Rajeev K. Varshney
Manish Roorkiwal
Mark E. Sorrells *Editors*

Genomic Selection for Crop Improvement

New Molecular Breeding Strategies for Crop Improvement

Springer

Genomic Selection for Crop Improvement

Rajeev K. Varshney • Manish Roorkiwal
Mark E. Sorrells
Editors

Genomic Selection for Crop Improvement

New Molecular Breeding Strategies for Crop Improvement

Springer

Editors

Rajeev K. Varshney
Center of Excellence in Genomics
Research Program - Genetic Gains
International Crops Research Institute
for the Semi-Arid Tropics (ICRISAT)
Patancheru, Telangana, India

Manish Roorkiwal
Center of Excellence in Genomics
Research Program - Genetic Gains
International Crops Research Institute
for the Semi-Arid Tropics (ICRISAT)
Patancheru, Telangana, India

Mark E. Sorrells
Department of Plant Breeding and Genetics
Cornell University
Ithaca, NY, USA

ISBN 978-3-319-63168-4 ISBN 978-3-319-63170-7 (eBook)
DOI 10.1007/978-3-319-63170-7

Library of Congress Control Number: 2017951639

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Today the world is facing an unprecedented challenge: how to feed a growing population predicted to reach over 9.1 billion people by 2050 on a resource base threatened by climate change and with limited options for bringing new arable land under cultivation. Associated challenges of high levels of women and child malnutrition in Asia and sub-Saharan Africa and environmental degradation add to the complexity threatening our future.

To meet these challenges, farmers need improved varieties of crops which give higher productivity and economic returns while withstanding risks induced by climate change such as high temperatures, changing spatial and temporal rainfall distribution, and emerging pests and diseases. These new varieties must also provide consumers, both rural and urban, with access to food that is highly nutritious and safe.

A key task before the agricultural research community is to integrate genomics into modern crop improvement to unlock the genetic diversity of food crops in ways that maximize the availability of improved varieties with the range of production and resilience traits (drought, heat, disease, and pest tolerance) alongside improved nutritional value. Modern genomics provides new tools for increasing both the yield and quality of crop products. Next-generation breeding will need to draw on genomics as the "best bet" for sustainably eradicating hunger, malnutrition, and poverty. Genomics coupled with advanced analytics and precision phenotyping can dramatically increase our capacity to utilize genetic diversity and develop highly nutritious, stress-tolerant crop varieties faster and cheaper than ever before and so response with urgency to the realization of the Sustainable Development Goals by 2030.

Despite rates of genetic gain leveling off in many cropping systems, significant efforts in genetic improvement have helped increase productivity and develop climate-resilient varieties. Next-generation sequencing technologies are reducing drastically the cost of genotyping and enabling genome-wide marker data to support the design, development, and delivery of robust and nutritious crop varieties.

Genomic Selection for Crop Improvement is a timely resource to fill the gap between genome science and crop breeding. In capturing the insights of global leaders on genomics and crop improvement, I am confident that this resource will advance our collective understanding and application of modern tools to unlock our wealth of crop genetic diversity to deliver resilience and profitably for farmers and nutritional value to consumers.

ICRISAT
Patancheru, Telangana, India
October 17, 2017

Dr. David Bergvinson

Preface

The past decade has seen a tremendous shift toward using next-generation sequencing (NGS) technologies for development of powerful tools to identify underlying genes for both simple and complex traits. The advent of NGS and high-throughput genotyping technologies have reduced the genotyping cost significantly and made it possible to use genome-wide marker data for prediction of phenotype to help reduce the cost of phenotyping. Integration of genomics tools with conventional breeding can forge new directions to meet environmental challenges efficiently in less time and more accurately. First-generation molecular breeding approaches (marker-assisted backcrossing (MABC) and marker-assisted recurrent selection (MARS)) require a lengthy process for developing genetic populations for identification of linked markers for a few simply inherited traits but failed to improve complex traits such as yield and drought tolerance due to their technical and genetic limitations. In the case of complex traits which are generally controlled by large number of genes/quantitative trait loci (QTLs) with small effect, "genomic selection (GS)" has gained momentum in plant breeding due to the decline in the genotyping cost. One of the strengths of GS lies in the ability to select an individual without phenotypic data (predicting the individual's breeding value) based on a prediction model trained with phenotypes and genotypes. However, practicing GS is not as simple as MABC and MARS and requires an understanding of complex statistical models. GS has been widely used in cattle breeding and more recently has gained popularity among plant breeders. This book is a timely effort to compile details about GS for users providing basic as well as advanced understanding. The content of this book will serve as a useful reference for users, covering the germplasm to be used, phenotyping evaluation, marker genotyping methods, and statistical models involved in genomic selection.

A total of 21 authors (Contributors) have contributed to the nine chapters of the book. The editors of this volume are grateful to all the authors for their contributions and for their commendable effort in summarizing the published/unpublished research work in a comprehensive, up-to-date manner. In addition, the cooperation they have extended in terms of timely completion and revision of chapters from time to time is well appreciated. While editing this book, the strong support received from many other colleagues (Drs. Aaron Lorenz, Isabel Vales, John M. Hickey, and José Crossa) to review the chapters is greatly appreciated. Their constructive comments and suggestions have been instrumental to further improve the chapters.

The editors also would like to thank their respective families for their cooperation and moral support as the editorial work for this book took away precious moments that they should have spent together with their families. RKV is thankful to Monika, his wife, for her constant encouragement and support and Prakhar (son) and Preksha (daughter) for their love and cooperation. Similarly, MR is grateful to his wife (Shweta) for her support and encouragement in doing editorial responsibilities in addition to research duties at International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), with special thanks to Divit (son) for his fondness. RKV and MR would also like to extend their sincerest thanks to Dr. David J. Bergvinson, Director General, ICRISAT, and Dr. Peter S. Carberry, Deputy Director General-Research, ICRISAT, for their help and support.

RKV and MR are also grateful to their colleagues from Center of Excellence in Genomics (CEG), Research Program - Genetic Gains, ICRISAT, and the collaborators for their direct/indirect suggestions during planning of the book. The cooperation and help received from Eric Stannard, Eric Hardy, and Rekha Udaiyar of Springer during various stages of the development and completion of this book project is gratefully acknowledged.

We hope that this book will be helpful and useful to students, young researchers, and crop specialists.

Patancheru, Telangana, India	Rajeev K. Varshney
Patancheru, Telangana, India	Manish Roorkiwal
Ithaca, NY, USA	Mark E. Sorrells

Contents

Contributors

Jared Crain Department of Plant Pathology, Kansas State University, Manhattan, KS, USA

José Crossa Biometric and Statistics Unit (BSU), International Maize and Wheat Improvement Center (CIMMYT), México, D.F, Mexico

Dorcus C. Gemenet International Potato Centre (CIP), Lima, Peru

Dario Grattapaglia EMBRAPA Genetic Resources and Biotechnology – EPqB, Brasilia, DF, Brazil

Universidade Católica de Brasília- SGAN, Brasília, DF, Brazil

Sachiko Isobe Laboratory of Plant Genomics and Genetics, Kazusa DNA Research Institute, Chiba, Japan

Ankit Jain International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Telangana, India

Awais Khan Plant Pathology and Plant-Microbe Biology Section, School of Integrative Plant Science, Cornell University, New York State Agricultural Experiment Station, Geneva, NY, USA

Aaron Lorenz University of Minnesota, Minneapolis, MN, USA

Abelardo Montesinos-López Departamento de Matemáticas, Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, Jalisco, Guadalajara, Mexico

Osval A. Montesinos-López Facultad de Telemática, Universidad de Colima, Colima, Colima, Mexico

Akihiro Nakaya Department of Genome Informatics, Graduate School of Medicine, Osaka University, Osaka, Japan

Liana Nice University of Minnesota, Minneapolis, MN, USA

Manish K. Pandey International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Telangana, India

Jesse Poland Department of Plant Pathology, Kansas State University, Manhattan, KS, USA

Jochen C. Reif Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Manish Roorkiwal International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Telangana, India

Jessica E. Rutkoski International Programs, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY, USA

Albert Wilhelm Schulthess Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Mark E. Sorrells Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA

Rajeev K. Varshney International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Telangana, India

Yusheng Zhao Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Chapter 1
Genomic Selection for Crop Improvement: An Introduction

Rajeev K. Varshney, Manish Roorkiwal, and Mark E. Sorrells

1.1 Introduction

Producing sufficient food to meet the demand of vastly growing population and eradication of rural poverty is one of the critically important issues that the world is facing. At the current pace, the world population is expected to cross the mark of nine billion people by 2050 adding further pressure to already exhausted food production systems. Considering the increasingly volatile climate, it will be difficult to maintain the crop production in conjugation with the demand, resulting in increased food prices affecting people who already spend the highest percentage of their disposable income on food. In addition to climate change, limited water resource availability and poor soil health have the potential to restrict food crop production. Furthermore, with increases in the world population, the availability of agricultural land is decreasing. Under these constraints, to meet the rising demand for food, agricultural production must increase by an estimated 50% without greatly increasing water usage or expanding the total land area dedicated to agriculture. Smallholder farmers, especially from underdeveloped and developing countries with limited access to agricultural inputs or agricultural markets, are likely to be affected by rising production costs and climate volatility. As per the United

R.K. Varshney (✉) • M. Roorkiwal
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru,
Telangana, India
e-mail: R.K.Varshney@CGIAR.ORG

M.E. Sorrells
Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York 14853,
USA

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7_1

Nations' estimates, more than 790 million people globally do not have access to sufficient nutritious food (https://sustainabledevelopment.un.org/sdg2), posing a threat to achieving the Sustainable Development Goal (SDG) target of zero hunger (universal access to safe, nutritious, and sufficient food at all times of the year).

In the event of these challenges, there is a need to look for new ways of breeding for food crops and other plant species by using modern technologies. Modern breeding approaches that have the capability to reduce breeding cycle time provide more precision in selection, and more efficient use of genetic variation can be exploited to increase the rate of genetic gains in breeding programs. The rapid decline in the cost of sequencing and genotyping has led to the development of new tools and strategies that can transform the way we breed plant species. In the past, the cost of genotyping restricted the regular use of markers in breeding. In most cases a limited number of markers for the target regions were used for selecting the lines based on presence or absence of agriculturally important alleles. Development of crop varieties using conventional breeding approaches has been effective but time-consuming and labor-intensive. Recent advances in the next-generation sequencing (NGS) technologies have been able to reduce the cost of genotyping and sequencing. This has enabled the use of the high-throughput and cost-effective high-density genotyping. These low-cost genotyping platforms have accelerated the use of markers in the breeding programs using genome-wide approaches (Varshney et al. 2014).

Integrating genomic tools with conventional breeding can have a major impact for dealing with current and future environmental challenges more efficiently. In such conditions, germplasm, genetic, and genomic resources are mandatory in all plant species for rapid genetic gains in productivity of these species using decision support tools. First-generation molecular breeding approaches (marker-assisted backcrossing, marker-assisted recurrent selection) followed a lengthy process for developing mapping populations for identification of markers linked to quantitative trait loci (QTL) for a few simple traits. The majority of economically important traits such as drought tolerance and yield are polygenic in nature and controlled by multiple genes with small effects. In order to improve complex traits, such as drought tolerance and yield, the modern breeding approach, genomic selection (GS) (Meuwissen et al. 2001), can be deployed which specifically aims at improving quantitative traits by using genome-wide marker data without requiring identification of markers associated with QTL for traits of interest. GS uses a "training population" of individuals that have been both phenotyped and genotyped to train a prediction model for calculating genomic estimated breeding values (GEBVs). Subsequently by using this model, GEBVs can be calculated for untested individuals from a "candidate population", and selection candidates (SCs) for making crosses or for advanced yield trials can be identified. Although GEBVs do not identify the function of the underlying QTLs/genes for the trait, they are an excellent selection criterion (Jannink et al. 2010). GS attempts to capture the total additive genetic variance with genome-wide marker coverage and effect estimates (Rutkoski et al. 2011). Therefore, selection of an individual without phenotypic data can be performed based on the individual's predicted breeding value.

The models can be used to calculate GEBVs that help the breeder to identify offspring that will be good parents in the next generation, based solely on genotypic information about an existing line. The use of GEBVs in the context of genome-wide prediction promises to help accelerate the rate of genetic gain in breeding.

The purpose of this book is to bring up-to-date information on GS breeding and its application for crop species improvement. The editors believe that this book can serve as ready reference for geneticists and crop breeders. This chapter introduces the book and provides a summary of different chapters included in the book.

1.2 Methodologies and Models for GS

The first step toward deploying GS in crop breeding is to define a training set, which should be closely related to the selection candidate population. Chapter 2 entitled "Training population design and resource allocation for genomic selection in crop breeding" provides detailed information about composition and optimization of training population design related to population and trait architecture. In this chapter, Aaron Lorenz and Liana Nice highlight the importance of the training population design for predicting the breeding value of lines. The chapter focuses on the process to select a calibration set (training population) for model training and optimizes the resource allocation for field trials. With the advent of new technologies, it has become possible to collect phenotyping data in a more precise manner with decreased error and increased efficiency and in larger quantities. NGS technologies are contributing to a continuing decrease in the genotyping cost and are enabling the prediction of breeding value using genome-wide marker profiling. This chapter also discusses the possible resource allocation in terms of the number of replications for calibration of GS models vs allocation of more plots for model training and allocation of plots within and across environment replication.

The Chap. 3 entitled "Derivation of linear models for quantitative traits by Bayesian estimation with Gibbs sampling" contributed by Akihiro Nakaya and Sachiko Isobe provides detailed information about construction of a prediction model using a linear model. Model parameters are determined using the Bayesian estimation with Gibbs sampling providing a theoretical background sufficient to implement practical software for the model construction. The chapter also provides a sample output by the implemented software. The chapter describes the different prediction models including linear models, one-marker model, two-marker model without interactions, and two-marker model with interactions to predict the trait value of a sample using the environment types and genotypes. Prediction of the trait values of samples based on their genetic and environmental factors is explained using a prediction model that describes the relationship between the explanatory factors observed in the samples and the trait values. This chapter suggests that defining a prediction model for the target trait enables the selection to be based on the predicted trait values, making it an essential part of genomic selection. When the number of markers is greater than the number of samples, the prediction model will be distorted. In order to address the issue related to model overfitting, detailed inspection of the prediction model is necessary, and strategy based on the linear mixed models and the Bayesian estimation may be useful in the prediction of trait values of samples.

Montesinos-López and colleagues highlighted recent advances in models for genomic-enabled prediction developed for ordinal categorical and count data in Chap. 4 entitled "Bayesian genomic-enabled prediction models for ordinal and count data". Authors used these two models on simulated as well as a real dataset using Bayesian framework suggesting that models used are a good alternative for analyzing ordinal and count data in the context of genomic-enabled prediction. Tested models have an advantage to perform an exact logistic or probit ordinal regression without having to do approximations to perform a logistic ordinal regression. Genotype (G) and environment (E) interaction is expected to affect the prediction accuracies, and therefore modelling G × E in the context of genomic-enabled prediction plays a central role in crop breeding for the selection of candidate genotypes. In order to best use GS models, understanding the data type being analyzed is important before deciding on the modelling approach to be employed.

1.3 GS in Field Crop Breeding

GS has been used or is being used in several crop breeding programs. This book includes three chapters on applications offering both constraints and opportunities of GS in crop breeding. The Chap. 5 entitled "Genomic selection for small grains improvement" by Rutkoski and colleagues presents an overview of GS efforts being undertaken in the small grain cereals. Authors in the chapter have explained different approaches for implementation of GS in applied breeding programs. A total of 40 GS studies have been undertaken so far in small grains including wheat, barley, oat, rye, durum wheat, perennial ryegrass, and intermediate wheatgrass. This chapter also discusses the factors affecting the GS prediction accuracies in small grains and highlights the applicability of GS for analyzing and predicting G × E. They have discussed various scenarios affecting gain from selection and cost relative to conventional breeding. Authors discussed the cost-benefit ratio for deploying GS in cereal crops.

In Chap. 6 entitled "Current status and prospects of genomic selection in legumes", Jain and colleagues from ICRISAT provide an update on molecular breeding in legumes and describe the ongoing GS efforts in some legume-breeding programs including soybean, alfalfa, pea, chickpea, and groundnut. Legumes have witnessed significant progress in the field of genomics and genetics in the past decade, and efforts to deploy MAS have yielded some success for developing superior legume varieties. However, as expected, MAS has not been that successful for addressing complex traits such as drought and yield and therefore, efforts to deploy GS in legume breeding were initiated. Authors have suggested that it is time for other legumes to start deployment of GS in those breeding programs to achieve a higher rate of genetic gain.

Hybrid breeding has been successful over varietal improvement in several crops. Schulthess and colleagues from the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany, describe the basic concepts of hybrid breeding and deployment of GS methods to simplify the philosophy underlying GS with hybrid breeding in Chap. 7, "Genomic selection in hybrid breeding". Authors have explained the basic concepts relevant to hybrid breeding including dominance, heterosis, combining abilities, and heterotic groups and patterns. The chapter also describes the deployment of GS for hybrid genotypes using cross-validated prediction accuracy, accommodating dominance effects within the GS model and other GS approaches employed in hybrid breeding. Deployment of GS in hybrid breeding is very challenging as many variables impact in hybrid breeding as compared to pureline breeding. Authors propose an integrated plan with multidisciplinary skills of breeders, scientists, and technicians before implementing GS in hybrid breeding.

1.4 GS for Improvement of Clonal Crops and Tree Species

Breeding in clonal crops and tree species is different from field crops. Therefore, Gemenet and Khan in Chap. 8 entitled "Opportunities and challenges to implementing genomic selection in clonally propagated crops" discuss issues related to deployment of GS for improving the rate of genetic gain in clonal crops. Authors highlight conventional breeding approaches for clonal crops that involve crossing and planting of true seed plants in different generations followed by evaluation of clones for several generations, making it a time- and resource-consuming process. Therefore, GS-based selection of true seed plants can expedite the breeding process. The chapter also describes the challenges including modelling of genetic effects and heritability, linkage disequilibrium between markers and QTLs, genetic architecture of traits, size of training population, and number of generations following training model to deploy GS in clonal crops. For instance, GS models generally handle additive effects and assume dominance and epistatic effects as part of the residual which is not the case for clonally propagated crops, as dominance and epistatic effects play an important role along with additive effects and need special consideration. Therefore, for clonal crops, GS models with the capability to include additive, dominance, and epistatic genetic effects need to be employed for analysis.

For tree species, Dario Grattapaglia from EMBRAPA Genetic Resources and Biotechnology, Brazil, provides perspectives of genomic selection and a comprehensive discussion on the factors relevant to GS in tree breeding in Chap. 9, "Status and perspectives of genomic selection in forest tree breeding". The chapter highlights the potential of GS in enhancing the rate of genetic gain in a tree breeding program by reducing the selection cycle. In the case of a tree breeding program, the long generation time typically necessary to complete a full breeding cycle can be reduced by genotyping young seedlings and predicting their phenotype instead of waiting for long a breeding cycle of 4–20 years or more. The authors have compiled and presented all GS experimental studies in forest trees along with their key attributes and performance of predictive abilities for different traits in the chapter.

1.5 Summary

As can be seen from the introduction of eight different chapters, GS, a modern breeding approach, is gaining popularity and becoming the choice for many breeders for improving complex traits. The book provides up-to-date information about models, methodologies, factors affecting prediction accuracy, and some examples of deployment of GS for crop improvement. This book will serve as reference for users that provides basic as well as advanced understanding about GS. The book is expected to serve as a useful review for users that explains the germplasm to be used, phenotyping, marker genotyping methods, and statistical models involved in GS. It also includes some examples of ongoing activities of genomic selection in cereal, legume, clonal crop, and tree species.

References

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–177

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Rutkoski JE, Heffner EL, Sorrells ME (2011) Genomic selection for durable stem rust resistance in wheat. Euphytica 179:161–173

Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLoS Biol 12(6):e1001883

Chapter 2
Training Population Design and Resource Allocation for Genomic Selection in Plant Breeding

Aaron Lorenz and Liana Nice

2.1 Introduction

Obtaining accurate and inexpensive estimates of genetic value is a fundamental goal for plant breeders. To obtain these estimates and choose new varieties, breeders continue to rely heavily on standard phenotyping practices for their crops and traits of interest. Series of phenotypic testing procedures employed by plant breeders can vary in scale, complexity, and relevance, both within and across breeding programs. Scale can range from early generation, single plant observations to large, prerelease strip trials. Similarly, the complexity of phenotyping traits within a breeding program can range from measuring flowering time, which can be reliably phenotyped in a single environment in many cases, to drought tolerance which can only be measured in a field setting if specific weather conditions occur or if specially designed water stress nurseries are available. While phenotyping followed by selection is the primary means of advancing lines, the time, cost, and environmental error associated with obtaining phenotypic values leave room for improvement. Advancements in phenotyping technologies have resulted in decreased error, fewer inefficiencies in the phenotyping process, or larger quantities of phenotypic data (Araus and Cairns 2014). An alternative yet complementary approach to reducing phenotyping expenditures involves implementing genomic selection using high-throughput molecular markers in breeding programs (Cabrera-Bosquet et al. 2012).

Initially, molecular markers were used in the context of marker-assisted selection (MAS). This approach typically requires identification of tightly linked or causal markers through mapping or cloning of quantitative trait loci (QTL) using mapping populations or discovery panels. These markers are then used in parallel

A. Lorenz (✉) • L. Nice
University of Minnesota, Minneapolis, MN 55455, USA
e-mail: lore0149@umn.edu

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7_2

with measured phenotypes to make selections in the breeding program (Johnson 2004). The development of genomic selection techniques has altered the relationship between markers and phenotypic data in breeding programs by introducing a new role for phenotyping. Instead of using phenotypic data for direct measurement of the phenotypic or breeding value of lines, phenotypic data in the context of genomic selection is used to estimate marker effects and develop marker-based predictive models. This is accomplished by developing calibration sets, or training populations, that have been both phenotyped and genotyped with dense, genome-wide markers. From this calibration set, a statistical model using all marker information simultaneously is applied to predict the breeding values of individuals that had not been phenotyped, known as the target population or prediction set. With accurate predictive models, breeders can minimize the number of individuals that are phenotyped and continue selection in environments that are not conducive to obtaining quality phenotypes, such as off-season nurseries. Both scenarios can effectively reduce the cost and/or time necessary for achieving the desired genetic gain.

As genotyping costs continue to decrease, genomic selection will play an increasingly important role in plant breeding. Research surrounding the hypothetical and empirical implementation of genomic selection is an active field of study, and the resulting techniques are being adopted by breeders in many crop species. This movement toward an increasingly data-rich breeding process leads to questions surrounding the application of statistics, experimental design, and quantitative genetics, to the selection of progenies for advancement and varietal release. While the implementation of genomic selection may not affect the methods used for phenotyping per se, breeders will need to consider training accurate genomic prediction models when designing field trials, which would involve at least two aspects: (1) selection of genotypes for field testing that are informative for model building (i.e., training population design) in addition to those being advanced toward variety release and (2) the allocation of field plots to genotypes. The objective of this chapter is to review and discuss studies related to these two important topics. It was our aim to provide the reader with a simple and hopefully intuitive introduction to these topics.

2.2 Training Population Design

A critical first step toward the use of genomic selection is the establishment of the training population (Jannink et al. 2010). Training population composition and the way in which it's established varies according to the role of genomic selection, whether it be rapid recurrent selection within a closed population, selection within a single biparental family, or selection among exotic plant accessions comprising a germplasm collection. Approaches to compiling training populations include the collection of new phenotypic data from targeted trials as well as the mining of historical phenotypic data available on genotyped lines. Once again, the choice between these two basic strategies depends upon the role of genomic selection in a crop improvement program. The most important consideration of training population design is the target population. In other words, the target population should be defined first and foremost, and then the training population is designed around the target population. There are two basic aims of training population design: (1) minimize costs associated with phenotyping by selecting smaller training populations, and (2) maximize prediction accuracy for the set of individuals being predicted. Balancing these goals should help breeders avoid poor prediction accuracies or wasted resources.

