THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Improved methods of predicting genetic merit in plant breeding programs using linear mixed models

Colleen Hunt

B.Sc.(Hons), M.Sc.(Math and Comp Sci)

0000-0001-8359-5318

A thesis submitted for the degree of Doctor of Philosophy at The University of Queensland in 2020

Queensland Alliance for Agriculture and Food Innovation

Abstract

Plant breeders are constantly faced with many challenges, particularly in testing of genotypes. These may include the variability among and within environments, seed availability and resources. Despite careful planning and management there are always uncontrolled factors that can be minimised by using appropriate statistical analysis techniques, which in some circumstances are very complex. The implementation of genomic prediction and subsequent selection has opened new avenues for in-depth exploration of better statistical methods for optimising plant breeding.

For many years Australian plant breeding trials have been analysed using techniques that include the genetic parentage and adjustments for the spatial arrangement of the genotypes in the field. This can be extended further to allow for inter-plot competition. Inter-plot competition is of particular value for trials that have two row plots. The added advantage of including pedigrees in the analysis allows for the possible detection of particular families that may be prone to competition effects.

The relationship between genotypes can be calculated using knowledge of the parent lines. This can be extended to also include the relationship calculated using marker information. We have developed a model that teases apart the parentage and the marker relationships to investigate possible increases in prediction accuracy. We observe that the difference in accuracy is largely affected by the environment (GxE) with some trials analysed optimally with only marker information and others best analysed with both pedigree and markers.

As a further improvement of the analysis, we investigate the effect of partitioning the genetic variance into additive and dominance effects while simultaneously allowing for the spatial field effects and GxE. We have found that including dominance has an effect on the accuracy of the additive effects, which in turn has an effect on selection. This study also showed that the presence of a dominance effect has a strong environmental interaction.

The final study considers the optimal combinations of testers and lines for early generation trials. The presence/magnitude of dominance and GxE has a detrimental effect on the selections of early generation hybrids that use only a single tester. We investigate this issue using trials that have two testers and compare results between the testers. Results vary between environments, in most cases the use of a single tester has limited capacity to genomically predict the performance of the lines crossed with a second tester.

Plant breeding programs require careful planning and construction of trials, with one of the most important aspects being the composition set of genotypes in each trial. This is inherently more complex for breeding programs in hybrid crops. All of the above knowledge can aid the design of a training set of genotypes that will help achieve the best genetic gain.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications included in this thesis

1. Hunt et al. (2013) **Hunt, C. H.**, Smith, A. B., Jordan, D. R., and Cullis, B. R. Predicting additive and non-additive genetic effects from trials where traits are affected by interplot competition, *Journal of Agricultural, Biological and Environmental Statistics*, 18(1), 2013.

2. Hunt et al. (2018) **C. Hunt**, F. Eeuwijk, E. Mace, B. Hayes, and D. Jordan, Development of Genomic Prediction in Sorghum, *Crop Science*, 58(2), 2018.

3. Hunt et al. (2020) **Hunt, C. H.**, Hayes, B. J., van Eeuwijk, F. A., Mace, E. S., and Jordan, D. R. Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships, *Theoretical and Applied Genetics*, 133(3), 2020.

Submitted manuscripts included in this thesis

No manuscripts submitted for publication

Other publications during candidature

Conference abstracts

1. **Colleen Hunt**, David Jordan, Alison Smith and Brian Cullis, Predicting additive and non-additive genetic effects from trials where traits are affected by interplot competition, *Australasian Biometrics Conference*, 1-5 December 2013, Mandurah, WA

2. **Colleen Hunt**, Brian Cullis, Emma Mace and David Jordan, The accuracy of mixed models for use in genomic selection *27th International Biometric Conference*, 6-11 July 2014, Florence, Italy

3. **Colleen Hunt**, Brian Cullis, Emma Mace and David Jordan, Potential of Genomic Selection for Sorghum *15th Australasian Plant Breeding Conference*, 27-19 October 2014, St Kilda, Victoria

4. **Colleen Hunt**, Ben Hayes, Fred van Eeuwijk, Emma Mace and David Jordan, Investigation of GxE in genomic prediction for yield in sorghum using trials from multiple years *TropAG*, November 2017, Brisbane.

5. **Colleen Hunt**, Emma Mace, Ben Hayes, Fred van Eeuwijk and David Jordan, GxE for Genomic Prediction Models *Australasian Biometrics Conference*, 26-30 November 2017, Kingscliff, NSW.

6. **Colleen Hunt**, Ben Hayes, Fred van Eeuwijk, Emma Mace and David Jordan, Genomic prediction for grain yield in sorghum *International Sorghum Conference*, April 2018, Cape Town, South Africa.

Contributions by others to the thesis

Colleen Hunt has been the primary author of all thesis chapters while gratefully acknowledging the suggestions and edits for all chapters: David Jordan, Fred van Eeuwijk, Ben Hayes, Emma Mace and Alan Cruickshank.

Published papers in chapters 3, 4 and 5 were co-authored by David Jordan, Brian Cullis, Fred van Eeuwijk, Ben Hayes, Alison Smith and Emma Mace. The breakdown of contributions has been noted in the page preceding each chapter.

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis

Research involving human or animal subjects

No animal or human subjects were involved in this research

Acknowledgments

Firstly, I would like to express my sincere gratitude to my primary supervisor Prof. David Jordan for his continuous support of my PhD studies, for his patience, motivation and knowledge, without whom this thesis would not have been completed.

I am grateful to my supervisors Prof. Fred van Eeuwijk from Wageningen University and Prof. Ben Hayes from University of Queensland for always giving their time to give helpful and constructive support, guidance and feedback.

I acknowledge the support and supervision of Dr. Emma Mace for giving me inspiration and helpful advice throughout the journey and for her constructive feedback.

Mr. Alan Cruickshank for always inspiring me with his forever cheery disposition and also for his constant support and providing helpful and constructive feedback.

I acknowledge the support of Prof. Brian Cullis for giving a lot of input and inspirational guidance during the early stages of my PhD. He provided the initial concept for chapter 3 and gave me many things to think about.

To all of my biometrical family especially Dr Helena Oakey and Dr Katia Stefanova, always there, always helpful and always nice.

Finally but definitely not least I acknowledge my top fan David John Sendy (Jr) for riding the roller coaster with me.

Financial support

No financial support was provided to fund this research

Keywords

mixed models, spatial effects, pedigrees, additive effects, dominance effects, genotype by environment interaction, genomic prediction, training dataset, hybrid prediction, hybrid breeding

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 010402, Biostatistics, 75%
ANZSRC code: 070305, Crop and Pasture Improvement (Selection and Breeding) 25%

Fields of Research (FoR) Classification

FoR code: 0104 Statistics 75%
FoR code: 0703 Crop and Pasture Production 25%

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviation	Meaning
DAF	Department of Agriculture and Fisheries (Queensland)
QAAFI	Queensland Alliance for Agriculture and Food Innovation
GRDC	Grains Research and Development Corporation
DArT	Diversity Arrays Technology
REML	Residual Maximum Likelihood
BLUP	Best Linear Unbiased Prediction
GxE	Genotype by Environment Interaction
GS	Genomic Selection
GP	Genomic Prediction
LD	Linkage Disequilibrium
GCA	General Combining Ability
SCA	Specific Combining Ability
MAS	Marker Assisted Selection
GWAS	Genome-Wide Association Study
QTL	Quantitative Trait Locus
GEBV	Genome Estimated Breeding Value

Chapter 1

Introduction

1.1 Plant Breeding Background

Plant breeding will play an essential role in feeding approximately 10 billion people sustainably by 2050 (Desa, 2019). There is a need to increase productivity without increases in resources such as land and water use, fertilizer and chemicals. Advances in statistics, quantitative and population genetics, molecular biology, genomics phenomics, offer the potential of transforming plant breeding programs toward a data-rich, evidence-based, and team-oriented process and away from the romantic tradition of an individual breeder as an artist (Cobb et al., 2019). As our understanding of factors such as the environment and temperature increases we are more readily able to adjust for these in our predictions of phenotypic traits.

The parameters breeding teams manipulate as part of the crop improvement process can be eloquently expressed in an equation commonly known as *the breeder's equation* (Mühlenbein, 1997; Frankham et al., 2011; Cobb et al., 2019). The equation calculates the response to selection (R) by multiplying the additive genetic variation within the population (σ_a), selection intensity (i), and heritability (h^2) with the number of years per cycle (t) on the denominator.

$$R = \sigma_a i h^2 / t \tag{1.1}$$

Each of the parameters in this equation can be developed by fitting statistical models with greater prediction accuracy.

Decreasing the number of years is an obvious parameter that will result in an increase in genetic gain per unit time. The development of genomic selection protocols have demonstrated that breeding cycles can be shortened by selecting parents purely on the basis of genomically predicted breeding values (Heffner et al., 2009; Gaynor et al., 2017). The use of inbred lines as parents is arguably only a by-product of the need to phenotypically identify new parents. With the advent of genomic selection, the use of inbred lines as parents for the next cycle of recombination and selection could be eliminated entirely (Heffner et al., 2009). For genomic prediction and selection to realise these promised benefits, the impacts on other factors in the breeders' equation must be positive or at least not totally negating the positive impact of shorter cycle time.

Genetic variance

The first step in establishing a breeding pipeline is the selection of elite parents as founders of the program (Cobb et al., 2019). Elite germplasm can be defined as a set of genotypes enriched for favourable alleles that improve breeding value (i.e., the mean performance of the progeny of a given parent) in a particular environment. Breeding values are used regularly in the context of animal breeding since the breeding product is not a sire itself, but rather its progeny. A breeding value uses pedigree or genome-wide marker data to borrow information from related lines in a phenotypic data set to estimate the additive value of an individual. While a BLUP value for phenotypic performance accounts for both the additive and non-additive genetic values of a line, a breeding value uses the relationship matrix to determine the additive value of a line, which is the primary source of genetic variance passed on to its offspring (Henderson, 1976). This is critical information for parental selection decisions and determining the relative superiority of a line.

Selection intensity

Generating and testing more selection candidates while holding the number of selected candidates constant lead to higher selection intensity (i) which in turn increases the rate of genetic gain. Selection intensity can also be increased by selecting fewer parents; however, it is usually more advisable to determine the number of parents to select based on whether the objective of the breeding program is long or short term genetic gain (Bernardo and Charcosset, 2006). Thus, increasing i by way of increasing population sizes requires that either budgets be increased, or a reduction in the cost of testing each selection candidate.

Genomic selection (Meuwissen et al., 2001) could be used to increase the total number of selection candidates with a fixed budget if genotyping is less costly than phenotyping. Sparse testing designs, where individual lines are unreplicated or partially replicated across locations, but relatives are randomized among locations to allow estimates of haplotype x environment effects, can reduce the replication of selection candidates within and across environments. This reduces field costs and would allow a larger number of selection candidates to be tested (Endelman et al., 2014; Roorkiwal et al., 2018). Studies by Lorenz (2013) and Riedelsheimer et al. (2013) found that the application of genomic prediction generally led to greater response to selection because phenotyping all selection candidates, even at reduced levels of replication, increased both the intensity of selection and its accuracy (heritability).

Heritability (selection accuracy)

Phenotyping is the most expensive component of a plant breeding operation (Reynolds et al., 2018). The value of improvements in phenotyping is usually expressed by citing increases in broad sense heritability (H^2). Prediction accuracy increases as the value of trait heritability increases (Holland et al., 2003; Zhang et al., 2017), since genetic gain is proportional to the genetic accuracy, which is the square root of the narrow-sense heritability (h^2). This has big implications for deciding how

to invest a breeding program's limited resources. For most breeding programs, the simplest way to increase heritability is to better sample the targeted population of environments by increasing the number of yield trial locations. This turns out to be a very expensive option and is limited by physical capacity and partnerships as much as it is by budgets. Thus, most innovations in phenotyping for greater heritability have focused on extracting more information from existing yield trials.

1.2 Sorghum breeding in Australia

Grain sorghum (*sorghum bicolor* (L.) Moench) is the main summer grain crop in the north eastern Australia, it is used as feed grain to the beef, dairy, pig and poultry industries. An export market of around 1 Mt exists, particularly to Japan, but the average amount exported is in the order of 300-500 kt (GRDC, 2020). Sorghum is Australia's fifth highest grain export below wheat, barley, canola and chick peas.

Grain sorghum in Australia is a hybrid crop adapted for mechanised production. In hybrid sorghum a cytoplasmic male sterility system is employed that involves crossing female (cytoplasmic male-sterile) lines with male (fertility restoring) is used to produce F1 hybrid cultivars.

The sorghum pre-breeding program is a joint venture of the Department of Agriculture and Fisheries (DAF) and the Queensland Alliance for Agriculture and Food Innovation (QAAFI) (Jordan et al., 2011, 2013). The primary aims of the program are to select superior parents for the development of new genetic material and selection of hybrids for promotion to the next stage of breeding. The sorghum breeding program effectively has two breeding pools (heterotic groups) and evaluates selections from the two pools in two separate trial series. One series is used to predict male performance by making experimental hybrids with one or few females (called testers) with many different males to be evaluated and selected. The female series follows the same process but with one to few male testers and many females instead.

Breeding trials are grown in multiple environments for assessment of GxE interactions and specific or general adaptation of genotypes. Approximately 60% of the Australian crop is grown in Queensland and the remainder in northern NSW (GRDC, 2017). The area of sorghum planted for grain in northern NSW is on average 160,000 ha and Queensland 470,000 ha annually. The growing region spans approximately 1300km from northern New South Wales to Central Queensland (Figure 1.1). The DAF/QAAFI pre-breeding trials span this area with the northern most site is Kilcummin ($-22°11', 147°57'$) in central Queensland and the southern most site is Liverpool Plains ($-31°56', 150°47'$) in northern New South Wales.

Statistical analysis of these breeding trials is an essential part of the breeding and selection process in order to predict the best genotypes and/or the best parents to satisfy the purpose of the trials. Breeding hybrids such as sorghum necessitates the need for sophisticated statistical analysis techniques in order to extract the optimum predicted performance of the hybrids, their parents and their interactions (both interactions among genes and between genotypes and environments) included in each trial series.

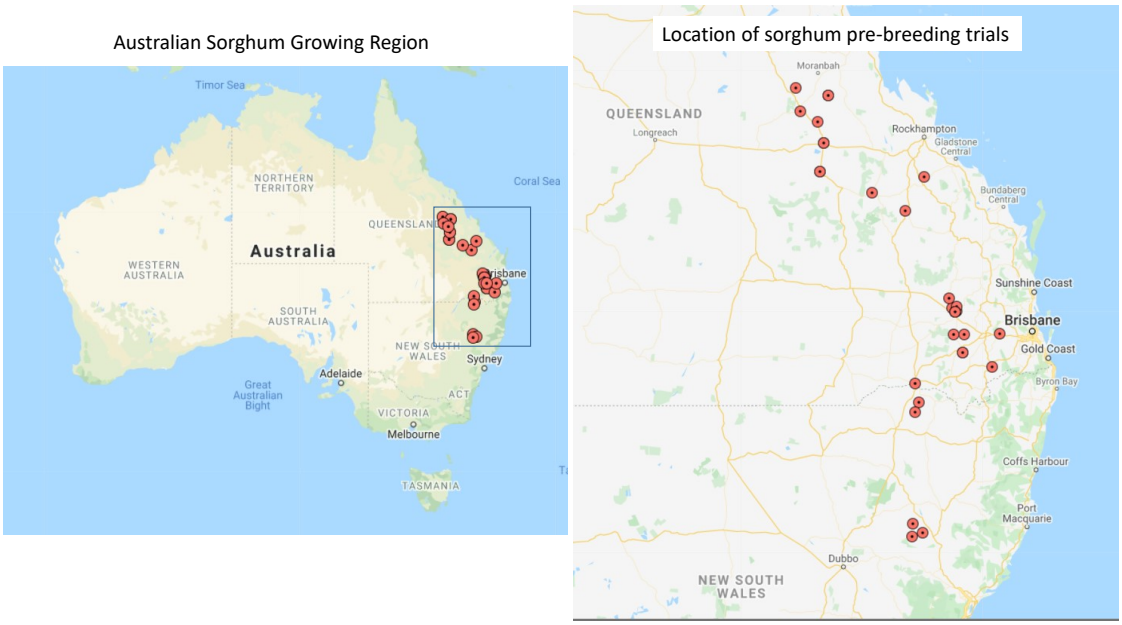

Figure 1.1: The Australian sorghum growing area and the trial locations for the DAF/QAAFI sorghum pre-breeding.

1.3 Structure of the thesis

The major aim of this thesis is in developing a greater understanding of factors leading to an increase in the response to selection, i.e. the genetic variance, selection intensity and heritability.

This thesis discusses the development and implementation of improved statistical analysis techniques to enable the comprehensive detection of both field trial effects and genetic effects. These techniques are required to gain better prediction of performance of sorghum genotypes and therefore a greater capacity to produce genetic gain in grain sorghum over time.

Following a review of the literature introducing concepts relating to molecular markers and statistical analysis methods for plant breeding trials, this thesis presents three published papers followed by a research chapter.

1. Incorporation of inter-plot competition effects into a model that includes pedigree information
2. Development of a fully functioning single stage multi environment analysis of genomic prediction using both pedigrees and markers
3. Investigating and discussing an efficient single stage linear mixed model for determining dominance effects over a large range of environments and the implications of GxE on dominance effects
4. Determining the most efficient use of testing resources allocated to early generation selection particularly where multiple testers can be used.

Chapter 2

Literature Review

2.1 Introduction

The aim of plant breeding is to create new genotypes that have improved phenotypic traits on the currently available genotypes through a cyclical process of crossing, evaluation and selection. Typically in each breeding cycle new sets of parents are identified based on information from phenotypic or genotypic evaluations. These new parents are inter-crossed to produce a large set of progeny or selection candidates that are evaluated in a series of trials. Based on their performance a set of these candidates are chosen to be used as new varieties or as parents of the next generation of selection candidates. In crops where hybrid cultivars are grown such as sorghum and maize, there is an additional step where the selection candidates from two different heterotic pools are crossed together. The phenotypes of resulting F1 progeny are used to evaluate the parents and also to identify new commercial varieties. In such schemes there is selection on both the average performance of lines as well as performance of specific hybrid combinations.

The following review outlines some of the current studies involved with implementing statistical methods for the improved analysis of plant breeding trials. Variation within trials is attributable to genetic and non-genetic sources. Non-genetic sources of variation occur as individual site specific sources of error due to design layout and plot positions. Genetic variation can be partitioned into additive, dominance and residual genetic components which allows the prediction of breeding values associated with parentage. Molecular markers and pedigree information can be incorporated into the analysis as a complimentary feature to the genetic parentage. Multiple trials can be combined into a single step multi-environment trial (MET) analysis that assesses the potential environment effects on each of the partitions of the genetic effects.

2.2 Plant breeding

Plant breeding uses principles from a variety of sciences to improve the genetic potential of plants. The process involves crossing parental plants containing different valuable genes to obtain the next

generation combining favourable characteristics from both parents (Finlay and Wilkinson, 1963; Wricke and Weber, 1986; Allard, 1999; Moose and Mumm, 2008). Breeders aim to improve their plants by making selections based on the performance of their data and also by using ancestral pedigree information and possibly more sophisticated genetic information based on DNA markers. Breeding involves the creation of genetically diverse populations, selection is performed with an aim to create new plants which have been adapted and perform well to specific desirable traits. The selection process is driven by the assessing performance in relevant target environments and using knowledge of genes and genomes. Progress is assessed based on gain under selection, which is a function of genetic variation, selection intensity, and time. There are three main aims of a plant breeding program, creation of new genetic material, selection of candidates to be used for further breeding and testing new candidates for future varietal release (Voss-Fels et al., 2019).

2.2.1 Hybrid breeding

A hybrid plant is formed by crossing two genetically different plants to produce hybrid progeny plants. In many cases such hybrids show superior performance to either of the parents, a phenomenon known as hybrid vigor (Shull, 1908) or heterosis. One theory on the cause of heterosis is the presence of dominance (Jones, 1917; Wright, 1934). The situation where all genes show dominance in the same direction is a phenomena called directional dominance. Accounting for dominance by way of a statistical model leads to the capability to directly predict hybrid performance and allow for selection of superior hybrids (Melchinger et al., 2007).

Many hybrid breeding systems are based around the inbred/hybrid model where pure breeding parental lines are produce through inbreeding, a process where a line is crossed with itself to produce offspring that have a high degree of homozygosity and therefore closely resembles its parents. Subsequently the resulting F1 hybrids produced by crossing these inbred lines are close to identical but heterozygous for the genes that differ between the parents and benefit from the associated heterosis. Heterosis is expected to increase with the genetic divergence between its parents (Melchinger, 1999).

In a plant breeding program it is possible to produce a large number of potential inbred parents resulting in a larger number of potential F1 hybrids. This combination problem has been simplified by plant breeders to some extent by grouping inbred parents into heterotic groups. Heterotic groups can be defined as sets of lines deriving from a common origin and displaying similar combining ability when crossed with lines from different origins. These heterotic groups are generally unrelated by pedigree and making crosses between them produce superior hybrids (Smith et al., 1999; Larièpe et al., 2017). Typically lines from one heterotic group are consistently used to make hybrids in combination with another heterotic pool. Cycles of selection of parents based on the performance in hybrid combination increase the complementary of these heterotic groups. The crosses made from lines from within the different heterotic groups, results in the thousands of potential between the heterotic group hybrids (Hallauer et al., 1988) which is beyond the capacity of most breeding programs to test. The problem of efficiently searching for elite combinations is therefore a major issue faced by most breeding programs.

2.2.2 General and specific combining ability

General combining ability (GCA) can be defined as the average performance of an inbred line based on the value that has been predicted by making crosses with other lines to form hybrids (Sprague and Tatum, 1942). In statistical terminology, GCA could be defined as the main effect of the inbred parental line of the hybrid and the interaction between the parents can be referred to as specific combining ability (SCA). In other terminology, GCA can be referred to as the additive effects, and SCA is the non-additive effects, which also encompasses the dominance effects.

Hallauer et al. (1988) considered a range of different strategies for corn breeding. They found that the correlations between inbred and hybrids performance for yield were generally low (less than 0.5). The complications are mostly due to the presence of dominance (Schrag et al., 2006). A method of measuring the value of lines in hybrid combination is needed.

The work of Sprague and Tatum (1942) supported use of a broad-base tester for preliminary screening for general combining ability, followed by testing in specific combinations. Hybrids are grown in two distinct trials using a scheme known as the North Carolina II mating design. In this design, each member of a group of parents used as males is mated to each member of another group of parents used as females (and vice versa) (Nduwumuremyi et al., 2013). The North Carolina II design is a factorial mating scheme used to evaluate inbred lines for combining ability.

2.2.3 Choice of parents of a hybrid

An important requirement of any successful hybrid breeding programme is the availability of efficient testers, which could effectively discriminate and classify inbred lines into appropriate heterotic groups for the development of high-yielding hybrids, see for example Dudley et al. (1991), Melchinger and Gumber (1998), Lee and Tollenaar (2007), Annor et al. (2020) and others. Crosses need to be made from inbreds that come from large heterotic groups and the number of possible crosses is the product of the number of inbreds within each heterotic group. This number of crosses is too large to realistically work with. To overcome this breeders make all crosses by crossing all inbred from one heterotic group with only a few tester lines from the opposing heterotic group. An effective tester should be able to rank inbred lines correctly for performance in hybrid combinations and efficient discrimination. One of the issues is to identify representative tester lines that can be used to accurately identify GCA and SCA.

Bernardo (1994) proposed a model for BLUP prediction where hybrid performance of a set of lines is combined with the genetic relatedness between a tested set of lines and an untested set of lines to predict untested hybrids. Bernardo (1994) found correlations between observed and predicted performance ranging from 0.65 to 0.80. He compared predictions based on genomic marker relationships with predictions based on pedigree relationships and found higher correlations for the marker based predictions. Models that result in BLUPs are useful for routine identification of single crosses prior to testing.

In a hybrid breeding program tester lines are chosen to enhance the objectives of the program and

the traits of interest. If the objective is to improve the performance of the population, then the testers should ideally contain a low frequency of favourable alleles at the loci where the population is in need of improvement. If additive gene action is of primary importance, then any tester will be effective. However, if dominance, is important the tester should be one that has a high frequency of recessive alleles at loci where improvement is needed (Dudley, 1997).

How many tester lines to use is a common problem when looking at general combining ability. The tester line could be good or bad depending on the environment and the hybrid combinations involved. This has implications for resource allocation, can we use less hybrids and still accurately predict the performance of the parent lines when crossed with a small number or an unbalanced number of tester lines. A good tester line should have high genetic variance to help facilitate a broad range of values for effective selections from the hybrid progeny. The impact of the environment and dominance is high across environments, therefore the best tester line may not be the same for all environments.

2.3 Molecular markers and their applications in plant breeding

Molecular markers enable detection of the variation between individuals at the level of the individual nucleotide sequence in the DNA. The development of molecular markers began in the 1980s with restriction fragment length polymorphism (RFLP) markers for construction of the first molecular map of the human genome (Botstein et al., 1980). Since the 1980s there has been a rapid increase in knowledge of plant genome sequences and the physiological and molecular role of various plant genes. This knowledge has revolutionized molecular genetics and its efficiency in plant breeding programmes (Paterson et al., 1991; Dudley, 1993; Lee, 1995; Staub et al., 1996; Mohan et al., 1997; Gupta and Varshney, 2000; Zamir, 2001; Moose and Mumm, 2008; Xu and Crouch, 2008; Crossa et al., 2010; Nadeem et al., 2018b). In the 1980s molecular markers could be used to detect and screen for genes making it possible to develop new breeding technologies (Tanksley, 1983). The detection of desirable genes is important for advancing the quality of newly produced genotypes. This information was used at that time to justify expensive and time-consuming genotyping activities

Since the 1980s, multiple different types of molecular markers have been identified and used in plant breeding applications including amplified fragment length polymorphisms (AFLP) (Vos et al., 1995), single nucleotide polymorphisms (SNP), randomly amplified polymorphic DNA (RAPD), micro satellites or simple sequence repeats (SSR) and diversity arrays technology markers (DArT) (Kilian et al., 2003). These new markers and new platforms have made use of technological advancements that have significantly reduced the cost per data point in comparison to the early implementation of markers in the 1980s (Ganal et al., 2019).

More recently advanced DNA sequencing technologies and platforms have produced new opportunities based on single-nucleotide polymorphisms (SNPs). High-throughput and large-scale genotyping of SNPs is now a routine tool in plant breeding in all major crop species including cereals (Rasheed et al., 2017). SNP genotyping has almost completely replaced other genotyping technologies due to their potential for high-throughput, high-speed data generation, repeatability, and cost effectiveness

(Kim et al., 2016; Ganal et al., 2019).

Applications of molecular markers include genetic diversity, mapping, marker assisted selection and genomic prediction. These methods have complemented breeding strategies by providing insight into the diversity of the genotypes used in crop improvement trials (Bazakos et al., 2017).

2.3.1 Diversity analysis

Genetic diversity refers to the genetic variability of a species, which can be quantified by a variety of metrics, e.g. genetic distances, population structure (Hokanson et al., 1998; Gilbert et al., 1999; Huang et al., 2002; Ferriol et al., 2003; Barkley et al., 2006). Molecular markers have been used in many studies to investigate genetic diversity and heterosis in plants (Xie et al., 2014). Molecular markers offer cost and time effective approaches to investigate the diversity in large germplasm collections (Lassois et al., 2016). In a plant breeding context, if a population of genotypes is diverse there will be a greater chance of finding specific genotypes that will be adapted to the required conditions, either an environment or a trait of interest.

Genetic diversity assessment is very helpful in the study of plant development using their genomic structure and genetic map. Genetic markers have been successfully applied in the determination of genetic diversity and the classification of genetic material (Naeem et al., 2015; Wang et al., 2017; Nadeem et al., 2018a). DArT and SNP markers are the most commonly used markers for the determination of genetic diversity in various crops (Baloch et al., 2017).

2.3.2 QTL analysis and genetic mapping and GWAS

Quantitative trait loci (QTL) are regions of DNA that can be associated with the phenotypic variation of complex traits such as yield (Kearsey et al., 1998). The location of the QTL is given relative to molecular marker positions. The detection of QTL is a method for determining chromosomal regions that influence particular traits in plants (Kearsey and Farquhar, 1998). A plant breeder can use this information to advantage by selecting for areas where there are favourable genes and speeding up the plant breeding process (Asins et al., 2010). QTLs are useful for indirect selection, where selections are made based on a marker effect instead of a trait of interest. This process allows for selection without phenotyping.

Methods for QTL mapping depend on the type of population and range from the simplest method of single-marker analysis, interval mapping, joint mapping, multiple regression and composite interval mapping. Structured populations include those where crosses are made between two parents and information is available for the parentage and the additive and dominance variation can be computed. Unstructured populations generally involve multiple parents that are not structured to fit a predefined distribution of genotypes.

More recently association mapping, which requires collections of germplasm instead of bi-parental populations, has also been developed as a method for identifying genes underlying quantitative traits. Genome wide association studies (GWAS) identifies significant associations between molecular

markers and a phenotypic trait and can be assessed by calculating the covariance between the within marker polymorphisms the phenotypic trait of interest (Jannink and Walsh, 2002; Zhang et al., 2016a). GWAS can be preferable to linkage mapping because there is no need to develop specialised populations, and a wide variety of different lines can be used, and hence has the capacity to capture greater genetic variation (Kraakman et al., 2004; Zhang et al., 2016b). GWAS is therefore considered more powerful but less accurate than bi-parental linkage mapping (Korte and Farlow, 2013). Nested association mapping (NAM), where multiple crosses are made between a single reference line and other diverse parental lines, combines the power of QTL mapping and association mapping, and represents a very useful resource for the dissection of genomic architecture of phenotypic traits and is being used increasingly in a wide variety of plant species (Yu et al., 2008).

Regardless of the approach used, QTL or GWAS analysis can be used to locate genes for traits of interest in specific circumstances. This type of work is essential in the improvement of complex quantitative traits of relevance to plant breeding programs. Many studies continue the process of improving plant breeding by using QTL and GWAS mapping for many traits, and for example in sorghum, over 6000 QTLs have been identified in many traits such as yield, maturity, height and integrated into the sorghum QTL Atlas (Mace et al., 2019).

2.3.3 Marker assisted selection

Marker assisted selection (MAS) involves the use of markers linked to phenotypic traits as an indirect selection tool for the trait without the requirement for phenotyping (Collard et al., 2005). MAS has been found to be most useful for traits that are not controlled by many genes and has been used as a breeding tool for the maintenance of recessive alleles during back-crossing and for expediting back-cross breeding in general (Bernardo et al., 2006). In sorghum, marker assisted selection has been used for a number of traits (e.g. stay-green, drought tolerance, cold tolerance and striga resistance) in recurrent selection and back-cross (introgression) breeding programs (see Ejeta and Knoll (2007), Burow et al. (2019) and others).

MAS has limitations however related to the cost of genotyping relative to phenotyping, the effect of the environment on the phenotype and the effect of other genetic components such as epistasis (Lande and Thompson, 1990; Ribaut and Ragot, 2007).

2.3.4 Genomic selection

Genomic selection (GS) is an approach that integrates molecular markers into the development of models for genetic evaluation, and was first developed by Meuwissen et al. (2001). It is a technique that has the ability to predict the genetic merit of genotypes using their genome-estimated breeding values (GEBVs) which have been predicted by using markers that cover the whole genome. GEBV is a prediction model that combines the phenotypic data with marker and pedigree data in order to increase the accuracy of prediction (Nadeem et al., 2018b). In this technique, genetic markers that cover the

whole genome are selected and utilized in a way that all QTLs are in LD with at least a single marker (Goddard and Hayes, 2007).

In the GS process for a particular trait, the breeding value of a parent and individual progeny is determined by summing the individual effects of each marker (based on high-quality phenotyping data from a training population). The advantage of this method is that the genetic gain per generation is estimated to be much higher than solely using phenotypic evaluation (Jonas and de Koning, 2016; Crossa et al., 2017). The statistical methods for the prediction approaches are discussed in section 2.5.

2.4 Statistical Approaches used for Plant Breeding

2.4.1 Introduction

Analysis of Variance (ANOVA) techniques have been by plant breeders used for over a century, since Fisher introduced the term variance in 1918 followed by the book entitled "The Design of Experiments" (Fisher, 1935). In the analysis of a plant variety trial, ANOVA is a method that partitions the total variation into sources due to varieties, environments, and variety by environment interaction and within-trial error variation. In the 1950s Henderson presented work on best linear unbiased estimates (BLUEs) and predictions (BLUPs) (Henderson, 1975) extending Fishers work. Henderson's work opened the door for the techniques that are commonly used today, in particular Henderson's set of 'mixed model equations' which were unsolvable until the REML algorithm (Patterson and Thompson, 1971). These so-called mixed models, that include both fixed and random effects, are preferable due to their capacity to handle missing data and estimate within-trial error variation (Smith et al., 2005). With the development of the statistical software package ASReml (Gilmour et al., 2009) and the R (R Core Team, 2018) package ASReml-R (Butler et al., 2009) we now have the required tools for undertaking complex mixed model analyses. These packages have been developed specifically for REML (Patterson and Thompson, 1971) estimation of mixed models and efficiently handle very large data sets across multiple sites. Smith and Cullis (2018) discuss the successful implementation of these procedures in the analysis of Australian plant breeding programs.

These analyses can be performed on quantitative traits that have been measured at the plot or plant level for any trial that has been designed as a rectangular array. The aim of these trials is to assess genotype performance as a pure stand measurement, that is the effect consisting purely from each stand alone genotype. Pure stand genotype predictions can be made after removing all of the non genetic effects such as field effects, effects due to neighbouring competition, Genotype by environment interaction effects and interactions with the other genotypes in the trial.

2.4.2 Field Trend

Since the 1920s scientists have been aware that the performance of genotypes are affected by their position in trials and have incorporated this into their designs (Fisher, 1935). Early designs focused on creating blocking and incorporating this into the analysis of experiments with general blocking as

treatments or replications. However, Papadakis (1937) observed that neighbouring plots behave in a similar way and hence correlated in their performance. This correlation is likely to decrease with distance. This idea was taken up in a series of studies including Wilkinson et al. (1983) Cullis and Gleeson (1991), Gilmour et al. (1997) and Besag and Higdon (1999). These papers discuss alternative ways of allowing for correlation or field trends. These effects can take the form of natural variation due to the neighbouring plots (known as the spatial location) and/or extraneous field variation that affects whole columns or rows, for example soil or moisture gradient and harvester/seeder effects (Gilmour et al., 1997).

Model-based analyses that focus on controlling spatial variation have been shown to result in substantial gains in response to selection (Cullis and Gleeson, 1991, 1989). Most of the current spatial approaches involve a direct modeling of trend using a correlation model, the basic premise being that plots that are closer together are more similar (more highly correlated) than plots that are further apart (Cullis and Gleeson, 1991, 1989). The approach of Gilmour et al. (1997) has been used successfully for the analysis of grain yield data from Australian cereal breeding programs for many years. Gilmour et al. (1997) recognised that it was necessary to incorporate both correlation and variation in the analysis of field trials where required. Gilmour et al. (1997) present diagnostics for assessing the presence of these components. A further extension is given by Stefanova et al. (2009) who describe analysis of individual trials using a method which includes terms in the linear mixed model to account for spatial variation and randomisation processes used in the design. Stefanova et al. (2009) introduce a new diagnostic process where the 3D variogram introduced by Gilmour et al. (1997) is displayed as a variogram slice complete with confidence intervals.

The methods of Gilmour et al. (1997) and Stefanova et al. (2009) involve user intervention in order to assess graphical diagnostics and in some instances there is a danger of over-fitting the spatial parameters. To address these issues a method was proposed by Velazco et al. (2017) who incorporated a method of two dimensional smoothing splines to model the natural spatial variation. Velazco et al. (2017) had success with the same linear mixed model approach as described about but replacing the auto-regressive variogram diagnostics with a spatial spline surface and gained model fitting flexibility.

2.4.3 Effects due to neighbouring competition

Neighbour competition effects can be defined as the effect that a neighbouring plant has on the effect of the phenotypic trait. This is a phenomenon that is present for many plant species (Keddy, 2001). The usual spatial models are not appropriate for traits measured in trials that exhibit inter-plot competition. An important example of this type in Australia is yield from sorghum breeding trials. Hunt and Jordan (2009) examined sorghum yield (in tonnes per hectare t/ha) for 36 such trials and found evidence of inter-plot competition in one third of those trials. They suggested that for this type of data a joint modeling approach that can accommodate both inter-plot competition and spatial trend is desirable.

