THE UNIVERSITY OF ADELAIDE

Application and Optimisation of Genomic Selection for Wheat Breeding

Adam NORMAN

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

School of Agriculture, Food and Wine
Plant Genetics, Genomics and Breeding

THE UNIVERSITY of ADELAIDE

June 2019

Contents

Publications

This thesis comprises a collection of research articles either published or in preparation for publication in peer reviewed journals. Statements of authorship are included in the chapters containing a research article.

NORMAN, A., TAYLOR, J., TANAKA, E., TELFER, P., EDWARDS, J., MARTINANT, J., & KUCHEL, H. (2017). Increased genomic prediction accuracy in wheat breeding using a large Australian panel. *Theoretical and Applied Genetics* **130**, 2543-2555.

NORMAN, A., TAYLOR, J., EDWARDS, J., KUCHEL, H. (2018). Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3: Genes—Genomes—Genetics* **8**, 2889-2899.

NORMAN, A., EDWARDS, J., KUCHEL, H. Increasing response to genomic selection through concurrent use of low and high density genotyping platforms. Unpublished, planned for submission in 2019.

Abstract

Plant breeding has a rich history of producing yield gains in bread wheat through the innovation and adoption of new technologies. This result is driven by extensive research, first in developing the technology, and second on establishing its application. Genomic selection is a recent technology which over the past decade has been the focus of extensive research effort. This research has been highly effective at developing the technology, and our attention should now pivot towards establishing and refining the parameters under which it should be applied. If genomic selection is to be successfully implemented in wheat breeding programmes breeders must be better informed on the optimal design of training strategies, and will also require cost-effective genotyping solutions. This body of work concentrates on delivering three overarching intended research outcomes: i) establish the achievable accuracy of genomic prediction in a large breeding population, ii) identify criteria for the optimal design of a genomic selection training strategy, and iii) investigate concepts and formulate methods for reducing the cost of implementing genomic selection. We present a dataset of unprecedented size in genomic selection studies, and utilise it to address these objectives.

In the first component of the project we confirmed the significant potential of genomic selection by producing high prediction accuracies in a large and representative set of breeding germplasm, and showed genomic selection to be more accurate than marker assisted selection in all 14 traits tested. It was also demonstrated that genomic relationship information can be incorporated into the analysis of phenotype data to significantly improve model accuracy. The second component investigated factors affecting genomic prediction accuracy and how these relationships could be exploited in order to efficiently design accurate training strategies. We found that prediction accuracy continued to respond to training set size well beyond sizes previously tested in the literature, and that this response was independent of the genetic complexity of the trait. The impact of relatedness on prediction accuracy was highlighted, and it was shown that accuracy could be improved by increasing relatedness between training and prediction sets, or by increasing the diversity in the training set. To reduce the cost of implementing genomic selection, we present two novel methodologies for accurately utilising a low density genotyping platform. These approaches were shown to significantly increase the rate of genetic gain compared to a high density platform, with the same total genotyping expenditure. They could also be used to lower the cost of genomic selection without sacrificing genetic gain.

The work presented here represents a significant resource which will inform pragmatic plant breeders on how to effectively and efficiently implement genomic selection in their programmes. The findings clarify uncertainties and overcome constraints associated with applying genomic selection, and can therefore be leveraged to facilitate increased rates of genetic gain in wheat breeding programmes around the world.

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the Universitys digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Programme Scholarship.

Signed:

Date: 5/6/2019

Acknowledgements

Over the past five years many people have assisted me not only in completing this PhD, but in guiding me into the plant breeding community where I now work in my dream job. These contributions range all the way from supervising my project, to a five minute conversation which provoked some thought or clarified a concept. I'm grateful for all of this support, and thank everyone who has provided some form of it.

Numerous people from the University of Adelaide helped me through my postgraduate studies and provided valuable assistance by giving feedback on drafts, discussing my project, and helping me complete the formal project milestones. In particular I would like to thank Jason Able, Diane Mather, Ken Chalmers, and Matthew Denton.

I owe considerable thanks to Brian Cullis and Alison Smith for their work in establishing the overarching project, and for providing me with statistical guidance and support. I consider myself very fortunate to have had such access to you and your expertise during my candidature. To Emi Tanaka, a considerable portion of my current skill set stems from the help you provided me in the world of statistics and R. At the beginning of this project I was extremely green in that space, and you bore the brunt of my questions. Thankyou for always being willing to help.

Thanks are due to the amazing AGT Roseworthy team for assistance in the data collection side of the project. What best highlights your contribution is that entering 11000 lines into disease nurseries, genotyping plates and yield plots was actually a very straightforward process. That has no right to be the case and is a testament to your proficiency. Moreover, you provided a great environment to work at everyday and I look forward to enjoying that in the future. Thankyou to the breeding team at AGT, past and present, (Paul Telfer, Stewart Coventry, Dion Bennett, Jason Reinheimer, Dini Ganesalingam, Russell Eastwood,

Britt Kalmeier, Meiqin Lu, and Tom Kapcejevs) for being available to bounce ideas around. Interacting with you has kept me focussed on delivering outputs for plant breeders. To James Walter, thanks for following me out to AGT! Without you I would have missed many important university emails, plus it was great having someone across the foyer to complain to about R troubles. I'm looking forward to working with you in the future.

To my supervisors Julian Taylor, James Edwards and Haydn Kuchel, I simply can't thank you enough. Julian, the help you gave during my fortnightly trips to the Waite campus was always worth the travel, but even if it wasn't I still would have come down just to hang out over a coffee. Thankyou for you for helping me understand the underlying theory of the project, and for your pragmatic approach to getting ideas up and running. I'm looking forward to working with you more in the future. James, I think you win the award for fastest turnarounds in reading drafts! Thanks for always looking at the fine details in my work, and for finding me some PhD time over the last two years. You've had a big role in teaching me applied plant breeding, and I look forward to repaying the time you spent on me in years to come. Haydn, I couldn't have asked for a better supervisor and mentor. You do a great job at balancing productivity and curiosity in the directions you give, which has given me a pragmatic yet ambitious mindset. I am thankful for the opportunities you have provided me, and am grateful for your continued efforts in teaching me how to leverage quantitative genetics to benefit the wheat industry and Australian farmers.

Thankyou to my friends and family for continuously asking when I'll be finished, and for feigning interest while regretting having asked about my project. Mum and Dad, thankyou for always encouraging me to pursue my interests, I'm very grateful for all the support you have given me. The years I spent studying full time would not have been possible without your help. Ella, my work-life balance has definitely improved since we met, thankyou for that! I think if I hadn't met you I'd actually have some difficulty adjusting to life without a PhD project, but with you around I'm simply looking forward to the extra time we have to spend together - let us enjoy it!

Chapter 1

Literature review

1.1 Introduction

With the global population forecast to exceed nine billion by 2050, FAO (2017) predict that food production levels must rise by 60%. Wheat is the second most important crop for human dietary intake (FAO, 2017), and improving its yields could therefore be a key component of achieving this production increase. Plant breeding has been successful in achieving significant yield gains in wheat since the beginning of the 20th century (Wrigley & Rathjen, 1981), and it is critical that this continues if future production is to meet demand. Genomic selection (GS), first proposed by (Meuwissen et al., 2001), has revolutionised dairy cattle breeding and has since generated significant interest in its potential to increase rates of genetic gain in plant breeding (Nakaya & Isobe, 2012). This review defines and analyses the components and dynamics of GS that will determine its effectiveness in a wheat breeding programme.

1.2 Overview of wheat breeding and genomic selection

Wheat breeding through artificial cross-fertilisation began in Australia late in the 19th century (Wrigley & Rathjen, 1981). In the early 20th century, formal wheat breeding developed as the process of making crosses and generating inbred lines, then selecting individuals for variety release (Figure 1.1). Since its inception, wheat breeding has experienced immense change, largely driven by the adoption of various technologies. For example, the use of improved statistical techniques in trial analysis and line selection has revolutionised the way genotypes are tested in the field.

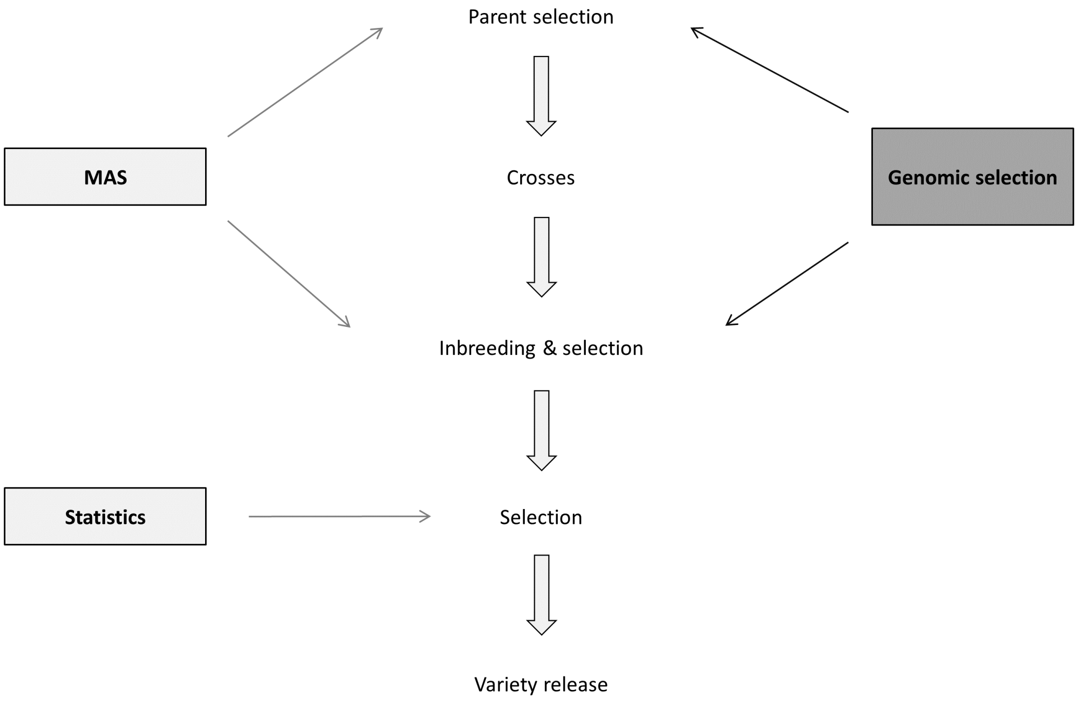

FIGURE 1.1: Core processes of wheat breeding (centre column), showing some technologies that have already been implemented (left) (MAS: marker assisted selection). Genomic selection (right) is a candidate technology yet to be adopted.

Since molecular markers have become available in plant breeding, their application has largely revolved around quantitative trait locus (QTL) and gene mapping, and marker assisted selection (MAS). MAS can assist in parent selection or be employed during the inbreeding stage (Figure 1.1), and is particularly effective when applied to qualitative traits under monogenic control. However, many important traits (e.g. yield) are quantitative in nature with complex polygenic control (Kuchel et al., 2007; Bennett et al., 2012; Maphosa et al., 2014), and it is difficult to improve such traits through MAS (Dekkers et al., 2002). GS is an emerging method in plant breeding, and is more suited to polygenic traits. GS consists of an initial stage of model development, and a subsequent stage of prediction and selection. A collection of lines related to the target germplasm (referred to as the training set) with both phenotype and marker data is employed in the model development stage to determine marker effects. These effects are incorporated into a model and used to predict genomic estimated breeding values (GEBVs) for individuals in the target germplasm. These individuals are therefore selected for crossing or progression to the next generation based solely on their genotype.

The high potential for GS to increase rates of genetic gain is demonstrated by its successfull adoption in the dairy industry. Over the last decade, it has led to genetic gains of up to double that of conventional breeding, and is currently revolutionising the structure of dairy breeding programmes (Bouquet & Juga, 2013). Heffner et al. (2009) reviewed the opportunity for applying GS in plant breeding and identified similarly high potential, discussing aspects such as statistical models, required marker density, maintaining genetic diversity, and genotype by environment (G x E) interactions. Numerous studies have since been undertaken focussing on GS methods and its potential in maize and wheat (Heffner et al., 2010; Jannink et al., 2010; Poland et al., 2012; Schulz-Streeck et al., 2013; Crossa et al., 2014, 2016; He et al., 2016, 2017; González-Camacho et al., 2018; Michel et al., 2018). On a world scale, the vast majority of maize breeding is undertaken by the private sector, and while detailed information is not publicly available, it is understood that commercial companies commonly use genomic prediction techniques in their breeding programmes (Cooper et al., 2014). Wheat breeding is largely private in Europe and Australia, but in North America there are both public and private programmes. Some private companies in North America are currently adopting GS, and several North American university breeding programmes have published research on the topic (Zhong et al., 2009; Heffner et al., 2011a; Juliana et al., 2017). Of the private wheat programmes in Europe and Australia, several are in the early to mid stages of applying GS (H. Kuchel, personal communication).

1.3 Review of statistical methods

While the theory of genomic prediction is built on the same foundational concepts as QTL analysis, there are several fundamental differences that should be acknowledged. This section interrogates the respective statistical models of the two approaches in order to highlight their differences.

1.3.1 QTL analysis

QTL analysis is the process of identifying genomic regions that are linked to traits of interest and estimating their effects; significant QTL can then be employed in MAS (Collard et al., 2005). The concept of mapping QTL on the genome has existed since the beginning of the 20th century where linkage between phenotypes was studied (Sax, 1923; Rasmusson, 1933). Initial approaches to QTL mapping with molecular markers, detailed by Soller et al. (1976), involved analysing individual markers one at a time. Lander & Botstein (1989) derived a simple interval mapping approach that used maximum likelihood for estimation of QTL. This method was shown to be more powerful than single marker regression methods as it distinguishes between weak marker-QTL linkage and small QTL effects. Haley & Knott (1992) developed a regression version of interval mapping which showed similar performance to the maximum-likelihood approach, but was simpler to implement computationally. However, this approach produced biased estimates of the residual variance which affected the power of QTL detection (Xu & Atchley, 1995).

A disadvantage with these approaches is that separate models are used to estimate the effects of individual QTL. This can result in the total variance explained by a QTL to be overestimated (Xu, 2003). To overcome this, a method combining multiple regression analysis and interval mapping was proposed (Jansen, 1993; Zeng, 1993, 1994). Known as composite interval mapping (CIM), this approach uses markers surrounding the interval of interest as covariates to account for the effects of other QTL, thus reducing residual variance (background noise) (Kao et al., 1999). A natural extension of CIM is multiple interval mapping (MIM), proposed by Kao et al. (1999). This approach uses multiple marker intervals simultaneously to detect and estimate multiple putative QTL. This increases the power and accuracy in QTL detection compared to CIM and earlier methods. Sillanpää & Arjas (1998) proposed a novel approach for mapping multiple QTL based on Bayesian hierarchical modelling. This method is similar to MIM and provides a more accurate representation of individual putative QTL by simultaneously estimating the effects of QTL elsewhere on the genome.

Although single marker and interval methods of QTL analysis have been popular, they are unfavourable for several reasons. Firstly, the genetic influence of the complete marker set is not accounted for when estimating the effect of putative QTL. In addition, the majority of approaches do not allow for more complex model structures, such as additional random or fixed components that often arise from plant experiments. Xu (2003) employed a genome wide Bayesian approach detailed by Meuwissen et al. (2001) to simultaneously estimate all marker effects. However, it fails to provide information on the significance of QTL, and so identifying the strongest QTL becomes a challenge. These problems are avoided with the whole genome average interval mapping (WGAIM) algorithm derived by Verbyla et al. (2007). The WGAIM approach integrates whole genome QTL detection and estimation with linear mixed modelling technology. In the initial stage of this algorithm, a base linear mixed model is extended by incorporating the complete set of marker intervals as random effects. An outlier detection method is then used to identify significant putative QTL. The QTL are moved to the fixed effects and the process is repeated until no significant putative QTL are detected. This approach has been shown to be more powerful than single marker and interval methods, and has been computationally implemented in the WGAIM package (Taylor et al., 2011) available in the R statistical computing environment (R Core Team, 2018). Verbyla et al. (2012) discuss a computationally efficient random formulation of WGAIM. This reduces bias in selection and effect estimation of putative QTL, and also decreases the occurrence of false positives.

1.3.2 Genomic prediction

A brief note on terminology, ”genomic prediction” is used here to describe the process of calculating the genomic predictions, where genomic selection refers to both predicting and selecting individuals in a breeding programme. Genomic prediction models generate GEBVs in two stages. Similar to whole genome QTL analysis, the first stage involves the simultaneous estimation of marker effects. Secondly, net predictions are calculated for each individual by summing the marker effects according to their marker-genotype. There are many variations of genomic prediction models; the majority of which reside in the first stage as there are numerous approaches to estimating marker effects. In their seminal study where GS was first proposed, Meuwissen et al. (2001) discuss several variations of genomic prediction models, and many have since been developed.

One common approach to genomic prediction is the ridge regression formulation which was first used in a whole genome QTL analysis context by Whittaker et al. (2000), and was also used for genomic prediction by Heffner et al. (2011b) and Piepho et al. (2012b). Ridge regression unrealistically assumes all marker effects to have equal variance; this assumption results in all marker effects being equally shrunk toward zero, and can cause large effect QTL to be underestimated (Nakaya & Isobe, 2012). Despite this, ridge regression remains a suitable approach for quantitative traits, where small effect QTL are prevalent (Heslot et al., 2012, 2013). The ridge regression approach can be extended by including a reproducing kernel Hilbert space (RKHS) (de Los Campos et al., 2009) in the formulation. Here, marker data is used to calculate the genetic relatedness between individuals. This extension has been applied in GS studies by Crossa et al. (2010) and Heslot et al. (2012). The ridge regression formulations generally require best linear unbiased predictions (BLUP) for the estimation of marker or genotype effects. For this reason, the term BLUP has been used quite broadly in the literature to describe various formulations of ridge regression (Lorenzana & Bernardo, 2009; Crossa et al., 2010).

Another approach to genomic prediction is to use a Bayesian formulation and assign a prior distribution to the variance of individual markers. Meuwissen et al. (2001) proposed two variations known as BayesA and BayesB. BayesA uses an inverse chi-squared prior distribution of individual marker variances that shrinks small marker effects close to zero and enhances important effects. Realistically, some genomic regions will possess no QTL for the trait of interest. The BayesB formulation accounts for this by allowing some markers to have zero effect. This uses a mixture prior distribution of individual marker variances that contains a fixed probability of zero variance. The Bayes-Cπ formulation is an extension of BayesB that estimates this probability (Lorenz et al., 2010; Heslot et al., 2012). Since their initial publication, numerous variations of these models have been proposed and used. Among these is an expectation maximisation (EM) algorithm for the BayesB model detailed by Hayashi & Iwata (2010), and used by Heslot et al. (2012). Empirical Bayes (E-Bayes) is a variation of BayesA that uses a maximisation algorithm to estimate the variance parameters and reduce computation time (Xu, 2007; Heslot et al., 2012). The least absolute shrinkage and selection operator (LASSO) and Elastic Net methods have also been used in Bayesian formulations to shrink marker effects (Crossa et al., 2010; Heslot et al., 2012).

1.3.3 Single and two-stage analyses

The process of generating genomic prediction calibrations can be described as having two components; the first being the computation of adjusted phenotype means, and the second the prediction of molecular marker effects using the adjusted means along with marker genotype data. These processes can either be completed in two separate analyses (two-stage), or together in one analysis (single-stage) (Schulz-Streeck et al., 2013). The adjusted means have important variance-covariance structures associated with them; in a two-stage analysis these structures are lost after the first stage as only the adjusted means are taken to the second. Single-stage analyses on the other hand, can utilise the entire variance-covariance structure. This may not be significant when phenotype data is produced in small replicated field trials, but plant breeding programmes employ large unbalanced multi-environment field trials which produce complex variance-covariance structures. In such trials, the difference between single- and two-stage analyses becomes larger (Piepho et al., 2012a).

1.4 Marker genotyping

For GS to be successfully implemented and economically viable on a commercial scale, a molecular marker platform that provides dense coverage of the genome is required (Heffner et al., 2010; Misztal & Legarra, 2017). Effective application of GS requires the training set to adequately represent the target germplasm. Therefore, if the target germplasm is very large, the training set may need to include thousands of individuals in order to represent it. Larger germplasm collections are capable of yielding more unique marker combinations due to a higher likelihood of recombination events occurring between tightly linked markers. Sufficiently dense marker genotyping will therefore maximise the genetic information represented by the markers, and hence captured in the model (Poland et al., 2012; Hickey et al., 2014). Once GS is applied within a breeding programme, a large number of individuals require marker genotyping per cycle (Heffner et al., 2009; Heslot et al., 2015). The cost of genotyping must therefore be suitably low for GS to be economically viable.

Fortunately, development in marker genotyping methods has been rapid; a decade ago low-throughput gel-based platforms were prevalent (Collard et al., 2005), where now high-throughput sequencer and chip based platforms can generate dense marker maps at realistic prices (Poland & Rife, 2012). Three types of platforms for high-throughput dense marker genotyping are diversity array technology (DArT), chip based single nucleotide polymorphism (SNP) arrays, and genotyping by sequencing (GBS). Chip based SNP arrays are based on oligo ligation assay methods, and have been verified as a reliable and cost-effective method for SNP genotyping in wheat (Akhunov et al., 2009). Here, SNPs are first identified in a diverse germplasm collection, before being transferred to a fixed assay to carry out genotyping (Poland & Rife, 2012). This method is therefore most effective when the germplasm used in SNP identification is representative of the germplasm of interest; when this is not the case, ascertainment bias, and a reduction in polymorphic markers can be expected (Heslot et al., 2013). GBS is a next generation sequencing method that combines these two stages into one, hence enabling simultaneous discovery of SNPs, and generation of their scores (Elshire et al., 2011). GBS has been successfully applied to hexaploid wheat, and also shown to be a suitable platform for use in GS (Poland et al., 2012). DArT is a hybridisation-based technique that scores the presence versus absence of DNA fragments in genomic representations, and has been shown to be an effective method for genotyping the hexaploid bread wheat genome (Akbari et al., 2006). DArT resembles chip based assays in that it requires polymorphisms to be identified in a separate stage prior to them being scored in the germplasm of interest.

Another approach of lowering genotyping cost is to use low and high density platforms concurrently and impute low density progeny data up to high density using parental data. This has been shown to be effective in animal and human genetics when imputing up to sequence data (Kong et al., 2008; Howie et al., 2009; Antolín et al., 2017), but until very recently little work has been carried out to develop suitable methods for highly structured plant populations (Gonen et al., 2018), as earlier methods faced issues with long computational time (Hickey et al., 2015). Gorjanc et al. (2017) showed that if accurate imputation methods were made available for plant breeders, they could be used to improve the overall impact achieved with GS.

1.5 Factors affecting accuracy of genomic estimated breeding values

True GEBV accuracy can be defined as the correlation between the GEBV and the true breeding value (Heffner et al., 2009; Nakaya & Isobe, 2012). However, as the true breeding value cannot be definitively known, accuracy is estimated by the correlation between the GEBV and observed phenotypic values (Piepho, 2009; Crossa et al., 2010; Schmidt et al., 2016). Erroneous phenotype data and GxE interactions will therefore contribute error to the estimated GEBV accuracy. This can reduce how informative the estimated accuracy is because a low estimated accuracy could be due to low true accuracy, or simply GxE between the training and validation environments. Statistical methods of improving the accuracy of phenotype data (i.e. spatial and multivariate analyses) are therefore valuable in increasing the relative contribution that true GEBV accuracy makes to the estimated GEBV accuracy. This section discusses both components of GEBV accuracy.

1.5.1 Prediction accuracy; contributing factors and a review of methodology

For GS to be effective in a plant breeding programme, a GEBV accuracy threshold must be met (Heffner et al., 2010). This accuracy threshold will be specific to individual breeding programmes and will vary with trait, breeding strategy, and the economics of phenotyping and genotyping. Many studies on GS in plant breeding have estimated accuracies using a range of methodologies (Table 1.5.1). Variations between methodologies arise from differences in statistical models; marker platforms; germplasm type and size; structure of training and validation sets; and management of genotype by environment interactions. While these factors make it difficult to directly compare accuracies across studies, they can be used to determine how relevant a study is to a particular breeding programme.

Size and type of germplasm collections

In order to maximise the gene pool available to breeders, breeding germplasm should be large and genetically diverse. This enables more diverse genetic combinations to be created, thus increasing the likelihood of achieving significant genetic gain (Bernardo, 2002).

This is a challenge for GS as the prediction model must adequately account for all marker effects and genetic backgrounds present in the germplasm. Researchers studying GS do not always have access to breeding germplasm, and different GS strategies utilise different population types. A wide range of germplasm types have therefore been used in studies reported thus far. Bi-parental populations were used by Heffner et al. (2011a) and Lorenzana & Bernardo (2009) in wheat and barley respectively, with populations ranging from 140 to 209 individuals in size. Collections of breeding lines ranging from 254 to 8416 individuals in size have been used in numerous studies with lines being sourced from programmes such as the International Maize and Wheat Improvement Centre (CIMMYT) (Crossa et al., 2010; Poland et al., 2012; Dawson et al., 2013; Lado et al., 2013; Crossa et al., 2016), the Cornell University breeding programme (Heslot et al., 2013), and the private European breeding company KWS (He et al., 2016, 2017). The largest training set used in the literature is that of Crossa et al. (2016) at 3052 lines, which remains significantly smaller than what would be available for large commercial programmes, but is significantly larger than the majority of populations used in GS studies. Also, the effect of training set size on prediction accuracy has not been explicitly investigated in training sets larger than 300 individuals, and so the optimum training set size remains unclear.

Structure of training and validation sets

The concept of GS in a plant breeding programme consists of developing a model in a training set, and using it to predict the performance of lines in the breeding programme. Consequently, the relatedness between the training set and the lines to be predicted has a significant impact on GEBV accuracy (Nakaya & Isobe, 2012). Structures of training and validation sets used in studies often differ to those in breeding programmes, but the fundamental relationship between relatedness and GEBV accuracy remains. The structure of validation and training sets used in studies therefore influences the GEBV accuracy achieved. The most common method of partitioning germplasm into training and validation sets is cross-validation. Here, germplasm is divided into a number of sets (10 for example), nine of which are combined for model development (training set), and GEBVs are then calculated for lines in the remaining set (validation set). Accuracy is estimated by correlating the GEBVs with phenotype data in the validation set. This process is carried out 10 times with a different set used for validation each time. Estimated accuracies of each repetition can then be averaged (Crossa et al., 2010). Cross-validation methodology can vary in the number of sets used, which determines the training:-validation ratio (Lorenzana & Bernardo, 2009; Lado et al., 2013). It has been suggested that a higher training: validation ratio is needed when the germplasm has higher genetic diversity, or when the trait of interest has lower heritability (Nakaya & Isobe, 2012).

Methodology also varies in the way the germplasm is divided into sets; most studies randomly assign sets (Lorenzana & Bernardo, 2009; Crossa et al., 2010; Heslot et al., 2012, 2013; Lado et al., 2013; He et al., 2016), but some have managed relatedness between training and validation sets by grouping closely related individuals within sets (Poland et al., 2012). Having lower relatedness between the training and validation sets is a closer scenario to that of a breeding programme where historical lines might be used to predict the performance of future lines.

Few studies have specifically investigated the response of GEBV accuracy to varying levels of relatedness between the training and validation sets (Scutari et al., 2016). When applied in a breeding programme, relatedness will likely fluctuate continuously. A comprehensive understanding of this response would allow breeders to comprehend the degree to which they can predict germplasm outside the training set.

Management of genotype by environment interactions

Genotype by environment (GxE) interactions (i.e. variability in the relative performance of individuals across environments) is a major challenge in plant breeding. This is particularly true in countries such as Australia, where key growth factors such as rainfall and soil type are subject to large temporal and spatial variation (Haldane, 1946; Cooper & DeLacy, 1994). Plant breeders have employed a range of methods to negate this challenge (Basford et al., 1991); chief among them is the grouping of environments into clusters that behave similarly (Horner & Frey, 1957). With GS, approaches to managing GxE interactions are much the same. Crossa et al. (2010) and Dawson et al. (2013) both used a grouping approach, such that a prediction model is generated for each environment cluster. Breeders could then assess predictions for each cluster, and base their selections on the relative importance of cluster. Another approach used in studies reviewed here is a global prediction, where data from multiple environments is incorporated into a single prediction model (Dawson et al., 2013; Heslot et al., 2013). This approach works best when environments are highly correlated, but is unlikely to be effective in managing large GxE interactions. Several studies performed GS across environments, taking training data from one set of environments, and validation data from another (Heffner et al., 2011a,b; Dawson et al., 2013; Lado et al., 2013; He et al., 2016). The point should be raised that this approach increases the degree to which environmental factors affect the estimation of GEBV accuracy (i.e. a low estimated accuracy may simply be the result of a large GxE interaction, and thus a poor correlation between environments). This can confound the interpretation of GEBV accuracies with other factors being investigated, such as relatedness between training and validation sets. It is therefore important to consider all contributing factors when assessing estimated accuracies. Heslot et al. (2012) analysed the same wheat data set used by Heffner et al. (2011b). Here, Heslot performed cross-validation within environment, where Heffner cross-validated across environment. Heffner's method also used a larger portion of the germplasm in model development. The contrast in approaches showed through in the accuracies. For yield, Heffner's ridge regression accuracy was 0.19 compared to Heslot's 0.36, highlighting the impact of GxE and cross-validation structure on accuracy estimation.

Statistical methods

Numerous statistical methods for calculating GEBVs have been studied to determine which are most accurate, and whether that accuracy is related to the trait being predicted. Crossa et al. (2010) compared a Bayesian method with RKHS regression and observed slightly higher accuracies with RKHS regression when predicting for yield. Also trialled was a BLUP method, which was consistently less accurate. The inclusion of pedigree data was also tested, and slightly improved accuracies with both models. It was suggested that as marker density increases, improvement from including the pedigree would decrease (Crossa et al., 2010).

Heffner et al. (2011a) used RR-BLUP and Bayes-Cπ on a range of quality traits in two bi-parental populations. Bayes-Cπ produced higher accuracies in the population derived from a wide cross, and RR-BLUP was superior in the population from a narrow cross. It was put forward that Bayes-Cπ is better suited to germplasm with larger QTL effects (such as those derived from wider crosses) as it allows some markers to contribute no effect, thus enabling the remaining larger marker effects to contribute more to the GEBV.

Lorenzana & Bernardo (2009) compared a BLUP approach with several Bayesian variations. BLUP consistently produced higher accuracy than the Bayesian approaches, which contrasts with the finding of Crossa et al. (2010). These studies used different genetic material and marker platforms, and methodology varied slightly, making it difficult to compare their findings.

Heslot et al. (2012) compared 10 statistical methods within a range of barley and wheat germplasm collections, predicting for yield and several other traits. In his study, the RKHS method consistently outperformed ridge regression, Bayesian LASSO and Bayes-Cπ, which all performed similarly. The E-Bayes method produced lower accuracies than each of the other methods.

Traits of interest

The accuracy of genomic prediction varies according to the genetic control of the trait of interest. Higher accuracies are commonly achieved when predicting qualitative traits with simple genetic control, as opposed to quantitative traits (Huang et al., 2006; Heffner et al., 2011b). This is true when using either genomic prediction, or a more traditional QTL approach such as multiple linear regression (MLR). This can be attributed to the fact that qualitative traits are predominantly controlled by few genes with large effect, which are easier to account for than the many genes of small effect that control quantitative traits. While qualitative traits are more easily predicted using MLR with several QTL (Butler et al., 2005; Sadeque & Turner, 2010; Heslot et al., 2012), it is not fully known if GS is more accurate than MLR in these instances. Almost all studies investigating GS in plant breeding have focused on grain yield, which is quantitative, but some studies have also considered a range of other traits, both quantitative and qualitative (Table 1.5.1).

TABLE 1.1: A summary of genomic selection studies in wheat and barley.

Species	Germplasm type (size)	Marker platform (# of markers)^1	Statistical model(s)^2	Traits	# of environments	GEBV accuracy	Reference
Wheat	Collection of CIMMYT^3 lines (599)^4	DArT (1279)	Pedigree-RKHS, Pedigree-BL, RKHS, BL, BLUP	Grain yield	4 (clusters)	RKHS: 0.45 – 0.60 P-RKHS: 0.48 – 0.61	Crossa et al. (2010)
Wheat	Collection of CIMMYT landrace accessions (8416)^4	DArT (23574-23946	GBLUP	9 agronomic and quality traits	2 (drought & heat)	0.34 – 0.60	Crossa et al. (2016)
Wheat	Collection of CIMMYT lines (384)	GBS SNP (13357)	RR	Grain yield, TKW^5, maturity, KPS^6	5	GY: 0.23 – 0.62 TKW: 0.76 – 0.84 Maturity: 0.40 – 0.58 KPS: 0.46 – 0.67	Lado et al. (2013)
Wheat	Collection of CIMMYT lines (622)	GBS SNP (34483)	BLUP	Grain yield	168 (historic data)	All years TP: 0.44 3-year window TP: 0.43 Random 16-fold CV: 0.56	Dawson et al. (2013)
Wheat	Bi-parental DH (209)	SSR, DArT, AFLP, TRAP, RFLP (484)	RR, Bayesian, MLR	9 grain quality traits	19	RR: 0.27 – 0.68 Bayesian: 0.31 – 0.67 MLR: 0.17 – 0.51	Heffner et al. (2011a)
Wheat	Bi-parental DH (174)	DArT (574)	RR, Bayesian, MLR	9 grain quality traits	19	RR: 0.37 – 0.63 Bayesian: 0.43 – 0.74 MLR: 0.27 – 0.48	Heffner et al. (2011a)
Wheat	Collection of Cornell University lines (365)	DArT (1544)	RR	Grain yield, plant height, maturity, sprouting	6	GY: 0.36 Height: 0.48 Maturity: 0.30 Sprouting: 0.47	Heslot et al. (2013)
Wheat	Collection of Cornell University lines (365)	GBS SNP (38412)	RR	Grain yield, plant height, maturity, sprouting	6	GY: 0.41 Height: 0.52 Maturity: 0.47 Sprouting: 0.57	Heslot et al. (2013)
Wheat	KWS elite winter lines (2325)	SNP array (12642)	RR, Bayes-Cπ, RKHS, Extended GBLUP	Grain yield	9	RR: 0.63 Extended GBLUP: 0.68 RKHS: 0.68 Bayes-Cπ: 0.62	He et al. (2016)

TABLE 1.1: A summary of genomic selection studies in wheat and barley.

Species	Germplasm type (size)	Marker platform (# of markers)[1]	Statistical model(s)[2]	Traits	# of environments	GEBV accuracy	Reference
Wheat	KWS elite winter lines (3816)	SNP array (3047 − 9153)	GBLUP, Extended GBLUP, RR-haplotype	Grain yield	10	GBLUP: 0.63 Extended GBLUP: 0.65 RR-haplotype: 0.64	He et al. (2017)
Wheat	DH breeding lines (840)	GBS (4598)	RR, W-BLUP, MW-BLUP	7 end use quality traits	10	RR: 0.30 − 0.53 W-BLUP: 0.35 − 0.61 MW-BLUP: 0.38 − 0.63	Michel et al. (2018)
Barley	KWS elite spring lines (training sets: 65 − 424)	SNP array (4095)	RR	12 quality and agronomic traits	2 − 7	0.14 − 0.58	Schmidt et al. (2016)
Barley	KWS elite winter lines (3816)	SNP array (4359)	RR	12 quality and agronomic traits	6	0.11 − 0.30	Schmidt et al. (2016)
Barley	Bi-parental population DH (150)	RFLP (223)	BLUP, e-Bayes	Grain yield, plant height, protein, malt extract, α-amylase activity	16	BLUP: 0.64 − 0.86 e-Bayes: 0.51 − 0.79	Lorenzana & Bernardo (2009)
Barley	Bi-parental population DH (140)	RFLP, AFLP (107)	BLUP, e-Bayes	Grain yield, plant height, protein, malt extract, α-amylase activity	9	BLUP: 0.61 − 0.85 e-Bayes: 0.58 − 0.86	Lorenzana & Bernardo (2009)
Wheat	Collection of CIMMYT lines (254)	DArT (1276)	BLUP	Grain yield, TKW, maturity	2 (drought irrigated)	GY (irrigated): 0.13 GY (drought): 0.18 TKW: 0.28 Maturity: 0.20	Poland et al. (2012)
Wheat	Collection of CIMMYT lines (254)	GBS SNP (1827)	BLUP	Grain yield, TKW, maturity	2 (drought & irrigated)	GY (irrigated): 0.25 GY (drought): 0.35 TKW: 0.26 Maturity: 0.34	Poland et al. (2012)
Wheat	Collection of CIMMYT lines (254)	GBS SNP (34749)	BLUP	Grain yield, TKW, maturity	2 (drought & irrigated)	GY (irrigated): 0.32 GY (drought): 0.42 TKW: 0.33 Maturity: 0.33	Poland et al. (2012)

TABLE 1.1: A summary of genomic selection studies in wheat and barley.