To aid in this decision process, we review a range of studies that explore composition and optimization of training population design. Windhausen et al. (2012) laid out four breeding scenarios under which genomic selection may be used: (1) training and target populations are segregating progenies from the same cross, (2) training and target populations include related and unrelated genotypes, (3) training and target sets include lines from a diverse germplasm collection, and (4) recurrent selection within a closed synthetic population. Literature on training population design for the first three scenarios will be reviewed. Literature on training population design for the case of synthetic populations is sparse at the present time; however one recently published study sheds light on this topic (Schopp et al. 2017). Following the discussion of breeding scenarios, we explore methods of training population selection and other considerations for training population design related to population and trait architecture.

2.2.1 Training and Target Populations Are Segregating Progenies from the Same Cross

The most straightforward way to conduct genomic selection is to create family-specific training populations. In this scenario, individuals from the same family, or biparental population, are used as both the training population and target population. This approach has been discussed extensively in the maize breeding literature (Bernardo and Yu 2007; Windhausen et al. 2012; Lorenz 2013; Jacobson et al. 2014), where large biparental families of inbred or doubled haploid lines are common, as well as the wheat breeding literature (Heffner et al. 2011). To perform genomic prediction, the entire family is genotyped, with a subset of these lines serving as the training population to train a model to predict the individuals that were not phenotyped. The genomic prediction model can also be used to predict future selection cycles created by intermating selected individuals within the family (Bernardo and Yu 2007; Combs and Bernardo 2013; Massman et al. 2013; Lorenz 2013). This breeding method is similar to marker-assisted recurrent selection in terms of family structure (Johnson 2004), and the first published studies on genomic selection for plant breeding used this approach (Whittaker et al. 2000; Bernardo and Yu 2007).

Within-family predictions are often accurate, and only modest population sizes and marker numbers are needed to achieve good prediction accuracy. High accuracy is possible because of the extensive linkage disequilibrium (LD) generated by the initial hybridization event (Lorenzana and Bernardo 2009; Zhao et al. 2012). This LD, which provides power for QTL mapping in biparental populations, also leads to accurate predictions in the context of genomic selection. Generally, as training population size increases within families, predictive ability increases until a maximum has been reached. When working with high heritability traits, the maximum prediction accuracy will be reached with a smaller training population size.

In an era when genotyping can be less expensive than phenotyping, selecting a subset of individuals to phenotype based on genotype data in order to reduce population size (and thus cost of phenotyping) while maintaining QTL detection power is a desirable goal. This is known as selective phenotyping. It has been shown that selective phenotyping for QTL detection can enhance mapping power and resolution depending on the number of QTL controlling a trait and their effect sizes (Jannink 2005; Sen et al. 2009). Although the increase in power for QTL mapping was minimal under optimized schemes, researchers have explored whether similar optimizations could be used in genomic prediction. Marulanda et al. (2015) simulated a biparental population with training population sets that varied based on a large number of parameters. The parameters examined included measures of collinearity among markers, LD, allele frequency, genetic relationships among lines, diversity indices, mixed model parameters, and phenotypic variance of the training population sets. While many of these factors varied with training population size, none of the parameters derived from marker data were associated with prediction accuracy. However, they did find that selection for enhanced phenotypic variation of the training set led to greater prediction accuracy in the case of smaller training populations. While marker-based optimization would be ideal, the authors proposed that a first round of phenotyping with little replication could be used for training population selection, followed by more intense phenotyping of the optimized set across multiple locations (Marulanda et al. 2015). Ultimately, the lack of population structure in a biparental cross allows for relatively good prediction from a random sample, as long as marker number and population size are large enough to adequately train the selection model. The use of genomic prediction to select non-phenotyped individuals within a single family, however, needs to be carefully considered as studies on resource allocation have suggested little to no benefit to only phenotyping a subset of a single family in order to develop a model to predict the remaining individuals in a family, unless family size is very large (Lorenz 2013; Endelman et al. 2014; Riedelsheimer and Melchinger 2013).

2.2.2 Training and Target Populations Include Related and Unrelated Genotypes

Realistically, models built from single biparental populations are limited in their applications outside of breeding systems with easy access to large population sizes and efficient doubled haploid technologies. The time required to develop and phenotype biparental populations diminishes the potential time savings of implementing genomic selection in place of phenotypic selection. Therefore, methods that combine data across multiple related and/or unrelated families would be valuable for breeders. This can be a challenge because many additional factors come into play when combining data across populations, and adding more individuals to the training population does not necessarily result in greater prediction accuracy as we will discuss below.

The inclusion of related and unrelated genotypes in training and target populations can be further broken down into two scenarios for our purposes here. One scenario includes the development and testing of large families, often consisting of DH lines, as is used in hybrid maize breeding. Families often consist of 150 progenies or more. Under this scenario, it would be possible and appropriate to pool together a few well-chosen families into a single training population. A second common scenario is the development of many, small families. This scenario is common in crops such as soybean and small grains, where crossing is followed by multiple generations of inbreeding followed by visual selection on simply inherited traits and on molecular markers tagging large-effect QTL. The number of progenies per family reaching the yield trial phase is typically small (~20–40) which excludes the possibility of within-family training populations as well as the pooling together of only a few families to form a training population. Rather, training populations would need to be formed by pooling together progenies that are derived from various pedigrees and genetic backgrounds, spanning levels of relatedness. If the populations have been genotyped, ancestral relationships among individuals in the training and the target populations can be used to optimize the selection of training set.

Numerous studies in both plant and animal breeding systems have shown that prediction accuracy suffers when training populations are not related to the target population (Pszczola et al. 2012; Windhausen et al. 2012; Ly et al. 2013; Technow et al. 2013; Albrecht et al. 2014; Lorenz and Smith 2015). Analysis of genomic selection in sheep showed that the strongest predictor of prediction accuracy of each individual was the strength of relationship between the individual being predicted and the top ten relatives in the training population (Clark et al. 2012). In contrast, the mean relationship of the training population to the individual being predicted was a weak predictor of prediction accuracy. Therefore, for an individual to be predicted well using genomic prediction, the training population must include several close relatives to that individual.

Along these same lines, results looking at pooling together large families (first scenario described above) to predict a specific target family have generally indicated that the best results are obtained when the families being pooled share one parent with the target family. The addition of families sharing one parent with the family-specific training population could increase model accuracy above the family-specific training population, especially if the target family is small in size (Schulz-Streeck et al. 2012; Jacobson et al. 2014). Lehermeier et al. (2014) found that the predictive ability of pooled half-sib training populations could achieve similar accuracy to family-specific training populations, but models built using 375 half-sib individuals were needed to reach the accuracy of models built using only 50 full-sib individuals. Riedelsheimer et al. (2013) found that half-sib training populations that shared one parent in common with the target population only reached 50% of the predictive ability of family-specific training populations. This study, however, only included a limited number of families (six), and in reality, breeding programs would likely include many more families from which to pool data.

The use of data from families unrelated to the target population (family) is more problematic. Training populations consisting of only individuals unrelated to the target population generally result in zero or near-zero prediction accuracy (Riedelsheimer et al. 2013; Jacobson et al. 2014; Lehermeier et al. 2014). Moreover, the addition of unrelated families to a family-specific training population can reduce prediction accuracy compared to the family-specific training population alone (Riedelsheimer et al. 2013; Jacobson et al. 2014) or have no effect despite increasing the training population size by up to sixfold (Zhao et al. 2012). Lorenz and Smith (2015) showed a decline in prediction accuracy when individuals less and less related to the target population were added to the training population. Model accuracy was maximized by using smaller training populations that were more closely related to the target population, and the addition of less related individuals (mostly from a different breeding program) reduced accuracy of predictions for all traits. High marker densities may enhance the sharing of information between families and improve prediction accuracy by pooling unrelated families (Hickey et al. 2014). Hickey et al. (2014) found that training populations consisting of families unrelated to the target family could produce models with accuracies reaching 0.70, but only with population sizes approaching 20,000 individuals and marker numbers greater than 10,000. It is possible that such training populations could be constructed within the seed industry, but to our knowledge, nothing in the public sector has yet come close to this scale.

2.2.3 Training and Target Populations Include Lines from a Diverse Germplasm Collection

Besides predicting the genetic value of progenies comprising an active breeding program, another role of genomic prediction includes the prediction of diverse accessions comprising a germplasm collection. Germplasm collections can be very large, containing up to hundreds of thousands of plant accessions. Advancements in genotyping have made it possible to genotype entire germplasm collections (Hearne et al. 2015; Song et al. 2015), opening up the possibility of predicting the performance of all accessions (Jarquin et al. 2016). Phenotyping entire collections, on the other hand, is often not feasible.

In this scenario, the training and target populations are essentially two subsets of the same population, and thus the training population should be selected to represent the entire population. Several studies have examined the performance of chosen statistical criteria and accompanying optimization algorithms in choosing informative training populations.

Two criteria for assessing population design derived from mixed linear model theory have been proposed: prediction error variance (PEV) and the generalized coefficient of determination (CD). The PEV quantifies the error of prediction of each random effect in the model. It is a function of the ratio of the model error to genetic variance, the number of times an individual is measured, the number of relatives of the individual included in the dataset, and the strength of their relationship. The CD is defined as the amount of variation in true contrasts of genetic values by predicted contrasts of genetic values, where the contrast is between each individual being predicted in the target population and target population mean (Laloë et al. 1996). Optimizing the reliability of these contrasts rather than of the predictions per se takes the covariances among the individuals comprising the target population into account and thus prevents the selection of closely related individuals for training population formation (Rincent et al. 2012). Because genetic variance is not included in the calculation of PEV, using this method may result in selecting a relatively narrow training population that contains many close relatives. These statistics are calculated for each individual in the target population, and the average value across the target population (i.e., PEVmean and CDmean) is the final optimization criteria.

Criteria related to minimizing PEV have been previously used to optimize data collection in animal breeding programs (Laloë and Phocas 2003; Kuehn et al. 2007). Rincent et al. (2012) expanded the use of these criteria for training population design and genomic selection in plant populations by implementing them in combination with a simple exchange algorithm. An exchange algorithm involves removal and replacement of one individual in the training population, followed by calculation of the optimization criteria (e.g., PEVmean, CDmean) for the newly formed training population. If the removal and replacement results in an improvement measured by the chosen criteria, then the newly added individual remains; else it is removed in place of another randomly sampled individual from the pool of candidates. Rincent et al. (2012) found that an optimization scheme based on a CDmean-optimized training population resulted in models of higher accuracy compared to random sampling. An optimized population of 100 individuals achieved the same prediction accuracy as a randomly selected population of 200, indicating large reduction in costs associated with phenotyping if this method is applied. The CDmean criteria typically outperformed PEVmean and other diversity criteria such as mean genetic relatedness of the selected training population measured by the genomic relationship matrix. Isidro et al. (2015) applied these same criteria to rice and wheat panels. These authors found that a simple, stratified sampling method that ensured representation of each subpopulation in the training set was superior for the highly structured rice population, whereas the CDmean method was superior for the minimally structured wheat population. This indicates that training population optimization does depend on the population, as well as the trait.

Akdemir et al. (2015) also showed a consistent benefit to optimizing training populations using relationship-based selection procedures. These authors focused on a principal component-based approach that increased computational efficiency and selected training populations with regard to a specified target population, rather than relationships within the training population itself. Their results suggest that such methods hold great potential to help choose maximally informative training populations. Software that implement these methods have been made available to the general user (Rincent et al. 2012; Akdemir et al. 2015).

2.2.4 Sources of Information and Population Genomic Architecture Influence Training Population Design

The overall theme of the literature reviewed above is that relationships between training and target populations are highly important for genomic prediction. It is clear that small training populations can be used, and are likely superior, if they are closely related to the target population. Very large training populations are needed if little to no relationship exists. Some researchers (Campos de los et al. 2013; Habier et al. 2013) have contributed a theoretical basis to the importance of relationships and their interaction with marker density and prediction model. By far the most common methods for performing genomic prediction are ridge regression best linear unbiased prediction (RR-BLUP) and genomic best linear unbiased prediction (G-BLUP). Although these two models are mathematically equivalent under the properties of the multivariate normal distribution (Habier et al. 2013), practitioners of breeding and genomic selection view the information sharing of these models from two different perspectives. From the RR-BLUP perspective, information is shared between training populations and target populations through the LD that exists between markers and QTL. Because of this, as marker-QTL LD increases, prediction accuracy is expected to increase. From the G-BLUP perspective, information is shared via the realized genomic relationships of the training and target individuals, which reflect the higher degree of resemblance of more closely related individuals. Prediction of selection candidates is a function of the weighted sum of phenotypes of individuals in the training population, with weights being proportional to the genomic relationships (Campos de los et al. 2013). Depending on the family structure and distribution of relationships, only a few close relatives could be heavily weighted in the calculation of the genomic predictions, or weights could be more uniformly distributed among individuals in training populations that are distantly related to the target population.

Ultimately, it is the genetic relationships at causal loci that influence the effectiveness of training populations to predict trait values in prediction sets and not genetic relationships calculated according to markers (Habier et al. 2013; Campos de los et al. 2013). The genomic relationship matrix, calculated using genome-wide markers, is an estimate of the genomic relationship matrix at the causal polymorphisms. Therefore, the accuracy of this estimation is what determines the effectiveness of G-BLUP (Campos de los et al. 2013). The resemblance between the genomic relationship matrix at causal polymorphisms and the estimated genomic relationship matrix based on markers is determined by marker-QTL LD, which in turn is determined by pedigree relationships of the population, population history and diversity, and marker density. Formula for calculating PEV and reliability of predictions using expected genomic relationships based on pedigree data was derived by Henderson (1975). Under these expectations, the reliability of predictions approach 1.0 as the population size goes to infinity. This is even the case if the training population is distantly related to the target population, although the number of individuals required to increase accuracy is much higher compared to the addition of more closely related individuals (Campos de los et al. 2013). Campos et al. (Campos de los et al. 2013) showed that the marker-QTL LD sets an upper limit to prediction accuracy. This limit is lowered when there is a lack of relationship between the training and target populations due to a decrease in marker-QTL LD. This is especially true for distantly related individuals where genomic relationships can be variable with respect to which markers are in high LD with QTL, leading to a major source of error in the G-BLUP model (Hill and Weir 2011). The expected value of realized or pedigree relationship decreases, while the variance of the realized relationship increases (Hill and Weir 2011).

Another way to look at this problem is by partitioning the information contained in the genomic relationship matrix into three components: (1) marker-QTL LD, which is an association between alleles among the population founders; (2) linkage or co-segregation of alleles created by pedigree relationships at QTL; and (3) additive genetic relationships captured by markers (Habier et al. 2013). Habier et al. (2013) used simulations and models to partition these three sources of information. First, they showed that large population sizes and high marker densities are needed to exploit the LD source of information. Secondly, the proportion of accuracy from shared additive genetic relationships is reduced if training populations are expanded by adding unrelated individuals. Accuracy due to LD, however, might be able to compensate for low relatedness if very large training population sizes and/or high marker densities are available. Still, Habier et al. (2013) present an example from cattle data where the increase in the accuracy from LD could not compensate for the loss of information from additive genetic relationships, and an overall decrease in accuracy was observed after the addition of unrelated individuals. However, in their maize example, additive genetic relationship accuracy was not changed by increasing training population size, possibly due to a stronger family structure with many more close relatives in the maize training population.

2.3 Resource Allocation for Phenotyping for Genomic Prediction Model Calibration: To Rep or Not to Rep

A key design aspect of breeding programs is the allocation of resources among breeding trials in terms of population size, number of replications, and locations. Allocation decisions are multifaceted, involving consideration of trait logistics, selection intensity, breeding stage of the materials being tested, and any associated genotyping costs. These decisions affect the genetic gain that is possible, as well as the power to detect QTL or accurately estimate marker effects. Considering selection in general, the fundamental trade-off is between achieving accurate estimates of genotypic value by increasing replication and sampling a greater number of individuals to increase the chance of identifying superior genotypes (Gauch and Zobel 1996). Bos (1983) explored the optimum replication scheme for breeding programs with respect to heritability. Because replication decreases phenotypic variance, it also increases heritability. However, this increase only occurs to a point, after which, fundamental changes to the experimental design would be needed to improve heritability (Gauch and Zobel 1996). Therefore, more replication generally results in better selection outcomes, with the exception of situations where heritability is high and selection intensity is relaxed (Bos 1983). Gauch and Zobel (1996) extended the scope of the Bos (1983) findings to consider the precision of data collected and the relative efficiency of data collected. They found that in experiments with high precision, adding replication beyond two is much less efficient than in lower precision experiments that retain efficiency at greater replication numbers.

When considering markers, the focus changes from identifying the genotypic value of individuals to estimating the additive genetic values of alleles. Knapp and Bridges (1990) identified sources of variation in a QTL mapping experiment and found that increasing population size instead of replication resulted in higher power to detect QTL, particularly when residual genetic variation existed in the population. Other studies reported similar findings, where larger population sizes generally result in higher power of QTL detection, and only moderately sized populations of 150–300 individuals benefit from replication (Schön et al. 2004). Because of the similarities between QTL mapping and MAS, resource allocation recommendations for QTL mapping seem to transfer well to the context of MAS. Moreau et al. (Moreau 2000) showed that larger population sizes resulted in maximum gain from selection when traits were controlled by 5–10 QTL and when genotyping costs were equal to phenotyping costs. The shift toward genomic selection has required a reevaluation of these resource allocation recommendations in the context of a cultivar development program.

In contrast to MAS, genomic selection aims to improve traits that are influenced by many more QTL. In addition, because MAS considers marker effects as fixed and statistical thresholds are used to determine which markers are used to calculate marker scores, the success of MAS is closely related to QTL detection power. Here, we will explore recent published literature that aims to address the resource allocation questions relevant to genomic selection breeding programs and are not sufficiently addressed by previous MAS studies. Specifically, we review: (1) the value of replication for calibrating genomic selection models, (2) allocation of plots to stages within the breeding cycle, and (3) allocation of plots to within versus across environment replication.

2.3.1 Replication and Plot Allocation for Calibrating Genomic Selection Models

To determine whether resource allocation recommendations for MAS can be extrapolated to genomic selection models, Lorenz (2013) compared the accuracies of genomic selection models (RR-BLUP) and MAS models (ordinary least squares, OLS) under varying resource allocation schemes. The factors studied included total plot budget, relative cost of genotyping in comparison to phenotyping, population size, number of replications, heritability, and percentage of phenotyped individuals. A very clear distinction in resource optimization between GS and MAS models was found. Prediction accuracy was always substantially lower with MAS, and the effect of replication was more apparent for MAS. Within a set budget, the addition of replications and a consequent reduction of total individuals screened lead to a decrease in accuracy with MAS. In contrast, the RR-BLUP model remained fairly constant across different resource allocation scenarios, with low heritability, high marker cost scenarios slightly favoring fewer individuals, and more replication. When the total number of individuals was varied, the accuracy of genomic selection models began to level off around 50–75 individuals, whereas MAS models took many more individuals to achieve moderate prediction accuracies and continued to improve as the numbers increased. These results suggest that the underlying considerations for MAS are different from genomic selection.

2.3.2 Allocation of Resources Across Preliminary and Advanced Breeding Tests

Breeding programs are generally structured with less replications in early generation screening, followed by greater replication, larger scale, and higher-cost trials in later generations (Bernardo 2010). Breeders must take this tiered structure of the breeding program into account when planning for genomic selection implementation. The stage at which genomic selection is implemented can affect genetic gain as well as costs. Bassi et al. (2016) compared a series of wheat breeding schemes that implemented genomic selection starting in generations F_2, F_3, F_4, or F_7. They found that without including phenotypic selection at some stage in the program, early generation F_2 implementation had the highest potential for gain per year, but also the highest genotyping costs. Longin et al. (2015) found that genomic selection without a stage of phenotypic selection would only be useful with very high prediction accuracies, possibly unrealistically high accuracies.

When accuracies are low, genomic selection can fill the role of a pretest, whereby a low selection intensity is applied to remove the lowest performing individuals (Longin et al. 2015). Most studies have focused on overall accuracy of genomic selection, without considering the effectiveness of these selection schemes to accurately remove the worst individuals or include the best. Endelman et al. (2014) proposed using a response to selection metric R_{max} based on the maximum genotypic value of selections instead of R_{mean} based on mean values for selection to analyze genetic gain in preliminary yield trials. Because the mean of the selected population decreases as more individuals are selected, the R_{mean} measure of genetic gain may encourage overly stringent selection in early generations that have less precise phenotypic estimates. Additional studies are needed to expand on the use of genomic selection for early generation screening.

In contrast to early generation genomic selection, Bassi et al. (2016) compared intermediate and later generation schemes. They found that implementation in the F_3 and F_4 was a good compromise between no stage of phenotypic selection and the minimal benefits of F_7 implementation. While F_7 implementation might be attractive to breeders because of its ease of implementation and lower genotyping costs, this scheme resulted in minimal benefit over phenotypic selection alone. Longin et al. (2015) concluded that for traits such as yield in wheat, with prediction accuracies of approximately 0.3, one stage of genomic selection followed by one stage of phenotypic selection provides the best compromise between genomic and phenotypic selection.

2.3.3 Across Environment Versus Within Environment Replication

For simplicity, much of the literature surrounding the topic of resource allocation focuses on trade-offs within single environments, but the distribution of plot resources across environments is a major consideration for breeders. Riedelsheimer and Melchinger (2013) attempted to tackle this issue by developing a resource allocation planning tool for distributing plot resources across and within environments. Their tool is limited to a single cycle of selection in biparental populations, and it requires some degree of estimation based on previous experimental data. Their calculations extend those developed by Daetwyler et al. (2008) to include considerations of multi-environment testing. They found that larger budgets favored more environments, with a lower proportion of plots being allocated to the training set. As the budget decreased, the training set became a larger proportion of the plots, and the number of environments tested decreased. Furthermore, they emphasize that under low-budget scenarios, the optimization has much less flexibility than under large-budget scenarios. Overall, their findings suggest relatively few environments are needed for high prediction accuracy.

Endelman et al. (2014) looked at the effect of spreading replicates across locations in preliminary yield trials under fixed budgets. They found that accuracy increased as individuals were replicated across locations, but under a fixed budget, the optimum accuracies were obtained without replication across locations, unless the budget forced a relatively small training population size. That is, each individual should be phenotyped in only one environment, and population size should be maximized to the extent the total number of plots across environments allows. Markers provide the connectivity between environments. In contrast, across environment phenotypic estimates based on phenotyping alone were poor when individuals were phenotyped in single environments. This reinforces the idea that shared marker information does provide the connectivity between individuals, providing potential cost savings for breeders implementing genomic selection.

2.3.4 Conclusions

The role of phenotyping in plant breeding is rapidly shifting from its previous sole purpose of providing information for making breeding line advancement, parent selection, and variety release decisions to providing the necessary data to train genomic prediction models to enable genomic selection. As this new role of phenotyping increases in relative importance, plant breeders need to rethink how they design field trials, allocate plot resources to genotypes, and which individuals are included in field trials. This review provides a short and simple introduction to this literature. We have two basic conclusions at this time: (1) Training population selection and design should take genetic relationships with the target population into consideration, and optimization criteria such as PEVmean and CDmean combined with exchange algorithms are useful methods for selecting training populations. (2) The number of individuals phenotyped should be maximized by allocating only one field plot to each genotype in most situations. Further research is needed to develop a comprehensive theoretical framework for phenotyping for genomic selection.