The joint modelling of inter-plot competition and fertility trends has been discussed by Stringer (2006). Stringer (2006) analysed a number of early stage sugarcane trials and found that including

inter-plot competition into the random effects of a linear moxed model provided a good fit to the data in many cases. The analysis of yield from a hybrid that is not surrounded by hybrids of differing genetic background is called a pure stand yield effect.

2.4.4 Genotype by Environment interactions

Genotype by environment interactions (G×E) or differential genotype responses to types of environments cause re-ranking and complicate selection within breeding programs. Multi-environment trials (METs) are commonly used in an attempt to produce an across site genotype effect that represents the average performance of the genotypes across a sample of environments (e.g. Cooper and DeLacy (1994),Annicchiarico (2002), Malosetti et al. (2013)).

Smith et al. (2001), Smith et al. (2002a), Smith et al. (2002b) and Smith and Cullis (2018) detail the model fitting techniques that are commonly used in Australian plant breeding programs. They fit mixed models to series of trials as a single stage analysis that simultaneously allow for spatial effects at each site and structure the G×E. These techniques identify the extent and complexity of G×E with respect to providing the most accurate analysis of genotypes within multiple environments. The resulting system of genetic correlations can be used to group trials that produce similar genotype rankings.

In some cases large MET data sets extending over many years have been used to estimate the frequency of particular types of environments. Comstock (1977) defined the concept of a target population of environments (TPE) associated with a breeding program as the complete set of types of environments in which cultivars can be grown within the geographical area targeted by a breeding program. Cooper and DeLacy (1994) discussed some of the complexities involved with finding genotype effects by fitting (G×E) in statistical models. They highlighted the need for greater understanding of the causes of (G×E) and how to manage it. Characterising environments in such ways as described in Comstock (1977) and continued by authors such as Chapman et al. (2000b) and Chenu et al. (2011) who incorporated the TPE concept into their investigations.

2.4.5 Pedigree based genotypic relationships

The variation between genotypes is controlled by genetics, the environment and the interaction between genotype and environment. The genetics part of the variation between genotypes can be quantified in absolute terms as a genetic variance or expressed as a proportion of total variation estimated as the heritability of the trait in that population, where population is the set of genotypes in the trial. The genetic variation can be attributed to a fixed *additive* component, a fixed *dominance* component and the remaining variation due to the interaction between genes (Cockerham, 1954; Kempthorne, 1954; Falconer and Mackay, 1996). Analyses that utilise this partitioning of the genetic variance are used to estimate genetic effects, or breeding values, rather than genotypic values.

Underlying quantitative genetic theory makes the strong assumptions that genotypes under consideration can be traced back to the same idealised base population and that all genotypes in the

base population are unrelated and unselected (Nyquist and Baker, 1991; Holland et al., 2003). These assumptions are rarely met in breeding populations. Henderson (1975) and Im et al. (1989) concluded that the analysis assumes no selection and all genotypes stem from an ideal base population. If complete records are not kept results will be biased (Sorensen and Kennedy, 1984; Van der Werf and de Boer, 1990; Schenkel et al., 2002). It may be easy to conclude that pedigree based analysis is not necessarily valid for typical breeding trial analysis.

In the case of plants it can be considered that the F2 population is a random mating population (Wricke and Weber, 1986). Lynch and Walsh (1998) argue that likelihood methods partially account for biases due to selection because the pedigree relationship matrix corrects for generational information. Bauer et al. (2006) found that the pedigree analysis outperformed those from a genotype analysis without pedigrees in barley.

Oakey et al. (2006) proposed a mixed model for field trial data in which genetic effects are partitioned into additive and non-additive components using an additive relationship matrix whilst error variation is simultaneously modeled using the spatial techniques of Gilmour et al. (1997) and Stefanova et al. (2009). The extension of their process to include multi-environment trials is given in Oakey et al. (2007), Beeck et al. (2010) and Cullis et al. (2010). These papers fully describe the partitioning of the genetic variance into additive and non-additive parts. This method has been used extensively, see for example Burgueno et al. (2007), Piepho et al. (2008), Crossa et al. (2010), to name a few.

A pedigree relationship matrix known as the **A** matrix can be defined as a symmetrical matrix with diagonal elements $a_{ii} = 1 + F_i$ and off-diagonal elements $a_{ij} = 2f_{ij}$ where F_i is the inbreeding coefficient of entry i and f_{ij} is the coefficient of parentage between entries i and j (Henderson, 1976). The inbreeding coefficient is the proportion of similarity a genotype will have when crossed with itself. The coefficient of parentage is the genetic distance between two genotypes calculated as the sum of all the coefficients for all common ancestors between the genotypes.

2.4.6 Genotypic dominance

Computing possible dominance effects is particularly important for hybrid crops such as sorghum, where cultivars are hybrid combinations of female and male plant parents. Smith et al. (1990) showed that crosses with the highest yield were between unrelated inbred pairs with a coefficient of parentage of zero. The coefficient of parentage does not quantify the variation between siblings that are resultant from unrelated parents. Coefficients of parentage only quantify relationships between crosses and families of crosses with known pedigree relationships.

Environment can differentially affect the performance of inbred lines and hybrids, altering the relationship between genetic diversity and heterosis (Betran et al., 2003). Multi-environment trials (METs) can be analysed using a mixed model approach such as one developed by Smith et al. (2001). These approaches assess variety performance by considering all varieties as independent genetic lines without allowing for the fact that they may or may not be parentally related.

The previous section has highlighted the advantages of fitting models that allow for additive and non-additive genetic effects by partitioning the genetic variance component via the use of a relationship matrix **A** (Oakey et al., 2006; Piepho et al., 2008; Crossa et al., 2010). The accuracy of these pedigree models can be improved by the addition of a dominance effect (Mäki-Tanila, 2007). Oakey et al. (2007) and Dias et al. (2018) discuss partitioning the genetic effects into additive, dominance and residual genetic parts in particular partitioning for dominance in METs. The partitioning is performed by calculating relationship matrices for the additive and the dominance terms and incorporating them into the genetic variance structure of the fitted mixed model. Authors such as Mäki-Tanila (2007) and Oakey et al. (2007) have expressed concern over the computational difficulties involved with calculating a dominance matrix and it's subsequent inversion.

The calculation of the dominance relationship matrix can be obtained via a Monte Carlo simulation approach where repeated random sampling was used to approximate the dominance relationships (Hunt et al., 2011). A dominance matrix **D** can be added into the previous pedigree models. Oakey et al. (2007) present formulae for computing the elements of **D** and noted that unlike for **A** there is no computationally efficient algorithm for computing these elements or more importantly obtaining $\mathbf{D}^{-1}$. These formulae have not yet been implemented into ASReml-R due to this problem. Therefore an alternate approach based on simulation has been implemented in the ASReml-R package Pedigree. This computes the elements of **A**, **D** (as well as other relationship matrices) using IBD probabilities for a given number of simulations based on the known pedigree information. Sufficient accuracy was achieved using $N = 2000$ simulations (Hunt et al., 2011).

2.5 Genomic prediction

2.5.1 Statistical approaches

A number of statistical approaches have been proposed for GS including ridge regression (Whittaker et al., 2000), Bayesian approaches (Meuwissen et al., 2001) and kinship relationship methods (de los Campos et al., 2010; Scutari et al., 2013). Ridge regression (Whittaker et al., 2000) is a method which uses a mixed model where marker effects are fitted as random and assumes that all markers have an equal variance. Bayesian approaches (known as BayesA and BayesB) allow markers to have unequal variance (Meuwissen et al., 2001; Gianola et al., 2009). The kinship relationship methods use the markers to estimate a relationship matrix (such as a kinship matrix) which is then used to estimate a variance parameter (de los Campos et al., 2010).

These statistical methods essentially fall into two groups; firstly those that recognise the marker effects as "prior" information (ridge regression and Bayesian methods) and estimate the genotype effects by using the marker effects directly, and secondly the kinship type of methods that use the markers indirectly through a genetic relationship context.The first type of method has to address the problem of dimensionality imposed by the large numbers of marker effects that are required. Taylor et al. (2012) overcomes this problem by introducing a mixed model variable selection method which

simultaneously selects and estimates effects associated with a large number of potential covariates. The kinship method overcomes the problems associated with having to predict more effects then observations through using a mixed model which computes the genotype effects directly by computing their relationships using the markers and thus does not compute the marker effects *per se* (Burgueno et al., 2012). By utilising the relationship between the genotype BLUPs and the marker blups via the relationship matrix this type of GS methodology could theoretically incorporate many thousands of markers (Strandén and Garrick, 2009).

The relative importance of the genetic structure of the training and selection populations and the genetic relationships of individuals in each population influences the effectiveness of the statistical approach used to estimate GEBVs (Albrecht et al., 2011; Jannink et al., 2010; Habier et al., 2013; Riedelsheimer et al., 2013; Zhao et al., 2013; Massman et al., 2013). Different statistical approaches to genomic selection vary in their capacity to make use of LD and relationship information (Jannink et al., 2010; Zhao et al., 2013). Massman et al. (2013) used a single cross as a training population for maize and concluded that this was not an advantageous training set. Albrecht et al. (2011) and Jannink et al. (2010) suggest using multiple unrelated populations in their training population. Albrecht et al. (2011) reported prediction accuracies at the population level and found that predictions are more accurate for closely related populations. Both Albrecht et al. (2011) and Jannink et al. (2010) conclude that it is not appropriate to perform model training without taking information from related families.

The design of effective training populations has emerged as an issue of critical importance to the deployment of GS in applied breeding programs (Heffner et al., 2009; Jannink et al., 2010; Nakaya and Isobe, 2012; Habier et al., 2013). In contrast to the situation in animal populations, genotype by environment (G$\times$E) interactions play a much larger role in genotype performance and hence the requirement to design or sample appropriate training environments relevant to the target population of environments (Comstock, 1977) is of critical importance (Burgueno et al., 2012; Heslot et al., 2013).

2.5.2 Additive and dominance relationship matrices

As previously noted information related to pedigree information is restricted. Firstly it is restricted by the assumption that there was equal genetic contribution from each parent and secondly by the quality of the pedigree information available (Smith et al., 1990). Molecular markers contain additional information about the relatedness of individuals not contained in the pedigree, for example within family information due to Mendelian sampling. Authors such as Smith et al. (1990) and Betran et al. (2003) make the direct comparison of information given by pedigrees with the information given by markers. Betran et al. (2003) noted that the marker data classification was similar to their pedigree information, however Smith et al. (1990) found that the marker based relationships were more accurate. The idea of creating a genomic relationship matrices for the additive and the dominance components of genetic variance was investigated by VanRaden (2008); Goddard and Hayes (2009); Vitezica et al. (2013); Aliloo et al. (2016); Muñoz et al. (2014); Dias et al. (2018) to name a few.

While the calculations for the pedigree additive matrix may be straight forward, the calculation and

inversion of a pedigree based dominance relationship matrix typically requires a set of hybrids with many combinations of males and females (Aliloo et al., 2016). Using markers to create a genomic dominance matrix has been discussed by authors such as Su et al. (2012) and Vitezica et al. (2013). They give derivations of a genomic dominance matrix which can be easily calculated. The availability of large numbers of markers can compensate for the lack of balance by considering gene action at the individual marker level. Vitezica et al. (2013) states that the genomic version of the dominance relationship matrix is not comparable to the pedigree-based version. However for the purpose of selection the focus is on accurate predictions of the additive effects.

2.5.3 Genotype by environment

In the context of molecular markers there has been much research into the idea of QTL by environment (QTL×E) interaction where QTL effects may differ between environments. In recent years, mixed model frameworks have been used to detect QTL×E effects while modeling the variance-covariance matrix (Piepho, 2000; Verbyla et al., 2003; Malosetti et al., 2004; van Eeuwijk et al., 2005; Boer et al., 2007; Mathews et al., 2008; Verbyla and Cullis, 2012). Verbyla et al. (2003) fitted QTL×E effects as random, while others considered these effects as fixed. Piepho and Pillen (2004) show that mixed models provide a highly flexible multi-environment QTL modeling framework, with attention for incomplete blocking, heterogeneity of error variance, inclusion of standard varieties, genetic correlations between environments, and pedigree relations. A simulation study showed that modeling the variance covariance matrix within a mixed model framework was more powerful in detecting fixed QTL and QTL×E effects than when fixed models were used (Piepho, 2005). Mathews et al. (2008) compared QTL results from a mixed model analysis that incorporates G×E within the model, to QTL results from combining single site QTL analyses. Mathews et al. (2008) found that there was not too much difference in the techniques using their data. However Mathews et al. (2008) also used a two stage approach.

Model simulation studies provide great insight into the intricate maze of complexity which arises from genotype performance in multiple environments. Chapman et al. (2003) gave a first example of complete integration of crop and genetic simulation models to create genotypes and predict realistic yields and G×E in a breeding program over long-term (>20 to 50 years). Simulation of selection for yield illustrated phenomena, such as the preferential fixation of alleles associated with the most adaptive traits. Limitations are that the precise genetic models of input traits were largely unknown (and still are). Chenu et al. (2009) used a robust physiological model to integrate genetic variability observed empirically and simulate G×E for crop yield.

Given that genotypes and QTL vary with environment it is natural to consider the expansion of these methods to encompass genomic selection. As previously stated in section 2.5, G×E interactions play such a large role in genotype performance and hence are of critical importance (Burgueno et al., 2012; Heslot et al., 2013). When predicting genotype performance using genetic information where no phenotyping has occurred, such as the case of genomic selection, it is difficult to predict genotype

performance beyond the range of the observed phenotypic environments. Burgueno et al. (2012) assessed mixed model MET analyses including marker and pedigree effect and using a range of G×E structures. Burgueno et al. (2012) state that MET models can boost predictive power in across-environment prediction and also showed that modeling G×E using information on molecular markers and/or pedigree gives better prediction accuracy than not using molecular markers and pedigree information. Borgognone et al. (2016) and Tolhurst et al. (2019) recently discussed fitting MET models that incorporated genomic additive relationship matrices. They discuss the superiority of these models over their pedigree counterparts.

Further research is required to examine the prediction assessment of modeling the non-additive genetic variances such as dominance and epistasis and their interactions with environments. There is also a requirement to analyse across a broader range of environments to assess the future use of genomic selection.

2.6 Synthesis

This literature review contains many studies on the improvement of the statistical analysis methods of plant breeding trials. There are two important considerations to take into account for the continual improvement of genotypes. Firstly the factors that affect the genotype performance in the field trial. These factors such as spatial effects, trend, competition and G×E need to be allowed for in the analysis to get the best possible predictions of the phenotypic effects of the genotypes in the field trials. Secondly the genetic structure of the genotypes themselves have influence over genotype performance. These structures can be allowed for by incorporating ancestry information or molecular marker information.

The following published papers and research chapter will show research that fills the gaps covered by the presented literature. Chapter 3 extends the existing method for allowing for inter-plot competition to include pedigree relationships. Chapter 4 is the first application of genomic prediction in sorghum breeding using both genomic and genetic relationships. Chapter 5 extends the current methods of including genomic additive effects into a multi-environment trial analyses by incorporating dominance effects. Finally, Chapter 6 dicusses the use of multiple testers in hybrid breeding programs, comparing these results to those from trials that use a single tester.

The following publication has been incorporated as Chapter 3.

Hunt et al. (2013)

Hunt, C. H., Smith, A. B., Jordan, D. R., and Cullis, B. R. (2013). Predicting additive and non-additive genetic effects from trials where traits are affected by interplot competition. *Journal of Agricultural, Biological and Environmental Statistics*, 18(1):53–63.

Author	Statement of contribution	%
Colleen Hunt	writing of text	70
	proof reading	50
	Theoretical derivations	25
	numerical calculations	100
	preparation of figures	100
	initial concept	10
Alison Smith	writing of text	10
	proof reading	20
David Jordan	writing of text	10
	proof reading	10
	Supervision, guidence	50
	initial concept	20
Brian Cullis	writing of text	10
	proof reading	20
	Theoretical derivations	75
	Supervision, guidence	50
	initial concept	70

Chapter 3

Predicting additive and non-additive genetic effects from trials where traits are affected by interplot competition

3.1 Abstract

There are two key types of selection in a plant breeding program, namely selection of hybrids for potential commercial use and the selection of parents for use in future breeding. Oakey et al. (2006) showed how both of these aims could be achieved using pedigree information in a mixed model analysis in order to partition genetic effects into additive and non-additive effects. Their approach was developed for field trial data subject to spatial variation. In this paper we extend the approach for data from trials subject to interplot competition. We show how the approach may be used to obtain predictions of pure stand additive and non-additive effects. We develop the methodology in the context of a single field trial using an example from an Australian sorghum breeding program.

3.2 Introduction

It is widely recognised that data from plant breeding trials often exhibit spatial variation due to the spatial location of plots in the field. Model-based analyses that focus on controlling spatial variation have been shown to result in substantial gains in response to selection. Most of the current spatial approaches involve a direct modelling of trend using a correlation model, the basic premise being that plots that are closer together are more similar (more highly correlated) than plots that are further apart. One such approach is that of Gilmour et al. (1997) which has been used successfully for the analysis of grain yield data from Australian cereal breeding programs for many years. However, these models are not appropriate for traits measured in trials that exhibit interplot competition. An important example of this type in Australia is yield from sorghum breeding trials. Hunt and Jordan (2009) examined sorghum yield (in tonnes per hectare t/ha) for 36 such trials and found evidence of interplot competition in

one third of those trials. They suggested that for this type of data a joint modelling approach that can accommodate both interplot competition and spatial trend is desirable.

Stringer (2006) discussed a number of approaches for the joint modelling of interplot competition and fertility trends. In one of these approaches interplot competition was modelled using a random effects analogue of the Besag and Kempton (1986) treatment interference model (TIM). In this model an individual variety is assumed to have both a direct effect (on the plots in which it is grown) and a neighbour effect (on adjacent plots). In the random effects setting these are regarded as (correlated) genetic effects so that competition is modelled at the genetic level. Stringer (2006) analysed a number of early stage sugarcane trials and found that the random effects treatment interference model (R-TIM) (or a reduced rank version there-of) provided a good fit to the data in many cases. In terms of hybrid yield performance the key trait of interest is yield in a pure stand, that is the yield from a hybrid that is not surrounded by hybrids of differing genetic background. Predictions of hybrid effects for this trait are easily obtained from the R-TIM as a simple linear combination of the predictions for direct and neighbour effects.

In Australian sorghum breeding programmes the aim is primarily to develop new (fully in-bred) parental lines for commercial companies to use within their hybrid breeding programmes. Oakey et al. (2006) demonstrated that this aim is best met using a statistical analysis in which pedigree information is incorporated. Oakey et al. (2006) proposed a mixed model for field trial data in which genetic effects are partitioned into additive and non-additive components using an additive relationship matrix whilst error variation is simultaneously modelled using the spatial techniques of Gilmour et al. (1997). In this paper we propose an extension of the approach in Oakey et al. (2006) that incorporates an R-TIM to accommodate interplot competition. The resultant model enables the partitioning of pure stand genetic effects into additive and non-additive components. Here we consider the analysis of a single trial. Extensions for the analysis of multiple trials will be considered elsewhere.

The paper is arranged as follows. First we introduce a motivating example (Section 3.3). In Section 3.4 we present a sequence of statistical models for the analysis of a single field trial. We commence with a base-line analysis then build to an analysis that incorporates pedigree information and accommodates both spatial variation and interplot competition. Results of the application of these methods to the example are given in Section 3.5.

3.3 Motivating example

Our motivating example is taken from the Queensland Department of Agriculture, Fisheries and Forestry sorghum breeding programme. This programme runs what is analogous to two separate pedigree breeding programmes, one for female parents and one for male parents. All field evaluation of lines within these programmes is undertaken using F1 hybrids of combinations between the two pedigree programmes. We consider a trial grown in 2008 at the Hermitage Research Station in Warwick Queensland. The trial is a preliminary yield trial for males (PYTM).

The trial contained 791 F1 hybrids, comprising 783 test hybrids, being the result of a cross between

a single unreleased female and 783 male parental lines, 6 commercial F1 hybrids and 2 checks, being F1 hybrids close to release. The experimental design for the trial was a resolvable p-rep design (Cullis et al., 2006). Test and check F1 hybrids were sown in either one or two plots in the trial, while most commercial F1 hybrids had additional replication.

The sorghum breeding programme plants trials in a rectangular array of plots in which we notionally index plots by two factors, namely Rows and Columns. Plots are 1.5 $\times$ 10m comprising two plot-rows of plants. Plots which are row-neighbours (ie, within the same column) share the longest plot boundary. The prevalence of midge neccesitates the inclusion of spray-out rows to allow for access of spraying machinery. These spray-out rows occur at regular intervals across the trial, in pairs every 10 rows. Thus rows $(11,12),(23,24),\ldots,(12n-1,12n)$, where n depends on the total number of rows in the trial, will be spray-out rows. The PYTM trial we consider as an example consisted of 64 rows and 20 columns. The resolvable blocks were aligned so that block 1 occupied rows 1 to 31 and block 2 occupied rows 32 to 64, for all columns. The sizes of the blocks differed due to the occurrence of extra spray-out rows in block two.

The genetic design was determined by the aim. The aim of the PYTM trial is to select roughly 10% of the F4 male parental lines for promotion to Advanced trials. The PYTM trial represents the first opportunity for selection on yield and therefore the breeder is interested in both family and individual performance. A total of 783 F4 males were crossed with one female. The 783 F4 males were distributed across 48 full-sib families. The number of lines per family varied from 1 to 70 with an average of 17.4. In addition to the phenotypic data from the trial there was pedigree information on 1778 lines, including 61 founder lines (ie. lines with unknown parents). The average inbreeding coefficient of the F1 hybrids was 0.07, ranging from 0 to 0.24, while the average inbreeding of the ancestral lines was 0.985. The genetic connectivity in the design was high with an average additive correlation of 0.499 between the F1 hybrids. The availability of pedigree information is fundamental to the analysis that follows.

3.4 Statistical Methods

3.4.1 Excluding information on pedigrees

We begin by considering the analysis which does not use information on pedigrees. Our approach uses the enhanced spatial modelling ideas found in Stefanova et al. (2009). They describe an approach to the analysis of individual trials using a "hybrid" approach which includes terms in the linear mixed model to account for spatial variation and randomisation processes used in the design.

The model for data vector $\mathbf{y}^{n\times 1} = vec\,(\mathbf{Y}^{r\times c})$ can be written as

$$\mathbf{y} = \mathbf{X}\tau + \mathbf{Z}_{g_{do}}\mathbf{u}_{g_{do}} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e} \tag{3.1}$$

where the vectors $\tau, \mathbf{u}_{g_{do}}, \mathbf{u}_p$ represent fixed effects, random variety direct effects and random non-genetic (or peripheral, ie design and additional) effects respectively. The vec() operator stacks the

columns (here $1,2,\ldots,c$) of its matrix argument into a vector of length $n = rc$, where r is the number of rows in the trial and c is the number of columns. The additional subscript o and d, for the vector of direct effects $\mathbf{u}_{g_{do}}$ has been used to distinguish that this vector contains direct effects for entries which are in the data-set, as opposed to entries which are in the pedigree but are not in the data-set. We shall denote the vector of the latter direct effects by $\mathbf{u}_{g_{dp}}$ (see section 3.2) and we also introduce neighbour effects in section 3.3.

All random effects are assumed to follow a Gaussian distribution, with mean zero and each of the three random effect vectors are assumed pairwise independent. Variance models used for the random and residual effects are given by

$$\begin{aligned} var(\mathbf{u}_{g_{do}}) &= \sigma^2_{g_{do}}\mathbf{I}_{m_o}, \\ var(\mathbf{u}_p) &= \oplus_{l=1}^{b}\sigma^2_{p_l}\mathbf{I}_{q_l}, \\ var(\mathbf{e}) &= \mathbf{R} = \sigma^2\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r \end{aligned} \tag{3.2}$$

where we use $\mathbf{I}_n$ to denote an identity matrix of order n. m_o represents the number of hybrids present in the data-set. The symbol $\otimes$ denotes the Kronecker product and is defined for example in the appendix in Smith et al. (2005). The symbol $\oplus$ denotes the direct sum and is a shorthand method for expressing a block diagonal matrix. For example,

$$\oplus_{l=1}^{b}\sigma^2_{p_l}\mathbf{I}_{q_l} = \begin{pmatrix} \sigma^2_{p_1}\mathbf{I}_{q_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \sigma^2_{p_2}\mathbf{I}_{q_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \sigma^2_{p_b}\mathbf{I}_{q_b} \end{pmatrix} \tag{3.3}$$

The variance models in (3.2) allow for a genetic variance component $(\sigma^2_{g_{do}})$, a maximum of b random non-genetic terms, with the l^{th} term $(l = 1\ldots b)$ having q_l effects and an associated variance component $(\sigma^2_{p_l})$, a residual variance parameter (σ^2) and residual (scaled) covariance structure expressed as a Kronecker product of two (scaled) covariance matrices for the factors which enumerate the two dimensions of the field layout (typically called rows and columns; the factor rows is assigned, by default to the largest dimension of the array). The correlation structure is modelled using a first order separable autoregressive process (AR1) in each direction. The submatrices $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_c$ are the scaled correlation matrices for columns and rows respectively and are functions of vectors of unknown parameters denoted by $\boldsymbol{\phi}_r$ and $\boldsymbol{\phi}_c$ respectively.

3.4.2 Including information on pedigrees

The extension of (3.1) to include pedigree information has been described in Oakey et al. (2006) for single trials and Oakey et al. (2007), Beeck et al. (2010) and Cullis et al. (2010) for multi-environment trials. These papers fully describe the partitioning of the genetic variance into additive and non-additive parts. This method has been used extensively, see for example Burgueno et al. (2007), Crossa et al. (2010), Piepho et al. (2008).

In the following we present a brief overview of the models described by Oakey et al (2006) and Oakey et al (2007) but extend their models to explicitly account for the partitioning of the vector of (total) genetic effects into two sub-vectors. That is, if we let $\mathbf{u}_{g_d}$ be the vector of (total) genetic direct effects, then we assume that $\mathbf{u}_{g_d} = (\mathbf{u}_{g_{dp}}^T, \mathbf{u}_{g_{do}}^T)^T$. The vector $\mathbf{u}_{g_{dp}}$ is the vector of genetic direct effects of entries in the pedigree but not present in the data-set and as before the vector $\mathbf{u}_{g_{do}}$ is the vector of genetic effects for entries in the pedigree and present in the data-set. These vectors are of length m_p and m_o respectively and $m = m_p + m_o$.

We use the genetic model for $\mathbf{u}_{g_d}$ which assumes that

$$\mathbf{u}_{g_d} = \mathbf{u}_{a_d} + \mathbf{u}_{e_d} \tag{3.4}$$

where $\mathbf{u}_{a_d}$ represents the vector of entry additive genetic direct effects and $\mathbf{u}_{e_d}$ represents the vector of residual genetic direct effects. Each of these vectors are partitioned conformably with $\mathbf{u}_{g_d}$ with respect to the present/not present in the current data-set (the third suffix, viz p for "parent" and o for "offspring" present in the data-set) . Our model including pedigree information is derived by replacing $\mathbf{u}_{g_{do}}$ in 3.1 with $\mathbf{u}_{g_d}$ of 3.4 so is given by

$$\mathbf{y} = \mathbf{X}\tau + \mathbf{Z}_{g_d}(\mathbf{u}_{a_d} + \mathbf{u}_{e_d}) + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e} \tag{3.5}$$

where $\mathbf{Z}_{g_d} = [\mathbf{0} \;\; \mathbf{Z}_{g_{do}}]$

We assume that each of the vectors of genetic direct effects namely $\mathbf{u}_{a_d}$ and $\mathbf{u}_{e_d}$ are (pairwise) independent and are Gaussian with zero mean, with variance matrices $\sigma_{a_{dd}}\mathbf{A}$, and $\sigma_{e_{dd}}\mathbf{I}_m$.

The matrix $\mathbf{A} = \{a_{ij}\}$ is the relationship matrix and its elements are given by $a_{ii} = 1 + F_i$ and $a_{ij} = 2f_{ij}$ where F_i is the inbreeding coefficient of entry i and f_{ij} is the coefficient of parentage between entries i and j. The inbreeding coefficient is the percentage of similarity a genotype will have when crossed with itself. The coefficient of parentage is the genetic distance between two genotypes calculated as the sum of all the coefficients for all common ancestors between the genotypes.

All computations including the matrix $\boldsymbol{A}^{-1}$ are computed in the R (R Core Team, 2018) package ASRemL-R (Butler et al., 2009). The matrix $\boldsymbol{A}^{-1}$ is calculated using the asreml.Ainverse function which uses the algorithms of Meuwissen and Luo (1992) and Henderson (1976) with modifications to adjust for selfing. Details are given in an unpublished report (Gilmour, pers comm.).

3.4.3 Including information on pedigrees and competition

To allow for inter-plot competition in the row direction we incorporate the random effects treatment interference model (R-TIM) of Stringer et al. (2011). Each entry is assumed to have a direct genetic effect (for each of the components) on the plot into which it was sown and a neighbour effect on the adjacent row-neighbour plots. Hence (3.5) can be extended as follows

$$\mathbf{y} = \mathbf{X}\tau + \mathbf{Z}_g(\mathbf{u}_a + \mathbf{u}_e) + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e} \tag{3.6}$$

where $\mathbf{u}_a = (\mathbf{u}_{a_d}^T, \mathbf{u}_{a_n}^T)^T$, and $\mathbf{u}_e = (\mathbf{u}_{e_d}^T, \mathbf{u}_{e_n}^T)^T$, where the subscripts d and n represent the direct and neighbour effects respectively. The associated genetic design matrix is given by $\mathbf{Z}_g = [\mathbf{Z}_{g_d} \; \mathbf{N}_g\mathbf{Z}_{g_d}]$, where $\mathbf{N}_g = \mathbf{I}_c \otimes \mathbf{N}_r$ and $\mathbf{N}_r$ is the within row first order neighbour incidence matrix.

Stringer et al. (2011) proposed two variance models for the R-TIM. In the first, more general model, the variance matrices for the vectors of genetic effects are given by

$$\begin{aligned} var(\mathbf{u}_a) &= \begin{pmatrix} \sigma_{a_{dd}} & \sigma_{a_{dn}} \\ \sigma_{a_{dn}} & \sigma_{a_{nn}} \end{pmatrix} \otimes \mathbf{A} = \mathbf{G}_a \otimes \mathbf{A} \\ var(\mathbf{u}_e) &= \begin{pmatrix} \sigma_{e_{dd}} & \sigma_{e_{dn}} \\ \sigma_{e_{dn}} & \sigma_{e_{nn}} \end{pmatrix} \otimes \mathbf{I_m} = \mathbf{G}_e \otimes \mathbf{I_m} \end{aligned}$$

The second form for the R-TIM corresponds to the model of Draper and Guttman (1980) in which the neighbour effects are assumed to be a scalar multiple of the direct effects. In terms of our notation this leads to reduced rank forms (with rank 1) for the variance matrices $\mathbf{G}_a$ and $\mathbf{G}_e$. The reduced rank model is essentially a factor analytic model with the specific variances set to zero (see Chapter 5 for a description of the factor analytic model). This model is more succinct than the 2x2 matrices described by $\mathbf{G}_a$ and $\mathbf{G}_e$ above since it results in a 2×1 matrix of loadings representing the direct and neighbour effects. This model can be fitted in ASRemL-R using the algorithm described in Thompson et al. (2003).

All models in this paper were fitted using the ASReml-R package (Butler et al., 2009). This provides residual maximum likelihood (REML) estimates of the variance parameters, empirical best linear unbiased estimates (E-BLUEs) of the fixed effects and empirical best linear unbiased predictions (E-BLUPs) of the random effects.

It is important to note that the design of this trial did not allow for the genetic relationships and therefore there may be a chance that neighbouring plots contain hybrids that are genetically related. In this case the yields may display similarities that are not due to interplot competition but rather to the genetic relationship. This demonstrates that it is vital that the pedigree relationships be allowed for in order to assess competition effects and appropriate pure stand yields.

3.5 Results and Discussion

Table 3.1 presents the summary of the sequence of models fitted to the PYTM trial. Our analysis commenced by fitting a baseline model following the approaches recommended by Gilmour et al. (1997) and modified by Stefanova et al. (2009). This model included direct (D) effects for both additive and non-additive genetic effects, as well as a Block term to respect the resolvability of the design, and lastly used a separable first order autoregressive variance model for the residuals.

The base-line spatial analysis for the PYTM trial resulted in the estimated variance parameters as given for Model 1 in Table 3.2. The negative auto-correlation (-0.11) for the row dimension is indicative of the existence of interplot competition (Stringer and Cullis, 2002). A standard tool for examining the adequacy of an assumed spatial model is the graph of the sample variogram Gilmour et al. (1997). In order to focus on the effect of competition we restrict our attention to the slice of the variogram corresponding to zero column separation. This is given in Figure 3.1 for row separations up

Table 3.1: Summary of the models fitted to the PYTM trial. The notation RR() denotes the Draper and Guttman variance model for the terms in brackets, D - direct effects, N - neighbour effects. All models also include a random Block term.

Model	Add	Nonadd	Other	logl	Test	P-value
1	D	D		-453.02		
2	D	D	Row,Column	-430.80		
2a	D	D	Column	-450.83	M2a v M2	0.000
2b	D	D	Row	-434.93	M2b v M2	0.002
3	RR(D,N)	RR(D,N)	Row,Column	-420.48		
3a	RR(D,N)	D	Row,Column	-422.41	M3a v M3	0.049
3b	D	RR(D,N)	Row,Column	-426.67	M3b v M3	0.000

Table 3.2: REML estimates of variance parameters (standard errors in parentheses) from three models fitted to PYTM data. Model 1: base-line spatial with pedigree information; Model 2: base-line spatial with pedigree information plus random row and column effects; Model 3: joint spatial and competition with pedigree information. Genetic parameters are above the line and non-genetic below. $\sigma^2_{p_1}$, $\sigma^2_{p_2}$ and $\sigma^2_{p_3}$ are the variance components for blocks, columns and rows respectively.

Variance parameter	Model 1 estimate	Model 2 estimate	Model 3 estimate
$\sigma_{a_{dd}}$	0.281 (0.140)	0.239 (0.121)	0.137 (0.112)
$\sigma_{a_{nn}}$			0.020 (0.026)
$\sigma_{a_{dn}}$			-0.052 (0.030)
$\sigma_{e_{dd}}$	0.241 (0.067)	0.248 (0.062)	0.120 (0.060)
$\sigma_{e_{nn}}$			0.006 (0.013)
$\sigma_{e_{dn}}$			-0.027(0.012)
σ^2	0.630 (0.056)	0.525 (0.050)	0.449 (0.045)
$\sigma^2_{p_1}$	0.312 (0.446)	0.307 (0.445)	0.304 (0.449)
$\sigma^2_{p_2}$		0.017 (0.010)	0.016 (0.010)
$\sigma^2_{p_3}$		0.092 (0.026)	0.087 (0.026)
ϕ_c	0.17 (0.046)	-0.01 (0.054)	0.05 (0.056)
ϕ_r	-0.11 (0.049)	-0.18 (0.052)	0.12 (0.088)

to 15. In the case of spatial trend (that is, with a positive auto-correlation) this graph should increase smoothly to a plateau. However the large spike at a row separation of one in Figure 3.1 (a) means that adjacent plots (one row apart) have a higher semi-variance than those that are further apart. This suggests that adjacent plots have a negative effect on each other.

