weg, W. E., Di Guardo, M., et al. (2015). Accuracy and responses of genomic selection on key traits in apple breeding. *Horticulture Research*, 2:15060.

[129] Muylle, H., Baert, J., Van Bockstaele, E., Moerkerke, B., Goetghebeur, E., and Roldàn-Ruiz, I. (2005a). Identification of molecular markers linked with crown rust (*Puccinia coronata* f. sp. *lolii*) resistance in perennial ryegrass (*Lolium perenne*) using AFLP markers and a bulked segregant approach. *Euphytica*, 143(1-2):135–144.

[130] Muylle, H., Baert, J., Van Bockstaele, E., Pertijs, J., and Roldàn-Ruiz, I. (2005b). Four QTLs determine crown rust (*Puccinia coronata* f. sp. *lolii*) resistance in a perennial ryegrass (*Lolium perenne*) population. *Heredity*, 95(5):348–357.

[131] Myers, W. (1939). Colchicine induced tetraploidy in perennial ryegrass *Lolium perenne* L. *Journal of Heredity*, 30(11):499–504.

[132] Öfversten, J., Jauhiainen, L., and Kangas, A. (2004). Contribution of new varieties to cereal yields in finland between 1973 and 2003. *The Journal of Agricultural Science*, 142(3):281–287.

[133] O'Kiely, P. and Flynn, V. (1987). Grass silage. `https://www.teagasc.ie/media/website/animals/beef/grass-silage.pdf`.

[134] O'Donoghue, C., Creamer, R., Crosson, P., Curran, T., Donnellan, T., Farrelly, N., Fealy, R., French, P., Geoghegan, C., Green, S., et al. (2015). Drivers of agricultural land use change in Ireland to 2025.

[135] Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457(7229):551–556.

[136] Peng, F. Y., Hu, Z., and Yang, R.-C. (2015). Genome-wide comparative analysis of flowering-related genes in Arabidopsis, Wheat, and Barley. *International Journal of Plant Genomics*, 2015.

[137] Pfeifer, M., Martis, M., Asp, T., Mayer, K. F., Lübberstedt, T., Byrne, S., Frei, U., and Studer, B. (2013). The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. *Plant Physiology*, 161(2):571–

582.

[138] Plummer, R., Hall, R., and Watt, T. (1990). The influence of crown rust (*Puccinia coronata*) on tiller production and survival of perennial ryegrass (*Lolium perenne*) plants in simulated swards. *Grass and Forage Science*, 45(1):9–16.

[139] Postel, S., Küfner, I., Beuter, C., Mazzotta, S., Schwedt, A., Borlotti, A., Halter, T., Kemmerling, B., and Nürnberger, T. (2010). The multifunctional leucine-rich repeat receptor kinase bak1 is implicated in *Arabidopsis* development and immunity. *European journal of cell biology*, 89(2):169–174.

[140] Potter, L. (1987). Effect of crown rust on regrowth, competitive ability and nutritional quality of perennial and Italian ryegrasses. *Plant Pathology*, 36(4):455–461.

[141] Price, T. (1987). Ryegrass rust in Victoria. *Plant Protect Quart*, 2:189.

[142] Project, I. R. G. S. et al. (2005). The map-based sequence of the rice genome. *Nature*, 436(7052):793–800.

[143] Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B., and Vandepoele, K. (2014). Plaza 3.0: an access point for plant comparative genomics. *Nucleic Acids Research*, page gku986.

[144] Pryce, J. and Daetwyler, H. (2012). Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Animal Production Science*, 52(3):107–114.

[145] Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589.

[146] Reheul, D. and Ghesquiere, A. (1996). Breeding perennial ryegrass with better crown rust resistance. *Plant breeding*, 115(6):465–469.

[147] Resende, M. F., Muñoz, P., Resende, M. D., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F., and Kirst, M. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics*, 190(4):1503–1510.

[148] Riedelsheimer, C., Endelman, J. B., Stange, M., Sorrells, M. E., Jannink, J.-L., and Melchinger, A. E. (2013). Genomic predictability of interconnected biparental maize populations. *Genetics*, 194(2):493–503.

[149] Roderick, H. and Thomas, B. (1997). Infection of ryegrass by three rust fungi (*Puccinia coronata*, *P. graminis* and *P. loliina*) and some effects of temperature on the establishment of the disease and sporulation. *Plant Pathology*, 46(5):751–761.

[150] Roderick, H., Thorogood, D., and Adomako, B. (2000). Temperature-dependent resistance to crown rust infection in perennial ryegrass, *Lolium perenne*. *Plant Breeding*, 119(1):93–95.

[151] Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., and Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3: Genes, Genomes, Genetics*, 6(9):2799–2808.