Species	Germplasm type (size)	Marker platform (# of markers)[1]	Statistical model(s)[2]	Traits	# of environments	GEBV accuracy	Reference
Wheat	Collection of CIMMYT lines (599)	DArT (1279	RR, BL, Elastic net, wBSR, Bayes-Cπ, e-Bayes, RKHS	Grain yield	4 (clusters)	RR: 0.38 − 0.51 BL: 0.37 − 0.50 Elastic net: 0.35 − 0.46 wBSR: 0.36 − 0.50 Bayes-Cπ: 0.58 − 0.51 e-Bayes: 0.36 − 0.49 RKHS: 0.43 − 0.59	Heslot et al. (2012)
Wheat	Collection of Cornell University lines (374)[7]	DArT (1158)	RR, BL, Elastic net, wBSR, Bayes-Cπ, e-Bayes, RKHS	Grain yield, maturity	2	RR: 0.36 − 0.45 BL: 0.35 − 0.44 Elastic net: 0.37 − 0.41 wBSR: 0.37 − 0.44 Bayes-Cπ: 0.34 − 0.44 e-Bayes: 0.26 − 0.41 RKHS: 0.28 − 0.55	Heslot et al. (2012)
Wheat	Partial diallel (eight crosses, five parents) (551)	SNP (319)	RR, BL, Elastic net, wBSR, Bayes-Cπ, e-Bayes, RKHS	Grain yield, plant height, TKW	6	RR: 0.53 − 0.64 BL: 0.52 − 0.66 Elastic net: 0.51 − 0.68 wBSR: 0.52 − 0.67 Bayes-Cπ: 0.53 − 0.66 e-Bayes: 0.51 − 0.67 RKHS: 0.58 − 0.73	Heslot et al. (2012)
Barley	Collection of Limagrain Europe elite lines (761)	SNP (338)	RR, BL, Elastic net, wBSR, Bayes-Cπ, e-Bayes, RKHS	Grain yield	24 (sparse)	RR: 0.53 BL: 0.55 Elastic net: 0.52 wBSR: 0.53 Bayes-Cπ: 0.53 e-Bayes: 0.53 RKHS: 0.60	Heslot et al. (2012)

TABLE 1.1: A summary of genomic selection studies in wheat and barley.

Species	Germplasm type (size)	Marker platform (# of markers)[1]	Statistical model(s)[2]	Traits	# of environments	GEBV accuracy	Reference
Barley	Collection of CAP[8] lines (911)	SNP (2146)	RR, BL, Elastic net, wBSR, Bayes-Cπ, e-Bayes, RKHS	β-glucan content	15 (sparse)	RR: 0.57 BL: 0.57 Elastic net: 0.57 wBSR: 0.57 Bayes-Cπ: 0.57 e-Bayes: 0.57 RKHS: 0.60	Heslot et al. (2012)
Wheat	Collection of Cornell University lines	DArT (1158)	RR, Bayes-A, Bayes-B, Bayes-Cπ	Grain yield, 12 agronomic and quality traits	6	RR: $0.20 - 0.75$ Bayes-A: $0.22 - 0.76$ Bayes-B: $0.22 - 0.75$ Bayes-Cπ: $0.17 - 0.76$	Heffner et al. (2011b)

[1] DArT (diversity array technology), SNP (single nucleotide polymorphism), GBS (genotyped by sequencing)

[2] RR (ridge regression), W-BLUP (ridge regression with large effect markers included as fixed effects), MW-BLUP (multi-trait W-BLUP model), RKHS (reproducing kernel Hilbert space regression), BL (Bayesian least absolute shrinkage and selection operator), BLUP (best linear unbiased prediction), MLR (multiple linear regression), e-Bayes (empirical Bayesian), Bayes (Bayesian), wBSR (weighted Bayesian shrinkage regression)

[3] International Maize and Wheat Improvement Centre

[4] The CIMMYT dataset used by Heslot et al. (2012) is the same as that used by Crossa et al. (2010)

[5] Thousand kernel weight

[6] Kernels per spike

[7] The Cornell University dataset used by Heslot et al. (2012) is the same as that used by Heffner et al. (2011b)

[8] Coordinated Agricultural Project

1.5.2 Relevance of reported accuracies

Studies reviewed here report estimated GEBV accuracies ranging from 0.13 to 0.86. As detailed above, there are many factors contributing to both estimated and true GEBV accuracy in GS studies. These factors are highly variable across studies, which makes comparisons difficult and affects how relevant a study is to breeding applications. When designing studies, consideration should be given to maximise the relevance to breeding programmes.

1.6 Genomic selection in a plant breeding programme

The potential for GS to be applied in plant breeding programmes has been a point of speculation for almost two decades (Meuwissen et al., 2001). In the past ten years, research effort into GS for plant breeding has increased dramatically (Lorenzana & Bernardo, 2009; Heffner et al., 2011b; Heslot et al., 2012; Dawson et al., 2013; Crossa et al., 2014; Hickey et al., 2014; He et al., 2016; Gaynor et al., 2017; Gorjanc et al., 2018). This section explores some important issues surrounding the application of GS.

1.6.1 Wheat breeding strategies utilising genomic selection

Several studies have proposed plant breeding strategies around GS that have the potential to significantly increase the rate of genetic gain (Heffner et al., 2009; Nakaya & Isobe, 2012). Figure 1.2 details the structure of a GS wheat breeding programme adapted from Heffner et al. (2009) and Morrell et al. (2011). The key difference from a conventional programme is that progeny of crosses are selected as parents for the next cross using GS, as opposed to phenotypic selection. This greatly reduces what is known as the generation interval, one of the key determinants of rate of genetic gain. As wheat is an inbred crop, lines must be selfed in order to produce homozygous individuals (or otherwise use technology such as doubled haploid). In conventional programmes, this inbreeding stage takes place before intense phenotypic assessment, although there is some overlap of the two. This can also be the case in GS programmes, with some level of inbreeding taking place before the GS stage (Heffner et al., 2010). Figure 1.2 illustrates a strategy where crossing designs (e.g. straight cross, top cross, or backcross) are carried out to produce an F1 generation which immediately undergoes genotyping and GS (Gaynor et al., 2017).

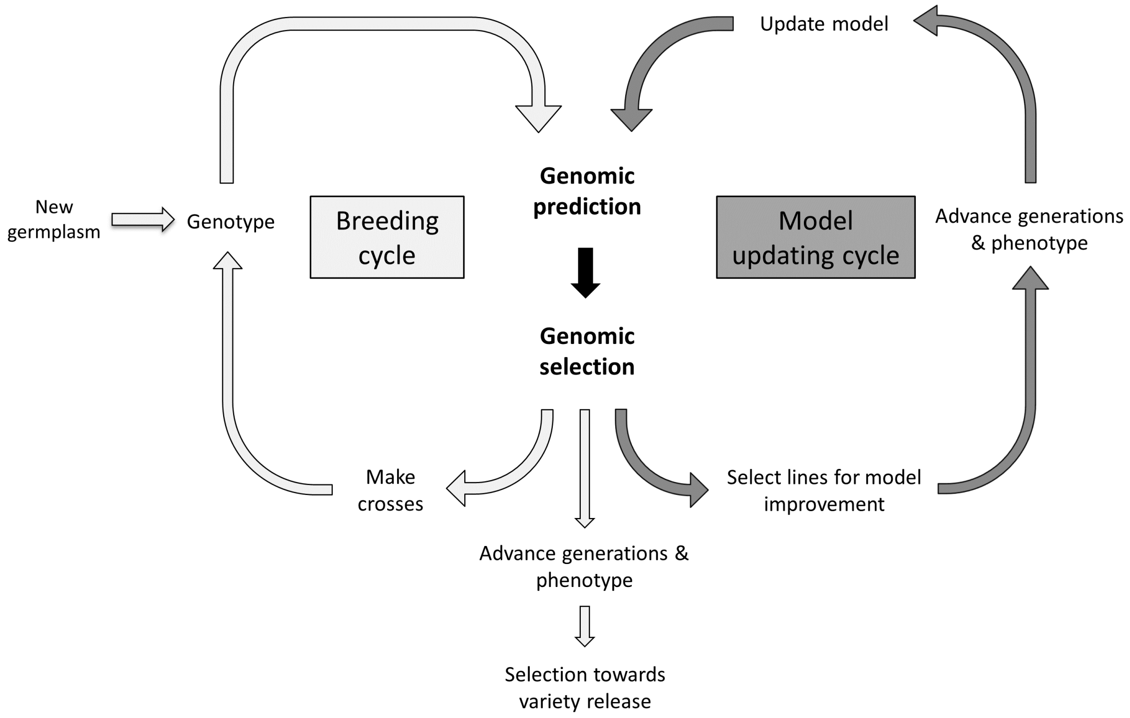
Update model
New germplasm
Genotype
Breeding cycle
Genomic prediction
Model updating cycle
Advance generations & phenotype
Genomic selection
Make crosses
Select lines for model improvement
Advance generations & phenotype
Selection towards variety release

FIGURE 1.2: Adapted from Heffner et al. (2009) and Morrell et al. (2011). Crossing cycle and generation interval is shortened by using genomic selection in place of phenotyping for selection of parents. The model is updated with data from new individuals to maintain prediction accuracy. New germplasm is introduced to increase genetic diversity.

As well as selecting parents for crossing, Figure 1.2 shows individuals would also be selected for progression in the breeding programme towards variety release. It should be noted that when selecting for parents breeders are interested in the individual's breeding value (only additive effects), but when selecting for progression in the programme breeders are interested in that individual's phenotypic performance. This is referred to as the genotypic value, and includes both additive and non-additive effects (Bernardo, 2002). Models described in the review of statistical methods section include only additive effects, and hence predict the breeding value. As wheat is an inbred crop, the genetic performance is comprised of additive and additive-by-additive interactions (epistasis) only. There is uncertainty around the importance of epistasis in the control of quantitative traits

(Hill et al., 2008; Cooper et al., 2009). The amount and size of epistatic effects will define how accurate an additive model is at predicting genotypic value. If epistasis scarcely contributes to the control of quantitative traits, as suggested by Oakey et al. (2006), then an additive model would likely suffice for predicting genotypic value. If the contribution from epistasis is large however, models that account for epistatic effects would be needed; Wang et al. (2012) presented such a model.

1.6.2 Maintaining prediction accuracy

Genomic prediction models use markers to represent the effect of QTL with which they are in linkage disequilibrium (LD). A QTL effect will only be accurately represented by its corresponding marker while the two remain in LD (Cooper et al., 2014). LD between markers and QTL can decay over multiple generations as recombination events create different linkage phases in some individuals which can be inherited in the next generation. Decay of LD therefore poses a threat to genomic prediction accuracy after several generations of GS (Muir, 2007). In order to negate this issue, the model should be periodically re-trained by including genotype and phenotype data from the new individuals in the training set (model updating cycle in Figure 1.2). It should also be noted that LD decay occurs most rapidly when LD between two loci is weak. LD is more intense between loci that have tighter physical linkage (Nakaya & Isobe, 2012); having tighter physical linkage between markers and QTL will therefore slow the rate of LD decay. Higher marker densities result in more markers having tighter physical linkage with QTL, and could therefore act as a buffer against loss of prediction accuracy over generations (Hickey et al., 2014).

1.6.3 Performing genomic selection: tools and approaches

Making selections is one of the most important processes in plant breeding. Conventional breeding programmes commonly perform between one and two rounds of selection for crossing per year, but with GS this number could theoretically exceed five (Morrell et al., 2011). Therefore, efficient selection methods will play an important role in increasing the efficiency and effectiveness of GS breeding programmes.

Selection indices

The concept of using selection indices to rank breeding lines has been reported since the 1930s (Smith, 1936), and is the process of combining performance data of multiple traits into one value based on the relative economic importance of traits (Lin, 1978). A common model for an index, known as the Smith-Hazel index (Smith, 1936; Hazel, 1943), also accounts for genetic and phenotypic correlations between traits. Heffner et al. (2011b) used the Smith-Hazel index in a GS study and showed indexing based on GEBV to be slightly more accurate than indexing based on phenotype, and suggested more research go into selection indices with GS. This selection index, as used by Heffner, is not an ideal model as it unreasonably assigns a multiplier weight to threshold traits that actually only need to be better than a defined cut-off (e.g. test weight). This results in threshold traits contributing to the index even after surpassing their threshold performance. However, this study along with successful applications of selection indices with GS in livestock breeding (Dekkers, 2007; Börner & Reinsch, 2012), does reveal potential for their application in GS plant breeding programmes. Future research in applying selection indices to plant breeding would be valuable, particularly in addressing better management of threshold traits and accounting for varying trait heritabilities. This topic is further discussed in section 5.4.1.

Maintaining genetic diversity

Issues have been raised around an increased rate of inbreeding and subsequent loss of genetic diversity when using GS (Daetwyler et al., 2007; Heffner et al., 2009; Jannink et al., 2010; Gorjanc et al., 2016; Gaynor et al., 2017). This can occur when superior individuals are constantly selected for crossing, causing alleles they carry to contribute increasingly to succeeding generations and alleles from inferior individuals to contribute less. This leads to an undesirable loss of genetic diversity within the germplasm. To combat this, Jannink et al. (2010) describes a method of placing weight on favourable alleles with low frequency, thus reducing the loss of favourable alleles and maintaining genetic diversity. However, this area requires continued research effort to ensure that high rates of genetic gain are maintained under GS over time.

1.7 Summary

Since their inception, the principal use of molecular markers in breeding programmes has been QTL analysis and MAS. Recent advances in dense marker genotyping and statistical methods have resulted in a shift of focus towards whole genome approaches such as GS. There are numerous factors that determine the effectiveness of GS in a wheat breeding programme; sufficient prediction accuracy is required, as are optimised breeding strategies. Many factors regulate true prediction accuracy, and methods for estimating prediction accuracy vary. The optimal methodology is therefore not clear among current literature; this needs to be determined in order for plant breeders to make informed decisions about the potential deployment of GS in their programmes. While several studies have recently increased the size of training sets up to 3052 individuals, the effect of training set size on prediction accuracy has not been explicitly investigated and so the optimum training set size remains unclear. Breeding programmes often have access to much larger populations, and these should be studied in regards to size and relatedness if we are to better understand the optimal design of GS training sets. If breeders can efficiently implement accurate GS with low cost genotyping strategies, then GS is likely to realise its potential and enable a significant increase in rate of genetic gain achieved by plant breeding programmes.

Bibliography

AKBARI, M., WENZL, P., CAIG, V., CARLING, J., XIA, L., YANG, S., USZYNSKI, G., MOHLER, V., LEHMENSIEK, A., KUCHEL, H., HAYDEN, M., HOWES, N., SHARP, P., VAUGHAN, P., RATHMELL, B., HUTTNER, E., & KILIAN, A. (2006). Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theoretical and applied genetics* **113**, 1409–1420.

AKHUNOV, E., NICOLET, C., & DVORAK, J. (2009). Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theoretical and applied genetics* **119**, 507–517.

ANTOLÍN, R., NETTELBLAD, C., GORJANC, G., MONEY, D., & HICKEY, J. (2017). A hybrid method for the imputation of genomic data in livestock populations. *Genetics Selection Evolution* **49**, 30.

BASFORD, K., KROONENBERG, P., & I., D. (1991). Three-way methods for multi-attribute genotype x environment data: an illustrated partial survey. *Field crops research* **27**, 131–157.

BENNETT, D., IZANLOO, A., REYNOLDS, M., KUCHEL, H., LANGRIDGE, P., & SCHNURBUSCH, T. (2012). Genetic dissection of grain yield and physical grain quality in bread wheat (*Triticum aestivum L.)* under water-limited environments. *Theoretical and Applied Genetics* **125**, 255–271.

BERNARDO, R. (2002). *Breeding for quantitative traits in plants*. Stemma Press Woodbury.

BÖRNER, V. & REINSCH, N. (2012). Optimising multistage dairy cattle breeding schemes including genomic selection using decorrelated or optimum selection indices. *Genetics, selection, evolution : GSE* **44**, 1.

BOUQUET, A. & JUGA, J. (2013). Integrating genomic selection into dairy cattle breeding programmes: a review. *Animal : an international journal of animal bioscience* **7**, 705–713.

BUTLER, J., BYRNE, P., MOHAMMADI, V., CHAPMAN, P., & HALEY, S. (2005). Agronomic performance of *Rht* alleles in a spring wheat population across a range of moisture levels. *Crop science* **45**, 939–947.

COLLARD, B., JAHUFER, M., BROUWER, J., & PANG, E. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* **142**, 169–196.

COOPER, M. & DELACY, I. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics* **88**, 561–572.

COOPER, M., MESSINA, C., PODLICH, D., TOTIR, L., BAUMGARTEN, A., HAUSMANN, N., WRIGHT, D., & GRAHAM, G. (2014). Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop and Pasture Science* **65**, 311–336.

COOPER, M., VAN EEUWIJK, F., HAMMER, G., PODLICH, D., & MESSINA, C. (2009). Modeling QTL for complex traits: detection and context for plant breeding. *Current opinion in plant biology* **12**, 231–240.

CROSSA, J., DE LOS CAMPOS, G., PÉREZ, P., GIANOLA, D., BURGUEÑO, J., ARAUS, J., MAKUMBI, D., SINGH, R., DREISIGACKER, S., YAN, J., ET AL. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**, 713–724.

CROSSA, J., JARQUÍN, D., FRANCO, J., PÉREZ-RODRÍGUEZ, P., BURGUEÑO, J., SAINT-PIERRE, C., VIKRAM, P., SANSALONI, C., PETROLI, C., AKDEMIR, D., ET AL. (2016). Genomic prediction of gene bank wheat landraces. *G3: Genes, Genomes, Genetics* **6**, 1819–1834.

CROSSA, J., PERÉZ, P., HICKEY, J., BURGUEÑO, J., ORNELLA, L., CERÓN-ROJAS, J., ZHANG, X., DREISIGACKER, S., BABU, R., LI, Y., BONNETT, D., & MATHEWS, K. (2014).

Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* **112**, 48–60.

DAETWYLER, H., VILLANUEVA, B., BIJMA, P., & WOOLLIAMS, J. (2007). Inbreeding in genome-wide selection. *Journal of Animal Breeding and Genetics* **124**, 369–376.

DAWSON, J., ENDELMAN, J., HESLOT, N., CROSSA, J., POLAND, J., DREISIGACKER, S., MANÈS, Y., SORRELLS, M., & JANNINK, J. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research* **154**, 12–22.

DE LOS CAMPOS, G., GIANOLA, D., & ROSA, G. (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of animal science* **87**, 1883–1887.

DEKKERS, J. (2007). Prediction of response to marker-assisted and genomic selection using selection index theory. *Journal of animal breeding and genetics* **124**, 331–341.

DEKKERS, J., HOSPITAL, F., ET AL. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* **3**, 22–32.

ELSHIRE, R., GLAUBITZ, J., SUN, Q., POLAND, J., KAWAMOTO, K., BUCKLER, E., & MITCHELL, S. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**, e19379.

FAO (2017). Food and Agriculture Organization of the United Nations: Food and agiculture data.

GAYNOR, R., GORJANC, G., BENTLEY, A., OBER, E., HOWELL, P., JACKSON, R., MACKAY, I., & HICKEY, J. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science* **56**, 1–15.

GONEN, S., WIMMER, V., GAYNOR, R., BYRNE, E., GORJANC, G., & HICKEY, J. (2018). A heuristic method for fast and accurate phasing and imputation of single-nucleotide polymorphism data in bi-parental plant populations. *Theoretical and Applied Genetics* pages 1–13.

GONZÁLEZ-CAMACHO, J., ORNELLA, L., PÉREZ-RODRÍGUEZ, P., GIANOLA, D., DREISIGACKER, S., & CROSSA, J. (2018). Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *The plant genome* **11**.

GORJANC, G., BATTAGIN, M., DUMASY, J., ANTOLIN, R., GAYNOR, R., & HICKEY, J. (2017). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Science* **57**, 216–228.

GORJANC, G., GAYNOR, R., & HICKEY, J. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics* **131**, 1953–1966.

GORJANC, G., JENKO, J., HEARNE, S., & HICKEY, J. (2016). Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC genomics* **17**, 30.

HALDANE, J. (1946). The interaction of nature and nurture. *Annals of eugenics* **13**, 197–205.

HALEY, C. & KNOTT, S. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

HAYASHI, T. & IWATA, H. (2010). EM algorithm for bayesian estimation of genomic breeding values. *BMC genetics* **11**, 3.

HAZEL, L. (1943). The genetic basis for constructing selection indexes. *Genetics* **28**, 476–490.

HE, S., REIF, J., KORZUN, V., BOTHE, R., EBMEYER, E., & JIANG, Y. (2017). Genome-wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to central europe. *Theoretical and Applied Genetics* **130**, 635–647.

HE, S., SCHULTHESS, A., MIRDITA, V., ZHAO, Y., KORZUN, V., BOTHE, R., EBMEYER, E., REIF, J., & JIANG, Y. (2016). Genomic selection in a commercial winter wheat population. *Theoretical and Applied Genetics* **129**, 641–651.

HEFFNER, E., JANNINK, J., IWATA, H., SOUZA, E., & SORRELLS, M. (2011a). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Science* **51**, 2597–2606.

HEFFNER, E., JANNINK, J., & SORRELLS, M. (2011b). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome* **4**, 65–75.

HEFFNER, E., LORENZ, A., JANNINK, J., & SORRELLS, M. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Science* **50**, 1681–1690.

HEFFNER, E., SORRELLS, M., & JANNINK, J. (2009). Genomic selection for crop improvement. *Crop Science* **49**, 1–12.

HESLOT, N., JANNINK, J., & SORRELLS, M. (2015). Perspectives for genomic selection applications and research in plants. *Crop Science* **55**, 1–12.

HESLOT, N., RUTKOSKI, J., POLAND, J., JANNINK, J., & SORRELLS, M. (2013). Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PLoS One* **8**, e74612.

HESLOT, N., YANG, H., SORRELLS, M., & JANNINK, J. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science* **52**, 146–160.

HICKEY, J., DREISIGACKER, S., CROSSA, J., HEARNE, S., BABU, R., PRASANNA, B., GRONDONA, M., ZAMBELLI, A., WINDHAUSEN, V., MATHEWS, K., ET AL. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science* **54**, 1476–1488.

HICKEY, J., GORJANC, G., VARSHNEY, R., & NETTELBLAD, C. (2015). Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a hidden Markov model. *Crop Science* **55**, 1934–1946.

HILL, W., GODDARD, M., & VISSCHER, P. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics* **4**, e1000008.

HORNER, T. & FREY, K. (1957). Methods for determining natural areas for oat varietal recommendations. *Agronomy Journal* **49**, 313–315.

HOWIE, B., DONNELLY, P., & MARCHINI, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529.

HUANG, X., CLOUTIER, S., LYCAR, L., RADOVANOVIC, N., HUMPHREYS, D., NOLL, J., SOMERS, D., & BROWN, P. (2006). Molecular detection of QTLs for agronomic and quality traits in a doubled haploid population derived from two Canadian wheats (*Triticum aestivum L.). Theoretical and Applied Genetics* **113**, 753–766.

JANNINK, J., LORENZ, A., & IWATA, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics* **9**, 166–177.

JANSEN, R. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

JULIANA, P., SINGH, R., SINGH, P., CROSSA, J., HUERTA-ESPINO, J., LAN, C., BHAVANI, S., RUTKOSKI, J., POLAND, J., BERGSTROM, G., & SORRELLS, M. (2017). Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theoretical and applied genetics* **130**, 1415–1430.

KAO, C., ZENG, Z., & TEASDALE, R. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

KONG, A., MASSON, G., FRIGGE, M., GYLFASON, A., ZUSMANOVICH, P., THORLEIFSSON, G., OLASON, P., INGASON, A., STEINBERG, S., RAFNAR, T., ET AL. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics* **40**, 1068.

KUCHEL, H., WILLIAMS, K., LANGRIDGE, P., EAGLES, H., & JEFFERIES, S. (2007). Genetic dissection of grain yield in bread wheat. I. QTL analysis. *Theoretical and Applied Genetics* **115**, 1029–1041.

LADO, B., MATUS, I., RODRÍGUEZ, A., INOSTROZA, L., POLAND, J., BELZILE, F., DEL POZO, A., QUINCKE, M., CASTRO, M., & VON ZITZEWITZ, J. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3: Genes— Genomes— Genetics* **3**, 2105–2114.

LANDER, E. & BOTSTEIN, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

LIN, C. (1978). Index selection for genetic improvement of quantitative characters. *Theoretical and applied genetics* **52**, 49–56.

LORENZ, A., HAMBLIN, M., & JANNINK, J. (2010). Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PloS one* **5**, e14079.

LORENZANA, R. & BERNARDO, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics* **120**, 151–161.

MAPHOSA, L., LANGRIDGE, P., TAYLOR, H., PARENT, B., EMEBIRI, L., KUCHEL, H., REYNOLDS, M., CHALMERS, K., OKADA, A., EDWARDS, J., ET AL. (2014). Genetic control of grain yield and grain physical characteristics in a bread wheat population grown under a range of environmental conditions. *Theoretical and Applied Genetics* **127**, 1607.

MEUWISSEN, T., HAYES, B., & GODDARD, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

MICHEL, S., KUMMER, C., GALLEE, M., HELLINGER, J., AMETZ, C., AKGÖL, B., EPURE, D., LÖSCHENBERGER, F., & BUERSTMAYR, H. (2018). Improving the baking quality of bread wheat by genomic selection in early generations. *Theoretical and Applied Genetics* **131**, 477–493.

MISZTAL, I. & LEGARRA, A. (2017). Invited review: efficient computation strategies in genomic selection. *animal* **11**, 731–736.

MORRELL, P., BUCKLER, E., & ROSS-IBARRA, J. (2011). Crop genomics: advances and applications. *Nature reviews. Genetics* **13**, 85–96.

MUIR, W. (2007). Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics* **124**, 342–355.

NAKAYA, A. & ISOBE, S. (2012). Will genomic selection be a practical method for plant breeding? *Annals of botany* **110**, 1303–1316.

OAKEY, H., VERBYLA, A., PITCHFORD, W., CULLIS, B., & KUCHEL, H. (2006). Joint modeling of additive and non-additive genetic line effects in single field trials. *Theoretical and applied genetics* **113**, 809–819.

PIEPHO, H. (2009). Ridge regression and extensions for genome-wide selection in maize. *Crop Science* **49**, 1165–1176.

PIEPHO, H., MÖHRING, J., SCHULZ-STREECK, T., & OGUTU, J. (2012a). A stage-wise approach for the analysis of multi-environment trials. *Biometrical journal* **54**, 844–860.

PIEPHO, H., OGUTU, J., SCHULZ-STREECK, T., ESTAGHVIROU, B., GORDILLO, A., & TECHNOW, F. (2012b). Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. *Crop Science* **52**, 1093–1104.

POLAND, J., ENDELMAN, J., DAWSON, J., RUTKOSKI, J., WU, S., MANES, Y., DREISIGACKER, S., CROSSA, J., SÁNCHEZ-VILLEDA, H., SORRELLS, M., ET AL. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome* **5**, 103–113.

POLAND, J. & RIFE, T. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* **5**, 92–102.

R CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RASMUSSON, J. (1933). A contribution to the theory of quantitative character inheritance. *Hereditas* **18**, 245–261.

SADEQUE, A. & TURNER, M. (2010). QTL analysis of plant height in hexaploid wheat doubled haploid population. *Thai Journal of Agricultural Science* **43**, 91–96.

SAX, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* **8**, 552–560.

SCHMIDT, M., KOLLERS, S., MAASBERG-PRELLE, A., GROSSER, J., SCHINKEL, B., TOMERIUS, A., GRANER, A., & KORZUN, V. (2016). Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. *Theoretical and Applied Genetics* **129**, 203–213.

SCHULZ-STREECK, T., OGUTU, J., & PIEPHO, H. (2013). Comparisons of single-stage and two-stage approaches to genomic selection. *Theoretical and applied genetics* **126**, 69–82.

SCUTARI, M., MACKAY, I., & BALDING, D. (2016). Using genetic distance to infer the accuracy of genomic prediction. *PLOS Genetics* **12**, 1–19.

SILLANPÄÄ, M. & ARJAS, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Theoretical and Applied Genetics* **148**, 1373–1388.

SMITH, H. (1936). A discriminant function for plant selection. *Annals of Eugenics* **7**, 240–250.

SOLLER, M., BRODY, T., & GENIZI, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47**, 35–39.

TAYLOR, J., VERBYLA, A., ET AL. (2011). R package wgaim: QTL analysis in bi-parental populations using linear mixed models. *Journal of Statistical Software* **40**, 1–18.

VERBYLA, A., CULLIS, B., & THOMPSON, R. (2007). The analysis of QTL by simultaneous use of the of the full linkage map. *Theoretical and Applied Genetics* **116**, 95–111.

VERBYLA, A., TAYLOR, J., & VERBYLA, K. (2012). RWGAIM: an efficient high-dimensional random whole genome average (QTL) interval mapping approach. *Genetics Research* **94**, 291–306.

WANG, D., EL-BASYONI, S., BAENZIGER, P., CROSSA, J., ESKRIDGE, K., & DWEIKAT, I. (2012). Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity* **109**, 313–319.

WHITTAKER, J., THOMPSON, R., & DENHAM, M. (2000). Marker-assisted selection using ridge regression. *Genetical research* **75**, 249–252.

WRIGLEY, C. & RATHJEN, A. (1981). Wheat breeding in australia. In Carr, S. & Carr, S., editors, *Plants and Man in Australia*, pages 96–135. Academic Press, New York.

XU, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789–801.

XU, S. (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**, 513–521.

XU, S. & ATCHLEY, W. (1995). A random model approach to interval mapping of quantitative trait loci. *Genetics* **141**, 1189–1197.

ZENG, Z. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 10972–10976.

ZENG, Z. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

ZHONG, S., DEKKERS, J., FERNANDO, R., & JANNINK, J. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* **182**, 355–364.

Chapter 2

An initial assessment of genomic prediction accuracy

2.1 Exegetical statement

For GS to be applied in a breeding programme, large populations require genotyping both for training the model and for selection. Even with low-cost genotyping strategies, this represents a significant financial investment that needs to be justified by the potential benefits. An understanding is therefore required of the accuracy that can be achieved when predicting relevant traits in relevant large-scale breeding germplasm. In addition, qualitative traits with simple genetic control may be better predicted with only several quantitative trait loci using marker assisted selection. Predictive ability of the two methods should therefore be assessed in traits of varying genetic complexity. This paper utilises a dataset of unprecedented size to address these questions by comparing cross-validation accuracies of genomic and marker assisted prediction in 14 traits of varying genetic control. We investigate the level of linkage disequilibrium in the dataset which is important in understanding how the prediction calibrations are capturing the genetic effects, and also characterise the genetic and residual correlations between traits which reveal to breeders where correlated response to selection may be occurring.

Statement of Authorship

Title of Paper	Increased genomic prediction accuracy in wheat breeding using a large Australian panel
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Theoretical and Applied Genetics (2017) 130:2543-2555 DOI 10.1007/s00122-017-2975-4

Principal Author

Name of Principal Author (Candidate)	Adam Norman		
Contribution to the Paper	Collected phenotype data. Involved in formulating the research objective and experimental design. Performed all analysis of phenotypic and genetic data. Wrote and prepared the manuscript. Acted as corresponding author.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24/8/2018

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Julian Taylor		
Contribution to the Paper	Construction of the genetic linkage and consensus maps. Contributed to statistical methodology. Reviewing and approving the manuscript. PhD co-supervisor of Adam Norman.		
Signature		Date	13/09/18

Name of Co-Author	Emi Tanaka		
Contribution to the Paper	Construction of the genetic linkage and consensus maps. Reviewing and approving the manuscript.		
Signature		Date	06/09/2018

Name of Co-Author	Paul Telfer		
Contribution to the Paper	Generation of several bi-parental populations used for genetic mapping. Reviewing and approving the manuscript.		
Signature		Date	6/9/18

Name of Co-Author	James Edwards		
Contribution to the Paper	Involved in formulating the research objective and experimental design. Reviewing and approving the manuscript. PhD co-supervisor of Adam Norman.		
Signature		Date	6/9/2018

Name of Co-Author	Jean-Pierre Martinant		
Contribution to the Paper	Development of the marker genotyping platform.		
Signature		Date	03/09/2018

Name of Co-Author	Haydn Kuchel		
Contribution to the Paper	Involved in formulating the research objective and experimental design. Reviewing and approving the manuscript. PhD principal supervisor of Adam Norman.		
Signature		Date	10/9/18

Theor Appl Genet
DOI 10.1007/s00122-017-2975-4

ORIGINAL ARTICLE

Increased genomic prediction accuracy in wheat breeding using a large Australian panel

Adam Norman[1,3] · Julian Taylor[1] · Emi Tanaka[2] · Paul Telfer[1,3] · James Edwards[1,3] · Jean-Pierre Martinant[4] · Haydn Kuchel[1,3]

Received: 12 May 2017 / Accepted: 19 August 2017
© The Author(s) 2017. This article is an open access publication

Abstract

Key message **Genomic prediction accuracy within a large panel was found to be substantially higher than that previously observed in smaller populations, and also higher than QTL-based prediction.**

Abstract In recent years, genomic selection for wheat breeding has been widely studied, but this has typically been restricted to population sizes under 1000 individuals. To assess its efficacy in germplasm representative of commercial breeding programmes, we used a panel of 10,375 Australian wheat breeding lines to investigate the accuracy of genomic prediction for grain yield, physical grain quality and other physiological traits. To achieve this, the complete panel was phenotyped in a dedicated field trial and genotyped using a custom Axiom$^{\mathrm{TM}}$ Affymetrix SNP array. A high-quality consensus map was also constructed, allowing the linkage disequilibrium present in the germplasm to be investigated. Using the complete SNP array, genomic prediction accuracies were found to be substantially higher than those previously observed in smaller populations and also more accurate compared to prediction approaches using a finite number of selected quantitative trait loci. Multi-trait genetic correlations were also assessed at an additive and residual genetic level, identifying a negative genetic correlation between grain yield and protein as well as a positive genetic correlation between grain size and test weight.

Introduction

Plant breeding has been successful in producing significant yield gains in wheat since the beginning of the twentieth century (Wrigley and Rathjen 1981); this has largely been driven by the innovation and adoption of new breeding technologies. Such progress is underpinned by extensive research, first in developing the technology, and second on establishing its application. If new technologies are to continue enabling plant breeding to deliver genetic gain to growers, innovative research must be undertaken in datasets that are relevant to the setting in which they will be applied.

Molecular markers are one technology that represent an invaluable research tool for understanding the genetic control of various traits. They have frequently been utilised in quantitative trait loci (QTL) mapping studies, and applied in breeding programmes through marker-assisted selection (MAS) (Koebner and Summers 2003; Collard and Mackill 2008). Early statistical modelling approaches to QTL mapping involved the analysis of individual markers through simple scanning procedures (Soller et al. 1976). In more

Communicated by Dr. Ian Mackay.

Marker data will be available for downloading as supplementary material 12 months after publication, or in advance from the authors subject to the terms of MTA.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-017-2975-4) contains supplementary material, which is available to authorized users.

✉ Adam Norman
adam.norman@agtbreeding.com.au

[1] School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Glen Osmond, SA, Australia

[2] National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, Australia

[3] Australian Grain Technologies Pty Ltd, Perkins Building, Roseworthy Campus, Roseworthy, SA, Australia

[4] Centre of Research, Limagrain Field Seeds Pty Ltd, Chappes, France

Published online: 08 September 2017

modern approaches, statistical methods have improved the efficiency and power of QTL detection through the simultaneous incorporation of markers from the whole genome in complex linear mixed models (Zhang et al. 2010; Verbyla et al. 2012). There has also been focus on whole genome QTL mapping in broader multiparent populations (Huang et al. 2012; Sannemann et al. 2015; Mackay et al. 2014), and diverse association panels (Neumann et al. 2011; Bentley et al. 2014; Zanke et al. 2014). The latter usually involves the use of genome-wide association studies (GWAS) and has become a valuable tool for broad validation of previously identified QTL as well as identification of QTL in the target breeding germplasm. For qualitative traits under simple genetic control, GWAS, and subsequent application of MAS has been shown to be an effective tool in breeding programmes (Xu and Crouch 2008). However, for more complex polygenic quantitative traits such as grain yield, there have been few examples of genetic improvement using MAS (Dekkers et al. 2002). This deficiency can be overcome by implementing a genomic selection (GS) method that uses a complete set of molecular marker effects for predicting the performance of quantitative polygenic traits (Meuwissen et al. 2001). Current research in this area suggests with sufficient prediction accuracy, GS can be successfully applied in a breeding programme to increase rates of genetic gain (Cooper et al. 2014; Schmidt et al. 2016). Recent studies investigating the accuracy of GS in wheat have used population sizes ranging from several hundred to several thousand individuals, and achieved prediction accuracies mostly in the range of 0.50–0.60 as measured by Pearson correlation coefficients (Heslot et al. 2012; Nakaya and Isobe 2012; Isidro et al. 2015; He et al. 2016).

In GWAS and QTL analysis, the use of physical and genetic maps has been widely adopted (Kang et al. 2010; Zhang et al. 2010). Recombination information from these maps could also be used in GS programmes to simulate the progeny of specific parents for the purpose of designing crosses (Podlich and Cooper 1998). Physical maps are becoming available for wheat (Pozniak 2016), but can be of limited value if the individuals sequenced are not closely related to the target germplasm. Additionally, physical maps do not incorporate recombination information, which reduces their value when we are interested in simulating progeny based on recombination probabilities in the germplasm of interest. Therefore, high-quality genetic maps built from relevant germplasm are a better resource for these applications. Examples of such maps in the literature include those produced using multi-parent advanced generation inter-cross (MAGIC) populations (Huang et al. 2012; Gardner et al. 2016), as well as consensus maps constructed from multiple bi-parental populations (Cavanagh et al. 2013; Wang et al. 2014). These maps can also be used to measure the extent of linkage disequilibrium (LD) between markers (Zhao et al. 2005; Chao et al. 2010). In the context of association mapping and genomic prediction, LD becomes vitally important as it influences the achievable mapping resolution (Huang et al. 2012), power and accuracy of QTL detection (Somers et al. 2007), and the accuracy of genomic prediction in a breeding programme after multiple generations (Muir 2007). The extent of LD is also known to vary significantly depending on the germplasm structure (Hao et al. 2011; Huang et al. 2012) and as a consequence, assessments of LD should be conducted on the genetic material being studied.