Figure 3.1 (panels (a) and (d)) present the diagnostic plots suggested by Stefanova et al. (2009). These are the row and column faces of the sample values of the empirical semi-variogram of the residuals from model 1 in Table 3.1. These plots are augmented with the mean and 95% point-wise coverage intervals of the faces of the empirical semi-variogram from a parametric boostrap sample of size 100. This proceedure is fully descibed in Stefanova et al. (2009), essentially the current model is simulated 100 times using the current variance components and the sample variogram is calculated for each simulation. The 2.5% and 97.5% percentiles are obtained and included in figure 3.1. There are clear and systematic discrepancies between the mean row and column faces of the parametric bootstrap

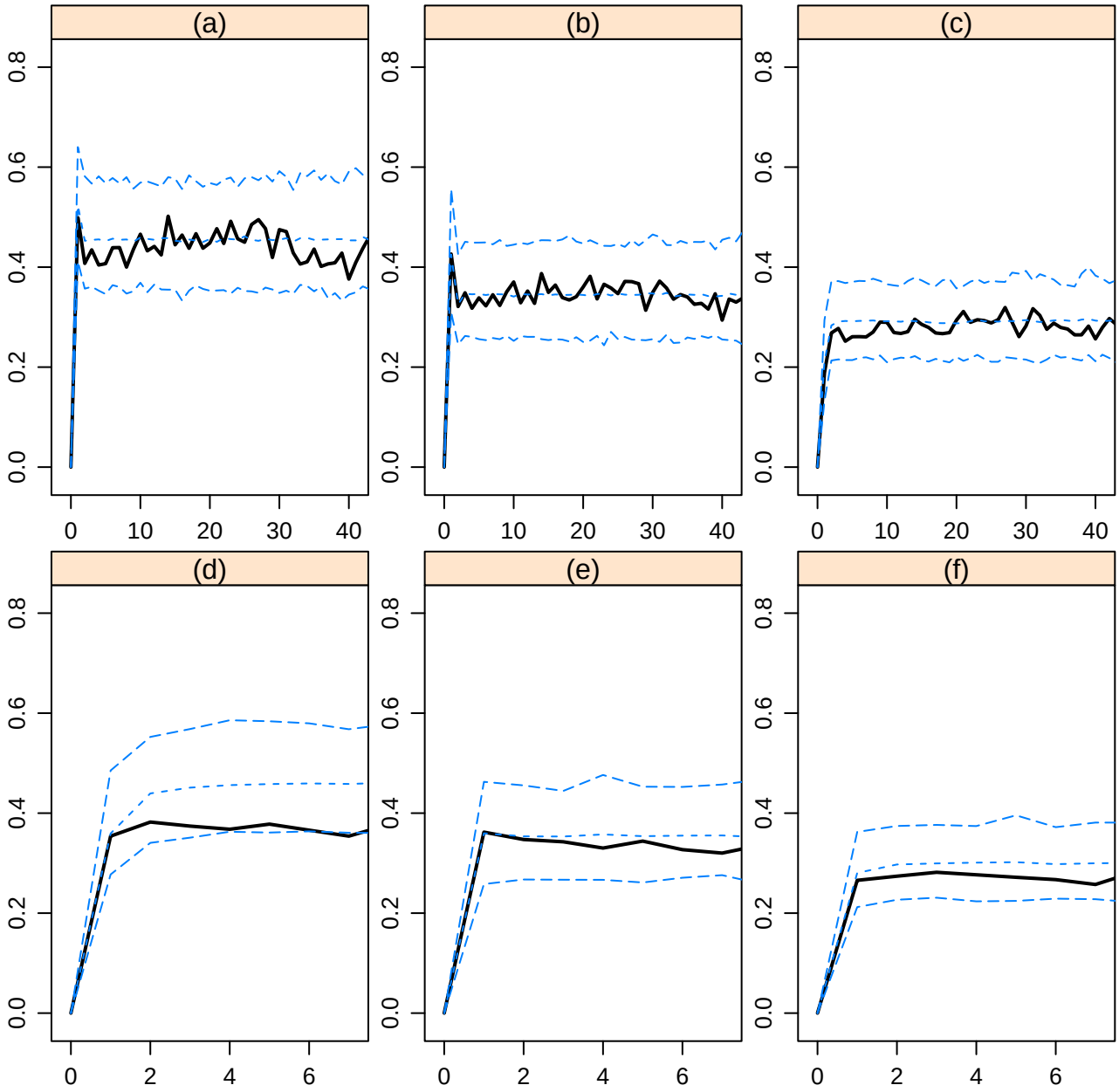

Figure 3.1: Plots of the row and column faces of the empirical semi-variogram for the residuals (solid line) for the PYTM trial from model 1 (panels (a) and (d)), model 2 (panels (b) and (e)) and model 3 (panels (c) and (f)). These plots are augmented with the mean and 95% point-wise coverage intervals of the row and column faces of the empirical semi-variogram from a parametric bootstrap sample of size 100

sample and the residuals from model 1. In both figures the mean is generally higher for all lags. This indicates the presence of both row and column effects.

Model 2 investigates this possibility by including random effects for both rows and columns. There is a substantial increase in the residual log-likelihood for model 2 over model 1. Models 2(a) and 2(b) drop the Row and Column terms respectively to formally test the need for these terms. Both terms are

deemed significant ($p < 0.05$) using Residual Maximum Likelihood Ratio Tests (REMLRTs). Figure 3.1 (panels (b) and (e)) present the diagnostic plots for the residuals from model 2. There is generally good agreement between the row and column faces of the empirical semi-variogram with the mean of the parametric bootstrap sample.

The noteworthy feature of these plots is the presence of a sharp "spike" at lag one for the row-face of the empirical semi-variogram. The REML estimate of the row autocorrelation parameter for model 2 was -0.18. This suggests that there is competition present in this direction (ie between neighbouring plots within the same column, sharing a common long boundary).

Our approach for modelling this (apparent) competition is to fit the reduced rank version of the R-TIM to both terms, denoting this by RR(D,N).

Model 3 provided a substantial improvement in fit over model 2, with both non-additive and additive competition deemed significant (using a REMLRT for models 3a and 3b vs model 3 respectively).

The diagnostic plots of the empirical semi-variogram of the residuals from model 3 are satisfactory (panels (c) and (f) in figure 3.1). Note that the level of these plots has dropped quite appreciably from the previous model (panels (b) and (e)), reflecting the amount of variation explained by the competition effects, this is also reflective in the reduction of σ^2 (see table 3.2). Also note that the large spike in the row-face of the empirical semi-variogram has been removed. The REML estimate of the row autocorrelation parameter for this model was 0.12, compared with -0.18 for model 2 (table 3.2).

Figure 3.2 presents a plot of the top 10% of the empirical BLUPS (E-BLUPs) of the pure stand yield ($\tilde{\mathbf{u}}_{a_d} + 2\tilde{\mathbf{u}}_{a_n}$) from model 3 versus the E-BLUPs of the direct effects for model 2 for the F4 male parents. The simple correlation coefficient, displayed in the top right panel shows a correlation of 0.56 for the top 10% of the E-BLUPs from these two models. This suggests that the selection of male parents from each model is noticeably different, in fact the top 10% of the E-BLUPs from both models only have 77% of the male parents in common. Additionally, the E-BLUPs of the pure stand effects are substantially smaller in magnitude than the E-BLUPs of the direct effects. This is due to the negative relationship between the direct and neighbour effects.

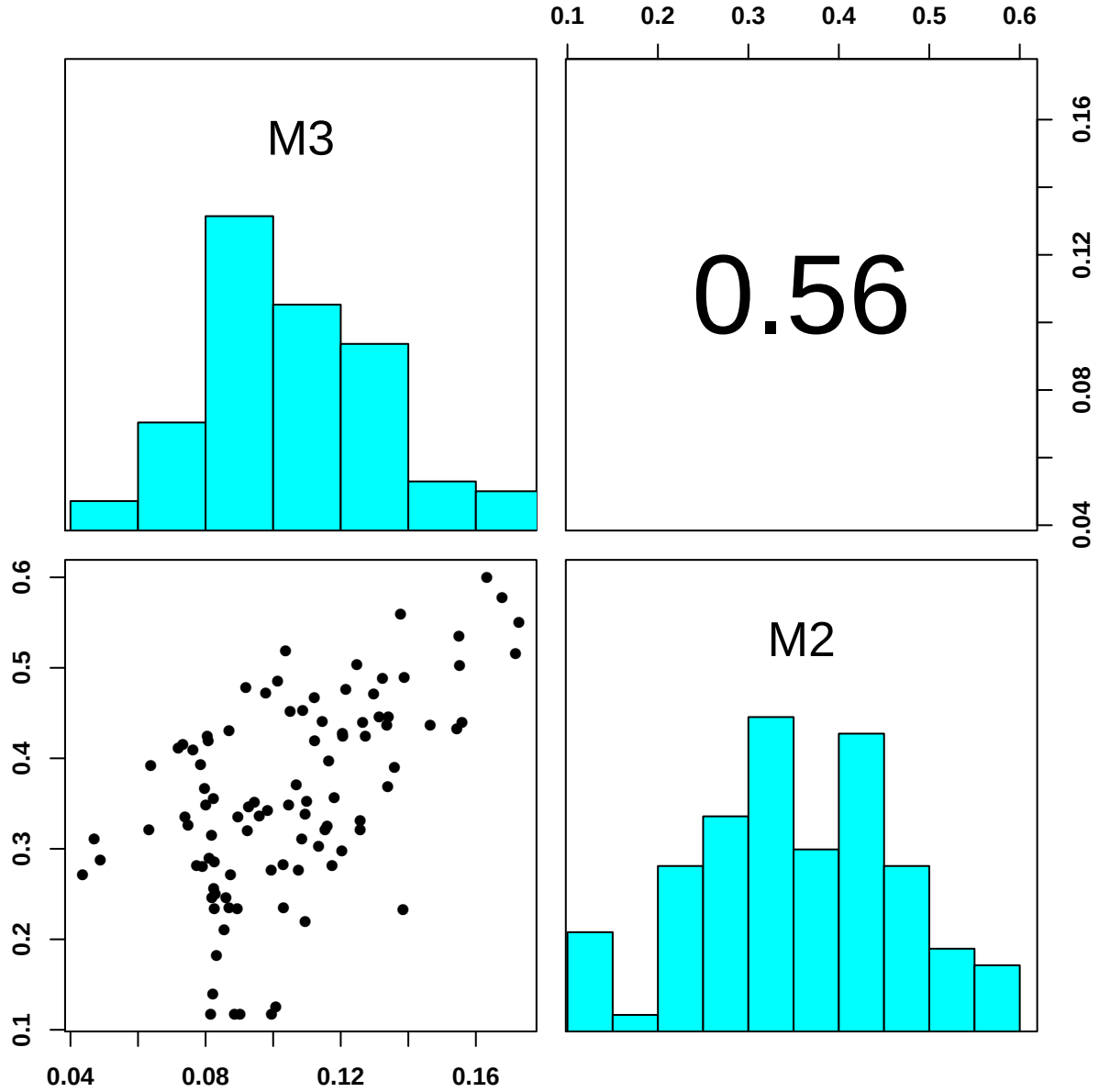

Figure 3.2: Pairwise scatter plot (lower left), simple correlation coefficient (upper right) and histograms (diagonals) of the E-BLUPs of the pure stand effects from model 3, and the E-BLUPs of the direct effects from model 2 for the top 10% of additive effects for the F4 male parents in the PYTM trial.

The following publication has been incorporated as Chapter 4.

Hunt et al. (2018)

Hunt, C., Eeuwijk, F., Mace, E., Hayes, B., and Jordan, D. (2018). Development of genomic prediction in sorghum. *Crop Science*, 58(2).

Author	Statement of contribution	%
Colleen Hunt	writing of text	70
	proof reading	40
	Theoretical derivations	80
	numerical calculations	100
	preparation of figures	100
	initial concept	40
Fred van Eeuwijk	writing of text	10
	proof reading	20
	Supervision, guidence	15
	Theoretical derivations	10
Emma Mace	writing of text	5
	proof reading	10
	Supervision, guidence	15
Ben Hayes	writing of text	5
	proof reading	10
	Supervision, guidence	25
	Theoretical derivations	10
David Jordan	writing of text	10
	proof reading	20
	Supervision, guidence	45
	initial concept	60

Chapter 4

Development of genomic prediction in sorghum

4.1 Abstract

Genomic selection can increase the rate of genetic gain in plant breeding programs by shortening the breeding cycle. Gain can also be increased through higher selection intensities, as the size of the population available for selection can be increased by predicting performance of non-phenotyped, but genotyped, lines. This paper demonstrates the application of genomic prediction in a sorghum breeding program and compares different genomic prediction models incorporating relationship information derived from molecular markers and pedigree information. These models were used to predict yield performance of genotypes from early stage sorghum breeding trials grown in four contrasting environments in Australia. In cross validation, the models using marker based relationships had higher selection accuracy than the selection accuracy for models that used pedigree based relationships. It was demonstrated that genotypes that have not been included in the trials could be predicted quite accurately using marker information alone. The accuracy of prediction declined as the genomic relationship of the predicted individual to the training population declined. We also demonstrate that the accuracy of genomic breeding values from the prediction error variance derived from the mixed model equations is a useful indicator of the accuracy of prediction. This will be useful to plant breeders, as the accuracy of the genomic predictions can be assessed with confidence before phenotypes are available. Four distinct environments were studied and shown to perform very differently with respect to the accuracy of predictions and the composition of estimated breeding values. This paper shows that there is considerable potential for sorghum breeding programs to benefit from the implementation of genomic selection.

4.2 Introduction

Sorghum is an important source of animal feed and forage in many areas of the world and is a staple food for half a billion of the world's poorest people. In most parts of the world productivity gain has been slow (FAO, 2009). It is critically important to improve the productivity of this crop in order to meet the world's need to double food production by 2050 in order to meet the demand generated by population growth and increasing affluence (FAO, 2009). Unfortunately the development of new sorghum varieties, incorporating positive alleles for key performance traits for different environments and different end-uses, is a long process. For example, the average time from the initial cross to preliminary yield testing within sorghum breeding programs is approximately 6 years for male parents and up to 8 years for female parents. Another 2-4 years are required before these lines are identified as parents for another cycle of selection (Rizal et al., 2014).

Genomic Selection (GS) (Meuwissen et al., 2001) is a technology that has been used widely in animal breeding, and is increasingly being used in plant breeding because of its potential for improving the rate, and reducing the cost, of achieving genetic gain per unit of time (Jannink et al., 2010; Habier et al., 2013; Daetwyler et al., 2013). The concept behind genomic selection is simple; genotype effects are estimated via the use of markers distributed across the genome.They are firstly estimated in a training population of representative individuals that are both phenotyped and genotyped. These data are subsequently used to create a prediction model which is then applied to new, non-phenotyped samples of the target population of individuals producing genomic estimated genotypic values which can then be used for selection.

If the cost of genotyping is lower than that of generating phenotypic data and prediction accuracy is sufficiently high, then this method may allow more selection candidates to be screened (than with phenotypic selection for example), increasing the intensity of selection (Heffner et al., 2009). More importantly, for increasing the rate of genetic gain, is the potential to conduct multiple cycles of selection without the need to phenotype, thereby enabling generation times to be substantially reduced and potentially increasing genetic gain per unit time and per unit cost (Heffner et al., 2009; Heslot et al., 2013).

In plant breeding, substantial gain can be achieved with the capacity to conduct selection without the need to produce pure lines by inbreeding. Gain is also achieved with the ability to conduct multi-location field trials to adequately sample the target population of environments (Heffner et al., 2009). GS has been assessed in wheat (Lopez-Cruz et al., 2015) and maize (Beyene et al., 2015), see Jonas and de Koning (2016) for a review of the implementation of GS in both crops and animals.

GS models exploit two types of information contained in the training data. Firstly they use markers that are in strong linkage disequilibrium (LD) with quantitative trait loci (QTL) controlling the traits of interest to select superior individuals (Fernando et al., 2007). This is analogous to conventional marker assisted selection, but requires the use of statistical methods to deal with the complexities that arise from attempting to predict marker effects when the number of markers greatly exceeds the number of phenotype observations. The second source of information from the training data is the relationships

between individuals. In practical applications of GS the individuals in the training population will be related to those in the selection population (Habier et al., 2007). In this circumstance, markers used for prediction will also capture additive genetic relationships between individuals (Ritland, 1996; de los Campos et al., 2009; Fernando, 1998; Habier et al., 2007). This information contributes to prediction accuracy in the same way as pedigree information is used to allow for non-independence of individuals included in a breeding trial (Oakey et al., 2007; Piepho et al., 2008; Crossa et al., 2010; Burgueno et al., 2012). In addition to pedigree based relationship information, genetic relationships based on markers are able to capture Mendelian sampling within families.

Mixed model methods have been proposed for genomic prediction such as ridge regression (Piepho et al., 2012), where markers are fitted as random effects, and relationship methods (VanRaden, 2008; de los Campos et al., 2010; Scutari et al., 2013), where markers are used to calculate a relationship matrix which is used to estimate a variance parameter. Genomic relationship based methods overcome problems associated with having to predict more effects than observations through using a mixed model which has the capacity to predict the genotype effects directly from the parameters in the model (Burgueno et al., 2012).

In this paper we present a linear mixed model analysis for prediction of genetic effects of individuals grown in breeding trials from a sorghum pre-breeding program operated by QAAFI, DAF and GRDC. Prediction models that involve pedigree and marker relationship matrices were compared in order to identify the most accurate prediction model using cross validation based selection accuracy. The resultant models were used to investigate the prediction accuracy for different families and the factors influencing these accuracies. Factors include relatedness between families based on pedigrees or markers and the number of progeny within each family. Cross validation was carried out by excluding data from whole families in order to assess the prediction accuracy of non-phenotyped lines. We discuss strategies for compiling training populations to be used for genomic selection in this crop.

4.3 Materials and Methods

4.3.1 Genetic materials and phenotypic data

The data for this study were from a set of preliminary yield trials grown in Queensland in 2008 as part of the sorghum pre-breeding program. Three of the trials were located in southern Queenland ($-28°20', 152°10'$) and one trial in central Queensland ($-24°24', 150°51'$) (see Figure 1.1 for a map of the Australian sorghum growing area). These trials represent the first yield testing stage of an integrated pre-breeding program where grain yield is measured in hybrid test cross combination with a single cytoplasmic male sterile tester female. Selected lines tested in hybrid combination from the first stage of testing are then evaluated in advanced trials in combination with multiple females at additional locations and in multiple years. Sorghum environments in Australia are highly variable. The trials used in this study are typical of the major sorghum growing environments in Australia.

Grain yield was collected from trials at four locations. The number of F_1 hybrids per trial ranged

from 738 to 791 (Table 4.1). The experimental design for each trial was a partially replicated design (Cullis et al., 2006) where the replicated entries can be resolved into two equal blocks. The hybrids in the breeding trials consisted of 6 commercial F_1 hybrids and 2 check hybrids, the remaining hybrids were the result of a cross between a single cytoplasmic male sterile female parent and between 730 and 783 F_4 male parental lines (different across trials). The experimental male parents resulted from 45 individual bi-parental crosses (hereafter referred to as families). The families included crosses between elite inbred lines with considerable shared ancestry as well as crosses between elite inbred lines and diverse germplasm not known to be related to individuals in the breeding program. A large population of F_2 plants were produced from each F_1 cross and particular plants were selected and advanced by single seed descent with some selection for maturity and height to the F_4 generation. The average number of progeny per family was 17 ranging from 1 up to 70 F_4 progenies per family. Experimental and check F_1 hybrids were sown in either one or two plots in the trial, with around 30% of the hybrids having two plots while most commercial F_1 hybrids had additional replication.

Table 4.1: Description of the field trials used in the analysis; including site mean yield (in t/ha), the total number of F_1 hybrids and the number of genotyped lines.

Site	Mean Yield	Total number of F_1 lines	Genotyped male parental lines
Biloela	2.38	780	537
Dalby	6.90	765	526
Dalby Box	6.29	738	506
Hermitage	10.48	791	544

This study focused on a subset of 544 genotyped lines from 31 families, all lines are homozygotes (Supp. Table S1). Of the total set of 791 unique male parental lines included in the trials, only 544 had genotypic data. Resources available for genotyping were limited and a number of the small families were excluded, along with lines from families with parents that had limited pedigree information. The material in the trials will be hereafter referred to as lines. The lines were grown in hybrid combination (testcross), with all lines being crossed to a single female tester. The program is a pre-breeding program focused on developing germplasm lines with high levels of general combining ability for yield and the female parent used to produce the test crosses was selected to represent typical female germplasm used in Australia.

4.3.2 Pedigree Data

Ancestral pedigree information was available for all 544 lines for up to 20 generations of ancestry (Supp. Figure S1). In total there were 443 unique ancestral lines included in the full pedigree file, including 61 founder lines with unknown parents. With the inclusion of the 544 lines present in the trials, the number of lines in the pedigree file totalled 987. The average inbreeding coefficient of the lines was 0.07, ranging from 0 to 0.24, the genetic connectivity in the design was high with an average additive correlation of 0.499 between the lines (Supp. Figure S1). The 544 lines used in this study had families with between 4 and 34 progeny (full-siblings) and each parent was used in 1 to 8 crosses

(Supp. Table S1). Full pedigree information was available for all the parents of the families except 2 parents (PI563516 and PI609489). Ten of the families were produced from crosses which had one parent that was unique to that cross. All lines within a family were full siblings but they also shared at least one parent (ie were half siblings) with one or more of the other families. The total number of full and half-siblings for each family ranged from 36 (R04127) to 228 (R04330).

4.3.3 Marker Data

DNA was extracted from the progeny of the 31 families using a modified CTAB-based extraction protocol, as detailed in Parh et al. (2008). The progeny, which constituted the male parental line from the hybrids, were genotyped with DArT markers, as detailed in Mace et al. (2008). Since all 544 lines are homozygotes, the dominant properties of DArT markers does not pose a problem.

In total, 581 DArT markers were polymorphic across the 544 male parental lines. The number of polymorphic DArT markers per chromosome ranged from 34 to 77 (Supp. Table S2), spanning 94% of the consensus map coverage (Mace et al., 2009). All redundant markers that were mono-morphic across the full population were removed. Overall, the amount of missing data was low; ninety percent of the markers had less than 5% missing values. The maximum missing data frequency per marker was 0.11, with an average of 0.03. LD between loci was calculated across the full population using a Pearson coefficient of correlation (Supp. Figure S2) and average LD within each linkage group ranged from 0.31 (LG5) and 0.49 (LG7).

Despite the relatively low marker density we have an average of a marker every 3cM and on average LD declines by 50% within 12cM. The level of marker density we used in this analysis is similar to that used by a number of studies (Habier et al., 2009; Wellmann et al., 2013; Zhang et al., 2015).

4.3.4 Statistical Models

The predicted performance of lines was analyzed using a mixed model approach as detailed in Stefanova et al. (2009). This was an integrated approach to analysing individual trials which included terms in the linear mixed model to account for spatial variation and randomisation processes used in the design. The extension of these models to include pedigree information has been described in Oakey et al. (2006) and Hunt et al. (2013). These papers fully describe the partitioning of the genetic variance into additive and non-additive parts.

A mixed model for a single trial where the vector $\boldsymbol{y}$ represented the phenotypic yield arranged as trial rows within trial columns can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_h\boldsymbol{u}_{h_g} + \boldsymbol{Z}_o\boldsymbol{u}_o + \boldsymbol{e} \tag{4.1}$$

The vectors $\boldsymbol{\tau}, \boldsymbol{u}_{h_g}, \boldsymbol{u}_o$ represent fixed effects, random effects for lines and random non-genetic (or peripheral, ie design and additional) effects respectively. $\boldsymbol{X}$, $\boldsymbol{Z}_h$, and $\boldsymbol{Z}_o$ are the design matrices for

the fixed effects, random genetic effects, and the random nongenetic effects, respectively, and e is the random residual term.

The fixed effects ($\boldsymbol{\tau}$) in the baseline model included a covariate for establishment at each site which was measured as the number of plants per plot. For each site the baseline spatial randomisation model included random effects ($\boldsymbol{u}_o$) for replicate, where replicate is a factor with two levels representing random effects between the resolvable replicated entries. Random effects ($\boldsymbol{u}_o$) also included a row effect for each site where row has levels equal to the number of rows in each site. The Hermitage and Biloela sites had extra rows of missing data inserted to account for spraying operations. Supp. Table S3 describes the non-genetic terms that were fitted for each site.

The variance model for $\boldsymbol{e}$ contained the Kronecker product of first order auto-regressive processes in the row ($\boldsymbol{\Sigma}_r$) and column ($\boldsymbol{\Sigma}_c$) directions respectively: $\text{var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{\Sigma}_r \otimes \boldsymbol{\Sigma}_c)$. The vector $\boldsymbol{u}_{h_g}$ is of length n representing the random effects for the n genotyped lines.

The nongenetic terms, including the residual effects $\boldsymbol{e}$ and the peripheral effects $\boldsymbol{u}_o$ as well as the fixed effects $\boldsymbol{\tau}$ were calculated using the total number of lines in the data (791). The genetic effects $\boldsymbol{u}_{h_g}$ were based on genotyped lines only (544). Lines without genotypic data were retained to preserve the spatial effects but did not contribute to the estimate of genetic variance parameters by inclusion of a fixed effect that distinguishes between genotyped and non-genotyped lines. It was assumed hereafter for ease of computation that all design matrices conformed to allow fo discrepancies in number of genotyped lines vs the number of phenotyped lines by the inclusion of zeros where no effect was present.

We propose an extension to (4.1) where genetic effects $\boldsymbol{u}_{h_g}$ are partitioned into 3 parts:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_h(\boldsymbol{u}_{h_m} + \boldsymbol{u}_{h_p} + \boldsymbol{u}_{h_e}) + \boldsymbol{Z}_o\boldsymbol{u}_o + \boldsymbol{e}. \tag{4.2}$$

where $\boldsymbol{u}_{h_m}$ is the additive effect for lines captured by markers, $\boldsymbol{u}_{h_p}$ is the additive effect for lines due to pedigree and $\boldsymbol{u}_{h_e}$ is the residual genetic effect.

We assume that each of the three vectors of genetic effects namely $\boldsymbol{u}_{h_m}$,$\boldsymbol{u}_{h_p}$ and $\boldsymbol{u}_{h_e}$ were (pairwise) independent and Gaussian with zero mean and variance matrices $\sigma^2_m\boldsymbol{A}_m$, $\sigma^2_a\boldsymbol{A}$ and $\sigma^2_e\boldsymbol{I}_{m_o}$ where σ^2_m, σ^2_a and σ^2_e are the marker based additive variance, pedigree based additive variance and residual genetic variance respectively.

Matrix $\boldsymbol{A}$ is defined as the additive relationship matrix with diagonal entries given by $1 + F_i$ and off diagonal entries given by $2f_{ij}$ where F_i is the inbreeding coefficient of entry i and f_{ij} is the coefficient of parentage between lines i and j. Since the lines in this study are all homozygotes the diagonal of $\boldsymbol{A}$ is 2. The inverse of this matrix, $\boldsymbol{A}^{-1}$ can be computed using the algorithm of Henderson (1976) and was computed in the R (R Core Team, 2018) package Asreml-R (Butler et al., 2009) using the function asreml.Ainverse. The computation method is fully described in Oakey et al. (2006).

The matrix $\boldsymbol{A}_m$ is the relationship matrix formed using markers with

$$\boldsymbol{A}_m = (\boldsymbol{M} - 2\boldsymbol{P})(\boldsymbol{M} - 2\boldsymbol{P})^T / (2 \sum_{i=1,\ldots,m} p_i(1-p_i)), \tag{4.3}$$

where $\boldsymbol{M}$ was the $n \times m$ marker matrix of n genotyped lines by m markers with values of -1 and 1 representing the two alleles and with missing values calculated by the average marker frequency across

all lines. $\boldsymbol{P}$ is a matrix with columns given by p_i, where p_i is the allele frequency of of the second allele of marker i. This formulation creates a matrix that is analogous to the relationship matrix $\boldsymbol{A}$ (VanRaden, 2008).

Fitted models used four different partitions of the genetic term $\boldsymbol{Z}_h\boldsymbol{u}_{h_g}$ and are labelled I, P, M and P+M;

I independent lines as given by equation (4.1),

P pedigree based relationship only ($\boldsymbol{Z}_h\boldsymbol{u}_{h_p} + \boldsymbol{Z}_h\boldsymbol{u}_{h_e}$),

M marker based relationship only ($\boldsymbol{Z}_h\boldsymbol{u}_{h_m} + \boldsymbol{Z}_h\boldsymbol{u}_{h_e}$)

P+M all 3 terms as in equation (4.2).

Model P and M are not sub models of each other and therefore cannot be directly compared. A Comparison was made between the baseline model I and model P+M to assess the difference in model fit between models P and M.

4.3.5 Accuracy of selection

Prediction error variance

In general the reliability of genotype i can be written as

$$r_i^2 = 1 - (PEV_i/\sigma_g^2) \tag{4.4}$$

where PEV is a vector of prediction error variances and σ_g^2 is the genetic variance. The prediction error variance can be defined as the fraction of the additive genetic variance not accounted by by the prediction. It is commonly written as $PEV = \text{var}(\tilde{\boldsymbol{u}} - \boldsymbol{u})$ i.e. the variance of the difference between the prediction and it's true value. It's calculation can be dervied from the set of equations known as the mixed model equations (MMEs)(Henderson, 1975).

The equation for a linear mixed model is typically written

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \tag{4.5}$$

The MMEs for the model expressed by equation (4.5) are

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix} \tag{4.6}$$

where $\text{var}(\boldsymbol{e}) = \boldsymbol{R}$ and $\text{var}(\boldsymbol{u}) = \boldsymbol{G}$. $\hat{\boldsymbol{\tau}}$ are the fixed effects (BLUEs) and $\tilde{\boldsymbol{u}}$ are the random effects (BLUPs). Let the coefficient matrix of equation (4.6) be $\boldsymbol{C}$, and write the solution to these equations

$$\begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{12}^T & \boldsymbol{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}. \tag{4.7}$$

The predicted error variance for the random effects in (4.5) are the diagonals of the variance/covariance matrix $PEV(\boldsymbol{u}) = \text{var}(\tilde{\boldsymbol{u}} - \boldsymbol{u}) = \boldsymbol{C}_{22}^{-1}$ (Butler et al., 2009; Welham et al., 2004).

Similarly, the *PEV* values can be calculated for the random genetic effects in equation 4.2. Since the random components are independant, $\boldsymbol{u}$ from equation 4.5 can be partitioned into the marker additive effects ($\boldsymbol{u}_{h_m}$), the pedigree additive effects ($\boldsymbol{u}_{h_p}$), the residual genetic effects ($\boldsymbol{u}_{h_e}$) and the periferal spatial effects ($\boldsymbol{u}_o$). The required values for *PEV* can be calculated by substituting in the respective genetic variance in the place of $\boldsymbol{G}$, i.e. $\sigma_a^2\boldsymbol{A}$ for model M, $\sigma_m^2\boldsymbol{A}_m$ for model P and $\sigma_m^2\boldsymbol{A}_m + \sigma_a^2\boldsymbol{A}$ for model P+M.

Full family validation sets

Cross-validation procedures that require the removal of random subsets of individuals are problematic for partially replicated data. These data have large discrepancies between the predicted standard errors for lines that were replicated and lines that have no replication. There is great danger in comparing random subsets of lines without taking care to ensure that the lines for removal have a comparable replication structure. Authors such as Würschum et al. (2017) have discussed across family cross-validation by removing large numbers of individuals from families.

To test the prediction capability of the best fit model, we examined validation data sets that involved the prediction of lines from entire families. From the 31 full sib families involved in this study we chose to remove each of 21 families that had 10 or more full siblings and balanced within family replication (see Supp. Table S1 for a detailed list of families and their respective number of siblings). Models P, M and P+M were fitted 21 times, each time the validation data were the lines from one family and the training data consisted of the lines from the remaining families. The models were fitted using fixed values for the random terms (genetic and peripheral terms).

Realised accuracies were calculated as correlations for the predicted breeding values for the lines that were removed against the predicted mean for each line from the full data model (using model I). Validation accuracies were made by dividing the average correlation across the 21 families by the square root of the heritability for the full data model. Since the predicted effects for removed line will always be the same within each family for the pedigree part of the model, validation correlations were only calculated for model M.

For each model the model based prediction accuracies of each line i for each site can be calculated as $\sqrt{1 - PEV_i/\boldsymbol{G}_{ii}}$ (Strandén and Garrick, 2009). The total genetic variance for each site is represented by $\boldsymbol{G}_{ii}$, where $\boldsymbol{G}_{ii}$ is the relevant genetic relationship matrix, for model M it is given by the iith diagonal element of $\sigma_m^2\boldsymbol{A}_m + \sigma_e^2\boldsymbol{I}$. PEV_i are the ith value of the diagonal elements of the inverse of the coefficient matrix multiplied by the error variance. The *PEV* values for the non-phenotyped lines from each removed family (validation sets) can be calculated for the removed lines using the Asreml-R predict function (Butler et al., 2009; Welham et al., 2004).

Expected accuracy for full set of phenotyped lines

Expected prediction accuracies can be calculated for each site and model using the complete set of phenotyped lines (ie full data model). If this prediction accuracy is low the capacity for genomic selection from that data will be limited.

Expected prediction accuracy was calculated for each line within each site/model combination from the diagonal elements of the inverse of the coefficient matrix (prediction error variance) (Hayes et al., 2009). These values are useful since they are available to plant breeders without the need to perform cross-validation, and furthermore are calculated for each individual line, and as such will reflect the relationship of that line to the reference population. Hayes et al. (2009) have shown that the expected prediction accuracy derived from BLUP models agree with realised accuracies from cross-validation.

Expected accuracy for each line i within each model and each site were calculated using the accuracy formula as above where the total genetic variance for each site would be, for model P+M as the average of the diagonals of the matrix given by $\sigma_m^2 \boldsymbol{A}_m + \sigma_a^2 \boldsymbol{A} + \sigma_e^2 \boldsymbol{I}$. The predicted error variances can be calculated for individual lines in the model using the Asreml-R predict function (Butler et al., 2009; Welham et al., 2004).

4.4 Results

4.4.1 Association between marker and pedigree relationships

The degree of relatedness of each line to the rest of the genotyped lines in the study, was calculated using either pedigree or marker information. These values can be scaled to values between 0 and 1 with 0 having no relationship and 1 being identical. A value for each line could be calculated by averaging across the rows of each of the matrices $\boldsymbol{A}$ and $\boldsymbol{A}_m$. These values were high for lines that were in a typical parentage for this set of lines, with lower values indicating that they were more diverse. The relatedness values derived from the $\boldsymbol{A}$ matrix showed distinct groupings of lines due to the assumptions inherent in pedigree based relationships (eg full siblings derived from the same cross are equally similar) whereas the $\boldsymbol{A}_m$ derived values were more evenly spread because they take into account Mendelian sampling within families (Figure 4.1). This indicated that lines that were derived from a pedigree that contained diverse siblings were more likely to be predicted accurately using marker information in comparison to using pedigree information, in which case all full siblings within a single family would have the same predictions.

Figure 4.2 shows a heatmap representation of the relationships between lines based on the relatedness given by the $\boldsymbol{A}$ matrix and the $\boldsymbol{A}_m$ matrix. The relationships due to pedigree had a blocked appearance since each family was identically related to all full siblings within each of the other families, in contrast to the less distinct block pattern based on the marker-based relationships, where Mendelian sampling increased the variance of relationship.