References

Akdemir D, Sanchez JI, Jannink J-L (2015) Optimization of genomic selection training populations with a genetic algorithm. Genet Sel Evol 47:38. doi:10.1186/s12711-015-0116-6

Albrecht T, Auinger H-J, Wimmer V et al (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor Appl Genet 127:1375–1386. doi:10.1007/s00122-014-2305-z

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19:52–61. doi:10.1016/j.tplants.2013.09.008

Bassi FM, Bentley AR, Charmet G et al (2016) Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.) Plant Sci 242:23–36. doi:10.1016/j.plantsci.2015.08.021

Bernardo R (2010) Breeding for quantitative traits in plants, 2nd edn. Stemma Press, Woodbury, MN

Bernardo R, Yu J (2007) Prospects for Genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090. doi:10.2135/cropsci2006.11.0690

Bos I (1983) The optimum number of replications when testing lines or families on a fixed number of plots. Euphytica 32:311–318. doi:10.1007/BF00021439

Cabrera-Bosquet L, Crossa J, von Zitzewitz J et al (2012) High-throughput Phenotyping and genomic selection: the Frontiers of crop breeding ConvergeF. J Integr Plant Biol 54:312–320. doi:10.1111/j.1744-7909.2012.01116.x

Campos de los G, Vazquez AI, Fernando R et al (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet 9:e1003608. doi:10.1371/journal.pgen.1003608

Clark SA, Hickey JM, Daetwyler HD, van der Werf JH (2012) The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol 44:4. doi:10.1186/1297-9686-44-4

Combs E, Bernardo R (2013) Accuracy of Genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6:0. doi: 10.3835/plantgenome2012.11.0030

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:e3395. doi:10.1371/journal.pone.0003395

Endelman JB, Atlin GN, Beyene Y et al (2014) Optimal Design of Preliminary Yield Trials with genome-wide markers. Crop Sci 54:48–59. doi:10.2135/cropsci2013.03.0154

Gauch HG, Zobel RW (1996) Optimal replication in selection experiments. Crop Sci 36:838–843. doi:10.2135/cropsci1996.0011183X003600040002x

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194:597–607. doi:10.1534/genetics.113.152207

Hearne S, Franco J, Chen J et al (2015) Genome wide assessment of maize Genebank diversity; synthesis of next generation technologies and GIS based approaches. San Diego, USA

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

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447. doi:10.2307/2529430

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

Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res 93:47–64. doi:10.1017/S0016672310000480

Isidro J, Jannink J-L, Akdemir D et al (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128:145–158. doi:10.1007/s00122-014-2418-4

Jacobson A, Lian L, Zhong S, Bernardo R (2014) General combining ability model for Genomewide selection in a Biparental cross. Crop Sci 54:895–905. doi:10.2135/cropsci2013.11.0774

Jannink J-L (2005) Selective Phenotyping to accurately map quantitative trait loci. Crop Sci 45:901–908. doi:10.2135/cropsci2004.0278

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–177. doi:10.1093/bfgp/elq001

Jarquin D, Specht J, Lorenz A (2016) Prospects of genomic prediction in the USDA soybean Germplasm collection: historical data creates robust models for enhancing selection of accessions. G3 GenesGenomesGenetics:g3.116.031443. doi:10.1534/g3.116.031443

Johnson R (2004) Marker-assisted selection. Plant Breeding Reviews. John Wiley & Sons, In

Knapp SJ, Bridges WC (1990) Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for Unreplicated and replicated progeny. Genetics 126:769–777

Kuehn LA, Notter DR, Nieuwhof GJ, Lewis RM (2007) Changes in connectedness over time in alternative sheep sire referencing schemes. J Anim Sci 86:536–544. doi:10.2527/jas.2007-0256

Laloë D, Phocas F (2003) A proposal of criteria of robustness analysis in genetic evaluation. Livest Prod Sci 80:241–256. doi:10.1016/S0301-6226(02)00092-1

Laloë D, Phocas F, Ménissier F (1996) Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. Genet Sel Evol 28:1–20. doi:10.1186/1297-9686-28-4-359

Lehermeier C, Krämer N, Bauer E et al (2014) Usefulness of Multiparental populations of maize (Zea mays L.) for genome-based prediction. Genetics 198:3–16. doi:10.1534/genetics.114.161943

Longin CFH, Mi X, Würschum T (2015) Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. Theor Appl Genet 128:1297–1306. doi:10.1007/s00122-015-2505-1

Lorenz AJ (2013) Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. G3 GenesGenomesGenetics 3:481–491. doi:10.1534/g3.112.004911

Lorenz AJ, Smith KP (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. Crop Sci 55:2657–2667. doi:10.2135/cropsci2014.12.0827

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161. doi:10.1007/s00122-009-1166-3

Ly D, Hamblin M, Rabbi I et al (2013) Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: a study in cassava. Crop Sci 53:1312–1325. doi:10.2135/cropsci2012.11.0653

Marulanda JJ, Melchinger AE, Würschum T (2015) Genomic selection in biparental populations: assessment of parameters for optimum estimation set design. Plant Breed 134:623–630. doi:10.1111/pbr.12317

Massman JM, Jung H-JG, Bernardo R (2013) Genomewide selection versus marker-assisted recurrent selection to improve grain yield and Stover-quality traits for cellulosic ethanol in maize. Crop Sci 53:58–66. doi:10.2135/cropsci2012.02.0112

Moreau L, Lemarie S, Charcosset A, Gallais A (2000) Economic efficiency of one cycle of marker-assisted selection. Crop Sci 40:329–337. doi:10.2135/cropsci2000.402329x

Pszczola M, Strabel T, Mulder HA, Calus MPL (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci 95:389–400. doi:10.3168/jds.2011-4338

Riedelsheimer C, Endelman JB, Stange M et al (2013) Genomic predictability of interconnected Biparental maize populations. Genetics 194:493–503. doi:10.1534/genetics.113.150227

Riedelsheimer C, Melchinger AE (2013) Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor Appl Genet 126:2835–2848. doi:10.1007/s00122-013-2175-9

Rincent R, Laloë D, Nicolas S et al (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize Inbreds (Zea mays L.) Genetics 192:715–728. doi:10.1534/genetics.112.141473

Schön CC, Utz HF, Groh S et al (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485–498

Schopp P, Muller D, Technow F, Melchinger A (2017) Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium. Genetics 205:441–454

Schulz-Streeck T, Ogutu JO, Karaman Z et al (2012) Genomic selection using multiple populations. Crop Sci 52:2453–2461. doi:10.2135/cropsci2012.03.0160

Sen Ś, Johannes F, Broman KW (2009) Selective genotyping and Phenotyping strategies in a complex trait context. Genetics 181:1613–1626. doi:10.1534/genetics.108.094607

Song Q, Hyten DL, Jia G et al (2015) Fingerprinting soybean Germplasm and its utility in genomic research. G3 GenesGenomesGenetics 5:1999–2006. doi:10.1534/g3.115.019000

Technow F, Bürger A, Melchinger AE (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for Heterotic groups. G3 GenesGenomesGenetics 3:197–203. doi:10.1534/g3.112.004630

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252. doi: null

Windhausen VS, Atlin GN, Hickey JM et al (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 GenesGenomesGenetics 2:1427–1436. doi:10.1534/g3.112.003699

Zhao Y, Gowda M, Liu W et al (2012) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124:769–776. doi:10.1007/s00122-011-1745-y

Chapter 3
Derivation of Linear Models for Quantitative Traits by Bayesian Estimation with Gibbs Sampling

Akihiro Nakaya and Sachiko Isobe

3.1 Introduction

Prediction of the trait values of samples based on their genetic and environmental factors is one of the vital steps in the genomic selection (GS) process used to identify valuable individuals for the use in the breeding process (Meuwissen et al. 2001). The GS strategies are different from other marker-assisted selection (MAS) strategies in that they use the predicted trait values. To obtain the estimated values of a trait, a prediction model that describes the relationship between the observed explanatory factors, e.g., the genetic and environmental factors, in the samples and the trait values is derived using a training dataset consisting of the candidates of the explanatory factors and the true values of the trait. Such a prediction model can be composed of the effects by the explanatory factors selected, and when these effects are thought to be additive, the trait value of a sample is estimated as their weighted summation. The models based on the summation are referred to as linear models and are suitable for focusing on the additive effects of the explanatory factors. Aside from the genetic factors obtained by the genome-wide markers, the Bayesian approach used in the derivation process is the key element for GS.

In this chapter, construction of a prediction model using a linear model and determination of the model parameters using the Bayesian estimation with Gibbs sampling are explained with an assuming dataset example for GS. Knowing the

A. Nakaya (✉)
Department of Genome Informatics, Graduate School of Medicine, Osaka University,
Yamadaoka, 2-2, Suita, Osaka 565-0871, Japan
e-mail: nakaya@gi.med.osaka-u.ac.jp

S. Isobe
Laboratory of Plant Genomics and Genetics, Kazusa DNA Research Institute,
Kazusa-Kamatari, 2-6-7, Kisarazu, Chiba 292-0818, Japan
e-mail: sisobe@kazusa.or.jp

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7_3

details of the model derivation, one can understand what is carried out during the prediction of the trait values. This will also provide a theoretical background sufficient to implement practical software for the model construction. A sample output by the implemented software is presented.

3.2 A Dataset for Genomic Selection

We assume N samples and M genetic markers in a given dataset. The ith sample ($i = 1, \ldots, N$) is associated with a numerical trait value y_i and the genotype of the jth genetic marker z_{ij} ($j = 1, \ldots, M$). A tuple of the values ($y_i, z_{i1}, \ldots, z_{iM}$) represents the data of the ith subject ($i = 1, \ldots, N$). If the samples are also associated with L values indicating whether they have the hth environment type x_{ih} ($h = 1, \ldots, L$), then the tuple of the values for the ith subject is ($y_i, x_{i1}, \ldots, x_{iL}$, $z_{i1}, \ldots, z_{iM}$). The dataset can be represented by a vector of the trait values, the profiles of the environment types in an $N \times L$ matrix, and the profiles of the genotypes in an $N \times M$ matrix. We assume there is no missing data.

Figure 3.1 shows an example of the dataset. The dataset consists of three parts: P, E, and G, respectively, corresponding to phenotypes (trait values), environment types ($L = 2$), and genotypes ($M = 50$) of the samples ($N = 30$). In this example, as shown by the histogram in part P, the baseline values of y_i follow $N(0.75, 0.05^2)$ for the upper 15 samples ($i = 1, \ldots, 15$) and $N(0.25, 0.05^2)$ for the lower 15 samples ($i = 16, \ldots, 30$). Here, $N(\mu, \sigma^2)$ is the normal distribution with mean μ and standard deviation σ (i.e., variance σ^2). The 30 × 2 matrix in part E shows two environment types of the 30 samples. The hth column ($h = 1$ or 2) of the matrix corresponds to the hth environment type. For the 10 samples ($i = 1, \ldots, 5$ and $16, \ldots, 20$), a small value (0.3) is intentionally added to y_i to reflect putative environmental effects. According to whether or not a sample is associated with an environment type, each element of the matrix in part E takes a value of 0 or 1, which are respectively colored in white and gray. In this example, we assume that the 10 samples ($i = 1, \ldots, 5$ and $16, \ldots, 20$) have environment type E1, and the other samples have type E2. Part G shows the genotypes of the samples. The element of the matrix in the ith row and jth column (z_{ij}) shows the genotype of the jth genetic marker of the ith sample. The genotype of the sample is represented using multiple colors. This example, which assumes that each genetic marker has two genotypes, uses two colors (white and gray) to indicate the genotypes of the samples. If a genetic marker has more than two genotypes, additional colors are used. For example, white, light gray, and dark gray would be used for a genetic marker with three genotypes. The 30 × 50 matrix in part G shows the genotypes of the 30 samples at the 50 genetic markers (G1 to G50). The genotypes are randomly generated except for the 10th and 40th genetic marker (G10 and G40), so that these two genetic markers are associated with the distribution of the trait values (the samples with higher trait values have the genotype in gray with high probability).

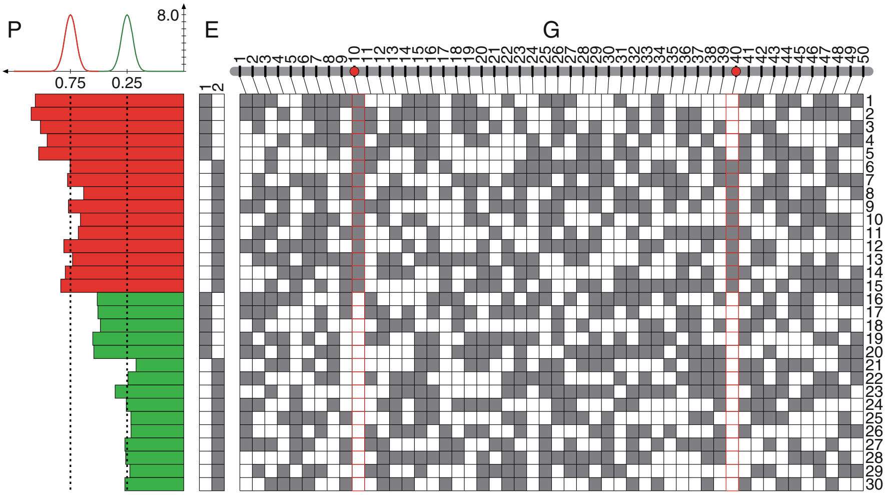


Fig. 3.1 A dataset example for genomic selection

3.3 Prediction Models

One goal of the GS is to predict the trait value of a sample using the obtained environment types and genotypes; for this purpose, a prediction model that can output prediction values for the trait is constructed.

3.3.1 Linear Models

One simple format of the prediction model for the target values (i.e., trait values) is a summation of the effects by the multiple factors (e.g., environmental factors, genetic factors, and interactions among them). A model with this format is referred to as a linear model. Here, note that a matrix, $\boldsymbol{A}'$, which is also denoted by $\boldsymbol{A}^T$, is the transpose of a matrix, $\boldsymbol{A}$, i.e., a matrix whose rows and columns are exchanged. Note also that variables in bold fonts will hereafter refer to vectors and matrices. A linear model for the target values of N samples is given as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \sum\nolimits_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j + \boldsymbol{e}. \tag{3.1}$$

Here, $\boldsymbol{y} = (y_1, \ldots, y_N)'$ represents the target values of the N samples. $\boldsymbol{X\beta}$ is the term representing the effects by the environmental factors (environment types), $\sum\nolimits_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j$ are the terms representing the summation of the effects by the genetic factors (marker genotypes), and $\boldsymbol{e} = (e_1, \ldots, e_N)'$ are the random residuals.

The target value of a sample is made up of two constituent values, one calculated as the summation of the values by fixed effects and the other as the summation of the values by random effects. The linear model given by Eq. 3.1 assumes that the target value of a sample can be explained by the summation of the values from these two types of effects, fixed effects and random effects. The reason why this model is referred to as a linear mixed model is that these two types of effects are included in it. In relation to Eq. 3.1, $\boldsymbol{X\beta}$ and $\sum_{j=1}^{M} \boldsymbol{Z}_j\boldsymbol{u}_j$, respectively, correspond to the fixed effects and the random effects. We assume that the dimensions of $\boldsymbol{X}$ are $N \times L$ and that $\boldsymbol{X}$ has full column rank, i.e., rank $(\boldsymbol{X}) = L$. $\boldsymbol{e} = (e_1, \ldots, e_N)'$ are the random residuals that could not be explained by those effects. Letting σ_e^2 be the variance of the random residuals, the target values follow the normal distributions as follows:

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_M, \sigma_e^2 \sim N\left(\boldsymbol{X\beta} + \sum_{j=1}^{M} \boldsymbol{Z}_j\boldsymbol{u}_j, \boldsymbol{R}\sigma_e^2\right), \tag{3.2}$$

where $\boldsymbol{R}$ is a known matrix of $N \times N$ dimensions. When we assume that the total samples are divided into several groups according to their observed attribute values, we can intuitively consider that the fixed effects represent the global baseline value and the intergroup deviations, while the random effects correspond to the intragroup deviations.

In relation to the linear models for the trait values, typically and naturally, environmental factors constitute the fixed effects, while the genetic factors constitute the random effects for the samples in the same environment group. In Eq. 3.1, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_L)'$ represents the effects of the environment types where β_0 is a variable for the global baseline and β_h is the effect of the hth environment type ($h = 1, \ldots, L$). The ith row of $\boldsymbol{X}$ is the vector of a dummy value for the global baseline and the L observed environment types of the ith sample $(1, x_{i1}, \ldots, x_{iL})$. Here, x_{ih} is set to 1 if the ith sample has the hth environment type, and otherwise it is 0 ($h = 1, \ldots, L$). The ith row of the product $\boldsymbol{X\beta}$ gives the portion of the trait value that is explained by the environmental factors, $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_L x_{iL}$. With respect to $\sum_{j=1}^{M} \boldsymbol{Z}_j\boldsymbol{u}_j$, $\boldsymbol{u}_j = \left(u_{j1}, \ldots, u_{jq_j}\right)'$ are the effects of the genotypes of the jth genetic marker (q_j is the number of the genotypes of the jth genetic marker), and the ith row of $\boldsymbol{Z}_j$ shows the observed genotypes of the ith sample, $\left(z_{ji1}, \ldots, z_{jiq_j}\right)$. Here, z_{jit} is set to 1 if the ith sample has the tth genotype for the jth genetic marker, and otherwise it is 0 ($t = 1, \ldots, q_j$). The ith row of the product $\boldsymbol{Z}_j\boldsymbol{u}_j$ gives the portion of the trait value that is explained by the jth genetic marker, $u_{j1}z_{ji1} + \cdots + u_{jq_j}z_{jiq_j}$. $\boldsymbol{X}$ and $\boldsymbol{Z}_j$ are referred to as design matrices. Note that the elements of $\boldsymbol{X}$ and $\boldsymbol{Z}_j$ are constants that indicate whether the samples have specific conditions (e.g., environment types and marker genotypes) in the given dataset. On the other hand, the elements of $\boldsymbol{\beta}$ and $\boldsymbol{u}_j$ are variables that must be determined so that the resulting linear model fits the dataset well and the trait values are predicted with high accuracy. Using the variables above, Eq. 3.1 can be rewritten as follows:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1L} \\ 1 & x_{21} & \vdots & x_{2L} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{NL} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_L \end{pmatrix} + \sum_{j=1}^{M} \begin{pmatrix} z_{j11} & \cdots & z_{j1q_j} \\ z_{j21} & \vdots & z_{j2q_j} \\ \vdots & \vdots & \vdots \\ z_{jN1} & \cdots & z_{jNq_j} \end{pmatrix} \begin{pmatrix} u_{j1} \\ \vdots \\ u_{jq_j} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{pmatrix}. \tag{3.3}$$

3.3.2 One-Marker Model

We first consider a simple example where no environmental factors ($L = 0$) and only a single genetic marker ($M = 1$) are included as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{e}, \tag{3.4}$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)'$, $\boldsymbol{X} = (1, \ldots, 1)'$, $\boldsymbol{\beta} = (\beta_0)'$, and $\boldsymbol{Z}_1 = \begin{pmatrix} 1 & \text{or} & 0 & \cdots \\ 0 & & 1 & \end{pmatrix}'$. Here, we consider that the genetic marker has two genotypes γ_1 and γ_2 (e.g., homozygous for an allele and heterozygous for two alleles), and a sample takes one of the two genotypes at the genetic marker. Then, the ith row vector of $\boldsymbol{Z}_1$ is one of $(1, 0)$ and $(0, 1)$, respectively, if its genotype is γ_1 and γ_2. In Eq. 3.4, $\boldsymbol{u}_1 = (u_{11}, u_{12})'$ indicates the effects of those two genotypes of the genetic marker. $\boldsymbol{e} = (e_1, \ldots, e_N)'$ are the random residuals. Then Eq. 3.4 can be rewritten as follows:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (\beta_0) + \begin{pmatrix} 1 & 0 \\ \multicolumn{2}{c}{\text{or}} \\ 0 & 1 \\ \multicolumn{2}{c}{\vdots} \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}. \tag{3.5}$$

Here, $y_i = \beta_0 + u_{11} + e_i$ if its genotype is γ_1, and $y_i = \beta_0 + u_{12} + e_i$ if γ_2. Letting $a_0 = \beta_0 + u_{11}$ and $a_1 = u_{12} - u_{11}$, we have:

$$y_i = a_0 + a_1 g_{1i} + e_i, \tag{3.6}$$

where $g_{1i} = 1$ if the genotype of the genetic marker in the ith sample is γ_2, and otherwise 0 ($i = 1, \ldots, N$).

Note that the number of the genotypes is not restricted to two. If the genetic marker has a total of q genotypes, the size of the design matrix $\boldsymbol{Z}_1$ is $N \times q$, and the

tth column of its ith row vector is set to 1 if the ith sample has the tth genotype at this genetic marker and otherwise is set to 0. $\boldsymbol{u}_1$ has q elements $(u_{11}, \ldots, u_{1q})$.

3.3.3 Two-Marker Model Without Interactions

In a similar way as for the one-marker model, we can consider another example where no environmental factors ($L = 0$) and two markers ($M = 2$) are included as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{Z}_2\boldsymbol{u}_2 + \boldsymbol{e}, \tag{3.7}$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)'$, $\boldsymbol{X} = (1, \ldots, 1)'$, and $\boldsymbol{\beta} = (\beta_0)'$. The terms $\boldsymbol{Z}_1\boldsymbol{u}_1$ and $\boldsymbol{Z}_2\boldsymbol{u}_2$, respectively, represent the effects by the two genetic markers. $\boldsymbol{e}$ are the random residuals. Here, we assume that the two genetic markers are mutually independent and their effects contribute to the trait value in an additive manner. There are no epistatic effects caused by their interactions. We also assume that the genetic markers, respectively, have two genotypes (γ_{11} and γ_{12} for one genetic marker while γ_{21} and γ_{22} for the other). Then, Eq. 3.7 can be rewritten as follows:

$$\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} (\beta_0) + \begin{pmatrix} 1 \quad 0 \\ \text{or} \\ 0 \quad 1 \\ \vdots \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} + \begin{pmatrix} 1 \quad 0 \\ \text{or} \\ 0 \quad 1 \\ \vdots \end{pmatrix} \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ \vdots \\ e_N \end{pmatrix}. \tag{3.8}$$

According to the genotypes of the two genetic markers, which are either $(\gamma_{11}, \gamma_{21})$, $(\gamma_{12}, \gamma_{21})$, $(\gamma_{11}, \gamma_{22})$, or $(\gamma_{12}, \gamma_{22})$, the trait value of a sample can be written as follows:

$$y_i = \beta_0 + u_{11} + u_{21} + e_i, \tag{3.9}$$

$$\begin{aligned} y_i &= \beta_0 + u_{12} + u_{21} + e_i \\ &= (\beta_0 + u_{11} + u_{21}) + (u_{12} - u_{11}) + e_i, \end{aligned} \tag{3.10}$$

$$\begin{aligned} y_i &= \beta_0 + u_{11} + u_{22} + e_i \\ &= (\beta_0 + u_{11} + u_{21}) + (u_{22} - u_{21}) + e_i, \end{aligned} \tag{3.11}$$

$$\begin{aligned} y_i &= \beta_0 + u_{12} + u_{22} + e_i \\ &= (\beta_0 + u_{11} + u_{21}) + (u_{12} - u_{11}) + (u_{22} - u_{21}) + e_i. \end{aligned} \tag{3.12}$$

Letting $a_0 = \beta_0 + u_{11} + u_{21}$, $a_1 = u_{12} - u_{11}$, and $a_2 = u_{22} - u_{21}$, we have

$$y_i = a_0 + a_1 g_{1i} + a_2 g_{2i} + e_i, \tag{3.13}$$

where $g_{ji} = 1$ if the genotype of the ith sample is γ_{j2}, and otherwise 0 ($j = 1$ or 2, $i = 1, \ldots, N$).