For GS to be applied effectively, plant breeders must have a sound understanding of the relationship between traits of interest as it enables optimisation of selection strategies through correlated response to selection (Bernardo 2002). Trait correlations in bread wheat have long been reported at the phenotypic level (Bhatt and Derera 1975; Fischer and Wood 1979). Advances in statistical techniques have since made it possible to draw genetic correlations between traits by separating the genetic variance from the residual error (Gilmour et al. 1997), and these have been reported for various physiological traits in bread wheat (Rebetzke and Richards 1999; Sukumaran et al. 2015). These approaches, coupled with the use of pedigree or molecular marker information, can also be used to separate the genetic variance into its additive and residual components, thus allowing genetic correlations to be drawn at the additive and residual genetic level (Rebetzke et al. 2013). These genetic correlations, particularly the additive, provide a more precise measure of trait relationships and facilitate better optimisation of selection strategies.

In the present study we use a panel of 10,375 lines from a commercial wheat breeding programme to: (1) assess the level of LD using a constructed high-quality genetic consensus map; (2) investigate genetic correlations between traits at an additive and residual genetic level; (3) investigate the improvement in selection accuracy that is achieved by incorporating a genomic relationship matrix into the analysis model; (4) investigate the improvement in genomic prediction accuracy that is achievable with a germplasm of this size and compare it to a simplified prediction approach based on selection of finite QTL.

Materials and methods

Plant material and phenotype data

A panel of diverse bread wheat lines was provided by Australian Grain Technologies Pty Ltd (AGT). The panel consists of lines from preliminary yield testing (PYT) and advanced yield testing (AYT) stages of the AGT breeding programmes. Online Resource 1 summarises the panel and its subsets. The PYT-South and AYT-South sets are comprised of lines bred for southern Australia, and the AYT-Other set represents lines from the north eastern, eastern, and western growing regions. PYT material is a combination of F_2 and F_5 derived lines, whereas AYT lines are derived from the F_5 generation or later. By including germplasm from both preliminary and advanced stages of the breeding programme, a set of unselected lines exist for each trait of interest. The panel was phenotyped in 2014 in a dedicated field trial at Roseworthy, South Australia (−34.52, 138.69), which was sown as a non-replicated randomised design with repeated grid checks (1 check per 11 plots). The trial was non-replicated as the large number of lines in the AWP made loading a replicated trial logistically infeasible. Dimensions of the trial were 476 rows by 24 ranges, and plot size was $3m^2$. The trial was managed according to best local practice including fertiliser applications to maximise grain yield and grain quality, and fungicide applications to control disease. Table 1 details the phenotyping methods and summarises the data for each trait, while Online Resource 2 highlights the phenotypic differences between the germplasm sets. Raw phenotype data are provided in Online Resource 3.

Genotype data

Genotyping platform

Marker genotyping was performed using a custom Axiom™ Affymetrix array containing 18,101 SNP markers. To build the customised array, SNPs were selected from previous variant identifications and SNP screenings in a range of genotyping platforms. The most prominent platform was a high-density Axiom™ array developed in the collaborative French BreedWheat project (Etienne Paux, personal communication) consisting of 420,000 diverse SNPs. This was used to genotype a panel of approximately 200 wheat accessions from a range of geographic regions (western Europe, eastern Europe, North America, Australia, and exotic sources) for use in SNP selection. To achieve adequate and even coverage of the genome, SNPs were clustered into 20,000 groups based on a linkage disequilibrium threshold of $r^2 = 0.96$. One SNP per group was then selected based on technical quality, information content, and to have a call rate greater than 70%. It was ensured that SNPs could be accurately read as co-dominant markers by confirming they generated clear allele clusters, and required fewer probes. A final selection was then carried out based on initial batches from the 20K array, and 18,101 of the most reliable and reproducible SNPs were selected. This final selection of SNPs was used to build the custom 18K Axiom™ 384 layout array from Affymetrix. Arrays were read using the GeneTitan Multi-Channel Instrument, and allele calls were made using Axiom™ Analysis Suite software by Affymetrix.

Consensus map

To provide an accurate assessment of LD between SNP markers in the AWP a consensus map was constructed using nine doubled haploid (DH) populations (Online Resource 1) genotyped on the custom Axiom™ Affymetrix array. The DH populations represent key families of Australian wheat germplasm and were chosen to maximise polymorphic markers across the genome. The individual SNP DH

Table 1 Summary of the phenotype data and the methods used for collection

Trait	Assessment method	Scale	Mean	SD
Growth habit	Visual	1–9; 1 = erect	2.4	1.0
Leaf width	Visual	1–9; 1 = narrow	4.8	1.4
Biomass	Visual	1–9; 1 = low biomass	6.9	1.3
NDVI	GreenSeeker[a]	NDVI	0.68	0.1
Physiological yellows	Visual	1–9; 1 = low expression	1.7	0.9
Relative maturity	Visual	Zadoks scale[b]	53	5.7
Greenness	Visual	1–9; 1 = pale green	5.7	1.5
Glaucousness	Visual	1–9; 1 = low expression	3.5	2.0
Leaf loss	Visual	1–9; 1 = low loss	4.6	1.7
Plant height	Visual	1–9; 1 = short	5.2	1.1
Grain yield	Machine harvester	kg/ha	5124	655
Test weight	Chondrometer	kg/hl	84.4	1.8
Thousand kernel weight	Image analysis	TKW	37.5	4.6
Grain protein	NIR[c]	Concentration (%)	11.1	0.9

Mean and standard deviation are calculated from the raw phenotype data

[a] Trimble (2016)

[b] Zadoks et al. (1974)

[c] Zeutec (2016)

linkage maps were constructed using a synergistic combination of the R/qtl (Broman and Sen 2009; Broman and Wu 2015) and R/ASMap (Taylor and Butler 2017) packages available in the R statistical computing environment (R Development Core Team 2015). Before construction, individual marker sets were thoroughly diagnostically checked and problematic lines and markers containing excessive segregation distortion or missing values were removed. For each DH population, markers were clustered and optimally ordered using the MSTmap (Wu et al. 2008) functionality available in R/ASMap. The individual constructed linkage maps were scrutinized and lines with excessive recombination or markers exhibiting large numbers of double crossovers removed. Chromosomal alignment of linkage maps occurred sequentially with initial alignment of the Kukri/RAC875 SNP map performed using legacy markers from the pre-existing Kukri/RAC875 SSR/DArT map (Bennett et al. 2012; Edwards 2012). All other DH SNP linkage maps were then aligned to the Kukri/RAC875 SNP map through commonality of markers. A summary of the final individual DH linkage maps and their common markers is given in Online Resource 4.

The complete set of nine DH linkage maps (marker names and positions) were then used in MergeMap (Wu et al. 2011) to form a consensus map. To ensure the greatest marker position accuracy, the population size for each bi-parental linkage map was also passed to MergeMap as a set of predefined weights. A total 13,747 markers were assigned to linkage groups and relative positions across the 21 chromosomes of the wheat genome. The MergeMap algorithm is known to inflate consensus map linkage group distances (Close et al. 2009; Cavanagh et al. 2013; Wang et al. 2014). Scaling of the consensus map in this research used a minimum mean square criterion. Let M_{ijk} be the position of the kth marker in the jth linkage group of the ith bi-parental linkage map and C_{jk} be the position of the equivalent marker in the jth linkage group of the consensus map. The optimal scaling factor R_j applied to the jth consensus linkage group was then derived using

$$\arg\min_{R_j \in \mathbb{R}} \sum_{i=1}^{9} N_{ij} \sum_{k=1}^{N_i} (C_{jk} R_j - M_{ijk})^2$$

The function is easily minimised by considering $R_j = \bar{D}_j / D_j^c$ where D_j^c is the length of the jth observed consensus linkage group and profiling $\bar{D}_j$ over a conservative window in the vicinity of the average length of jth linkage groups from the bi-parental linkage maps. This procedure was repeated for all 21 chromosomes and the consensus map was scaled accordingly. Assessment of LD was then based on these scaled positions within each of the chromosomes. Table 2 summarises the consensus map by detailing individual chromosomes, chromosome groups and genomes, while final scaled (as well as unscaled) consensus map positions for the 13,747 markers are given in Online Resource 4.

Imputation

Before imputation, markers were omitted if they had a minor allele frequency less than 1%. The remaining markers in the SNP array had a low missing call rate of 1%. The substantial numerical dimensions of the complete SNP array made it computationally impractical to impute missing allele scores using algorithms based on unclustered and unsorted markers (Rutkoski et al. 2013). To reduce this computational burden, chromosomal identifications of the markers from the consensus map were used to subset the SNP marker set. The remaining 4354 markers with no consensus map chromosomal assignment were then linked to these subsets using LD. For each chromosome subset, the K-nearest neighbour (KNN) method (Troyanskaya et al. 2001) implemented in the R package pedicure (Butler 2016) was used to impute missing allele calls from the weighted average of the data points at the nearest 10 markers. The complete marker matrix of 10,375 lines by 17,181 markers from herein was defined as **M**.

Statistical methods

Statistical modelling

An initial baseline linear mixed model was used to provide a preliminary assessment of the genetic variation of the traits collected from the Roseworthy trial. For a given vector of trait observations, $\mathbf{y} = (y_1, \ldots, y_n)$, the linear mixed model had the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{Z}_g\mathbf{g}_t + \mathbf{e} \qquad (1)$$

Here, $\boldsymbol{\tau}$ is a vector of fixed effects, with associated design matrix $\mathbf{X}$, and contained an intercept and potential coefficients for covariates in $\mathbf{X}$ explaining trends across the experimental layout. Non-genetic variation associated with the design of the experiment, such as blocks in the experimental area, was accounted for through the random effects $\mathbf{u}$ with indicator design matrix $\mathbf{Z}$ with $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{I})$. Other remaining sources of non-genetic environmental variation were modelled through the residual error $\mathbf{e}$ which was assumed to have the form $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{R})$ with $\mathbf{R} = \boldsymbol{\Sigma}_r(\rho_r) \otimes \boldsymbol{\Sigma}_c(\rho_c)$ defining a two-dimensional separable AR1 ⊗ AR1 correlation structure in the rows and column direction of the experiment (Gilmour et al. 1997). In the baseline model the total genetic variation of the 10,375 AWP lines was captured using the random effects $\mathbf{g}_t$ with indicator design matrix $\mathbf{Z}_g$ which maps AWP lines to the appropriate random effects in $\mathbf{g}_t$. These effects were assumed to have the distribution

Table 2 Summary of the consensus linkage map

	Total markers	Map positions	Markers per map position	Genetic length	Mean interval[a]
1A	838	308	2.7	129	0.42
1B	905	250	3.6	136	0.55
1D	222	112	2.0	137	1.22
2A	777	226	3.4	128	0.57
2B	1074	286	3.8	147	0.51
2D	204	109	1.9	159	1.46
3A	909	267	3.4	156	0.58
3B	1175	282	4.2	145	0.51
3D	246	120	2.1	152	1.27
4A	652	276	2.4	168	0.61
4B	490	184	2.7	113	0.61
4D	237	120	2.0	129	1.08
5A	922	350	2.6	190	0.54
5B	1057	340	3.1	172	0.51
5D	236	147	1.6	198	1.35
6A	590	208	2.8	127	0.61
6B	893	237	3.8	114	0.48
6D	209	101	2.1	142	1.40
7A	1068	319	3.3	164	0.51
7B	814	221	3.7	147	0.66
7D	229	140	1.6	171	1.22
Genome A	5756	1954	2.9	1062	0.54
Genome B	6408	1800	3.6	974	0.54
Genome D	1583	849	1.9	1088	1.28
Group 1	1965	670	2.9	403	0.60
Group 2	2055	621	3.3	434	0.70
Group 3	2330	669	3.5	453	0.68
Group 4	1379	580	2.4	410	0.71
Group 5	2215	837	2.6	560	0.67
Group 6	1692	546	3.1	383	0.70
Group 7	2111	680	3.1	482	0.71
Total	13,747	4603	3.0	3124	0.68

[a] Mean interval (cM) between unique map positions

$\mathbf{g}_t \sim N(\mathbf{0}, \sigma_t^2\mathbf{I})$ and the set of effects $(\mathbf{u}, \mathbf{g}_t, \mathbf{e})$ were considered to be mutually independent.

For each of the traits, the baseline model (1) was then extended by partitioning the total genetic effects into additive marker and residual genetic effects to form the marker linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{Z}_g(\mathbf{M}\mathbf{g}_m + \mathbf{g}_p) + \mathbf{e} \tag{2}$$

where $\mathbf{g}_m$ is a vector of marker effects and $\mathbf{g}_p$ is a vector of residual genetic effects. The effects were assumed to be distributed $\mathbf{g}_m \sim N(\mathbf{0}, \sigma_a^2\mathbf{I})$ and $\mathbf{g}_p \sim N(\mathbf{0}, \sigma_p^2\mathbf{I})$ with $(\mathbf{u}, \mathbf{g}_m, \mathbf{g}_p, \mathbf{e})$ mutually independent. The large number of markers in **M**, coupled with the substantial number of lines in the population made the fitting of (2) computationally prohibitive. For this reason an alternative formulation using the approach of Strandén and Garrick (2009) was sought. Let $\mathbf{g}_a$ define a set of additive genotype effects with $\mathbf{g}_a = \mathbf{M}\mathbf{g}_m$ then the genotype linear mixed model used was

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{Z}_g(\mathbf{g}_a + \mathbf{g}_p) + \mathbf{e} \tag{3}$$

where $\mathbf{g}_a \sim N(\mathbf{0}, \sigma_a^2\mathbf{G})$ and $\mathbf{G} = \mathbf{M}\mathbf{M}^T$ is a $10,375 \times 10,375$ additive relationship matrix. For the purpose of providing an appropriate scaling, **G** was replaced by $\mathbf{G}_s = \mathbf{M}\mathbf{M}^T/r$ with $r = \text{trace}(\mathbf{G})/10{,}375$ (Forni et al. 2011). An eigen decomposition of $\mathbf{G}_s$ revealed only positive eigenvalues indicating $\mathbf{G}_s$ was positive definite and could be safely inverted.

Estimation of the parameters for the linear mixed models (1) and (3) occurred iteratively. Fixed effect estimates and predictions of random effects were determined through direct solving of the mixed model equations (Henderson 1953).

Variance parameters were estimated using residual maximum likelihoood (REML) (Patterson and Thompson 1971). From the fitted baseline model (1) broad sense heritabilities were then calculated for each of the traits using REML estimates of the variance parameters, namely $H^2 = \sigma_t^2/(\sigma_t^2 + \sigma^2)$. For the fitted additive genotype model (3) the broad sense heritability was calculated by replacing the total genetic variability in H^2 by $\sigma_t^2 = \sigma_a^2 + \sigma_p^2$. Narrow sense heritabilities were also calculated using $h^2 = \sigma_a^2/(\sigma_t^2 + \sigma^2)$.

Genomic prediction

Using mixed model results, genomic best linear unbiased predictions of the additive genetic effects $\mathbf{g}_a$ and predictions of the residual genetic effects $\mathbf{g}_p$ in (3) were immediately determined for each trait using

$$\begin{aligned}\tilde{\mathbf{g}}_a &= \sigma_a^2\mathbf{G}_s\mathbf{Z}_g^T\mathbf{P}\mathbf{y}\\ \tilde{\mathbf{g}}_p &= \sigma_p^2\mathbf{Z}_g^T\mathbf{P}\mathbf{y}\end{aligned} \tag{4}$$

where $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}^{-1}$ and $\mathbf{H} = \sigma^2\mathbf{R} + \sigma_u^2\mathbf{Z}\mathbf{Z}^T + \mathbf{Z}_g(\sigma_a^2\mathbf{G}_s + \sigma_p^2\mathbf{I})\mathbf{Z}_g^T$. The additive genetic effects, $\tilde{\mathbf{g}}_a$ reflect the breeding value of lines estimated from phenotpyic and genetic information. Both $\tilde{\mathbf{g}}_a$ and $\tilde{\mathbf{g}}_p$ were used to investigate the additive and residual genetic relationships between the analysed Roseworthy traits.

From the marker linear mixed model (2), predicted marker effects were immediately calculated using

$$\tilde{\mathbf{g}}_m = \sigma_a^2\mathbf{M}^T\mathbf{Z}_g^T\mathbf{P}\mathbf{y} = \mathbf{M}^T\mathbf{G}_s^{-1}\tilde{\mathbf{g}}_a \tag{5}$$

This result ensured the marker effects were efficiently derived from the additive genetic values for the lines given by (4). Inversion of $\mathbf{G}_s$ would usually be computationally expensive but was very efficient using the highly parallelised Basic Linear Algebra Subprograms available in the Intel$^{\text{TM}}$ Math Kernel Libraries. Given a new set of lines with marker data $\mathbf{M}^*$ genotyped across identical markers in $\mathbf{M}$, genomic predictions for the new lines can then be determined using the simple equation $\tilde{\mathbf{g}}^* = \mathbf{M}^*\tilde{\mathbf{g}}_m$, utilizing the complete set of predicted marker effects.

To evaluate the power of the genomic prediction approach using the results derived from the full additive genotype linear mixed model (3), it was compared to a simplified prediction approach based on finite selection of putative QTL. To provide a mechanism for selecting important markers linked to a QTL for each of the traits, the complete set of marker outlier statistics were calculated using the formula derived in Verbyla et al. (2007). For any given trait, the kth marker outlier statistic is

$$t_k = \frac{\tilde{g}_{m;k}^2}{\text{var}\,(\tilde{g}_{m;k})}$$

where $\tilde{g}_{m;k}$ is the kth marker effect obtained directly from (5) with its variance extracted from the diagonal components of the variance matrix $\text{var}\,(\tilde{\mathbf{g}}_m) = \mathbf{M}^T\mathbf{G}_s^{-1}\,\text{var}\,(\tilde{\mathbf{g}}_a)\mathbf{G}_s^{-1}\mathbf{M}$. In most modern linear mixed modelling software $\text{var}\,(\tilde{\mathbf{g}}_a)$ is usually available from the fitted additive genotype model in (3), ensuring efficient computing of the variance of the predicted marker effects. For each of the traits, the largest one and five marker outlier statistics were identified iteratively using a consensus map exclusion window of 25cM either side of any selected marker. The selected markers were then extracted from $\mathbf{M}$, denoted $\mathbf{M}_1$ and $\mathbf{M}_5$, respectively, placed in the baseline model (1) as an additive set of QTL fixed effects

$$\mathbf{y} = \mathbf{Z}_g\mathbf{M}_j\boldsymbol{\beta}_j + \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{Z}_g\mathbf{g}_p + \mathbf{e} \tag{6}$$

where $j = (1, 5)$ and $\boldsymbol{\beta}_j$ are the QTL fixed effect parameters for the selected markers in $\mathbf{M}_j$. In this model, $\mathbf{g}_t$ has been replaced with a residual genetic effect $\mathbf{g}_p$ as the inclusion of markers strongly linked to QTL will absorb genetic variation. The genetic value of the lines were then calculated directly from the equation $\tilde{\mathbf{g}}_a = \mathbf{M}_j\hat{\boldsymbol{\beta}}_j$, where $\hat{\boldsymbol{\beta}}_j$ are estimates of the QTL fixed effects extracted from the fitted model of (6). Similarly, given a new set of lines with marker data for the selected markers, $\mathbf{M}_j^*$, QTL-based predictions for the new lines can be calculated using $\tilde{\mathbf{g}}^* = \mathbf{M}_j^*\hat{\boldsymbol{\beta}}_j$.

Prediction accuracy

To provide an informative comparison with genomic prediction results discussed in the plant research literature, the predictive ability of the fitted additive genotype model (3), as well as of predictions obtained using selected QTL effects estimated from the fitted model of (6), was calculated for each of the traits using fivefold cross-validation. The cross-validation method initially randomly partitioned the AWP lines into five equal subsets. Let $(\mathbf{g}_a^{(i)}, \mathbf{g}_p^{(i)})$ be the additive and residual genetic effects of the AWP lines in the ith subset (validation set) and $(\mathbf{g}_a^{(-i)}, \mathbf{g}_p^{(-i)})$ the additive and residual genetic effects of the AWP lines remaining in the other four (training set). The cross-validation for each prediction method was conducted sequentially for each of the folds $i = 1, \ldots, 5$. For the genomic prediction approach incorporating the additive relationship matrix, $(\mathbf{g}_a^{(-i)}, \mathbf{g}_p^{(-i)})$ were fitted as additive and residual genetic effects in the additive genotype model, the additive genetic values for $\tilde{\mathbf{g}}_a^{(-i)}$ were derived using (3) and marker effects, $\tilde{\mathbf{g}}_m^{(-i)}$, were calculated using (5). The AWP lines in the ith validation set were then predicted using $\tilde{\mathbf{g}}_a^{(i)} = \mathbf{M}_j^{(i)}\tilde{\mathbf{g}}_m^{(-i)}$. For prediction methods using selected QTL, $\mathbf{g}_p^{(-i)}$ was fitted in (6) and QTL effects $\hat{\boldsymbol{\beta}}_j^{(-i)}$ were extracted and used to calculate predictions for the validation set of AWP lines using $\tilde{\mathbf{g}}_a^{(i)} = \mathbf{M}_j^{(i)}\hat{\boldsymbol{\beta}}_j^{(-i)}$. Prediction accuracies were calculated by correlating the validation set predictions obtained from each cross-validation fold, $\{\tilde{\mathbf{g}}_a^{(i)}; i = 1, \ldots, 5\}$, to their full additive genetic values ($\tilde{\mathbf{g}}_a$) extracted from the additive genotype model containing the complete set of lines. To enable the comparison of these results to those of previous studies, validation set predictions were also correlated to their corresponding total genetic values obtained from the baseline model, and divided by the square root of the heritability of the baseline model (Heffner et al. 2011b; Estaghvirou et al. 2013; Battenfield et al. 2016). Comparing predictions to both total and additive genetic values enabled an assessment of prediction accuracy to be made for line selection and parental value, respectively.

Computations

All linear mixed modelling was conducted using the ASReml-R package (Butler et al. 2009) available in the R statistical computing environment (R Core Team 2017). Trait models containing the full additive relationship matrix took an average of 60 h computational time to converge on a Windows 10 box with a quad core Intel$^{\text{TM}}$ i7-6700K (4.00Ghz) with 64Gb RAM.

Results

Linkage disequilibrium

Linkage disequilibrium was assessed by calculating r^2 values between marker pairs within each consensus map chromosome (Fig. 1). In the full panel, the median r^2 for marker pairs with proximity less than 2 cM is just 0.12, and this steadily decreases as the distance between a pair of markers increases. However, there is significant variation in the r^2 value between markers in very close proximity, with some

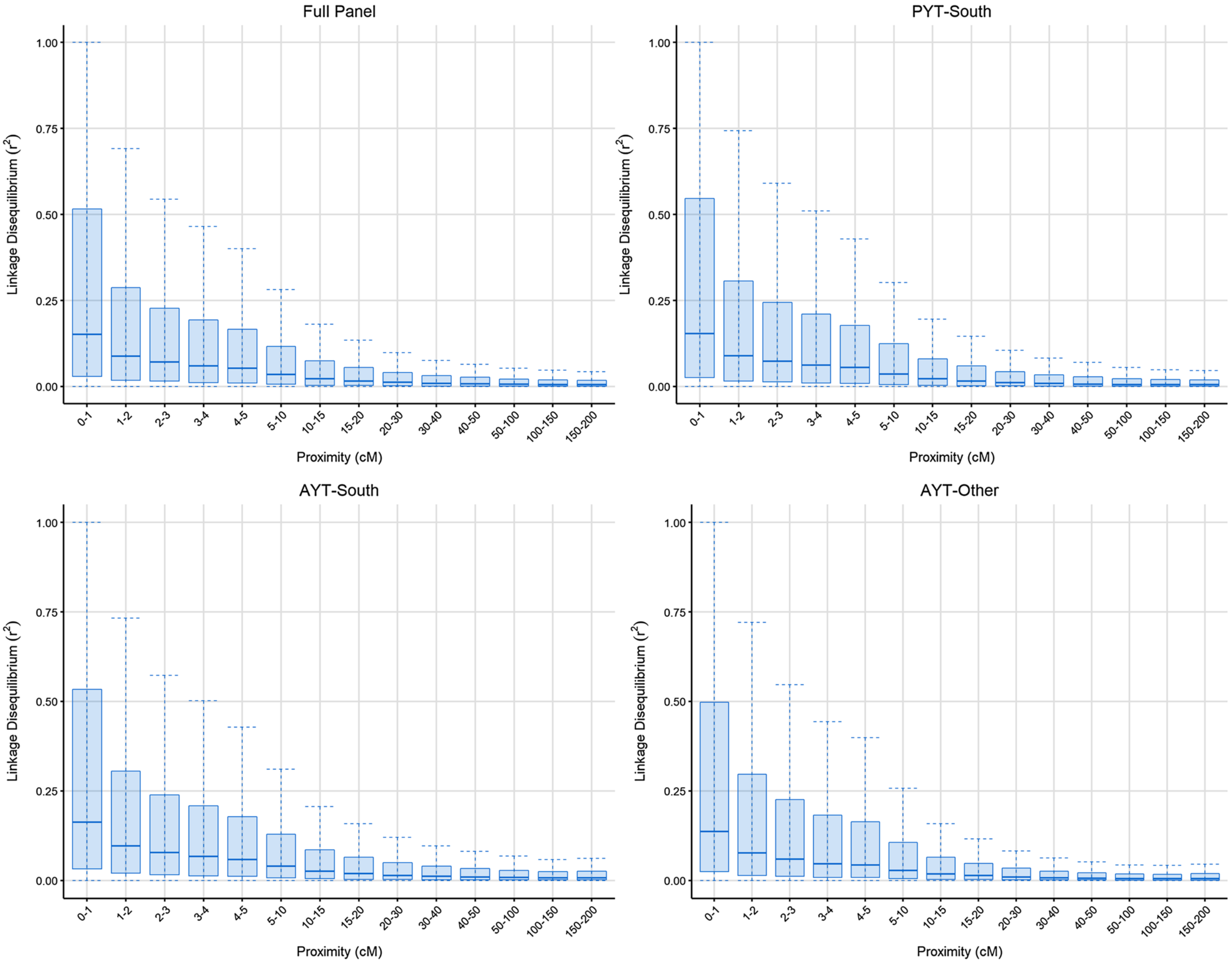

Fig. 1 Boxplots comparing linkage disequilibrium (r^2) of marker pairs with their proximity on the consensus map

being in complete LD with each other. The boxplots clearly show that this variation decreases rapidly with increasing distance, and plateaus off after proximity exceeds 20 cM. The broad pattern of LD decay was very similar for each of the germplasm sets, but there were subtle differences for close marker pairs (<2 cM) with AYT-South showing slightly higher LD than PYT-South, which itself was higher than AYT-Other.

Genetic trait correlations

From each of the traits, the additive genetic values and residual genetic values were extracted from their respective fitted additive genotype models and used to understand genetic relationships between the traits. Table 3 presents the pairwise additive and residual genetic correlations between traits analysed in the 2014 Roseworthy field trial. The two correlation measures largely agreed, with a correlation of 0.79 across the 91 trait pairs. Of the 91 trait pairs, 74 had correlations in the same direction, and those that differed in direction were all near zero. Additive genetic correlations were overall stronger than residual genetic with an absolute mean of 0.26 compared to 0.14. Notable correlations include the well-known strong negative relationship between grain yield and grain protein, with an additive correlation of −0.55 and a residual genetic of −0.30. A negative relationship was also observed between grain protein and test weight (additive correlation −0.22, residual genetic −0.43). Strong positive relationships were observed between test weight and thousand kernel weight (TKW) (additive correlation: 0.37, residual genetic 0.52), and relative maturity score and biomass (additive correlation 0.76, residual genetic 0.42).

A comparison of additive and baseline models

All traits collected from the Roseworthy experiment were analysed and results from the fitted baseline models and additive genotype linear mixed models are compared in Table 4. Additive models had significantly higher log-likelihood (model fit) for all traits, with an average improvement of 44% over the equivalent baseline models. The additive model also improved broad sense heritability for all traits, with an average increase of 24%. Narrow sense heritabilities of the additive models were comparable with the broad sense heritability from the equivalent baseline models, being just 0.5% lower on average. The proportion of the genetic variance that was additive averaged 81% across all traits, and ranged from 58% (NDVI) to 91% (grain size). There was a strong positive relationship between the improvement in model fit obtained with the additive model and narrow sense heritability ($r = 0.86$).

Prediction accuracy

Table 5 presents the fivefold cross-validation accuracies of the genomic predictions and QTL-based predictions for all 14 traits. Prediction accuracy was assessed by correlating genomic and QTL-based predictions to both the additive genetic values from the full additive genotype model (shown to be the model of best fit for every trait, Table 4), and the total genetic values from the baseline model. When comparing genomic predictions to total genetic values, prediction accuracies were varied with a range between 0.55 (yellows) and 0.85 (TKW). As expected, comparing these predictions to the additive genetic values produced higher and more consistent prediction

Table 3 Pairwise genetic correlations between traits from the Roseworthy experiment

	Bm.	Gl.	GP	GY	Gr.	GH	PH	LL	LW	Mat.	NDVI	TW	TKW	Yl.
Biomass	–	−0.24	−0.39	0.49	−0.44	−0.45	0.10	0.69	0.49	0.76	0.51	0.15	0.19	−0.34
Glaucousness	−0.18	–	0.41	−0.01	0.73	−0.23	−0.04	−0.41	0.24	−0.30	−0.28	−0.04	0.13	0.41
Grain protein	−0.14	0.15	–	−0.55	0.50	0.10	0.02	−0.40	−0.08	−0.39	−0.34	−0.22	−0.23	0.35
Grain yield	0.27	−0.03	−0.30	–	0.06	−0.06	0.11	0.01	0.10	0.19	0.16	0.28	0.23	−0.19
Greenness	−0.14	0.23	0.22	−0.15	–	−0.14	−0.05	−0.61	0.02	−0.43	−0.45	−0.02	−0.01	0.49
Growth habit	−0.15	0.07	0.20	−0.30	0.11	–	0.15	−0.36	−0.59	−0.48	0.25	−0.10	−0.30	−0.25
Plant height	0.19	−0.14	−0.04	−0.27	−0.08	0.05	–	−0.10	−0.01	−0.10	0.12	0.09	−0.04	−0.16
Leaf loss	0.32	−0.16	−0.22	0.28	−0.26	−0.23	−0.05	–	0.33	0.75	0.31	0.10	0.23	−0.18
Leaf width	0.22	−0.04	−0.05	0.30	−0.03	−0.11	0.06	0.13	–	0.43	0.04	0.05	0.37	0.09
Maturity	0.42	−0.19	−0.30	0.36	−0.19	−0.32	0.01	0.45	0.19	–	0.24	0.24	0.29	−0.20
NDVI	0.34	−0.13	−0.01	0.43	−0.10	0.12	0.20	0.14	0.05	0.08	–	0.04	0.06	−0.51
Test weight	0.12	−0.10	−0.43	0.29	−0.09	−0.17	0.01	0.02	0.02	0.21	0.12	–	0.37	0.00
TKW	0.14	−0.08	−0.33	0.39	−0.08	−0.20	0.06	0.12	0.15	0.35	−0.06	0.52	–	0.11
Yellows	−0.11	0.04	0.00	−0.25	−0.04	−0.07	−0.15	0.06	−0.01	−0.06	−0.04	−0.01	−0.05	–

Additive genetic correlations are in the upper triangle and residual genetic are in the lower triangle

Table 4 Comparison of the baseline and genomic mixed linear models

	Baseline model		Genomic model			
	H^2	Log *l*	H^2	h^2	Log *l*	Add. var. (%)[a]
Biomass	0.56	−4113	0.75	0.56	−2401	75
Glaucousness	0.81	−12,424	0.89	0.76	−8370	86
Grain protein	0.57	−1119	0.75	0.62	1517	82
Grain yield	0.44	−76,861	0.63	0.45	−75,322	72
Greenness	0.64	−9271	0.75	0.58	−6479	77
Growth habit	0.71	−4148	0.89	0.78	−1781	88
Plant height	0.74	−5212	0.91	0.81	−2655	89
Leaf loss	0.67	−10,067	0.83	0.69	−7648	82
Leaf width	0.71	−7888	0.86	0.75	−4674	87
Maturity	0.92	−24,045	0.98	0.91	−20,562	93
NDVI	0.45	25, 269	0.62	0.36	26, 160	58
Test weight	0.75	−10,566	0.91	0.82	−7546	90
TKW	0.79	−21,047	0.93	0.85	−17076	91
Yellows	0.73	−3662	0.82	0.53	−2418	65

Broad sense heritabilities are presented for each model, and narrow sense for the genomic model as there is no term in the base model to capture the additive genetic variance. Model fit is compared through the log-likelihood measure

[a] Proportion of the variance accounted for by the model that is additive

Table 5 Fivefold cross-validation accuracy of genomic and QTL prediction models (one and five QTL)

	Genomic		One QTL		Five QTL	
	Additive [a]	Total [b]	Additive	Total	Additive	Total
Biomass	0.97	0.72	0.26	0.20	0.46	0.48
Glaucousness	0.98	0.82	0.49	0.45	0.76	0.68
Grain protein	0.97	0.84	0.16	0.16	0.59	0.54
Grain yield	0.97	0.71	0.19	0.16	0.64	0.51
Greenness	0.98	0.80	0.54	0.44	0.78	0.65
Growth habit	0.96	0.75	0.36	0.30	0.59	0.50
Plant height	0.96	0.76	0.28	0.24	0.48	0.43
Leaf loss	0.97	0.77	0.41	0.37	0.55	0.54
Leaf width	0.98	0.81	0.26	0.24	0.54	0.46
Maturity	0.96	0.77	0.26	0.25	0.59	0.55
NDVI	0.96	0.56	0.20	0.15	0.42	0.31
Test weight	0.96	0.80	0.10	0.11	0.43	0.39
TKW	0.97	0.85	0.38	0.33	0.52	0.49
Yellows	0.97	0.55	0.17	0.15	0.63	0.41

[a] Correlation between the predicted values and the additive genetic values from the full genomic model

[b] Correlation between the predicted values and the total genetic values from the baseline model, divided by the square root of the broad sense heritability

accuracies with all traits falling between 0.96 and 0.98. Using one QTL to predict performance was much less accurate with traits ranging between 0.11 (test weight) and 0.45 (glaucousness) when comparing to total genetic values, and between 0.10 (test weight) and 0.54 (greenness) when comparing to additive genetic values. The five QTL model yielded prediction accuracies ranging from 0.31 (NDVI) to 0.68 (glaucousness) when compared to total genetic values, and between 0.42 (NDVI) and 0.78 (greenness) when compared to additive genetic values. There was a strong positive relationship ($r = 0.84$) between genomic prediction accuracy calculated using total genetic values and the proportion of genetic variance that was additive for the trait. This relationship was negligible for genomic prediction accuracies calculated using additive genetic values values ($r = -0.13$).

Discussion

Previous applications of GS have predominantly used wheat germplasm collections of approximately 500 individuals (Crossa et al. 2010; Heslot et al. 2012, 2013; Dawson et al. 2013; Lado et al. 2013), while two recent studies used panels containing over 3000 individuals (He et al. 2016, 2017). This research has been invaluable in promoting the concept of GS in wheat, and providing a framework for future research. The natural progression is to work with larger datasets that provide more direct relevance to large-scale breeding programmes. In this study we used a panel of 10,375 wheat breeding lines to investigate the genomic prediction accuracy achievable in germplasm of this size and nature. We also compare these prediction accuracies to those achieved with models using a finite number of QTL, which are reflective of the style of marker-assisted selection already being used within wheat breeding programmes. We also assessed the extent of LD present in the germplasm and investigated genetic correlations between traits.

Significant LD within a training set leads to low genetic resolution and results in prediction calibrations which break down quickly in a breeding programme (Hickey et al. 2014). The panel presented here contains very low levels of LD compared to multi-parent advanced inter-cross (MAGIC) populations (Huang et al. 2012), and is more comparable to diverse germplasm collections (Chao et al. 2010; Sukumaran et al. 2015). This information, along with the high prediction accuracies we observed, highlights that our calibration successfully exploited short haplotype effects rather than long. Hickey et al. (2014) suggested that this type of calibration would retain prediction accuracy over multiple generations of inter-crossing, which future work will investigate.