[152] Saatchi, M., McClure, M. C., McKay, S. D., Rolf, M. M., Kim, J., Decker, J. E., Taxis, T. M., Chapple, R. H., Ramey, H. R., Northcutt, S. L., et al. (2011). Accuracies of genomic breeding values in american angus beef cattle using k-means clustering for cross-validation. *Genetics Selection Evolution*, 43(1):1.

[153] Sallam, A., Endelman, J., Jannink, J.-L., and Smith, K. (2015). Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *The Plant Genome*, 8(1).

[154] Salvi, S. and Tuberosa, R. (2015). The crop QTLome comes of age. *Current opinion in biotechnology*, 32:179–185.

[155] Sampoux, J.-P., Baudouin, P., Bayle, B., Béguier, V., Bourdon, P., Chosson, J.-F., Deneufbourg, F., Galbrun, C., Ghesquière, M., Noël, D., et al. (2011). Breeding perennial grasses for forage usage: An experimental assessment of trait changes in diploid perennial ryegrass (*Lolium perenne* L.) cultivars released in the last four decades. *Field Crops Research*, 123(2):117–129.

[156] Schein, R. D. and Rotem, J. (1965). Temperature and humidity effects on uredospore viability. *Mycologia*, 57(3):397–403.

[157] Schejbel, B., Jensen, L., Xing, Y., and Lübberstedt, T. (2007). QTL analysis of crown rust resistance in perennial ryegrass under conditions of natural and artificial infection. *Plant Breeding*, 126(4):347–352.

[158] Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S.,

Liang, C., Zhang, J., Fulton, L., Graves, T. A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115.

[159] Schubiger, F. X., Baert, J., Cagas, B., Cernoch, V., Chosson, J. F., Czembor, E., Eickmeyer, F., Feuerstein, U., Hartmann, S., Jakesova, H., et al. (2010). The eucarpia multi-site rust evaluation–results 2007. In *Sustainable use of Genetic Diversity in Forage and Turf Breeding*, pages 331–340. Springer.

[160] Shinozuka, H., Cogan, N. O., Spangenberg, G. C., and Forster, J. W. (2012). Quantitative trait locus (QTL) meta-analysis and comparative genomics for candidate gene prediction in perennial ryegrass (*Lolium perenne* L.). *BMC Genetics*, 13(1):101.

[161] Simeão Resende, R. M., Casler, M. D., and Vilela de Resende, M. D. (2014). Genomic selection in forage breeding: accuracy and methods. *Crop Science*, 54(1):143–156.

[162] Skøt, L., Humphreys, J., Humphreys, M. O., Thorogood, D., Gallagher, J., Sanderson, R., Armstead, I. P., and Thomas, I. D. (2007). Association of candidate genes with flowering time and water-soluble carbohydrate content in *Lolium perenne* (L.). *Genetics*, 177(1):535–547.

[163] Skøt, L., Humphreys, M. O., Armstead, I., Heywood, S., Skøt, K. P., Sanderson, R., Thomas, I. D., Chorlton, K. H., and Hamilton, N. R. S. (2005). An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.). *Molecular Breeding*, 15(3):233–245.

[164] Smit, H. J., Tas, B. M., Taweel, H. Z., Tamminga, S., and Elgersma, A. (2005). Effects of perennial ryegrass (*Lolium perenne* L.) cultivars on herbage production, nutritional quality and herbage intake of grazing dairy cows. *Grass and Forage Science*, 60(3):297–309.

[165] Spindel, J., Begum, H., Akdemir, D., Collard, B., Redona, E., Jannink, J., and McCouch, S. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, 116(4):395.

[166] Studer, B., Byrne, S., Nielsen, R. O., Panitz, F., Bendixen, C., Islam, M. S., Pfeifer, M., Lübberstedt, T., and Asp, T. (2012). A transcriptome map of perennial ryegrass (*Lolium perenne* L.). *BMC Genomics*, 13(1):140.

[167] Studer, B., Jensen, L. B., Hentrup, S., Brazauskas, G., Kölliker, R., and Lübberstedt, T. (2008). Genetic characterisation of seed yield and fertility traits in perennial ryegrass (*Lolium perenne* L.). *Theoretical and Applied Genetics*, 117(5):781–791.

[168] Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J.-L., and Sorrells, M. E. (2017). Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *The Plant Genome*.

[169] Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). Mega6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, page mst197.

[170] Tang, Y., Horikoshi, M., and Li, W. (2016). ggfortify: unified interface to visualize statistical results of popular R packages. *The R Journal*, 8(2):478–489.

[171] Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., Chabert-Martinello, M., Magnin-Robert, J.-B., Marget, P., Aubert, G., et al. (2015). Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. *Frontiers in Plant Science*, 6:941.

[172] Team, R. C. (2015). A language and environment for statistical computing. Vienna, Austria. 2014.