3.3.4 Two-Marker Model with Interactions

We can include the effects by the interactions among the markers, by adding a term to Eq. 3.7 as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{Z}_2\boldsymbol{u}_2 + \boldsymbol{W\rho} + \boldsymbol{e}, \tag{3.14}$$

where $\boldsymbol{W}$ is the design matrix for the interactions among the pairs of the genetic markers and $\boldsymbol{\rho}$ is the vector of the effects by these interactions. Each column of $\boldsymbol{W}$ corresponds to a pair of genotypes of the two genetic markers. When the two genetic markers have, respectively, two genotypes (γ_{11}/γ_{12} and γ_{21}/γ_{22}) as in Eq. 3.7, an example of $\boldsymbol{W\rho}$ can be written as follows:

$$\boldsymbol{W\rho} = \begin{pmatrix} (\gamma_{11} \times \gamma_{21})_1 & (\gamma_{11} \times \gamma_{22})_1 & (\gamma_{12} \times \gamma_{21})_1 & (\gamma_{12} \times \gamma_{22})_1 \\ (\gamma_{11} \times \gamma_{21})_2 & (\gamma_{11} \times \gamma_{22})_2 & (\gamma_{12} \times \gamma_{21})_2 & (\gamma_{12} \times \gamma_{22})_2 \\ \vdots & \vdots & \vdots & \vdots \\ (\gamma_{11} \times \gamma_{21})_N & (\gamma_{11} \times \gamma_{22})_N & (\gamma_{12} \times \gamma_{21})_N & (\gamma_{12} \times \gamma_{22})_N \end{pmatrix} \begin{pmatrix} \rho_{11} \\ \rho_{12} \\ \rho_{21} \\ \rho_{22} \end{pmatrix}. \tag{3.15}$$

Here, the columns of $\boldsymbol{W}$ correspond to the combinations of the genotypes of the genetic markers. In this matrix, $(\gamma_{1t_1} \times \gamma_{2t_2})_i$, which is for the ith sample, is 1 if the first genetic marker has the t_1th genotype and also the second genetic marker has the t_2th genotype, and otherwise it is 0 (t_1, $t_2 = 1$ or 2 in this example). The elements of $\boldsymbol{\rho} = (\rho_{11}, \rho_{12}, \rho_{21}, \rho_{22})'$ are the effects by the combinations of the genetic markers. In a similar way, we can arbitrarily include the additional terms in Eq. 3.7; for example, we can include the combinations between an environmental factor and a genetic factor, in addition to the combinations among genetic factors as given by Eq. 3.14. Both the terms for the genetic factors and the terms for their interactions are given by the products of the design matrix and the vector of the effects. Therefore, they cannot be distinguished from each other merely by using the numerical dataset described in the model: some interpretations specific to the context are required for the modeling. In this sense, even with the interactions among environmental and genetic factors, the values of the target trait can be described using the linear mixed model given by Eq. 3.1.

3.4 Decomposition of Variance and Contributions of Genetic Markers

The linear model given by Eq. 3.1 is rewritten as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \sum\nolimits_{j=1}^{M} \boldsymbol{Z}_j\boldsymbol{u}_j + \boldsymbol{e} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{Z}_2\boldsymbol{u}_2 + \cdots + \boldsymbol{Z}_M\boldsymbol{u}_M + \boldsymbol{e}. \tag{3.16}$$

For the jth genetic marker, $\boldsymbol{Z}_j\boldsymbol{u}_j$ is a vector of the genetic values of the samples. Letting δ_{ji} be the value associated with the ith sample, we have:

$$\boldsymbol{Z}_j\boldsymbol{u}_j = \begin{pmatrix} 1\,\text{or}\,0 & \cdots & \cdots & 1\,\text{or}\,0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1\,\text{or}\,0 & \cdots & \cdots & 1\,\text{or}\,0 \end{pmatrix} \begin{pmatrix} u_{j1} \\ u_{j2} \\ \vdots \\ u_{jq_j} \end{pmatrix} = \begin{pmatrix} \delta_{j1} \\ \delta_{j2} \\ \vdots \\ \delta_{jN} \end{pmatrix}. \tag{3.17}$$

Here, only one element of the row vector of $\boldsymbol{Z}_j$ is 1 (the rest of the elements are 0) because a sample can have exclusively one genotype at the jth genetic marker. Missing data is not considered. If all the pairs of the terms are not correlated (independent and additive), i.e., their covariance is zero, the total variance $V(\boldsymbol{y})$ can be given by the summation of their variances:

$$V(\boldsymbol{y}) = V(\boldsymbol{X\beta}) + V(\boldsymbol{Z}_1\boldsymbol{u}_1) + V(\boldsymbol{Z}_2\boldsymbol{u}_2) + \cdots + V(\boldsymbol{Z}_M\boldsymbol{u}_M) + V(\boldsymbol{e}). \tag{3.18}$$

For the jth genetic marker, $V(\boldsymbol{Z}_j\boldsymbol{u}_j)$ is the variance of the values, $\delta_{j1}, \delta_{j2}, \cdots, \delta_{jN}$ in Eq. 3.17. $V(\boldsymbol{X\beta}) = 0$, if there are no environmental effects. Since the total variance is obtained by $V(\boldsymbol{y}) = \sum\nolimits_{i=1}^{N} (y_i - \mu)^2/N$ (μ is the mean of y_i), the contribution of the jth genetic marker is evaluated by

$$r_j = V\left(\boldsymbol{Z}_j\boldsymbol{u}_j\right)/V(\boldsymbol{y}). \tag{3.19}$$

This ratio shows the proportion of the variance explained by the jth genetic marker. Note again that we are assuming the markers are not correlated and there are no genetic interactions, i.e., linkage disequilibrium and epistatic effects are not assumed.

Since the denominator of Eq. 3.19 is constant in a set of samples, the ratio r_j is determined by the product of $\boldsymbol{Z}_j$ and $\boldsymbol{u}_j$. As shown in Eq. 3.17, the former represents the distribution of the genotypes in the samples, while the latter represents the effects of the genotypes. Therefore, if a portion of the elements of $\boldsymbol{u}_j$ are diverse and distributed in the samples, $V(\boldsymbol{Z}_j\boldsymbol{u}_j)$ increases and the ratio has a higher value. On the other hand, if all the row vectors of $\boldsymbol{Z}_j$ are identical or all the elements of $\boldsymbol{u}_j$ are identical, the ratio is 0, showing that there are no differences among the samples in relation to the genetic marker.

Although the ratio of the variances by Eq. 3.19 provides an index for the estimation of the effects of a genetic marker, it can fail to detect a genotype with a strong effect but with a low frequency in the samples. To remedy this problem, the variance of the elements of $\boldsymbol{u}_j = \left(u_{j1}, u_{j2}, \ldots, u_{jq_j}\right)'$ is evaluated as follows:

$$v_j = \sum\nolimits_{k=1}^{q_j} \left(u_{jk} - u_{j\mu}\right)^2 / q_j, \tag{3.20}$$

where $u_{j\mu} = \sum\nolimits_{k=1}^{q_j} u_{jk}/q_j$. Once $\boldsymbol{u}_j$ is obtained, v_j is not dependent on the samples, while r_j is dependent on the distribution of the genotypes in the samples.

Even with a low frequency in the samples, a genotype highly correlated to the trait of interest is important, especially in the selection process for the breeding purposes. As described in the later sections, the variance given by Eq. 3.20 is one of the key values for elucidating the factors that contributed to the trait.

3.5 Breeding Values and Heritability

To focus on the effects that are inherited from one generation to the next, a trait value P (also referred to as a phenotype or a phenotypic value) of a sample can be decomposed in a symbolic manner as follows:

$$P = G + E, \tag{3.21}$$

where G (representing a genetic value or a genotypic value) is the portion that is determined by the genetic factors, while E (representing an environmental value or environmental deviation) is the portion that is not explained by the genetic factors and is attributed to the nongenetic factors (i.e., environmental factors). Here, P is the observed value in a sample, and G for the sample is the expected value of the trait in samples with the same genetic background as the sample. E is the random residual. The deviation from the mean of the trait values in the samples can be used for the actual value of P for a sample. The deviation is partitioned into the effects of the genetic factors and the environmental factors. The portion by the genetic factors G is further divided as follows:

$$G = A + D + I, \tag{3.22}$$

where A shows the deviations by the additive genetic effects, D shows the deviations by the dominant genetic effects, and I shows the deviations by the effects of interactions among genetic markers, also referred to as epistatic effects. Here, A captures the effects of alleles in an additive manner. D captures the nonlinear effects (i.e., deviations from the linear effects) caused by the combinations of alleles (heterozygous genotypes) within a genetic marker, while I captures the nonlinear effects (i.e., deviations from the linear effects) by the combinations of alleles or genotypes among multiple genetic markers. D and I, respectively, correspond to the deviations by the intra- and inter-marker interactions among alleles. Here, interactions among G and E are ignored. From Eqs. 3.21 and 3.22, we have:

$$P = G + E = (A + D + I) + E. \tag{3.23}$$

If there are no interactions among the terms of Eq. 3.23, the variance can be decomposed as follows:

$$V(P) = V(G) + V(E) = V(A) + V(D) + V(I) + V(E). \tag{3.24}$$

In the traditional GS, we usually consider that only the additive effect A is transmitted to the progenies in the succeeding generations and only this effect is referred to as a breeding value (BV). The pairs of alleles of the heterozygous genotypes (dominant effects) and the combinatorial patterns of alleles among the genetic markers (epistatic effects) are not entirely passed on to the offspring, and therefore the effects that are expected to be the same in the progeny, i.e., the additive effects, are emphasized.

The ratios of the variance of G (genotypic value) and A (additive genotypic value) against the variance of P in samples are, respectively, referred to as the broad-sense heritability and the narrow-sense heritability. The broad-sense heritability is defined as follows:

$$H^2 = V(G)/V(P). \tag{3.25}$$

The narrow-sense heritability is defined as follows:

$$h^2 = V(A)/V(P). \tag{3.26}$$

The genetic value of a sample is defined by the phenotypic value that can be attributed to the genetic factors including nonadditive allele interactions. On the other hand, the breeding value of a sample is defined by the phenotypic value that can be attributed to only the additive genetic factors.

In this way, the GS is usually based on the effects of additive genotypic values. One reason why such a strategy has been adopted might be that the decomposition of phenotypic variance has been considered in a symbolic and abstract manner as in Eq. 3.24, where the entity of the genetic factors is actually ambiguous. However, the datasets obtained by the high-throughput sequencers (also known as the next-generation sequencers (NGS)) and high-density microarrays provide the genotypes of the samples at single base-pair resolution across their whole-genome sequences. Consequently, such high-resolution data make it possible to evaluate the genotypic values based on the actual variations of DNA sequences, e.g., single nucleotide variations (SNVs), short insertion and deletions (Indels), and structural variations such as copy number variations (CNVs) instead of putative genetic factors. Especially in relation to the SNVs, for example, the types of the alleles of a genetic marker are restricted to those that are represented symbolically by their nucleobases (A, C, G, and T), and therefore the dominant effect by a heterozygous genotype at a position in the DNA sequence will be expected to found again in the progeny. Actually, such point mutations may alter the characteristics of the proteins that they correspond to. If the trait is controlled by the point mutations, the problem will be simple. The inter-marker interactions among plural genetic markers may have the same characteristics because the probability that we will find the same genotype patterns in the progeny is not zero even if it is low.

A genetic marker represents a chromosomal region which it belongs to, and its genotype partially shows the way the region has been inherited from the ancestors. However, even if the identical genotype at a genetic marker is found in the samples, the genetic factors which were near to the genetic marker are not always the same in those samples. It depends on how they are genetically related as determined by the manner of crossing. The identical allele constituting a genotype at a genetic marker can be found in the samples when it has been inherited from the independent ancestors having it or it has been caused by spontaneous mutations (identity by state (IBS)). The identical allele can be found in the samples also when it has been inherited from the common ancestors (identity by descent (IBD)). If the founder of the chromosomal regions can be traced back in the ancestors, detection of the dominant effects by the accompanying genetic factors can be expected even in their offspring for such regions. An increase in the number of the genetic markers makes it possible to capture the genetic factors by the patterns of their genotypes (haplotypes) instead of the genotype of a single genetic marker; however, improvement of the resolution introduces nonadditive effects into the datasets. Linkage disequilibrium (LD) among the genetic markers in a chromosome (cis-interactions) must be taken into consideration in addition to their genetic relationships across chromosomes (trans-interactions).

3.6 Determination of Model Parameters by Least-Squares Estimation

The least-squares estimation approach can analytically determine the parameters in the models. The random residuals, i.e., the deviations of the predicted values from the observed real values, are minimized in the samples. The resulting models identify the underlying causal factors in the given datasets and characterize the target values by using those factors. We consider, for example, a one-marker model by Eq. 3.6 for a dataset with $N=2$ samples and $M=1$ genetic marker. When the target values and the 0/1-encoded genotypes are given, respectively, by $\boldsymbol{y}=(y_1, y_2)'=(2, 1)'$ and $\boldsymbol{t}_1=(g_{11}, g_{12})'=(1, 0)'$, the relationships among the parameters in the model are given as follows:

$$\begin{array}{l} 2 = a_0 + a_1 \times 1 + e_1 \\ 1 = a_0 + a_1 \times 0 + e_2 \end{array}. \tag{3.27}$$

Letting the random residuals $\boldsymbol{e} = (e_1, e_2)'$ be $\mathbf{0}$, the equations can be uniquely solved $(a_0, a_1) = (1, 1)$. However, if an additional sample is given $(y_3, g_{13}) = (2.1, 1)$, for example, the equations cannot be solved under the assumption that the random residuals are $\mathbf{0}$. Some portion of the target values must be assigned to the random residuals under the restriction that they are minimized, 0.1 from y_3 to e_3 in this case. If any genetic markers in a model are not associated with the target values, by setting all the parameters except the random residuals to $\mathbf{0}$ and assigning all the values to the random residuals, we have a trivial solution $\boldsymbol{y} = \boldsymbol{e}$. Thus, the problem can be solved by minimization of the random residuals.

3.6.1 Least-Squares Estimation for One-Marker Model

From Eq. 3.6, we have $e_i = y_i - a_0 - a_1 g_{1i}$. Letting J denote the squared error as follows, a_0 and a_1 are determined so that J is minimized:

$$J = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - a_0 - a_1 g_{1i})^2. \tag{3.28}$$

Letting $\partial J / \partial a_0 = -2 \sum_{i=1}^{N} (y_i - a_0 - a_1 g_{1i}) = 0$, we have:

$$a_0 = \sum_{i=1}^{N} \frac{y_i}{N} - a_1 \sum_{i=1}^{N} \frac{g_{1i}}{N}. \tag{3.29}$$

Similarly, letting $\partial J / \partial a_1 = 2 \sum_{i=1}^{N} (y_i - a_0 - a_1 g_{1i}) \partial (y_i - a_0 - a_1 g_{1i}) / \partial a_1 = 0$ and replacing a_0 with Eq. 3.29, we have:

$$\frac{\partial J}{\partial a_1} = 2 \sum_{i=1}^{N} \left\{ y_i - \left(\sum_{i=1}^{N} \frac{y_i}{N} - a_1 \sum_{i=1}^{N} \frac{g_{1i}}{N} \right) - g_{1i} a_1 \right\} \left\{ - \left(g_{1i} - \sum_{i=1}^{N} \frac{g_{1i}}{N} \right) \right\} \\ = 0. \tag{3.30}$$

Manipulating Eq. 3.30, we have:

$$\frac{\partial J}{\partial a_1} = -2\left\{\sum_{i=1}^{N}\left(y_i - \sum_{i=1}^{N}\frac{y_i}{N}\right)\left(g_{1i} - \sum_{i=1}^{N}\frac{g_{1i}}{N}\right) - a_1\sum_{i=1}^{N}\left(g_{1i} - \sum_{i=1}^{N}\frac{g_{1i}}{N}\right)^2\right\} = 0. \tag{3.31}$$

Equation 3.31 can be written as $\partial J/\partial a_1 = -2N(\sigma_{y1} - a_1\sigma_{11}) = 0$. Here, σ_{y1} is the covariance between $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)'$ and $\boldsymbol{t}_1 = (g_{11}, g_{12}, \ldots, g_{1N})'$. σ_{11} is the variance of $\boldsymbol{t}_1$. If $\sigma_{11} \neq 0$, we have:

$$a_1 = \frac{\sigma_{y1}}{\sigma_{11}} = \frac{\sum_{i=1}^{N}\left\{\left(y_i - \sum_{i=1}^{N}\frac{y_i}{N}\right)\left(t_{1i} - \sum_{i=1}^{N}\frac{g_{1i}}{N}\right)\right\}/N}{\sum_{i=1}^{N}\left(g_{1i} - \sum_{i=1}^{N}\frac{g_{1i}}{N}\right)^2/N}. \tag{3.32}$$

The prediction value for the ith sample is given as follows:

$$p_i = a_0 + a_1 g_{1i}. \tag{3.33}$$

3.6.2 Least-Squares Estimation for Two-Marker Model Without Interactions

From Eq. 3.13, we have $e_i = y_i - a_0 - a_1 g_{1i} - a_2 g_{2i}$. In a similar way as for the one-marker model, the squared error J is given as follows and minimized:

$$J = \sum_{i=1}^{N}(y_i - a_0 - a_1 g_{1i} - a_2 g_{2i})^2. \tag{3.34}$$

Letting $\partial J/\partial a_0 = -2\sum_{i=1}^{N}(y_i - a_0 - a_1 g_{1i} - a_2 g_{2i}) = 0$, we have:

$$a_0 = \sum_{i=1}^{N}\frac{y_i}{N} - a_1\sum_{i=1}^{N}\frac{g_{1i}}{N} - a_2\sum_{i=1}^{N}\frac{g_{2i}}{N}. \tag{3.35}$$

Replacing a_0 with Eq. 3.35 in Eq. 3.13, we have:

$$\begin{aligned} y_i &= \sum_{k=1}^{N}\frac{y_k}{N} - a_1\sum_{k=1}^{N}\frac{g_{1k}}{N} - a_2\sum_{k=1}^{N}\frac{g_{2k}}{N} + a_1 g_{1i} + a_2 g_{2i} + e_i \\ &= \sum_{k=1}^{N}\frac{y_k}{N} + a_1\left(g_{1i} - \sum_{k=1}^{N}\frac{g_{1k}}{N}\right) + a_2\left(g_{2i} - \sum_{k=1}^{N}\frac{g_{2k}}{N}\right) + e_i. \end{aligned} \tag{3.36}$$

Using Eq. 3.36, Eq. 3.34 can be rewritten as follows:

$$J = \sum_{i=1}^{N} \left\{ y_i - \sum_{k=1}^{N} \frac{y_k}{N} - a_1\left(g_{1i} - \sum_{k=1}^{N} \frac{g_{1k}}{N}\right) - a_2\left(g_{2i} - \sum_{k=1}^{N} \frac{g_{2k}}{N}\right) \right\}^2. \tag{3.37}$$

Then, we have:

$$\begin{aligned}\frac{\partial J}{\partial a_1} &= -2\sum_{i=1}^{N} \left\{ \left(y_i - \sum_{k=1}^{N} \frac{y_k}{N} - a_1\left(g_{1i} - \sum_{k=1}^{N} \frac{g_{1k}}{N}\right) - a_2\left(g_{2i} - \sum_{k=1}^{N} \frac{g_{2k}}{N}\right)\right) \right. \\ &\quad \left. \times \left(g_{1i} - \sum_{k=1}^{N} \frac{g_{1k}}{N}\right) \right\} \\ &= -2\left\{ \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{N} \frac{y_k}{N}\right)\left(g_{1i} - \sum_{k=1}^{N} \frac{g_{1k}}{N}\right) \right. \\ &\quad - a_1 \sum_{i=1}^{N} \left(g_{1i} - \sum_{k=1}^{N} \frac{g_{1k}}{N}\right)^2 \\ &\quad \left. - a_2 \sum_{i=1}^{N} \left(g_{2i} - \sum_{k=1}^{N} \frac{g_{2k}}{N}\right)\left(g_{1i} - \sum_{k=1}^{N} \frac{g_{1k}}{N}\right) \right\}.\end{aligned} \tag{3.38}$$

Therefore, we also have:

$$\frac{\partial J}{\partial a_1} = -2N\left(\sigma_{y1} - a_1\sigma_{11} - a_2\sigma_{12}\right) \tag{3.39}$$

and

$$\frac{\partial J}{\partial a_2} = -2N\left(\sigma_{y2} - a_1\sigma_{12} - a_2\sigma_{22}\right). \tag{3.40}$$

Here, σ_{y1} is the covariance between $\boldsymbol{y}$ and $\boldsymbol{t}_1 = (g_{11}, g_{12}, \ldots, g_{1N})'$, σ_{y2} is the covariance between $\boldsymbol{y}$ and $\boldsymbol{t}_2 = (g_{21}, g_{22}, \ldots, g_{2N})'$, and σ_{11} and σ_{22} are the variance of $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$. σ_{12} and σ_{21} are the covariance between $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ ($\sigma_{12} = \sigma_{21}$). Letting $\partial J/\partial a_1 = 0$ and $\partial J/\partial a_2 = 0$, we have:

$$\left\{ \begin{array}{lcl} a_1\sigma_{11} + a_2\sigma_{12} & = & \sigma_{y1} \\ a_1\sigma_{21} + a_2\sigma_{22} & = & \sigma_{y2} \end{array} \right. . \tag{3.41}$$

If $\sigma_{11}\sigma_{22} - \sigma_{12}^2 \neq 0$, we have:

$$a_1 = \frac{\sigma_{y1}\sigma_{22} - \sigma_{y2}\sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \text{ and } a_2 = \frac{\sigma_{y2}\sigma_{11} - \sigma_{y1}\sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}. \tag{3.42}$$

Equation 3.41 can also be described using matrices as follows:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sigma_{y1} \\ \sigma_{y2} \end{pmatrix}. \tag{3.43}$$

Letting $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ and its determinant $\det\boldsymbol{\Sigma} = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}$, if $\det\boldsymbol{\Sigma} \neq 0$, there exists the inverse of $\boldsymbol{\Sigma}$, given by $\boldsymbol{\Sigma}^{-1} = \frac{1}{\det\boldsymbol{\Sigma}} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{pmatrix}$. Then we have:

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \sigma_{y1} \\ \sigma_{y2} \end{pmatrix} = \frac{1}{\det\boldsymbol{\Sigma}} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{pmatrix} \begin{pmatrix} \sigma_{y1} \\ \sigma_{y2} \end{pmatrix}. \tag{3.44}$$

Equation 3.44 is equivalent to Eq. 3.42.

3.6.3 Number of Parameters

As shown in the examples in the previous sections, the number of parameters that must be determined increases with the increase in the number of the genetic markers in the model. Intuitively, N independent equations are required to uniquely determine N parameters. In the typical datasets we use for the GS and the MAS, the number of the genetic markers p is much greater than that of the samples N; therefore, they cannot provide a sufficient number of equations for the parameters that are associated with the genetic markers. We thus cannot uniquely determine the parameters. This situation is sometimes referred to as the $N \ll p$ problem.