The additive and residual genetic correlations between 91 trait combinations show that while the two measures commonly mirror each other, they do at times differ (glaucousness–greenness, leaf loss–maturity). A negative relationship between grain protein and grain yield has frequently been identified at a phenotypic level (Brooks et al. 1982; Jenner et al. 1991; Simmonds 1995; Oury and Godin 2007), and here we extend this understanding by showing the relationship exists at both an additive and residual genetic level. The same applies for the strong positive relationship between test weight and TKW, where phenotypic correlations were previously demonstrated by (Sharma and Anderson 2004; Rharrabti et al. 2003). Negative correlations between grain protein and test weight, as observed here, are common when plants are stressed during grain fill (Sadras et al. 2002) as the Roseworthy experiment was. The positive additive and residual genetic correlations between grain yield and relative maturity score were caused by the dry finish to the season, which favoured early maturing lines.

Incorporating the genomic relationship matrix into the linear mixed models vastly improved the model fit for all traits. This translates to more genetic variation of the trait being captured by the model, and also more accurate partitioning of variance into genetic (subsequently partitioned into additive and residual genetic) and residual error sources. The strong positive correlation between improvement in model fit and narrow sense heritability demonstrates that the additive relationship matrix improves the model by more accurately capturing additive genetic variance. Traits with a high proportion of additive genetic variance will, therefore, benefit most from the inclusion of the marker relationship matrix in the model.

The efficacy of genomic prediction is typically assessed by means of cross-validation, where predictions of the validation set are correlated to the corresponding phenotypic estimated breeding values (Crossa et al. 2010; Lado et al. 2013). These phenotypic values (in this case a best linear unbiased prediction) represent both additive and residual genetic variance, whereas the genomic prediction represents only additive genetic variance. This discrepancy between the two values results in lower perceived prediction accuracies that are skewed according to the proportion of trait variance that is additive. The results presented in Table 5 demonstrate this as the genomic prediction accuracies produced by correlating predictions to total genetic values and dividing by the square root of heritability were significantly lower than those produced by correlating to additive genetic values, and were also strongly related to the proportion of genetic variance that is additive. Correlating cross-validation predictions directly to the additive genetic values, therefore, provides a purer measure of prediction accuracy as both values contain only additive genetic variance, which prevents the proportion of additive variance from confounding the measure. Breeders can then use the prediction accuracy of a given trait (as measured by correlating to additive genetic values) to judge how effective GS will be for selecting lines with high breeding value (parents), and use both the prediction accuracy and the proportion of additive variance to judge how effective GS will be for selecting lines with high phenotypic performance (varieties). The concept of separating these two breeding objectives was investigated by Gaynor et al. (2017) and was found to significantly increase the rate of genetic gain.

Genomic prediction accuracy was very high for all traits when comparing to additive genetic values. This suggests that genomic selection is promising for all traits when the breeder is interested in additive genetic variance, i.e. when selecting parents. When assessed against total genetic values, cross-validation accuracies for grain yield, maturity, TKW, plant height and grain protein were all higher than those reported in previous studies (Crossa et al. 2010; Heffner et al. 2011b; Heslot et al. 2012, 2013; Poland et al.

2012; Dawson et al. 2013; Lado et al. 2013; He et al. 2016). The prediction accuracy improvement is likely due to larger population size of this study compared to those previous (between 254 and 2325). In addition, previous studies sometimes sourced phenotype data from multiple environments which introduce genotype by environment (GxE) variation and decrease prediction accuracy. In this study we used just one environment to remove the confounding effect of GxE and gain a more direct assessment of genomic prediction accuracy in the most optimal scenario. However, the prediction accuracies observed here were still higher than previous cross-validation accuracies produced within one environment, showing that larger population size is important in achieving high prediction accuracy.

QTL-based predictions calculated from five selected QTL were more accurate for all traits than those utilizing one QTL, while the use of genomic prediction was significantly more accurate than both. This result is in line with previous comparisons between QTL-based prediction and genomic prediction in different traits. Rutkoski et al. (2012) found that genome-wide prediction models outperformed targeted marker models for most traits related to Fusarium head blight, while Heffner et al. (2011a) showed that genomic predictions were significantly more accurate than QTL-based predictions for grain quality traits. The research presented here demonstrates that this trend holds true for grain yield, physical grain quality, and physiological traits. The traits that were most accurately predicted by QTL were greenness and glaucousness. These two traits expressed several large effect QTL (Online Resource 5) which explain their high prediction accuracy (Desta and Ortiz 2014). NDVI showed low QTL-based prediction accuracy as there were no moderate or large effect QTL influencing the trait (Online Resource 5).

The dataset used in this study represents an unprecedented resource for studying the efficacy and application of genomic selection in bread wheat. We showed that incorporating a genomic additive relationship matrix into the linear mixed model significantly improved the model fit and increased trait heritability. The fivefold cross-validation produced higher genomic prediction accuracies than those from previous studies which used smaller populations. We also showed that for all traits assessed in this research, genomic prediction was significantly more accurate than QTL-based prediction, but as expected the improvement was smaller for qualitative traits. This panel will be used in future work to investigate the effects of population size, population structure, and GxE interaction on genomic prediction accuracy.

Addendum Marker data will be available for downloading as supplementary material 12 months after publication, or in advance from the authors subject to the terms of a material transfer agreement.

Author contribution statement AN: manuscript preparation, phenotypic data generation, analysis of phenotypic and genetic data. JT: construction of the genetic linkage and consensus maps; PhD co-supervisor of Adam Norman. ET: construction of the genetic linkage and consensus maps. PT: generation of several bi-parental populations used in the genetic mapping. JE: PhD co-supervisor of Adam Norman; direction and content of research and the article. JPM: development of SNP genotyping platform. HK: PhD principal supervisor of Adam Norman; direction and content of research and the article.

Acknowledgements We would like to thank Australian Grain Technologies for their contribution in managing the Roseworthy field experiment and funding the research. We acknowledge funding from the Australian Research Council, and the Grains Research and Development Corporation.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Battenfield S, Guzmán C, Gaynor R, Singh R, Peña R, Dreisigacker S, Fritz A, Poland J (2016) Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. Plant Genome. doi:10.3835/plantgenome2016.01.0005

Bennett D, Izanloo A, Reynolds M, Kuchel H, Langridge P, Schnurbusch T (2012) Genetic dissection of grain yield and physical grain quality in bread wheat (Triticum aestivum L.) under water-limited environments. Theor Appl Genet 125(2):255–271

Bentley A, Scutari M, Gosman N, Faure S, Bedford F, Howell P, Cockram J, Rose G, Barber T, Irigoyen J et al (2014) Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. Theor Appl Genet 127(12):2619–2633

Bernardo R (2002) Breeding for quantitative traits in plants. Stemma Press, Woodbury

Bhatt G, Derera N (1975) Genotype x environment interactions for, heritabilities of, and correlations among quality traits in wheat. Euphytica 24(3):597–604

Broman K, Sen S (2009) A guide to QTL mapping with R/ qtl. Springer, New York

Broman K, Wu H (2015) qtl: tools for analayzing QTL experiments. R package version 1.36-6. http://www.CRAN.R-project.org/package=qtl

Brooks A, Jenner C, Aspinall D (1982) Effects of water deficit on endosperm starch granules and on grain physiology of wheat and barley. Funct Plant Biol 9(4):423–436

Butler D (2016) Package 'pedicure': pedigree tools. https://www.asreml.org

Butler D, Cullis B, Gilmour A, Gogel B (2009) ASReml-R reference manual. Queensland Department of Primary Industries, Queensland

Cavanagh C, Chao S, Wang S, Huang B, Stephen S, Kiani S, Forrest K, Saintenac C, Brown-Guedira G, Akhunova A, See D, Bai G, Pumphrey M, Tomar L, Wong D, Kong S, Reynolds M, da Silva M, Bockelman H, Talbert L, Anderson J, Dreisigacker S, Baenziger S, Carter A, Korzun V, Morrell P, Dubcovsky J, Morell M, Sorrells M, Hayden M, Akhunov E (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Nat Acad Sci 110(20):8057–8062. doi:10.1073/pnas.1217133110

Chao S, Dubcovsky J, Dvorak J, Luo M, Baenziger S, Matnyazov R, Clark D, Talbert L, Anderson J, Dreisigacker S et al (2010) Population-and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.). BMC Genom 11(1):727

Close T, Bhat P, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson J, Wanamaker S, Bozdag S, Roose M, Moscou M, Chao S, Varshney R, Szűcs P, Sato K, Hayes P, Matthews D, Kleinhofs A, Muehlbauer G, DeYoung J, Marshall D, Madishetty K, Fenton R, Condamine P, Graner A, Waugh R (2009) Development and implementation of high-throughput SNP genotyping in barley. BMC Genom 10(1):1–13. doi:10.1186/1471-2164-10-582

Collard B, Mackill D (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond B Biol Sci 363(1491):557–572

Cooper M, Messina C, Podlich D, Totir L, Baumgarten A, Hausmann N, Wright D, Graham G (2014) Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. Crop Pasture Sci 65(4):311–336

Crossa J, de Campos G, Pérez P, Gianola D, Burgueño J, Araus J, Makumbi D, Singh R, Dreisigacker S, Yan J et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2):713–724

Dawson J, Endelman J, Heslot N, Crossa J, Poland J, Dreisigacker S, Manès Y, Sorrells M, Jannink J (2013) The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crops Res 154:12–22

Dekkers J, Hospital F et al (2002) The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet 3(1):22–32

Desta Z, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19(9):592–601

Edwards J (2012) A genetic analysis of drought related traits in hexaploid wheat. Ph.D. thesis, The University of Adelaide

Estaghvirou S, Ogutu J, Schulz-Streeck T, Knaak C, Ouzunova M, Gordillo A, Piepho H (2013) Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. BMC Genom 14(1):860

Fischer R, Wood J (1979) Drought resistance in spring wheat cultivars. III.* Yield associations with morpho-physiological traits. Crop Pasture Sci 30(6):1001–1020

Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43(1):1

Gardner K, Wittern L, Mackay I (2016) A highly recombined, high-density, eight-founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments. Plant Biotechnol J 14(6):1406–1417. doi:10.1111/pbi.12504

Gaynor R, Gorjanc G, Bentley A, Ober E, Howell P, Jackson R, Mackay I, Hickey J (2017) A two-part strategy for using genomic selection to develop inbred lines. Crop Sci 56:1–15. doi:10.2135/cropsci2016.09.0742

Gilmour A, Cullis B, Verbyla A (1997) Accounting for natural and extraneous variation in the analysis of field experiments. J Agric Biol Environ Stat 2(3):269–293

Hao C, Wang L, Ge H, Dong Y, Zhang X (2011) Genetic diversity and linkage disequilibrium in Chinese bread wheat (*Triticum aestivum* L.) revealed by SSR markers. PLoS One 6(2):e17279

He S, Schulthess A, Mirdita V, Zhao Y, Korzun V, Bothe R, Ebmeyer E, Reif J, Jiang Y (2016) Genomic selection in a commercial winter wheat population. Theor Appl Genet 129:641–651. doi:10.1007/s00122-015-2655-1

He S, Reif J, Korzun V, Bothe R, Ebmeyer E, Jiang Y (2017) Genome-wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to central europe. Theor Appl Genet 130:635–647. doi:10.1007/s00122-016-2840-x

Heffner E, Jannink J, Iwata H, Souza E, Sorrells M (2011a) Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci 51(6):2597–2606

Heffner E, Jannink J, Sorrells M (2011b) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4(1):65–75

Henderson CR (1953) Estimation of variance and covariance components. Biometrics 9:226–252

Heslot N, Yang H, Sorrells M, Jannink J (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52(1):146–160

Heslot N, Rutkoski J, Poland J, Jannink J, Sorrells M (2013) Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One 8(9):e74612

Hickey J, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna B, Grondona M, Zambelli A, Windhausen V, Mathews K et al (2014) Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci 54(4):1476–1488

Huang B, George A, Forrest K, Kilian A, Hayden M, Morell M, Cavanagh C (2012) A multiparent advanced generation intercross population for genetic analysis in wheat. Plant Biotechnol J 10(7):826–839

Isidro J, Jannink J, Akdemir D, Poland J, Heslot N, Sorrells M (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128(1):145–158

Jenner C, Ugalde T, Aspinall D (1991) The physiology of starch and protein deposition in the endosperm of wheat. Funct Plant Biol 18(3):211–226

Kang H, Sul J, Service S, Zaitlen N, Kong S, Freimer N, Sabatti C, Eskin E et al (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42(4):348–354

Koebner R, Summers R (2003) 21st century wheat breeding: plot selection or plate detection? Trends Biotechnol 21(2):59–63

Lado B, Matus I, Rodríguez A, Inostroza L, Poland J, Belzile F, del Pozo A, Quincke M, Castro M, von Zitzewitz J (2013) Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 3(12):2105–2114

Mackay I, Bansept-Basler P, Barber T, Bentley A, Cockram J, Gosman N, Greenland A, Horsnell R, Howells R, O'Sullivan D et al. (2014) An eight-parent multiparent advanced generation intercross population for winter-sown wheat: creation, properties, and validation. G3 4(9):1603–1610

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

Muir W (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J Anim Breed Genet 124(6):342–355

Nakaya A, Isobe S (2012) Will genomic selection be a practical method for plant breeding? Ann Bot 110(6):1303–1316

Neumann K, Kobiljski B, Denčić S, Varshney R, Börner A (2011) Genome-wide association mapping: a case study in bread wheat (*Triticum aestivum* L.). Mol Breed 27(1):37–58

Oury F, Godin C (2007) Yield and grain protein concentration in bread wheat: how to use the negative relationship between the two characters to identify favourable genotypes? Euphytica 157(1–2):45–57

Patterson H, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58(3):545–554

Podlich D, Cooper M (1998) QU-GENE: a simulation platform for quantitative analysis of genetic models. Bioinformatics 14(7):632–653

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M et al (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5(3):103–113

Pozniak C (2016) IWGSC whole genome shotgun sequencing of chinese spring: towards a reference sequence of wheat. In: Plant and animal genome XXIV conference, plant and animal genome

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

R Development Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org (**ISBN: 3-900051-07-0**)

Rebetzke G, Richards R (1999) Genetic improvement of early vigour in wheat. Crop Pasture Sci 50(3):291–302

Rebetzke G, Rattey A, Farquhar G, Richards R, Condon A (2013) Genomic regions for canopy temperature and their genetic association with stomatal conductance and grain yield in wheat. Funct Plant Biol 40(1):14–33

Rharrabti Y, Villegas D, Royo C, Martos-Núñez V, Garcıa Del Moral L (2003) Durum wheat quality in mediterranean environments: II. influence of climatic variables and relationships between quality parameters. Field Crops Res 80(2):133–140

Rutkoski J, Benson J, Jia Y, Brown-Guedira G, Jannink J, Sorrells M (2012) Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. Plant Genome 5(2):51–61

Rutkoski JE, Poland J, Jannink JL, Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy. G3 3(3):427–439

Sadras V, Roget D, O'Leary G (2002) On-farm assessment of environmental and management factors influencing wheat grain quality in the Mallee. Crop Pasture Sci 53(7):811–820

Sannemann W, Huang B, Mathew B, Léon J (2015) Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept. Mol Breed 35(3):1–16

Schmidt M, Kollers S, Maasberg-Prelle A, Großer J, Schinkel B, Tomerius A, Graner A, Korzun V (2016) Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. Theor Appl Genet 129:203–213. doi:10.1007/s00122-015-2639-1

Sharma D, Anderson W (2004) Small grain screenings in wheat: interactions of cultivars with season, site, and management practices. Crop Pasture Sci 55(7):797–809

Simmonds N (1995) The relation between yield and protein in cereal grain. J Sci Food Agric 67(3):309–315

Soller M, Brody T, Genizi A (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor Appl Genet 47(1):35–39

Somers D, Banks T, DePauw R, Fox S, Clarke J, Pozniak C, McCartney C (2007) Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. Genome 50(6):557–567

Strandén I, Garrick D (2009) Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci 92(6):2971–2975

Sukumaran S, Dreisigacker S, Lopes M, Chavez P, Reynolds M (2015) Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. Theor Appl Genet 128(2):353–363

Taylor J, Butler D (2017) R package ASMap: efficient genetic linkage map construction and diagnosis. J Stat Softw 79(6):1–29. doi:10.18637/jss.v079.i06

Trimble (2016) GreenSeeker crop sensing system. http://www.trimble.com/Agriculture/greenseeker.aspx

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R (2001) Missing value estimation methods for dna microarrays. Bioinformatics 17(6):520–525

Verbyla A, Taylor J, Verbyla K (2012) RWGAIM: an efficient high-dimensional random whole genome average (QTL) interval mapping approach. Genet Res 94(06):291–306

Verbyla AP, Cullis BR, Thompson R (2007) The analysis of QTL by simultaneous use of the of the full linkage map. Theor Appl Genet 116:95–111

Wang S, Wong D, Forrest K, Allen A, Chao S, Huang B, Maccaferri M, Salvi S, Milner S, Cattivelli L, Mastrangelo A, Whan A, Stephen S, Barker G, Wieseke R, Plieske J, International Wheat Genome Sequencing Consortium, Lillemo M, Mather D, Appels R, Dolferus R, Brown-Guedira G, Korol A, Akhunova A, Feuillet C, Salse J, Morgante M, Pozniak C, Luo M, Dvorak J, Morell M, Dubcovsky J, Ganal M, Tuberosa R, Lawley C, Mikoulitch I, Cavanagh C, Edwards K, Hayden M, Akhunov E (2014) Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. Plant Biotechnol J 12(6):787–796. doi:10.1111/pbi.12183

Wrigley C, Rathjen A (1981) Wheat breeding in australia. In: Carr S, Carr S (eds) Plants and Man in Australia. Academic Press, New York, pp 96–135

Wu Y, Bhat P, Close T, Lonardi S (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genet 4(10):e1000212. doi:10.1371/journal.pgen.1000212

Wu Y, Close T, Lonardi S (2011) Accurate construction of consensus genetic maps via integer linear programming. IEEE/ACM Trans Comput Biol Bioinf 8(2):381–394. doi:10.1109/TCBB.2010.35

Xu Y, Crouch J (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48(2):391–407

Zadoks J, Chang T, Konzak C et al (1974) A decimal code for the growth stages of cereals. Weed Res 14(6):415–421

Zanke C, Ling J, Plieske J, Kollers S, Ebmeyer E, Korzun V, Argillier O, Stiewe G, Hinze M, Neumann K et al (2014) Whole genome association mapping of plant height in winter wheat (*Triticum aestivum* L.). PloS one 9(11):e113287

Zeutec (2016) SpectraAlyzer grain. https://goo.gl/tv3hPM

Zhang Z, Ersoz E, Lai C, Todhunter R, Tiwari H, Gore M, Bradbury P, Yu J, Arnett D, Ordovas J et al (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42(4):355–360

Zhao H, Nettleton D, Soller M, Dekkers J (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genet Res 86(01):77–87

Chapter 3

Criteria for the optimal design of a genomic selection training strategy

3.1 Exegetical statement

For plant breeders to most effectively implement GS in their programmes, an understanding on how to manage factors affecting prediction accuracy is required. It is understood (from work in small populations) that larger training set sizes produce higher prediction accuracies, but it is not known if this trend holds true at very large training sets. If it exists, a point of diminishing returns could be exploited to reduce the cost of producing an effective training set. Relatedness between training and predictions sets is understood to affect prediction accuracy, but again the extent to which this is true has not been well characterised. This will influence how much representative material is required in the training set and how diverse it should be, which could have flow on effects for the required marker density. While Chapter 2 addressed the "why" of GS, this chapter shifts focus to the "how" of GS training. It makes use of the dataset presented in Chapter 2 to inform breeders on how to optimally design their training strategy in regard to training set size, genetic relatedness, genetic diversity, predicting across breeding cohorts, and marker density.

Statement of Authorship

Title of Paper	Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	G3 Genes Genomes Genetics DOI 10.1534/g3.118.200311

Principal Author

Name of Principal Author (Candidate)	Adam Norman		
Contribution to the Paper	Collected phenotype data. Involved in formulating the research objective and experimental design. Performed all analysis of phenotypic and genetic data. Wrote and prepared the manuscript. Acted as corresponding author.		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24/8/2018

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Julian Taylor		
Contribution to the Paper	Contributed to statistical methodology. Reviewing and approving the manuscript. PhD co-supervisor of Adam Norman.		
Signature		Date	13/09/2018

Name of Co-Author	James Edwards		
Contribution to the Paper	Involved in formulating the research objective and experimental design. Reviewing and approving the manuscript. PhD co-supervisor of Adam Norman.		
Signature		Date	6/9/2018

Name of Co-Author	Haydn Kuchel		
Contribution to the Paper	Involved in formulating the research objective and experimental design. Reviewing and approving the manuscript. PhD principal supervisor of Adam Norman.		
Signature		Date	10/9/18

GENOMIC SELECTION

Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy

Adam Norman, Julian Taylor, James Edwards, and Haydn Kuchel
School of Agriculture, Food & Wine, University of Adelaide
ORCID ID: 0000-0002-0794-4907 (A.N.)

ABSTRACT Genomic selection applied to plant breeding enables earlier estimates of a line's performance and significant reductions in generation interval. Several factors affecting prediction accuracy should be well understood if breeders are to harness genomic selection to its full potential. We used a panel of 10,375 bread wheat (*Triticum aestivum*) lines genotyped with 18,101 SNP markers to investigate the effect and interaction of training set size, population structure and marker density on genomic prediction accuracy. Through assessing the effect of training set size we showed the rate at which prediction accuracy increases is slower beyond approximately 2,000 lines. The structure of the panel was assessed via principal component analysis and K-means clustering, and its effect on prediction accuracy was examined through a novel cross-validation analysis according to the K-means clusters and breeding cohorts. Here we showed that accuracy can be improved by increasing the diversity within the training set, particularly when relatedness between training and validation sets is low. The breeding cohort analysis revealed that traits with higher selection pressure (lower allelic diversity) can be more accurately predicted by including several previous cohorts in the training set. The effect of marker density and its interaction with population structure was assessed for marker subsets containing between 100 and 17,181 markers. This analysis showed that response to increased marker density is largest when using a diverse training set to predict between poorly related material. These findings represent a significant resource for plant breeders and contribute to the collective knowledge on the optimal structure of calibration panels for genomic prediction.

KEYWORDS
Genomic prediction
Wheat breeding
Marker density
Training set design
GenPred
Shared Data Resources
Genomic Selection

For breeders to make the best use of genomic selection, several factors influencing genomic prediction accuracy should be well understood from empirical breeding germplasm datasets in order to optimize rates of genetic gain. Also, before breeding programs divert finite resources toward the implementation of genomic selection, a number of potentially derailing features of diversity based genetic analysis deserve further attention.

Genomic selection involves estimating a large number of marker effects using a set of training lines, and then using these to predict the value of a separate set of lines (Meuwissen *et al.*, 2001). Three major factors that affect the accuracy at which lines can be predicted are training set size, marker density, and population structure, which have been studied previously in wheat populations up to 8,416 lines in size (Nakaya and Isobe 2012; Crossa *et al.*, 2016; He *et al.*, 2017). Larger training sets were shown to increase prediction accuracy within bi-parental populations by Heffner *et al.* (2011a), where training sets consisted of 24 to 96 lines, and also in multifamily populations by Heffner *et al.* (2011b), where training sets ranged from 96 to 288 lines in size. This result was corroborated by Isidro *et al.* (2015) and Michel *et al.* (2017) where training sets up to 300 lines in size were tested. Training sets consisting of up to 3,052 lines have been used in other studies, but not to directly investigate the effect of training set size. Larger training sets give higher prediction accuracy as increased sample size reduces bias and decreases the variance of marker effect estimates (Liu *et al.*, 2011; de los Campos *et al.*, 2013). Of the studies investigating

Copyright © 2018 Norman *et al.*
doi: https://doi.org/10.1534/g3.118.200311
Manuscript received April 11, 2018; accepted for publication June 29, 2018; published Early Online July 3, 2018.
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Supplemental material available at Figshare: https://figshare.com/s/287c2c7f1623008487a5.
Australian Grain Technologies Pty Ltd, 20 Leitch Rd Roseworthy, South Australia, Australia, 5371, E-mail: adam.norman@adelaide.edu.au

the effect of training set size, none reached the point where further increases in size would not continue to increase prediction accuracy. Here we address this question using uniquely large training sets (n = 8,300). This research therefore provides the most relevant results to large scale breeding programs which typically work with tens of thousands of lines.

Another factor for breeding programs to consider is the required marker density. Prediction accuracy increases with marker density due to more quantitative trait loci (QTL) being in LD with a marker (Heffner *et al.*, 2009; Desta and Ortiz 2014). Solberg *et al.* (2008) showed using simulated data that increasing single nucleotide polymorphism (SNP) density from one to eight SNP per cM resulted in a 25% increase in prediction accuracy. Heffner *et al.* (2011b) used a multifamily wheat dataset to show a 10% increase in prediction accuracy was achieved when moving from 192 to 1,158 markers. However, most of this increase occurred from 192 to 384 markers, indicating that the response to increased marker density would eventually reach a plateau (de los Campos *et al.*, 2013). The point at which this plateau occurs is determined by the genetic diversity within the population, and the relatedness between the training and prediction sets. Hickey *et al.* (2014) showed in a maize simulation study that fewer markers are required when there is high relatedness between training and prediction sets, as they share long haplotype effects and large linkage blocks. The study also found that increasing the size and diversity of the training set was only beneficial when using a large number of markers. Heffner *et al.* (2011a) investigated the response of prediction accuracy to marker density using bi-parental wheat populations, and found a positive response up to 256 markers but a decrease when increasing to 384. As explained by Hickey *et al.* (2014), large numbers of markers can result in the model being overfitted, where non-genetic effects are attributed to the markers. While this improves the model fit, it decreases the accuracy of predicting independent data sets which do not share the non-genetic effects (Jannink *et al.*, 2010). Previous studies have investigated the required marker density in wheat using small empirical datasets of up to 1,158 markers, while other species have been studied using simulated datasets. The current study uses a much larger empirical dataset to extend previous findings into the range where responses can plateau.

Discrete groups of lines with contrasting origin often have differences in allele frequency (population structure) due to selection or parentage (founder effects) (Isidro *et al.*, 2015). This can be problematic as differences in observed phenotypic performance between the two groups may be associated with the markers differing in allele frequency, regardless of whether they are linked to the QTL responsible for the trait variance (Price *et al.*, 2010). The underlying structure of a population is commonly assessed and accounted for using principal component analysis (PCA) of the complete genetic marker set (Patterson *et al.*, 2006; Bentley *et al.*, 2014; Daetwyler *et al.*, 2014). This is an effective method for identifying and visualizing the genetic structure of diverse germplasm panels.

The extent and nature of genetic structure within and across training and validation sets influences the achievable prediction accuracy, and is therefore of interest to breeders when designing training sets. When the training set contains lines closely related to those being predicted, accuracy is higher due to shared long haplotype effects (Daetwyler *et al.*, 2013). Ben Hassen *et al.* (2018) recently observed this relationship in a small rice germplasm set. However, these large linkage blocks are quickly broken up by recombination events, and so crossing cycles can rapidly decrease prediction accuracy (Hickey *et al.*, 2014). If marker density is adequate, increased diversity in the training set will lead to calibration by linkage disequilibrium where short haplotype effects are exploited; this is more stable over multiple generations of crossing (Hickey *et al.*, 2014). However, distant relationships increase noise and bias in the genomic relationship matrix, which in turn reduces the power of prediction (Lund *et al.*, 2016). This study uses multiple breeding cohorts from a commercial breeding program in a unique cross-validation design to investigate the interaction of these opposing effects in an applied scenario, which will inform breeders on optimal training set design.

In this research we study the optimal design of genomic selection training sets by using a panel of 10,375 wheat lines to investigate the effect that training set size, marker density, and genetic structure have on genomic prediction accuracy. We also examine the interaction between marker density and population structure.

MATERIALS AND METHODS

Plant material and associated data

This study utilizes an association panel of 10,375 bread wheat lines, sourced from preliminary and advanced yield testing programs of Australian Grain Technologies Pty Ltd (AGT). The panel was phenotyped in a dedicated field trial at Roseworthy, South Australia (-34.52, 138.69) in the 2014 growing season. We studied data from a single site in order to remove the potentially confounding effect of genotype by environment interaction (GxE). As described in Norman *et al.* (2017), the trial was sown as a non-replicated randomized design with repeated grid checks (1 check per 11 plots), as the large number of lines made a replicated trial logistically infeasible. Dimensions of the trial were 476 rows by 24 ranges, and plot size was $3m^2$. The trial was managed according to best local practice which included fertilizer applications to maximize grain yield and grain quality, and fungicide applications to control disease. Grain yield was measured with a machine harvester and thousand kernel weight (TKW) through image analysis. Both glaucousness and relative maturity were assessed visually, glaucousness on a 1-9 scale (1 = low expression) and relative maturity using the Zadoks scale (Zadoks *et al.*, 1974). These four traits were selected for the current study as they display sufficient phenotypic variation, represent varying levels of genetic control, and experience different selection pressure in a breeding program. Glaucousness has simple genetic control (Bennett *et al.*, 2012a; Norman *et al.*, 2017) and was not actively selected for in this breeding program. Maturity is predominantly controlled by several large effect genes (Snape *et al.*, 2001; Cane *et al.*, 2013) and is selected for mid range performance suitable for the Australian environment. TKW is quantitative (Huang *et al.*, 2006; Sun *et al.*, 2009; Bennett *et al.*, 2012b), and lines are heavily selected to perform above a threshold. Grain yield is a highly complex trait (Kuchel *et al.*, 2007; Bennett *et al.*, 2012b; Maphosa *et al.*, 2014) and lines are strongly selected to yield as high as possible.

Marker genotyping was performed using a custom Axiom™ Affymetrix array containing 18,101 single nucleotide polymorphism (SNP) markers. Markers with minor allele frequency (MAF) lower than 0.01 were removed. Further details on the development of the genotyping platform and preparation of the marker data are provided in Norman *et al.* (2017).

Statistical modeling

One step genomic prediction model: In this research we followed the statistical modeling approach similar to Norman *et al.* (2017). Initially, the phenotypic data from the full Roseworthy trial as well as the complete genotypic marker data was used to form a one-step genomic prediction linear mixed model. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a vector of trait observations then the linear mixed model had the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{Z}_g\mathbf{g} + \mathbf{e} \tag{1}$$

where $\boldsymbol{\tau}$ was a vector of fixed effects with associated design matrix $\mathbf{X}$, and contains an intercept and coefficients for covariates in $\mathbf{X}$ explaining potential trends or known environmental anomalies across the layout of the trial. Extraneous non-genetic variation due to the experimental design such as blocks were captured using random effects $\mathbf{u}$ with design matrix $\mathbf{Z}$ where the effects were assumed to be distributed $\mathbf{u} \sim N(0, \sigma_u^2\mathbf{I})$. To ensure dependence between trait observations was appropriately modeled, the residual error, $\mathbf{e}$, was assumed to be distributed $\mathbf{e} \sim N(0, \sigma^2\mathbf{R})$ where $\mathbf{R} = \mathbf{R}_r(\rho_r) \otimes \mathbf{R}_c(\rho_c)$ was parameterised as a separable AR1 ⊗ AR1 (AR1 = auto-regressive of order 1) correlation structure in the row and column dimensions of the experimental layout (Gilmour *et al.*, 1997). In (1) the n_g length vector of total genetic effects g were defined by the genetic model

$$\mathbf{g} = \mathbf{a} + \mathbf{p} \tag{2}$$

where **a** and **p** were the additive and residual genetic effects respectively with joint distribution

$$\begin{bmatrix} \mathrm{a} \\ \mathrm{p} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2\mathrm{K} & 0 \\ 0 & \sigma_p^2\mathrm{I} \end{bmatrix}\right)$$

Here, $\mathbf{K} = \mathbf{M}\mathbf{M}^T/s$ where $\mathbf{M}$ is the complete marker matrix and s is a scaling constant defined by $s = \left(\sum_{j=1}^{n_g} d_{jj}\right)/n_g$ where d_{jj} is the jth diagonal element of $\mathbf{M}\mathbf{M}^T$ (Forni *et al.*, 2011). The matrix $\mathbf{K}$ is known as the additive relationship or kinship matrix (VanRaden 2008) and can be viewed as a full rank variance matrix detailing the additive connectivity between the genotyped lines. The constant s ensures the genetic variance parameters σ_a^2 and σ_p^2 are numerically comparable and interpretable.

Parameter estimation in the one-step genomic prediction linear mixed model (1) was achieved through an iterative algorithm. Best linear unbiased estimators (BLUEs) of the fixed effects and best linear unbiased predictions (BLUPs) of the random effects were obtained from solutions to the mixed model equations (MMEs) (Henderson 1953). Estimates of the variance parameters are then obtained through an average information algorithm (Gilmour *et al.*, 1995) implemented through maximizing the residual maximum likelihood (REML) derived in Patterson and Thompson (1971). From these solutions the genomic best linear unbiased predictions (GBLUPs) of the additive genetic effects **a** can be written as

$$\tilde{\mathbf{a}} = \sigma_a^2\mathbf{K}\mathbf{Z}_g^T\mathbf{P}\mathbf{y} \tag{3}$$

where $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}^{-1}$ and $\mathbf{H} = \mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{R} + \sigma_u^2\mathbf{Z}\mathbf{Z}^T + \sigma_a^2\mathbf{Z}_g\mathbf{K}\mathbf{Z}_g^T + \sigma_p^2\mathbf{Z}_g\mathbf{Z}_g^T$. These GBLUPs $\tilde{\mathbf{a}}$ represent the relative genetic merit of the lines and are commonly called estimated breeding values.

Cross validation: For each cross-validation scenario conducted, training data sets were created by setting the validation set records from the phenotypic data to missing and appropriately subsetting the genetic marker data to include training set lines only. A training set model was fitted using an adaptation of the linear mixed model defined in (1) with non-genetic parameters fixed at their estimates from the full model. Marker effects were then predicted using the methods described in Norman *et al.* (2017), namely

$$\tilde{\mathbf{q}}_t = \mathbf{M}_t^{\mathrm{T}}\mathbf{K}_t^{-1}\tilde{\mathbf{a}}_t \tag{4}$$

where $\mathbf{M}_t$ and $\mathbf{K}_t$ were the genetic marker data and additive relationship matrix respectively for the training set of lines and $\tilde{\mathbf{a}}_t$ were GBLUPs for training lines calculated using (3). Genomic predictions for lines in the validation set were then determined using

$$\tilde{\mathbf{a}}_v = \mathbf{M}_v\tilde{\mathbf{q}}_t \tag{5}$$

where $\mathbf{M}_v$ is the genetic marker data for the validation set and $\tilde{\mathbf{q}}_t$ is defined in (4).

For cross-validation scenarios in section 2.6 where the number of markers is reduced below the number of lines used in the training set, the model (1) cannot be used due to rank deficiency in the relationship matrix. Consequently, an alternative formulation was adopted for the genetic effects defined in (2), namely

$$\mathbf{g}_t = \mathbf{M}_t^*\mathbf{q}_t + \mathbf{p}_t \tag{6}$$

where $\mathbf{M}_t^*$ is the genetic marker data with reduced numbers of markers for the training set, $\mathbf{q}_t$ represents a vector of marker effects with assumed distribution $\mathbf{q}_t \sim N(0, \sigma_a^2\mathbf{I})$ and $\mathbf{p}_t$ are the residual genetic effects defined in (2). The iterative estimation algorithm proceeds similarly to the previous section and marker effect predictions for the training set were determined directly using

$$\tilde{\mathbf{q}}_t = \sigma_a^2\mathbf{M}_t^{*T}\mathbf{Z}_g^T\mathbf{P}\mathbf{y}$$

GBLUPs of the additive effects for the validation lines were then immediately determined using an analogous equation to (5), namely $\tilde{\mathbf{a}}_v = \mathbf{M}_v^*\tilde{\mathbf{q}}_t$.

Computations: All statistical analysis was carried out in the R Statistical Computing Environment (R Core Team 2017). Linear mixed models were fitted using the flexible linear mixed modeling package ASReml-R (Butler *et al.*, 2009) available as an R package and downloadable from *www.vsni.co.uk/software/asreml*.

Impact of training set size on prediction accuracy

The effect of training set size on genomic prediction accuracy was assessed through an extended five-fold cross-validation analysis. First, the full panel was randomly divided into five folds each containing 2,075 lines. Four of these folds acted as a training set (8,300 lines) which was used to predict the remaining fold (validation set). The training set was then randomly sampled to sizes of 250, 500, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000 and 7,000, where each acted as a training set to predict that fold's validation set which remained at a fixed size of 2,075 lines. These subsets were sampled without replacement resulting in varying levels of replication for the different sizes. Within each fold there were 33 reps at 250, 16 of 500, 8 of 1,000, 4 of 2,000, 2 of 3,000, 2 of 4,000, and 1 of 5,000 and above. All training models were fitted according to (1) where marker effects were then calculated by (4), and used to form genomic predictions of lines in the validation set according to (5). All training sets within each fold were used to predict the same validation set. Relative prediction accuracies were calculated by correlating the genomic predictions to the corresponding additive GBLUP values from the full data set model. For the remainder of this paper, the term prediction accuracy is used to describe the capacity of the comprised training sets to predict line performance as described by the maximal model.

Impact of population structure on prediction accuracy

To investigate how genomic prediction training sets can be optimally designed, the panel was partitioned using two different approaches for the purposes of training and cross-validation. In the first method, K-means clustering was used to partition based on genetic similarity. This was used as a surrogate for assessing calibration within and between germplasm pool (breeding program). In the second method, the germplasm was partitioned by cross-year to examine the effect on prediction accuracy of including multiple historical 'breeding cohorts' (historical lines/data). Online Resource 1 details which lines belong to each cluster and breeding cohort.