[173] Team, R. C. (2016). R: A language and environment for statistical computing. vienna: R foundation for statistical computing; 2014.

[174] Thorogood, D., Paget, M., Humphreys, M., Turner, L., Armstead, I., and Roderick, H. (2001). QTL analysis of crown rust resistance in perennial ryegrass-implications for breeding. *International Turfgrass Society Research Journal*, 9:218–223.

[175] Thorwarth, P., Ahlemeyer, J., Bochard, A.-M., Krumnacker, K., Blümel, H., Laubach, E., Knöchel, N., Cselényi, L., Ordon, F., and Schmid, K. J. (2017). Genomic prediction ability for yield-related traits in german winter barley elite material. *Theoretical and Applied Genetics*, pages 1–15.

[176] Tomaszewski, C., Heslop-Harrison, J. P., Anhalt, U. C., and Barth, S. (2010). Fine mapping of quantitative trait loci for biomass yield in perennial ryegrass. In

Sustainable use of Genetic Diversity in Forage and Turf Breeding, pages 461–464. Springer, Dordrecht.

[177] Valk, P., Proveniers, M., Pertijs, J., Lamers, J., Dun, C., and Smeekens, J. (2004). Late heading of perennial ryegrass caused by introducing an *Arabidopsis* homeobox gene. *Plant Breeding*, 123(6):531–535.

[178] Van Inghelandt, D., Reif, J. C., Dhillon, B. S., Flament, P., and Melchinger, A. E. (2011). Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. *Theoretical and applied genetics*, 123(1):11–20.

[179] Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., Barry, K., Lucas, S., Harmon-Smith, M., Lail, K., et al. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, 463(7282):763–768.

[180] Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G., van Eck, H. J., and van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theoretical and Applied Genetics*, 130(1):123–135.

[181] Waller, R. A. and Sale, P. W. G. (2001). Persistence and productivity of perennial ryegrass in sheep pastures in south-western Victoria: a review. *Animal Production Science*, 41(1):117–144.

[182] Wang, R., Albani, M. C., Vincent, C., Bergonzi, S., Luan, M., Bai, Y., Kiefer, C., Castillo, R., and Coupland, G. (2011). Aa TFL1 confers an age-dependent response to vernalization in perennial *Arabis alpina*. *The Plant Cell*, 23(4):1307–1321.

[183] Whittaker, J. C., Thompson, R., and Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetical research*, 75(02):249–252.

[184] Wilken, D. (1993). Lolium. *The Jepson Manual. Higher Plants of California, University of California Press, Berkeley*, 1400.

[185] Wilkins, P. (1978). Specialisation of crown rust on highly and moderately resistant plants of perennial ryegrass. *Annals of Applied Biology*, 88(1):179–184.

[186] Wilkins, P. (1991). Breeding perennial ryegrass for agriculture. *Euphytica*, 52(3):201–214.

[187] Wilkins, P. and Humphreys, M. (2003). Progress in breeding perennial

forage grasses for temperate agriculture. *The Journal of Agricultural Science*, 140(02):129–150.

[188] Williamson, M. L. (2008). Differential responses of tillers to floral induction in perennial ryegrass (*Lolium perenne* L.): implications for perenniality: a thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Plant Biology at Massey University, Palmerston North, New Zealand.

[189] Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H.-J., Wang, Y., and Schön, C.-C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, 195(2):573–587.

[190] Wray, N. and Visscher, P. (2008). Estimating trait heritability. *Nature Education*, 1(1):29.

[191] Xing, Y. and Zhang, Q. (2010). Genetic and molecular bases of rice yield. *Annual review of plant biology*, 61:421–442.

[192] Yamada, T., Jones, E., Cogan, N., Vecchies, A., Nomura, T., Hisano, H., Shimamoto, Y., Smith, K., Hayward, M., and Forster, J. (2004). QTL analysis of morphological, developmental, and winter hardiness-associated traits in perennial ryegrass. *Crop Science*, 44(3):925–935.

[193] Yamaguchi, A., Kobayashi, Y., Goto, K., Abe, M., and Araki, T. (2005). TWIN SISTER OF FT (TSF) acts as a floral pathway integrator redundantly with FT. *Plant and Cell Physiology*, 46(8):1175–1189.

[194] Yoo, S. J., Chung, K. S., Jung, S. H., Yoo, S. Y., Lee, J. S., and Ahn, J. H. (2010). BROTHER OF FT AND TFL1 (BFT) has TFL1-like activity and functions redundantly with TFL1 in inflorescence meristem development in *Arabidopsis*. *The Plant Journal*, 63(2):241–253.

[195] Zhang, Z., Ding, X., Liu, J., Zhang, Q., and de Koning, D.-J. (2011). Accuracy of genomic prediction using low-density marker panels. *Journal of dairy science*, 94(7):3642–3650.