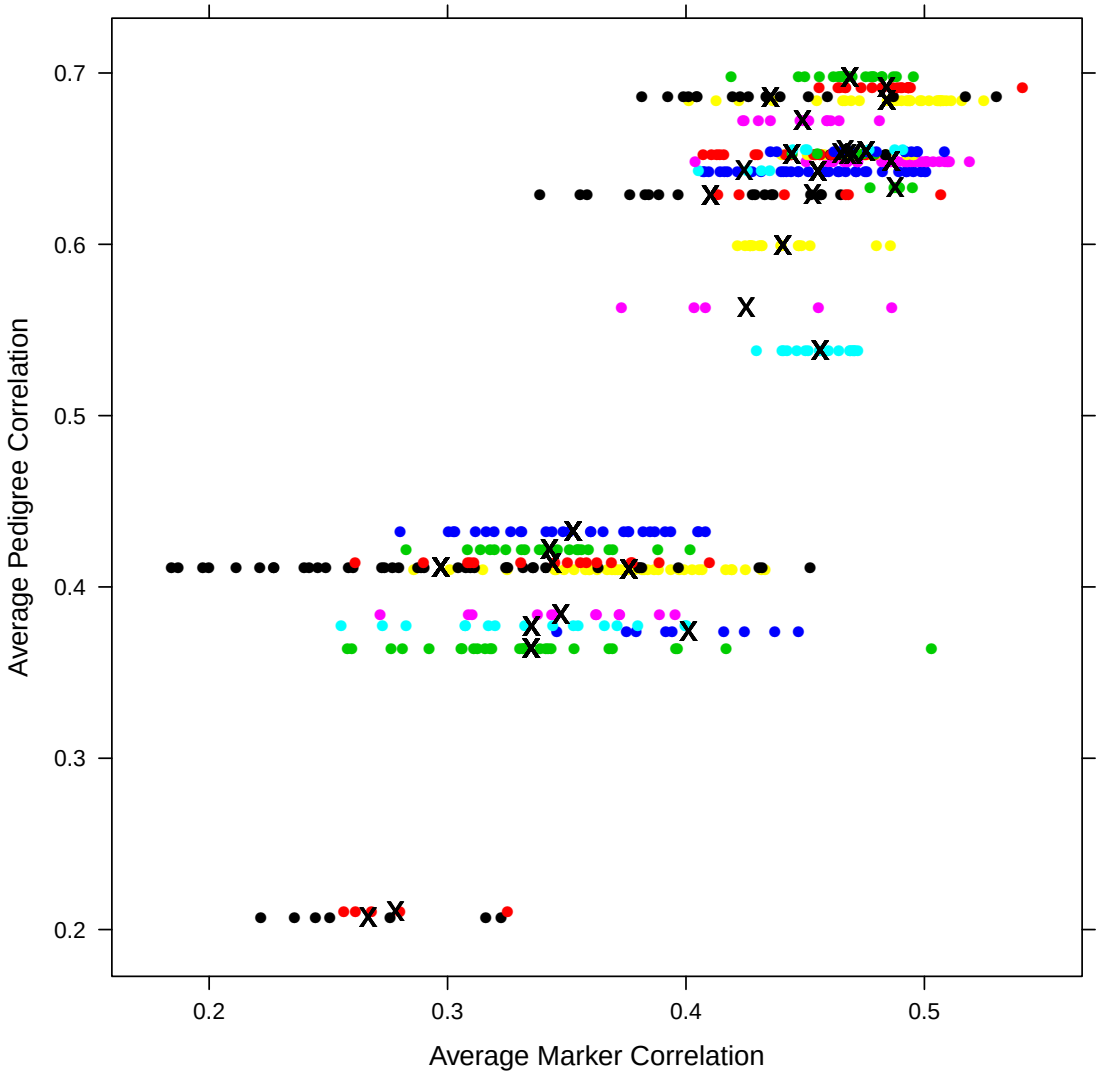

Figure 4.1: Scatterplot of average relatedness between families based on ancestral and marker relationship matrices. Black crosses represent the mean for each family.

4.4.2 Fit of alternate linear mixed models

All sites showed a decrease in the residual genetic variance (σ_e^2) when comparing models P, M and P+M against model I (Table 4.2). The percentage decrease was different for each site with the Dalby Box site showing a decrease in σ_e^2 up to 92%, indicating that the model including both pedigree based and marker based relationships together explain the majority of the total genetic variance for that site. The decrease of residual genetic variance at the other three sites was lower, with a decrease of 32 to 49%.

For model P+M, the percentage of each component that contributed to the total genetic effect varied from site to site. By considering the variance components from model P+M for each site we calculated the percentage of total genetic variation contributed by markers and by pedigrees for each site. Dalby Box had the highest percentage of total genetic variation contributed by marker based additive variance (50%), followed by Biloela (46%) and Hermitage (17%). The Dalby site had 0% genetic variation

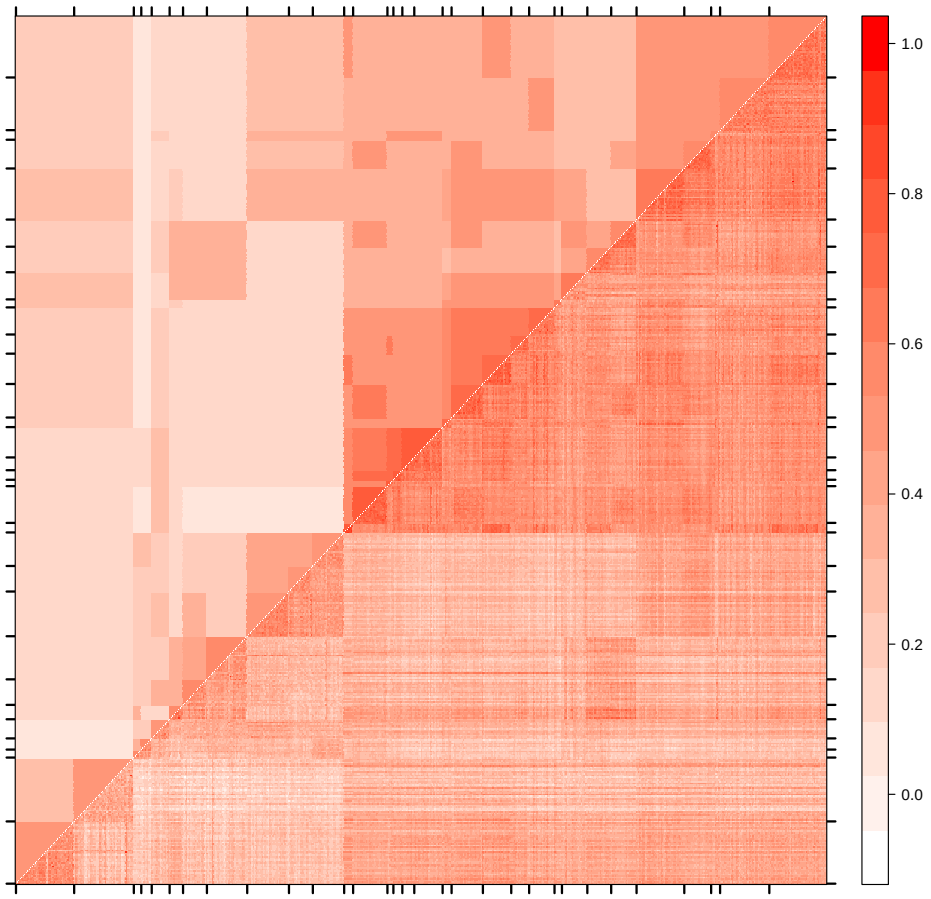

Figure 4.2: Heatmap of the relationships between lines based on pedigrees (upper Triangle) and markers (lower triangle). Values from the $\boldsymbol{A}$ and $\boldsymbol{A}_m$ matrices are represented on a scale of values between 0 and 1; axes tick marks and black or grey bars indicate the different families.

contributed by marker based additive variance, but the highest percentage of total genetic variation contributed by pedigree based additive variance (59%), followed by Dalby Box (45%), Hermitage (26%) and Biloela (10%). The different site rankings of total genetic variation contributed by marker based additive variance and pedigree based additive variance indicated that each site had different partitioning of total genetic variance into its respective additive and residual genetic terms.

REML log likelihoods were calculated for each of the models for each site (Table 4.2). Tests of significance can be performed using a REML log likelihood ratio test where twice the difference in REML log likelihood (Table 4.2) can be compared to a χ^2_p where p is the difference in the number of parameters. At the Biloela site, both model P and model M showed significant improvement over the model that did not use pedigree or marker data to generate relationship information (model I). However model P+M only showed an improvement over model P (based on 6.42 compared to χ^2_1), therefore model M can be considered the best fit model for Biloela. For the Dalby site, where σ^2_m was negligible, there was no significant difference between model P+M and model P which indicated no advantage to including marker based relationships in the model applied to this site; hence model P can be considered to be the best fit model for Dalby. The best fit model for the Dalby Box site was model

Table 4.2: Variance components from fitting models I, P, M and P+M to each site. σ_a^2 is the pedigree based additive variance, σ_m^2 is the marker based additive variance and σ_e^2 is the residual genetic variance, a blank in the table indicates that term was not present in the model. REML log likelihoods are presented as difference in REML log likelihood from model P+M and calculated AIC values are presented with the lowest AIC value in bold font.

		Variance component			REML	
Model	Site	σ_a^2	σ_m^2	σ_e^2	logl	AIC
I	Biloela			0.146	-6.78	27.6
P	Biloela	0.046		0.111	-3.21	22.4
M	Biloela		0.094	0.090	-1.16	18.3
P+M	Biloela	0.035	0.080	0.075	0	**18.0**
I	Dalby			0.074	-4.25	24.5
P	Dalby	0.079		0.040	0	**18.0**
M	Dalby		0.024	0.066	-3.43	24.9
P+M	Dalby	0.079	0.000	0.040	0	20.0
I	Dalby Box			0.242	-17.22	48.4
P	Dalby Box	0.328		0.040	-5.02	26.0
M	Dalby Box		0.312	0.094	-3.62	23.2
P+M	Dalby Box	0.206	0.218	0.015	0	**18.0**
I	Hermitage			0.360	-7.43	24.9
P	Hermitage	0.142		0.262	-1.46	15.4
M	Hermitage		0.154	0.276	-1.74	15.2
P+M	Hermitage	0.110	0.089	0.237	0	**14.0**

P+M which had significant increase in REML log likelihood over both model P and model M. Finally, the Hermitage site showed that model P+M was not a significant improvement over either models P or M, both of which were a significant improvement over model I. The Akaike information criterion (AIC) values in Table 4.2 showed that model P+M was the best fit model for Biloela, Dalby Box and Hermitage and model P was the best fit model for Dalby.

4.4.3 Full family validation sets

Average cross validation accuracies across the 21 analyses of removed families are between 0.12 and 0.27 (Table 4.3). This is to be expected based on the low heritabilites for yield in each site and also since the genomic predicted values are based only on the additive partition of the total genetic prediction of each line.

Table 4.3 shows the average expected prediction accuracy across the 21 runs for each model at each site. For all sites except Dalby the prediction accuracy for model M was higher than the prediction accuracy for model P and model P+M does not appear to have an increased accuracy over model M. The accuracy for Dalby was higher for Model P than for Model M and Model P+M is the same as Model P since $\sigma_m^2 = 0$ in model P+M (Table 4.2).

There was minimal increase in prediction accuracy between model M and model P+M for all sites except Dalby. Hermitage showed lower prediction accuracy for these non-phenotyped additive effects which may be due to the lower contribution of additive variance to the total genetic variance as seen in

Table 4.2.

Table 4.3: Heritability for the analysis of the full set of lines, Cross Validation accuracy and Expected prediction accuracy for each site and each model averaged across 21 cross validation runs using standard errors from lines from whole families that have been removed in each run.

Site	Heritability (%)	Cross Validation accuracy	Expected Prediction Accuracy		
			Model P	Model M	Model P+M
Biloela	35	0.20	0.25	0.43	0.43
Dalby	17	0.24	0.40	0.24	0.40
Dalby Box	31	0.27	0.49	0.59	0.59
Hermitage	45	0.12	0.29	0.36	0.34

4.4.4 Accuracy of selection for phenotyped lines

The average prediction accuracies for each fitted model using the full set of phenotyped lines are detailed in Table 4.4. There was minimal improvement in prediction accuracy between models I and P for sites Biloela and Hermitage, indicating that model P was not more accurate than model I for those sites. However, model M showed an increase in accuracy over model I for all sites. In the case of Biloela, Dalby Box and Hermitage the accuracy of model P+M, was not better than model M. This was expected for Biloela and Hermitage since the REML log likelihood test showed no significant improvement in model fitting between model M and model P+M for those sites (Table 4.2), however this was not the case for Dalby Box which showed significant REML log likelihood improvement of model P+M over model M but the selection accuracy is the same. For the Dalby site model P was the most accurate and marker based relationships only showed a small increase in the prediction accuracy over model I and since σ^2_m is 0 for model P+M (Table 4.2), the selection accuracy for model P+M is equivalent to model P.

Average standard errors were calculated using lines within each family for both model P and model M. These standard errors were plotted against the average relatedness for each family (as in Figure 4.1) using the marker relationship matrix (Figures 4.3 and 4.4). Figure 4.3 showed that in general a family that was less related on average to the population of lines (i.e. with a low average relatedness) had a higher standard error than those families with higher values of relatedness. There were generally more points with high standard errors when they were formed using the pedigree model. However, when applying model P, there were high standard errors for some families that had high values of relatedness (Figure 4.3A).

Figure 4.4 showed that standard errors were not strongly affected by the number of full and half siblings included in each family, and that high standard error occurs when the average family relatedness is low. The numbers plotted in Figure 4.4 represent individual families sorted from lowest number of full and half siblings (1) through to the family with the highest number of full and half siblings (31). (see Supp. Table S1 for list of families and their corresponding ID number)

Table 4.4: Expected prediction accuracy for each site and each model using the full set of phenotyped lines and average standard error for all phenotyped lines.

Site	Model I	Model P	Model M	Model P+M
Biloela	0.59	0.59	0.64	0.64
Dalby	0.41	0.57	0.45	0.57
Dalby Box	0.56	0.67	0.70	0.70
Hermitage	0.67	0.68	0.69	0.68

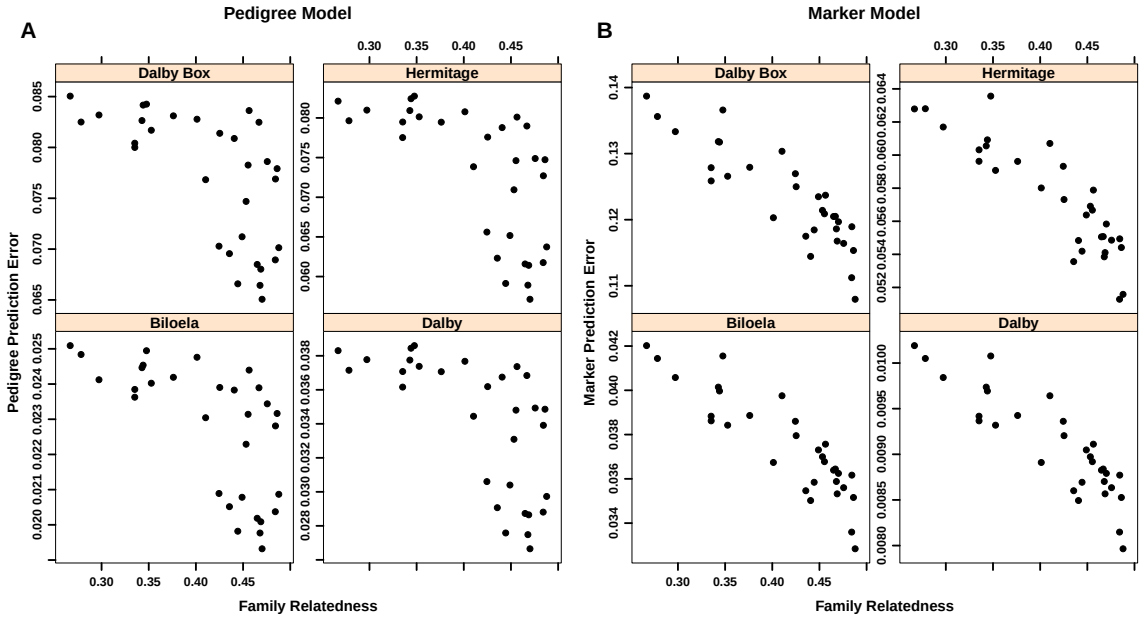

Figure 4.3: Average within-family standard error vs. average marker relatedness for each fully phenotyped family. Each point represents a single family. (A) The pedigree model (Model P), and (B) the markermodel (Model M).

4.5 Discussion

This study has applied single stage mixed models to four trials that were part of an existing sorghum pre-breeding program. The accuracy of prediction of the test cross hybrids using relationships information based on pedigree or marker data was strongly influenced by the degree to which a particular parent line of a test cross hybrid was representative of the training population.

Prediction accuracy is improved by including both marker and pedigree information

Our results support previous studies, for example Crossa et al. (2010) and Burgueno et al. (2012), where the inclusion of pedigree based and marker based relationships can provide improved prediction accuracy over models based on either marker based or pedigree based relationships alone.

However, the utility of pedigree information is frequently restricted, firstly by the assumption that there is equal genetic contribution from each parent and secondly by the quality of the pedigree information available. Pedigree information will rapidly decline in relevance to the selection candidates

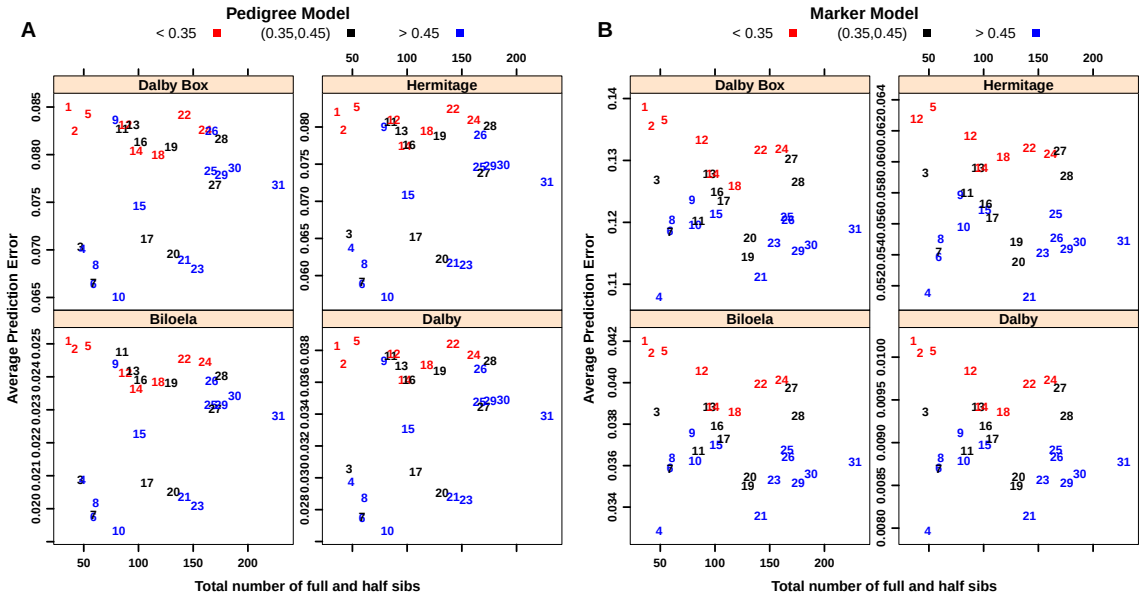

Figure 4.4: Average within-family standard error vs. the total number of full and half siblings per family. Each point represents a single family numbered 1 through 31, with 1 being the family with the lowest number of full and half siblings and 31 having the highest. (A) The pedigree model (Model P), and (B) the marker model (Model M). Colors represent the average marker relatedness for (A) pedigree or (B) marker.

as the link to the training population declines through successive rounds of genomic selection (Wolc et al., 2011). Phenotyping will be required to update the lines in the training population to retain the relationship between pedigree and performance.

Genotype by environment (G×E) interactions are commonly observed in sorghum breeding trials, for example in a set of 23 trials spanning 5 years Jordan et al. (2012) observed genetic correlations in yield between sites varying from -0.15 to 0.97 with an average between site correlation of 0.28. In other sorghum breeding trial analyses we have observed high G×E when analysing the data without relationship information. However when using a model that incorporates pedigree information we observe less G×E for the additive partition of the genetic variance (unpublished studies). In the current study four sites were analysed separately and shown to exhibit differing results in terms of the fitting of pedigree and/or marker in statistical models. Our results indicate the need for further exploration in this area for multi-environment analysis and investigations of G×E using models that include both pedigree and markers. This will require a larger number of trials than were available for the current study.

The relationship between individuals and the training population influences prediction accuracy

Habier et al. (2013) demonstrated that genomic prediction exploits two sources of information, relationships between individuals and linkage disequilibrium (LD) between markers and genes influencing the trait. Although these components are not independent, the contribution of relationships to predictions

declines more rapidly with generations of GS compared with the contribution of LD, particularly if marker density is high (Habier et al., 2013; Jannink et al., 2010).

We observed that the standard error for marker based predictions increased as lines become less related to lines in the training population (Figure 4.3). In the case of the pedigree based predictions an increase in relatedness can have high or low prediction accuracy. This is particularly relevant in cases where families have parents with limited or no pedigree history and that are less related to the training population as a whole. Such families are expected to have lower prediction accuracy due to both relationships and LD, for example, family R04001, which had a parent with unknown pedigree and low similarity to the other families in the training set with an average correlation of 0.29, had the highest prediction error in all sites; this can be seen in Figure 4.3, where the lowest value on the *x*-axis has the highest value on the *y*-axis. Prediction approaches using pedigrees or markers had higher prediction errors as they diverge from having a sufficient number of close lines. Figure 4.3A demonstrated that for standard pedigree based genomic best linear unbiased prediction, the prediction accuracy was lower for families that were less similar to the typical family. Habier et al. (2007, 2013) found that prediction accuracies are strongly affected by the number of close relatives in the training population. However in the current study the prediction errors and relatedness did not appear to be affected by the numbers of full and half siblings within each family. This is likely explained by the fact that the pedigrees of the individuals in the training population had a high degree of inter-relationships, and that each individual had some relatedness to all other individuals with, on average, each individual being a half or full sibling to 20% of the total number of individuals (see Supp. Figure S1 and Supp. Table S1). Large families that were less related had a tendency to have slightly inflated degree of relatedness due to making a greater contribution to the full population of lines. Those larger less related families were also less accurate when tested using cross validation. It appears from our results that the smaller more related families are more desirable than large unrelated families. This is an important finding for the design of training and selection populations.

Feasibility of GS in sorghum

The prediction accuracies observed in this study were generally high using relatively low marker density and high average LD, indicating the potential utility of GS in sorghum (Table 4.4). Our approach, and most published plant breeding examples to date, have made use of existing multi-environment breeding trial data rather than examining purpose designed training and selection populations. As a result, care should be taken before extending these results, as the close relationship between the predicted individuals and the training set (e.g. presence of large numbers of full and half siblings) will not be the case in most applications of GS, particularly when GS is implemented over multiple generations. Such inflation of prediction accuracies, due to the high level of pedigree interrelatedness between the individuals in the training population, was also observed by Ly et al. (2013). Future implementation of GS will need to consider that the individuals in the selection population will not be as closely related as those presented in this study (i.e. not full siblings) and therefore the prediction

accuracy will be lower.

G×E presents a major challenge to the deployment of GS in sorghum. Where between site correlations are low, the capacity to predict genotype performance across sites will be low. Care will need to be taken to create a representative set of environments that have known environment types that encompass as many as possible of the environments used in the subsequent selection population. This is an area where the application of crop simulation modelling to identify environmental types (Chapman et al., 2000a; Heslot et al., 2013) would have considerable utility.

Our results support the initial conservative deployment of a GS approach where the training population is a subset of a larger selection population, which is genotyped and a subset is phenotyped, with selections made on the predicted performance of the entire population. Such an approach would use GS to increase genetic gain by increasing selection intensity while forgoing returns from GS that could be achieved by conducting multiple generations of crossing and selection without phenotyping. This is feasible since the cost of genotyping is currently less than the cost of phenotyping.

The following publication has been incorporated as Chapter 5.

Hunt et al. (2020)

Hunt, C. H., Hayes, B. J., van Eeuwijk, F. A., Mace, E. S., and Jordan, D. R. (2020). Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships. *Theoretical and Applied Genetics*. https://doi.org/10.1007/s00122-019-03526-7

Author	Statement of contribution	%
Colleen Hunt	writing of text	70
	proof reading	60
	Theoretical derivations	80
	numerical calculations	100
	preparation of figures	100
	initial concept	50
Ben Hayes	writing of text	5
	proof reading	10
	Supervision, guidance	25
	Theoretical derivations	10
	initial concept	10
Fred van Eeuwijk	writing of text	10
	proof reading	10
	Supervision, guidance	15
	Theoretical derivations	10
Emma Mace	writing of text	5
	proof reading	10
	Supervision, guidance	15
David Jordan	writing of text	10
	proof reading	10
	Supervision, guidance	45
	initial concept	40

Chapter 5

Multi-Environment analysis of sorghum breeding trials using additive and dominance genomic relationships

5.1 Abstract

Sorghum is an important hybrid crop that is grown extensively in many sub-tropical and tropical regions including Northern NSW and Queensland in Australia. The highly varying weather patterns in the Australian summer months mean that sorghum hybrids exhibit a great deal of variation in yield between locations. To ultimately enable prediction of the outcome of crossing parental lines, both additive effects on yield performance and dominance interaction effects need to be characterised. This paper demonstrates that fitting a linear mixed model that includes both types of effects calculated using genetic markers in relationship matrices improves predictions. Genotype by environment interactions were investigated by comparing FA1 (single factor analytic structure) and FA2 (two factor analytic) structures. The GxE causes a change in hybrid rankings between trials with a difference of up to 25% of the hybrids in the top 10% of each trial. The prediction accuracies increased with the addition of the dominance term (over and above that achieved with an additive effect alone) by an average of 15% and a maximum of 60%. The percentage of dominance of the total genetic variance varied between trials with the trials with higher broad-sense heritability having the greater percentage of dominance. The inclusion of dominance in the factor analytic models improves the accuracy of the additive effects. Breeders selecting high yielding parents for crossing need to be aware of effects due to environment and dominance.

5.2 Introduction

The phenomena of hybrid vigour, sometimes called heterosis, has been exploited to improve yields in a variety of crops such as maize, sorghum, sunflower, canola, rice and wheat through the deployment

of F_1 hybrids produced by crossing genetically diverse inbred parent lines. The phenomena of hybrid vigour arises because of a range of factors particularly directional dominance where favourable alleles for fitness at a locus are dominant to unfavourable alleles. Because hybrids between inbred parent lines are heterozygous at loci that are polymorphic between the parents, the performance of the hybrid is the result of both the additive contributions of both parents and the directional dominance resulting from the interaction in a heterozygous locus, as well as mechanisms such as epistasis. So in hybrid crops, additive, dominance and epistatic components of genetic variance contribute to differences in yield between cultivars. Typically, breeding programs for hybrid crops identify elite cultivars by assessing the performance of large numbers of different combinations of offspring of inbred parent lines crossed to tester lines in multi-environment trials. The advantage of including dominance effects in the analysis of hybrid crop data has been recently discussed for maize (Dias et al., 2018), rice (Cui et al., 2019) and hybrid wheat (Würschum et al., 2018) to name a few. Investigating the impact of dominance genetic contribution has never been done in sorghum, and therefore, this is one of the objectives of the current study.

For genetic evaluation of cultivars with best linear unbiased prediction (BLUP), including marker based additive relationships has been used extensively and proven to have improved predictive performance over pedigree based additive relationships (Hayes et al., 2009; Habier et al., 2010; de los Campos et al., 2013; Heslot et al., 2012). Pedigree relationships assume that all siblings share an equal proportion of the genome (the expected relationship), while marker based relationships have the advantage that they can estimate Mendelian sampling within siblings (realised relationships at the genome level).

Oakey et al. (2007) and Dias et al. (2018) discuss partitioning the genetic effects into additive, dominance and residual genetic parts (in particular partitioning for dominance) in multi environment trials. They found that including dominance in their models improved the statistical fit and the accuracy of the predicted values. The partitioning can be performed by calculating relationship matrices for both additive and dominance relationships and incorporating them into the genetic variance structure of a fitted linear mixed model. Markers can be used to calculate the relationship matrices for both additive and dominance components (Vitezica et al., 2013; VanRaden, 2008; Aliloo et al., 2016; Muñoz et al., 2014; Dias et al., 2018). The calculation and inversion of a pedigree based dominance relationship matrix typically requires a set of hybrids with many combinations of males and females (Aliloo et al., 2016). The availability of large numbers of markers can compensate for the lack of balance by considering gene action at the individual marker level.

Genotype by environment interactions (GxE), that is differential genotype responses to types of environments, can cause re-ranking and complicate selection within breeding programs. Dominance may be a component of this GxE. For example, Betran et al. (2003) found that for maize dominance was greater in environments that had experienced drought stress. Dias et al. (2018) found that including the additive and dominance terms in a GxE model improves the accuracy when considering genomic predictions. Plant breeders use one of two strategies to manage GxE, either they ignore it and select for broad adaptation or they attempt to exploit it by selecting for both broad and specific adaptation. In

both cases multi-environment trials (METs) are commonly used in an attempt to produce an across trial genotype effect that represents the average performance of the genotypes across a sample of environments. In best-practice design and analysis of MET data, breeders will attempt to account for spatial variation in each field and for genetic correlations among the trials.

Smith and Cullis (2018) detail the model fitting techniques that are commonly used in Australian plant breeding programs. They fit mixed models to series of trials as a single stage analysis that simultaneously allow for spatial effects at each trial and fit correlated variance structures to the genotype by environment interactions. These techniques identify the extent and complexity of GxE with respect to providing the most accurate analysis of hybrids within multiple environments. Smith and Cullis (2018) also present a factor analytic selection tool (FAST) which examines measures of overall performance and stability across environments. The FAST method is applicable to MET analyses where the first order FA loadings are positive and represent the majority of the explained variation.

Authors such as Oakey et al. (2016), Borgognone et al. (2016) and Tolhurst et al. (2019) discuss using marker based additive relationship matrices in a mixed model MET analysis incorporating spatial effects and factor analytic variance structures for both the additive effects and residual genetic effects. In hybrid crops such as sorghum these non additive residual genetic effects can be partly accounted for by dominance but models still need to accommodate for possible residual genetic effects.

The changes in genotype rankings in different environments are driven by changes in the importance of different traits that contribute to yield. Differences in the genetic architecture of these component traits can therefore potentially alter the importance of the different components of genetic variance. It has been observed for example that environment can differentially affect the performance of inbred lines and hybrids, altering the relationship between genetic diversity and heterosis (Betran et al., 2003).

In this paper we investigate the change in additive hybrid predictions for yield of sorghum after including hybrid dominance in the model, using both additive and dominance relationship matrices among the lines derived from markers. We examine the trial by hybrid interactions for both additive and dominance and discuss changes in prediction accuracy and hybrid rankings across trials.

5.3 Materials and Methods

5.3.1 Description of experimental data

We considered a set of sixteen trials from the 2015 and 2016 sorghum pre-breeding program conducted by the Queensland Department of Agriculture and Fisheries and the Queensland Alliance for Agriculture and Food Innovation. Sorghum hybrids are made using the cytoplasmic male sterility system this means that there are effectively two heterotic groups, restorers (male parents) and maintainers (female parents). The trials used in this study are known as advanced yield trials for males. One aim of these trials is to identify elite male parents, with high general combining ability (i.e. additive genetic value) for release to commercial breeding companies. The males, or restorer parents, have more genetic

diversity than the maintainer parents, thus the hybrids in the male breeding trials also have a broad diversity.

The trials contained a total of 1424 hybrids, the 2015 trials contained 691 unique hybrids and 2016 trials contained 925 unique hybrids with 192 hybrids being common across both years. The 1424 hybrids were comprised of 1401 test hybrids and 23 commercial sorghum varieties and the 1401 test hybrids were produced by crossing 867 F_4 or F_5 males with 2 inbred females across all trials with an extra inbred female in 2015. There were 111 males that were crossed with all 3 females and the total crosses for each female were 247, 668 and 486. Not all trials contain the same number of hybrids due to either lack of seed quantity or restrictions on the size of the available land. The breakdown of the number of hybrids and dimensions of each trial are given in Table 5.1. The three inbred females were chosen to contrast in their sensitivity to drought stress.

Trials were designed with partial replication (Cullis et al., 2006), where approximately 30% of the hybrids were replicated and the remaining hybrids had a single replicate. Hybrids were laid out using a spatial row-column design with the replicated hybrids resolved into two equal blocks. These designs enabled the trials to be analysed using linear mixed models with random genetic effects and including spatial effects for each trial. For 4 trials, Emerald 2015, Hermitage 2015, Blackville 2016 and Hermitage 2016, the prevalence of midge made it necessary to spray sorghum trials therefore within the trial design allowance was made for access to machinery every tenth row. This complicated the design by the creation of entire rows of missing data. All trials were planted as two row plots, 5 m long and 1.5 m wide with field layouts and raw mean yields as described in Table 5.1.

The data of interest here are the yields expressed in tonnes hectare (t/ha) obtained from harvesting in the year after the crop was planted.

Table 5.1: Description of the trials: location, number of hybrids, males, rows, columns and raw mean yield for each trial in the dataset.

Trial	Year	Location	Hybrids	Males	Rows	Columns	Mean Yield (t/ha)
Blackville	2016	NSW	732	414	44	24	8.1
Capella	2015	Nth Qld	594	373	30	26	3.0
Croppa Creek	2016	NSW	710	402	40	24	5.8
Dalby Box	2015	Sth Qld	645	404	30	28	7.3
Dalby Box	2016	Sth Qld	852	534	40	28	6.6
Emerald	2015	Nth Qld	612	377	25	38	3.1
Emerald	2016	Nth Qld	836	523	34	40	2.6
Gatton	2015	Sth Qld	474	329	30	21	6.3
Hermitage	2015	Sth Qld	652	407	34	28	7.0
Hermitage	2016	Sth Qld	926	591	40	36	7.2
Jimbour	2015	Sth Qld	626	387	30	28	4.3
Jimbour	2016	Sth Qld	748	422	40	25	5.3
Liverpool Plains	2015	NSW	636	397	30	28	6.9
Orion	2016	Nth Qld	707	400	48	24	2.9
Pirrinuan	2016	Sth Qld	878	548	40	28	6.5
Spring Ridge	2016	NSW	891	561	40	30	6.2

5.3.2 Genetic information

Genotypic data in the form of 26K SNP markers was available for 565 of the 866 male lines and the 3 female testers. The difference of 301 lines is a result of a lack of genetic data for 301 of the phenotyped males. For each female there were 239, 545 and 340 genotyped males with 173 males in common between all 3 female testers. The male lines had an interconnected pedigree structure with 255 unique parents, each line had as least one half sibling. The markers form a physical map with 10 chromosomes, each is between 55Mbp and 78Mbp in length. A consensus map has been used to predict the centimorgan distances using the physical distances, the lengths are between 112cM and 228cM. The marker distance between genotypes had a minimum of 0.18 cM and a maximum of 0.91 cM with an average distance of 0.63 cM. There were between 1532 and 4509 markers in each linkage group with the average LD within linkage groups between 0.054 and 0.069 for the male lines and between 0.041 and 0.073 for the hybrids (ESM Table B.1). Genotypes for the 1124 hybrids in the trials were created by combining the marker values for the male and female parents of each hybrid. At each loci the markers were coded as “00” and “11” for the homozygotes and “01” for the heterozygotes.

5.4 Statistical methods

Linear mixed models were fitted which allowed for the investigation of significant GxE within each additive and dominance partition.

The multi-environment linear mixed model for data vector $\boldsymbol{y}^{n\times 1}$ can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \tag{5.1}$$

where the vectors $\boldsymbol{\tau}, \boldsymbol{u}_g, \boldsymbol{u}_p$ represent fixed effects, random effects for hybrids and random non-genetic (or peripheral, ie design and additional) effects respectively. The 16 trials are stacked into a vector of length n, where n is the number of observational units in the whole dataset across trials, in this case the observational unit is a single field plot. The matrix $\boldsymbol{X}$ is the design matrix for the fixed effects and the matrices $\boldsymbol{Z}_g$ and $\boldsymbol{Z}_p$ are the design matrices for the genetic and peripheral terms. Spatial effects for each trial were found by analysing each trial individually. Each trial included a fixed covariate to adjust yield for establishment which was measured as number of plants per plot. The peripheral random effects $\boldsymbol{u}_p$ consisted of blocking parameters for replicate and row and the natural spatial AR1 auto regression terms for both column and row directions (see Gilmour et al. (1997) for a discussion on spatial field adjustments). The residual term $\boldsymbol{e}$ is assumed normal with zero mean and different variances for each trial and the peripheral effects $\boldsymbol{u}_p$ are allowed to vary between each individual trial. We assume that the random effects $\boldsymbol{u}_g$, $\boldsymbol{u}_p$ and $\boldsymbol{e}$ are mutually independent.