Impact of underlying population structure: K-means clustering was performed on a marker based genetic dissimilarity matrix using the K-means functionality inside the R statistical computing environment (R Core Team 2017). The sum of squares within clusters was assessed when setting the number of clusters between 2 and 50, which showed the variance plateaus when there were more than five. The number of clusters was therefore set at five. In order to achieve clusters of equal size, 1,500 lines were randomly selected from each to be used in the cross-validation analysis. With these 7,500 lines, four cross-validation designs (detailed in Figure 1) were then used to achieve i) equal representation of all clusters in both the training and validation sets ('all clusters'), ii) representing the same cluster in both the training and validation set ('within cluster'), iii) representing one cluster in the training set and one different cluster in the validation set ('between cluster - narrow training'), and iv) representing four clusters in the training set and the remaining one cluster in the validation set ('between cluster - broad training'). Within each cluster, lines were randomly sampled without replacement to produce subsets. This allowed all training sets in each of the four designs to contain 1,000 lines, and all validation sets to contain 500 lines. In each design these subsets were rotated to all possible combinations in order to provide replication. All training models were fitted according to (1) where marker effects were then calculated by (4), and used to form genomic predictions of lines in the validation set according to (5). Prediction accuracies were calculated by correlating the genomic predictions to the corresponding additive GBLUP values from the full data set model.

Impact of breeding cohort: Here, lines from four different breeding cohorts were selected from the PYT-South subset of breeding lines. The cohorts were randomly selected from the second yield testing stage of the south breeding program from years 2010 to 2013, and each cohort contained 996 lines. Three cross-validation designs were used to assess i) one cohort year (training set) used to predict the following cohort year (validation set), ii) two cohort years (training set) used to

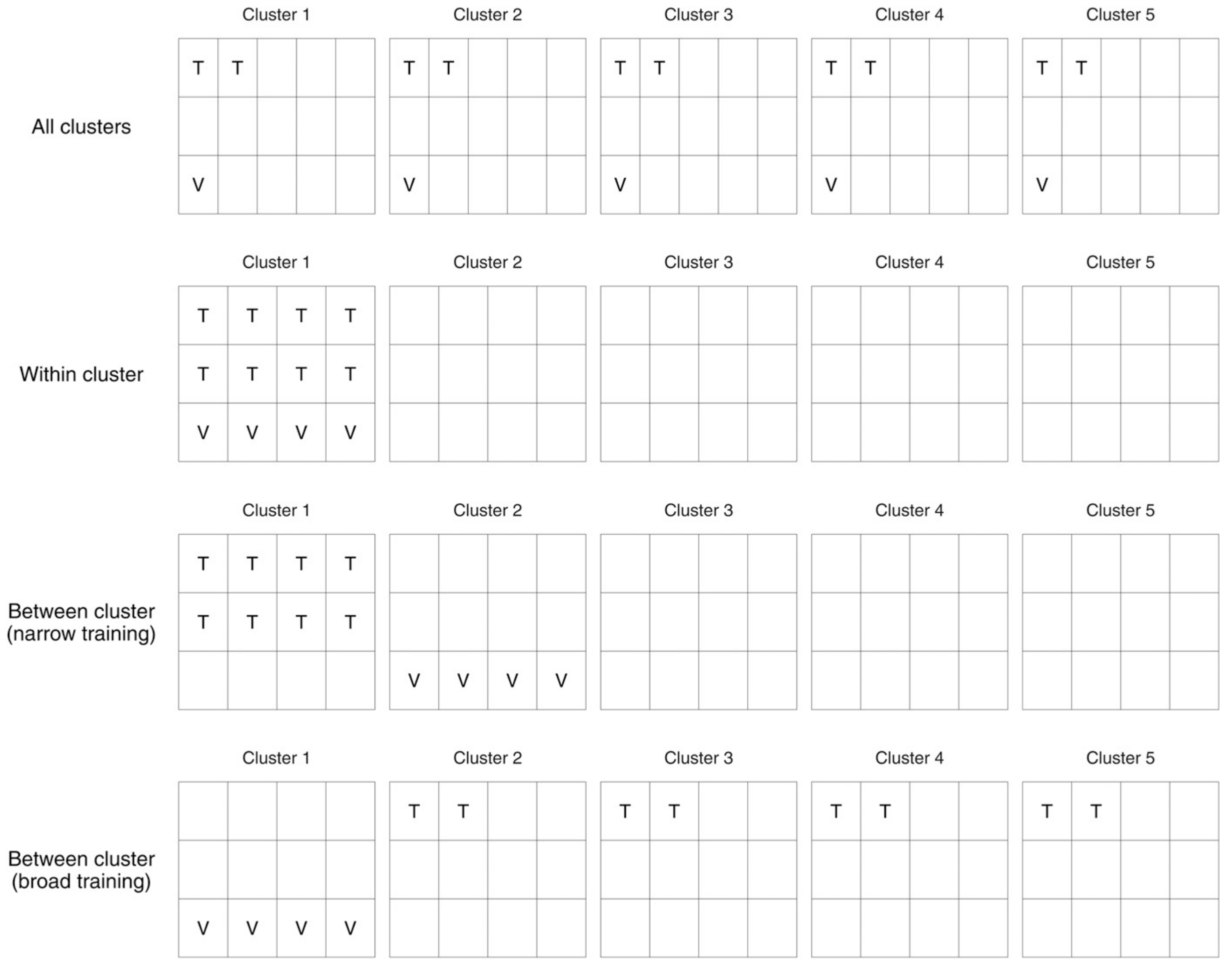

Figure 1 Description of the four cross-validation designs used to assess the impact of underlying population structure. The partitions within each cluster were formed by randomly sampling without replacement. Replication was achieved by rotating partitions within each design to provide all combinations of partitions and clusters. All designs had consistent training and validation set sizes of 1,000 and 500 respectively.

predict the following cohort year (validation set), and iii) three cohort years (training set) used to predict the following cohort year (validation set). As in the K-means clustering method, lines were randomly sampled without replacement within each cohort year to produce subsets. This allowed all training and validation sets in each of the three designs to contain 996 lines. In design ii) the training sets were made up of one 498 line subset from each of the two cohort years, and in design iii) they consisted of one 332 line subset from each of the three cohort years. Cross-validations were performed according to the same methods used in the K-means clustering method.

Marker density analysis

Marker subsets of varying size (100, 500, 1,000, 3,000, 5,000, 10,000, 13,639 and 17,181) were selected in order to assess the effect of marker density on prediction accuracy, and its interaction with population structure. The 13,639 markers on the consensus map from Norman *et al.* (2017) were selected as the first subset, from which markers for the lower densities were selected with the criteria of being evenly distributed on the genome, as well as having high minor allele frequency (MAF). To achieve this, markers were first allocated into linkage map bins of varying size for each target density, and those with the highest MAF within each bin were selected. Table 1 summarizes each marker subset and genetic maps of each subset are plotted in Online Resource 2. Online Resource 3 details which markers belonged to each subset.

Random five-fold cross validation: The effect of marker density on prediction accuracy was assessed with random five-fold cross validation, where training sets consisted of 8,300 lines and validation sets 2,075 lines. The cross validation was repeated for each marker density. Training models for marker densities containing fewer markers than lines were fitted according to (6) where marker effects were determined directly. For densities containing more markers than lines, training models were fitted according to (1) and marker effects were then calculated through (4). Marker effects from either method were then used to formulate genomic predictions of lines in the training set according to (5), and prediction accuracies were calculated by correlating the predictions to additive GBLUP values from the full model.

K-means clustering: The response of prediction accuracy to marker density was assessed in different population structures by repeating the K-means clustering method for each marker density. As in section 2.6.1, training models for densities containing fewer markers than lines were fitted according to (6), and those containing more markers than lines were fitted according to (1). Genomic predictions were calculated according to (5), and correlated to GBLUP values from the full model to determine prediction accuracy.

Data availability

File S1 specifies the breeding cohorts used for analysis. File S2 contains genetic map plots of each marker subset. File S3 specifies which markers were included in each subset, and the genetic map position of each marker. File S4 contains all genetic marker data, and file S5 contains all phenotype data. Supplemental material available at Figshare: https://figshare.com/s/287c2c7f1623008487a5.

RESULTS

Impact of training set size on prediction accuracy

Figure 2 details the effect of training set size on genomic prediction accuracy for the four traits analyzed. A similar trend was observed at each trait with accuracy increasing substantially from training set size of 250 to 2,000. A correlation with the maximal model of 0.95 was achieved with training set sizes of between 3,950 and 7,650 (for traits glaucousness and relative maturity respectively). Glaucousness was the most accurate trait at all sizes, and maturity the least. The difference in accuracy between traits was more pronounced at smaller training set sizes (0.59 to 0.79 at size 250, 0.96 to 0.98 at size 8,300). Grain yield showed the most variation between replications of each training set size (indicated by the shading of upper and lower quartiles), and glaucousness the least.

■ **Table 1 Summary of the marker selections using the consensus map**

Number of markers	Unique map positions	Markers per map position	Mean interval[a]	Mean MAF[b]
100	100	1.00	31.2	0.49
500	500	1.00	6.25	0.44
1000	1000	1.00	3.12	0.40
3000	3000	1.00	1.04	0.34
5000	4580	1.09	0.68	0.32
10000	4590	2.18	0.68	0.29
13639	4593	2.97	0.68	0.26

[a] Mean interval (cM) between unique map positions.
[b] Mean minor allele frequency across the full panel.

Impact of population structure on prediction accuracy

Figure 3 details the structure of lines included in each of the population structure analyses. Sub-plots **A** and **B** display components one and two, and one and three respectively from a PCA performed on the lines included in the K-means cluster analysis. Sub-plots **C** and **D** represent similar plots from a PCA performed on the lines included in the breeding cohort analysis, where lines are colored according to their cohort year. There is a clear distinction between the K-means clusters, while the genetic dissimilarity between the breeding cohorts is less pronounced.

Impact of underlying population structure: Figure 4 summarizes prediction accuracies from the K-means clustering method of assessing population structure impacts on prediction accuracy. 'All clusters' and 'within cluster' accuracies were similarly high for glaucousness and grain size, whereas for grain yield 'all clusters' was slightly higher and for relative maturity slightly lower. For all traits, predicting between cluster with a broad training set was more accurate than predicting between cluster with a narrow training set, but both were significantly less accurate than 'all clusters' and 'within cluster'.

Impact of breeding cohort: Figure 5 presents prediction accuracies from the breeding cohort method of assessing the impact of population structure on prediction accuracy. This shows that as more cohort years were represented in the training set, prediction accuracy increased significantly for grain yield, and slightly for relative maturity. Glaucousness and TKW however, had relatively stable prediction accuracy regardless of how many cohort years were represented in the training set. Prediction accuracies were highest for TKW and glaucousness, with relatively maturity being slightly lower and grain yield lower again.

Impact of marker density on prediction accuracy

Table 1 summarizes each marker selection using the consensus map to calculate unique positions, markers per map position and mean

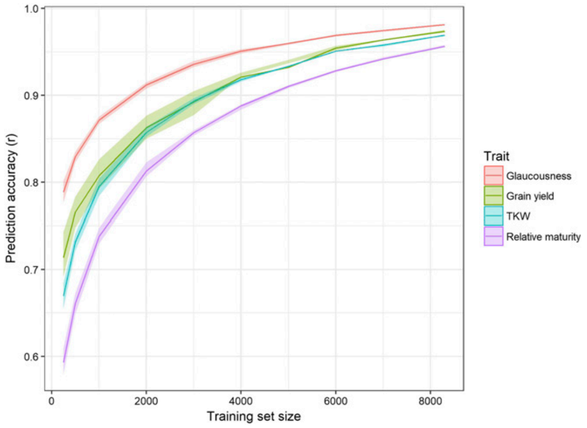

Figure 2 Genomic prediction accuracies from five-fold random cross-validation with varying training set sizes. Shading represents upper and lower quartiles. Prediction accuracy is defined as the correlation between genomic predictions of the validation set and their corresponding additive GBLUP values from the maximal model. TKW represents thousand kernel weight.

interval. This shows only a slight increase in the number of map positions at selections containing more than 5,000 markers. Therefore, at the selections with more than 5,000 markers, the mean position interval plateaus off and markers per map position increases. The mean MAF of the markers at each selection starts very high at 0.49 for the 100 marker selection, and steadily decreases to 0.26 for the 13,639 selection.

The effect of marker density on prediction accuracy was assessed in the first instance through random five-fold cross validation, the results of which are summarized in Figure 6. All four traits showed a sharp increase in accuracy before reaching a plateau at approximately 5,000 markers, with only a marginal increase in prediction accuracy when increasing from 5,000 to 17,181 markers. All traits showed the highest prediction accuracy when all available markers were used. Glaucousness, relative maturity and grain yield all had similar response curves, but TKW had a more pronounced increase in accuracy with marker number, particularly when increasing from 1,000 to 3,000 markers.

Effect of interaction Between marker density and population structure on prediction accuracy

The K-means clustering analysis was repeated for each marker density in order to investigate the interaction between population structure and marker density (Figure 7). Similar to the five-fold cross validation analysis, prediction accuracies increased sharply up to approximately 3,000 markers before plateauing. Similar responses were observed for 'all clusters' and 'within cluster' prediction structures across all traits. Between cluster prediction saw greater response to increased marker density, particularly with broad training when increasing from 100 to 1,000 markers. Relative maturity saw a slight decrease in prediction accuracy when marker number was increased beyond 5,000.

DISCUSSION

If plant breeders are to effectively apply genomic selection in their breeding programs, they require a sound understanding of factors affecting prediction accuracy in large scale germplasm datasets. In the present study we utilized a panel of 10,375 lines sourced from an active breeding program to investigate the effect and interaction of

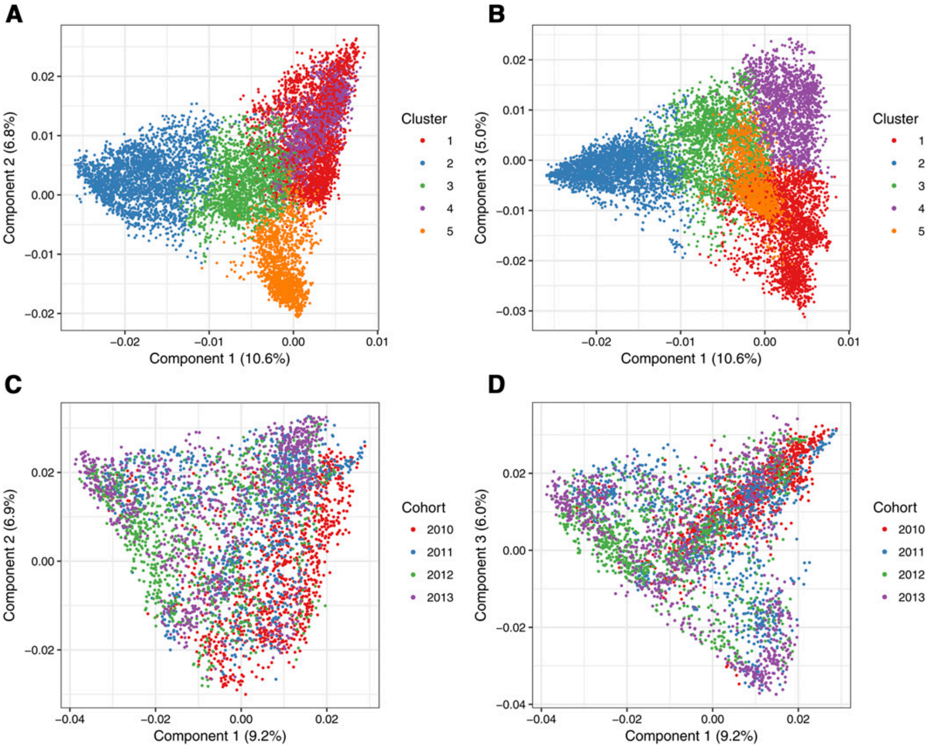

Figure 3 Pairwise plots of components from two principal component analyses (PCA). A First and second components of the PCA performed on lines included in the K-means clustering method, with lines colored according to which cluster they belonged. B First and third components of the PCA performed on lines included in the K-means clustering method, with lines colored according to which cluster they belonged. C First and second components of the PCA performed on lines included in the breeding cohort method, with lines colored according to which cohort they belonged. D First and third components of the PCA performed on lines included in the breeding cohort method, with lines colored according to which cohort they belonged.

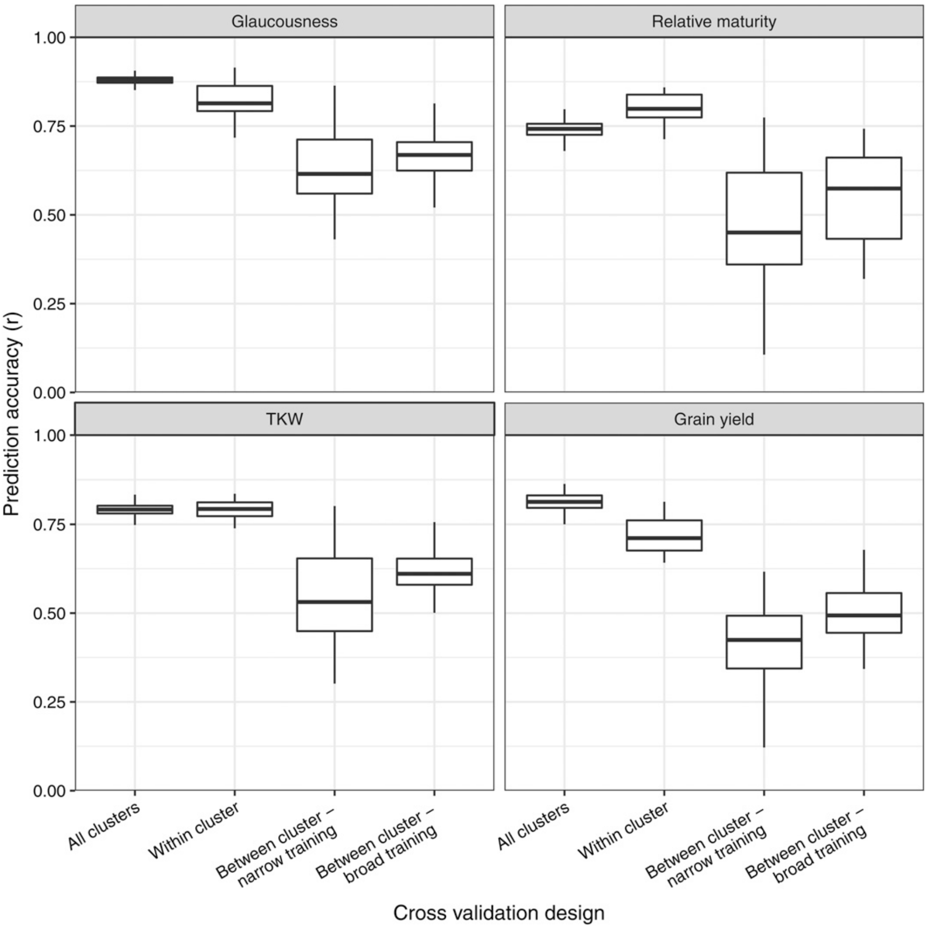

Figure 4 Boxplots showing prediction accuracies from the K-means clustering method for each category of training and validation set combinations, detailed in section 2.5.1. Prediction accuracy was calculated by correlating predictions of the validation set to the corresponding additive GBLUP values from the full model with all lines included. TKW represents thousand kernel weight.

training set size, population structure and marker density on prediction accuracy. The findings presented here will assist breeders in optimizing their programs, allowing them to make the most effective and efficient use of their resources when implementing genomic selection.

Effect of training set size on prediction accuracy

An important factor influencing genomic prediction accuracy is the size of the training set used to develop the prediction calibration (Nakaya and Isobe 2012). However, research questions pertaining to this have previously proven difficult to address, as the large number of lines required to locate the point of diminishing returns has an often prohibitively high cost of genotyping. The data set analyzed here provides a unique opportunity to investigate the effect of population size on prediction accuracy in bread wheat. Prediction accuracy increased substantially when the training set size was increased from 250 lines to approximately 2,000, after which the rate of increase slowed. While an acceptable prediction accuracy would be determined by the breeder on a case by case basis, if we look at an accuracy of 0.95 as an example, this is achieved at a training set size of 3,930 and 7,450 for glaucousness and relative maturity respectively. This result confirms previous findings from smaller populations (Heffner *et al.*, 2011a, b; Isidro *et al.*, 2015), and extends the relationship to larger training sets showing there is a point at which accuracy begins to plateau in response to increased training set size. Plant breeders should take this result into account when weighing up the benefit of including additional lines in a training set. While there were differences between traits in the level of accuracy achieved, the trend in response to training set size was consistent for all traits despite their differences in genetic complexity. This suggests that response in prediction accuracy to training set size is not dependent on the complexity and genetic architecture of the trait.

The difference in prediction accuracy between traits was more pronounced at smaller training set sizes. This was also driven by the genetic complexity of the trait, as more lines are needed to provide the high number of allelic observations required to accurately predict small effect QTL (Gilmour 2007). Prediction accuracies in this analysis varied more within the smaller training set sizes than the large, particularly for grain yield. This indicates population structure was present and the variation in accuracy was likely caused by the presence or absence of highly related lines across training and validation sets. (Poland *et al.*, 2012). In the next section we investigate how the relatedness between training and validation sets affects the resultant accuracy.

Effect of population structure on prediction accuracy

K-means clustering produced five genetically distinct clusters, which is demonstrated in Figure 3. Prediction accuracy within and between

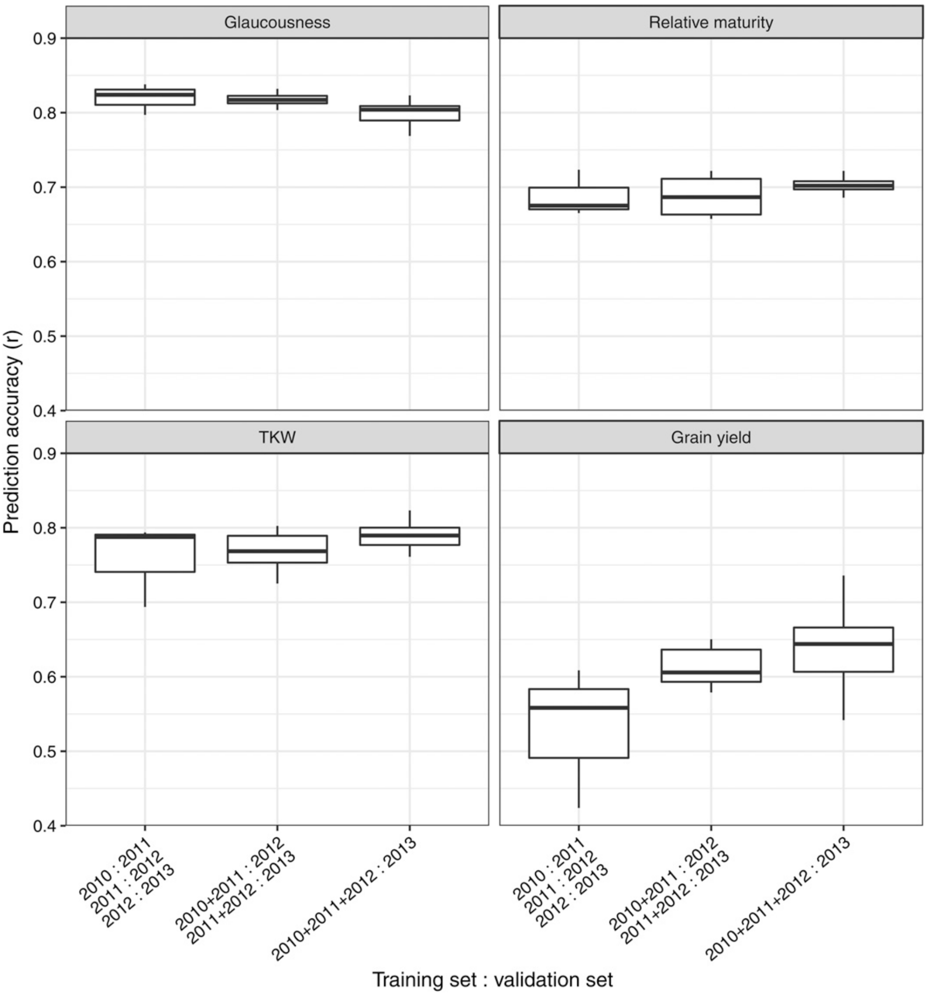

Figure 5 Boxplots summarizing the prediction accuracies from the breeding cohort method, as detailed in section 2.5.2. At each trait the first boxplot represents one cohort year used as a training set to predict the subsequent cohort year (validation set). The second represents two consecutive cohort years used as training to predict the subsequent cohort year, and third represents three consecutive cohort years used to predict the subsequent.

clusters was tested using structured cross-validation with training sets containing 1,000 lines and validation sets containing 500 lines. The breeding cohorts were less distinct as they were all sourced from the southern breeding program. The accuracy of predicting one cohort using a training set sourced from one, two or three prior cohorts was tested using training and validation sets of the same size as those in the K-means clustering method. This unique assessment is representative of how genomic prediction would be applied in a commercial breeding program.

In the K-means cluster method, 'all clusters' and 'within cluster' prediction accuracies were similar for glaucousness and TKW. The training sets of both prediction structures directly represent the clusters in their respective validation set, the only difference being that 'all clusters' uses all five clusters whereas 'within cluster' uses just one. This result therefore suggests the broadness of the training and validation sets has little effect on prediction accuracy when the training set contains at least some lines that are highly representative of those being predicted. For relative maturity however, 'within cluster' prediction accuracy was slightly higher than 'all clusters', and the reverse was observed for grain yield. There are several large effect photoperiod and vernalisation genes that control maturity (Snape *et al.*, 2001; Cane *et al.*, 2013), and the predominating genes differ between clusters (data not shown). The higher accuracy when predicting maturity within cluster was therefore likely to be caused by the key large effect genes having greater representation in the training set. For grain yield on the other hand, the increased diversity was beneficial as 'all clusters' showed higher prediction accuracy than 'within cluster'. This is because there was more and comparable phenotypic diversity represented within both the training and validation sets for 'all clusters'.

For all traits, predicting a single cluster using a broad training set produced higher accuracies than predicting with narrow training, but was substantially less accurate than 'within cluster' and 'all clusters'. This shows that prediction accuracy is significantly higher when the training set contains close relatives of lines in the validation set, but accuracy can also be increased by including more genetic diversity in the training set. Breeders should therefore design genetically diverse training sets that are highly related to the prediction set in order to maximize genetic response to genomic selection. This is corroborated by the results of the breeding cohort cross-validation, where prediction accuracy was improved for grain yield and relative maturity by including more cohort years (and therefore more diversity) in the training set. With increased

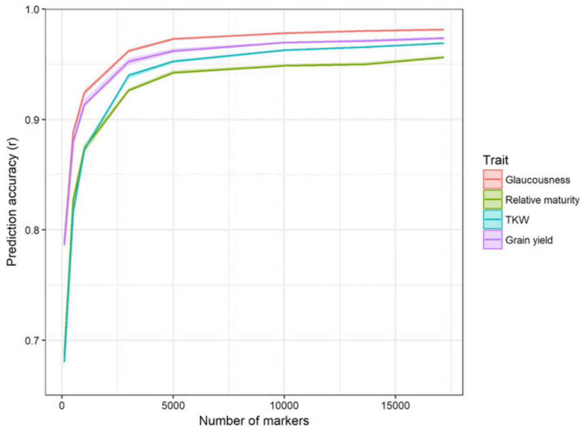

Figure 6 Plot showing the effect of marker density on prediction accuracy for each trait. Prediction accuracy was assessed by performing random five-fold cross-validation for each selection of markers, and correlating predictions of the validation set to the corresponding additive GBLUP values from the full model with all lines included. Marker subsets were selected to be evenly distributed over the genome and to have high minor allele frequency.

genetic diversity and high SNP density, the training set can better capture short haplotype effects that are relevant to the validation set. This type of calibration is based on short haplotype effects and linkage disequilibrium information, and is suggested by Hickey *et al.* (2014) to be less susceptible to breaking down after multiple breeding cycles.

The breeding cohort analysis is the most representative of how genomic selection would be applied in a breeding program, predicting the current cohort using previous cohorts. The increase in prediction accuracy with more cohorts in the training set was most pronounced for grain yield, and supports previous findings in rye (Auinger *et al.*, 2016). Muir (2007) observed through simulation of animal breeding that continued selection over multiple generations eventually reduced prediction accuracy. The difference between that study and the present is the longer generation intervals of wheat breeding and consequently the fewer number of generations represented. The results presented here show that incorporating more breeding cohorts in the training set is beneficial in a conventional breeding program with a long generation interval. A recent study by Gorjanc *et al.* (2018) investigates the response in a rapid cycling program which uses genomic selection to quickly identify parents.

While grain yield undergoes continual and intense selection within the breeding program, relative maturity and TKW are threshold traits and therefore change less over time, which results in them benefiting less from the inclusion of additional cohort years in the training set. Glaucousness undergoes no direct selection meaning genetic change will only occur through correlated response, and it therefore sees little benefit from adding more cohort years to the training set.

Marker density

The effect of marker density on prediction accuracy was assessed with a random five-fold cross validation analysis performed with various marker densities. All traits experienced a strong response to increases in marker density up to 5,000 markers, showing that this was sufficient for generating a relatively accurate prediction calibration within this panel. This number is significantly higher than the plateau point of previous

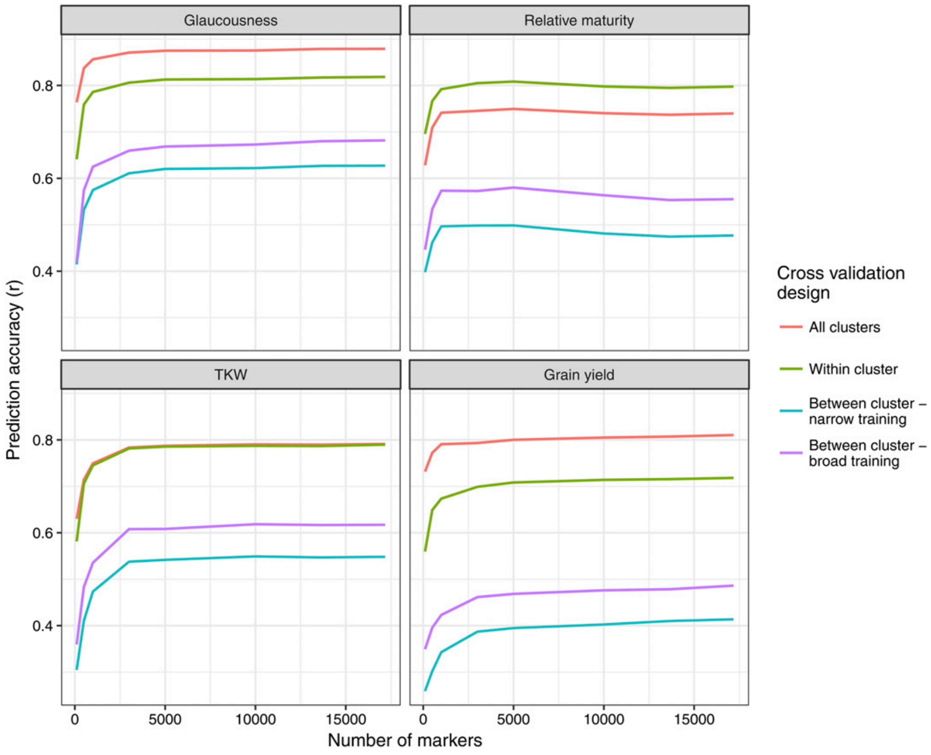

Figure 7 Plots showing the interactive response of prediction accuracy to marker density and population structure. The K-means clustering method detailed in section 2.5.1 was repeated for each selection of markers. Marker subsets were selected to be evenly distributed over the genome and to have high minor allele frequency.

studies in smaller populations (Heffner *et al.*, 2011b), as high marker densities only facilitate finer resolution and more accurate estimates of QTL effects when combined with large population size and low linkage disequilibrium (Huang *et al.*, 2012). TKW benefited from increased marker density more than the other traits, which could be explained by its quantitative genetic nature requiring more markers to accurately estimate its many small QTL effects (Zhang *et al.*, 2015). However, grain yield is also a highly quantitative trait and it saw a similar response curve to the more qualitative traits glaucousness and relative maturity.

The interactive effect of marker density and population structure on prediction accuracy was assessed by repeating the K-means cluster analysis with various marker densities. The density at which prediction accuracy plateaued was slightly lower than that observed in the random five-fold cross validation. This is consistent with previous studies using smaller data sets where additional markers benefited prediction accuracy more when larger training sets were used (Heffner *et al.*, 2011a,b). Prediction accuracy responded more to increased marker density when predicting between clusters, particularly when there was more genetic diversity in the training set. This is consistent with the findings of Hickey *et al.* (2014), where in a simulated maize data set the required marker density was lower when closely related material was shared between training and validation sets. The study also showed there was greater response to increased marker density when the training set contained more diversity, which corroborates our findings. The slight decrease in prediction accuracy at high marker densities for relative maturity is likely due to excess markers overfitting the model (Heslot *et al.*, 2012). A similar result was seen in Heffner *et al.* (2011a), where higher marker densities resulted in lower prediction accuracy in bi-parental wheat populations.

Conclusions

Here we used a wheat panel of unprecedented size to investigate several key factors affecting genomic prediction accuracy that previously have not been explored at this scale. We showed there is a point at which prediction accuracy begins to plateau in response to training set size, and that this response is independent from the genetic complexity of the trait. The population structure analyses showed that relatedness between training and validation sets has a large effect on prediction accuracy, but importantly when relatedness is low, as is often the case when applying genomic selection, prediction accuracy can be increased by increasing diversity in the training set. We also found that traits under higher selection pressure can be more accurately predicted by including several previous breeding cohorts in the training set. This was shown for up to three previous cohorts, but further work should be done to explore how stable this trend is across different breeding programs and more cohorts. By assessing the interaction between marker density and population structure, we showed the response to increased marker density is larger when using a diverse training set and predicting from poorly related training sets. The work presented herein provides a framework for pragmatic plant breeders to optimally design their genomic selection training strategy to achieve high selection accuracy and subsequent rates of genetic gain.

ACKNOWLEDGMENTS

The authors thank Brian Cullis and Alison Smith for their guidance and substantial work in establishing the project, and Emi Tanaka for her work on the project. We also thank Australian Grain Technologies for their contribution in managing the Roseworthy field experiment and funding the research. We acknowledge funding from the Australian Research Council, and the Grains Research and Development Corporation.

LITERATURE CITED

Auinger, H., M. Schönleben, C. Lehermeier, M. Schmidt, V. Korzun *et al.*, 2016 Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale L.*). Theor. Appl. Genet. 129: 2043–2053. https://doi.org/10.1007/s00122-016-2756-5

Ben Hassen, M., T. Cao, J. Bartholomé, G. Orasen, C. Colombi *et al.*, 2018 Rice diversity panel provides accurate genomic predictions for complex traits in the progenies of biparental crosses involving members of the panel. Theor. Appl. Genet. 131: 417–435. https://doi.org/10.1007/s00122-017-3011-4

Bennett, D., A. Izanloo, J. Edwards, H. Kuchel, K. Chalmers *et al.*, 2012a Identification of novel quantitative trait loci for days to ear emergence and flag leaf glaucousness in a bread wheat (*Triticum aestivum L.*) population adapted to southern Australian conditions. Theor. Appl. Genet. 124: 697–711. https://doi.org/10.1007/s00122-011-1740-3

Bennett, D., A. Izanloo, M. Reynolds, H. Kuchel, P. Langridge *et al.*, 2012b Genetic dissection of grain yield and physical grain quality in bread wheat (*Triticum aestivum L.*) under water-limited environments. Theor. Appl. Genet. 125: 255–271. https://doi.org/10.1007/s00122-012-1831-9

Bentley, A., M. Scutari, N. Gosman, S. Faure, F. Bedford *et al.*, 2014 Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. Theor. Appl. Genet. 127: 2619–2633. https://doi.org/10.1007/s00122-014-2403-y

Butler, D., B. Cullis, A. Gilmour, and B. Gogel, 2009 *ASReml-R reference manual*, Queensland Department of Primary Industries, Queensland, Australia.

Cane, K., H. Eagles, D. Laurie, B. Trevaskis, N. Vallance *et al.*, 2013 *Ppd-B1* and *Ppd-D1* and their effects in southern Australian wheat. Crop Pasture Sci. 64: 100–114. https://doi.org/10.1071/CP13086

Crossa, J., D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño *et al.*, 2016 Genomic prediction of gene bank wheat landraces. G3: Genes, Genomes. Genetics 6: 1819–1834.