Let $f(\boldsymbol{g}_i) = a_0 + a_1 g_{1i} + \ldots + a_M g_{Mi}$ be the formula for the prediction model. Letting $\boldsymbol{a} = (a_0, a_1, \ldots, a_M)'$ be the vector of the coefficients and $\boldsymbol{g}_i = (g_0, g_{1i}, \ldots, g_{Mi})$ be the vector of the observed values (e.g., 0/1-encoded genotypes) for the ith sample ($f(\boldsymbol{g}_i) = \boldsymbol{g}_i\boldsymbol{a}$) with a dummy variable $g_0 \equiv 1$, we have N predictions for N samples. Letting $\boldsymbol{G} = (\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_N)'$, those predictions are written as $\boldsymbol{Ga}$. Then, the squared error J is given as follows:

$$J = (\boldsymbol{Ga} - \boldsymbol{y})'(\boldsymbol{Ga} - \boldsymbol{y}). \tag{3.45}$$

Equations 3.28 and 3.34 in the previous sections can be written in this format. Here, $\boldsymbol{y}'\boldsymbol{Ga} = (\boldsymbol{y}'\boldsymbol{Ga})' = \boldsymbol{a}'\boldsymbol{G}'\boldsymbol{y}$ because it is a scalar, and we then have:

$$J = (\boldsymbol{a}'\boldsymbol{G}' - \boldsymbol{y}')(\boldsymbol{Ga} - \boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{G}'\boldsymbol{Ga} - 2\boldsymbol{a}'\boldsymbol{G}'\boldsymbol{y} + \boldsymbol{y}'\boldsymbol{y}. \tag{3.46}$$

The derivative of J with respect to $\boldsymbol{a}$ is given as follows:

$$\partial J/\partial \boldsymbol{a} = 2\boldsymbol{G}'\boldsymbol{Ga} - 2\boldsymbol{G}'\boldsymbol{y}. \tag{3.47}$$

From Eq. 3.47, we have:

$$\boldsymbol{G}'\boldsymbol{G}\boldsymbol{a} = \boldsymbol{G}'\boldsymbol{y}. \tag{3.48}$$

This equation is referred to as the normal equation. When $\boldsymbol{G}'\boldsymbol{G}$ is a regular matrix, there exists its inverse and $\boldsymbol{a}$ is uniquely determined in an analytical manner as follows:

$$\boldsymbol{a} = (\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{y}. \tag{3.49}$$

However, as noted above, the number of the parameters is greater than that of the equations given by $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{a}$, and hence the parameters cannot be uniquely determined. Instead of solving the equations in an analytical manner, therefore, methods based on model fitting to the data are used so that the squared errors are minimized. Although the number of the parameters is greater than that required to solve the equations, all the parameters do not contribute to the values of the dependent variable, $\boldsymbol{y}$. Therefore, in parallel with determining the values of the parameters, the selection of the parameters is explicitly and implicitly carried out. As a result, the parameters for the independent variables that do not contribute to the dependent variable will converge to zero and be excluded from the prediction model.

3.7 Determination of Model Parameters by Bayesian Estimation

The Bayesian approach can determine the parameters in the models. As we have seen in the previous sections, the random residuals associated with the prediction model can be assumed to follow some distribution, e.g., the normal distribution. The characteristics of the distribution of the random residuals such as the mean and the variance are important for evaluation of model fitting to a given dataset rather than the values assigned to the samples. The random residuals can be thus evaluated by using the distribution of a random variable. In a similar way, if we represent the parameters in the prediction model by using the distributions of random variables, we can intuitively introduce the Bayesian approach into the construction of the prediction model. The ranges with high probability can estimate the likely values for the parameters in a stochastic manner. A part of this section in relation to the Gibbs sampling is based on Wang et al. (1993).

3.7.1 Basic Concepts of Bayes' Theorem

The joint probability can be written as follows:

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A). \tag{3.50}$$

Here, $P(A|B)$ and $P(B|A)$ are the conditional probabilities of A and B, respectively, given B and A. $P(A)$ and $P(B)$ are the marginal probabilities. If $P(B) \neq 0$, we have:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \propto P(B|A)P(A). \tag{3.51}$$

This relationship is referred to as the Bayes' theorem and shows that the probability of A given B is proportional to the product of the probability of A and the probability of B given A. Using this relationship, we can estimate the posterior probability $P(A|B)$ from the prior probability $P(A)$. The conditional probability $P(B|A)$ shows the probability of obtaining B under the condition A, which is referred to as the likelihood.

In Eq. 3.50, the probabilities are assigned to discrete events represented by A and B. This concept can be extended to the continuous distributions of a parameter θ as follows:

$$\pi(\theta|D) \propto f(D|\theta)\pi(\theta). \tag{3.52}$$

Here, a parameter θ is expressed by using a distribution of probability instead of a single value. The value of θ with a high probability is likely to be the true value of θ. In Eq. 3.52, the product of the prior distribution $\pi(\theta)$ and the likelihood of obtaining the dataset D under the condition θ, i.e., $f(D|\theta)$, provide the evaluation for the posterior distribution $\pi(\theta|D)$.

3.7.2 Estimation of Parameters of Normal Distributions

Here, we consider the case that the dataset D consists of N observations of x_i $(i = 1, 2, \ldots, N)$. We also assume the condition that the distribution of x_i follows a normal distribution with the mean μ and the variance σ^2. Here, μ and σ^2 are not known, and they are the targets of the prediction. If the N observations are independent of each other, the likelihood of obtaining the dataset D is given by the product of the probability density functions as follows:

$$f(D|\mu,\sigma^2) = \prod\nolimits_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right). \tag{3.53}$$

If we set values for μ and σ^2, the probability of the event that we obtained the dataset D is given by $f(D|\mu, \sigma^2)$. This value is referred to as the likelihood under the condition that the values follow the normal distribution. The plausible values for μ and σ^2 can be determined so that the likelihood is maximized.

3.7.2.1 Estimation of Mean with Known Variance

Suppose we are estimating μ under the condition that σ^2 is known. As the prior probability distribution for μ, we can use a normal distribution with a mean μ_0 and a variance σ_0^2 as follows:

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right). \tag{3.54}$$

If σ_0^2 is a very large value, the distribution is near to a flat distribution, showing that there is no prior knowledge. Then, we have:

$$\pi(\mu|D) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right). \tag{3.55}$$

Manipulating Eq. 3.55 by using $(x_i - \mu)^2 = (x_i - \bar{x} + \bar{x} - \mu)^2$ where $\bar{x} = \sum_{i=1}^{N} x_i$, we have:

$$\pi(\mu|D) \propto \exp\left(-\left(\mu - \frac{N\sigma_0^2\, \bar{x} + \mu_0\sigma^2}{N\sigma_0^2 + \sigma^2} \right)^2 \Big/ \left(2\frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2} \right) \right). \tag{3.56}$$

Letting

$$\mu_1 = \frac{N\sigma_0^2\bar{x} + \mu_0\sigma^2}{N\sigma_0^2 + \sigma^2} \text{ and } \sigma_1^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}, \tag{3.57}$$

we have:

$$\pi(\mu|D) \propto \exp\left(-\frac{(\mu - \mu_1)^2}{2\sigma_1^2} \right). \tag{3.58}$$

This shows that the posterior probability distribution is again a normal distribution and has its maximum value at μ_1. Therefore, μ_1 is adopted for the estimated value of μ. This method is referred to as the maximum a posteriori probability (MAP) estimation. Letting $\tau = 1/\sigma^2$ and $\tau_0 = 1/\sigma_0^2$ in Eq. 3.57, we have:

$$\tau_1 = \frac{1}{\sigma_1^2} = \frac{1}{\sigma^2}N + \frac{1}{\sigma_0^2} = \tau N + \tau_0 \text{ and } \mu_1 = \frac{\tau N}{\tau N + \tau_0}\bar{x} + \frac{\tau_0}{\tau N + \tau_0}\mu_0. \tag{3.59}$$

τ is the inverse of the variance and is referred to as the precision. Equation 3.59 shows that the precision is improved by the summation of the precision of the samples observed ($\tau N > 0$) while the mean is updated to the weighted average of the mean in the samples observed ($\bar{x}$) and that of the prior probability distribution (μ_0).

3.7.2.2 Estimation of Variance with Known Mean

On the other hand, suppose we are estimating σ^2 ($\sigma^2 \neq 0$) under the condition that μ is known. From Eq. 3.53, we have:

$$f\left(D|\sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^{N/2}} \exp\left(-\frac{\sum_{i=1}^{N}\left(x_i-\mu\right)^2}{2\sigma^2}\right). \tag{3.60}$$

As the prior distribution for σ^2, we can use a distribution with parameters ν_0 (the number of chi-squared degrees of freedom) and s_0^2 (the scaling parameter) as follows:

$$\pi\left(\sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^{1+\nu_0/2}} \exp\left(-\frac{\nu s_0^2}{2\sigma^2}\right). \tag{3.61}$$

Then the posterior probability distribution is as follows:

$$\pi\left(\sigma^2|D\right) \propto \frac{1}{\left(\sigma^2\right)^{N/2}} \exp\left(-\frac{\sum_{i=1}^{N}\left(x_i-\mu\right)^2}{2\sigma^2}\right) \times \frac{1}{\left(\sigma^2\right)^{1+\nu/2}} \exp\left(-\frac{\nu_0 s_0^2}{2}\right). \tag{3.62}$$

Manipulating Eq. 3.62, we have:

$$\pi(\sigma^2|D) \propto \frac{1}{\left(\sigma^2\right)^{1+(\nu_0+N)/2}} \exp\left(-\frac{\left(\nu_0+N\right)\frac{\nu_0 s_0^2+\sum_{i=1}^{N}\left(x_i-\mu\right)^2}{\nu_0+N}}{2\sigma^2}\right). \tag{3.63}$$

Letting

$$\nu_1 = \nu_0 + N \text{ and } s_1^2 = \frac{\nu_0 s_0^2 + \sum_{i=1}^{N}\left(x_i-\mu\right)^2}{\nu_0+N}, \tag{3.64}$$

we have:

$$\pi\left(\sigma^2|D\right) \propto \frac{1}{\left(\sigma^2\right)^{1+\nu_1/2}} \exp\left(-\frac{\nu_1 s_1^2}{2\sigma^2}\right). \tag{3.65}$$

This shows that the posterior probability is the same distribution as the prior probability distribution. This distribution is referred to as the scaled inverse chi-squared distribution, and the probability density function is given as follows:

$$f(x;\nu,s^2) = \frac{(s^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \frac{\exp\left(-\frac{\nu s^2}{2x}\right)}{x^{1+\nu/2}}. \tag{3.66}$$

Here, $\Gamma(x)$ is the gamma function and defined for positive real numbers as follows:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt, (x > 0), \tag{3.67}$$

where $\Gamma(1) = 1$ and $\Gamma(x+1) = x\Gamma(x)$ for positive real numbers. $\Gamma(x)$ can be considered to be the extension of the factorial of natural numbers, $f(n) = n!$.

The scaled inverse chi-squared distribution is derived from the normal distribution. When the values of the random variable x are independent and follow $N(0, s^2)$, summations of the squares of the values of the random variable, $z^2 = \sum_{i=1}^{\nu} x_i^2$, follow the chi-squared distribution with ν degrees of freedom. The probability density function of the chi-squared distribution ($x \geq 0$) is given as follows:

$$f(x;\nu) = \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)} \frac{\exp\left(-\frac{x}{2}\right)}{x^{1-\nu/2}}. \tag{3.68}$$

Letting $y = \nu s^2/x$ (i.e., $x = \nu s^2/y$), the probability density function of y is given as follows:

$$g(y;\nu,s^2) = f(x;\nu)\left|\frac{dx}{dy}\right| = \frac{(s^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \frac{\exp\left(\frac{-\nu s^2}{2y}\right)}{y^{1+\nu/2}}. \tag{3.69}$$

This distribution is referred to as the scaled inverse chi-squared distribution. Here, s^2 is referred to as a scaling parameter. This shows that the values of $\nu s^2/z^2$ follow the scaled inverse chi-squared distribution. By $\nu s^2/z^2 = 1/\sum_{i=1}^{\nu} \frac{(x_i/s)^2}{\nu}$, this implies that if the values of x follow $N(0, s^2)$, the inverses of the variances of the values of x follow the scaled inverse chi-squared distribution, Scaled-inv-$\chi^2(\nu, s^2)$. Letting $s^2 = \sum_{i=1}^{N} (x_i - \mu)^2$ in Eq. 3.64, we have:

$$\nu_1 = \nu_0 + N \text{ and } s_1^2 = \frac{N}{\nu_0 + N} s^2 + \frac{\nu_0}{\nu_0 + N} s_0^2. \tag{3.70}$$

Equation 3.70 shows that the degrees of freedom are improved by the number of the samples observed ($N > 0$), while the scaling parameter is updated to the weighted average of the scaling parameter in the samples observed (s^2) and that of the prior probability distribution (s_0^2).

3.7.2.3 Natural Conjugate Prior Probability Distributions

When the posterior probability distribution (or simply the posterior) is in the same family as the corresponding prior probability distribution (or simply the prior) in relation to a given likelihood function, they are referred to as conjugate distributions. Such a prior probability distribution is referred to as a (natural) conjugate prior probability distribution. For example, the probability distribution of a normal distribution has the maximum at the mean, and the estimator can be easily obtained from the posterior probability distribution. However, this is not the case with the arbitrary distributions. Generally, the characteristics of the posterior probability distributions are not clear if the prior and posterior probability distributions are not conjugate. When the posterior probability distribution is not well known, numerical integration is required to obtain its characteristics (e.g., the mean and the variance). However, this is not always easy for the unknown distributions or complex distributions (e.g., high-dimensional distributions with multiple variables).

3.7.3 Estimation of Linear Mixed Models by Sampling

One method for obtaining the characteristics of the complex posterior probability distributions is to approximate the distribution by sampling the distribution. This sampling procedure is also sometimes referred to as simulation. Markov chain Monte Carlo (MCMC) methods are a class of methods that simulate the distributions by iterated sampling. Monte Carlo (MC) methods are a class of methods for numerical analysis using random numbers. Markov chain Monte Carlo methods are categorized into Monte Carlo methods and are a class of algorithms that carry out the sampling of the probability distributions based on the Markov chain that has the target distribution as its equilibrium distribution. Here, a Markov chain is a stochastic process that has the Markov property (also known as the property of memorylessness), i.e., the conditional probability distribution of the future can be determined only by the current state. The Gibbs sampling method is one of the MCMC methods and is applicable when the conditional distribution of each variable (full conditional distribution) is given.

3.7.3.1 Parameters of Linear Mixed Models for Bayesian Estimation

Let us consider the linear model given by Eq. 3.1 again:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \sum\nolimits_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j + \boldsymbol{e}. \tag{3.71}$$

We obtain the parameters of this model, $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \sigma_e^2\right)$, by Bayes' estimation. Here, $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M)$ and $\boldsymbol{v} = \left(\sigma_1^2, \ldots, \sigma_M^2\right)$. σ_j^2 $(j = 1, \ldots, M)$ is the variance of $\boldsymbol{u}_j = \left(u_{j1}, \ldots, u_{jq_j}\right)$, and σ_e^2 is the variance of $\boldsymbol{e} = (e_1, \ldots, e_N)$. All the parameters are unknown. Note that σ_j^2 is the variance of the effects of the genotypes, not the variance of the genotypic values. If we obtain the estimation of $\boldsymbol{u}_j$, however, we can derive the variance of the genotypic values. In the succeeding sections, the probability distributions required for Gibbs sampling, i.e., the prior probability distributions (3.7.3.2), the joint posterior probability distributions (3.7.3.3), and the conditional probability distributions for the parameters (3.7.3.4–3.7.3.6), will be introduced in detail. Then, an implementation of Gibbs sampling using those distributions will be introduced (3.7.3.7). Finally, the results of estimation of the linear mixed model for the example given in Fig. 3.1 will be presented.

3.7.3.2 Prior Probability Distributions

Consider the case that the prior probability distributions for the parameters are as follows. For the fixed effects, we will assume a flat prior probability distribution for the naive prior probability distribution, which represents the case that we are not given enough knowledge, as follows:

$$\pi(\boldsymbol{\beta}) \propto \text{constant}. \tag{3.72}$$

When q_j is the number of the genotypes for the jth genetic marker $(j = 1, \ldots, M)$ and $\boldsymbol{u}_j$ follows a q_j-dimensional multivariate normal distribution, $\boldsymbol{u}_j \,|\, \boldsymbol{G}_j, \sigma_j^2 \sim N_{q_j}(\boldsymbol{0}, \boldsymbol{G}_j\sigma_j^2)$, where $\boldsymbol{G}_j$ is a known matrix indicating the relationships among $u_{j1}, \ldots, u_{jq_j}$ (the effects by the genotypes), the prior probability distribution is given as follows:

$$\begin{aligned} \pi\left(\boldsymbol{u}_j | \boldsymbol{G}_j, \sigma_j^2\right) &\propto \frac{1}{\left(\sqrt{\sigma_j^2}\right)^{q_j}} \exp\left(-\frac{1}{2}\boldsymbol{u}_j{}' \left(\boldsymbol{G}_j \sigma_j^2\right)^{-1} \boldsymbol{u}_j\right) \\ &\propto \frac{1}{\left(\sqrt{\sigma_j^2}\right)^{q_j}} \exp\left(-\frac{1}{2\sigma_j^2}\boldsymbol{u}_j{}' \boldsymbol{G}_j^{-1} \boldsymbol{u}_j\right). \end{aligned} \tag{3.73}$$

Letting $s_e^2 = \left(\boldsymbol{y} - \boldsymbol{X\beta} - \sum_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j\right)' \left(\boldsymbol{y} - \boldsymbol{X\beta} - \sum_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j\right)/N$ and $s_j^2 = \boldsymbol{u}_j' \boldsymbol{G}_j^{-1} \boldsymbol{u}_j / q_j$, the prior probability distribution for the variances is given using the scaled inverse chi-squared distributions as follows:

$$\pi\left(\sigma_e^2 | \nu_e, s_e^2\right) \propto \frac{1}{\left(\sigma_e^2\right)^{1+\nu_e/2}} \exp\left(-\frac{\nu_e s_e^2}{2\sigma_e^2}\right) \tag{3.74}$$

and

$$\pi\left(\sigma_j^2 | \nu_j, s_j^2\right) \propto \frac{1}{\left(\sigma_j^2\right)^{1+\nu_j/2}} \exp\left(-\frac{\nu_j s_j^2}{2\sigma_j^2}\right). \tag{3.75}$$

Letting $\nu_e = 0$ and $\nu_j = 0$ in Eqs. 3.74 and 3.75 for the naive prior probability distributions, we have:

$$\pi\left(\sigma_e^2\right) \propto \frac{1}{\sigma_e^2} \tag{3.76}$$

and

$$\pi\left(\sigma_j^2\right) \propto \frac{1}{\sigma_j^2}. \tag{3.77}$$

The joint prior density of $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \sigma_e^2\right)$ is given by the product of the density given by Eqs. 3.72, 3.73, 3.76, and 3.77 as follows:

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta}) \times \prod_{j=1}^{M} \pi\left(\boldsymbol{u}_j | \boldsymbol{G}_j, \sigma_j^2\right) \times \prod_{j=1}^{M} \pi\left(\sigma_j^2\right) \times \pi\left(\sigma_e^2\right). \tag{3.78}$$

3.7.3.3 Joint Posterior Probability Distributions

If we assume that $\boldsymbol{R} = \boldsymbol{I}$ (the identity matrix) in Eq. 3.2, i.e., the samples are independent to each other and the random residuals of all the samples have the identical variance, the likelihood associated with the N samples is given as follows:

$$\pi(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{\sigma_e^2}} \exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{y} - \boldsymbol{X\beta} - \sum_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j\right)' \left(\boldsymbol{y} - \boldsymbol{X\beta} - \sum_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j\right)\right\}. \tag{3.79}$$

By using Eqs. 3.78 and 3.79, the joint posterior probability distribution is given as follows:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{y}) &= \pi(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{v},\sigma_e^2|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})\\ &\propto \frac{1}{(\sigma_e^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{y}-\boldsymbol{X\beta}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)'\left(\boldsymbol{y}-\boldsymbol{X\beta}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\\ &\times\text{constant}\times\prod\nolimits_{j=1}^{M}\left\{\frac{1}{\left(\sigma_j^2\right)^{q_j/2}}\exp\left(-\frac{1}{2\sigma_j^2}\boldsymbol{u}_j{}'\boldsymbol{G}_j^{-1}\boldsymbol{u}_j\right)\right\}\times\prod\nolimits_{j=1}^{M}\frac{1}{\sigma_j^2}\times\frac{1}{\sigma_e^2}\\ &\propto \frac{1}{\left(\sigma_e^2\right)^{N/2+1}}\exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{y}-\boldsymbol{X\beta}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)'\left(\boldsymbol{y}-\boldsymbol{X\beta}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\\ &\times\prod\nolimits_{j=1}^{M}\left\{\frac{1}{\left(\sigma_j^2\right)^{q_j/2+1}}\exp\left(-\frac{1}{2\sigma_j^2}\boldsymbol{u}_j{}'\boldsymbol{G}_j^{-1}\boldsymbol{u}_j\right)\right\}.\end{aligned}\tag{3.80}$$

Here, $\boldsymbol{\theta}=\left(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{v},\sigma_e^2\right)$. $\boldsymbol{y}$ are the trait values observed.