We can partition the genetic effects $\boldsymbol{u}_g$ from (5.1) into three parts $\boldsymbol{u}_g = \boldsymbol{u}_a + \boldsymbol{u}_d + \boldsymbol{u}_e$ as described in Oakey et al. (2007). $\boldsymbol{u}_a$ represents the vector of hybrid additive genetic effects, $\boldsymbol{u}_d$ represents the vector of hybrid dominance genetic effects and $\boldsymbol{u}_e$ represents the vector of residual genetic effects which are

not already defined by the additive and dominance partitions. The model written in (5.1) can now be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_{g_a}\boldsymbol{u}_a + \boldsymbol{Z}_{g_d}\boldsymbol{u}_d + \boldsymbol{Z}_{g_e}\boldsymbol{u}_e + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e}. \tag{5.2}$$

We assume that each of the three vectors of genetic effects namely $\boldsymbol{u}_a$, $\boldsymbol{u}_d$ and $\boldsymbol{u}_e$ are (pairwise) independent and are Gaussian with zero mean, with variance matrices $\boldsymbol{S}_a \otimes \boldsymbol{A}_m$, $\boldsymbol{S}_d \otimes \boldsymbol{D}_m$ and $\boldsymbol{S}_e \otimes \boldsymbol{I}$ where each $\boldsymbol{S}$ matrix is the 16×16 trial by genetic variance/covariance matrix for additive, dominance and residual terms respectively. Various parameterizations of $\boldsymbol{S}$ where considered. These include a compound symmetry structure (CS), where all trials have the same variance and all pairs of trials have the same covariance (Patterson et al., 1977); a diagonal structure (DIAG) which where trials are uncorrelated. Smith et al. (2001) and Piepho (1998) consider a factor analytic structure (FAk) with $k = 1$ or 2 factors so that the genotype effects in each environment are dependent on a set of random factors f_r such that $\boldsymbol{u}_a = \boldsymbol{f}_1\boldsymbol{\lambda}_1 + \boldsymbol{f}_2\boldsymbol{\lambda}_2 + ... + \boldsymbol{\delta}$ where $\boldsymbol{\lambda}_i$ are called loadings and $\boldsymbol{\delta}$ is the vector of residuals for the model. The variance of $\boldsymbol{u}_a$ can be expressed as $\boldsymbol{S} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$ where $\boldsymbol{\Lambda}$ is a $16 \times k$ matrix of loadings where $k = 1$ for an FA1 model and $k = 2$ for an FA2 model and $\boldsymbol{\Psi}$ is a diagonal matrix of specific variances for each trial. Smith and Cullis (2018) discuss the use of FA structures in MET analyses using multiplicative mixed models when the genetic variance has been partitioned.

The matrix $\boldsymbol{A}_m$ is the additive relationship matrix for hybrids and the matrix $\boldsymbol{D}_m$ is the dominance relationship matrix for hybrids. We use the subscript m to distinguish these from their respective pedigree counterparts. Both matrices were calculated using the genome-wide SNPs using methods described in VanRaden (2008); Vitezica et al. (2013).

$$\boldsymbol{A}_m = \frac{\boldsymbol{W}\boldsymbol{W}^T}{\sum_{i=1}^{j}(2p_iq_i)} \tag{5.3}$$

$$\boldsymbol{D}_m = \frac{\boldsymbol{M}\boldsymbol{M}^T}{\sum_{i=1}^{j}(2p_iq_i)^2} \tag{5.4}$$

where j is the total number of SNPs. For the additive matrix, $\boldsymbol{A}_m$, $\boldsymbol{W}$ is a matrix containing values equal to $-2p_i$, $(q_i - p_i)$ and $2q_i$ for "00", "01" and "11" respectively. p_i is the allele frequency of the most frequent allele ("00") for each individual SNP and can be calculated for the ith SNP as $p_i = freq(11) + freq(10)/2$ and $q_i = 1 - p_i$.

For the dominance matrix $\boldsymbol{D}_m$, the matrix $\boldsymbol{M}$ is a matrix containing values equal to $-2p_i^2$, $2p_iq_i$ and $-2q_i^2$ for "00", "01" and "11" respectively.

5.5 Model testing

The random components of linear mixed models can be tested for significance using a REML log-likelihood ratio test as long as the components included in the models are nested within each other. Typically a test can be performed when a random term is added to a model without subtracting any other random terms. The fixed components of each model must be the same. In the case of factor

analytic models, they can be compared for higher orders of the same terms, i.e. an FA1 can be compared to an FA2 since an FA1 is nested within an FA2.

The REML log-likelihood ratio test for testing model A against model B where model A is directly nested within model B is defined as

$$2(REMLll_B - REMLll_A) \sim \chi^2_r \tag{5.5}$$

where $REMLll_B$ is the REML log-likelihood for model B and $REMLll_A$ is the REML log-likelihood for model A. This log-likelihood test asymptotically follows a chi-squared distribution with degrees of freedom given by r, which is defined as the difference in number of variance parameters between models A and B.

Eight models were fitted and compared to test the significance of adding the dominance term into the model. Compound symmetry (CS), diagonal (DIAG), first order factor analytic (FA1) and a second order factor analytic (FA2) models were each fitted with and without a dominance partition for the genetic variance. All eight models retain the spatial terms for each trial. The FA1.A model, which had an FA1 additive term is a baseline model for testing an additive main effect. Model FA2.A had an FA2 additive term, comparing this model to FA1.A tests the significance of the second factor. The FA2.A factor analytic model explains significantly more hybrid by trial genetic variation than the FA1.A factor analytic model. Model FA1.AD had FA1 terms for both additive and dominance, tests of significance of these against FA1.A tested if a single dominance effect is significant against no dominance effect. FA2.AD had an FA2 structure for both additive and dominance, comparing this to FA1.AD tests the significance of GxE in both terms and comparing this to FA2.A tests the significance of including GxE for both additive and dominance against GxE for additive alone.

Models were also compared using the Akaike information criteria (AIC) (Akaike, 1974), which was calculated for each model as $AIC = 2(v - REMLll)$, where $REMLll$ is the REML log-likelihood of the fitted model and v is the total number of variance parameters in the model. Models with the lowest AIC values can be considered to be more parsimonious given the number of variance parameters they contain.

Prediction accuracies for the additive effects for each trial were calculated using the predicted error variances as described in section 4.3.5. The accuracy was calculated as the square root of reliability, $\sqrt{1 - \text{PEV}/\sigma^2_a \boldsymbol{A}_m}$ (Strandén and Garrick, 2009). $\sigma^2_a \boldsymbol{A}_m$ is the additive variance matrix for each trial.

All models were fitted using the R (R Core Team, 2018) package ASReml-R (Butler et al., 2009). The standard errors of difference were calculated using the ASReml-R predict function (Butler et al., 2009; Welham et al., 2004).

5.6 Results

5.6.1 Modelling the genetic terms

Fitted models using 8 different structures for the trial by hybrid terms are shown in Table 5.2. To examine the importance of including a dominance effect, the genetic variance was fitted with and

without the dominance partition. The four trial by hybrid structures considered were compound symmetry (CS), diagonal (DIAG), one component factor analytic (FA1) and two component factor analytic (FA2). The residual genetic term was found to be zero for many trials so was considered at the trial level only. The residual genetic partition was modelled with a DIAG structure regardless of the terms fitted for the additive and dominance partitions.

The AIC values for the compound symmetry models (CS.A and CS.AD) were comparable to those from the DIAG models (DIAG.A and DIAG.AD). However the DIAG.AD had the highest AIC value. This indicated that fitting a main hybrid effect and trial by hybrid interaction term such as in a compound symmetry model is a more appropriate model than fitting additive, dominance and residual genetic variances to each trial individually (without trial by hybrid interactions). CS.AD was significant when compared to CS.A, this indicated that dominance significantly improved the CS model. All factor analytic models out perform the DIAG and CS models showing that it is best to allow for the genetic variances and covariances to vary between trials.

The REML log likelihoods were found to increase significantly when fitting FA2 models for both the additive models and the dominance models (FA2.A versus FA1.A; FA2.AD versus FA1.AD). The addition of dominance was significant over the additive model (FA1.AD versus FA1.A and FA2.AD versus FA2.A). The model that included FA2 additive and FA2 dominance effects (FA2.AD) was the best fit model for these data based on the AIC values and the significant REML log likelihood increase.

The variance explained (VAF) was calculated for each trial as the sum of the squared FA loadings for each genetic component in each model (Table 5.2). The average VAF for the additive term (FA2.A) increased when fitting an FA2 model with both additive and dominance (FA2.AD). The best fit model based on AIC (FA2.AD) showed the largest values of VAF for both additive and dominance this indicated that allowing for GxE by fitting FA2 models had a strong influence on the variance explained for both additive and dominance.

Table 5.2: Number of genetic terms (n), REML log likelihoods, Akaike information criterion (AIC) and percentage variance explained (VAF) for models with and without dominance using compound symmetry (CS), DIAG, FA1 and FA2 structures for the trial by genetic variance/covariance matrices.

Model	n	REML log-likelihood	AIC	%VAF Additive	%VAF Dominance
DIAG.AD	48	−3516.62	7129.24	-	-
DIAG.A	32	−3527.07	7118.14	-	-
CS.A	4	−3483.75	6975.50	-	-
CS.AD	6	−3478.10	6968.20	-	-
FA1.A	48	−3328.13	6752.26	51	-
FA1.AD	80	−3276.62	6713.24	67	74
FA2.A	64	−3261.72	6651.44	60	-
FA2.AD	112	−3206.74	6637.48	79	89

5.6.2 Genetic variances

Variance components from models using compound symmetry and FA2 structures for the trial by hybrid terms for the additive, dominance and residual genetic terms are detailed in Table 5.3. The CS models include main effects for all partitions of the genetic variance and is included in the value for total genetic variance. The percentage of additive variance to the total genetic variance was 58% for CS.A and 50% for CS.AD, and CS.AD had 8% dominance variance. The additive main effect was 38% for CS.A and 33% for CS.AD, and the dominance main effect for CS.AD was 7.5%. The results for the best fit model FA2.AD showed that the additive genetic variance is between 42% (Pirrinuan 2016) and 91% (Cappella 2015 and Emerald 2016) of the total genetic variance, with an average proportion of 70%. The dominance variance as a proportion of the total genetic variance ranged from 8% (Cappella 2015) to 35% (Blackville 2016) with an average proportion of 25%. For 11 trials the residual genetic variance was 0, this indicated that only the additive and dominance partitions were needed for these trials. For all trials the additive genetic variance decreases when the dominance term is included in the model (Table 5.3), similarly the standard errors also decreased for all trials except Orion 2016.

Table 5.4 shows the prediction accuracies for each trial. For most trials the accuracy increased after adding dominance to the model. A few trials showed a small decrease, these trials show very little to no significant dominance variation.

5.6.3 Between trial correlations and assessment of GxE

The second order factor analytic structures applied to the additive and dominance genetic terms contain a set of estimated loadings for model FA2.AD (Figure 5.1, see ESM Figure S1 for FA2.A model). These loadings allow calculation of pairwise trial correlations (Figure 5.2, see ESM Figure S2 for FA2.A model). The trial correlations for the additive effects for lines range from -0.63 (Emerald 2016 versus Springridge 2016) to 0.99 (Blackville 2016 versus Dalby Box 2015). For the dominance effects the between trial correlations range from -0.99 (Jimbour 2015 versus Croppa Creek 2016) to 0.98 (Springridge 2016 versus Capella 2015). The average between trial correlation for additive effects was 0.59 and the average between trial correlation for dominance effects was 0.14. These results indicate that the dominance component had a larger spread of between trial correlations.

Plots of the rotated loadings from the FA2.AD model showed all first order loadings were positive for the additive partition and highly variable for the dominance partition (Figure 5.1). This result indicated that the trials were more variable for the dominance partition of the genetic variance and possibly were more associated with more GxE than the additive partition. Heatmaps of the pairwise trial correlations showed the spread of colour from -1 (blue) up to 1 (red) was more apparent in the heatmap of the dominance effects (Figure 5.2).

To further investigate the impact of GxE we considered the ranking of hybrids within each trial. Table 5.5 shows the percentage of hybrids that are in both the top 10% of the predicted yield across trials and the top 10% of the predicted yields for each individual trial. These were calculated using the additive effects from the FA2 model with additive only (FA2.A) and the FA2 model with both additive

Table 5.3: REML estimates of the genetic variance terms from the compound symmetry models (CS.A and CS.AD), and the FA2 models (FA2.A and FA2.AD). Genetic variances with standard error in brackets are given for the additive, dominance and residual genetic terms. (*)For the CS model the total includes the hybrid main effect.

	FA2.A			FA2.AD			
Trial	σ_a^2	σ_e^2	Total	σ_a^2	σ_d^2	σ_e^2	Total
CS Model	0.086 (0.006)	0	0.148*	0.079 (0.007)	0.013 (0.006)	0	0.158*
Blackville 2016	0.205 (0.040)	0	0.205	0.157 (0.039)	0.085 (0.057)	0	0.242
Capella 2015	0.035 (0.018)	0	0.035	0.033 (0.017)	0.003 (0.005)	0	0.036
Croppa Creek 2016	0.105 (0.038)	0	0.105	0.094 (0.028)	0.028 (0.009)	0	0.122
Dalby Box 2015	0.272 (0.088)	0	0.272	0.200 (0.043)	0.060 (0.035)	0.037 (0.065)	0.297
Dalby Box 2016	0.127 (0.045)	0	0.127	0.096 (0.038)	0.041 (0.022)	0	0.137
Emerald 2015	0.197 (0.074)	0.097(0.031)	0.294	0.158 (0.068)	0.063 (0.057)	0.077 (0.031)	0.298
Emerald 2016	0.171 (0.052)	0	0.171	0.159 (0.021)	0.015 (0.003)	0	0.174
Gatton 2015	0.423 (0.139)	0	0.423	0.342 (0.079)	0.079 (0.042)	0	0.421
Hermitage 2015	0.570 (0.170)	0.024(0.049)	0.594	0.473 (0.161)	0.190 (0.038)	0	0.663
Hermitage 2016	0.455 (0.144)	0	0.455	0.342 (0.133)	0.162 (0.017)	0	0.504
Jimbour 2015	0.304 (0.101)	0.041(0.031)	0.345	0.271 (0.093)	0.076 (0.018)	0.019 (0.031)	0.366
Jimbour 2016	0.078 (0.025)	0	0.078	0.059 (0.022)	0.027 (0.025)	0	0.086
Liverpool Plains 2015	0.213 (0.087)	0.140(0.054)	0.353	0.198 (0.083)	0.049 (0.044)	0.111 (0.055)	0.358
Orion 2016	0.060 (0.018)	0	0.060	0.048 (0.020)	0.017 (0.005)	0	0.065
Pirrinuan 2016	0.113 (0.044)	0	0.113	0.058 (0.022)	0.056 (0.025)	0.023 (0.022)	0.137
Spring Ridge 2016	0.140 (0.013)	0	0.140	0.101 (0.011)	0.050 (0.030)	0	0.151

Table 5.4: Prediction accuracy for the additive genetic variance from the FA2 additive model (FA2.A) and the FA2 dominance model (FA2.AD). The values are presented as percentages.

Trial	FA2.A	FA2.AD
Blackville 2016	52	57
Capella 2015	11	11
Croppa Creek 2016	37	67
Dalby Box 2015	57	57
Dalby Box 2016	36	45
Emerald 2015	76	80
Emerald 2016	67	77
Gatton 2015	62	64
Hermitage 2015	78	79
Hermitage 2016	67	63
Jimbour 2015	67	68
Jimbour 2016	71	76
Liverpool Plains 2015	22	35
Orion 2016	79	75
Pirrinuan 2016	62	71
Spring Ridge 2016	40	57

and dominance (FA2.AD). Both models showed a large range in values across trials, for model 4 the percentage of similarities ranged from 25% to 80% with model 2 having a larger range of percentages from 13% to 97%.

5.6.4 Changes in Hybrid Selection

These observed changes in top10% rankings showed that selections of hybrids from individual trials are variable compared to selecting hybrids from an across trial prediction. These results showed that the top 10% of hybrids change between trials and they change when using models with and without dominance. The percentages change between each trial and model, for example only 35% of the across trial predictions yield in the top 10% at Blackville 2016, but after including dominance 56% yield in the top 10%. 52% of the across trial predictions are in the top 10% of the Orion 2016 trial with additive only but this reduces to 27% when adding in dominance.

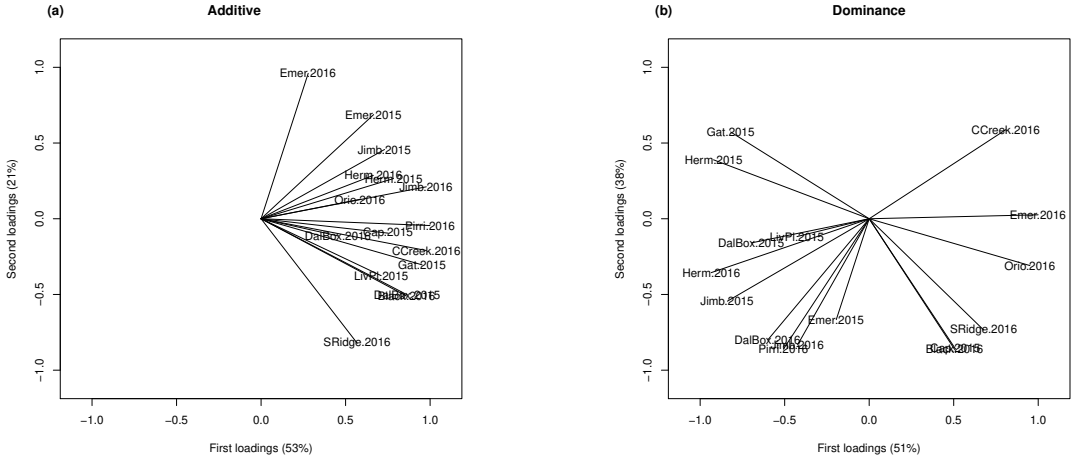

Figure 5.1: The rotated loadings from the FA2.AD analysis, (a) loadings from the additive partition and (b) loadings from the dominance partition.

The first order FA loadings for the additive partition were positive for both FA2.A and FA2.AD models, where they explain 48% and 53% for each model respectively. Given these results, the FAST method of Smith and Cullis (2018) was applied to both FA2.A and FA2.AD (Figure 5.3). FA2.A showed higher values for the root mean square deviation (RMSD) indicating that the overall performance of hybrids deviated more from the average than those from the FA2.AD analysis. Furthermore the colouring of the female parent showed that the inclusion of dominance in the model created a separation of the 3 parents. This showed that parents can differ in their stability across environments.

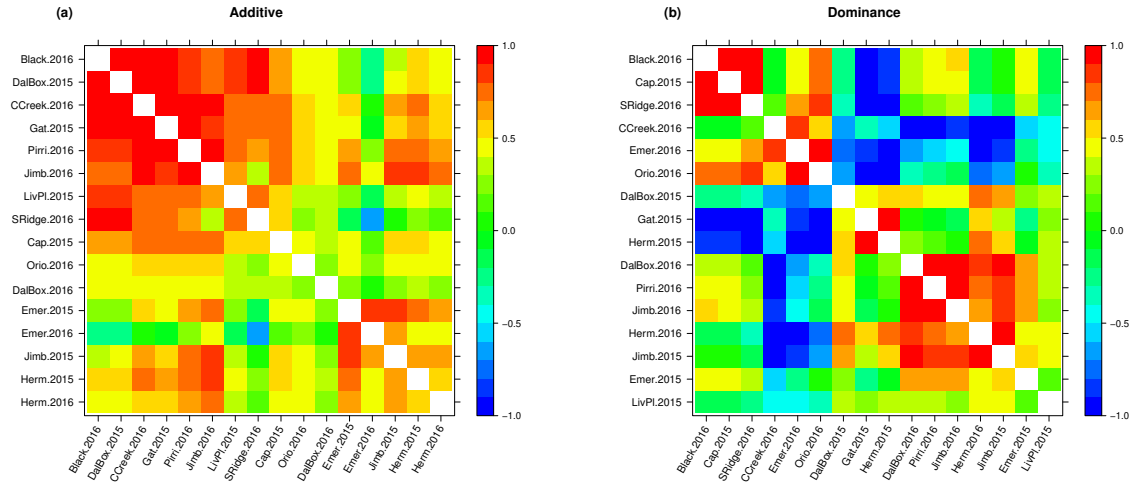

Figure 5.2: Heatmaps showing correlations from the FA2.AD analysis, (a) between trial correlations for the additive partition and (b) between trial correlations for the dominance partition.

Preferable hybrids for selection will be those with the smallest amount of deviation (small RMSD) and highest overall performance. Figure 5.3 showed that the selected hybrids based on this criteria changes between analyses with and without dominance. For the FA2.A analysis the preferable hybrids are coloured in green, whereas in FA2.AD they are red. (see also Appendix B Figure B.3).

5.7 Discussion

Our results demonstrated that partitioning genetic variance into additive, dominance and residual genetic variances was a significantly better model for these data than just considering additive effects and residual genetic effects without the dominance partition. We showed that the GxE effects accounted for a significant amount of trait variation since an FA2 model fitted better than an FA1 model for both the additive and dominance terms.

Despite their small numbers, the testers used in this study were specifically chosen to expose the variation present in the male parents on different ways in different environments. In particular the female testers were known to exhibit different effects for stay-green, which is a drought resistance trait that is expressed when water stress occurs during the grain filling period. By partitioning the genetic variance into additive, dominance and residual genetic parts, more accurate effects for hybrids can be examined across environments.

Table 5.5: Percentage of hybrids in common between the top 10% ranking of the across trial effects and the additive effect for FA2 models without dominance (FA2.A) and with dominance (FA2.AD).

Trial	FA2.AD (%)	FA2.A (%)	Difference (%)
Blackville 2016	56	35	-21
Capella 2015	47	45	-3
Croppa Creek 2016	45	57	13
Dalby Box 2015	44	35	-9
Dalby Box 2016	42	27	-15
Emerald 2015	51	59	8
Emerald 2016	25	36	11
Gatton 2015	49	49	0
Hermitage 2015	62	67	5
Hermitage 2016	42	44	2
Jimbour 2015	60	66	6
Jimbour 2016	80	97	17
Liverpool Plains 2015	52	38	-14
Orion 2016	27	52	25
Pirrinuan 2016	69	75	6
Spring Ridge 2016	26	13	-13

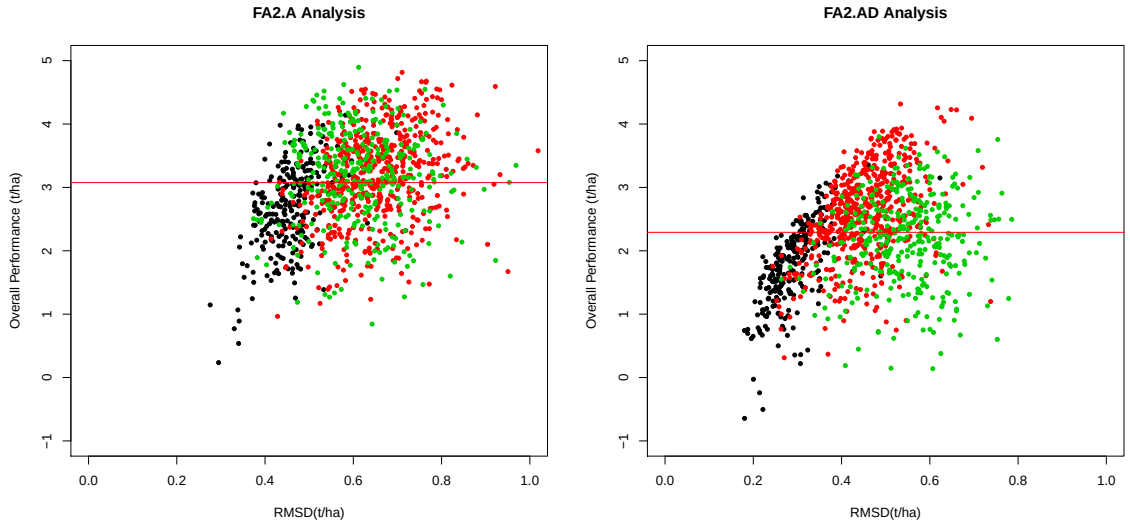

Figure 5.3: Additive overall performance versus root mean square deviation (RMSD; a stability measure) for FA2.A on the left and FA2.AD on the right. Colours represent the three Female parents.

5.7.1 Partitioning additive and dominance effects increases prediction accuracy

Partitioning additive and dominance effects increases the statistical fit of the data as shown by comparing models with and without dominance. The additive genetic variance decreased when dominance was added to the model. This indicated that a model that does not include dominance over estimates the contribution of additive genetic variation. The lower standard errors in the dominance model also confirmed that the addition of dominance gave more accurate additive effects. Generally, the prediction accuracy based only on the additive effects increased with the addition of the dominance

term. This has consequences for genetic and genomic evaluations - if dominance is not included in the model for these evaluations, the resulting estimated breeding values may be biased. This has considerable implications for breeding programs using breeding values to select parents for crossing. So including dominance in the model for genomic predictions should result in more accurate selection of lines on breeding values.

5.7.2 Dominance effects span wider correlations

When considering joint analysis of a multi-environment trial possible GxE can be considered for each of the partitions of genetic variance separately. We observed that the pattern of GxE that was exhibited by the additive and dominance partitions was quite different; between trial correlations for the additive effects had a high average correlation while the correlations for the dominance effects spanned from negative to positive across trials with a low average correlation. This may be an artifact of these type of models that include no hybrid main effect, and the additive proportion of the genetic variance to some degree includes these main effects. This does not pose a problem when the aim of the analysis is to predict the average across trial additive effect for hybrids. The inclusion of a dominance effect in the model can possibly pick up some of the additive effect that may have not been accounted for when fitting a model without dominance effects.

The hybrids change in ranking between trials and also between models with or without dominance. This means if breeders were to select hybrids using only predictions from individual trials, their selection would vary between trials. There is a danger that breeders can be either choosing low yielding lines or discarding high yielding lines for other environments. Selection should take into account the environment type as well as the percentage of dominance present at each trial.

The dominance model proposed here has demonstrated the importance of GxE in the performance of sorghum hybrids. The dominance partition of the genetic variance is highly influenced by the environment. Across trial multi-environment analyses can have different results depending on the environments that have been represented by the trials in the analysis. Selection should be made by combining as many trials as possible to predict an average performance of a hybrid in a target population. Using a factor analytic variance covariance structure enables accurate across trial prediction by enabling the use of pairwise between site genetic correlations. The resulting predictions will be representative of the average environment demonstrated by the trials in the analysis.

Chapter 6

Identifying efficient strategies for preliminary evaluation in hybrid breeding programs

6.1 Introduction

In hybrid breeding programs the evaluation of F1 hybrid combinations has two purposes, the first is to identify parents with good performance on average when combined with other parents in hybrids (called general combining ability, GCA) and the second is identify specific superior hybrid combinations (called specific combining ability, SCA). Typically in the early stages of a hybrid testing program it is impossible to test large numbers of combinations so breeders focus on identifying lines with high levels general combining ability and subsequently search for superior specific combinations. Typical hybrid sorghum breeding programs involve early-stage selections on the basis of performance with a single elite tester.

Hybrid breeders are focused on the identification of lines with optimal general combining ability, the phenomenon displayed only when inbred lines are crossed with each other, complementing each other in desired traits. For practical reasons it is impossible to perform all cross combinations therefore, in a standard F1 breeding scheme, the first step is testing for general combining ability, by testing of a large number of lines with a single tester line (Rudolf-Pilih et al., 2019). Similarly a pre-breeding program developing improved germplasm to be used by hybrid programs will aim for improvement in general combining ability. Based on progeny performance, the best inbreds are chosen for specific combining ability (SCA) testing, defined as the specific interaction between the two parents of the hybrid. A major challenge with this approach is achieving adequate testing of the inbreds to evaluate their likely performance in all pairwise possible combinations (Hallauer et al., 2010).

Heterotic groups can be defined as sets of lines deriving from a common origin and displaying similar combining ability when crossed with lines from different origins (other heterotic groups). These heterotic groups are generally unrelated to one another by pedigree and crosses between them produce superior hybrids (Melchinger and Gumber, 1998; Meena et al., 2017). In sorghum, hybrids are made using the cytoplasmic male sterility system this means that there are effectively two heterotic

groups, restorers (male parents) and maintainers (female parents). Hybrids are grown in two distinct trials using a scheme known as the North Carolina II mating design. Early generation, or preliminary trials typically involve crossing lines from a heterotic group with a one or two tester lines from a complementary heterotic group. The number of testers used is limited by resource constraints given that adding each new tester creates thousands of potential hybrids within each heterotic group $(n_1 \times n_2)$, where n_1 is the number for one group and n_2 is the other group (Guo et al., 2019). Free from resource constraints, it would be ideal to test all combinations of possible parents early in the hybrid breeding. The advantage of early evaluation of all potential single crosses is to identify the best parental combination immediately after progeny development (Kadam et al., 2016). Selection of progenies only on the basis of a single cross tester leaves open the possibility that some unique parental combinations never made and tested could be superior in performance and become commercial hybrids. Despite these advantages, field testing of all potential single crosses of inbred progenies is completely impractical for a mature hybrid breeding program.

An effective tester for the early stages of hybrid breeding programs should be able to rank inbred lines correctly for performance in hybrid combinations and increase the differences between test-crosses (relative to standard to standard errors) for efficient discrimination (Annor et al., 2020). A good tester should have the capacity to reveal high genetic variance between hybrids but must also be representative of the heterotic group in order to make effective selections. Given the impact of GxE and dominance across environments (Hunt et al., 2020), the best tester may not the same for all environments. In these circumstances there is a need to find an optimal strategy to identify the best single tester or combination of multiple testers given the breeder has limited knowledge on the future characteristics of specific environments.

With the advances in genomic prediction (Meuwissen et al., 2001) it may be possible to predict hybrid performance from untested hybrids based on their relationship to the hybrids in a training data set. Several studies have indicated the usefulness of genomic selection to predict hybrids in maize (Albrecht et al., 2011, 2014; Fritsche-Neto et al., 2018). However, most of the experimental studies have focused on predictions based mainly on a single tester scenario (Albrecht et al., 2014). Therefore, the most critical point is the choice of a tester to evaluate the lines general combining ability. However when the phenotypic evaluation of lines is performed with a single tester, the effects of general and specific combining ability cannot be separated (Albrecht et al., 2011). Hence, the real breeding values of the parents may be masked by the interaction with the tester, then predictions obtained within the same group but with a different tester can disappointingly low (Albrecht et al., 2014; Fritsche-Neto et al., 2018). For hybrids there is a need to generate genomic predictions using covariance matrices for both additive and dominance relationships (Guo et al., 2019). For early generation breeding trials that have only a limited number of tester parents the calculation of dominance is problematic and therefore hybrid prediction is restricted to general combining ability.

There is a need for an optimal strategy for producing and testing representative hybrids in early generation trials when resources are limited by costs and management. Guo et al. (2019) compared strategies for choosing a small number of hybrids as a training set for predicting the larger hybrid

population. In this study we will compare differing proportions of hybrid combinations while keeping the total number of hybrids the same. The aim is to compare the predictions of hybrids based on a single tester with those that involve two testers.

This chapter investigates the optimum allocation of hybrid combinations in early generation trials to identify general combining ability using genetic and genomic relationships where resources are constrained. The aim is to consider genomic and pedigree predictions for a range of different combinations of testers within two distinct heterotic groups including the analyses of the hybrids from each single tester

6.2 Materials and Methods

6.2.1 Phenotype Data

We considered a set of twelve trials from the 2016, 2017 and 2018 sorghum pre-breeding program conducted by the Queensland Department of Agriculture and Fisheries and the Queensland Alliance for Agriculture and Food Innovation. The trials used in this study are structured to evaluate the two heterotic groups of sorghum and are designated advanced yield trials for males (AYTM) and advanced yield trials for females (AYTF). The main aim of these trials is to identify elite male or female parents, with high general combining ability (i.e. additive genetic value) for release to commercial breeding companies. The males, or restorer parents, have more genetic diversity than the maintainer parents, thus the hybrids in the male breeding trials also have greater diversity.

The trials evaluated a total of 1389 inbred lines, comprising of 850 female B lines and 539 male R lines. The AYTF trials have a total of 1351 genotyped hybrids and the AYTM trials have a total of 946 genotyped hybrids (Table 6.1). Trials were designed with partial replication (Cullis et al., 2006), where between 30% and 50% of the hybrids were replicated and the remaining hybrids had a single replicate. Hybrids were laid out using a spatial row-column design with the replicated hybrids resolved into two equal blocks. These designs enabled the trials to be analysed using linear mixed models with random genetic effects and including spatial effects for each trial. Table 6.1 shows the numbers of lines that were crossed to both testers in each trial of both the AYTM and AYTF series over the trials considered.

The data of interest here are the yields expressed in tonnes hectare (t/ha) obtained from harvesting in the year after the crop was planted.

6.2.2 Pedigree and Genotype Data

Ancestral pedigree information was available for all genotyped hybrids for up to 20 generations of ancestry. For the AYTM data there were 1767 unique ancestral lines included in the full pedigree file, including the 539 genotyped male lines and the 2 female testers. Also included in the pedigrees were 120 founder lines with unknown parents. With the inclusion of the 946 hybrids present in the trials, the number of lines in the pedigree file totalled 2713. The average inbreeding coefficient of the lines was 0.36, ranging from 0 to 1.96, the genetic connectivity in the design was high with an average

Table 6.1: Number of genotyped hybrids, number of hybrids from each tester and the number of lines crossed to both testers for AYTF and AYTM trials

	AYTF				AYTM			
	Hybrids	Tester 1	Tester 2	Both Testers	Hybrids	Tester 1	Tester 2	Both Testers
2016.Black	482	258	225	197	706	405	302	295
2016.CCreek	471	253	219	186	686	394	293	282
2016.DBox	493	262	232	205	740	425	316	299
2016.Herm	506	268	239	212	784	452	333	316
2016.Jimb	475	255	221	192	723	413	311	304
2016.Orion	440	236	205	167	678	390	289	280
2016.Pirri	459	249	211	175	754	425	330	312
2016.SRidge	500	265	236	214	763	434	330	308
2017.Black	838	336	503	289	458	244	215	187
2017.Maca	831	341	491	268	443	235	209	178
2017.SRidge	847	356	492	269	470	252	219	188
2018.Pampas	835	303	533	272	462	244	219	189

additive correlation of 0.662 between the lines. For the AYTF data there were 2223 unique ancestral lines included in the full pedigree file, including the 850 genotyped female lines and the 2 male testers. Also included in the pedigrees were 69 founder lines with unknown parents. With the inclusion of the 1351 hybrids present in the trials, the number of lines in the pedigree file totalled 3574. The average inbreeding coefficient of the lines was 0.46, ranging from 0 to 1.96, the genetic connectivity in the design was high with an average additive correlation of 0.585 between the lines.

Genotypic data in the form of 18783 SNP markers were available for all 1389 parent lines including both testers from each trial series (946 genotyped hybrids for AYTM and 1351 for AYTF). Genotypes for the hybrids in the trials were created by combining the marker values for the male and female parents of each hybrid. At each locus the markers were coded as “00” and “11” for the homozygotes and “01” for the heterozygotes.

Generally the AYTF hybrids have closer relationships than the AYTM hybrids within each tester. The relatedness between testers is lower for the pedigree data than the marker data for both AYTF and AYTM (see Appendix Figure C for PCA of heterotic groups).

6.2.3 Statistical Models

For each of the 12 trials within the male and female data 2 linear mixed models were fitted using the method described in Hunt et al. (2018). The first model incorporated markers and the second incorporated pedigree information. The fitted model was written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_{h_a}\boldsymbol{u}_{h_a} + \boldsymbol{Z}_{h_e}\boldsymbol{u}_{h_e} + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \tag{6.1}$$

The vectors $\boldsymbol{\tau}, \boldsymbol{u}_{h_a}\boldsymbol{u}_{h_e}, \boldsymbol{u}_p$ represent fixed effects, additive random effects for hybrids, residual (non-additive) genetic random effects for hybrids and random non-genetic (or peripheral, ie design and additional) effects respectively. $\boldsymbol{X}$, $\boldsymbol{Z}_{h_a}$, $\boldsymbol{Z}_{h_e}$ and $\boldsymbol{Z}_p$ are the design matrices for the fixed effects, additive and residual genetic effects (effects not accounted for by the additive term), and the random non-genetic effects, respectively, and $\boldsymbol{e}$ is the random residual term.

The fixed effects ($\boldsymbol{\tau}$) included in the baseline model included a covariate for establishment at each trial which is a function of both the number of plants per plot and the distribution of gaps in plots with less than the target number of plants. For each site the baseline spatial randomisation model included random effects ($\boldsymbol{u}_p$) for replicate, where replicate is a factor with 2 levels representing random effects between the resolvable replicated entries. Random effects ($\boldsymbol{u}_p$) also included a row effect for each site where row has levels equal to the number of rows in each site. Both trials at 5 locations (2016.Black, 2016.Herm, 2016.Orion, 2017.Black, 2018.Pampas) had extra rows of missing data inserted to account for bulk crop rows that allowed spraying operations without damaging test plots. The variance model for $\boldsymbol{e}$ contained the Kronecker product of first order auto-regressive processes in the row (AR1_r) and column (AR1_c) directions respectively.