Daetwyler, H., M. Calus, R. Pong-Wong, G. de los Campos, and J. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193: 347–365. https://doi.org/10.1534/genetics.112.147983

Daetwyler, H., U. Bansal, H. Bariana, M. Hayden, and B. Hayes, 2014 Genomic prediction for rust resistance in diverse wheat landraces. Theor. Appl. Genet. 127: 1795–1803. https://doi.org/10.1007/s00122-014-2341-8

de los Campos, G., J. Hickey, R. Pong-Wong, H. Daetwyler, and M. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–345. https://doi.org/10.1534/genetics.112.143313

Desta, Z., and R. Ortiz, 2014 Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 19: 592–601. https://doi.org/10.1016/j.tplants.2014.05.006

Forni, S., I. Aguilar, and I. Misztal, 2011 Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43: 1. https://doi.org/10.1186/1297-9686-43-1

Gilmour, A., 2007 Mixed model regression mapping for QTL detection in experimental crosses. Comput. Stat. Data Anal. 51: 3749–3764. https://doi.org/10.1016/j.csda.2006.12.031

Gilmour, A., R. Thompson, and B. Cullis, 1995 Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51: 1440–1450. https://doi.org/10.2307/2533274

Gilmour, A., B. Cullis, and A. Verbyla, 1997 Accounting for natural and extraneous variation in the analysis of field experiments. J. Agric. Biol. Environ. Stat. 2: 269–293. https://doi.org/10.2307/1400446

Gorjanc, G., R. C. Gaynor, and J. M. Hickey, 2018 Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theor. Appl. Genet.* https://doi.org/10.1007/s00122–018–3125–3

He, S., J. Reif, V. Korzun, R. Bothe, E. Ebmeyer *et al.*, 2017 Genome-wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to central europe. Theor. Appl. Genet. 130: 635–647. https://doi.org/10.1007/s00122-016-2840-x

Heffner, E., M. Sorrells, and J. Jannink, 2009 Genomic selection for crop improvement. Crop Sci. 49: 1–12. https://doi.org/10.2135/cropsci2008.08.0512

Heffner, E., J. Jannink, H. Iwata, E. Souza, and M. Sorrells, 2011a Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci. 51: 2597–2606. https://doi.org/10.2135/cropsci2011.05.0253

Heffner, E., J. Jannink, and M. Sorrells, 2011b Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4: 65–75. https://doi.org/10.3835/plantgenome2010.12.0029

Henderson, C. R., 1953 Estimation of variance and covariance components. Biometrics 9: 226–252. https://doi.org/10.2307/3001853

Heslot, N., H. Yang, M. Sorrells, and J. Jannink, 2012 Genomic selection in plant breeding: a comparison of models. Crop Sci. 52: 146–160. https://doi.org/10.2135/cropsci2011.06.0297

Hickey, J., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu *et al.*, 2014 Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 54: 1476–1488. https://doi.org/10.2135/cropsci2013.03.0195

Huang, B., A. George, K. Forrest, A. Kilian, M. Hayden *et al.*, 2012 A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol. J. 10: 826–839. https://doi.org/10.1111/j.1467-7652.2012.00702.x

Huang, X., S. Cloutier, L. Lycar, N. Radovanovic, D. Humphreys *et al.*, 2006 Molecular detection of QTLs for agronomic and quality traits in a doubled haploid population derived from two Canadian wheats (*Triticum aestivum L.*). Theor. Appl. Genet. 113: 753–766. https://doi.org/10.1007/s00122-006-0346-7

Isidro, J., J. Jannink, D. Akdemir, J. Poland, N. Heslot *et al.*, 2015 Training set optimization under population structure in genomic selection. Theor. Appl. Genet. 128: 145–158. https://doi.org/10.1007/s00122-014-2418-4

Jannink, J., A. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9: 166–177. https://doi.org/10.1093/bfgp/elq001

Kuchel, H., K. Williams, P. Langridge, H. Eagles, and S. Jefferies, 2007 Genetic dissection of grain yield in bread wheat. I. QTL analysis. Theor. Appl. Genet. 115: 1029–1041. https://doi.org/10.1007/s00122-007-0629-7

Liu, Z., F. Seefried, F. Reinhardt, S. Rensing, G. Thaller *et al.*, 2011 Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genet. Sel. Evol. 43: 19. https://doi.org/10.1186/1297-9686-43-19

Lund M, Van den Berg I, Ma P, Brøndum R, Su G (2016) How to improve genomic predictions in small dairy cattle populations. Animal 10(6):1042–1049

Maphosa, L., P. Langridge, H. Taylor, B. Parent, L. Emebiri *et al.*, 2014 Genetic control of grain yield and grain physical characteristics in a bread wheat population grown under a range of environmental conditions. Theor. Appl. Genet. 127: 1607–1624. https://doi.org/10.1007/s00122-014-2322-y

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

Michel, S., C. Ametz, H. Gungor, B. Akgöl, D. Epure *et al.*, 2017 Genomic assisted selection for enhancing line breeding: merging genomic and phenotypic selection in winter wheat breeding programs with preliminary yield trials. Theor. Appl. Genet. 130: 363–376. https://doi.org/10.1007/s00122-016-2818-8

Muir, W., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124: 342–355. https://doi.org/10.1111/j.1439-0388.2007.00700.x

Nakaya, A., and S. Isobe, 2012 Will genomic selection be a practical method for plant breeding? Ann. Bot. (Lond.) 110: 1303–1316. https://doi.org/10.1093/aob/mcs109

Norman, A., J. Taylor, E. Tanaka, P. Telfer, J. Edwards *et al.*, 2017 Increased genomic prediction accuracy in wheat breeding using a large Australian panel. Theor. Appl. Genet. 130: 2543–2555. https://doi.org/10.1007/s00122-017-2975-4

Patterson, H., and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554. https://doi.org/10.1093/biomet/58.3.545

Patterson, N., A. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2: e190. https://doi.org/10.1371/journal.pgen.0020190

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu *et al.*, 2012 Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5: 103–113. https://doi.org/10.3835/plantgenome2012.06.0006

Price, A., N. Zaitlen, D. Reich, and N. Patterson, 2010 New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11: 459–463. https://doi.org/10.1038/nrg2813

R Core Team, (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Snape, J., R. Sarma, S. Quarrie, L. Fish, G. Galiba *et al.*, 2001 Mapping genes for flowering time and frost tolerance in cereals using precise genetic stocks. Euphytica 120: 309–315. https://doi.org/10.1023/A:1017541505152

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. Meuwissen 2008 Genomic selection using different marker types and densities. J. Anim. Sci. 86: 2447–2454. https://doi.org/10.2527/jas.2007-0010

Sun, X., K. Wu, Y. Zhao, F. Kong, G. Han *et al.*, 2009 QTL analysis of kernel shape and weight using recombinant inbred lines in wheat. Euphytica 165: 615–624. https://doi.org/10.1007/s10681-008-9794-2

VanRaden, P., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423. https://doi.org/10.3168/jds.2007-0980

Zadoks, J., T. Chang, and C. Konzak 1974 A decimal code for the growth stages of cereals. Weed Res. 14: 415–421. https://doi.org/10.1111/j.1365-3180.1974.tb01084.x

Zhang, Z., M. Erbe, J. He, U. Ober, N. Gao *et al.*, 2015 Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3: Genes, Genomes. Genetics 5: 615–627.

Communicating editor: J.-L. Jannink

Chapter 4

Increasing response to genomic selection through concurrent use of low and high density genotyping platforms

4.1 Exegetical statement

Implementing GS in a large scale breeding programme comes with a considerable genotyping cost, which is often prohibitively high. Even in a case where the programme can afford the cost, more genetic gain could be made if genotyping were cheaper as it would enable selection within larger breeding populations. Low density genotyping platforms are available at prices lower than the high density platforms commonly used for GS, but training a GS model with less markers results in a significant decrease in prediction accuracy. An opportunity presents itself here to utilise the genetic position of markers along with high density parental data and low density progeny data to exploit identity by descent and improve the accuracy of cost effective, low density genotyping platforms. This paper presents two novel methodologies which achieve this. The first imputes missing markers on low density genotyped lines up to high density, and the second incorporates genetic position, low density progeny data, high density parental data, and high density marker effects in a single prediction step without any imputation being performed. The accuracy of each method is assessed, and the resultant response to selection is compared to single platform strategies under varying cost scenarios.

Statement of Authorship

Title of Paper	Increasing response to genomic selection through concurrent use of low and high density genotyping platforms
Publication Status	□ Published □ Accepted for Publication □ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Unpublished, planned for submission in 2019.

Principal Author

Name of Principal Author (Candidate)	Adam Norman		
Contribution to the Paper	Involved in formulating the research objective and experimental design. Performed all simulations and analysis. Developed the methodology presented in the paper. Wrote and prepared the manuscript. Acted as corresponding author.		
Overall percentage (%)	90%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	3/6/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	James Edwards		
Contribution to the Paper	Involved in formulating the research objective and experimental design. Reviewing and approving the manuscript. PhD co-supervisor of Adam Norman.		
Signature		Date	3/6/2019

Name of Co-Author	Haydn Kuchel		
Contribution to the Paper	Involved in formulating the research objective and experimental design. Contributed to the development of the methodology presented in the paper. Reviewing and approving the manuscript. PhD principal supervisor of Adam Norman.		
Signature		Date	3/6/2019

4.3 Abstract

Genomic selection has been established as a highly valuable breeding tool through two decades of extensive research. However, its application incurs a significant additive genotyping cost, which can be prohibitively high when implementing at a commercial scale. Solutions which lower the cost of application are therefore required, and an obvious potential solution is a low cost, low density genotyping platform. While such platforms are appropriately priced for large scale implementation, they come with a considerable sacrifice in prediction accuracy. Here we present two methodologies for effectively utilising a low density genotyping platform by incorporating high density parental data. We assessed the resultant genomic prediction accuracy and showed both methods to be highly accurate. We also investigated the response to selection achieved under each strategy and showed that with the same genotyping expenditure, using a low density genotyping platform containing 1000 markers at 20% the cost of a high density platform, the imputation method resulted in a 45% increase in response to selection over only using a high density platform. This work demonstrates that the methods presented here can be used by plant breeders to more efficiently increase the rate of genetic gain achieved in their programme.

4.4 Introduction

Over the past two decades extensive research has been carried out establishing, and improving, the methods of genomic selection in numerous species (Meuwissen et al., 2001; Nakaya & Isobe, 2012). It has been shown that genomic selection represents a valuable tool with which wheat breeders can significantly increase the rate of genetic gain in their programmes (based on theory and simulation studies) (Nakaya & Isobe, 2012; Gaynor et al., 2017). For the potential of genomic selection to be fully realised in an applied setting, large numbers of individuals must be genotyped at high density. This is due to the positive relationships between prediction accuracy and training set size, and between prediction accuracy and relatedness (Heffner et al., 2011; Isidro et al., 2015; Norman et al., 2018). The training set should therefore be large and updated regularly with germplasm that is closely related to the current breeding populations. In addition, breeders aim to maximise the size of their breeding populations as this enables higher selection intensity and therefore genetic gain. New breeding populations are generated in each breeding cycle and must be genotyped if genomic selection is to be carried out. The cost of genotyping this large number of individuals at sufficient density is often prohibitively high. Even in the case where a programme can afford the genotyping cost, more genetic gain could be made if genotyping were cheaper as it would permit larger breeding populations to be genotyped, which in turn allows a higher selection intensity. The positive effect that selection accuracy and selection intensity have on genetic gain is given by the well known breeders equation $R = ih^2\sigma$ where R is response to selection, i is selection intensity, h^2 is narrow sense heritability (selection accuracy), and σ represents the standard deviation of the trait (i.e. phenotypic trait variation) (Wricke & Weber, 1986).

A simple approach one could take to reduce the cost of genotyping is to use a platform with fewer markers. While this would result in a significant decrease in prediction accuracy due to short haplotype effects not being captured in the prediction calibration (Heffner et al., 2011; Hickey et al., 2014; Norman et al., 2018), it could still result in higher rates of genetic gain if the decrease in cost enables a sufficient increase in selection intensity. The concept of incorporating the genetic position of markers into quantitative trait loci (QTL) analysis has been widely applied in interval mapping methods (Lander & Botstein, 1989; Haley & Knott, 1992; Zeng, 1994). A logical and obvious extension of this inheritance inference process is to apply it to genomic selection to impute the inheritance of missing marker data. Consequently the relationship between parents and progeny can be used to infer the inheritance of marker alleles that may not actually have been assayed. In other words, marker position and parent-progeny relatedness can be used to impute from a low density platform (LDP) platform to a high density platform (HDP) by utilising LDP data on the progeny alongside HDP parental data. A less obvious approach to utilising LDP data is to incorporate the same linkage and relationship information into a single prediction step where progeny are predicted using LDP data along with HDP parental data and HDP marker effects. This approach has not been presented previously and may be more or less accurate than traditional imputation. If either of these approaches can be carried out with sufficient accuracy they could be used to improve the overall impact achieved with genomic selection (Gorjanc et al., 2017). Methodology and software is readily available for imputation in animal and human genetics (Kong et al., 2008; Howie et al., 2009; Hickey et al., 2011; Antolín et al., 2017) where the population structure is different from inbred plant species, and the target is often imputing sequence data (Van Raden et al., 2015). In plant species such as wheat however, identity by descent information can be utilised by tracing chromosome segments of the two parents to the progeny, which can be very powerful in imputation. Guo & Beavis (2011) showed accurate imputation to be possible in a maize nested association mapping population, but until very recently no imputation methodology had been designed to utilise the structure of bi-parental populations, and many methods faced issues with computation time (Hickey et al., 2015). Gonen et al. (2018) recently published methodology and software for phasing and imputing SNP array data in diploid plant species, which accurately imputed simulated bi-parental populations.

Here we present two methods for using an LDP in conjunction with HDP parental data, and compare the resultant genomic prediction accuracies with those from using only a HDP, and those using only an LDP. The first method iterates over intervals of LDP skeletal markers and imputes alleles at the missing HDP markers in the progeny using linkage information. The second method utilises HDP parental data in the process of calculating genomic predictions on the progeny, which have only LDP data. This is done by loading the prediction calibration onto the LDP markers at each parent based on their linkage phase, then predicting the individuals based on the likelihood of inheritance from each parent. Predictions are therefore calculated without imputing any alleles at missing markers. The rate of genetic gain was then calculated for various costs and accuracies in order to establish the value of using such approaches.

4.5 Materials and Methods

4.5.1 Data simulation

Initially, ten fixed wheat genotype lines with varying relatedness were extracted from the genetic map and marker data published in Norman et al (2017). Allelic information for the lines was restricted to markers only present in the consensus map. The lines were then used to simulate genetic marker data for four cross types including straight cross (A x B), backcross (AB x A), top cross (AB x C), and four-way cross (AB x CD). 5000 recombinant inbred lines (five selfing generations) were simulated for each cross type using the R package simcross (Broman, 2016) available in the R Statistical Computing Environment (R Core Team, 2018). To enable imputation and prediction methods to be assessed in the scenario where we have HDP marker data for parents and LDP data for the lines of interest, the marker subsets from Norman et al. (2018) (100, 500, 1000, 3000, 5000, 10000) were used as the LDP marker sets, and the 13639 markers on the consensus map formed the HDP marker set. The LDP marker sets were selected with the criteria of being evenly positioned on the genome, and having high minor allele frequency. For each marker subset, markers present on both the HDP and LDP sets acted as the skeletal markers, and those present on only the HDP set were treated as missing data at the lines of interest. Genetic marker effects for grain yield were estimated for all 13639 markers by repeating the genomic analysis of the field trial from (Norman et al., 2017) using only these markers. These genetic marker effects were used with the full marker data to calculate the "true" additive genetic values of the simulated lines.

4.5.2 Prediction methods

This section details two novel methodologies for calculating genomic predictions for lines of interest with LDP marker data by utilising HDP marker data on their parents. The imputation method involves imputing the alleles of the HDP markers on the lines of interest, then using the imputed marker data set to calculate genomic predictions. The direct prediction method directly calculates genomic predictions without imputing any allelic values.

The simulated lines detailed in section 4.5.1 were predicted using each method, and the accuracy of the prediction was calculated as the correlation coefficient (r) between the predictions and the "true" additive genetic values. Accuracy of the imputation method was also assessed by the correlation coefficient (r) between the true and imputed genotypes (Calus et al., 2014).

Imputation method

The imputation approach described here allows missing markers to be imputed for a line with LDP data using the HDP data of parents. It is suited for biallelic markers and requires alleles to be coded in -1 and 1 format. Figure 4.1 notates the information required to impute Z, the allelic value of the high density marker at the line of interest.

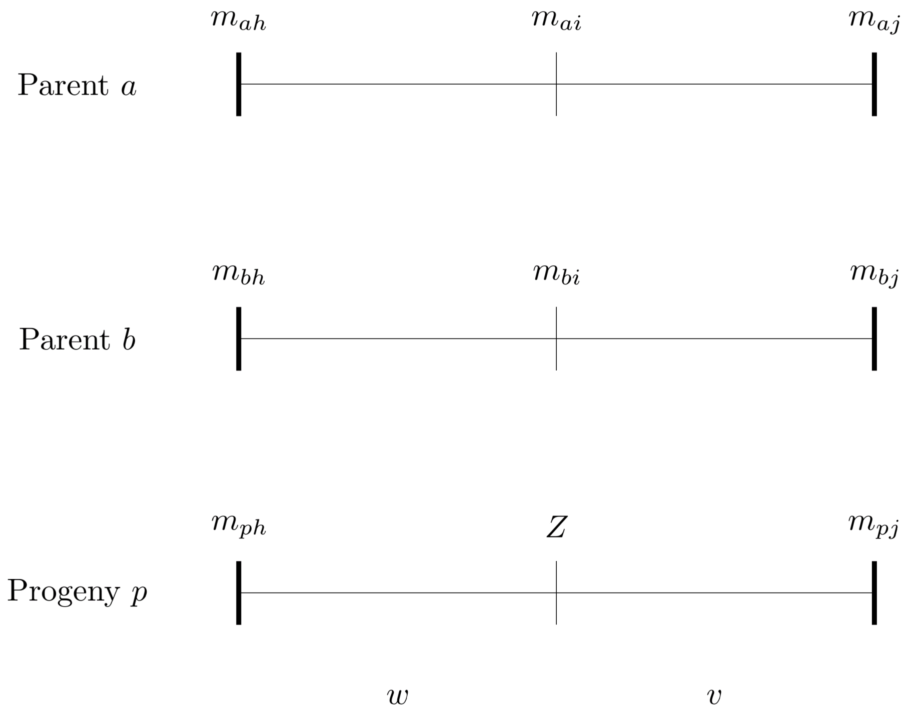

FIGURE 4.1: Notation used in the imputation method equations for parents a and b, and progeny p. m_h and m_j denote the allelic values of the left and right skeletal flanks respectively. m_i represents the parental allelic values at the high density marker to be imputed, and Z denotes the imputed allelic value of the progeny at this high density marker. w represents the cM distance between the high density marker to be imputed and the left skeletal flank, and v denotes the cM distance to the right skeletal flank.

For parents a and b, s_a and s_b represent the proportional similarity at the loci of interest between the line of interest and each parent, and are respectively defined by

$$s_a = 1 - \frac{|m_{ph} - m_{ah}| + |m_{pj} - m_{aj}|}{|m_{ph} - m_{ah}| + |m_{pj} - m_{aj}| + |m_{ph} - m_{bh}| + |m_{pj} - m_{bj}|} \tag{4.1}$$

$$s_b = 1 - \frac{|m_{ph} - m_{bh}| + |m_{pj} - m_{bj}|}{|m_{ph} - m_{ah}| + |m_{pj} - m_{aj}| + |m_{ph} - m_{bh}| + |m_{pj} - m_{bj}|}$$

This equation will not work appropriately when $|m_{ph} - m_{ah}| + |m_{pj} - m_{aj}| + |m_{ph} - m_{bh}| + |m_{pj} - m_{bj}| = 0$, as this infers that the progeny has zero similarity with either parent thus indicating an error in the data. In this case $s_a = s_b = 0.5$.

The imputed allele Z is then defined by

$$Z = (2(1 - \theta_w \theta_v) - 1)z \tag{4.2}$$

where

$$\begin{aligned} z = {} & m_{ai} s_a (1 - (\gamma |m_{ph} - m_{ah}| + \beta |m_{pj} - m_{aj}|)) \\ & + m_{bi} s_b (1 - (\gamma |m_{ph} - m_{bh}| + \beta |m_{pj} - m_{bj}|)) \end{aligned} \tag{4.3}$$

with $\gamma = 1 - (w/(w+v))$ and $\beta = 1 - (v/(w+v))$. Here, if $w + v = 0$ (markers are co-located), we reassign the denominator with 1. In (4.2) $\theta_w = \frac{1}{2}\tanh(\frac{w}{50})$ is the recombination probability calculated using the Kosambi genetic distance mapping function for the left flanking interval distance w. θ_v follows similarly. The use of these probabilities in this context is to adjust for potential double recombinations or interference between the marker being imputed and the skeletal marker.

The GEBVs of a set of lines can then be calculated by

$$GEBV = \mathbf{Mu}$$

where $\mathbf{M}$ is a marker matrix containing all individuals to be predicted with complete allelic information (imputed or otherwise) on the full set of markers, and $\mathbf{u}$ is a vector of marker effects for the trait being predicted.

Direct prediction method

The direct prediction method enables GEBVs to be calculated without imputing any missing alleles. Imputation methodologies are subject to patent claims in some countries (Hayes & Goddard, 2007), whereas the direct prediction method is not. Similar to the imputation method it is suited to biallelic markers and requires alleles to be coded in -1 and 1 format.

Figure 4.2 describes the notation used in the equations where the skeletal marker j is flanked by n_i HDP markers in the left interval and n_k HDP markers in the right interval. Markers have allelic values m, and known marker effects e.

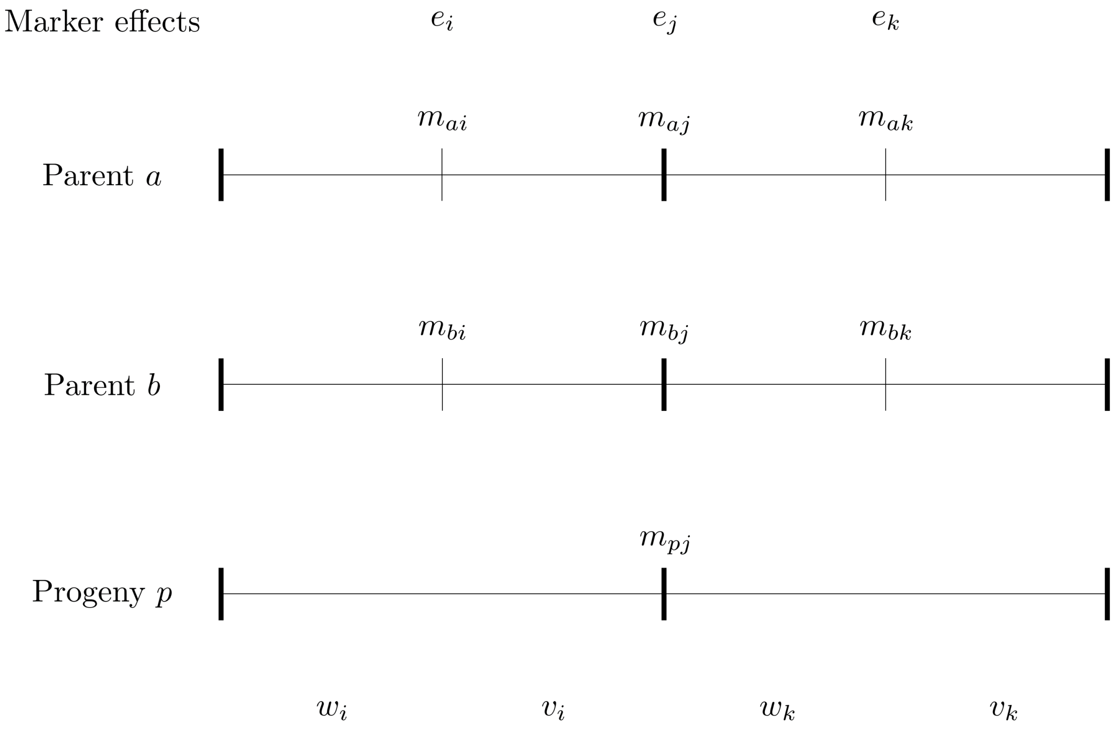

FIGURE 4.2: Notation used in the direct prediction equations to describe allelic values m at parents a and b and progeny p for the skeletal marker j, which is flanked by n_i HDP markers in the left interval and n_k HDP markers in the right interval. Marker effects are shown by e, and intervals (cM distance) are given w_i, v_i, w_k and v_k.

K_{aj} and K_{bj} represent the allelic similarity at the jth skeletal marker between the line of interest and parents a and b respectively, and are defined by

$$K_{aj} = \frac{(m_{aj}m_{pj}) + 1}{2} \tag{4.4}$$

$$K_{bj} = \frac{(m_{bj}m_{pj}) + 1}{2}$$

The HDP marker effects are consolidated onto the skeletal markers by splitting each effect value and loading it onto its two flanking markers. The effect value is split according to its proportional distance to each flank. Let E_{aj} and E_{bj} represent the consolidated allelic effect at the jth skeletal marker for parents a and b respectively. For n_i HDP markers in the left interval and n_k HDP markers in the right interval, E_{aj} and E_{bj} are respectively defined by

$$E_{aj} = e_j m_{aj} + \sum_{i=1}^{n_i} e_i m_{ai} \left(1 - \frac{v_i}{w_i + v_i}\right) + \sum_{k=1}^{n_k} e_k m_{ak} \left(1 - \frac{w_k}{w_k + v_k}\right) \tag{4.5}$$

$$E_{bj} = e_j m_{bj} + \sum_{i=1}^{n_i} e_i m_{bi} \left(1 - \frac{v_i}{w_i + v_i}\right) + \sum_{k=1}^{n_k} e_k m_{bk} \left(1 - \frac{w_k}{w_k + v_k}\right)$$

The consolidated effects E from (4.5) are then used with the allelic similarities K from (4.4) to directly calculate the predicted GEBVs. For n skeletal markers, the GEBV of the line of interest is defined by

$$GEBV = \sum_{j=1}^{n} \left(\frac{K_{aj}}{K_{aj} + K_{bj}} E_{aj} + \frac{K_{bj}}{K_{aj} + K_{bj}} E_{bj} \right) \tag{4.6}$$

In the case where one or both parents are a heterozygous F1, (4.4), (4.5) and (4.6) are expanded back to the four grandparents. If one parent is fixed then its parents are represented by two clones of itself.

LDP-GBLUP method

With the LDP-GBLUP method genomic prediction calibrations were recalculated for each marker density using the methodology described in Norman et al. (2018). Each LDP calibration was then used to calculate genomic predictions by multiplying the matrix of marker effects by the LDP genotype data.

Computations

All analyses were carried out in the R Statistical Computing Environment (R Core Team, 2018). Both the imputation and direct prediction methods were wrapped into R functions. The imputation function took 13 minutes to impute 5000 individuals, while the direct prediction approach was faster, taking just 10 seconds to predict 5000 individuals for one trait (Windows PC with a 4.00Ghz processor). Linear mixed models for the LDP-GBLUP method were fitted using the ASReml-R software (Butler et al., 2009), which is available as an R package and downloadable from *www.vsni.co.uk/software/asreml*.

4.5.3 Response to selection

A theoretical analysis of response to selection was conducted to investigate the relative rates of genetic gain between low and high density genomic selection strategies. Assuming normal distribution of the trait (which for grain yield is a reasonable assumption when the population has not been truncated (Kuchel et al., 2007), as was the case in this population (Norman et al., 2017)), relative genetic gain was calculated under varying genotyping cost ratios (LDP/HDP) and relative LDP prediction accuracies. This was done by increasing the population size in the LDP strategy relative to the genotyping cost ratio, and decreasing the LDP selection proportion to achieve the same resultant population size following selection using each strategy.

Using the well known response to selection equation $R = ih^2\sigma$ (Wricke & Weber, 1986) and a given HDP selection proportion α, the response to selection under the HDP strategy is given by

$$R_{HDP} = \frac{f(\Phi^{-1}(1-\alpha))}{\alpha} h^2 \sigma p \tag{4.7}$$

where $\Phi^{-1}(p)$ represents the quantile function and $f(\Phi^{-1}(p))$ gives the density. The HDP genomic prediction accuracy is given by p, and represents the correlation between the prediction and the "true" additive genetic value.

For a given LDP/HDP cost ratio c, and prediction accuracy relative to the HDP prediction r, the response to selection under the LDP strategy is given by

$$R_{LDP} = \frac{f(\Phi^{-1}(1-\alpha c))}{\alpha c} h^2 \sigma p r \tag{4.8}$$

Genetic gains were calculated for cost ratios ranging from 0.05 to 0.90, relative LDP prediction accuracies between 0.50 and 1.00, and an initial selection proportion of 0.20. It is important to note that for the imputation and direct prediction methods the cost ratio must include the cost of HDP parental data. This will vary between breeding programmes as population sizes vary, but would have only a small impact on cost. A breeding population of 5000 individuals may commonly have between 50 and 150 parents. Narrow sense heritability and standard deviation values were obtained from the grain yield analysis by Norman et al. (2017) ($h^2 = 0.45$, $\sigma = 429.88$). A HDP genomic prediction accuracy of 0.92 was used. This was based on the genomic prediction accuracy of grain yield from Norman et al. (2018) with a training set size of 4000 individuals. Statistical analyses were carried out in the R Statistical Computing Environment (R Core, 2017). The quantile function qnorm() and the density function dnorm() were used.

4.5.4 Data availability

Supplemental files are available at FigShare. File S1 contains all genetic marker data of the original parents and all simulated individuals. File S2 contains the pedigree information used in simulation, imputation and direct prediction. File S3 specifies which markers were included in each subset, and the genetic map position of each marker. File S4 contains the genetic markers effects used to calculate the “true” additive genetic values, and the genomic predictions.

4.6 Results

4.6.1 Prediction method accuracy

The prediction methods were assessed by predicting the simulated lines of each cross type. Accuracy was calculated as the correlation between predicted and “true” additive genetic values. The imputation method was further assessed by correlating the imputed and true genotype data.

Figure 4.3 shows the accuracy of imputing high density markers with the imputation method for the different cross types and marker densities. Of the cross types, straight is clearly the most accurate to impute while backcross is slightly better than top cross, and four-way is significantly less accurate. Straight and top cross showed a similar response to marker density, increasing in accuracy very sharply from 100 to 500 markers, then moderately from 500 to 1000. In contrast, backcross increased at a more linear rate, with the rate of increase only slowing slightly after 1000 markers. The proportion of markers that were polymorphic was higher for the more complex cross types. Straight crosses had on average 40.1% polymorphic markers, backcross 39.4%, top cross 54.9%, and four-way 65.9%.

Genomic predictions of the simulated progeny were calculated using the three LDP prediction methods, and their accuracy was measured as the correlation coefficient (r) against the “true” additive genetic values. Figure 4.4 details the prediction accuracies obtained with each method in each cross type. The imputation method was most accurate across all cross types, with direct prediction being slightly less accurate. The difference between imputation and direct prediction was largest in the four-way cross. The LDP-GBLUP method was significantly less accurate than imputation and direct prediction in all cross types for marker numbers less than 3000, but was comparable at marker numbers greater than 3000. Of the three methods, LDP-GBLUP showed the strongest response to marker density with an average increase of 0.22 from 100 to 1000 markers, and 0.08 from 1000 to 3000. When comparing the accuracy of each cross type, backcross is marginally more accurate than

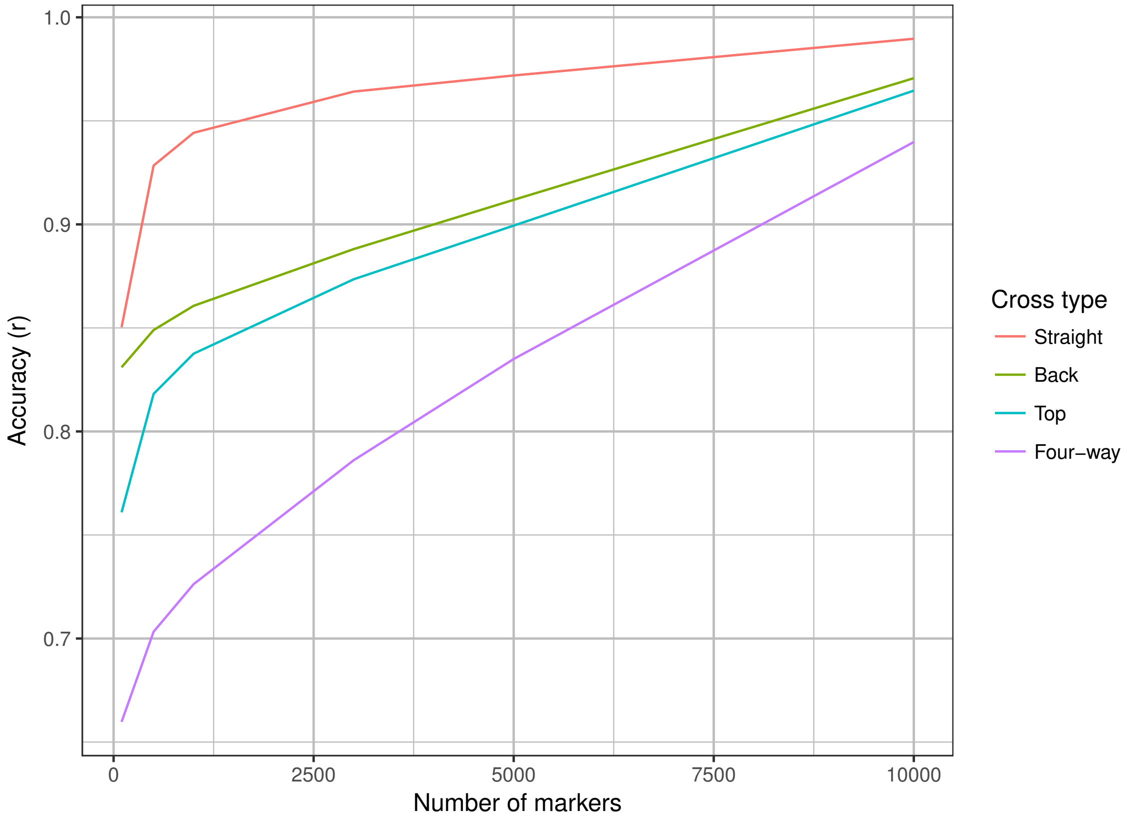

FIGURE 4.3: Imputation accuracy of the four simulated cross types calculated as the correlation coefficient (r) between an individual's imputed and actual allele scores. Low density marker numbers ranged from 100 to 10000.

top cross while straight cross is less accurate than both at marker numbers up to 1000, but more accurate thereafter. Four-way was the least accurate for all methods at all marker densities.

4.6.2 Response to selection

Figure 4.5 demonstrates the response to selection achievable when using HDP and LDP genomic selection strategies. In this example the HDP strategy increased the population mean by 249kg/ha. The increase using the LDP strategy ranged from 259 to 474kg/ha when the prediction accuracy relative to the HDP strategy was 1.00, and from 130 to 237kg/ha when the relative accuracy was 0.50. If the cost ratio is 0.80, a relative prediction accuracy of at least 0.92 is required for the LDP strategy to be more effective than HDP, and if the ratio is 0.20 the required accuracy is 0.65. If we use the relative prediction

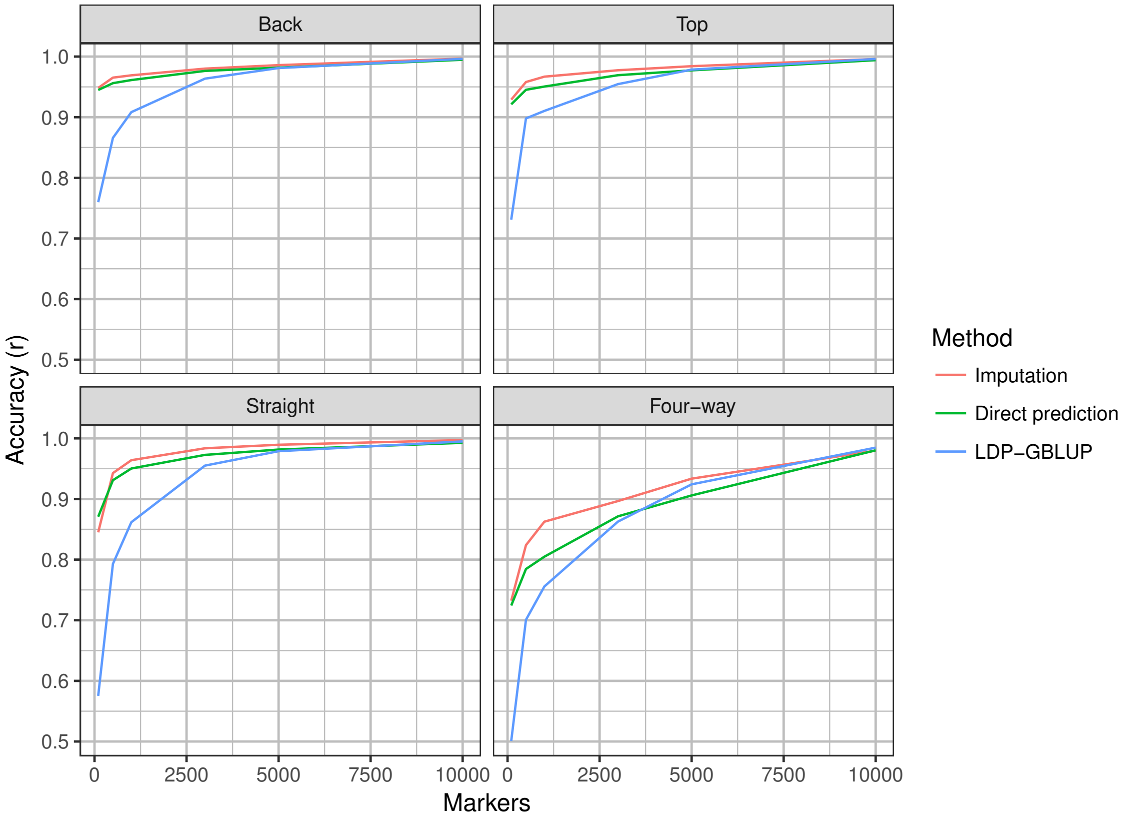

FIGURE 4.4: Accuracy of the three low density prediction methods relative to high density prediction values. Four cross types were simulated and tested with low density marker numbers ranging from 100 to 10000. Accuracy was calculated as the correlation coefficient (*r*) between an individual's prediction for grain yield and its true additive genetic value.

accuracies achieved with each LDP method in the scenario where a LDP genotyping platform contains 1000 markers and costs 20% of the HDP, then the imputation method would produce a 45% increase in response to selection over the HDP method, direct prediction a 42% increase, and LDP-GBLUP a 32% increase.