3.7.3.4 Conditional Probability Distribution for $\boldsymbol{\beta}$

By setting a variable unknown under the conditions that all other variables are known, as seen in the simple examples, we obtain the conditional posterior probability distributions. Letting $\boldsymbol{u}$, $\boldsymbol{v}$, σ_e^2, and $\boldsymbol{y}$ be known, we have:

$$\pi\left(\boldsymbol{\beta}|\boldsymbol{u},\boldsymbol{v},\sigma_e^2,\boldsymbol{y}\right)\propto\exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{X\beta}-\boldsymbol{y}+\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)'\left(\boldsymbol{X\beta}-\boldsymbol{y}+\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}.\tag{3.81}$$

Note that $\boldsymbol{y}-\boldsymbol{X\beta}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j=\boldsymbol{X\beta}-\boldsymbol{y}+\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j$. Then, we have:

$$\begin{aligned}\boldsymbol{X\beta}-\boldsymbol{y}+\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j &= \boldsymbol{XX}^{-1}\left(\boldsymbol{X\beta}-\boldsymbol{y}+\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\\ &= \boldsymbol{X}\left\{\boldsymbol{\beta}-\boldsymbol{X}^{-1}\left(\boldsymbol{y}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}.\end{aligned}\tag{3.82}$$

Using $(\boldsymbol{AB})'=\boldsymbol{B}'\boldsymbol{A}'$ for matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, the argument of the exponential function in Eq. 3.81 can be rewritten as follows:

$$
\begin{aligned}
\pi\left(\boldsymbol{\beta}|\boldsymbol{u},\boldsymbol{v},\sigma_e^2,\boldsymbol{y}\right) &\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{X\beta}-\boldsymbol{y}+\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)'\left(\boldsymbol{X\beta}-\boldsymbol{y}+\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\\
&= \exp\left\{-\frac{1}{2\sigma_e^2}\left[\boldsymbol{X}\left\{\boldsymbol{\beta}-\boldsymbol{X}^{-1}\left(\boldsymbol{y}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\right]'\left[\boldsymbol{X}\left\{\boldsymbol{\beta}-\boldsymbol{X}^{-1}\left(\boldsymbol{y}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\right]\right\}\\
&= \exp\left\{-\frac{1}{2\sigma_e^2}\left\{\boldsymbol{\beta}-\boldsymbol{X}^{-1}\left(\boldsymbol{y}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}'\boldsymbol{X}'\boldsymbol{X}\left\{\boldsymbol{\beta}-\boldsymbol{X}^{-1}\left(\boldsymbol{y}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\right\}\\
&= \exp\left\{-\frac{1}{2}\left\{\boldsymbol{\beta}-\boldsymbol{X}^{-1}\left(\boldsymbol{y}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}'\frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma_e^2}\left\{\boldsymbol{\beta}-\boldsymbol{X}^{-1}\left(\boldsymbol{y}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)\right\}\right\}.
\end{aligned}
\tag{3.83}
$$

Using $(\boldsymbol{AB})^{-1}=\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ for matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have $\overline{\boldsymbol{\beta}}=\boldsymbol{X}^{-1}\left(\boldsymbol{y}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\left(\boldsymbol{y}-\sum\nolimits_{j=1}^{M}\boldsymbol{Z}_j\boldsymbol{u}_j\right)$. Then we have:

$$
\pi\left(\boldsymbol{\beta}|\boldsymbol{u},\boldsymbol{v},\sigma_e^2,\boldsymbol{y}\right) \propto \exp\left(-\frac{1}{2}\left\{\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right\}'\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left\{\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right\}\right),
$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma_e^2$ (i.e., $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}=\boldsymbol{X}'\boldsymbol{X}/\sigma_e^2$). Thus, we obtain:

$$
\boldsymbol{\beta}\mid\boldsymbol{u},\boldsymbol{v},\sigma_e^2,\boldsymbol{y}\ \sim\ N(\bar{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}}).
\tag{3.84}
$$

3.7.3.5 Conditional Probability Distribution for $\boldsymbol{u}_j$

In a similar way, we consider the case in which all variables except $\boldsymbol{u}_j$ are known $(j=1,\ldots,M)$. Letting $\boldsymbol{u}_{-j}$ be the vector obtained by removing $\boldsymbol{u}_j$ from $\boldsymbol{u}$ and $\boldsymbol{p}_j=\boldsymbol{y}-\boldsymbol{X\beta}-\sum\nolimits_{k=1,k\neq j}^{M}\boldsymbol{Z}_k\boldsymbol{u}_k$, we have:

$$
\begin{aligned}
&\pi(\boldsymbol{u}_j|\boldsymbol{\beta},\boldsymbol{u}_{-j},\boldsymbol{v},\sigma_e^2,\boldsymbol{y})\\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{Z}_j\boldsymbol{u}_j-\boldsymbol{p}_j\right)'\left(\boldsymbol{Z}_j\boldsymbol{u}_j-\boldsymbol{p}_j\right)\right\}\times\exp\left(-\frac{1}{2\sigma_j^2}\boldsymbol{u}_j{}'\boldsymbol{G}_j^{-1}\boldsymbol{u}_j\right)\\
&= \exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{u}_j-\boldsymbol{Z}_j^{-1}\boldsymbol{p}_j\right)'\boldsymbol{Z}_j'\boldsymbol{Z}_j\left(\boldsymbol{u}_j-\boldsymbol{Z}_j^{-1}\boldsymbol{p}_j\right)\right\}\times\exp\left(-\frac{1}{2\sigma_j^2}\boldsymbol{u}_j{}'\boldsymbol{G}_j^{-1}\boldsymbol{u}_j\right)\\
&= \exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{u}_j-\boldsymbol{Z}_j^{-1}\boldsymbol{p}_j\right)'\boldsymbol{Z}_j'\boldsymbol{Z}_j\left(\boldsymbol{u}_j-\boldsymbol{Z}_j^{-1}\boldsymbol{p}_j\right)-\frac{1}{2\sigma_j^2}\boldsymbol{u}_j{}'\boldsymbol{G}_j^{-1}\boldsymbol{u}_j\right\}.
\end{aligned}
\tag{3.85}
$$

Here, note that $\boldsymbol{Z}_j'\boldsymbol{Z}_j$ and $\boldsymbol{G}_j^{-1}$ are symmetric (i.e., $\boldsymbol{G}_j$ is also symmetric), and letting $\boldsymbol{Z}_j=(z_1,z_2,\cdots,z_n)$ be a vector of row vectors, the (m,n)th and (n,m)th element of $\boldsymbol{Z}_j{}'\boldsymbol{Z}_j$ are, respectively, $(\boldsymbol{Z}_j{}'\boldsymbol{Z}_j)_{mn} = z_n{}'z_m$ and $(\boldsymbol{Z}_j{}'\boldsymbol{Z}_j)_{nm} = z_m{}'z_n$. $\boldsymbol{G}_j^{-1}$ is a multiple of a covariance matrix. Here, (m, n)th element is in the mth row and in the nth column. Generally, in relation to a summation of two quadratic forms, we have:

$$\begin{aligned} &\left(\boldsymbol{u} - \bar{\boldsymbol{u}}\right)'\boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{u} - \bar{\boldsymbol{u}}\right) + \boldsymbol{u}'\boldsymbol{G}\boldsymbol{u} \\ &= \left\{\boldsymbol{u} - (\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G})^{-1}\boldsymbol{Z}'\boldsymbol{Z}\bar{\boldsymbol{u}}\right\}'(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G})\left\{\boldsymbol{u} - (\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G})^{-1}\boldsymbol{Z}'\boldsymbol{Z}\bar{\boldsymbol{u}}\right\} \\ &\quad + \bar{\boldsymbol{u}}'\left((\boldsymbol{Z}'\boldsymbol{Z})^{-1} + \boldsymbol{G}^{-1}\right)^{-1}\bar{\boldsymbol{u}}. \end{aligned} \tag{3.86}$$

Thus, we have:

$$\begin{aligned} &\pi(\boldsymbol{u}_j|\boldsymbol{\beta}, \boldsymbol{u}_{-j}, \boldsymbol{v}, \sigma_e^2, \boldsymbol{y}) \\ &= \exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{u}_j - \boldsymbol{Z}_j^{-1}\boldsymbol{p}_j\right)'\boldsymbol{Z}_j{}'\boldsymbol{Z}_j\left(\boldsymbol{u}_j - \boldsymbol{Z}_j^{-1}\boldsymbol{p}_j\right) - \frac{1}{2\sigma_j^2}\boldsymbol{u}_j{}'\boldsymbol{G}_j^{-1}\boldsymbol{u}_j\right\} \\ &= \exp\left\{-\frac{1}{2}(\boldsymbol{u}_j - \bar{\boldsymbol{u}}_j)'\left(\frac{\boldsymbol{Z}_j{}'\boldsymbol{Z}_j}{\sigma_e^2} + \frac{\boldsymbol{G}_j^{-1}}{\sigma_j^2}\right)(\boldsymbol{u}_j - \bar{\boldsymbol{u}}_j) - \frac{1}{2}\bar{\boldsymbol{u}}_j{}'\left(\left(\frac{\boldsymbol{Z}_j{}'\boldsymbol{Z}_j}{\sigma_e^2}\right)^{-1} + \left(\frac{\boldsymbol{G}_j^{-1}}{\sigma_j^2}\right)^{-1}\right)^{-1}\bar{\boldsymbol{u}}_j\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{u}_j - \bar{\boldsymbol{u}}_j)'\left(\frac{\boldsymbol{Z}_j{}'\boldsymbol{Z}_j}{\sigma_e^2} + \frac{\boldsymbol{G}_j^{-1}}{\sigma_j^2}\right)(\boldsymbol{u}_j - \bar{\boldsymbol{u}}_j)\right\}. \end{aligned} \tag{3.87}$$

Here,

$$\begin{aligned} \bar{\boldsymbol{u}}_j &= \left(\frac{\boldsymbol{Z}_j{}'\boldsymbol{Z}_j}{\sigma_e^2} + \frac{\boldsymbol{G}_j^{-1}}{\sigma_j^2}\right)^{-1}\boldsymbol{Z}_j{}'\boldsymbol{Z}_j\boldsymbol{Z}_j^{-1}\boldsymbol{p}_j \\ &= \left(\frac{\boldsymbol{Z}_j{}'\boldsymbol{Z}_j}{\sigma_e^2} + \frac{\boldsymbol{G}_j^{-1}}{\sigma_j^2}\right)^{-1}\boldsymbol{Z}_j{}'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum\nolimits_{k=1,k\neq j}^{M}\boldsymbol{Z}_k\boldsymbol{u}_k\right), \end{aligned} \tag{3.88}$$

and

$$\boldsymbol{\Sigma}_j^{-1} = \left(\frac{\boldsymbol{Z}_j{}'\boldsymbol{Z}_j}{\sigma_e^2} + \frac{\boldsymbol{G}_j^{-1}}{\sigma_j^2}\right) \text{ and } \boldsymbol{\Sigma}_j = \left(\boldsymbol{Z}_j{}'\boldsymbol{Z}_j + \boldsymbol{G}_j^{-1}\frac{\sigma_e^2}{\sigma_j^2}\right)^{-1}\sigma_e^2. \tag{3.89}$$

Thus, we have:

$$\boldsymbol{u}_j \mid \boldsymbol{\beta}, \boldsymbol{u}_{-j}, \boldsymbol{v}, \sigma_e^2, \boldsymbol{y} \;\sim\; N(\bar{\boldsymbol{u}}_j, \boldsymbol{\Sigma}_j). \tag{3.90}$$

3.7.3.6 Conditional Probability Distributions for σ_j^2 and σ_e^2

We then consider the case in which all variables except σ_j^2 are known ($j = 1, \ldots, M$). Letting $\boldsymbol{v}_{-j}$ be the vector obtained by removing $\boldsymbol{v}_j$ from $\boldsymbol{v}$, we have:

$$\pi(\sigma_j^2|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}_{-j}, \boldsymbol{y}) \propto \frac{1}{(\sigma_j^2)^{q_j/2+1}} \exp\left(-\frac{1}{2\sigma_j^2} \boldsymbol{u}_j{}' \boldsymbol{G}_j^{-1} \boldsymbol{u}_j\right). \tag{3.91}$$

Thus, we have:

$$\sigma_j^2 \mid \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{y} \quad \sim \quad \text{Scaled–inv–}\chi^2(q_j, \boldsymbol{u}_j{}' \boldsymbol{G}_j^{-1} \boldsymbol{u}_j q_j). \tag{3.92}$$

In a similar way, considering the case in which all variables except σ_e^2 are known, we have:

$$\pi(\sigma_e^2|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{y})$$
$$\propto \frac{1}{(\sigma_e^2)^{N/2+1}} \exp\left\{-\frac{1}{2\sigma_e^2}\left(\boldsymbol{y} - \boldsymbol{X\beta} - \sum\nolimits_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j\right)' \left(\boldsymbol{y} - \boldsymbol{X\beta} - \sum\nolimits_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j\right)\right\}. \tag{3.93}$$

Thus, we have:

$$\sigma_e^2 \mid \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{y} \sim \text{Scaled}-\text{inv}-\chi^2(N, s_e^2), \tag{3.94}$$

where $s_e^2 = (\boldsymbol{y} - \boldsymbol{X\beta} - \sum_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j)' (\boldsymbol{y} - \boldsymbol{X\beta} - \sum_{j=1}^{M} \boldsymbol{Z}_j \boldsymbol{u}_j)/N$

3.7.3.7 Gibbs Sampling

Based on the probability distributions in the preceding sections, we can implement a program that carries out the Gibbs sampling procedure. By Eqs. 3.84, 3.87, 3.92, and 3.94, we have the conditional probability distributions for $\boldsymbol{\beta}$, $\boldsymbol{u}_j$, σ_j^2, and σ_e^2. Setting a portion of the parameters in Eq. 3.78 as known, we obtain the multivariate normal distributions for the coefficients for the fixed and random effects ($\boldsymbol{\beta}$ and $\boldsymbol{u}_j$) and the scaled inverse chi-squared distributions for the variances of their elements (σ_j^2 and σ_e^2).

We can obtain the values by sampling (simulating) using the conditional probability distributions. The values are obtained as the random numbers generated by the given normal distribution and the scaled inverse chi-squared distributions. Starting with arbitrary initial values for the parameters (for the round 0), the following steps are iterated T rounds until the values converge ($j = 1, \ldots, M$):

1. Update $\boldsymbol{\beta}$ by Eq. 3.84 under the condition that $\boldsymbol{u}$, $\boldsymbol{v}$, and σ_e^2 are known.
2. Update $\boldsymbol{u}_j$ by Eq. 3.87 under the condition that $\boldsymbol{\beta}$, $\boldsymbol{u}_{-j}$, $\boldsymbol{v}$, and σ_e^2 are known.
3. Update σ_j^2 by Eq. 3.92 under the condition that $\boldsymbol{\beta}$, $\boldsymbol{u}$, $\boldsymbol{v}_{-j}$, and σ_e^2 are known.
4. Update σ_e^2 by Eq. 3.94 under the condition that $\boldsymbol{\beta}$, $\boldsymbol{u}$, and $\boldsymbol{v}$ are known.

Here, $\boldsymbol{u}_{-j}$ and $\boldsymbol{v}_{-j}$ are, respectively, the vector obtained by removing the jth element $\boldsymbol{u}_j$ from $\boldsymbol{u}$, and the vector obtained by removing the jth element $\boldsymbol{v}_j$ from $\boldsymbol{v}$. Accumulating the values sampled and obtaining the frequencies of those values (the histograms of the frequency distributions), the posterior probability distributions of the variables are estimated.

In the rounds above, normally distributed random numbers are obtained by using, for example, the Box-Muller method (Box and Muller 1958), which transforms uniformly distributed random numbers to normally distributed random numbers. Random numbers following the scaled inverse chi-squared distribution are obtained by using the gamma distribution. When $X \sim \text{Scaled-inv-}\chi^2(\nu, \tau^2)$, we can generate the random numbers that follow the distribution using the inverse gamma distribution and the gamma distribution as follows:

$$X \sim \text{Inv–Gamma}(\nu/2, \nu\tau^2/2) \text{ and } 1/X \sim \text{Gamma}\Big(\nu/2, 2/(\nu\tau^2)\Big) \tag{3.95}$$

The random numbers following the gamma distribution can be procedurally generated (Tanizaki 2008).

To avoid dependency on the initial values, the initial rounds are discarded as the burn-in period. Also, to avoid dependency among the nearby rounds, the values are adopted to construct the distributions (i.e., frequencies of the values sampled) at intervals (thinning). Using the values obtained, the target distributions are estimated.

3.7.3.8 An Example of Parameter Estimation by Gibbs Sampling

Consider the case of a linear mixed model with five genetic markers (G10, G20, G30, G40, and G50) with two environmental factors (E1 and E2) as follows:

$$\boldsymbol{y} = \boldsymbol{X}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix} + \boldsymbol{Z}_{10}\begin{pmatrix}u_{11}\\ u_{12}\end{pmatrix} + \boldsymbol{Z}_{20}\begin{pmatrix}u_{21}\\ u_{22}\end{pmatrix} + \boldsymbol{Z}_{30}\begin{pmatrix}u_{31}\\ u_{32}\end{pmatrix} + \boldsymbol{Z}_{40}\begin{pmatrix}u_{41}\\ u_{42}\end{pmatrix}$$
$$+ \boldsymbol{Z}_{50}\begin{pmatrix}u_{51}\\ u_{52}\end{pmatrix} + \boldsymbol{e}. \tag{3.96}$$

The results of the estimation of the parameters by the Gibbs sampling explained in the preceding sections can be visualized as in Fig. 3.2. The line graphs in the right panels present the values of the parameters sampled ($T = 1000$ rounds), showing

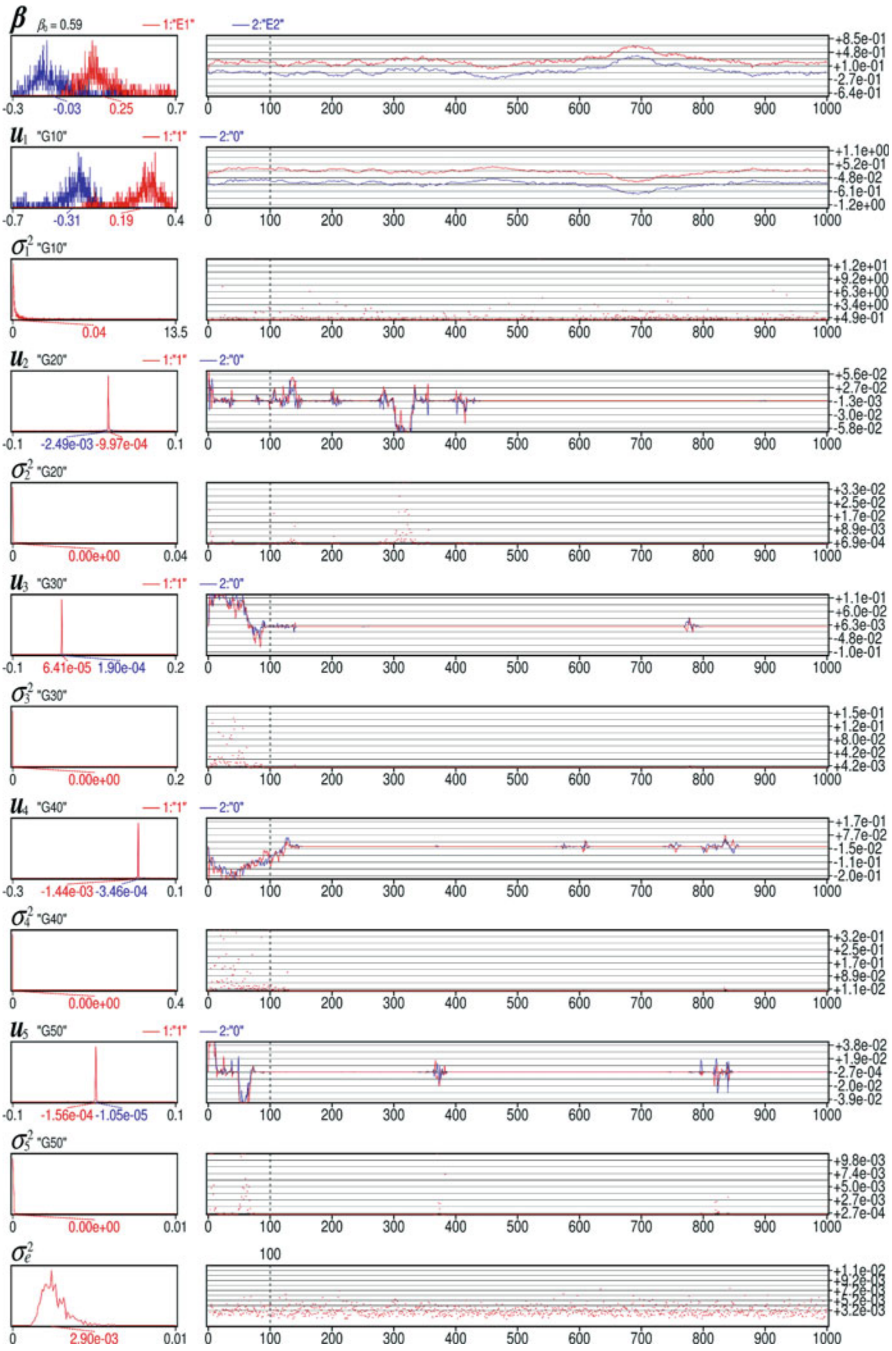


Fig. 3.2 Estimation of the parameters by Gibbs sampling

that the values converge after the initial disturbance. The histograms in the left panels present the estimated distributions (posterior probability distributions) of the parameters by using the sampled values obtained in the right panels. For each of $\boldsymbol{u}_j$ ($j = 1, \ldots, 5$), the red and blue bars, respectively, correspond to u_{j1} and u_{j2}. The red and blue bars, respectively, correspond to β_1 and β_2 for $\boldsymbol{\beta}$. The first hundred rounds were discarded as the burn-in period (i.e., the size of the burn-in period is 100). Thinning was not carried out in this example, and all the values after the burn-in period were adopted to construct the distributions.

Among the five genetic markers, the 10th genetic marker (G10) has the difference between the effects of the two genotypes against the target quantitative trait. All the genetic markers except for G10 have variances equal to zero or less than a sufficiently small number. Their effects of the genotypes ($\boldsymbol{u}_j$) also have similar characteristics. Thus, the genetic markers with such small effects can be excluded from the linear model, matching the expectation initially made for the dataset in Fig. 3.1.

If we exclude the genetic markers with small variance (σ_j^2) and coefficients ($\boldsymbol{u}_j$) other than the 10th genetic marker ($\sigma_1^2 = 0.04$) from the model, we have:

$$\boldsymbol{y} = \boldsymbol{X}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix} + \boldsymbol{Z}_{10}\begin{pmatrix}u_{11}\\ u_{12}\end{pmatrix} + \boldsymbol{e}. \tag{3.97}$$

Using the values obtained by the Gibbs sampling (Fig. 3.2), we have:

$$\boldsymbol{y} = \boldsymbol{X}\begin{pmatrix}0.59\\ 0.25\\ -0.03\end{pmatrix} + \boldsymbol{Z}_{10}\begin{pmatrix}0.19\\ -0.31\end{pmatrix} + \boldsymbol{e}. \tag{3.98}$$

If a sample has the genotype encoded as 1 (gray color in Fig. 3.1) in the design matrix at the 10th genetic marker, the value by the model is 0.59 + 0.19 = 0.78. On the other hand, if a sample has the genotype encoded as 0 in the design matrix at the 10th genetic marker, the value by the model is 0.59−0.31 = 0.28. The contributions of the environmental factors, E1 and E2, are, respectively, 0.25 and −0.03 and cause a difference in the trait value of 0.25 + 0.03 = 0.28, meaning that the samples in E1 have a gain of 0.28 relative to E2. Using the model obtained, we can predict the trait value of a sample.

3.8 Summary and Conclusions

A typical framework of GS requires prediction of the trait values of the samples to determine the candidates for the selected individuals. Finding the genetic markers associated with a target trait might not be sufficient for GS, even if the genetic markers are distributed in a genome-wide manner. Therefore, deriving the prediction model of the target trait is an essential part of GS, since it enables the selection to be based on the predicted trait values. Once such a prediction model is established, the samples with the identical genetic characteristics can be evaluated even if they do not have the trait values observed. When, however, the number of genetic markers is excessive relative to the number of the samples, or the number of the samples is not sufficient for the number of the genetic markers, several assumptions for the estimation of the model parameters—e.g., the genetic markers are independent of each other—break and the prediction model will be distorted. Linkage disequilibrium (LD) actually appears among the genetic markers in a dataset with a small number of samples. Although it is not clear that such dependence among the genetic markers reflects the true biological or genetic mechanisms, the number of the genetic markers sharing the same or a similar genotype pattern increases at least in those small datasets, detracting from the prediction ability. The problems related to overfitting must also be taken into account to assure the generalization ability. The accuracy in the prediction of the hidden trait values of the samples that are independent from the model construction can be evaluated by methods such as the cross-validation test and the bootstrapping test. Although it will be necessary to carry out further inspection of the prediction model for practical use, the strategy based on the linear mixed models and the Bayesian estimation will be useful as a fundamental step in the prediction of trait values of samples based on their genetic and environmental factors.

References

Box GEP, Muller ME (1958) A note on the generation of random normal deviates. Ann Math Statist 29(2):610–611

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

Tanizaki H (2008) A simple gamma random number generator for arbitrary shape parameters. Econ Bull 3(7):1–10

Wang CS, Rutledge JJ, Gianola D (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet Sel Evol 25(1):41–62

Chapter 4
Bayesian Genomic-Enabled Prediction Models for Ordinal and Count Data

Osval A. Montesinos-López, Abelardo Montesinos-López, and José Crossa

4.1 Introduction

Animal and plant breeding have been revolutionized by genomic-enabled prediction models. This tool is powerful for predicting the genomic merit of animals and plants based on high-density single nucleotide polymorphism (SNP) marker panels, and it has been implemented for genomic prediction for the predisposition to some diseases in human health (Yang and Tempelman 2012). However, most existing genomic-enabled prediction models assume normality (in the phenotype and error), linearity in the model parameters, and a constant variance. To translate these models for a non-Gaussian context is a complex task because integrating over the random effects is intractable (McCulloch and Searle 2001). For this reason, researchers normally approach non-Gaussian phenotypes in three ways: (a) they assume normality in the phenotypes, (b) they approximate to normality the non-normal response transforming the phenotype, or (c) they model the appropriate distribution of the phenotype using generalized linear mixed models (GLMMs) (Stroup 2015).