The non-genetic terms, including the residual effects $\boldsymbol{e}$ and the peripheral effects $\boldsymbol{u}_p$ as well as the fixed effects $\boldsymbol{\tau}$ were calculated using the total number of lines in the data. The genetic effects $\boldsymbol{u}_{h_g}$ and $\boldsymbol{u}_{h_e}$ were based on genotyped hybrids only. Hybrids without genotypic data were retained to preserve the spatial effects but did not contribute to the estimate of genetic variance parameters by inclusion of a fixed effect with 2 factor levels that distinguishes between genotyped and non-genotyped lines. It was assumed hereafter for ease of computation that all design matrices conformed to allow for the discrepancies in number of genotyped hybrids versus number of phenotyped hybrids by the inclusion of zeros where no effect is present.

The additive genetic effects $\boldsymbol{u}_{h_a}$ and the residual genetic effects $\boldsymbol{u}_{h_e}$ were independent and Gaussian with zero mean and variance matrices given by $\sigma^2_m\boldsymbol{A}_m$, for a marker based relationship or $\sigma^2_a\boldsymbol{A}$ for a pedigree based relationship and $\sigma^2_e\boldsymbol{I}_m$ for the residual genetic term, where σ^2_m, σ^2_a and σ^2_e are the marker based additive variance, pedigree based additive variance and residual genetic variance respectively.

The relationship matrices $\boldsymbol{A}_m$, the relationship matrix formed using markers, and $\boldsymbol{A}$, the relationship matrix formed using pedigrees were calculated as described in Hunt et al. (2018) and Hunt et al. (2020).

6.2.4 Prediction Accuracy

To look at differences between testers in the two trial series a cross validation procedure was conducted. This involved removing hybrids from the data so that the parent lines were combined in different proportions. The fitted model was re-run for each of the combinations listed in Table 6.2. In order to preserve the residual and spatial errors involved in each trial, the genotypes were partitioned into 2 parts, the validation set included the removed hybrids and the training set included the remaining hybrids. The variance component for the residual and spatial terms were fixed so that they were the same in the analysis of each run. Pediction error variances were calculated using the method in section 4.3.5.

Combinations that include all the data (1/1) and those where only a single tester is present(1/0 and 0/2) have a single representation so there is only one set of data fitted for each. All the other combinations were run using random samples for each combination in Table 6.2, the results represent

the average of repeating the analysis using 10 different random samples of lines within each tester.

Table 6.2: Proportions of each tester used in the analysis.

Tester 1	Tester 2					
	1	0.8	0.6	0.4	0.2	0
1	All data	✓	✓	✓	✓	single tester
0.8	✓	✓	✓	✓		
0.6	✓	✓	✓			
0.4	✓	✓				
0.2	✓					
0	single tester					No data

6.3 Results

6.3.1 Genetic Variances

The genetic variance from models where all hybrids are present and those where only hybrids from each single tester were included in the data are shown in Table 6.3, for female AYTF trials and Table 6.4 for the males AYTM trials. Generally the hybrids have different genetic variances within each tester with the larger genetic variance varying between testers. For example in the AYTF trials 2016.DBox tester 2 had a larger genetic variance than tester 1, but for 2016.Jimb tester 1 was larger than tester 2 (Figure 6.1). The genetic variance in marker and pedigree models also varied between trials, most of the trials were similar, for example 2017.SRidge had a genetic variance between the testers of 0.102 and 0.89 for the markers and 0.101 and 0.108 for the pedigrees. Other trials were very different, for example 2018.Pampas had genetic variances for the testers of 0.167 and 0.189 for the markers and 0.238 and 0.294 for the pedigrees.

For the male AYTM trials (Table 6.2) there is a similar agreement between the marker analyses and the pedigree analyses but the variation among testers was still variable between trials. Tester 1 had the larger genetic variance at 2016.SRidge and 2018.Pampas, and tester 2 had the larger genetic variance at 2016.DBox and 2016.Pirri. Figure 6.2 showed the distributions of the predicted BLUPs for the hybrids within each of the testers along with the distribution of all of the hybrids for marker and pedigree analyses.

6.3.2 Female trials

The result from fitting 16 combinations of inbred parents for both marker and pedigree analyses indicate that the superior predictions of untested hybrids come from the combinations where both testers were present.

Figure 6.3 shows the correlations between predictions of untested hybrids against their corresponding predictions from the analysis of the full set of Hybrids. The combinations with the highest correlations are those with green cells. The best combination of lines varies between trials. For

Table 6.3: Genetic variance of all the hybrids and within each tester for the AYTF trials from both the marker and pedigree models.

	Markers - GBLUP			Pedigree - PBLUP		
Trial	Genetic Var	Tester 1	Tester 2	Genetic Var	Tester 1	Tester 2
2016.Black	0.231	0.156	0.267	0.236	0.197	0.246
2016.CCreek	0.088	0.075	0.131	0.143	0.106	0.150
2016.DBox	0.120	0.070	0.175	0.126	0.079	0.144
2016.Herm	0.720	0.591	0.509	0.734	0.634	0.652
2016.Jimb	0.097	0.131	0.039	0.142	0.108	0.055
2016.Orion	0.034	0.097	0.000	0.027	0.070	0.016
2016.Pirri	0.061	0.059	0.058	0.057	0.043	0.047
2016.SRidge	0.125	0.063	0.148	0.105	0.062	0.107
2017.Black	0.076	0.035	0.118	0.065	0.042	0.083
2017.Maca	0.104	0.102	0.092	0.102	0.099	0.088
2017.SRidge	0.094	0.102	0.089	0.090	0.101	0.108
2018.Pampas	0.331	0.167	0.189	0.346	0.238	0.294

Table 6.4: Genetic variance of all the hybrids and within each tester for the AYTM trials from both the marker and pedigree models.

	Markers - GBLUP			Pedigree - PBLUP		
Trial	Genetic Var	Tester 1	Tester 2	Genetic Var	Tester 1	Tester 2
2016.Black	0.112	0.117	0.134	0.166	0.203	0.155
2016.CCreek	0.068	0.066	0.086	0.091	0.111	0.112
2016.DBox	0.112	0.055	0.107	0.110	0.038	0.106
2016.Herm	0.519	0.446	0.538	0.508	0.495	0.536
2016.Jimb	0.078	0.109	0.028	0.103	0.137	0.025
2016.Orion	0.053	0.050	0.046	0.057	0.057	0.050
2016.Pirri	0.084	0.051	0.107	0.106	0.089	0.131
2016.SRidge	0.092	0.116	0.100	0.103	0.137	0.082
2017.Black	0.047	0.040	0.046	0.038	0.032	0.041
2017.Maca	0.102	0.069	0.074	0.113	0.075	0.066
2017.SRidge	0.037	0.000	0.052	0.031	0.002	0.052
2018.Pampas	0.072	0.071	0.058	0.064	0.047	0.035

example using the data from the 2016 Orion trial the best fitted analyses were those that contain a majority of Tester 2, which was also the tester that has the highest genetic variance for that trial. Similarly for 2016 Spring Ridge and 2017 Blackville the best fitted analyses were those containing a majority of Tester 1, which generated the higher genetic variance for both those trials. Generally the highest R-squared values were when combinations of both testers which were represented by the green cells in the centre of each plot in Figure 6.3. For the analyses involving pedigrees the low errors when using a single tester were even more prominent than in the analyses involving markers. Generally the analyses that use marker and pedigree information agreed with respect to which combinations were the most accurate for each trial.

Figure 6.4 shows the full set of BLUPs for the hybrids plotted against the predicted hybrid effects for the hybrids when they were removed from the data. For both trials it was shown that the correlations of the predictions that involved a combination of both testers were superior. The pedigree analyses

agree with the marker analyses, in general the plots where the points that have only a single colour were not in as good agreement with the full data BLUPs as those where both testers are represented.

The average prediction error variances of the predicted hybrids are shown for all tester combinations for all trials in Figure 6.5. The tester combination with the highest PEV was the lowest one on the y-axis. For the marker analysis, for 7 of the 12 trials the combination with the highest PEV was not a combination that included 100% of one tester. For 2017.Black, 2017.SpringRidge and 2018.Pampas the best combination involved 100% of tester 2 and for 2016.Orion the best combination was all of the hybrids with 100% of tester 1. Only one site, 2016.Jimb was superior with just a single tester. For the pedigree analysis 10 of the 12 trials had the highest PEV for partial combinations of testers and only 2017.Black and 2016.SRidge had the highest PEV for the analysis that used 100% of a single tester.

6.3.3 Male trials

The best combination of lines for the Male trials varied between trials in a similar way to the Female trials but the use of a single tester was not as distinct as with the female trials (Figure 6.6). Figure 6.3 shows distinct bands of low values (in blue) on the right side (100% tester 1) and across the top (100% tester 2). In contrast Figure6.6 showed higher values in those positions without obvious banding. For the 2016.Pirri trial the least accurate combinations were those with partial frequencies for both testers. 2016.DBox, 2016.Orion and 2017.Maca showed the lowest accuracy for combinations that used 100% of Tester 2. 2016.Black and 2016.Jimb were lower for Tester 1. The pedigree analyses were, in general, less accurate based on the correlations in Figure 6.6. Two trials in particular, 2016.Pirri and 2016.DBox showed an obvious decrease in correlation for the pedigree analysis compared to the marker analysis.

The plots of the full analysis BLUPs versus the predicted hybrid BLUPs (Figure 6.7) were similar for the marker and pedigree analyses. The trial 2017.SRidge showed the zero genetic variance for tester 1 and there was an obvious drift of the values as more of the second tester was included. The pedigree results showed some separation between the testers indicating that a main effect for tester has not been accounted for in the analysis. The two female testers were closer genetically than the male lines they have been crossed with, the marker analysis more accurately predicted the hybrids in this case.

For the marker analysis of the male trials, the tester combination with the highest PEV included 100% of one of the testers in 8 sites (Figure 6.8). For all but 2 sites (2016.Jimb and 2017.SRidge) the combination with the lowest prediction error was 100% of tester 1 and 0% of tester 2, indicated by the lighter colours at the bottom right corner of each panel. The standard errors for the pedigree analyses were distinctly different from the marker analyses for the male trials. Overall the male trials had higher PEVs indicated by the darker colours. For 8 of the 12 trials marker analysis showed distinctly higher PEV for 100% of tester 1 and 0% tester 2, whereas the pedigree analysis did not show this.

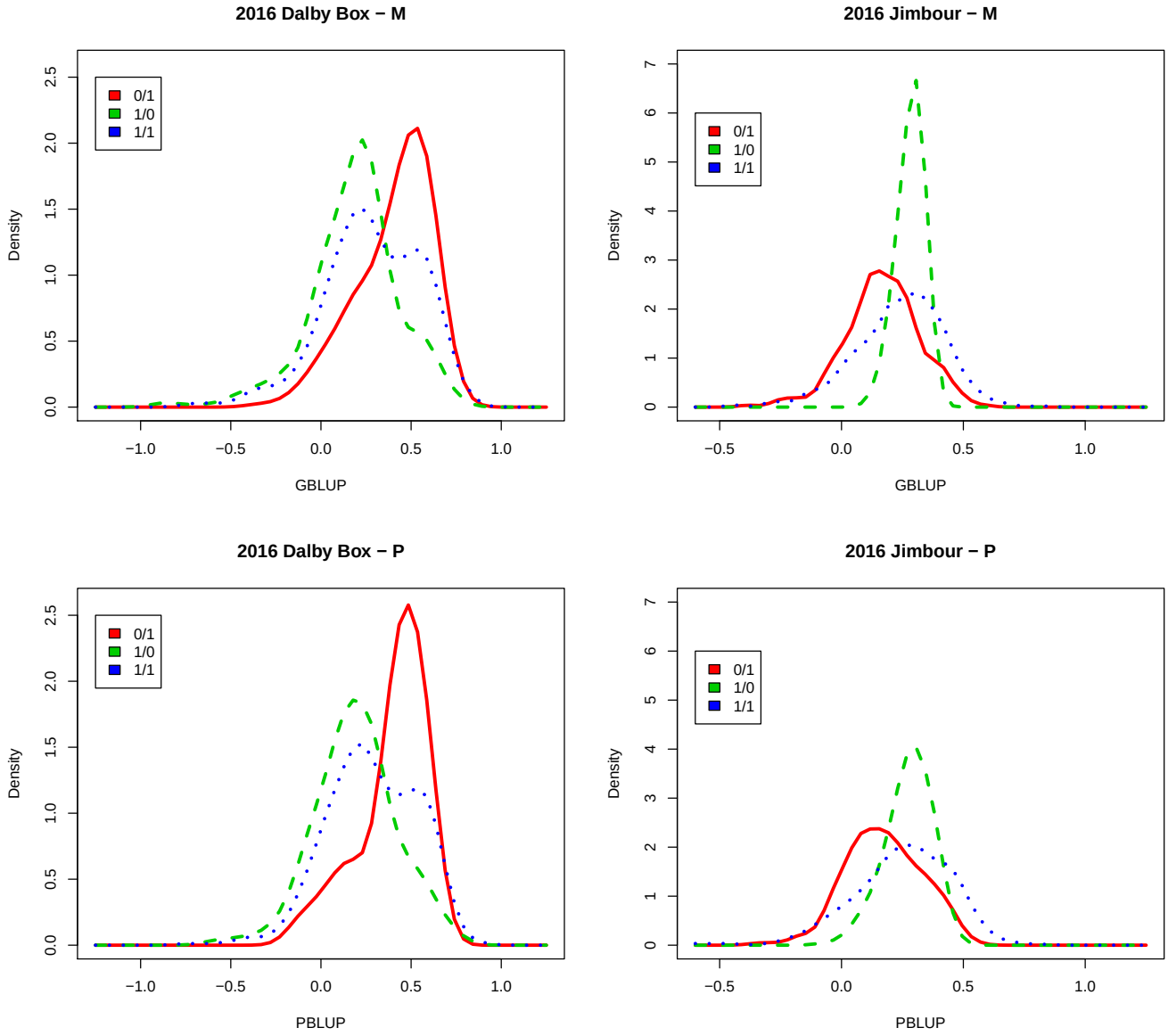

Figure 6.1: Density plots for AYTF trials 2016 Dalby Box and 2016 Jimbour, marker model above and pedigree model below. Red shows the distribution of the predicted values for hybrids from tester 1, green shows the predicted values for hybrids from tester 2 and the blue dotted line is the density for the pedicted values from the analysis of all hybrids.

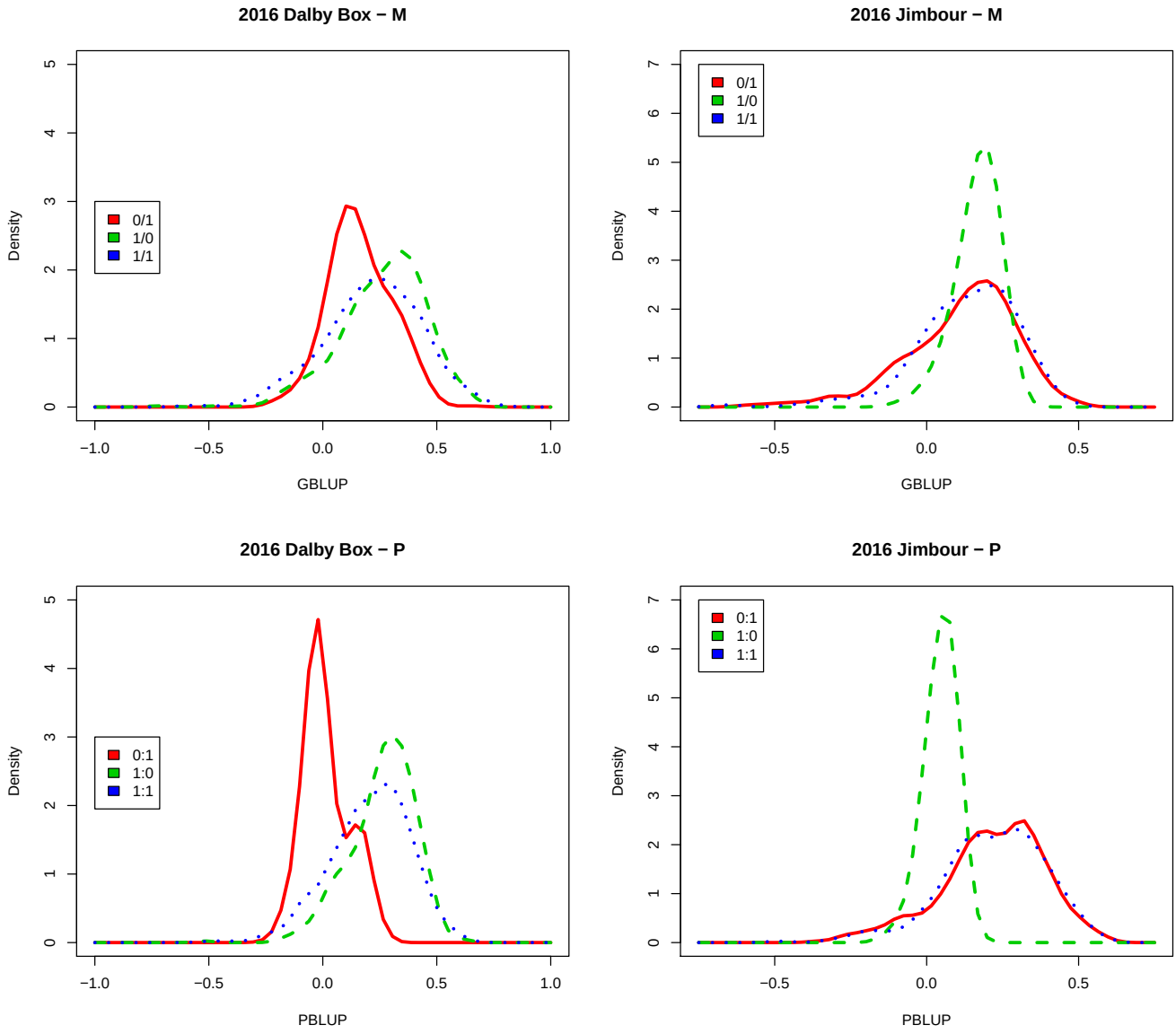

Figure 6.2: Density plots for AYTM trials 2016 Dalby Box and 2016 Jimbour, marker model above and pedigree model below. Red shows the distribution of the predicted values for hybrids from tester 1, green shows the predicted values for hybrids from tester 2 and the blue dotted line is the density for the pedicted values from the analysis of all hybrids.

6.4 Discussion

Results in this chapter have shown that in hybrid breeding programs where one or two testers are used to identify lines with high general combining ability, resources can be used more efficiently by using unbalanced combinations of multiple testers. The optimal number of testers and the optimal combination is dependent on the heterotic group and the environment. Breeders can use genomic prediction to their advantage to assess the performance of untested hybrids that involve combinations of inbreds and tester lines that are included in the training set of hybrids (Kadam et al., 2016). Hybrid trials that use a single tester vary in their capacity to predict the general combining ability of the test lines. Where specific combining ability is important any estimates of GCA based on a single tester will be confounded. This situation is further complicated if hybrid ranking is influenced by environment (Hunt et al., 2020).To decrease this risk the optimal strategy would be to use a greater number of testers but this greatly increases the resources required.

In this study we have considered the optimal use of two testers when resources are constrained in two different heterotic groups tested in multiple environments.

Accuracy can be increased without increasing resources required

The analyses have shown that in a trial with an unbalanced set of hybrid combinations, untested hybrids can be accurately predicted using either genetic or genomic information. It is possible to predict the hybrid performance of an inbred in combination with a different tester based on the relationship of the untested hybrid to those that have been phenotyped. As previously stated the use of a single tester is inherently inefficient for predicting GCA due to the confounding effects of SCA.

The study shows that in general in individual trials a set of inbred lines are more effectively evaluated for GCA using multiple testers even when the total number of plots in the trial remains the same. The study showed that the prediction accuracies of untested hybrids were very high (up to 90%) when all lines are present and crossed to either one or two testers even when the number of lines crossed to both testers is small.

Another advantage of designing preliminary trials with multiple testers is the capacity for early identification of the superior performing hybrids.

Heterotic groups vary in their prediction capacity

The results have shown that the value of using multiple testers depends on the SCA variance of the heterotic group and the genetic distance between the chosen testers.

In the current study the male heterotic pool is more diverse than the female heterotic pool and the two female testers used to test the male heterotic pool are more similar than the two selected male testers. This limits the inferences that can be made. For the female trials the most accurate predictions were estimated using both male testers in combination. The male trials showed higher difference in genetic variances generated by the two female testers but the correlations between the untested

predictions and tested predictions were generally higher. The results indicate that the optimum number and diversity of testers will depend on the material being evaluated.

Genotype by Environment Interaction impacts tester effectiveness for prediction

The analysis of 12 trials using two methods of estimating relationships (markers and pedigrees) and in two heterotic groups (male and female) has shown that the optimum combination of testers varies dramatically between trials. The standard method of analysis of plant breeding trials is to combine individual trials into a single multi-environment analysis and calculate across trial genetic effects (e.g. Cooper and DeLacy (1994),Annicchiarico (2002), Malosetti et al. (2013)). Using a single tester in preliminary trials to produce average across environment predictions is surprisingly common (Albrecht et al., 2014). However this study has highlighted that estimates of line performance can vary significantly between environments and testers due to the interaction of GxE and dominance (Hunt et al., 2020).

The standard practice to analyse plant breeding trials by combining many trials into a single MET analysis also allows for the determination of correlation between trials. Using the information gained from the between trial correlations together with the relationship between hybrids allows for prediction of unphenotyped hybrids in specific environments. This study has shown that it would be potentially misleading to combine trials that contain hybrids that do not share testers due to a degree of confounding between tester performance and between trial correlations. This difficulty can be overcome by using multiple testers in all trials. This would enhance the use of MET analysis by having representative genetic material in all trials and therefore increase the prediction ability of untested lines in untested environments. This also attempts to adjust for the change of genetic material over time as addressed by Albrecht et al. (2014).

Further increase in prediction accuracy would result from allowing different proportions of testers in different environments but retaining some proportion of each tester.

The choice of tester usually involves qualities apart from their yield capacity. In sorghum traits such as stay-green strongly influenced the performance of the tester lines in some environments. Clustering environments into groups of trials that share common effects due to stay green would allow different testers to be used in different stay green environments to obtain accurate across site hybrid predictions (Velazco et al., 2019).

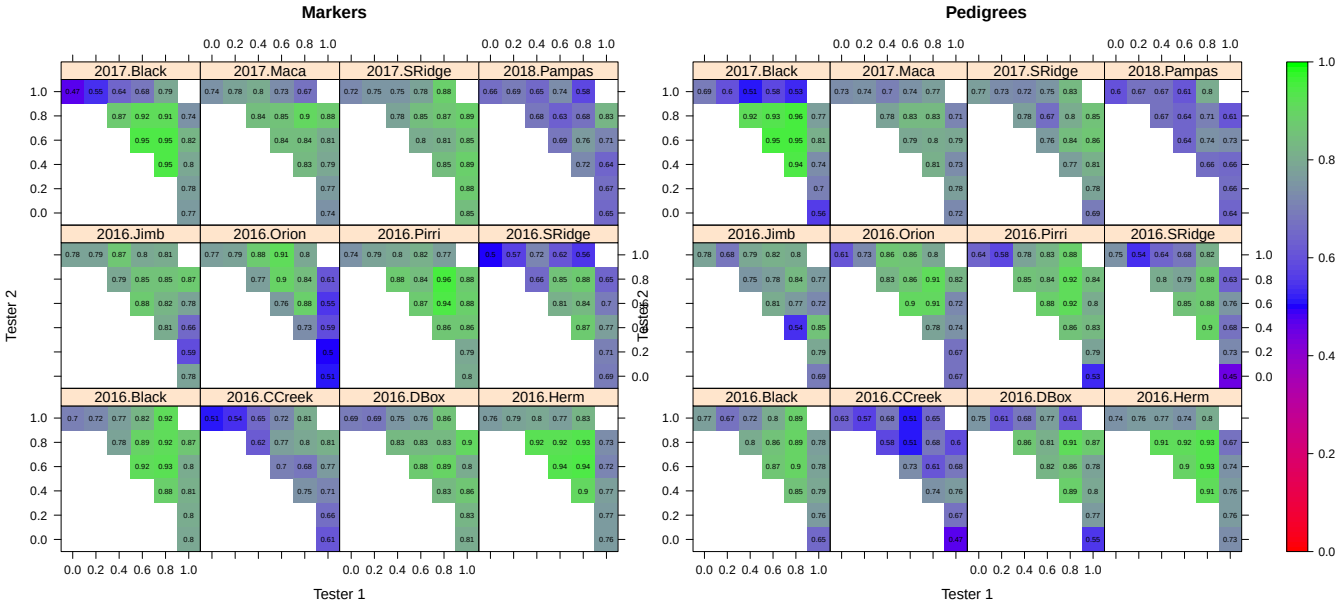

Figure 6.3: AYTF trials correlations between genomic predictions of removed data and the analysis of the full data set for each combination of testers Markers on the Left and Pedigree on the right. The most accurate combinations are those with green cells.

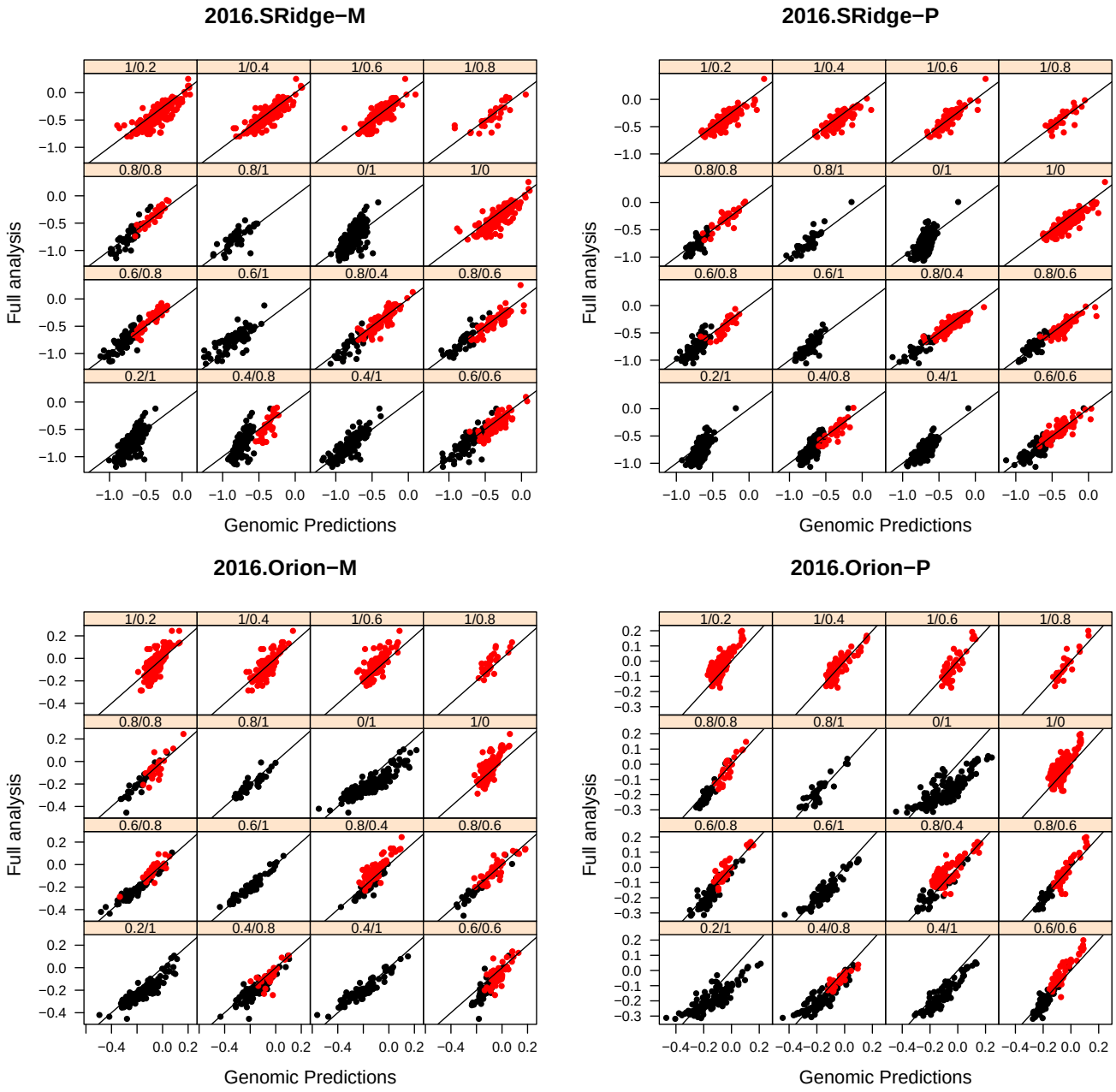

Figure 6.4: AYTF BLUPs from the analysis of all data versus BLUPs from the genomic predictions of the removed data for 2016 Orion and 2016 Spring Ridge. Black represents Tester 1 and red is Tester 2.

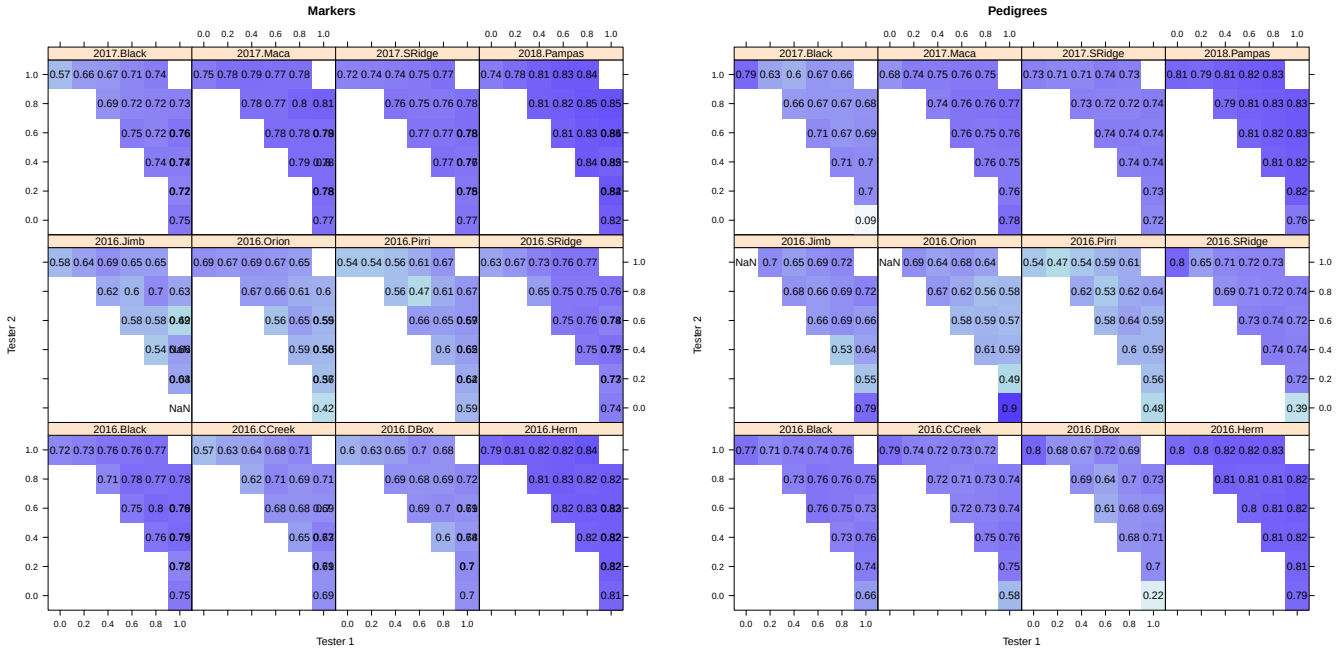

Figure 6.5: AYTF Predicted error variance for each tester combination, x-axis is the frequency of tester 1, y-axis is the frequency o tester 2; the numbers are the PEV values.

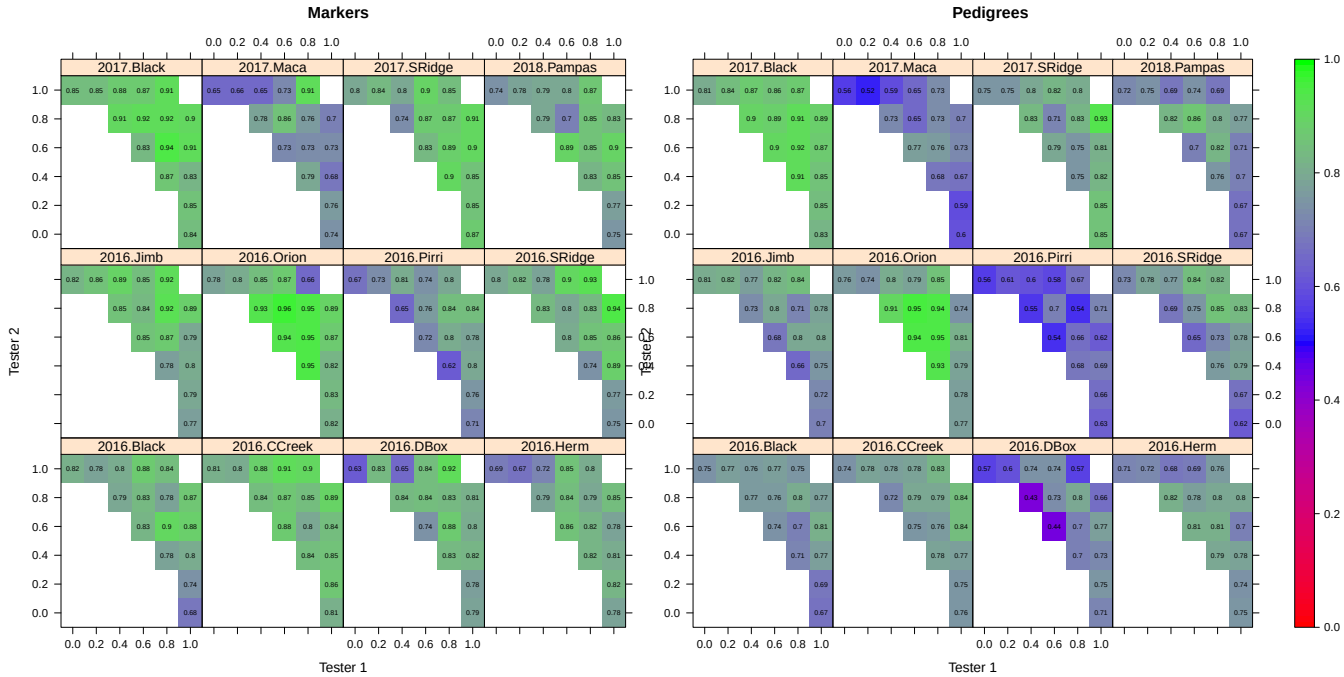

Figure 6.6: AYTM trials correlations between genomic predictions of removed data and the analysis of the full data set for each combination of testers Markers on the Left and Pedigree on the right. The most accurate combinations are those with green cells.