4.7 Discussion

It has repeatedly been shown that genomic selection has considerable potential as a breeding tool as it can significantly increase the rate of genetic gain achieved by a breeding programme (Meuwissen et al., 2001; Gaynor et al., 2017). However, large breeding populations must be genotyped if genetic gain is to be maximised, both for training the prediction calibration and selecting within a target population. The cost of genotyping is therefore

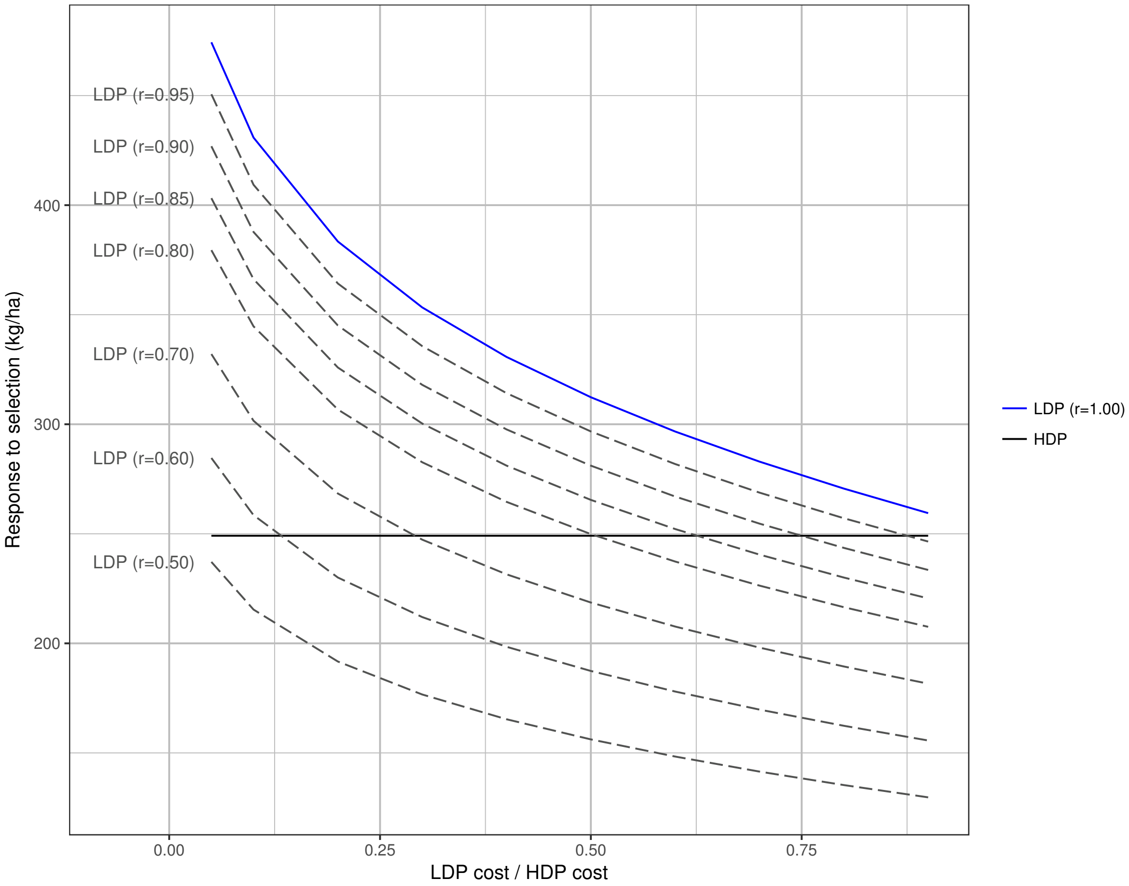

FIGURE 4.5: Response to selection (in kg/ha) achieved by high density platform (HDP) and low density platform (LDP) genomic selection strategies. Multiple LDP scenarios are shown with varying relative genotyping cost and relative prediction accuracies. Genotyping expenditure was the same under each strategy and initial population size was increased according to the genotyping cost ratio.

critical as larger populations can be genotyped if the cost is lower. Larger breeding populations facilitate higher selection intensity which in turn increases the rate of genetic gain. Here we present two methodologies for utilising an LDP along with HDP parental data, and assess their accuracy against only using an LDP and only using a HDP. We also investigated how the relative cost and accuracy of an LDP/HDP strategy interact to affect the rate of genetic gain.

Accuracy of the imputation method was first assessed by comparing the imputed allele values to the actual and then by using the resultant genomic predictions, thus enabling a comparison to the other methods. The imputation accuracies observed for each cross type highlight the importance of clear identity by descent, which is more accurately traced when both parents are homozygous. Straight cross was therefore the most accurately imputed as both parents are homozygous, followed by back and top cross where one parent is homozygous and the other is a heterozygous F1. Four-way crosses proved difficult to impute with high accuracy as the probability of both F1 parents being homozygous at a given allele is significantly lower, in which case the descent of that allele is more ambiguous. The difference in accuracy achieved is also influenced by the proportion of polymorphic markers within a cross, as monomorphic markers are easily imputed as the fixed allele. Backcrosses had more monomorphic markers than top crosses so while both have one homozygous and one heterozygous parent, backcross was more easily imputed. Also, it further explains the low accuracy of the four-way cross type as it had the highest proportion of polymorphic markers. In regards to number of markers used, imputation accuracy drops away significantly at marker numbers less than 500. The average interval was 6.2cM at 500 markers, and 31cM at 100 markers. When imputing a centrally positioned marker this translates to a 0.002 probability of a double recombination at 500 markers, and 0.05 at 100 markers. This increased probability explains the steep decline in accuracy at low marker density.

When comparing the prediction accuracy achieved with each method imputation was the most accurate, showing a slight improvement over direct prediction and a significant improvement over LDP-GBLUP. The imputation method iterates over skeletal marker intervals and therefore tracks the phase using two skeletal markers, where the direct prediction method iterates over each skeletal marker and loads marker effects onto each skeletal marker based on the relative probability of recombination. The imputation approach of utilising interval information is more effective at tracing identity by descent, and is also better equipped to deal with the more complex four-way crosses. Both the imputation and direct prediction methods use the marker effects from the HDP training model for all marker densities, whereas the LDP-GBLUP method uses marker effects from a training model with only the LDP markers. This means the LDP-GBLUP method has coarser resolution and less ability to identify short haplotype effects and accurately predict after genetic recombination has separated the training and prediction sets (Hickey et al., 2014; Norman et al., 2018). This also explains why the LDP-GBLUP method showed greater response to marker density than imputation and direct prediction, as at low marker density

(100 to 1000 markers) the training model is only able to capture long haplotype effects, of which there are few in a quantitative trait such as grain yield (Kuchel et al., 2007; Bennett et al., 2012; Maphosa et al., 2014).

The difference in prediction accuracy between straight, back and top crosses was much smaller than that of imputation accuracy. Despite this, an order was observed where backcrosses were most accurate and closely followed by top and straight. Similar to imputation accuracy the four-way cross type showed much lower accuracy. This follows the order of genetic diversity between the cross types. Despite there being fewer polymorphic markers identified, a straight cross is in fact more diverse than a top cross. When the heterozygous F1 is crossed to the homozygous parent in the top cross the resulting individuals are more closely related to the final parent, whereas individuals from the straight cross segregate equally toward each parent. The decrease in prediction accuracy at very low marker densities was therefore more pronounced in the straight cross than top and backcross. Shrinkage in the prediction calculations at very low densities proved advantageous, and thus contributed to the differences observed between cross types. Here, the predictions are shrunk more when the chance of double recombination is high, or when there is more heterozygosity in the parents. At low marker densities the back and top crosses are therefore shrunk more than the straight crosses as they have heterozygous parents as well as a higher chance of double recombination, where in the straight cross the calculation is more extreme in its prediction and therefore more wrong when a double recombination occurs. In the four-way cross the imputation and direct prediction methods are less able to trace identity by descent when both parents are heterozygous F1s. This explains why these methods perform more similar to the LDP-GBLUP method in the four-way corsses than in the other cross types, and the high level of genetic diversity in the four-way cross explains why the prediction accuracy of this cross type is lower across all three methods.

The response to selection analysis uses the HDP strategy as the base line, and shows how a LDP strategy compares under various cost and accuracy scenarios. The HDP strategy was calculated to increase the population mean by 249kg/ha, while that of the LDP strategy ranged from 130 to 474kg/ha depending on the scenario. The response to selection of the LDP strategy varies significantly with the relative cost and accuracy, and it is therefore important to apply realistic figures when interpreting the potential gains. We used the example of a 1000 marker genotyping platform with 20% the cost of the HDP, and applied the mean accuracy for each LDP method to calculate the response to selection of each method. These results showed imputation and direct prediction to both significantly outperform the LDP-GBLUP method, which highlights the benefit of methods which utilise LDP data in conjunction with HDP parental data. Even at the lowest marker density the relative accuracies reported here are comfortably sufficient for the imputation and direct prediction methods to give higher response to selection than the HDP strategy. The analysis also showed that even using the inferior LDP without imputation or direct prediction (LDP-GBLUP) still resulted a significant improvement in genetic gain over HDP. This demonstrates the importance of increasing the initial population size and selection intensity of a breeding programme, and shows that it outweighs the prediction accuracy trade off identified in this study.

The cost of genotyping sufficiently large populations to deploy genomic selection in a breeding programme is often prohibitively high. While the discussion around utilising a lower cost LDP strategy has thus far focused on the improved response to selection that can be achieved with the same genotyping expenditure, it is also useful to interpret the results in terms of how much cost can be saved without sacrificing genetic gain. Using the previously described scenario of a 1000 marker LDP that is 20% of the cost of a HDP, it was calculated that the response to selection achieved under the HDP strategy can be equalled using the imputation method with a saving in genotyping expenditure of 76.5%. For many breeding programmes the cost of implementing genomic selection using a HDP would be prohibitively high, and this result demonstrates that the methods presented here represent an opportunity to overcome that cost barrier.

The objective of this work was to provide two novel methodologies of concurrently utilising low and high density genotyping platforms, and assess their efficacy relative to using each density platform individually. The findings show that the imputation method produced the highest accuracy, which was followed very closely by the direct prediction method. With the same genotyping expenditure both methods achieved significantly more genetic gain than individual use of either genotype density platform. Alternatively, both methods could also be used to significantly decrease the cost of implementing a genomic selection strategy without sacrificing the genetic gain achieved. Methods such as these can be employed by plant breeders to more efficiently achieve high rates of genetic gain in their programmes. With recent progress in sequencing the wheat genome (Rimbert et al., 2018) this work could be extended to incorporate sequence information.

Bibliography

ANTOLÍN, R., NETTELBLAD, C., GORJANC, G., MONEY, D., & HICKEY, J. (2017). A hybrid method for the imputation of genomic data in livestock populations. *Genetics Selection Evolution* **49**, 30.

BENNETT, D., IZANLOO, A., REYNOLDS, M., KUCHEL, H., LANGRIDGE, P., & SCHNURBUSCH, T. (2012). Genetic dissection of grain yield and physical grain quality in bread wheat (*Triticum aestivum L.)* under water-limited environments. *Theoretical and Applied Genetics* **125**, 255–271.

BROMAN, K. (2016). *simcross: Simulate Experimental Crosses*. R package version 0.2-20.

BUTLER, D., CULLIS, B., GILMOUR, A., & GOGEL, B. (2009). ASReml-R reference manual. *Queensland Department of Primary Industries, Queensland, Australia* .

CALUS, M., BOUWMAN, A., HICKEY, J., VEERKAMP, R., & MULDER, H. (2014). Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. *Animal* **8**, 17431753.

GAYNOR, R., GORJANC, G., BENTLEY, A., OBER, E., HOWELL, P., JACKSON, R., MACKAY, I., & HICKEY, J. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science* **56**, 1–15.

GONEN, S., WIMMER, V., GAYNOR, R., BYRNE, E., GORJANC, G., & HICKEY, J. (2018). A heuristic method for fast and accurate phasing and imputation of single-nucleotide polymorphism data in bi-parental plant populations. *Theoretical and Applied Genetics* pages 1–13.

GORJANC, G., BATTAGIN, M., DUMASY, J., ANTOLIN, R., GAYNOR, R., & HICKEY, J. (2017). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Science* **57**, 216–228.

GUO, B. & BEAVIS, W. (2011). *In silico* genotyping of the maize nested association mapping population. *Molecular breeding : new strategies in plant improvement* **27**, 107–113.

HALEY, C. & KNOTT, S. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

HAYES, B. & GODDARD, M. (2007). Artificial selection method and reagents.

HEFFNER, E., JANNINK, J., & SORRELLS, M. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome* **4**, 65–75.

HICKEY, J., DREISIGACKER, S., CROSSA, J., HEARNE, S., BABU, R., PRASANNA, B., GRONDONA, M., ZAMBELLI, A., WINDHAUSEN, V., MATHEWS, K., ET AL. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science* **54**, 1476–1488.

HICKEY, J., GORJANC, G., VARSHNEY, R., & NETTELBLAD, C. (2015). Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a hidden Markov model. *Crop Science* **55**, 1934–1946.

HICKEY, J., KINGHORN, B., TIER, B., WILSON, J., DUNSTAN, N., & VAN DER WERF, J. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genetics Selection Evolution* **43**, 12.

HOWIE, B., DONNELLY, P., & MARCHINI, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529.

ISIDRO, J., JANNINK, J., AKDEMIR, D., POLAND, J., HESLOT, N., & SORRELLS, M. (2015). Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics* **128**, 145–158.

KONG, A., MASSON, G., FRIGGE, M., GYLFASON, A., ZUSMANOVICH, P., THORLEIFSSON, G., OLASON, P., INGASON, A., STEINBERG, S., RAFNAR, T., ET AL. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics* **40**, 1068.

KUCHEL, H., WILLIAMS, K., LANGRIDGE, P., EAGLES, H., & JEFFERIES, S. (2007). Genetic dissection of grain yield in bread wheat. I. QTL analysis. *Theoretical and Applied Genetics* **115**, 1029–1041.

LANDER, E. & BOTSTEIN, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

MAPHOSA, L., LANGRIDGE, P., TAYLOR, H., PARENT, B., EMEBIRI, L., KUCHEL, H., REYNOLDS, M., CHALMERS, K., OKADA, A., EDWARDS, J., ET AL. (2014). Genetic control of grain yield and grain physical characteristics in a bread wheat population grown under a range of environmental conditions. *Theoretical and Applied Genetics* **127**, 1607.

MEUWISSEN, T., HAYES, B., & GODDARD, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

NAKAYA, A. & ISOBE, S. (2012). Will genomic selection be a practical method for plant breeding? *Annals of botany* **110**, 1303–1316.

NORMAN, A., TAYLOR, J., EDWARDS, J., & KUCHEL, H. (2018). Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3: Genes, Genomes, Genetics* **8**, 2889–2899.

NORMAN, A., TAYLOR, J., TANAKA, E., TELFER, P., EDWARDS, J., MARTINANT, J., & KUCHEL, H. (2017). Increased genomic prediction accuracy in wheat breeding using a large Australian panel. *Theoretical and Applied Genetics* **130**, 2543–2555.

R CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RIMBERT, H., DARRIER, B., NAVARRO, J., KITT, J., CHOULET, F., LEVEUGLE, M., DUARTE, J., RIVIERE, N., EVERSOLE, K., ET AL. (2018). High throughput SNP discovery and genotyping in hexaploid wheat. *PloS one* **13**, e0186329.

VAN RADEN, P., SUN, C., & OCONNELL, J. (2015). Fast imputation using medium or low-coverage sequence data. *BMC genetics* **16**, 82.

WRICKE, G. & WEBER, E. (1986). *Quantitative genetics and selection in plant breeding*. Walter de Gruyter.

ZENG, Z. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

Chapter 5

General discussion

This project was established in response to a knowledge gap regarding the application of GS in wheat breeding. At the time the project was conceived GS was well understood from a theoretical perspective and a number of studies had demonstrated its potential in small research populations. However, there was a significant lack of information on how GS should best be implemented in a wheat breeding programme. Therefore, the intended research outcomes of this project were to: i) establish the achievable accuracy of genomic prediction in a large breeding population, ii) identify criteria for the optimal design of a GS training strategy, and iii) formulate methods for reducing the cost of implementing GS.

5.1 Potential efficacy of genomic prediction

In addressing the first research objective of establishing the achievable accuracy of genomic prediction, three major conclusions were drawn:

1. The high genomic prediction accuracies observed in this study demonstrate significant potential for GS to improve wheat breeding. This was shown using a novel approach for assessing prediction accuracy where genomic predictions were correlated to additive genetic value. This method revealed that prediction accuracies did not vary substantially across traits, but the proportion of genetic variance that is additive did.

2. Genomic prediction was significantly more accurate than QTL-based prediction for all traits tested. This shows that even for qualitative traits under simple genetic control, GS is more accurate than marker assisted selection. While marker assisted

selection may remain preferable in certain situations due to the immediate cost advantage, the price of genotyping for GS can be shared across all predictions made on a line during its selection lifetime, which will significantly influence which strategy should be chosen.

3. Incorporating genomic relationship information in the analysis of phenotype data significantly improves the linear mixed model accuracy. This translates to a more accurate assessment of individuals, and allows both performance and breeding values to be calculated. Response to selection can therefore be increased as selections are made based on the appropriate value (performance value for commercial selection, breeding value for parent selection).

These results confirm the significant potential of GS, and justify further investigation into how it can be applied in a way that optimises genetic gain by balancing accuracy and cost effectiveness.

5.2 Application and optimisation of genomic selection

The second research objective, identifying criteria for the optimal design of a GS training strategy, resulted in seven key conclusions which fall under four categories:

1. Training set size

    a. Accuracy was always higher with a larger training set, but the rate of accuracy increase slowed at larger sizes. This confirms that training sets significantly larger than those previously studied are required if the efficacy of GS is to be maximised in a breeding programme. While increasing the size of the training set continued to improve prediction accuracy up to the largest sizes tested, the smaller response observed at large training sizes could be exploited in traits that are expensive to phenotype as breeders can use these results to calculate if the value gained by increasing the training size outweighs the expense of phenotyping additional lines.

    b. The response of accuracy to training set size was independent of the genetic complexity of the trait.

2. Germplasm relatedness

    a. Prediction accuracy is improved when the relatedness between training and validation sets is higher.

    b. Prediction accuracy can be increased with additional genetic diversity in the training set. This was particularly true when there was lower relatedness between training and validation sets.

3. Including more breeding cohorts

    a. Traits under heavy selection pressure can be more accurately predicted by training with additional previous breeding cohorts.

4. Marker density

    a. Response to marker density is higher when the training set is more diverse.

    b. The required marker density is higher when predicting more distant breeding generations.

The conclusions drawn from the second research objective inform plant breeders on how they can optimise their GS training strategies. They highlight that the optimal marker density fluctuates under different scenarios. The genotyping platform used in a programme is a long-term decision that cannot be frequently changed without substantial cost. The most demanding requirements of a breeding programme must therefore be considered when selecting a genotyping platform, in addition to other criteria such as cost. Methods for reducing the overall cost of genotyping are explored in the third research objective.

5.3 Maximising genetic gain per unit cost

The third research objective focused on reducing the cost of implementing an accurate GS strategy. Two novel methodologies for utilising both a low and high density genotyping platform were developed and tested, allowing three conclusions to be made:

1. Imputation from a low density genotyping platform to high density genotype data was the most accurate method of utilising a low density platform. In some programmes this cost reduction would enable GS to be applied where it otherwise

would not be financially possible. In programmes that have stronger financial resources this cost reduction can be used to increase population size and thus genetic gain. The direct prediction method developed in this study was slightly less accurate, but is an effective alternative that is not subject to patent claims (Hayes & Goddard, 2007).

2. Imputation and direct prediction strategies both achieve substantially higher response to selection over a single platform strategy at the same cost.

3. Using only a low density platform achieved higher response to selection than using a more expensive HD platform. This is an example of genetic gain being increased by sacrificing prediction accuracy in favour of population size. While the loss in accuracy could be greater in certain scenarios which could affect the final outcome, this result should serve as a reminder to remain focussed on achieving genetic gain, and not solely on prediction accuracy.

5.4 Utilising genomic selection in a breeding programme

Research studies such as this have been required in order to characterise the overall potential of GS and to understand the intricacies involved in its application. This specific knowledge of GS in wheat is now sufficient for breeding programmes to leverage the technology. The knowledge we have gained in this study can be used to identify the most effective and efficient strategies of deploying GS. By implementing GS effectively, breeding programmes can be improved in three ways: increasing population size, earlier selection for economically important traits, and increasing selection accuracy. This section discusses these points in the context of improving a fixed line wheat breeding programme by implementing GS with a low density/high density genotyping strategy via imputation or direct prediction. Not considered here are other major alterations such as doubled haploid, rapid generation cycling, and shuttle breeding.

5.4.1 Earlier selection for economically important traits

A conventional wheat breeding programme consists of a cascade of correlated tests due to the staggered availability of phenotype data. Initially breeders select for highly heritable traits that are easy to measure on single plants such as plant height, phenology, and foliar disease resistance. Complex traits and those expensive or less accurate to assess are left until fixed line seed has been multiplied, with phenotyping and selection being carried out sequentially over several years (Bernardo, 2002; Collard & Mackill, 2008). Tandem selection such as this is necessary in a conventional programme due to the phenotyping timeline, but it is well understood that the method is ineffective when selecting for negatively correlated traits (Yan & Frégeau-Reid, 2008). Additionally, tandem selection is most effective when selecting in order of trait importance (Pešek & Baker, 1969), which is not possible when selecting visually in the population stage before any yield testing can be carried out. Hazel & Lush (1942) showed index selection to be more effective than tandem selection and independent culling. However, selection indices require data for all traits undergoing selection, which is not feasible when using standard phenotyping at the early stages of a breeding programme. Genomic predictions are valuable here as they are simultaneously available for all traits of interest (providing a prediction calibration is available), thus enabling index selection to be carried out from early in the breeding programme before the population is truncated. Therefore, implementing GS provides earlier selection for economically important traits, and allows the most effective selection method to be utilised.

The selection indices employed should recognise the difference between threshold traits (end use quality, physical grain quality) and those requiring continual improvement (grain yield) (Goddard, 1983; Itoh & Yamada, 1988; Groen et al., 1994; Byrne et al., 2016), and also differentiate breeding value from commercial value (Yan & Frégeau-Reid, 2008; Muñoz et al., 2014; Crossa et al., 2017). Thresholds should be set according to known control varieties in order to provide context to the trait predictions and index values. For scenarios where phenotype data is available for multiple traits, models which incorporate this and account for trait correlations could be utilised (Scutari et al., 2014).

5.4.2 Increasing population size

Population size is a key factor influencing genetic gain as it facilitates higher selection intensity and genetic variance (Bernardo, 2002; Witcombe & Virk, 2001). The simplest means of increasing population size is by increasing the scale of the breeding programme. However, this requires a corresponding increase in the operating budget to match phenotyping capacity to phenotyping demand, and is therefore not a desirable solution. Implementing GS at the point where fixed lines are derived (spike or plant selection) can achieve larger population size through deriving additional fixed lines that are then selected with GS prior to preliminary yield testing. The cost of genotyping the additional lines could be partially offset by reducing the number of individuals in the initial yield testing stage. By selecting for multiple traits at this point (via selection indices), the effective population size is further increased as the number of lines that will fail to meet economic thresholds during advanced testing will be reduced.

5.4.3 Increasing selection accuracy

By implementing genomic selection, breeding programmes have dense marker data available on their breeding lines. The findings in Chapter 2, which are supported by Van Raden (2008); de Los Campos et al. (2009); Hayes et al. (2009), show that this marker data can also be utilised to improve the accuracy of breeding trial analyses. In this case, historical phenotype data can also be incorporated into the analysis as the marker based relationship matrix provides a means of linking the historical and current datasets. This concept can be employed for traits that are first phenotyped during advanced trials, which would increase confidence in the small amount of direct phenotype data available (Hayes et al., 2009). Partially replicated trials are an effective means of maximising the number of individuals that can be phenotypically tested relative to the number of plots sown (Cullis et al., 2006). By incorporating marker based relationship into these analyses, the level of replication can be reduced dramatically allowing for either more individuals or more environments to be tested for the same resources. In a traditional breeding programme, replication in preliminary yield trials grown at three locations may result in a plot to entry ratio of six (i.e. two replicates at three sites) (Bernardo, 2002). In contrast, within a GS context a considerably more sparse approach could be used, and with the aid of genomic assisted design and analysis it may be possible to achieve similar selection accuracy with plot to entry ratios approaching one (Hill & Weir, 2012; Cowling et al., 2015). This conceptual framework can be utilised during initial yield testing where increasing population

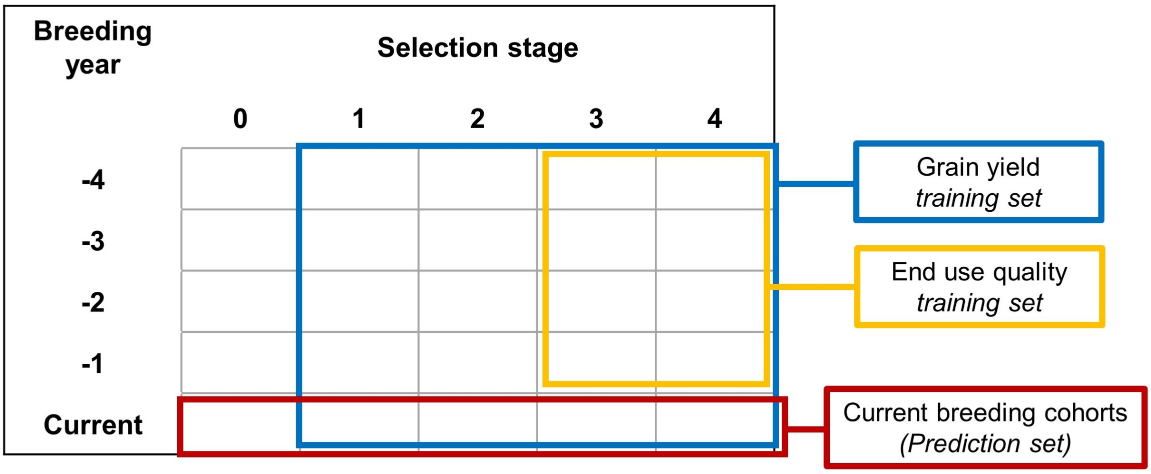
Breeding year
Selection stage
0
1
2
3
4
-4
-3
-2
-1
Current
Grain yield *training set*
End use quality *training set*
Current breeding cohorts *(Prediction set)*

FIGURE 5.1: Selection stage 0 represents individuals or populations immediately prior to yield testing, and stages 1-4 represent the subsequent rounds of phenotyping and selection. The current breeding year contains the breeding cohorts of interest that are to be selected, with previous years showing the historical data used for training prediction calibrations. Training sets for grain yield and end use quality would consist of previous breeding cohorts that have been phenotyped. The training set for grain yield may or may not include the current breeding cohorts depending on whether the prediction calibration was built before or after harvest. Current breeding cohorts can then be predicted.

size is more beneficial than increasing accuracy due to the wide genetic variance being assessed and the relatively low selection intensity being performed.

5.4.4 Model training strategy

Many of these implementations of genomic selection are enabled by the existence of global prediction calibrations. The training set for these calibrations (Figure 5.1) should consist of previous breeding cohorts and associated data, and should be updated each season when new phenotype data is generated in order to maximise relatedness between the training set and current breeding cohorts (Jannink et al., 2010).

In order to maintain allelic representation in the training set, a small number of lines with poor predictions should be carried forward from fixed line derivation to be phenotyped and included in the training set (Hickey et al., 2014). In scenarios where insufficient historic data exists (young breeding programme, new breeding objective) an alternative approach to global calibrations could be live training where a subset of the current breeding cohort is phenotyped and used as the training set, allowing the remainder of the cohort to then be predicted. As found in Chapter 3, high prediction accuracies can be achieved with small training sets if relatedness between training and prediction sets is very high. In this case each family would be directly represented in the training set through full siblings, which achieves a very high degree of relatedness.

5.5 Future work and opportunities

5.5.1 Multiple environment datasets

The work carried out here was done using data from a single environment trial in order to remove the confounding effect of genotype-by-environment interaction, and thus enable an accurate assessment of each test variable. This improved our ability to assess prediction accuracies, as poorly correlating environments would deleteriously affect the measurement of predictive ability. In practice, breeding programmes must make selections for their target environment which often consists of a wide range of contrasting environment types. Therefore, to improve the relevance of this work to applied plant breeding, future work should focus on extending the research questions addressed here to a multi-environment dataset. Ideally this would be done in a single stage analysis (Oakey et al., 2016), but this remains a computational challenge when working with large datasets (Schulz-Streeck et al., 2013). Further developments in analysis tools which address this challenge would therefore assist researchers in extending this work.

5.5.2 Rapid generation cycle genomic selection

The main application of the work presented here is selection within a population of fixed lines early in the breeding programme, but another application of GS with high potential is in rapid cycle crossing (Bernardo, 2010; Gaynor et al., 2017). Here complex F1 lines are genotyped immediately after being produced and GS is performed to select parents for direct use in the next crossing block. This practice would reduce generation interval from approximately five years to three months or lower (Zheng et al., 2013), and thus has significant potential to increase the rate of genetic gain. Challenges of this approach include: i) high risk involved with crossing lines based only on a genomic prediction and no phenotype data (due to potentially inaccurate genomic predictions), ii) prediction calibrations are required for all traits, as no phenotypic selection is used to assist in selecting parents,

iii) difficulties in genotyping, and predicting heterozygous individuals, iv) erosion of genomic prediction accuracy caused by the decay of linkage disequilibrium between markers and QTL, and v) logistical challenges associated with the short timeframes required for genotype data production and selection. If these issues can be addressed, rapid cycle crossing with GS could revolutionise early generation plant breeding.

5.6 Conclusion

Genomic selection is a promising emerging technology that offers much, but requires considerable research for its potential to be realised. It is this need which motivated the studies encompassed in this thesis. Herein research is presented which considers several potentially derailing issues, and how plant breeders can implement a GS programme most optimally. This was broken down into three intended research outcomes; to establish the achievable accuracy of GS, identify optimal training criteria, and reduce the cost of implementing GS.

In the first component of this study we confirmed the significant potential of GS by predicting breeding germplasm with very high accuracy, and also showed that incorporating the genomic relationship matrix into phenotypic analysis improves model accuracy substantially. With the high potential confirmed, the second and third components would focus on enabling this potential to be realised. The findings of the second component highlight to breeders how important an appropriate training strategy is, as large differences in accuracy were observed when varying the training set size, relatedness, variance, and genetic diversity. Breeders can refer to these findings to assist them in designing an effective training strategy which suits their programme. In the third component we provide two novel methodologies for accurately utilising a low density genotyping platform which can be used to reduce the cost of implementing GS. These approaches can be used either to lower the cost of GS application without sacrificing genetic gain, or to significantly increase the rate of genetic gain with equal cost.

In summary, this body of work contributes knowledge which plant breeders can use when designing and implementing GS within their breeding programmes. The findings clarify previous uncertainties and overcome several key constraints in the application of GS, and can therefore be applied to enable increased rates of genetic in wheat breeding programmes around the world.

Bibliography

BERNARDO, R. (2002). *Breeding for quantitative traits in plants*. Stemma Press Woodbury.

BERNARDO, R. (2010). Genomewide selection with minimal crossing in self-pollinated crops. *Crop Science* **50**, 624–627.

BYRNE, T., SANTOS, B., AMER, P., MARTIN-COLLADO, D., PRYCE, J., & AXFORD, M. (2016). New breeding objectives and selection indices for the australian dairy industry. *Journal of dairy science* **99**, 8146–8167.

COLLARD, B. & MACKILL, D. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **363**, 557–572.

COWLING, W., STEFANOVA, K., BEECK, C., NELSON, M., HARGREAVES, B., SASS, O., GILMOUR, A., & SIDDIQUE, K. (2015). Using the animal model to accelerate response to selection in a self-pollinating crop. *G3: Genes, Genomes, Genetics* **5**, 1419–1428.

CROSSA, J., PÉREZ-RODRÍGUEZ, P., CUEVAS, J., MONTESINOS-LÓPEZ, O., JARQUÍN, D., DE LOS CAMPOS, G., BURGUEÑO, J., GONZÁLEZ-CAMACHO, J., PÉREZ-ELIZALDE, S., BEYENE, Y., ET AL. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in plant science* **22**, 961–975.

CULLIS, B., SMITH, A., & COOMBES, N. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 381.

DE LOS CAMPOS, G., GIANOLA, D., & ROSA, G. (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of animal science* **87**, 1883–1887.

GAYNOR, R., GORJANC, G., BENTLEY, A., OBER, E., HOWELL, P., JACKSON, R., MACKAY, I., & HICKEY, J. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science* **56**, 1–15.

GODDARD, M. (1983). Selection indices for non-linear profit functions. *Theoretical and Applied Genetics* **64**, 339–344.

GROEN, A., VOLLEMA, A., BRASCAMP, E., ET AL. (1994). A comparison of alternative index procedures for multiple generation selection on non-linear profit. *Animal Science* **59**, 1–9.

HAYES, B. & GODDARD, M. (2007). Artificial selection method and reagents.

HAYES, B., VISSCHER, P., & GODDARD, M. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research* **91**, 47–60.

HAZEL, L. & LUSH, J. (1942). The efficiency of three methods of selection. *Journal of Heredity* **33**, 393–399.

HICKEY, J., DREISIGACKER, S., CROSSA, J., HEARNE, S., BABU, R., PRASANNA, B., GRONDONA, M., ZAMBELLI, A., WINDHAUSEN, V., MATHEWS, K., ET AL. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science* **54**, 1476–1488.

HILL, W. & WEIR, B. (2012). Variation in actual relationship among descendants of inbred individuals. *Genetics research* **94**, 267–274.

ITOH, Y. & YAMADA, Y. (1988). Linear selection indices for non-linear profit functions. *Theoretical and Applied Genetics* **75**, 553–560.

JANNINK, J., LORENZ, A., & IWATA, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics* **9**, 166–177.

MUÑOZ, P., RESENDE, M., GEZAN, S., RESENDE, M., DE LOS CAMPOS, G., KIRST, M., HUBER, D., & PETER, G. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* **198**, 1759–1768.

OAKEY, H., CULLIS, B., THOMPSON, R., COMADRAN, J., HALPIN, C., & WAUGH, R. (2016). Genomic selection in multi-environment crop trials. *G3: Genes, Genomes, Genetics* **6**, 1113–1126.

PEŠEK, J. & BAKER, R. (1969). Desired improvement in relation to selection indices. *Canadian journal of plant science* **49**, 803–804.

SCHULZ-STREECK, T., OGUTU, J., & PIEPHO, H. (2013). Comparisons of single-stage and two-stage approaches to genomic selection. *Theoretical and applied genetics* **126**, 69–82.

SCUTARI, M., HOWELL, P., BALDING, D. J., & MACKAY, I. (2014). Multiple quantitative trait analysis using bayesian networks. *Genetics* **198**, 129–137.

VAN RADEN, P. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science* **91**, 4414–4423.

WITCOMBE, J. & VIRK, D. (2001). Number of crosses and population size for participatory and classical plant breeding. *Euphytica* **122**, 451–462.

YAN, W. & FRÉGEAU-REID, J. (2008). Breeding line selection based on multiple traits. *Crop Science* **48**, 417–423.

ZHENG, Z., WANG, H., CHEN, G., YAN, G., & LIU, C. (2013). A procedure allowing up to eight generations of wheat and nine generations of barley per annum. *Euphytica* **191**, 311–316.

Appendix A

Chapter 2 Supplementary material

The marker data used in this study was made available via the electronic version of this article 12 months after the publication date.

A.1 Supplementary material 1

Plant material. Summary of the genetic material used in the study.
https://goo.gl/Tzzs4J

A.2 Supplementary material 2

Phenotype distribution plots. Distribution plots of raw phenotype data from the Roseworthy field trial. Germplasm is displayed in groups of AYT-South, AYT-Other, and PYT-South.
https://goo.gl/wd6Edr

A.3 Supplementary material 3

Phenotype data. Raw phenotype data from the Roseworthy field trial.
https://goo.gl/BEUs89

A.4 Supplementary material 4

Genetic map. Genetic map positions for all markers. Includes all nine bi-parental maps, the consensus map, and the unscaled map.
https://goo.gl/RtZmAL

A.5 Supplementary material 5

Trait QTL summary. Table summarising the markers used in multiple linear regression for each trait.
https://goo.gl/DJnmvZ

Appendix B

Chapter 3 Supplementary material

All material is available online via Figshare at: https://goo.gl/HdJy7f

B.1 Supplementary material 1

Clusters and breeding cohorts. Specifies which breeding lines belong to each cluster and breeding cohort.