The first approach is justified for large sample sizes by the central limit theorem. However, empirical and simulation studies have shown that this first approach produces highly biased results for small and moderate sample sizes (Stroup 2012,

O.A. Montesinos-López
Facultad de Telemática, Universidad de Colima, Colima 28040, Colima, Mexico

A. Montesinos-López
Departamento de Matemáticas, Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, Jalisco, Guadalajara 44430, Mexico

J. Crossa (✉)
Biometric and Statistics Unit (BSU), International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600, D.F., México
e-mail: j.crossa@cgiar.org

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7_4

2015). Transformations originally were proposed for variance stabilization of non-normal data to obtain a homogeneous variance (Bartlett 1947); this approach is still popular in many agricultural disciplines. The transformed phenotypes are assumed normally distributed variables and are implemented with the traditional linear model. However, often these remedial measures produce a great loss of accuracy and power (Stroup 2015), mostly in small sample sizes.

GLMMs unify models characterized as being linear on the systematic component (model predictors). For this reason, they are appropriate for normal and non-normal data with heterogeneous variance and even correlated observations (Nelder and Wedderburn 1972). GLMMs are very popular in many areas (finance, healthcare, biostatistics, etc.) but, to date, still underutilized in the agricultural research community. Empirical and simulation studies on small sample investigations show that GLMMs produce more accuracy and power than approaches (a) and (b) previously described. Also, for implementing GLMMs there are textbooks and software available, although implementation of approaches (a) and (b) are the dominant approaches in agricultural research (Stroup 2015). The use of GLMMs in genomic-enabled prediction is new because their implementation is not straightforward given that the number of observations usually is smaller than the number of covariates. In addition, the joint involvement of biological processes and pathways complex dependence structures are observed among markers and lines.

In the pre-genomic era, the use of models for non-normal data is not new. Wright (1934) developed the threshold concept to map a normally distributed underlying variable to the observed categorical phenotypes, and the ordinal categorical phenotype is assumed to be the visible expression of an underlying continuous variable (de Maturana et al. 2009). Gianola (1980, 1982) and Gianola and Foulley (1983) proposed a probit (threshold) model for ordinal categorical traits in animal breeding and Gonzalez-Recio, Forni (González-Recio and Forni 2011) and Villanueva et al. (2011) for binary trials. Authors Wang et al. (2013) and Montesinos-López et al. (2015a) extended to threshold model for more than two ordinal categories to deal with $p \gg n$ in the genomic era. Also, de los Campos and Perez-Rodriguez (2013) developed the BGLR package for genomic-enabled prediction for normal, binary, ordinal, and censored data. A log transformation is often used for counts to satisfy normality rather than being modeled on the basis of a count distribution. This transformation for count data is inefficient when there are zeros as observations, because with only one observation with zero, the entire data set needs to be shifted by adding an arbitrary value (usually 1) before transformation. Also, many times this transformation performs poorly, except when dispersion is small and mean counts are large (O'Hara and Kotze 2010). Next we present a review of the existing methods for genome-enabled prediction models for ordinal and count data that give a better idea of the need to develop this type of models.

Kizilkaya et al. (2014), in their paper titled "Reduction in accuracy of genomic prediction for ordered categorical data compared to continuous observations," pointed out that methods used to analyze continuously distributed traits are not optimal for analyzing categorical traits; therefore, it is important to develop appropriate methods for categorical ordinal data. Many low heritability traits have ordered categorical scores, such as susceptibility or resistance to a disease and reproductive traits such as calving difficulty (Kizilkaya et al. 2014). The goal of this study was to quantify reductions in accuracy for ordinal categorical traits relative to continuous traits (Kizilkaya et al. 2014).

For the above reasons, Kizilkaya et al. (2011) used a BayesC threshold model to analyze the ordinal categorical trait infectious bovine keratoconjunctivitis in Angus beef cattle. The same model was used for the genome-wide association analysis of pregnancy in Brangus heifers and first service conception (González-Recio and Forni 2011) and for insect bite hypersensitivity (Schurink et al. 2012).

Kizilkaya et al. (2014) show that genome-wide analysis of ordinal categorical data produced substantially lower accuracy of genomic expected breeding values (GEBV) than the analysis of a continuous phenotype. Kizilkaya et al. (2014) also found that a 2.25 larger training population size for ordinal categorical phenotypes analyzed using a threshold model is required to achieve an accuracy equal to or greater than that for continuous phenotypes for a training population size of 1000 animals. However, using a linear model (assuming normality), a more than 2.25-fold increase in the size of the training population would be required to achieve the same accuracy as a continuous trait with 1000 observations for analyzing an ordinal categorical phenotype. They also found that GEBV accuracy increased significantly when the training population size and heritability increased for the threshold model and for all number of categories in the ordinal categorical data (Kizilkaya et al. 2014).

Kizilkaya et al. (2014) also concluded that when analyzing categorical data, the threshold model had higher accuracies than the linear model (which assumes normality in the phenotype). The research of Varona et al. (1999) also reached similar conclusions when comparing linear and threshold models in conventional pedigree-based evaluations (EBV) using simulated data sets for calving difficulty. A study of Ramirez-Valverde et al. (2001) also supports this finding; they compared EBV accuracy of threshold animal, threshold sire-maternal grandsire, linear animal, and linear sire-maternal grandsire models for calving difficulty in beef cattle and determined that EBV accuracy of the threshold model was 10% higher than EBV accuracy of the linear model for animal and sire-maternal grandsire models. Casella et al. (2007) analyzed litter size using linear and threshold models and found better goodness of fit and predictive ability for EBV in a threshold model than in a linear model (Kizilkaya et al. 2014). These results are in agreement with those reported by Villanueva et al. (2011), who developed a version of the BayesB method for dichotomous traits and concluded that the threshold BayesB method improves prediction accuracy when dealing with disease-resistant dichotomous phenotypes, compared with accuracies obtained with the linear model. The threshold model showed an increase in accuracy of up to 16%, as well as significant advantages when heritability and disease prevalence were low and individuals were genotyped but not measured (testing set).

Kizilkaya et al. (2014) concluded that bias in predictions is reduced in the threshold model when heritability and training population size increase. Although this bias is considerable, it is worse when the data are categorical ordinal but analyzed as if they were continuous using a linear model. Kizilkaya et al. (2014)

also point out that linear model analyses perform as well as threshold model analyses when the number of categories is large. However, when training populations are small, the accuracies of GEBV for ordinal categorical phenotypes analyzed by the threshold model are higher than those analyzed with a linear model applied to the ordinal data.

On the other hand, Wang et al. (2013) showed that Bayesian threshold methods (BayesTA, BayesTB, and BayesTCπ) performed better than the corresponding normal Bayesian methods (BayesA, BayesB, and BayesCπ) in all cases (with 20, 50, 200, and 500 QTL), except in the case of 20 QTL, where BayesB, BayesCπ, BayesTB, and BayesTCπ gave almost the same accuracies. Wang et al. (2013) also found that BayesTB, BayesTCπ, BayesB, and BayesCπ were sensitive to the number of QTL, and their accuracies decreased rapidly when the number of simulated QTL increased from 20 to 200. In contrast, BayesTA and BayesA accuracies did not change (were not sensitive) to the number of simulated QTL.

Wang et al. (2013) also found that when the incidence of the binary trait decreased from 50% to 5%, the accuracies of GEBV decreased consistently. But the three Bayesian threshold methods (BayesTA, BayesTB, and BayesTCπ) performed better than the corresponding normal Bayesian methods in all cases. BayesTB and BayesTCπ produced similar accuracies, and their advantage over BayesB and BayesCπ increased as incidence decreased. Wang et al. (2013) also found that as the number of phenotypic categories increased, the accuracies of GEBV for all the Bayesian methods increased, but the superiority of the three BayesT methods over the corresponding normal Bayesian methods decreased as the number of categories increased, and with eight or more categories, the three BayesT methods completely lost their advantage. BayesA was the most sensitive of all methods to the number of categories, while BayesTA was not sensitive to the number of categories.

Wang et al. (2013) found that the accuracies in generation 2 improved by 30.4%, 2.4%, and 5.7% for BayesTA, BayesTB, and BayesTCπ, respectively, when the number of categories = 2, incidence = 0.3, number of QTL = 50, and heritability = 0.3. They also concluded that the performance of the methods (threshold and normal) significantly is affected by the genetic architecture underlying the traits since the accuracies of all methods declined with the decrease of the heritability when increasing the number of QTL.

The threshold model above mentioned were developed and applied in the context of animal breeding. However, the study by Montesinos-López et al. (2015a) titled "Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding," was conducted, in the context of plant breeding. Montesinos-López et al. (2015a) extended the so-called genomic best linear unbiased predictor (GBLUP) model for Gaussian phenotypes to ordinal data with probit link (TGBLUP). The main contributions of Montesinos-López et al. (2015a) are summarized as:

(a) Real data were used, not simulated data as in Wang et al. (2013) and Kizilkaya et al. (2014).

(b) They take into account genotype × environment interaction (G × E) interaction.
(c) They provide a very clear description of the threshold model and provides R code for its implementation.
(d) They provide an alternative metric (Brier score) for assessing prediction accuracy for categorical ordinal outcomes.
(e) They take into account epistatic additive × additive terms, even though this did not help much to increase prediction accuracy.

Montesinos-López et al. (2015a) found that models that take into account G × E have higher prediction accuracy than those that ignore the G × E term. Relative to models based on main effects only, models that include G × E gave gains in prediction accuracy between 9% and 14%.

Due to their connection to odds ratios and since it provides regression coefficients that are more interpretable, the ordinal logistic regression model is often preferred over the ordinal probit model in statistical applications (Zucknick and Richardson 2014). However, only the Bayesian probit ordinal regression (BPOR) model is frequently implemented in genomic-enabled prediction (when $p \gg n$), given that Bayesian methods that introduce sparseness through additional priors on the model size are very well suited to this problem. Due to the lack of a Bayesian logistic ordinal regression (BLOR) model analogous to the BPOR model that uses a data augmentation approach, Montesinos-López et al. (2015b) proposed the BLOR with logit link, without taking into account G × E interaction. This BLOR model was developed using the Pólya-Gamma data augmentation approach that produces a Gibbs sampler with similar full conditional distributions as the BPOR model, with the advantage that the BPOR model is a particular case of the BLOR model. The authors evaluated the proposed BLOR model using three sets of data. Results from Montesinos-López et al. (2015b) indicate that BLOR model is an alternative for analyzing ordinal data in the context of genomic-enabled prediction with the probit or logit link.

For count data, only two models for genomic-enabled prediction were found. The first one was proposed by Montesinos-López et al. (2015c) titled "Genomic prediction models for count data," extended the GBLUP to count data (CGGLUP), and allows modeling count data without assuming that the data are normally approximated and without using transformation, which many times produces estimations and predictions outside of nonnegativity, which makes no sense for count data. However, this model does not take into account G × E interaction. For this reason, Montesinos-López et al. (2016) extended this model to incorporate G × E interaction. They found that the Bayesian negative binomial regression (BNBR) model with G × E improved prediction accuracy compared to the normal model and to a normal model that uses log-transformed responses.

For both models (BLOR and BNBR), there are Bayesian implementations that estimate the posterior distribution of the required parameters via the Markov chain Monte Carlo (MCMC) algorithm with Gibbs sampling; however, full conditionals for the above models do not have analytic solutions. Therefore, this approach is not useful for large data sets that are commonly used in genomic selection.

For the above reasons, in this chapter, we provide an extension of the BLOR taking into account G × E interaction for ordinal categorical phenotypes; we also provide details of the derivation and implementation of the BNBR model for count data proposed by Montesinos-Lopez et al. (Montesinos-López et al. 2016) that include the G × E term. The full conditionals of the parameters required in both models were obtained analytically using the Pólya-Gamma augmentation approach (Polson et al. 2013) that allows the implementation of an efficient Gibbs sampler for the BLOR and BNBR models with G × E. These models could be very useful for genomic-enabled prediction in plant breeding because they take into account G × E interaction and are powerful enough to deal with large numbers of covariates and small numbers of observations. We illustrate our proposed method with simulation and a real data set.

4.2 Materials and Methods

4.2.1 Data Sets

4.2.1.1 Gray Leaf Spot and *Septoria* Data Sets

Gray leaf spot (GLS), caused by *Cercospora zeae-maydis*, is a foliar disease of global importance in maize production. The disease was evaluated using an ordinal scale [1 (no disease), 2 (low infection), 3 (moderate infection), 4 (high infection), 5 (complete infection)] in three environments (Mexico, Harare, and Colombia). Of the 278 maize lines evaluated, only 240 were the same in the three environments. For this reason, we used only the 240 lines to illustrate our methods with real data. The use of a Poisson random variable for analyzing ordered categorical responses is not new; for example, Vazquez et al. (2009) compared Poisson and threshold models for genetic analysis of clinical mastitis in the US Holsteins. These data are part of a data set that was previously analyzed (Crossa et al. 2011; González-Camacho et al. 2012) under the assumption of normality and using a threshold model for ordinal data (Montesinos-López et al. 2015a; Montesinos-López et al. 2015b). Genotypes of all 240 lines used were obtained using the 55 k single nucleotide polymorphism (SNP) Illumina platform. SNPs with $>10\%$ missing values or a minor allele frequency of ≤ 0.05 were excluded from the data. After line-specific quality control (applying the same quality control to each line separately), the maize data still contained 46,347 SNPs, which were used in the analysis.

4.2.1.2 Fusarium Head Blight Data

From a total of 297 spring wheat lines from CIMMYT evaluated for resistance to Fusarium head blight (FHB), 182 were used for implementing the models for count data because only for these lines had complete marker information. The phenotyping was measured in three environments (El Batan 2012, El Batan 2014, and Ecuador 2014). In each environment the genotypes were arranged in a randomized complete block design. The response variable FHB severity data were collected shortly before maturity by counting symptomatic spikelets on ten randomly selected spikes in each plot. DNA samples were genotyped using an Illumina 9 K SNP chip with 8632 SNPs (Cavanagh et al. 2013). After filtering the markers for 0.05 minor allele frequency (MAF) and deleting markers with more than 10% of no calls, the final set of SNPs was 1635 SNPs (Montesinos-López et al. 2016). This data set was only used for the models for count data.

4.2.2 Statistical Models

We use y_{ijt} to represent the response for the tth replication of the jth line in the ith environment with $i = 1, \ldots, I; j = 1, 2, \ldots, J, t = 1, 2, \ldots, n_{ij}$, and we propose the following linear predictor that takes into account $G \times E$:

$$\eta_{ij} = E_i + g_j + gE_{ij} \tag{4.1}$$

where E_i represents the environment i and is assumed fixed, g_j is the marker effect of genotype j, gE_{ij} is the interaction between genotypes and environments, $I = 3$ (Colombia, Zimbabwe, and Mexico), $J = 240$ (i.e., the number of lines under study), and n_{ij} represent the number of replicates of each line in each environment. The number of observations in environment i is $n_i = \sum_{j=1}^{J} n_{ij}$, while the total number of observations is $n = \sum_{i=1}^{I} n_i$. Rewriting the linear predictor (Eq. 4.1) as

$$\eta_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta} + b_{1j} + b_{2ij} \tag{4.2}$$

with $\boldsymbol{x}_i^T = [x_{i1}, x_{i2}, x_{i3}]$, where x_{i1}, x_{i2}, and x_{i3} are indicator variables that take the value of 1 if the observed environment i is 1, 2, or 3, respectively, and 0 otherwise, $\boldsymbol{\beta}^T = [\beta_{1,} \beta_{2,}, \beta_{3,}]$ because three is the number of environments under study, $\boldsymbol{x}_i^T \boldsymbol{\beta} = E_i$, $b_{1j} = g_j$ and $b_{2ij} = gE_{ij}$. Three models are proposed using the linear predictor given in Eqs. (4.1) and (4.2).

Model BLOR In this model, the linear predictor is $\eta_{ij(c)} = \gamma_c - \eta_{ij}$, where $\eta_{ij(c)}$ denotes the c^{th} link ($c = 1, 2, \ldots, C - 1$) for the fixed and random effects combination, γ_c is the threshold (intercept) for the c^{th} link, and η_{ij} is exactly as defined in Eq. (4.2). Distributions: $y_{ijt(1)}, y_{ijt(2)}, \ldots, y_{ijt(C)} | \eta_{ij} \sim \text{Multinomial}(1, \pi_{ij(1)}, \pi_{ij(2)}, \ldots, \pi_{ij(C)})$. $\boldsymbol{b}_1 = (b_{11}, \ldots, b_{1J})^T$

$\sim N(\mathbf{0}, \boldsymbol{G}_1\sigma^2_{b_1}), \boldsymbol{b}_{2i} = (b_{2i1}, \ldots, b_{2iJ})^{\boldsymbol{T}}, \boldsymbol{b}_2 = (\boldsymbol{b}^T_{21}, \ldots, \boldsymbol{b}^T_{2I})^T \sim N(\mathbf{0}, \boldsymbol{G}_2\sigma^2_{b_2})$. Link function: cumulative logit $\{\ \eta_{ij(c)} = \log\left(\frac{\pi_{ij(c)}}{1-\pi_{ij(c)}}\right), c = 1, 2, \ldots, C-1)\}$, since there are C categories, a total of $C-1$ link functions are required to fully specify the model. $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ were assumed known, with $\boldsymbol{G}_1$ computed from marker $\boldsymbol{W}$ data (for $m = 1, \ldots,$ q markers) as $\boldsymbol{G}_1 = \frac{\boldsymbol{WW}^T}{q}$**;** this matrix is called the genomic relationship matrix (GRM) (VanRaden 2008). The $\boldsymbol{G}_1$ matrix is a covariance matrix that contains the similarity between individuals based on marker information, rather than on expected similarity based on pedigree, that can help to improve prediction accuracy. While $\boldsymbol{G}_2$ is computed as $\boldsymbol{G}_2 = \boldsymbol{I}_{\boldsymbol{I}} \otimes \boldsymbol{G}_1$ of order IJxIJ and $\otimes$ denotes the Kronecker product, $\boldsymbol{I}_{\boldsymbol{I}}$ means that we assume independence between environments (Montesinos-López et al. 2016).

Model BNBR Linear predictor as given in Eq. (4.1). Distributions: $y_{ijt}|\eta_{ij}$~NB(μ_{ij}, r), NB stands for negative binomial distribution with r being the dispersion parameter (shape parameter), $\mu_{ij} = E(y_{ijt}|\eta_{ij}) = \exp(\eta_{ij})$. Note that the BNBR has expected value $\mu_{ij} = \frac{r\pi_{ij}}{(1-\pi_{ij})}$ and variance $\frac{r\pi_{ij}}{(1-\pi_{ij})^2} = \mu_{ij} + \frac{\mu^2_{ij}}{r}$, with the variance greater than the mean (Montesinos-López et al. 2016).

Model Pois Everything is the same as in **model BNBR**, except that $y_{ijt}|\eta_{ij}$~Poisson (μ_{ij}). Since according to Zhou et al. (2012) and Teerapabolarn and Jaioun (2014), the $\lim_{r\to\infty} NB(\mu_{ij}, r) = Pois(\mu_{ij})$, **model Pois** was implemented using the same method as **model BNBR**, but fixing r to a large value, depending on the mean count. We used $r = 1000$, which is reasonable when the mean count is less than 50 (see Fig. 4.1). However, for mean counts between 50 and 200, we suggest using $r = 5000$, and for counts larger than 200, we suggest a value of $r = 10000$ or larger. Figure 4.1 supports these suggestions where we plot the mean and variance of

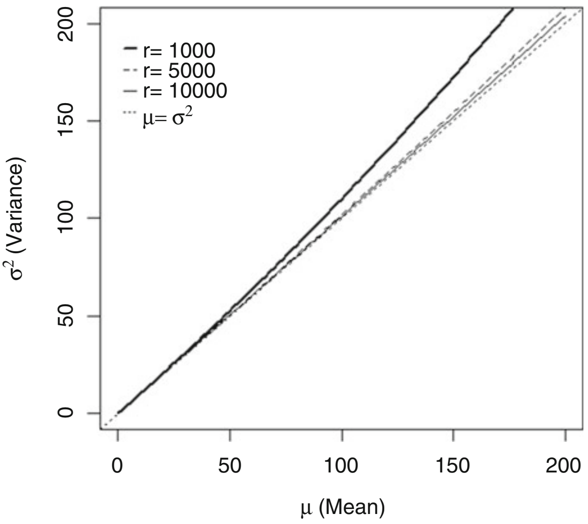


Fig. 4.1 Plot of the mean count versus the variance of NB distribution as a function of the scale parameter (r). Good approximations are obtained when the mean and variance are very similar; in the plot, they should follow the diagonal that plots $\mu = \sigma^2$ (Extracted from Montesinos-López et al. 2016)

model BNBR as a function of the scale parameter r, with three values of r (1000, 5000, 10,000). Acceptable approximations to the **model Pois** with the **model BNBR** occur when the mean and variance are very similar. For this reason, good approximations are those that follow the diagonal in Fig. 4.1, where $\mu = \sigma^2$. The mean count and variances are very similar for mean counts of less than 50 with $r = 1000$; however, when the mean count is larger than 50 and less than 200, we should use $r = 5000$, and for counts greater than 200, we suggest using a value of $r = 10{,}000$ or larger. In our applications with simulated and real data, the mean count is less than 50; for this reason, we used a value of $r = 1000$. Next, we provide details of the derivation of the full conditional distribution for each model (Montesinos-López et al. 2016).

4.2.2.1 Bayesian Logistic Ordinal Regression (BLOR)

Let $\boldsymbol{y}_{ij} = \left[y_{ij1}, \ldots, y_{ijn_{ij}}\right]^T$, $\boldsymbol{y}_i = \left[\boldsymbol{y}_{i1}^T, \ldots, \boldsymbol{y}_{iJ}^T\right]^T$, and $\boldsymbol{y} = \left[\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_I^T\right]^T$; in this model, the response variable y_{ijt} represents an assignment into one of C mutually exclusive and exhaustive categories that follow an order. Therefore, the ordinal logistic regression model can be written in terms of a latent response variable l_{ijt} as

$$l_{ijt} = \boldsymbol{x}_i^T \boldsymbol{\beta} + b_{1j} + b_{2ij} + \varepsilon_{ijt} \tag{4.3}$$

where l_{ijt} are called “liabilities,” $\varepsilon_{ijt} \sim L(0, 1)$, where $L(.)$ denotes the logistic distribution, and the remaining terms are as defined in Eq. (4.2). Since l_{ijt} are unobservable and can be measured indirectly by an observable ordinal variable y_{ijt}, then y_{ijt} can be defined by

$$y_{ijt} = \begin{cases} 1 & if\ -\infty < l_{ijt} < \gamma_1, \\ 2 & if \quad \gamma_1 < l_{ijt} < \gamma_2, \\ & \quad \vdots \\ C & if \quad \gamma_{C-1} < l_{ijt} < \infty \end{cases}$$

This means that l_{ijt} is divided by thresholds into C intervals, corresponding to C ordered categories. The first threshold, γ_1, defines the upper bound of the interval corresponding to observed outcome 1. Similarly, threshold γ_{C-1} defines the lower bound of the interval corresponding to observed outcome C. Threshold γ_c defines the boundary between the interval corresponding to observed outcomes $c - 1$ and c for (c=1,2,.., C -1). Threshold parameters are $\boldsymbol{\gamma}^T = (\gamma_{min} < \gamma_1 < \cdots < \gamma_{C-1} < \gamma_{max})$ with $\gamma_{min} = -\infty$ and $\gamma_{max} = \infty$.