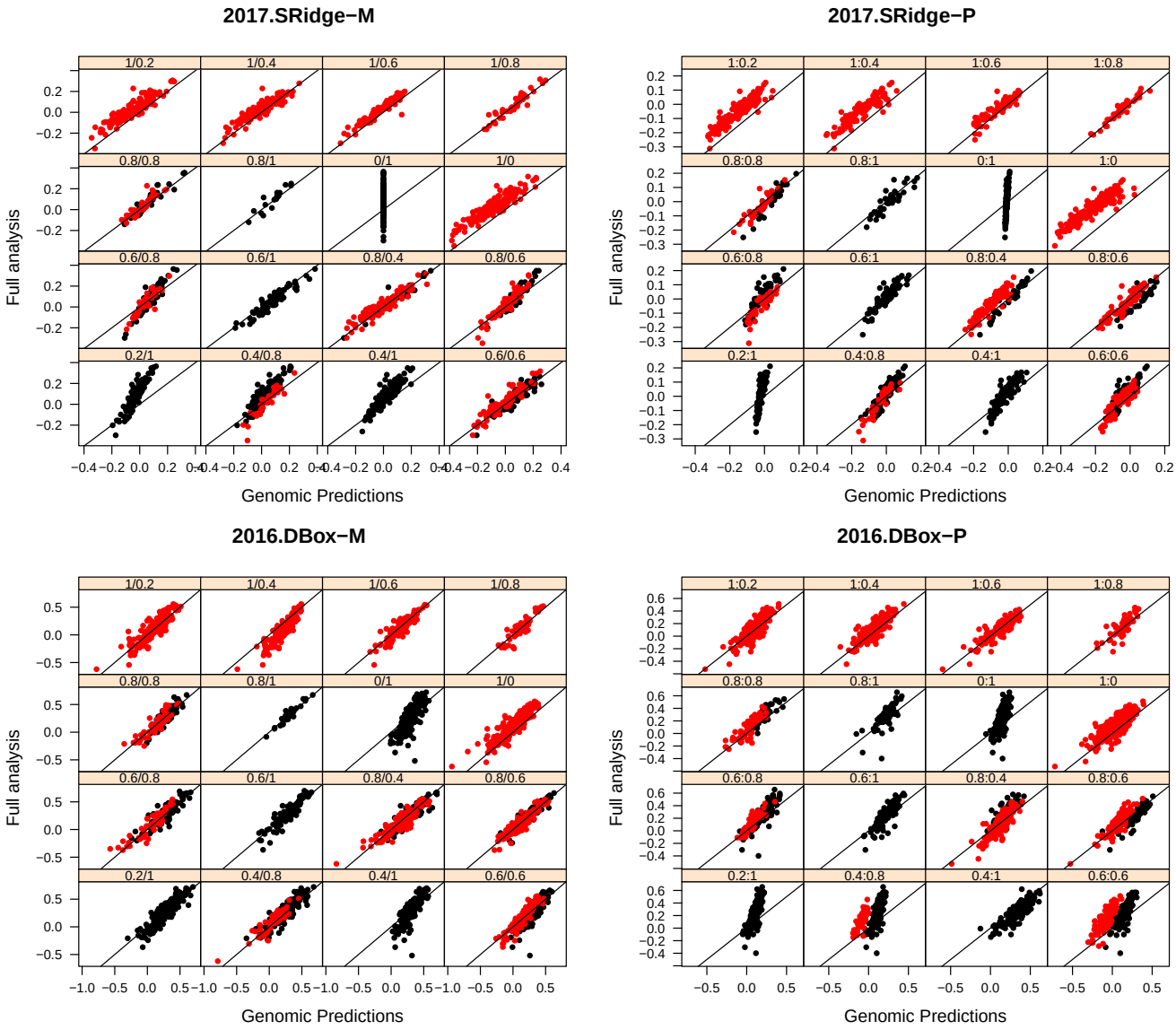

Figure 6.7: AYTM BLUPs from the analysis of all data versus BLUPs from the genomic predictions of the removed data for 2016 DBox and 2017 Spring Ridge. Black represents Tester 1 and red is Tester 2.

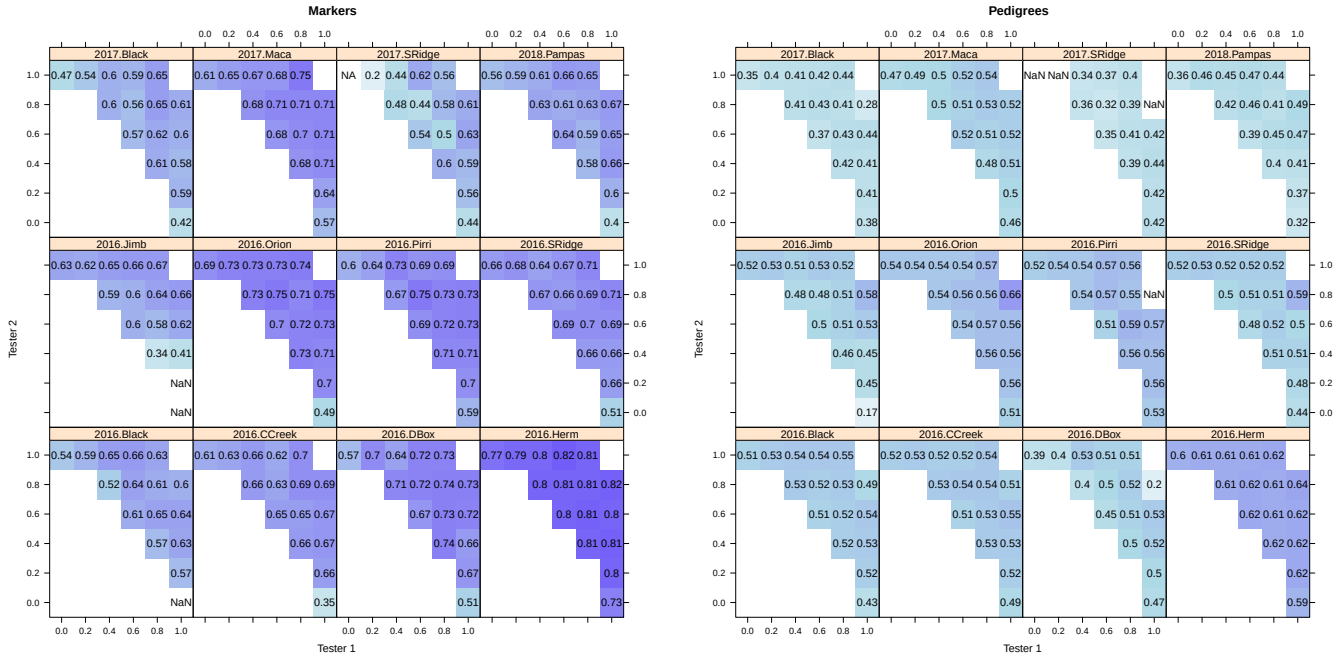

Figure 6.8: AYTM Predicted error variance for each tester combination, x-axis is the frequency of tester 1, y-axis is the frequency o tester 2; the numbers are the PEV values.

Chapter 7

Conclusion

The aim of this thesis is to improve the quality of plant breeding programs by increasing the value of the parameters that are directly proportional to the response to selection, i.e. genetic variance, selection intensity and accuracy of selection.

Firstly the genetic variance needs to be accurately calculated by analysing the phenotypic data using a linear mixed model that accounts for all of the variation, or error, in each trial. Chapter 3 demonstrated that a model can be improved by allowing for all of the extraneous and natural error in the field as well as allowing for inter-plot competition. The model was shown to be further enhanced by partitioning the genetic variance into additive and non-additive parts by using pedigree information. By using a better fitting model the genetic variance calculation was improved. This type was of linear mixed model was used in all subsequent chapters in the thesis. Chapter 4 highlighted the need for both ancestral pedigree information as well as molecular marker information to allow the relationships between the genotypes to be allowed for in the improvement of the genetic variance. Chapter 5 extended the model to further partition the genetic variance into additive, dominance and residual genetic parts.

The selection intensity can be increased by having the capacity to run trials that are unbalanced between trials and also within the hybrid selection within trials. Chapters 4 and 6 showed that genomic prediction of untested lines was accurately predicted by using linear mixed models including both pedigree and marker relationship matrices and data where genotypes have been removed then predicted and compared to the predictions that were made when the full data were analysed. By implementing genomic prediction into a breeding program there is a capacity to test a greater number of breeding lines than is possible in phenotyping alone (Chapter 4). Furthermore, with hybrid breeding there is a capacity to increase the number of hybrids by multitudes by creating hybrids from unbalanced parental crosses (Chapter 6).

Selection accuracy (heritability) was increased by fitting models that have the capacity to encapsulate all the relevant information from the data. It was shown in all research chapters that linear mixed models have the capacity to improve the accuracy of the predicted values in either a single trial or combinations of multi-environment trials.

In every breeding trial the genotypes, the environment and the management are unique. It follows that the statistical analysis of each trial is thus unique. Not all trials have extensive spatial effects, nor do they have significant competition. The genetic properties of each trial are also unique to the set of genotypes that are grown in the trial. The capacity to accurately predict genomic predictions is reliant on the set of genotypes and their properties. The environmental also has an enormous effect on the trial and must be taken into account when making any substantial conclusions about the predicted results. It is imperative to always take each situation into account and thus all of the results discussed in this thesis must be read in conjunction with the situations described here (the genotypes, their genetic make-up and the environments).

7.1 Implications and future work

7.1.1 Predicting additive and non-additive genetic effects from trials where traits are affected by inter-plot competition

Phenotypic yield from sorghum breeding trials is possibly subject to inter-plot competition, a phenomena where plot yields have a negative impact due to the influence of neighbouring plots. This is particularly important for trials that have 2 row plots, i.e. all rows in the trials have a neighbouring row of a different genotype. Statistical methods exist for removing this influence for the analysis of independent genotypes.

In the case of hybrids such as sorghum, there is a need to expand the existing methods to accommodate additive and non-additive genetic variance. This study introduced a method for removing the inter-plot competition from additive and non-additive partitions of genetic variances. The method allows for the computation of a pure stand yield for the additive genetic effect by fitting the correlation between each plot and it's respective neighbouring plots in the row direction. The results showed the competition model was superior to models that do not allow for competition.

Studies indicate that genotype competition occurs in around one third of all sorghum trials. The method presented here is only for a single trial analysis. Further work needs to be made to incorporate the capacity to fit competition in a multi-environment trial analysis.

7.1.2 Development of genomic prediction in sorghum

This study demonstrated that genomic prediction in sorghum trials using a single stage mixed model approach is feasible. Within this analytical framework we observed that the inclusion of pedigree information can improve prediction accuracy but is likely this improvement will decline as marker density increases. More critically we found that when small strongly interlinked families were used for GS, the impact of family size on prediction accuracy was reduced, however the similarity of a particular line to the average genotype in the training population had a large effect on prediction accuracy. From the perspective of practical deployment of genomic selection within current sorghum breeding programs in Australia, genotype by environment interactions will be the most important limiting factor. In the

short term we conclude that using a conservative approach where all of the lines within a selection population are genotyped and only a subset are phenotyped, is the most likely to be effective. In this circumstance genomic prediction improves genetic gain solely by increasing selection intensity rather that reducing generation time. More aggressive approaches involving multiple generations of selection without phenotyping require more research in order to deal with the complications posed by genotype by environment interaction.

7.1.3 Multi-Environment analysis of sorghum breeding trials using additive and dominance genomic relationships

Trials with high mean yield tend to have a higher broad-sense heritability, this might result in a better capacity to predict dominance variation. Another factor is due to trials with smaller total genetic variances having smaller or negligible residual genetic variance that cannot be partitioned into additive and dominance. It is advisable to use the results from these higher yielding trials for further investigation into hybrid dominance effects.

The results of this study must be considered in the light of the limited number of testers used. With this limitation in mind we have shown that including dominance in a linear mixed model can improve the predictability of hybrids across environments. The variation of the female testers also provides crucial information for testing males in different conditions. The additive proportion of the genetic variance is affected by the inclusion of dominance in the model with the dominance effects exhibiting a wider range of between trial correlations.

Cross prediction involving hybrid sampling is difficult when the hybrids are unbalanced across environment and male lines are not balanced within female testers. Some of these issues can be addressed by using the GxE analysis to group trials into environment categories and using these for sampling hybrids for use in cross prediction. This paper is a step towards cross prediction where predictions can be made using additive effects or dominance effects across correlated trials.

To implement genomic selection into a sorghum breeding program it is essential to discover factors that contribute to the genetic variance and include them in the statistical model. By fitting a model that partitions the genetic variance into its additive and dominance parts we can accurately calculate genomic performance and generate effects in different environments.

7.1.4 Identifying efficient strategies for preliminary evaluation in hybrid breeding programs

This work has highlighted the value of multiple tester parent lines in early generation hybrid breeding. Limitations of the current data availability has not allowed us to investigate prediction accuracy for hybrids that have more than 2 testers. Future work will involve trial analyses from multiple testers to address the question of the optimal number of testers and further explore the interactions of non-additive genetic variance and test environments.

Bibliography

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–722.

Albrecht, T., Auinger, H.-J., Wimmer, V., Ogutu, J. O., Knaak, C., Ouzunova, M., Piepho, H.-P., and Schön, C.-C. (2014). Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theoretical and Applied Genetics*, 127(6):1375–1386.

Albrecht, T., Wimmer, V., Auinger, H.-J., Erbe, M., Knaak, C., Ouzunova, M., Simianer, H., and Schön, C.-C. (2011). Genome-based prediction of testcross values in maize. *Theoretical and Applied Genetics*, 123(2):339–350.

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., and Hayes, B. J. (2016). Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. *Genetics Selection Evolution*, 48(1):8.

Allard, R. W. (1999). *Principles of plant breeding*. John Wiley & Sons.

Annicchiarico, P. (2002). *Genotype x environment interactions: challenges and opportunities for plant breeding and cultivar recommendations*. Number 174. Food & Agriculture Org.

Annor, B., Badu-Apraku, B., Nyadanu, D., Akromah, R., and Fakorede, M. A. (2020). Identifying heterotic groups and testers for hybrid development in early maturing yellow maize (zea mays) for sub-saharan africa. *Plant Breeding*.

Asins, M. J., Bernet, G. P., Villalta, I., and Carbonell, E. A. (2010). Qtl analysis in plant breeding. In *Molecular Techniques in Crop Improvement*, pages 3–21. Springer.

Baloch, F. S., Alsaleh, A., Shahid, M. Q., Çiftçi, V., de Miera, L. E. S., Aasim, M., Nadeem, M. A., Aktaş, H., Özkan, H., and Hatipoğlu, R. (2017). A whole genome dartseq and snp analysis for genetic diversity assessment in durum wheat from central fertile crescent. *Plos one*, 12(1).

Barkley, N. A., Roose, M. L., Krueger, R. R., and Federici, C. T. (2006). Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (ssrs). *Theoretical and Applied Genetics*, 112(8):1519–1531.

Bauer, A. M., Reetz, T. C., and Léon, J. (2006). Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. *Crop Science*, 46(6):2685–2691.

Bazakos, C., Hanemian, M., Trontin, C., Jiménez-Gómez, J. M., and Loudet, O. (2017). New strategies and tools in quantitative genetics: how to go from the phenotype to the genotype. *Annual Review of Plant Biology*, 68:435–455.

Beeck, C., Cowling, W., Smith, A., and Cullis, B. (2010). Analysis of yield and oil from a series of canola breeding trials. Part I: Fitting factor analytic models with pedigree information. *Genome*, 53:992–1001.

Bernardo, R. (1994). Prediction of maize single-cross performance using rflps and information from related hybrids. *Crop Science*, 34(1):20–25.

Bernardo, R. and Charcosset, A. (2006). Usefulness of gene information in marker-assisted recurrent selection: a simulation appraisal. *Crop Science*, 46(2):614–621.

Bernardo, R., Moreau, L., and Charcosset, A. (2006). Number and fitness of selected individuals in marker-assisted and phenotypic recurrent selection. *Crop Science*, 46(5):1972–1980.

Besag, J. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):691–746.

Besag, J. and Kempton, R. A. (1986). Statistical analysis of field experiments using neighbouring plots. *Biometrics*, 42:231–251.

Betran, F., Ribaut, J., Beck, D., and De Leon, D. G. (2003). Genetic diversity, specific combining ability, and heterosis in tropical maize under stress and nonstress environments. *Crop Science*, 43(3):797–806.

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., Sehabiague, P., Makumbi, D., Magorokosho, C., Oikeh, S., et al. (2015). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Science*, 55(1):154–163.

Boer, M. P., Wright, D., Feng, L., Podlich, D. W., Luo, L., Cooper, M., and van Eeuwijk, F. A. (2007). A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. *Genetics*, 177(3):1801–1813.

Borgognone, M. G., Butler, D. G., Ogbonnaya, F. C., and Dreccer, M. F. (2016). Molecular marker information in the analysis of multi-environment trials helps differentiate superior genotypes from promising parents. *Crop Science*, 56(5):2612–2628.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314.

Burgueno, J., Crossa, J., Cornelius, P. L., McLaren, G., Trethowan, R., and Krishnamachari, A. (2007). Modeling additve × environment and additive × additve × environment using genetic covariances of relatives of wheat genotypes. *Crop Science*, 47:311–320.

Burgueno, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. *Crop Science*, 52:707–719.

Burow, G., Chopra, R., Hughes, H., Xin, Z., and Burke, J. (2019). Marker assisted selection in sorghum using kasp assay for the detection of single nucleotide polymorphism/insertion deletion. In *Sorghum*, pages 75–84. Springer.

Butler, D. G., Cullis, B. R., Gilmour, A. R., and Gogel, B. J. (2009). ASReml-R reference manual release 3. Technical report, QLD Department of Primary Industries and Fisheries, Brisbane, QLD.

Chapman, S., Cooper, M., Butler, D., and Henzell, R. (2000a). Genotype by environment interactions affecting grain sorghum. i. characteristics that confound interpretation of hybrid yield. *Crop and Pasture Science*, 51(2):197–208.

Chapman, S., Cooper, M., Podlich, D., and Hammer, G. (2003). Evaluating plant breeding strategies by simulating gene action and dryland environment effects. *Agronomy Journal*, 95(1):99–113.

Chapman, S. C., Cooper, M., Hammer, G. L., and Butler, D. G. (2000b). Genotype by environment interactions affecting grain sorghum. II. frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. *Australian Journal of Agricultural Research*, 51:209–221.

Chenu, K., Chapman, S. C., Tardieu, F., McLean, G., Welcker, C., and Hammer, G. L. (2009). Simulating the yield impacts of organ-level quantitative trait loci associated with drought response in maize: a gene-to-phenotype modeling approach. *Genetics*, 183(4):1507–1523.

Chenu, K., Cooper, M., Hammer, G., Mathews, K. L., Dreccer, M., and Chapman, S. C. (2011). Environment characterization as an aid to wheat improvement: interpreting genotype–environment interactions by modelling water-deficit patterns in north-eastern australia. *Journal of Experimental Botany*, 62(6):1743–1755.

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., Hagen, T., Quinn, M., and Ng, E. H. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeders equation. *Theoretical and Applied Genetics*, 132(3):627–645.

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39(6):859.

Collard, B., Jahufer, M., Brouwer, J., and Pang, E. (2005). An introduction to markers, quantitative trait loci (qtl) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, 142:169–196. 10.1007/s10681-005-1681-5.

Comstock, R. E. (1977). Quantitative genetics and the design of breeding programs. In *Proceedings of the International Conference on Quantitative Genetics*, pages 705–718. Iowa State University Press: Ames, IA.

Cooper, M. and 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.

Crossa, J., de los Campos, G., Perez, P., Gianola, D., Burgueno, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., and Braun, H.-J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186:713–724.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, 22(11):961–975.

Cui, Y., Li, R., Li, G., Zhang, F., Zhu, T., Zhang, Q., Ali, J., Li, Z., and Xu, S. (2019). Hybrid breeding of rice via genomic selection. *Plant Biotechnology Journal*.

Cullis, B., Smith, A., Beeck, C., and Cowling, W. (2010). Analysis of yield and oil from a series of canola breeding trials. Part II: Exploring VxE using factor analysis. *Genome*, 53:1002–1016.

Cullis, B. R. and Gleeson, A. C. (1989). Efficiency of neighbour analysis for replicated field trials in Australia. *Journal of Agricultural Science, Cambridge*, 113:233–239.

Cullis, B. R. and Gleeson, A. C. (1991). Spatial analysis of field experiments - an extension to two dimensions. *Biometrics*, 47:1449–1460.

Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological and Environmental Statistics.*, 11:381–393.

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics*, 193(2):347–365.

de los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel hilbert spaces methods. *Genetics Research*, 92(4):295.

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

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375–385.

Desa, U. (2019). World population prospects 2019: Highlights. *New York (US): United Nations Department for Economic and Social Affairs*.

Dias, K. O. D. G., Gezan, S. A., Guimarães, C. T., Nazarian, A., e Silva, L. d. C., Parentoni, S. N., de Oliveira Guimaraes, P. E., de Oliveira Anoni, C., Pádua, J. M. V., de Oliveira Pinto, M., et al. (2018). Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. *Heredity*, 121(1):24.

Draper, N. R. and Guttman, I. (1980). Incorporating overlap effects from neighbouring units into response surface models. *Applied Statistics*, 29:128–134.

Dudley, J. (1993). Molecular markers in plant improvement: manipulation of genes affecting quantitative traits. *Crop Science*, 33(4):660–668.

Dudley, J., Maroof, M. S., and Rufener, G. (1991). Molecular markers and grouping of parents in maize breeding programs. *Crop Science*, 31(3):718–723.

Dudley, J. W. (1997). Quantitative genetics and plant breeding. *Advances in Agronomy*, 59:1–23.

Ejeta, G. and Knoll, J. E. (2007). Marker-assisted selection in sorghum. In *Genomics-Assisted Crop Improvement*, pages 187–205. Springer.

Endelman, J. B., Atlin, G. N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M. E., and Jannink, J.-L. (2014). Optimal design of preliminary yield trials with genome-wide markers. *Crop Science*, 54(1):48–59.

Falconer, D. S. and Mackay, T. (1996). *Introduction to Quantitative Genetics*. Longman Scientific and Technical, 4th edition.

FAO (2009). How to feed the world in 2050. In *Food and Agriculture Organization of the United Nations, 12-13 October, Rome, Italy*.

Fernando, R. (1998). Genetic evaluation and selection using genotypic, phenotypic and pedigree information. In *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, volume 26, pages 329–336, Animal Breeding and Genetics Unit, University of New England, Armidae NSW 2351, Australia .

Fernando, R. L., Habier, D., Stricker, C., Dekkers, J. C. M., and Totir, L. R. (2007). Genomic selection. *Acta Agriculturae Scandinavica, Section A - Animal Science*, 57(4):192–195.

Ferriol, M., Pico, B., and Nuez, F. (2003). Genetic diversity of a germplasm collection of cucurbita pepo using srap and aflp markers. *Theoretical and Applied Genetics*, 107(2):271–282.

Finlay, K. W. and Wilkinson, G. N. (1963). The analysis of adaptation in a plant breeding programme. *Australian Journal of Agricultural Research*, 14:742–754.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Frankham, R., Ballou, J. D., Eldridge, M. D., Lacy, R. C., Ralls, K., Dudash, M. R., and Fenster, C. B. (2011). Predicting the probability of outbreeding depression. *Conservation Biology*, 25(3):465–475.

Fritsche-Neto, R., Akdemir, D., and Jannink, J.-L. (2018). Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. *Theoretical and Applied Genetics*, 131(5):1153–1162.

Ganal, M. W., Plieske, J., Hohmeyer, A., Polley, A., and Röder, M. S. (2019). High-throughput genotyping for cereal research and breeding. In *Applications of Genetic and Genomic Research in Cereals*, pages 3–17. Elsevier.

Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., Mackay, I. J., and Hickey, J. M. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science*, 57(5):2372–2386.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, 183(1):347–363.

Gilbert, J., Lewis, R., Wilkinson, M., and Caligari, P. (1999). Developing an appropriate strategy to assess genetic variability in plant germplasm collections. *Theoretical and Applied genetics*, 98(6-7):1125–1131.

Gilmour, A. R., Cullis, B. R., and Verbyla, A. P. (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. R., Cullis, B. R., Welham, S. J., Gogel, B. J., and Thompson, R. (2009). ASREML, reference manual - release 3. Technical report, VSN International.

Goddard, M. and Hayes, B. (2007). Genomic selection. *Journal of Animal breeding and Genetics*, 124(6):323–330.

Goddard, M. and Hayes, B. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programs. *Nature Science Reviews*, 10:381–391.

GRDC (2017). Sorghum. *Grains Research and Development Corporation, Grownotes*.

GRDC (2020). Industry at a glance. *Grains Research and Development Corporation*.

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., McMullen, M. D., Holland, J. B., Szalma, S. J., Wisser, R. J., et al. (2019). Optimal designs for genomic selection in hybrid crops. *Molecular Plant*, 12(3):390–401.

Gupta, P. K. and Varshney, R. K. (2000). The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica*, 113(3):163–185.

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

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

Habier, D., Fernando, R. L., and Garrick, D. J. (2013). Genomic BLUP decoded: A look into the black box of genomic prediction. *Genetics*, 194(3):597+.

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., and Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in german holstein cattle. *Genetics Selection Evolution*, 42(1):5.

Hallauer, A. R., Carena, M. J., and Miranda Filho, J. d. (2010). *Quantitative Genetics in Maize Breeding*, volume 6. Springer Science & Business Media.

Hallauer, A. R., Russell, W. A., and Lamkey, K. R. (1988). *Corn Breeding*, volume 18. Agronomy Publications 259.

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

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

Henderson, C. (1976). A simple method for computing the inverse of a numerator relationship matrix used in the prediction of breeding values. *Biometrics*, 32:69–83.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–477.

Heslot, N., Jannink, J.-L., and Sorrells, M. E. (2013). Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. *Crop Science*, 53(3):921–933.

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science*, 52(1):146–160.

Hokanson, S., Szewc-McFadden, A., Lamboy, W., and McFerson, J. (1998). Microsatellite (ssr) markers reveal genetic identities, genetic diversity and relationships in a malus× domestica borkh. core subset collection. *Theoretical and Applied Genetics*, 97(5-6):671–683.

Holland, J. B., Nyquist, W. E., and Cervantes-Martínez, C. T. (2003). Estimating and interpreting heritability for plant breeding: An update. *Plant Breeding Reviews*, 22:9–112.

Huang, X., Börner, A., Röder, M., and Ganal, M. (2002). Assessing genetic diversity of wheat (triticum aestivum l.) germplasm using microsatellite markers. *Theoretical and Applied Genetics*, 105(5):699–707.

Hunt, C., Eeuwijk, F., Mace, E., Hayes, B., and Jordan, D. (2018). Development of genomic prediction in sorghum. *Crop Science*, 58(2).

Hunt, C. and Jordan, D. (2009). Competition effects in sorghum breeding trials. In *14th Australasian Plant Breeding Conference, SABRAO Journal of Breeding and Genetics*, Cairns, Australia.

Hunt, C. H., Butler, D. G., and Cullis, B. R. (2011). Analysis of sorghum breeding trials using pedigree information. In *1st Australasian Applied Statistics Conference (Genstat and ASRemL), Palm Cove, Australia*.

Hunt, C. H., Hayes, B. J., van Eeuwijk, F. A., Mace, E. S., and Jordan, D. R. (2020). Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships. *Theoretical and Applied Genetics*, 133(3):1009–1018.

Hunt, C. H., Smith, A. B., Jordan, D. R., and Cullis, B. R. (2013). Predicting additive and non-additive genetic effects from trials where traits are affected by interplot competition. *Journal of Agricultural, Biological and Environmental Statistics*, 18(1):53–63.

Im, S., Fernando, R. L., and Gianola, D. (1989). Likelihood inferences in animal breeding under selection: missing-data theory view point. *Genetic Selection and Evolution*, 21:399–414.

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

Jannink, J.-L. and Walsh, B. (2002). Association mapping in plant populations. *Quantitative Genetics, Genomics and Plant Breeding*, pages 59–68.

Jonas, E. and de Koning, D. J. (2016). Goals and hurdles for a successful implementation of genomic selection in breeding programme for selected annual and perennial crops. *Biotechnology and Genetic Engineering Reviews*, pages 1–25.

Jones, D. F. (1917). Dominance of linked factors as a means of accounting for heterosis. *Genetics*, 2(5):466.

Jordan, D., Hunt, C., Cruickshank, A., Borrell, A., and Henzell, R. (2012). The relationship between the stay-green trait and grain yield in elite sorghum hybrids grown in a range of environments. *Crop Science*, 52(3):1153–1161.

Jordan, D., Mace, E., Borrell, A., Cruickshank, A., Chapman, S., van Oosterom, E., and Hammer, G. (2013). An integrated approach to sorghum crop improvement in australia. In *4th International Conference on Integrated Approaches to Improve Crop Production under Drought-Prone Environments*, Perth, WA, Australia.

Jordan, D. R., Mace, E. S., Cruickshank, A. W., Hunt, C. H., and Henzell, R. G. (2011). Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. *Crop Science*, 51:1444–1457.

Kadam, D. C., Potts, S. M., Bohn, M. O., Lipka, A. E., and Lorenz, A. J. (2016). Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. *G3: Genes, Genomes, Genetics*, 6(11):3443–3453.

Kearsey, M. and Farquhar, A. (1998). Qtl analysis in plants; where are we now? *Heredity*, 80(2):137–142.

Kearsey, M. J., Pooni, H. S., et al. (1998). *The Genetical Analysis of Quantitative Traits.* Stanley Thornes (Publishers) Ltd.

Keddy, P. A. (2001). Studying competition. In *Competition. Population and Community Biology Series*, volume 26, pages 1–59. Springer, Dordrecht.

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London. Series B-Biological Sciences*, 143(910):103–113.

Kilian, A., Huttner, E., Wenzl, P., Jaccoud, D., Carling, J., Caig, V., Evers, M., Heller-Uszynska, K., Cayla, C., Patarapuwadol, S., et al. (2003). The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement. In *Proceedings of the international congress in the wake of the double helix: from the green revolution to the gene revolution*, pages 443–461.

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.-S., and Paterson, A. H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science*, 242:14–22.

Korte, A. and Farlow, A. (2013). The advantages and limitations of trait analysis with gwas: a review. *Plant Methods*, 9(1):1–9.

Kraakman, A. T., Niks, R. E., Van den Berg, P. M., Stam, P., and Van Eeuwijk, F. A. (2004). Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics*, 168(1):435–446.

Lande, R. and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743–756.

Larièpe, A., Moreau, L., Laborde, J., Bauland, C., Mezmouk, S., Décousset, L., Mary-Huard, T., Fiévet, J. B., Gallais, A., Dubreuil, P., and Charcosset, A. (2017). General and specific combining abilities in a maize (zea mays l.) test-cross hybrid panel: relative importance of population structure and genetic divergence between parents. *Theoretical and Applied Genetics*, 130(2):403–417.

Lassois, L., Denancé, C., Ravon, E., Guyader, A., Guisnel, R., Hibrand-Saint-Oyant, L., Poncet, C., Lasserre-Zuber, P., Feugey, L., and Durel, C.-E. (2016). Genetic diversity, population structure, parentage analysis, and construction of core collections in the french apple germplasm based on ssr markers. *Plant Molecular Biology Reporter*, 34(4):827–844.

Lee, E. and Tollenaar, M. (2007). Physiological basis of successful breeding strategies for maize grain yield. *Crop Science*, 47:S–202.

Lee, M. (1995). DNA markers and plant breeding programs. In *Advances in Agronomy*, volume 55, pages 265–344. Elsevier.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.-L., Singh, R. P., Autrique, E., and DE Los Campos, G. (2015). Increased prediction accuracy in wheat breeding trials using a marker $\times$ environment interaction genomic selection model. *G3: Genes, Genomes, Genetics*, pages g3–114.

Lorenz, A. J. (2013). Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. *G3: Genes, Genomes, Genetics*, 3(3):481–491.

Ly, D., Hamblin, M., Rabbi, I., Melaku, G., Bakare, M., Gauch, H. G., Okechukwu, R., Dixon, A. G., Kulakow, P., and Jannink, J.-L. (2013). Relatedness and genotype$\times$ environment interaction affect prediction accuracies in genomic selection: A study in cassava. *Crop Science*, 53(4):1312–1325.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinaeur Associates, Sunderland, Massachusetts.

Mace, E., Innes, D., Hunt, C., Wang, X., Tao, Y., Baxter, J., Hassall, M., Hathorn, A., and Jordan, D. (2019). The sorghum QTL atlas: a powerful tool for trait dissection, comparative genomics and crop improvement. *Theoretical and Applied Genetics*, 132(3):751–766.

Mace, E. S., Rami, J.-F., Bouchet, S., Klein, P. E., Klein, R. R., Kilian, A., Wenzl, P., Xia, L., Halloran, K., and Jordan, D. R. (2009). A consensus genetic map of sorghum that integrates multiple

component maps and high-throughput diversity array technology (DArT) markers. *BMC Plant Biology*, 9(1):13.

Mace, E. S., Xia, L., Jordan, D. R., Halloran, K., Parh, D. K., Huttner, E., Wenzl, P., and Kilian, A. (2008). DArT markers: diversity analyses and mapping in sorghum bicolor. *BMC Genomics*, pages 9–26.

Mäki-Tanila, A. (2007). An overview on quantitative and genomic tools for utilising dominance genetic variation in improving animal production. *Agricultural and Food Sience*, 16:188–198.

Malosetti, M., Ribaut, J.-M., and van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*, 4:44.

Malosetti, M., Voltas, J., Romagosa, I., Ullrich, S., and Van Eeuwijk, F. (2004). Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica*, 137(1):139–145.

Massman, J. M., Gordillo, A., Lorenzana, R. E., and Bernardo, R. (2013). Genomewide predictions from maize single-cross data. *Theoretical and Applied Genetics*, 126(1):13–22.

Mathews, K. L., Malosetti, M., Chapman, S., McIntyre, L., Reynolds, M., Shorter, R., and van Eeuwijk, F. (2008). Multi-environment QTL mixed models for drought stress adaptation in wheat. *Theoretical and Applied Genetics*, 117(7):1077–1091.

Meena, A. K., Gurjar, D., Patil, S., and Kumhar, B. L. (2017). Concept of heterotic group and its exploitation in hybrid breeding. *International Journal of Current Microbiology and Applied Sciences*, 6:61–73.

Melchinger, A. (1999). Genetic diversity and heterosis. *In JG Coors and S. Pandey (ed.) The Genetics and Exploitation of Heterosis in Crops. ASA, CSSA, and SSSA, Madison, WI.*, pages 99–118.

Melchinger, A. E. and Gumber, R. K. (1998). Overview of heterosis and heterotic groups in agronomic crops. *Concepts and Breeding of Heterosis in Crop Plants*, 25:29–44.

Melchinger, A. E., Piepho, H.-P., Utz, H. F., Muminović, J., Wegenast, T., Törjék, O., Altmann, T., and Kusterer, B. (2007). Genetic basis of heterosis for growth-related traits in arabidopsis investigated by testcross progenies of near-isogenic lines reveals a significant role of epistasis. *Genetics*, 177(3):1827–1837.

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

Meuwissen, T. H. E. and Luo, Z. (1992). Computing inbreeding coefficients in large populations. *Genetics, Selection and Evolution*, 24:269–293.

Mohan, M., Nair, S., Bhagwat, A., Krishna, T., Yano, M., Bhatia, C., and Sasaki, T. (1997). Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding*, 3(2):87–103.

Moose, S. P. and Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology*, 147(3):969–977.

Mühlenbein, H. (1997). The equation for response to selection and its use for prediction. *Evolutionary Computation*, 5(3):303–346.

Muñoz, P. R., Resende, M. F., Gezan, S. A., Resende, M. D. V., de los Campos, G., Kirst, M., Huber, D., and Peter, G. F. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics*, 198(4):1759–1768.

Nadeem, M. A., Habyarimana, E., Çiftçi, V., Nawaz, M. A., Karaköy, T., Comertpay, G., Shahid, M. Q., Hatipoğlu, R., Yeken, M. Z., Ali, F., et al. (2018a). Characterization of genetic diversity in turkish common bean gene pool using phenotypic and whole-genome dartseq-generated silicodart marker information. *PloS one*, 13(10).

Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., Yıldız, M., Hatipoğlu, R., Ahmad, F., Alsaleh, A., Labhane, N., et al. (2018b). Dna molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnology & Biotechnological Equipment*, 32(2):261–285.

Naeem, M., Ghouri, F., Shahid, M., Iqbal, M., Baloch, F., Chen, L., Allah, S., Babar, M., and Rana, M. (2015). Genetic diversity in mutated and non-mutated rice varieties. *Genetics and Molecular Research*, 14(4):17109–17123.

Nakaya, A. and Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding. *Annals of Botany*, 110:1303–1316.

Nduwumuremyi, A., Tongoona, P., and Habimana, S. (2013). Mating designs: helpful tool for quantitative plant breeding analysis. *Journal of Plant Breeding and Genetics*, 1(3):117–129.

Nyquist, W. E. and Baker, R. J. (1991). Estimation of heritability and prediction of selection response in plant populations. *Critical Reviews in Plant Sciences*, 10(3):235–322.

Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C., and Waugh, R. (2016). Genomic selection in multi-environment crop trials. *G3: Genes, Genomes, Genetics*, 6(5):1313–1326.

Oakey, H., Verbyla, A., Cullis, B., Wei, X., and Pitchford, W. (2007). Joint modelling of additive and non-additive (genetic line) effects in multi-environment trials. *Theoretical and Applied Genetics*, 114:1319–1332.

Oakey, H., Verbyla, A., Pitchford, W., Cullis, B., and Kuchel, H. (2006). Joint modelling of additive and non-additive genetic line effects in single field trials. *Theoretical and Applied Genetics.*, 113:809–819.

Papadakis, J. S. (1937). Methode statistique pour des experiences sur champ. Bulletin scientifique, Institut d'Amelioration des Plantes a Thessaloniki (Grece).