B.2 Supplementary material 2

Genetic map plots. Genetic map plots of each marker subset used. These consisted of 100, 500, 1000, 3000, 5000, 10000 and 13639 markers.

B.3 Supplementary material 3

Marker subsets. Specifies which markers were included in each subset, and the genetic map position of each marker.

B.4 Supplementary material 4

Genetic marker data. Full marker data of the 10,375 individuals and 17,181 markers.

B.5 Supplementary material 5

Phenotype data. Phenotype data for the four traits from the 2014 field experiment conducted at Roseworthy, South Australia.

Appendix C

Chapter 4 Supplementary material

All material is available online via Figshare at: https://goo.gl/PvJtEz

C.1 Supplementary material 1

Genetic marker data. Full marker data of the original parents and all simulated individuals.

C.2 Supplementary material 2

Pedigrees. Pedigree information used in simulation, imputation and direct prediction.

C.3 Supplementary material 3

Genetic map and marker subsets. Specifies which markers were included in each subset, and the genetic map position of each marker.

C.4 Supplementary material 4

Marker effects. The genetic marker effects used to calculate the "true" additive genetic values, and the genomic predictions.

Bibliography

AKBARI, M., WENZL, P., CAIG, V., CARLING, J., XIA, L., YANG, S., USZYNSKI, G., MOHLER, V., LEHMENSIEK, A., KUCHEL, H., HAYDEN, M., HOWES, N., SHARP, P., VAUGHAN, P., RATHMELL, B., HUTTNER, E., & KILIAN, A. (2006). Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theoretical and applied genetics* **113**, 1409–1420.

AKHUNOV, E., NICOLET, C., & DVORAK, J. (2009). Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theoretical and applied genetics* **119**, 507–517.

ANTOLÍN, R., NETTELBLAD, C., GORJANC, G., MONEY, D., & HICKEY, J. (2017). A hybrid method for the imputation of genomic data in livestock populations. *Genetics Selection Evolution* **49**, 30.

AUINGER, H., SCHÖNLEBEN, M., LEHERMEIER, C., SCHMIDT, M., KORZUN, V., GEIGER, H., PIEPHO, H., GORDILLO, A., WILDE, P., BAUER, E., ET AL. (2016). Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale L.*). *Theoretical and Applied Genetics* **129**, 2043–2053.

BASFORD, K., KROONENBERG, P., & I., D. (1991). Three-way methods for multi-attribute genotype x environment data: an illustrated partial survey. *Field crops research* **27**, 131–157.

BATTENFIELD, S., GUZMÁN, C., GAYNOR, R., SINGH, R., PEÑA, R., DREISIGACKER, S., FRITZ, A., & POLAND, J. (2016). Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. *The Plant Genome* **9**.

BEN HASSEN, M., CAO, T., BARTHOLOMÉ, J., ORASEN, G., COLOMBI, C., RAKOTOMALALA, J., RAZAFINIMPIASA, L., BERTONE, C., BISELLI, C., VOLANTE, A., ET AL. (2018). Rice diversity panel provides accurate genomic predictions for complex traits in the progenies of biparental crosses involving members of the panel. *Theoretical and Applied Genetics* **131**, 417–435.

BENNETT, D., IZANLOO, A., EDWARDS, J., KUCHEL, H., CHALMERS, K., TESTER, M., REYNOLDS, M., SCHNURBUSCH, T., & LANGRIDGE, P. (2012a). Identification of novel quantitative trait loci for days to ear emergence and flag leaf glaucousness in a bread wheat (*Triticum aestivum L.)* population adapted to southern Australian conditions. *Theoretical and Applied Genetics* **124**, 697–711.

BENNETT, D., IZANLOO, A., REYNOLDS, M., KUCHEL, H., LANGRIDGE, P., & SCHNURBUSCH, T. (2012b). Genetic dissection of grain yield and physical grain quality in bread wheat (*Triticum aestivum L.)* under water-limited environments. *Theoretical and Applied Genetics* **125**, 255–271.

BENTLEY, A., SCUTARI, M., GOSMAN, N., FAURE, S., BEDFORD, F., HOWELL, P., COCKRAM, J., ROSE, G., BARBER, T., IRIGOYEN, J., ET AL. (2014). Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. *Theoretical and applied genetics* **127**, 2619–2633.

BERNARDO, R. (2002). *Breeding for quantitative traits in plants*. Stemma Press Woodbury.

BERNARDO, R. (2010). Genomewide selection with minimal crossing in self-pollinated crops. *Crop Science* **50**, 624–627.

BHATT, G. & DERERA, N. (1975). Genotype x environment interactions for, heritabilities of, and correlations among quality traits in wheat. *Euphytica* **24**, 597–604.

BÖRNER, V. & REINSCH, N. (2012). Optimising multistage dairy cattle breeding schemes including genomic selection using decorrelated or optimum selection indices. *Genetics, selection, evolution : GSE* **44**, 1.

BOUQUET, A. & JUGA, J. (2013). Integrating genomic selection into dairy cattle breeding programmes: a review. *Animal : an international journal of animal bioscience* **7**, 705–713.

BROMAN, K. (2016). *simcross: Simulate Experimental Crosses*. R package version 0.2-20.

BROMAN, K. & SEN, S. (2009). *A Guide to QTL Mapping with R/**qtl***. Springer-Verlag.

BROMAN, K. & WU, H. (2015). ***qtl**: Tools for Analayzing QTL Experiments.* R package version 1.36-6.

BROOKS, A., JENNER, C., & ASPINALL, D. (1982). Effects of water deficit on endosperm starch granules and on grain physiology of wheat and barley. *Functional Plant Biology* **9**, 423–436.

BUTLER, D. (2016). *Package 'pedicure': pedigree tools.*

BUTLER, D., CULLIS, B., GILMOUR, A., & GOGEL, B. (2009). ASReml-R reference manual. *Queensland Department of Primary Industries, Queensland, Australia* .

BUTLER, J., BYRNE, P., MOHAMMADI, V., CHAPMAN, P., & HALEY, S. (2005). Agronomic performance of *Rht* alleles in a spring wheat population across a range of moisture levels. *Crop science* **45**, 939–947.

BYRNE, T., SANTOS, B., AMER, P., MARTIN-COLLADO, D., PRYCE, J., & AXFORD, M. (2016). New breeding objectives and selection indices for the australian dairy industry. *Journal of dairy science* **99**, 8146–8167.

CALUS, M., BOUWMAN, A., HICKEY, J., VEERKAMP, R., & MULDER, H. (2014). Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. *Animal* **8**, 17431753.

CANE, K., EAGLES, H., LAURIE, D., TREVASKIS, B., VALLANCE, N., EASTWOOD, R., GORORO, N., KUCHEL, H., & MARTIN, P. (2013). *Ppd-B1* and *Ppd-D1* and their effects in southern Australian wheat. *Crop and Pasture Science* **64**, 100–114.

CAVANAGH, C., CHAO, S., WANG, S., HUANG, B., STEPHEN, S., KIANI, S., FORREST, K., SAINTENAC, C., BROWN-GUEDIRA, G., AKHUNOVA, A., SEE, D., BAI, G., PUMPHREY, M., TOMAR, L., WONG, D., KONG, S., REYNOLDS, M., DA SILVA, M., BOCKELMAN, H., TALBERT, L., ANDERSON, J., DREISIGACKER, S., BAENZIGER, S., CARTER, A., KORZUN, V., MORRELL, P., DUBCOVSKY, J., MORELL, M., SORRELLS, M., HAYDEN, M., & AKHUNOV, E. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences* **110**, 8057–8062.

CHAO, S., DUBCOVSKY, J., DVORAK, J., LUO, M., BAENZIGER, S., MATNYAZOV, R., CLARK, D., TALBERT, L., ANDERSON, J., DREISIGACKER, S., ET AL. (2010). Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum L.). BMC genomics* **11**, 727.

CLOSE, T., BHAT, P., LONARDI, S., WU, Y., ROSTOKS, N., RAMSAY, L., DRUKA, A., STEIN, N., SVENSSON, J., WANAMAKER, S., BOZDAG, S., ROOSE, M., MOSCOU, M., CHAO, S., VARSHNEY, R., SZŰCS, P., SATO, K., HAYES, P., MATTHEWS, D., KLEINHOFS, A., MUEHLBAUER, G., DEYOUNG, J., MARSHALL, D., MADISHETTY, K., FENTON, R., CONDAMINE, P., GRANER, A., & WAUGH, R. (2009). Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* **10**, 1–13.

COLLARD, B., JAHUFER, M., BROUWER, J., & PANG, E. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* **142**, 169–196.

COLLARD, B. & MACKILL, D. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **363**, 557–572.

COOPER, M. & DELACY, I. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics* **88**, 561–572.

COOPER, M., MESSINA, C., PODLICH, D., TOTIR, L., BAUMGARTEN, A., HAUSMANN, N., WRIGHT, D., & GRAHAM, G. (2014). Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop and Pasture Science* **65**, 311–336.

COOPER, M., VAN EEUWIJK, F., HAMMER, G., PODLICH, D., & MESSINA, C. (2009). Modeling QTL for complex traits: detection and context for plant breeding. *Current opinion in plant biology* **12**, 231–240.

COWLING, W., STEFANOVA, K., BEECK, C., NELSON, M., HARGREAVES, B., SASS, O., GILMOUR, A., & SIDDIQUE, K. (2015). Using the animal model to accelerate response to selection in a self-pollinating crop. *G3: Genes, Genomes, Genetics* **5**, 1419–1428.

CROSSA, J., DE LOS CAMPOS, G., PÉREZ, P., GIANOLA, D., BURGUEÑO, J., ARAUS, J., MAKUMBI, D., SINGH, R., DREISIGACKER, S., YAN, J., ET AL. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**, 713–724.

CROSSA, J., JARQUÍN, D., FRANCO, J., PÉREZ-RODRÍGUEZ, P., BURGUEÑO, J., SAINT-PIERRE, C., VIKRAM, P., SANSALONI, C., PETROLI, C., AKDEMIR, D., ET AL. (2016). Genomic prediction of gene bank wheat landraces. *G3: Genes, Genomes, Genetics* **6**, 1819–1834.

CROSSA, J., PERÉZ, P., HICKEY, J., BURGUEÑO, J., ORNELLA, L., CERÓN-ROJAS, J., ZHANG, X., DREISIGACKER, S., BABU, R., LI, Y., BONNETT, D., & MATHEWS, K. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* **112**, 48–60.

CROSSA, J., PÉREZ-RODRÍGUEZ, P., CUEVAS, J., MONTESINOS-LÓPEZ, O., JARQUÍN, D., DE LOS CAMPOS, G., BURGUEÑO, J., GONZÁLEZ-CAMACHO, J., PÉREZ-ELIZALDE, S., BEYENE, Y., ET AL. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in plant science* **22**, 961–975.

CULLIS, B., SMITH, A., & COOMBES, N. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 381.

DAETWYLER, H., BANSAL, U., BARIANA, H., HAYDEN, M., & HAYES, B. (2014). Genomic prediction for rust resistance in diverse wheat landraces. *Theoretical and Applied Genetics* **127**, 1795–1803.

DAETWYLER, H., CALUS, M., PONG-WONG, R., DE LOS CAMPOS, G., & HICKEY, J. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* **193**, 347–365.

DAETWYLER, H., VILLANUEVA, B., BIJMA, P., & WOOLLIAMS, J. (2007). Inbreeding in genome-wide selection. *Journal of Animal Breeding and Genetics* **124**, 369–376.

DAWSON, J., ENDELMAN, J., HESLOT, N., CROSSA, J., POLAND, J., DREISIGACKER, S., MANÈS, Y., SORRELLS, M., & JANNINK, J. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research* **154**, 12–22.

de Los Campos, G., Gianola, D., & Rosa, G. (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of animal science* **87**, 1883–1887.

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

Dekkers, J. (2007). Prediction of response to marker-assisted and genomic selection using selection index theory. *Journal of animal breeding and genetics* **124**, 331–341.

Dekkers, J., Hospital, F., et al. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* **3**, 22–32.

Desta, Z. & Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in plant science* **19**, 592–601.

Edwards, J. (2012). *A Genetic Analysis of Drought Related Traits in Hexaploid Wheat*. PhD thesis, The University of Adelaide.

Elshire, R., Glaubitz, J., Sun, Q., Poland, J., Kawamoto, K., Buckler, E., & Mitchell, S. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**, e19379.

Estaghvirou, S., Ogutu, J., Schulz-Streeck, T., Knaak, C., Ouzunova, M., Gordillo, A., & Piepho, H. (2013). Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. *BMC genomics* **14**, 860.

FAO (2017). Food and Agriculture Organization of the United Nations: Food and agiculture data.

Fischer, R. & Wood, J. (1979). Drought resistance in spring wheat cultivars. iii.* yield associations with morpho-physiological traits. *Crop and Pasture Science* **30**, 1001–1020.

FORNI, S., AGUILAR, I., & MISZTAL, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics Selection Evolution* **43**, 1.

GARDNER, K., WITTERN, L., & MACKAY, I. (2016). A highly recombined, high-density, eight-founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments. *Plant biotechnology journal* **14**, 1406–1417.

GAYNOR, R., GORJANC, G., BENTLEY, A., OBER, E., HOWELL, P., JACKSON, R., MACKAY, I., & HICKEY, J. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science* **56**, 1–15.

GILMOUR, A. (2007). Mixed model regression mapping for QTL detection in experimental crosses. *Computational statistics & data analysis* **51**, 3749–3764.

GILMOUR, A., CULLIS, B., & VERBYLA, A. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 269–293.

GILMOUR, A., THOMPSON, R., & CULLIS, B. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450.

GODDARD, M. (1983). Selection indices for non-linear profit functions. *Theoretical and Applied Genetics* **64**, 339–344.

GONEN, S., WIMMER, V., GAYNOR, R., BYRNE, E., GORJANC, G., & HICKEY, J. (2018). A heuristic method for fast and accurate phasing and imputation of single-nucleotide polymorphism data in bi-parental plant populations. *Theoretical and Applied Genetics* pages 1–13.

GONZÁLEZ-CAMACHO, J., ORNELLA, L., PÉREZ-RODRÍGUEZ, P., GIANOLA, D., DREISIGACKER, S., & CROSSA, J. (2018). Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *The plant genome* **11**.

GORJANC, G., BATTAGIN, M., DUMASY, J., ANTOLIN, R., GAYNOR, R., & HICKEY, J. (2017). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Science* **57**, 216–228.

GORJANC, G., GAYNOR, R., & HICKEY, J. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics* **131**, 1953–1966.

GORJANC, G., JENKO, J., HEARNE, S., & HICKEY, J. (2016). Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC genomics* **17**, 30.

GROEN, A., VOLLEMA, A., BRASCAMP, E., ET AL. (1994). A comparison of alternative index procedures for multiple generation selection on non-linear profit. *Animal Science* **59**, 1–9.

GUO, B. & BEAVIS, W. (2011). *In silico* genotyping of the maize nested association mapping population. *Molecular breeding : new strategies in plant improvement* **27**, 107–113.

HALDANE, J. (1946). The interaction of nature and nurture. *Annals of eugenics* **13**, 197–205.

HALEY, C. & KNOTT, S. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

HAO, C., WANG, L., GE, H., DONG, Y., & ZHANG, X. (2011). Genetic diversity and linkage disequilibrium in Chinese bread wheat (*Triticum aestivum L.)* revealed by SSR markers. *PLoS One* **6**, e17279.

HAYASHI, T. & IWATA, H. (2010). EM algorithm for bayesian estimation of genomic breeding values. *BMC genetics* **11**, 3.

HAYES, B. & GODDARD, M. (2007). Artificial selection method and reagents.

HAYES, B., VISSCHER, P., & GODDARD, M. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research* **91**, 47–60.

HAZEL, L. (1943). The genetic basis for constructing selection indexes. *Genetics* **28**, 476–490.

HAZEL, L. & LUSH, J. (1942). The efficiency of three methods of selection. *Journal of Heredity* **33**, 393–399.

HE, S., REIF, J., KORZUN, V., BOTHE, R., EBMEYER, E., & JIANG, Y. (2017). Genome-wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to central europe. *Theoretical and Applied Genetics* **130**, 635–647.

HE, S., SCHULTHESS, A., MIRDITA, V., ZHAO, Y., KORZUN, V., BOTHE, R., EBMEYER, E., REIF, J., & JIANG, Y. (2016). Genomic selection in a commercial winter wheat population. *Theoretical and Applied Genetics* **129**, 641–651.

HEFFNER, E., JANNINK, J., IWATA, H., SOUZA, E., & SORRELLS, M. (2011a). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Science* **51**, 2597–2606.

HEFFNER, E., JANNINK, J., & SORRELLS, M. (2011b). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome* **4**, 65–75.

HEFFNER, E., LORENZ, A., JANNINK, J., & SORRELLS, M. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Science* **50**, 1681–1690.

HEFFNER, E., SORRELLS, M., & JANNINK, J. (2009). Genomic selection for crop improvement. *Crop Science* **49**, 1–12.

HENDERSON, C. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226–252.

HESLOT, N., JANNINK, J., & SORRELLS, M. (2015). Perspectives for genomic selection applications and research in plants. *Crop Science* **55**, 1–12.

HESLOT, N., RUTKOSKI, J., POLAND, J., JANNINK, J., & SORRELLS, M. (2013). Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PLoS One* **8**, e74612.

HESLOT, N., YANG, H., SORRELLS, M., & JANNINK, J. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science* **52**, 146–160.

HICKEY, J., DREISIGACKER, S., CROSSA, J., HEARNE, S., BABU, R., PRASANNA, B., GRONDONA, M., ZAMBELLI, A., WINDHAUSEN, V., MATHEWS, K., ET AL. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science* **54**, 1476–1488.

HICKEY, J., GORJANC, G., VARSHNEY, R., & NETTELBLAD, C. (2015). Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a hidden Markov model. *Crop Science* **55**, 1934–1946.

HICKEY, J., KINGHORN, B., TIER, B., WILSON, J., DUNSTAN, N., & VAN DER WERF, J. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genetics Selection Evolution* **43**, 12.

HILL, W., GODDARD, M., & VISSCHER, P. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics* **4**, e1000008.

HILL, W. & WEIR, B. (2012). Variation in actual relationship among descendants of inbred individuals. *Genetics research* **94**, 267–274.

HORNER, T. & FREY, K. (1957). Methods for determining natural areas for oat varietal recommendations. *Agronomy Journal* **49**, 313–315.

HOWIE, B., DONNELLY, P., & MARCHINI, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529.

HUANG, B., GEORGE, A., FORREST, K., KILIAN, A., HAYDEN, M., MORELL, M., & CAVANAGH, C. (2012). A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant biotechnology journal* **10**, 826–839.

HUANG, X., CLOUTIER, S., LYCAR, L., RADOVANOVIC, N., HUMPHREYS, D., NOLL, J., SOMERS, D., & BROWN, P. (2006). Molecular detection of QTLs for agronomic and quality traits in a doubled haploid population derived from two Canadian wheats (*Triticum aestivum L.). Theoretical and Applied Genetics* **113**, 753–766.

ISIDRO, J., JANNINK, J., AKDEMIR, D., POLAND, J., HESLOT, N., & SORRELLS, M. (2015). Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics* **128**, 145–158.

ITOH, Y. & YAMADA, Y. (1988). Linear selection indices for non-linear profit functions. *Theoretical and Applied Genetics* **75**, 553–560.

JANNINK, J., LORENZ, A., & IWATA, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics* **9**, 166–177.

JANSEN, R. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

JENNER, C., UGALDE, T., & ASPINALL, D. (1991). The physiology of starch and protein deposition in the endosperm of wheat. *Functional Plant Biology* **18**, 211–226.

JULIANA, P., SINGH, R., SINGH, P., CROSSA, J., HUERTA-ESPINO, J., LAN, C., BHAVANI, S., RUTKOSKI, J., POLAND, J., BERGSTROM, G., & SORRELLS, M. (2017). Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theoretical and applied genetics* **130**, 1415–1430.

KANG, H., SUL, J., SERVICE, S., ZAITLEN, N., KONG, S., FREIMER, N., SABATTI, C., ESKIN, E., ET AL. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348–354.

KAO, C., ZENG, Z., & TEASDALE, R. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

KOEBNER, R. & SUMMERS, R. (2003). 21st century wheat breeding: plot selection or plate detection? *Trends in biotechnology* **21**, 59–63.

KONG, A., MASSON, G., FRIGGE, M., GYLFASON, A., ZUSMANOVICH, P., THORLEIFSSON, G., OLASON, P., INGASON, A., STEINBERG, S., RAFNAR, T., ET AL. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics* **40**, 1068.

KUCHEL, H., WILLIAMS, K., LANGRIDGE, P., EAGLES, H., & JEFFERIES, S. (2007). Genetic dissection of grain yield in bread wheat. I. QTL analysis. *Theoretical and Applied Genetics* **115**, 1029–1041.

LADO, B., MATUS, I., RODRÍGUEZ, A., INOSTROZA, L., POLAND, J., BELZILE, F., DEL POZO, A., QUINCKE, M., CASTRO, M., & VON ZITZEWITZ, J. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3: Genes— Genomes— Genetics* **3**, 2105–2114.

LANDER, E. & BOTSTEIN, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

LIN, C. (1978). Index selection for genetic improvement of quantitative characters. *Theoretical and applied genetics* **52**, 49–56.

LIU, Z., SEEFRIED, F., REINHARDT, F., RENSING, S., THALLER, G., & REENTS, R. (2011). Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genetics Selection Evolution* **43**, 19.

LORENZ, A., HAMBLIN, M., & JANNINK, J. (2010). Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PloS one* **5**, e14079.

LORENZANA, R. & BERNARDO, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics* **120**, 151–161.

LUND, M., VAN DEN BERG, I., MA, P., BRØNDUM, R., & SU, G. (2016). How to improve genomic predictions in small dairy cattle populations. *animal* **10**, 1042–1049.

MACKAY, I., BANSEPT-BASLER, P., BARBER, T., BENTLEY, A., COCKRAM, J., GOSMAN, N., GREENLAND, A., HORSNELL, R., HOWELLS, R., OSULLIVAN, D., ET AL. (2014). An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. *G3: Genes— Genomes— Genetics* **4**, 1603–1610.

MAPHOSA, L., LANGRIDGE, P., TAYLOR, H., PARENT, B., EMEBIRI, L., KUCHEL, H., REYNOLDS, M., CHALMERS, K., OKADA, A., EDWARDS, J., ET AL. (2014). Genetic control of grain yield and grain physical characteristics in a bread wheat population grown under a range of environmental conditions. *Theoretical and Applied Genetics* **127**, 1607.

MEUWISSEN, T., HAYES, B., & GODDARD, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

MICHEL, S., AMETZ, C., GUNGOR, H., AKGÖL, B., EPURE, D., GRAUSGRUBER, H., LÖSCHENBERGER, F., & BUERSTMAYR, H. (2017). Genomic assisted selection for enhancing line breeding: merging genomic and phenotypic selection in winter wheat breeding programs with preliminary yield trials. *Theoretical and Applied Genetics* **130**, 363–376.

MICHEL, S., KUMMER, C., GALLEE, M., HELLINGER, J., AMETZ, C., AKGÖL, B., EPURE, D., LÖSCHENBERGER, F., & BUERSTMAYR, H. (2018). Improving the baking quality of bread wheat by genomic selection in early generations. *Theoretical and Applied Genetics* **131**, 477–493.

MISZTAL, I. & LEGARRA, A. (2017). Invited review: efficient computation strategies in genomic selection. *animal* **11**, 731–736.

MORRELL, P., BUCKLER, E., & ROSS-IBARRA, J. (2011). Crop genomics: advances and applications. *Nature reviews. Genetics* **13**, 85–96.

MUIR, W. (2007). Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics* **124**, 342–355.

MUÑOZ, P., RESENDE, M., GEZAN, S., RESENDE, M., DE LOS CAMPOS, G., KIRST, M., HUBER, D., & PETER, G. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* **198**, 1759–1768.

NAKAYA, A. & ISOBE, S. (2012). Will genomic selection be a practical method for plant breeding? *Annals of botany* **110**, 1303–1316.

NEUMANN, K., KOBILJSKI, B., DENČIĆ, S., VARSHNEY, R., & BÖRNER, A. (2011). Genome-wide association mapping: a case study in bread wheat (*Triticum aestivum L.*). *Molecular Breeding* **27**, 37–58.

NORMAN, A., TAYLOR, J., EDWARDS, J., & KUCHEL, H. (2018). Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3: Genes, Genomes, Genetics* **8**, 2889–2899.

NORMAN, A., TAYLOR, J., TANAKA, E., TELFER, P., EDWARDS, J., MARTINANT, J., & KUCHEL, H. (2017). Increased genomic prediction accuracy in wheat breeding using a large Australian panel. *Theoretical and Applied Genetics* **130**, 2543–2555.

OAKEY, H., CULLIS, B., THOMPSON, R., COMADRAN, J., HALPIN, C., & WAUGH, R. (2016). Genomic selection in multi-environment crop trials. *G3: Genes, Genomes, Genetics* **6**, 1113–1126.

OAKEY, H., VERBYLA, A., PITCHFORD, W., CULLIS, B., & KUCHEL, H. (2006). Joint modeling of additive and non-additive genetic line effects in single field trials. *Theoretical and applied genetics* **113**, 809–819.

OURY, F. & GODIN, C. (2007). Yield and grain protein concentration in bread wheat: how to use the negative relationship between the two characters to identify favourable genotypes? *Euphytica* **157**, 45–57.

PATTERSON, H. & THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

PATTERSON, N., PRICE, A., & REICH, D. (2006). Population structure and eigenanalysis. *PLoS genet* **2**, e190.

PEŠEK, J. & BAKER, R. (1969). Desired improvement in relation to selection indices. *Canadian journal of plant science* **49**, 803–804.

PIEPHO, H. (2009). Ridge regression and extensions for genome-wide selection in maize. *Crop Science* **49**, 1165–1176.

PIEPHO, H., MÖHRING, J., SCHULZ-STREECK, T., & OGUTU, J. (2012a). A stage-wise approach for the analysis of multi-environment trials. *Biometrical journal* **54**, 844–860.

PIEPHO, H., OGUTU, J., SCHULZ-STREECK, T., ESTAGHVIROU, B., GORDILLO, A., & TECHNOW, F. (2012b). Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. *Crop Science* **52**, 1093–1104.

PODLICH, D. & COOPER, M. (1998). QU-GENE: a simulation platform for quantitative analysis of genetic models. *Bioinformatics* **14**, 632–653.

POLAND, J., ENDELMAN, J., DAWSON, J., RUTKOSKI, J., WU, S., MANES, Y., DREISIGACKER, S., CROSSA, J., SÁNCHEZ-VILLEDA, H., SORRELLS, M., ET AL. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome* **5**, 103–113.

POLAND, J. & RIFE, T. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* **5**, 92–102.

POZNIAK, C. (2016). IWGSC whole genome shotgun sequencing of chinese spring: Towards a reference sequence of wheat. In *Plant and Animal Genome XXIV Conference*. Plant and Animal Genome.

PRICE, A., ZAITLEN, N., REICH, D., & PATTERSON, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463.

R CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R DEVELOPMENT CORE TEAM (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RASMUSSON, J. (1933). A contribution to the theory of quantitative character inheritance. *Hereditas* **18**, 245–261.

REBETZKE, G., RATTEY, A., FARQUHAR, G., RICHARDS, R., & CONDON, A. (2013). Genomic regions for canopy temperature and their genetic association with stomatal conductance and grain yield in wheat. *Functional Plant Biology* **40**, 14–33.

REBETZKE, G. & RICHARDS, R. (1999). Genetic improvement of early vigour in wheat. *Crop and Pasture Science* **50**, 291–302.

RHARRABTI, Y., VILLEGAS, D., ROYO, C., MARTOS-NÚÑEZ, V., & GARCIA DEL MORAL, L. (2003). Durum wheat quality in mediterranean environments: II. influence of climatic variables and relationships between quality parameters. *Field Crops Research* **80**, 133–140.

RIMBERT, H., DARRIER, B., NAVARRO, J., KITT, J., CHOULET, F., LEVEUGLE, M., DUARTE, J., RIVIERE, N., EVERSOLE, K., ET AL. (2018). High throughput SNP discovery and genotyping in hexaploid wheat. *PloS one* **13**, e0186329.

RUTKOSKI, J., BENSON, J., JIA, Y., BROWN-GUEDIRA, G., JANNINK, J., & SORRELLS, M. (2012). Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. *The Plant Genome* **5**, 51–61.

RUTKOSKI, J., POLAND, J., JANNINK, J., & SORRELLS, M. (2013). Imputation of unordered markers and the impact on genomic selection accuracy. *G3: Genes— Genomes— Genetics* **3**, 427–439.

SADEQUE, A. & TURNER, M. (2010). QTL analysis of plant height in hexaploid wheat doubled haploid population. *Thai Journal of Agricultural Science* **43**, 91–96.

SADRAS, V., ROGET, D., & O'LEARY, G. (2002). On-farm assessment of environmental and management factors influencing wheat grain quality in the Mallee. *Crop and Pasture Science* **53**, 811–820.

SANNEMANN, W., HUANG, B., MATHEW, B., & LÉON, J. (2015). Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept. *Molecular Breeding* **35**, 1–16.

SAX, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* **8**, 552–560.

SCHMIDT, M., KOLLERS, S., MAASBERG-PRELLE, A., GROSSER, J., SCHINKEL, B., TOMERIUS, A., GRANER, A., & KORZUN, V. (2016). Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. *Theoretical and Applied Genetics* **129**, 203–213.

SCHULZ-STREECK, T., OGUTU, J., & PIEPHO, H. (2013). Comparisons of single-stage and two-stage approaches to genomic selection. *Theoretical and applied genetics* **126**, 69–82.

SCUTARI, M., HOWELL, P., BALDING, D. J., & MACKAY, I. (2014). Multiple quantitative trait analysis using bayesian networks. *Genetics* **198**, 129–137.

SCUTARI, M., MACKAY, I., & BALDING, D. (2016). Using genetic distance to infer the accuracy of genomic prediction. *PLOS Genetics* **12**, 1–19.

SHARMA, D. & ANDERSON, W. (2004). Small grain screenings in wheat: interactions of cultivars with season, site, and management practices. *Crop and Pasture Science* **55**, 797–809.

SILLANPÄÄ, M. & ARJAS, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Theoretical and Applied Genetics* **148**, 1373–1388.

SIMMONDS, N. (1995). The relation between yield and protein in cereal grain. *Journal of the Science of Food and Agriculture* **67**, 309–315.

SMITH, H. (1936). A discriminant function for plant selection. *Annals of Eugenics* **7**, 240–250.

SNAPE, J., SARMA, R., QUARRIE, S., FISH, L., GALIBA, G., & SUTKA, J. (2001). Mapping genes for flowering time and frost tolerance in cereals using precise genetic stocks. *Euphytica* **120**, 309–315.

SOLBERG, T., SONESSON, A., WOOLLIAMS, J., ET AL. (2008). Genomic selection using different marker types and densities. *Journal of animal science* **86**, 2447–2454.

SOLLER, M., BRODY, T., & GENIZI, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47**, 35–39.

SOMERS, D., BANKS, T., DEPAUW, R., FOX, S., CLARKE, J., POZNIAK, C., & MCCARTNEY, C. (2007). Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. *Genome* **50**, 557–567.

STRANDÉN, I. & GARRICK, D. (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of dairy science* **92**, 2971–2975.

SUKUMARAN, S., DREISIGACKER, S., LOPES, M., CHAVEZ, P., & REYNOLDS, M. (2015). Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. *Theoretical and applied genetics* **128**, 353–363.

SUN, X., WU, K., ZHAO, Y., KONG, F., HAN, G., JIANG, H., HUANG, X., LI, R., WANG, H., & LI, S. (2009). QTL analysis of kernel shape and weight using recombinant inbred lines in wheat. *Euphytica* **165**, 615–624.

TAYLOR, J. & BUTLER, D. (2017). R package ASMap: Efficient genetic linkage map construction and diagnosis. *Journal of Statistical Software* **79**, 1–29.

TAYLOR, J., VERBYLA, A., ET AL. (2011). R package wgaim: QTL analysis in bi-parental populations using linear mixed models. *Journal of Statistical Software* **40**, 1–18.

TRIMBLE (2016). GreenSeeker crop sensing system. URL: https://goo.gl/RPo6nH.

TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D., & ALTMAN, R. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics* **17**, 520–525.

VAN RADEN, P. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science* **91**, 4414–4423.

VAN RADEN, P., SUN, C., & OCONNELL, J. (2015). Fast imputation using medium or low-coverage sequence data. *BMC genetics* **16**, 82.

VERBYLA, A., CULLIS, B., & THOMPSON, R. (2007). The analysis of QTL by simultaneous use of the of the full linkage map. *Theoretical and Applied Genetics* **116**, 95–111.

VERBYLA, A., TAYLOR, J., & VERBYLA, K. (2012). RWGAIM: an efficient high-dimensional random whole genome average (QTL) interval mapping approach. *Genetics Research* **94**, 291–306.

WANG, D., EL-BASYONI, S., BAENZIGER, P., CROSSA, J., ESKRIDGE, K., & DWEIKAT, I. (2012). Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity* **109**, 313–319.

WANG, S., WONG, D., FORREST, K., ALLEN, A., CHAO, S., HUANG, B., MACCAFERRI, M., SALVI, S., MILNER, S., CATTIVELLI, L., MASTRANGELO, A., WHAN, A., STEPHEN, S., BARKER, G., WIESEKE, R., PLIESKE, J., INTERNATIONAL WHEAT GENOME SEQUENCING CONSORTIUM, LILLEMO, M., MATHER, D., APPELS, R., DOLFERUS, R., BROWN-GUEDIRA, G., KOROL, A., AKHUNOVA, A., FEUILLET, C., SALSE, J., MORGANTE, M., POZNIAK, C., LUO, M., DVORAK, J., MORELL, M., DUBCOVSKY, J., GANAL, M., TUBEROSA, R., LAWLEY, C., MIKOULITCH, I., CAVANAGH, C., EDWARDS, K., HAYDEN, M., & AKHUNOV, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnol Journal* **12**, 787–796.

WHITTAKER, J., THOMPSON, R., & DENHAM, M. (2000). Marker-assisted selection using ridge regression. *Genetical research* **75**, 249–252.

WITCOMBE, J. & VIRK, D. (2001). Number of crosses and population size for participatory and classical plant breeding. *Euphytica* **122**, 451–462.

WRICKE, G. & WEBER, E. (1986). *Quantitative genetics and selection in plant breeding*. Walter de Gruyter.

WRIGLEY, C. & RATHJEN, A. (1981). Wheat breeding in australia. In Carr, S. & Carr, S., editors, *Plants and Man in Australia*, pages 96–135. Academic Press, New York.

WU, Y., BHAT, P., CLOSE, T., & LONARDI, S. (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genetics* **4**, e1000212.

WU, Y., CLOSE, T., & LONARDI, S. (2011). Accurate construction of consensus genetic maps via integer linear programming. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**, 381–394.

XU, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789–801.

XU, S. (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**, 513–521.

XU, S. & ATCHLEY, W. (1995). A random model approach to interval mapping of quantitative trait loci. *Genetics* **141**, 1189–1197.

XU, Y. & CROUCH, J. (2008). Marker-assisted selection in plant breeding: from publications to practice. *Crop Science* **48**, 391–407.

YAN, W. & FRÉGEAU-REID, J. (2008). Breeding line selection based on multiple traits. *Crop Science* **48**, 417–423.

ZADOKS, J., CHANG, T., KONZAK, C., ET AL. (1974). A decimal code for the growth stages of cereals. *Weed res* **14**, 415–421.

ZANKE, C., LING, J., PLIESKE, J., KOLLERS, S., EBMEYER, E., KORZUN, V., ARGILLIER, O., STIEWE, G., HINZE, M., NEUMANN, K., ET AL. (2014). Whole genome association mapping of plant height in winter wheat (*Triticum aestivum L.*). *PloS one* **9**, e113287.

ZENG, Z. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 10972–10976.

ZENG, Z. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

ZEUTEC (2016). SpectraAlyzer grain. URL: https://goo.gl/tv3hPM.

ZHANG, Z., ERBE, M., HE, J., OBER, U., GAO, N., ZHANG, H., SIMIANER, H., & LI, J. (2015). Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. *G3: Genes, Genomes, Genetics* **5**, 615–627.

ZHANG, Z., ERSOZ, E., LAI, C., TODHUNTER, R., TIWARI, H., GORE, M., BRADBURY, P., YU, J., ARNETT, D., ORDOVAS, J., ET AL. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics* **42**, 355–360.

ZHAO, H., NETTLETON, D., SOLLER, M., & DEKKERS, J. (2005). Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genetical research* **86**, 77–87.

ZHENG, Z., WANG, H., CHEN, G., YAN, G., & LIU, C. (2013). A procedure allowing up to eight generations of wheat and nine generations of barley per annum. *Euphytica* **191**, 311–316.

ZHONG, S., DEKKERS, J., FERNANDO, R., & JANNINK, J. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* **182**, 355–364.