Parh, D. K., Jordan, D. R., Aitken, E. A. B., Mace, E. S., Jun-ai, P., McIntyre, C. L., and Godwin, I. D. (2008). QTL analysis of ergot resistance in sorghum. *Theoretical and Applied Genetics.*, 117:369–382.

Paterson, A. H., Tanksley, S. D., and Sorrells, M. E. (1991). DNA markers in plant improvement. In *Advances in Agronomy*, volume 46, pages 39–90. Elsevier.

Patterson, H. D., Silvey, V., Talbot, M., and Weatherup, S. T. C. (1977). Variability of yields of cereal varieties in U.K. trials. *Journal of Agricultural Science, Cambridge*, 89:238–245.

Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 31:545–554.

Piepho, H. (2005). Statistical tests for QTL and QTL-by-environment effects in segregating populations derived from line crosses. *Theoretical and Applied Genetics*, 110(3):561–566.

Piepho, H., Ogutu, J., Schulz-Streeck, T., Estaghvirou, B., Gordillo, A., and Technow, F. (2012). Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. *Crop Science*, 52(3):1093–1104.

Piepho, H.-P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics*, 97:195–201.

Piepho, H.-P. (2000). A mixed-model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. *Genetics*, 156(4):2043–2050.

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161:209–228.

Piepho, H.-P. and Pillen, K. (2004). Mixed modelling for QTL$\times$ environment interaction analysis. *Euphytica*, 137(1):147–153.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., and He, Z. (2017). Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Molecular Plant*, 10(8):1047–1064.

Reynolds, D., Frederic, B., Welcker, C., Bostrom, A., Ball, J., Cellini, F., Lorence, A., Chawade, A., Khafif, M., Noshita, K., Mller-Linow, M., Zhou, J., and Tardieu, F. (2018). What is cost-efficient phenotyping? optimizing costs for different scenarios. *Plant Science*, 282.

Ribaut, J.-M. and Ragot, M. (2007). Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. *Journal of Experimental Botany*, 58(2):351–360.

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

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetics Research*, 67(2):175–185.

Rizal, G., Karki, S., Alcasid, M., Montecillo, F., Acebron, K., Larazo, N., Garcia, R., Slamet-Loedin, I. H., and Quick, W. P. (2014). Shortening the breeding cycle of sorghum, a model crop for research. *Crop Science*, 54(2):520–529.

Roorkiwal, M., Jarquin, D., Singh, M. K., Gaur, P. M., Bharadwaj, C., Rathore, A., Howard, R., Srinivasan, S., Jain, A., Garg, V., et al. (2018). Genomic-enabled prediction models using multi-environment trials to estimate the effect of genotype× environment interaction on prediction accuracy in chickpea. *Scientific Reports*, 8(1):1–11.

Rudolf-Pilih, K., Petkovšek, M., Jakse, J., Štajner, N., Murovec, J., and Bohanec, B. (2019). New hybrid breeding method based on genotyping, inter-pollination, phenotyping and paternity testing of selected elite f1 hybrids. *Frontiers in Plant Science*, 10:1111.

Schenkel, F. S., Schaeffer, L. R., and Boettcher, P. J. (2002). Comparison between estimation of breeding values and fixed effects using bayesian and empirical blup estimation under selection on parents and missing pedigree information. *Genetics Selection Evolution*, 34(1):41–60.

Schrag, T., Melchinger, A., Sørensen, A., and Frisch, M. (2006). Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. *Theoretical and Applied Genetics*, 113(6):1037–1047.

Scutari, M., Mackay, I., and Balding, D. (2013). Improving the efficiency of genomic selection. *Statistical Applications in Genetics and Molecular Biology*, 12(4):517–527.

Shull, G. H. (1908). The composition of a field of maize. *Journal of Heredity*, (1):296–301.

Smith, A., Cullis, B., Luckett, D., Hollamby, G., and Thompson, R. (2002a). *Exploring variety-environment data using random effects AMMI models with adjustments for spatial field trend. Part II: Applications. In Quantitative Genetics, Genomics and Plant Breeding. M. Kang (Ed.)*, pages 337–352. CABI Publishing, U.K.

Smith, A., Cullis, B., and Thompson, R. (2002b). *Exploring variety-environment data using random effects AMMI models with adjustments for spatial field trend. Part I: Theory. In Quantitative Genetics, Genomics and Plant Breeding. M. Kang (Ed.)*, pages 323–336. CABI Publishing, U.K.

Smith, A. B. and Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, 214(8):143.

Smith, A. B., Cullis, B. R., and Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, 57:1138–1147.

Smith, A. B., Cullis, B. R., and Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultural Science, Cambridge*, 143:449–462.

Smith, J., Duvick, D., Smith, O., Grunst, A., and Wall, S. (1999). Effect of hybrid breeding on genetic diversity in maize. *Genetics and Exploitation of Heterosis in Crops*, pages 119–126.

Smith, O., Smith, J., Bowen, S., Tenborg, R., and Wall, S. (1990). Similarities among a group of elite maize inbreds as measured by pedigree, F1 grain yield, grain yield, heterosis, and RFLPs. *Theoretical and Applied Genetics*, 80(6):833–840.

Sorensen, D. A. and Kennedy, B. W. (1984). Estimation of genetic variances from unselected and selected populations. *Journal of Animal Science*, 59(5):1213–1223.

Sprague, G. F. and Tatum, L. A. (1942). General vs. specific combining ability in single crosses of corn 1. *Agronomy Journal*, 34(10):923–932.

Staub, J. E., Serquen, F. C., and Gupta, M. (1996). Genetic markers, map construction, and their application in plant breeding. *HortScience*, 31(5):729–741.

Stefanova, K. T., Smith, A. B., and Cullis, B. R. (2009). Enhanced diagnostics for the spatial analysis of field trials. *Journal of Agricultural, Biological and Environmental Statistics*, 14:392–410.

Strandén, I. and Garrick, D. (2009). Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of Dairy Science*, 92(6):2971–2975.

Stringer, J. (2006). *Joint modelling of spatial variability and interplot competition to improve the efficiency of plant improvement*. PhD thesis, The University of Qld, Brisbane.

Stringer, J., Cullis, B., and Thompson, R. (2011). Joint modeling of spatial variability and within-row interplot competition to increase the efficiency of plant improvement. *Journal of Agricultural, Biological, and Environmental Statistics*, 16:269–281.

Stringer, J. K. and Cullis, B. R. (2002). Application of spatial analysis techniques to adjust for fertility trends and identify interplot competition in early stage sugarcane selection trials. *Australian Journal of Agricultural Research*, 53:911–918.

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., and Lund, M. S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PloS one*, 7(9).

Tanksley, S. D. (1983). Molecular markers in plant breeding. *Plant Molecular Biology Reporter*, 1(1):3–8.

Taylor, J. D., Verbyla, A. P., Cavanagh, C., and Newberry, M. (2012). Variable selection in linear mixed models using an extended class of penalties. *Australian & New Zealand Journal of Statistics*, 54(4):427–449.

Thompson, R., Cullis, B. R., Smith, A. B., and Gilmour, A. R. (2003). A sparse implementation of the Average Information algorithm for factor analytic and reduced rank variance models. *Australian and New Zealand Journal of Statistics*, 45:445–460.

Tolhurst, D. J., Mathews, K. L., Smith, A. B., and Cullis, B. R. (2019). Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model. *Journal of Animal Breeding and Genetics*, 136(4):279–300.

Van der Werf, J. H. and de Boer, I. J. (1990). Estimation of additive genetic variance when base populations are selected. *Journal of Animal Science*, 68(10):3124–3132.

van Eeuwijk, F. A., Malosetti, M., Yin, X., Struik, P. C., and Stam, P. (2005). Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. *Australian Journal of Agricultural Research*, 56:883–894.

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423.

Velazco, J. G., Jordan, D. R., Mace, E. S., Hunt, C. H., Malosetti, M., and Van Eeuwijk, F. A. (2019). Genomic prediction of grain yield and drought-adaptation capacity in sorghum is enhanced by multi-trait analysis. *Frontiers in Plant Science*, 10.

Velazco, J. G., Rodríguez-Álvarez, M. X., Boer, M. P., Jordan, D. R., Eilers, P. H., Malosetti, M., and van Eeuwijk, F. A. (2017). Modelling spatial trends in sorghum breeding field trials using a two-dimensional p-spline mixed model. *Theoretical and Applied Genetics*, 130(7):1375–1392.

Verbyla, A. P. and Cullis, B. R. (2012). Multivariate whole genome average interval mapping: QTL analysis for multiple traits and/or environments. *Theoretical and Applied Genetics*, 125(5):933–953.

Verbyla, A. P., Eckermann, P. J., Thompson, R., and Cullis, B. R. (2003). The analysis of quantitative trait loci in multi-environment trials using a multiplicative mixed model. *Australian Journal of Agricultural Research*, 54:1395–1408.

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, 195(4):1223–1230.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van De Lee, T., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., et al. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, 23(21):4407–4414.

Voss-Fels, K. P., Stahl, A., and Hickey, L. T. (2019). Q&A: modern crop breeding for future food security. *BMC biology*, 17(1):18.

Wang, X., Luo, G., Yang, W., Li, Y., Sun, J., Zhan, K., Liu, D., and Zhang, A. (2017). Genetic diversity, population structure and marker-trait associations for agronomic and grain traits in wild diploid wheat triticum urartu. *BMC Plant Biology*, 17(1):112.

Welham, S. J., Cullis, B. R., Gogel, B. J., Gilmour, A. R., and Thompson, R. (2004). Prediction in mixed linear models. *Australian and New Zealand Journal of Statistics*, 46:325–347.

Wellmann, R., Preuß, S., Tholen, E., Heinkel, J., Wimmers, K., and Bennewitz, J. (2013). Genomic selection using low density marker panels with application to a sire line in pigs. *Genetics Selection Evolution*, 45(1):28.

Whittaker, J. C., Thompson, R., and Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetics Research*, 75(2):249–252.

Wilkinson, G. N., Eckert, S. R., Hancock, T. W., and Mayo, O. (1983). Nearest neighbour (NN) analysis of field experiments (with discussion). *Journal of the Royal Statistical Society, Series B*, 45:151–211.

Wolc, A., Arango, J., Settar, P., Fulton, J. E., OSullivan, N. P., Preisinger, R., Habier, D., Fernando, R., Garrick, D. J., and Dekkers, J. C. (2011). Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genetics Selection Evolution*, 43(23):10–1186.

Wricke, G. and Weber, E. (1986). *Quantitative genetics and selection in plant breeding*. Walter de Gruyter.

Wright, S. (1934). Physiological and evolutionary theories of dominance. *The American Naturalist*, 68(714):24–53.

Würschum, T., Liu, G., Boeven, P. H., Longin, C. F. H., Mirdita, V., Kazman, E., Zhao, Y., and Reif, J. C. (2018). Exploiting the Rht portfolio for hybrid wheat breeding. *Theoretical and Applied Genetics*, 131(7):1433–1442.

Würschum, T., Maurer, H. P., Weissmann, S., Hahn, V., and Leiser, W. L. (2017). Accuracy of within-and among-family genomic prediction in triticale. *Plant Breeding*, 136(2):230–236.

Xie, F., He, Z., Esguerra, M. Q., Qiu, F., and Ramanathan, V. (2014). Determination of heterotic groups for tropical Indica hybrid rice germplasm. *Theoretical and Applied Genetics*, 127(2):407–417.

Xu, Y. and Crouch, J. H. (2008). Marker-assisted selection in plant breeding: From publications to practice. *Crop Science*, 48(2):391–407.

Yu, J., Holland, J. B., McMullen, M. D., and Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, 178(1):539–551.

Zamir, D. (2001). Improving plant breeding with exotic genetic libraries. *Nature Reviews Genetics*, 2(12):983–989.

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., Cui, Z., Ruan, Y., Burgueño, J., San Vicente, F., et al. (2017). Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. *Frontiers in Plant Science*, 8:1916.

Zhang, P., Zhong, K., Shahid, M. Q., and Tong, H. (2016a). Association analysis in rice: from application to utilization. *Frontiers in Plant Science*, 7:1202.

Zhang, P., Zhong, K., Tong, H., Shahid, M. Q., and Li, J. (2016b). Association mapping for aluminum tolerance in a core collection of rice landraces. *Frontiers in Plant Science*, 7:1415.

Zhang, X., Pérez-Rodríguez, P., Semagn, K., Beyene, Y., Babu, R., López-Cruz, M., San Vicente, F., Olsen, M., Buckler, E., Jannink, J.-L., Prassanna, B., and Crossa, J. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and gbs snps. *Heredity*, 114(3):291–299.

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Science*, 53:802–810.

Appendix A

Electronic Supplementary Material Development of genomic prediction in sorghum

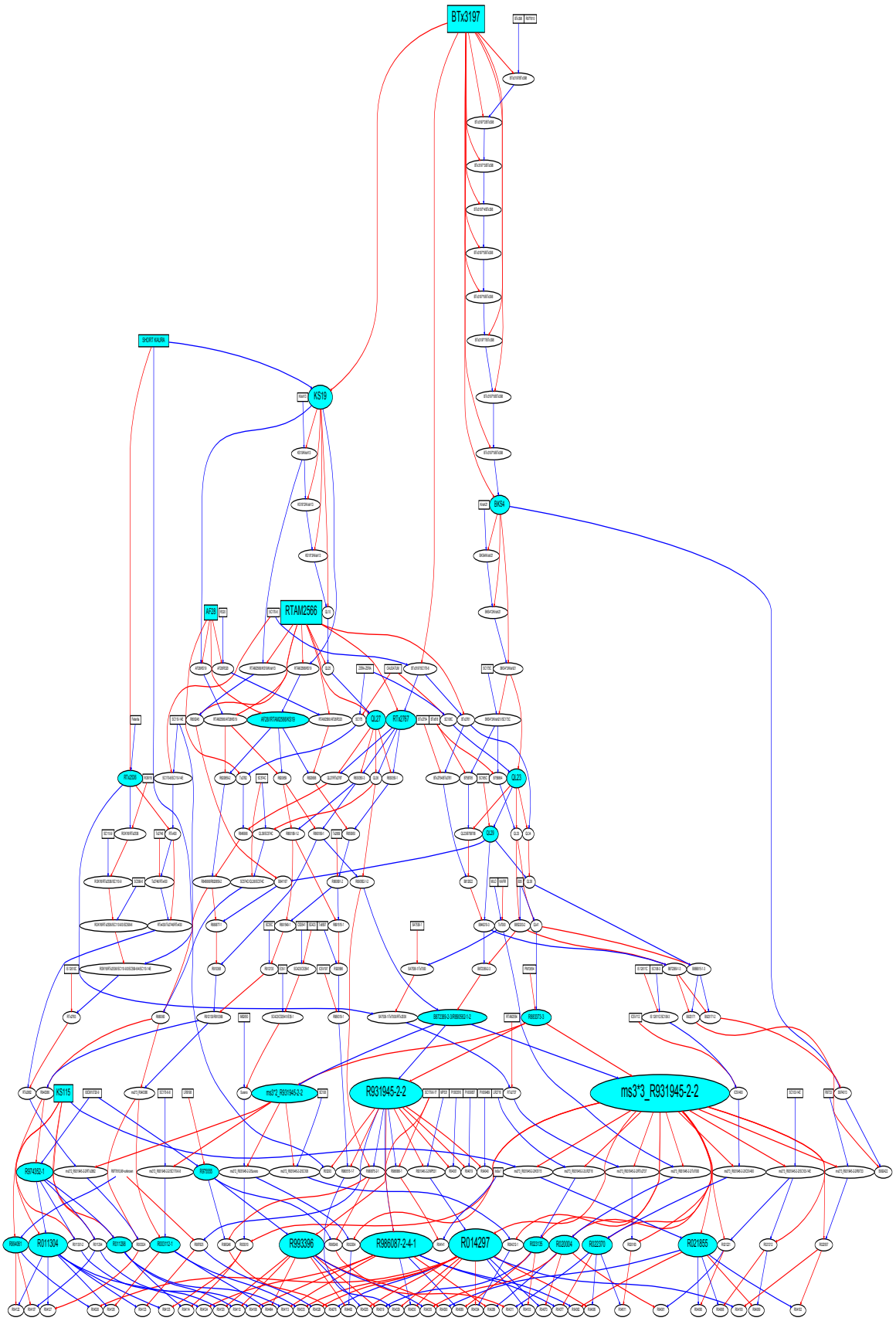

Figure A.1: Family tree for 25 generations of parents for the male parental lines used in the 2008 PYTM trials. Ancestral parents are presented at the top of the figure with their offspring descending below. Blue lines indicate the parent was used as a male and red lines indicate use as a female parent. Male parent lines included in this study are shown at the very bottom of the figure

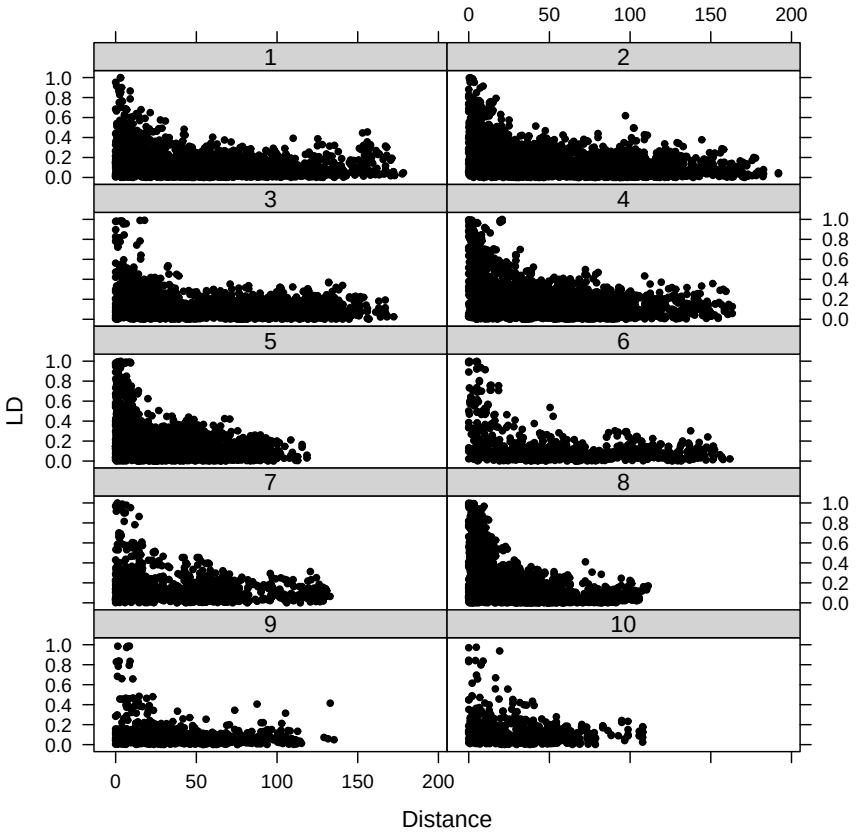

Figure A.2: Linkage Disequilibrium (LD) calculated as a Pearson coefficient of correlation versus distance between pairs of markers (in cM).

Table A.1: Properties of the pedigrees used in the study including the number of progeny, female and male parents, the total number of progenies derived from both the female and male parents and number of times each parent is used in a cross.

Family ID	Family[a] Name	Progeny[b]	Female Parent	Female Crosses[c]	Female Progeny[d]	Male Parent	Male Crosses[e]	Male Progeny[f]	Total Progeny[g]
1	R04127	6	R003324	1	6	R011304	3	30	36
2	R04126	5	R003112-1	2	12	R011304	3	30	42
3	R04102	4	R022557	2	11	R021855	5	36	47
4	R04511	4	R020163	1	4	R022370	3	45	49
5	R04114	9	R003010	1	9	R011298	3	45	54
6	R04088	5	R021212	2	23	R021855	5	36	59
7	R04089	18	R021855	5	36	R021212	2	23	59
8	R04090	5	R021221	2	25	R021855	5	36	61
9	R04492	6	R993396	6	73	R002934	1	6	79
10	R04081	20	R020004	4	57	R021221	2	25	82
11	R04494	7	R993396	6	73	R003112-1	2	12	85
12	R04001	27	R931945-2-2	2	61	PI 563516	1	27	88
13	R04040	34	R931945-2-2	2	61	PI 609489	1	34	95
14	R04120	25	R993396	6	73	R011301-2	1	25	98
15	R04334	4	R014297	8	97	R004212-1	1	4	101
16	R04329	5	R014297	8	97	R980515-1-7	1	5	102
17	R04100	7	R022557	2	11	R014297	8	97	108
18	R04112	13	R993396	6	73	R011298	3	45	118
19	R04335	9	R020004	4	57	R993396	6	73	130
20	R04103	13	R023135	2	35	R014297	8	97	132
21	R04095	19	R014297	8	97	R022370	3	45	142
22	R04108	11	R986087-2-4-1	8	131	R011294	1	11	142
23	R04073	17	R020004	4	57	R014297	8	97	154
24	R04124	19	R986087-2-4-1	8	131	R011304	3	30	161
25	R04389	22	R023135	2	35	R986087-2-4-1	8	131	166
26	R04362	4	R021855	5	36	R986087-2-4-1	8	131	167
27	R04325	13	R014297	8	97	R993396	6	73	170
28	R04113	23	R986087-2-4-1	8	131	R011298	3	45	176
29	R04377	22	R022370	3	45	R986087-2-4-1	8	131	176
30	R04289	11	R986087-2-4-1	8	131	R020004	4	57	188
31	R04330	19	R014297	8	97	R986087-2-4-1	8	131	228

[a]For full pedigree tree see ESM figure S1
[b]Number of Individual lines from each family
[c]Number of families that have the female parent as a parent of their family
[d]Number of lines within the dataset with the female parent as a parent
[e]Number of families that have the male parent as a parent of their family
[f]Number of lines within the dataset with the male parent as a parent
[g]Total number of full and half siblings from the family

Table A.2: A summary of the number of polymorphic DArT markers per linkage group (LG), the distance in cM where the LD has decayed by half.

LG	# DArT markers	LD (cM)[a]	LD decay (cM)
SBI-01	68	0.41	10.8
SBI-02	77	0.37	16.8
SBI-03	59	0.35	10.2
SBI-04	72	0.45	23.8
SBI-05	76	0.31	12.3
SBI-06	40	0.37	13.7
SBI-07	47	0.49	12.6
SBI-08	74	0.43	9.7
SBI-09	34	0.46	6.7
SBI-10	34	0.39	5.7

[a]See Supp Figure S2 for LD versus distance plot

Table A.3: Significant fixed terms and spatial error terms included in all fitted models. Line.out refers to the Lines that have been phenotyped but not genotyped, stand is a covariate to adjust for unequal numbers of plants within each trial plot due to establishment, lincol is a linear trend for column used at Biloela only. The random effects for all sites consist of Replicate, Row and AR1 spatial terms for each direction C indicates Column and R indicates Row, AR1(R) was not significant for Hermitage so the identity ID was used.

Site	Fixed terms	Random terms
Biloela	stand + Line.out + lincol	Rep + Row + AR1(C):AR1(R)
Dalby	stand + Line.out	Rep + Row + AR1(C):AR1(R)
Dalby Box	stand + Line.out	Rep + Row + AR1(C):AR1(R)
Hermitage	stand + Line.out	Rep + Row + AR1(C):ID(R)

Appendix B

Electronic Supplementary Material
Multi-Environment analysis of sorghum breeding trials using additive and dominance genomic relationships

---

Table B.1: A summary of the number of polymorphic DArT markers per linkage group (LG), lengths of each chromosome in cM and in Mbp, and the average LD for each linkage group for the males and the hybrids.

LG	# DArT markers	length (cM)	Length (Mbp)	Average LD (males)	Average LD (hybrids)
Chr01	4509	184.4	73.7	0.054	0.054
Chr02	3501	228.3	77.7	0.059	0.065
Chr03	3709	168.5	74.4	0.058	0.061
Chr04	2759	169.5	68.0	0.056	0.069
Chr05	1954	119.6	62.2	0.061	0.066
Chr06	2510	165.6	62.2	0.063	0.068
Chr07	1818	132.6	64.2	0.060	0.067
Chr08	1532	111.8	55.3	0.070	0.064
Chr09	2059	135.0	59.4	0.056	0.057
Chr10	2234	111.9	61.0	0.062	0.059

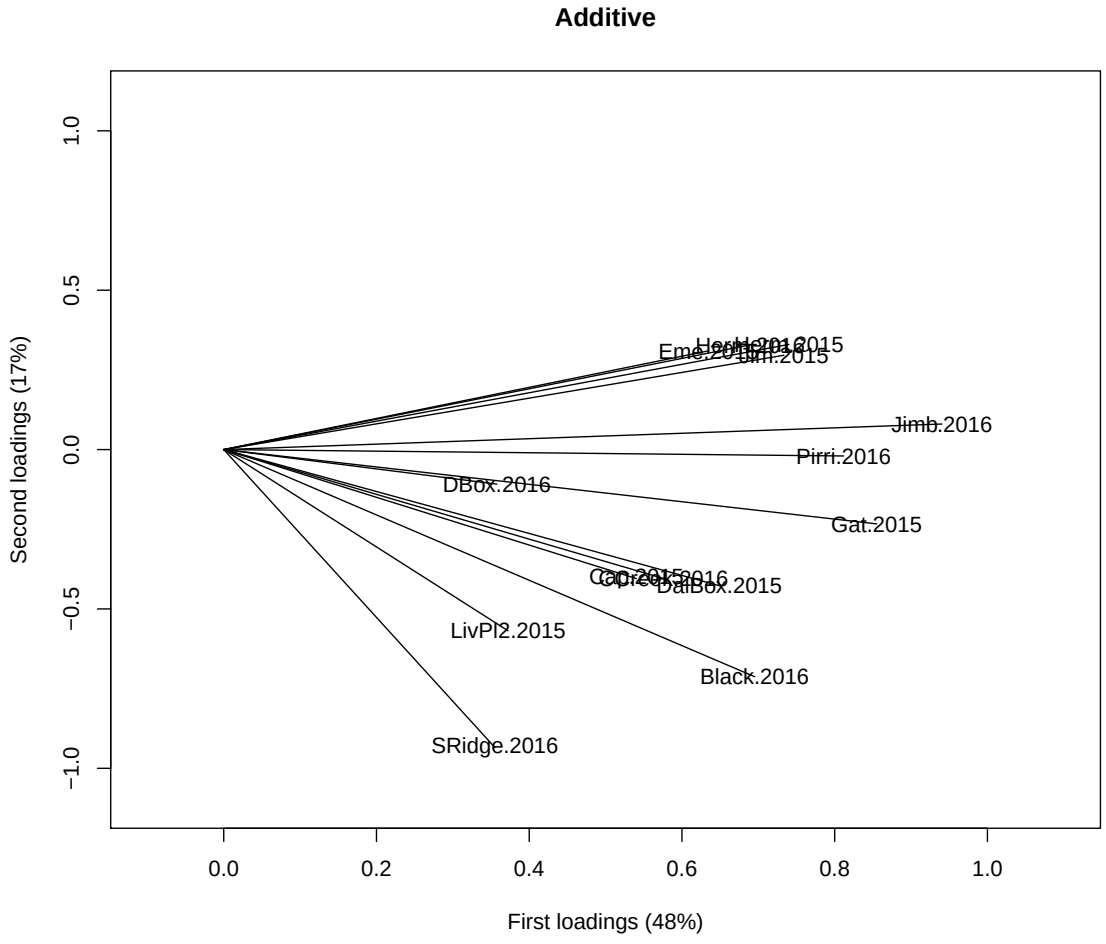

Figure B.1: FA2 loadings for additive partition from model FA2.A

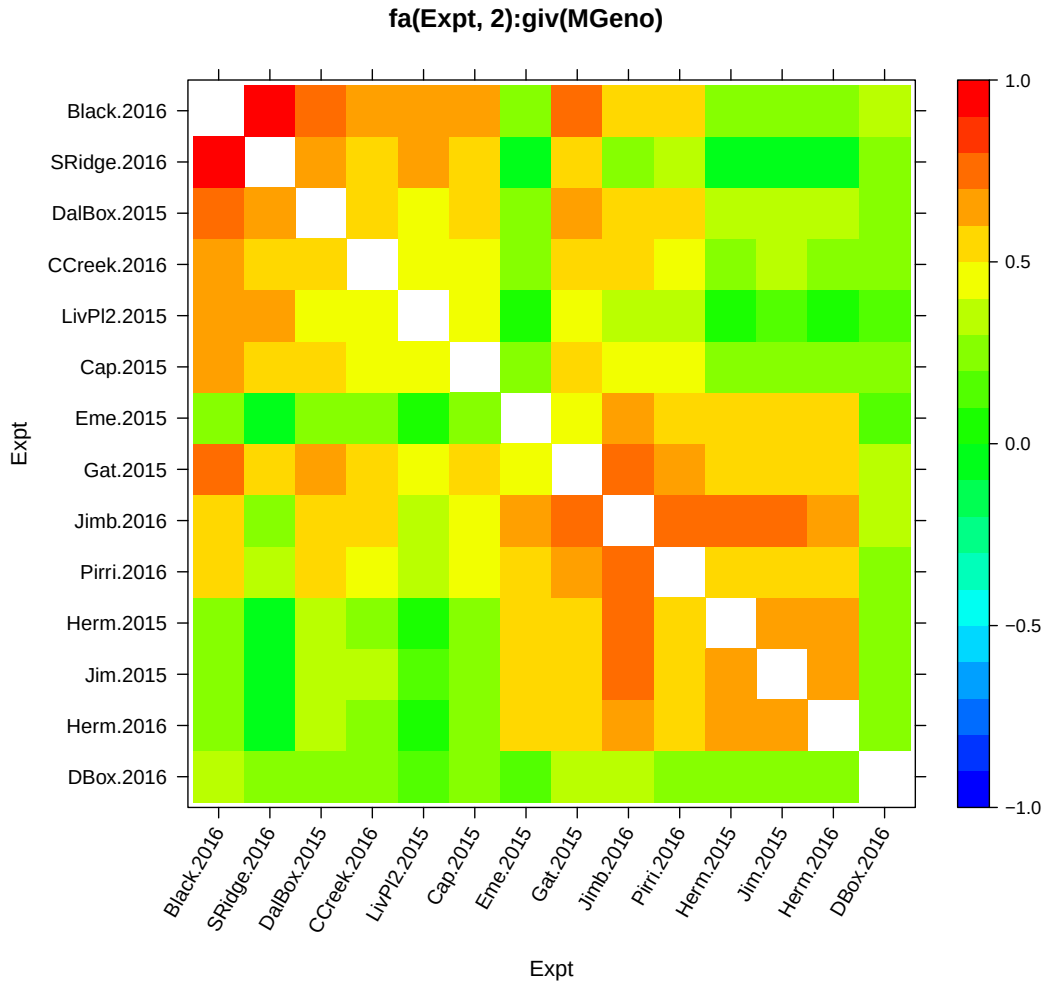

Figure B.2: Between trial correlatins for additive partition from model FA2.A

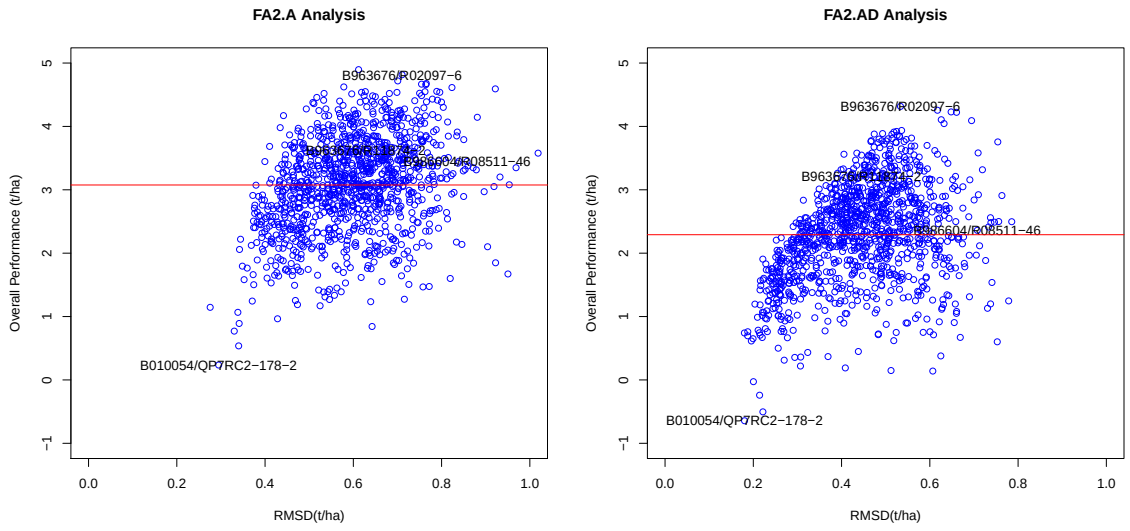

Figure B.3: Additive overall performance versus root mean square deviation (RMSD; a stability measure) for FA2.A on the left and FA2.AD on the right.

Appendix C

Supplimentary material - Identifying efficient strategies for preliminary evaluation in hybrid breeding programs

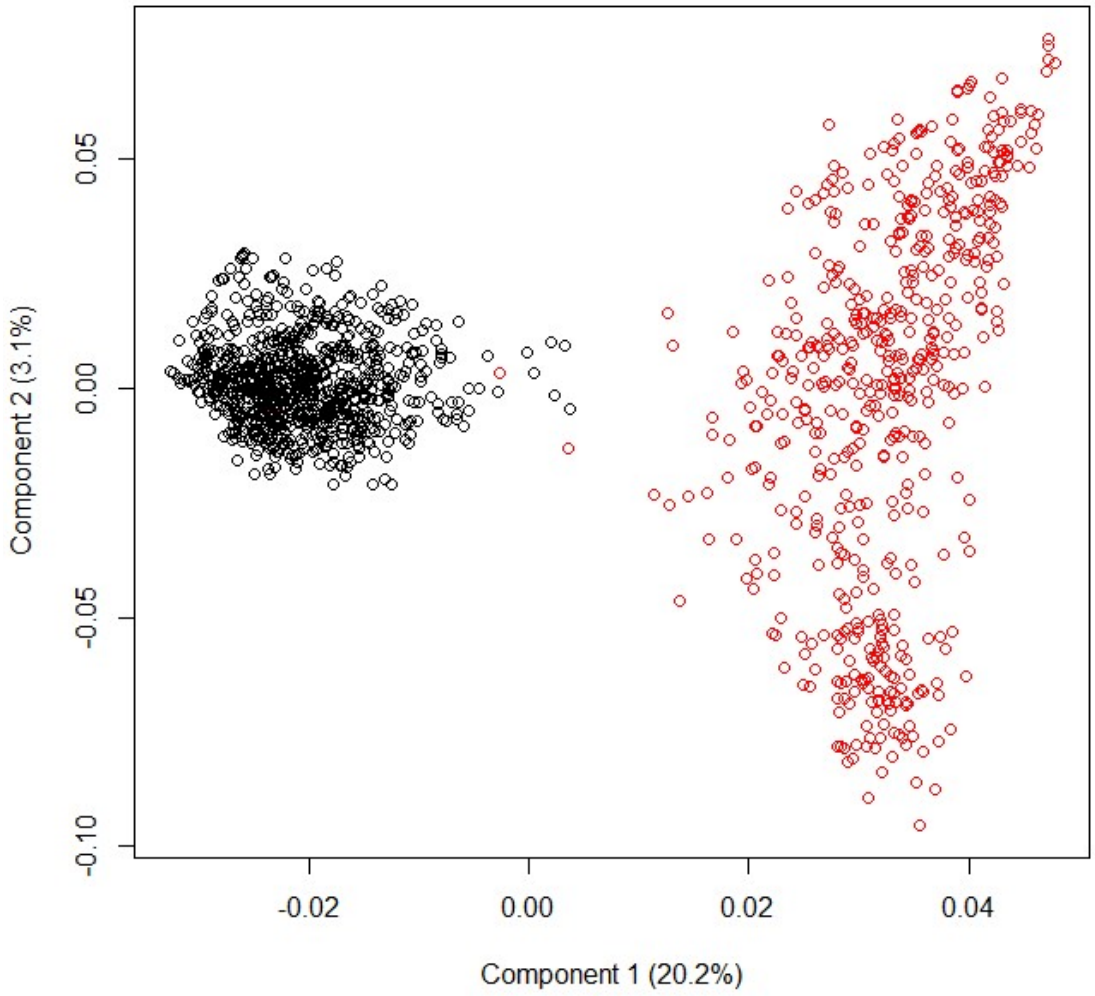

Figure C.1: PCA analysis of Male and Female heterotic groups using genomic data. Females are in black and Males are red.