Because that the error term ε_{ijt} of the latent response l_{ijt} is distributed as $L(0, 1)$, the cumulative response probability for the c category of the ordinal outcome y_{ijt} is

$$\begin{aligned} P(y_{ijt} \leq c|\boldsymbol{\beta},\boldsymbol{b}_1,\boldsymbol{b}_2) = \pi_{ij(c)} = P(l_{ijt} \leq \gamma_c|\boldsymbol{\beta},\boldsymbol{b}_1,\boldsymbol{b}_2) = \mathrm{P}(\boldsymbol{x}_i^T\boldsymbol{\beta} + b_{1j} + b_{2ij} + \varepsilon_{ijt} \leq \gamma_c) \\ = \mathrm{P}(\varepsilon_{ijt} \leq \gamma_c - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij}), \text{for } c = 1, 2, \ldots, C-1. \\ = \frac{\exp(\gamma_c - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij})}{1 + \exp(\gamma_c - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij})} \end{aligned} \tag{4.4}$$

Similarly, Eq. (4.4) can be written as a cumulative logit model:

$$\log\left(\frac{\pi_{ij(c)}}{1 - \pi_{ij(c)}}\right) = \gamma_c - \boldsymbol{x_i^T\beta} - b_{1j} - b_{2ij}, \text{for } c = 1, 2, \ldots, C-1.$$

Using the inverse link for this model, $P(y_{ijt} = c|\boldsymbol{\beta}, \boldsymbol{b}_1, \boldsymbol{b}_2) = \pi_{ij(c)}$ can be calculated as follows:

$$\begin{aligned} \pi_{ij(c)} &= P(\gamma_{c-1} < l_{ijt} < \gamma_c) \\ &= \frac{\exp(\gamma_c - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij})}{1 + \exp(\gamma_c - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij})} - \frac{\exp(\gamma_{c-1} - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij})}{1 + \exp(\gamma_{c-1} - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij})}. \end{aligned}$$

Since we have latent variables l_{ijt} distributed as $L\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + b_{1j} + b_{2ij}, 1\right)$ and we observe $y_{ijt} = c$ if, and only if, $\gamma_{c-1} < l_{ijt} < \gamma_c$, then the joint posterior density of the parameter vector and latent variable becomes

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_1, \boldsymbol{b}_2, \sigma_\beta^2, \sigma_{b_1}^2, \sigma_{b_2}^2 \boldsymbol{l}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{l}, \boldsymbol{\gamma}) f(\boldsymbol{\theta_T})$$

where $f(\boldsymbol{\theta_T}) = f(\boldsymbol{l}|\boldsymbol{\beta}, \boldsymbol{b}_1, \boldsymbol{b}_2) f(\boldsymbol{\gamma}) f\left(\boldsymbol{\beta}|\sigma_\beta^2\right) f(\boldsymbol{b}_1|\sigma_{b_1}^2) f(\boldsymbol{b}_2|\sigma_{b_2}^2) f(\sigma_\beta^2) f(\sigma_{b_1}^2) f(\sigma_{b_2}^2)$. Then assuming a scaled independent inverse chi-square $\chi^{-2}(\nu_h, S_h)$ prior for $\sigma_{b_h}^2$ for $h = 1, 2$, a normal prior distribution for $f\left(\boldsymbol{\beta}|\sigma_\beta^2\right) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0\sigma_\beta^2)$, a normal prior distribution for $f(\boldsymbol{b}_1|\sigma_{b_1}^2) \sim N_J(0, \boldsymbol{G}_1\sigma_{b_1}^2)$, a normal prior distribution for $f(\boldsymbol{b}_2|\sigma_{b_2}^2) \sim N_{IJ}(0, \boldsymbol{G}_2\sigma_{b_2}^2)$, and also a $\chi^{-2}(\nu_\beta, S_\beta)$ prior was given for σ_β^2 (Gianola 2013). Following Sorensen et al. (1995), a prior for the $C - 1$ unknown thresholds has been given as order statistics from $\mathrm{U}(\gamma_{min}, \gamma_{max})$ distribution,

$$P(\boldsymbol{\gamma}) = (C-1)!\left(\frac{1}{\gamma_{max} - \gamma_{min}}\right)^{C-1} I(\boldsymbol{\gamma} \in \boldsymbol{T})$$

where $\boldsymbol{T} = \{(\gamma_1, \ldots, \gamma_{max})|\gamma_{min} < \gamma_1 < \cdots < \gamma_{C-1} < \gamma_{max}\}$.

The full conditional posterior distributions are provided below and in Appendix A are all details of these derivations.

4.2.2.2 Liabilities and ω_{ijt}

The fully conditional posterior distribution of liability l_{ijt} is a truncated normal distribution and its density is

$$f(l_{ijt}|ELSE) = \frac{\phi\left(\sqrt{\omega_{ijt}}(l_{ijt} - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij})\right)}{\Phi(\sqrt{\omega_{ijt}}(\gamma_c - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij})) - \Phi\left(\sqrt{\omega_{ijt}}(\gamma_{c-1} - \boldsymbol{x}_i^T\boldsymbol{\beta} - b_{1j} - b_{2ij})\right)} \tag{4.5}$$

For simplicity, *ELSE* is the data and the parameters, except for the one in question. ϕ and Φ are the density and distribution function of a standard normal random variable, and the full conditional posterior distribution of ω_{ijt} is

$$f(\omega_{ijt}|ELSE) \sim PG(2, \, - l_{ijt} + \boldsymbol{x}_i^T\boldsymbol{\beta} + b_{1j} + b_{2ij}) \tag{4.6}$$

where *PG* stands for the Pólya-Gamma distribution.

4.2.2.3 Regression Coefficients ($\boldsymbol{\beta}$)

The full conditional posterior of $\boldsymbol{\beta}$ is

$$f(\boldsymbol{\beta}|\boldsymbol{ELSE}) \sim N_p(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\Sigma}}_0) \tag{4.7}$$

where $\tilde{\boldsymbol{\Sigma}}_0 = (\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \boldsymbol{X}^T\boldsymbol{D}_{\boldsymbol{\omega}}\boldsymbol{X})^{-1}$, and $\tilde{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\Sigma}}_0(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \boldsymbol{X}^T\boldsymbol{D}_\omega \sum_{h=1}^{2} \boldsymbol{Z}_h\boldsymbol{b}_h + \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{l})$. With $\boldsymbol{l} = [\boldsymbol{l}_1^T, \ldots, \boldsymbol{l}_I^T]^T$, $\boldsymbol{l_i} = [\boldsymbol{l}_{i1}^T, \ldots, \boldsymbol{l}_{iJ}^T]^T$, $\boldsymbol{l_{ij}} = \left[l_{ij1}, \ldots, l_{ijn_{ij}}\right]^T$, $\boldsymbol{X_{ij}} = \left[1_{n_{ij}}^T \bigotimes \boldsymbol{x}_i\right]^T$, $\boldsymbol{X_i} = [\boldsymbol{X}_{i1}^T, \ldots, \boldsymbol{X}_{iJ}^T]^T$, $\boldsymbol{X} = [\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_I^T]^T$, $\boldsymbol{D}_{\omega ij} = \text{diag}\left(\omega_{ij1}, \ldots, \omega_{ijn_{ij}}\right)$, $\boldsymbol{D}_{\omega i} = \text{diag}(\boldsymbol{D}_{\omega i1}, \ldots, \boldsymbol{D}_{\omega iJ})$, $\boldsymbol{D}_\omega = \text{diag}(\boldsymbol{D}_{\omega 1}, \ldots, \boldsymbol{D}_{\omega I})$, $\boldsymbol{b}_1 = [b_{11}, \ldots, b_{1J}]^T$, $\boldsymbol{b}_{2i} = [b_{2i1}, \ldots, b_{2iJ}]^T$, $\boldsymbol{b}_2 = [\boldsymbol{b}_{21}^T, \ldots, \boldsymbol{b}_{2I}^T]^T$, $\boldsymbol{Z}_{1i} = \begin{bmatrix} 1_{n_{1i1}} & 0 & \cdots & 0 \\ 0 & 1_{n_{1i2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_{n_{1iJ}} \end{bmatrix}$, $\boldsymbol{Z}_1 = [\boldsymbol{Z}_{11}^T, \ldots, \boldsymbol{Z}_{1I}^T]^T$, and $\boldsymbol{Z}_2 = \boldsymbol{Z}_1{}^* \sim \boldsymbol{X}$, where $^*\sim$indicates the horizontal Kronecker product between $\boldsymbol{Z}_1$ and $\boldsymbol{X}$. The horizontal Kronecker product performs a Kronecker product of $\boldsymbol{Z}_1$ and $\boldsymbol{X}$ and creates a new matrix by stacking these row vectors into a matrix. $\boldsymbol{Z}_1$ and $\boldsymbol{X}$ must have the same number of rows, which is also the same number of rows in the resulting matrix. The number of columns in the resulting matrix is equal to the product of the number of columns in $\boldsymbol{Z}_1$ and $\boldsymbol{X}$. It is important to point out that if we use a prior for $\boldsymbol{\beta} \propto$ Constant (improper uniform distribution), then in $\tilde{\boldsymbol{\Sigma}}_0$ and $\tilde{\boldsymbol{\beta}}_0$ we need to make $\boldsymbol{0}$ the term $\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}$.

4.2.2.4 Polygenic Effects ($\boldsymbol{b}_h$)

The full conditional distribution of $\boldsymbol{b}_h$ with $h = 1\,, 2$, is given as

$$f(\boldsymbol{b}_h|ELSE) \sim N\Big(\widetilde{\boldsymbol{b}}_h = \boldsymbol{F_h}(\boldsymbol{Z}_h^T\boldsymbol{D}_\omega\boldsymbol{l} - \boldsymbol{Z}_h^T\boldsymbol{D}_\omega\boldsymbol{\eta}^h), \boldsymbol{F_h} = (\sigma_{b_h}^{-2}\boldsymbol{G}_h^{-1} + \boldsymbol{Z}_h^T\boldsymbol{D}_\omega\boldsymbol{Z}_h^T)^{-1}\Big) \tag{4.8}$$

with $\boldsymbol{\eta}^1 = \boldsymbol{X\beta} + \boldsymbol{Z}_2\boldsymbol{b}_2$ and $\boldsymbol{\eta}^2 = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{b}_1$.

4.2.2.5 Variance of Polygenic Effects ($\sigma_{b_h}^2$)

Next, the full conditional posterior of $\sigma_{b_h}^2$ is

$$f(\sigma_{b_h}^2|ELSE) \sim \chi^{-2}\Big(\widetilde{\nu}_h = \nu_h + n_{b_h}, \widetilde{S}_b = (\boldsymbol{b}_h^T\boldsymbol{G}_h^{-1}\boldsymbol{b}_h + \nu_h S_h)/\nu_b + n_{b_h}\Big) \tag{4.9}$$

with $n_{b_1} = J$ and $n_{b_2} = IJ$.

4.2.2.6 Threshold Effects (γ_c)

The density of the full conditional posterior distribution of the cth threshold, γ_c, is

$$f(\gamma_c|ELSE) = \frac{1}{\min\{\min(l_{ijt}|y_{ijt} = c+1), \gamma_{c+1}, \gamma_{max}\} - \max\{\max(l_{ijt}|y_{ijt} = c), \gamma_{c-1}, \gamma_{min}\}} \tag{4.10}$$

4.2.2.7 Variance of Regression Coefficients

The full conditional posterior of σ_β^2 is

$$f\Big(\sigma_\beta^2|ELSE\Big) \sim \chi^{-2}\Big(\tilde{\nu}_\beta = \nu_\beta + p, \tilde{S}_\beta = \Big[(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \nu_\beta S_\beta\Big]/\nu_\beta + p\Big) \tag{4.11}$$

4.2.2.8 The Gibbs Sampler for model BLOR

The Gibbs sampler is implemented by sampling repeatedly from the following loop:

Step 1. Sample liabilities (l_{ijt}) from the truncated normal distribution in (4.5).
Step 2. Sample ω_{ijt} values from the Pólya-Gamma distribution in (4.6).
Step 3. Sample the regression coefficients ($\boldsymbol{\beta}$) from the normal distribution in (4.7).
Step 4. Sample the polygenic effects ($\boldsymbol{b}_h$) for $h = 1\,,2$, from the normal distribution in (4.8).
Step 5. Sample the variance effect ($\sigma^2_{b_h}$) for $h = 1\,,2$,from the scaled inverted χ^2 distribution in (4.9).
Step 6. Sample the thresholds (γ_c) from the uniform distribution in (4.10).
Step 7. Sample the variance of regression coefficients (σ^2_β) from the scaled inverted χ^2 distribution in (4.11).
Step 8. Return to step 1 or terminate if chain length is adequate to meet convergence diagnostics.

Ignoring the polygenic effects ($\boldsymbol{b}_h$), the Gibbs sampler given above can be used only by deleting steps 4 and 5. If all marker effects are considered fixed effects and included in the design matrix, $\boldsymbol{X}$, with a prior $\boldsymbol{\beta} \sim N_p(\mathbf{0}, \boldsymbol{I}_p\sigma^2_\beta)$ for the beta regression coefficients, we end up with a threshold Bayesian ridge regression. This is the ridge estimator of Hoerl and Kennard (1970) for ordinal categorical data, since the posterior expectation of $\boldsymbol{\beta}$ is equal to $E(\boldsymbol{\beta}|\ ELSE) = \left(\boldsymbol{X}^T D_\omega \boldsymbol{X} + \boldsymbol{I}_p \sigma_\beta^{-2}\right)^{-1} \boldsymbol{X}^T D_\omega \boldsymbol{l}$ with pseudo-response $\boldsymbol{l}$. Also, note that setting each $\omega_{ijt} = 1$, the Gibbs sampler given above for the BLOR with the logistic link is reduced to the Gibbs sampler with the probit for the BPOR link proposed by Albert and Chib (1993). Therefore, our proposed BLOR model is more general and includes the Gibbs sampler for the BPOR model as a particular case as implemented in BGLR package.

4.2.3 Bayesian Mixed Negative Binomial Regression

Note that under the **model BNBR**, because $\mu_{ij} = E(y_{ijt}|\eta_{ij}) = \exp(\eta_{ij})$, conditionally on b_{1j} and b_{2ij}, the probability that the random variable Y_{ijt} takes the value y_{ijt} is equal to

$$\Pr\left(Y_{ijt} = y_{ijt}|\eta_{ij}\right) = \binom{y_{ijt} + r - 1}{y_{ijt}} \left(1 - \frac{\mu_{ij}}{r + \mu_{ij}}\right)^r \left(\frac{\mu_{ij}}{r + \mu_{ij}}\right)^{y_{ijt}} \quad \text{for } y_{ijt} = 0, 1, 2, \ldots$$

$$= \frac{\Gamma\left(y_{ijt} + r\right)}{y_{ijt}!\Gamma(r)} \frac{\left[\exp\left(\eta_{ij}^{*}\right)\right]^{y_{ijt}}}{\left[1 + \exp\left(\eta_{ij}^{*}\right)\right]^{y_{ijt}+r}}, \quad y_{ijt} = 0, 1, 2, \ldots \tag{4.12}$$

We arrive at Eq. (4.12) because we make $\frac{\mu_{ij}}{r+\mu_{ij}}=r\mu_{ij}$

$$r\left(r+\mu_{ij}\right)=\frac{\mu_{ij/r}}{1+\mu_{ij/r}}=\frac{\exp\left(\eta_{ij}\right)\exp(-\log(r))}{1+\exp\left(\eta_{ij}\right)\exp(-\log(r))}=\frac{\exp\left(\eta_{ij}-\log(r)\right)}{1+\exp\left(\eta_{ij}-\log(r)\right)}=\frac{\exp\left(\eta_{ij}^{*}\right)}{1+\exp\left(\eta_{ij}^{*}\right)},$$

with $\eta_{ij}^{*}=\boldsymbol{x}_i^T\boldsymbol{\beta}^{*}+b_{1j}+b_{2ij},\boldsymbol{\beta}^{*}=\left[\beta_1^{*},\beta_2^{*},\beta_3^{*}\right]$ and $\beta_i^{*}=\beta_i-\log(r)$. Therefore, in Eq. (4.12) we have the connection between the probability distribution of the response (Y_{ijt}) induced by the assumed relation between the linear predictor (η_{ij}) and the expected value of Y_{ijt} (μ_{ij}) under **model BNBR**. Then we can rewrite the Pr(Y_{ijt}=y_{ijt}|η_{ij}) given in Eq. (4.12) as

$$\frac{\Gamma\left(y_{ijt} + r\right)}{y_{ijt}!\Gamma(r)} 2^{-y_{ijt}-r}\exp\left(\frac{y_{ijt} - r}{2}\eta_{ij}^{*}\right) \int_0^{\infty} \exp\left[-\frac{\omega_{ijt}\left(\eta_{ij}^{*}\right)^2}{2}\right] f\left(\omega_{ijt}, y_{ijt} + r, 0\right) d\omega_{ijt} \tag{4.13}$$

Expression (4.13) was obtained using the following equality given by Polson et al. (2013): $\frac{\left(e^{\psi}\right)^a}{\left(1 + e^{\psi}\right)^b} = 2^{-b}e^{\kappa\psi}\int_0^{\infty} e^{-\frac{\omega_{ijt}\psi^2}{2}} f\left(\omega_{ijt}; b, 0\right) d\omega_{ijt}$, where $\kappa = a - b/2$ and $f(., b, 0)$ denotes the density of the Pólya-Gamma distribution (ω_{ijt}) with parameters b and $c = 0$ ($PG(b, c = 0)$) (see Definition 1 in Polson et al. 2013). From here, conditioning on ω_{ijt} ~$PG(y_{ijt} + r, c = 0)$, we have that

$$\Pr\left(Y_{ijt} = y_{ijt}|\eta_{ij}, \omega_{ijt}\right) = \frac{\Gamma\left(y_{ijt} + r\right)}{y_{ijt}!\Gamma(r)} 2^{-y_{ijt}-r}\exp\left(\frac{y_{ijt} - r}{2}\eta_{ij}^{*}\right)\exp\left[-\omega_{ijt}\left(\eta_{ij}^{*}\right)^2/2\right] \tag{4.14}$$

To be able to get the full conditional distributions, here we provide the prior distributions, $f(\boldsymbol{\theta})$, for all the unknown model parameters $\boldsymbol{\beta}^{*}$, $\sigma_{\beta^{*}}^2$, $\boldsymbol{b}_1$, $\sigma_{b_1}^2$, $\boldsymbol{b}_2$, σ_{b2}^2 , and r. We assume the following prior between the parameters, that is,

$$f(\boldsymbol{\theta}) = f\left(\boldsymbol{\beta}^{*}|\sigma_{\boldsymbol{\beta}^{*}}^2\right) f(\sigma_{\beta^{*}}^2) f(\boldsymbol{b}_1|\sigma_{b_1}^2) f(\sigma_{b1}^2) f(\boldsymbol{b}_2|\sigma_2^2) f(\sigma_{b2}^2) f(r)$$

We assign conditionally conjugate but weakly informative prior distributions to the parameters because we have no prior information. The prior specification in terms of $\boldsymbol{\beta}^{*}$ instead of $\boldsymbol{\beta}$ is for convenience. We adopt proper priors with known hyper-parameters whose values we specify in model implementation to guarantee proper posteriors. We assume that $f\left(\boldsymbol{\beta}^*\middle|\sigma^2_{\beta^*}\right) \sim N_I\left(\boldsymbol{\beta}_0, \sum_0 \sigma^2_\beta\right)$, $f\left(\sigma^2_{\beta^*}\right) \sim \chi^{-2}$ $\left(\nu_\beta, S_\beta\right)$ where $\chi^{-2}\left(\nu_\beta, S_\beta\right)$ denote a scaled inverse chi-square distribution with shape ν_β and scale S_β parameters, $f(\boldsymbol{b}_1|\sigma^2_{b1}) \sim N_J(\mathbf{0}, \boldsymbol{G}_1\sigma^2_{b1})$, $f\left(\sigma^2_{b1}\right) \sim \chi^{-2}(\nu_{b1}, S_{b1})$, $f\left(\boldsymbol{b}_2|\sigma^2_{b2}\right) \sim N_{IJ}(\mathbf{0}, \boldsymbol{G}_2\sigma^2_{b2})$, $f\left(\sigma^2_{b2}\right) \sim \chi^{-2}(\nu_{b2}, S_{b2})$, and f $(r) \sim G(a_0, 1/b_0)$. Next we combine (4.14) using all data with priors to get the full conditional distribution for parameters $\boldsymbol{\beta}^*$, $\sigma^2_{\beta^*}$, $\boldsymbol{b}_1$, σ^2_{b1}, $\boldsymbol{b}_2$, σ^2_{b2} , and r.

4.2.4 Full Conditional Distributions

The full conditional distribution of $\boldsymbol{\beta}^*$ is given as

$$f(\boldsymbol{\beta}^*|ELSE) \sim N\left(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\Sigma}}_0\right) \tag{4.15}$$

where $\tilde{\boldsymbol{\Sigma}}_0 = \left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \boldsymbol{X}^T\boldsymbol{D}_{\boldsymbol{\omega}}\boldsymbol{X}\right)^{-1}$, $\tilde{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\Sigma}}_0\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \boldsymbol{X}^T\boldsymbol{D}_\omega\sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h + \right.$ $\left.\boldsymbol{X}^T\boldsymbol{\kappa}\right)$, $\boldsymbol{\kappa_{ij}} = \frac{1}{2}\left[y_{ij1} - r, \ldots, y_{ijn_{ij}} - r\right]^T$, $\boldsymbol{\kappa_i} = \left[\boldsymbol{\kappa}_{i1}^T, \ldots, \boldsymbol{\kappa}_{iJ}^T\right]^T$, and $\boldsymbol{\kappa} = \left[\boldsymbol{\kappa}_1^T, \ldots, \boldsymbol{\kappa}_I^T\right]^T$, where $\boldsymbol{X}$, $\boldsymbol{Z}_h$ and $\boldsymbol{D}_\omega$ are defined exactly as in **Model BLOR**. When the prior for $\boldsymbol{\beta}^*$ $\propto$constant, the posterior distribution of $\boldsymbol{\beta}^*$ is also normally distributed $N\left(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\Sigma}}_0\right)$, but we set the term $\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}$ to zero in both $\tilde{\boldsymbol{\Sigma}}_0$ and $\tilde{\boldsymbol{\beta}}$. In Appendix B are given in details the derivations of the full conditional distributions for the **BNBR model**.

The full conditional distribution of ω_{ijt} is

$$f\left(\omega_{ijt}|ELSE\right) \sim PG(y_{ijt} + r, \boldsymbol{x}_i^T\boldsymbol{\beta}^* + b_{1j} + b_{2ij}) \tag{4.16}$$

The full conditional distribution of $\boldsymbol{b}_h$, with $h = 1, 2$, is given as

$$f(\boldsymbol{b}_h|ELSE) \sim N\left(\tilde{\boldsymbol{b}}_h,\ \boldsymbol{F_h}\right) \tag{4.17}$$

If $\boldsymbol{\eta}^1 = \boldsymbol{X}\ \ \boldsymbol{\beta}^* + \boldsymbol{Z}_2\boldsymbol{b}_2$, then $\boldsymbol{F}_1 = \left(\sigma_{b_1}^{-2}\boldsymbol{G}_1^{-1} + \boldsymbol{Z}_1^T\boldsymbol{D}_\omega\boldsymbol{Z}_1\right)^{-1}$, $\tilde{\boldsymbol{b}}_1 = \boldsymbol{F}_1\left(\boldsymbol{Z}_1^T\boldsymbol{\kappa} - \boldsymbol{Z}_1^T\boldsymbol{D}_\omega\boldsymbol{\eta}^1\right)$, and then $\boldsymbol{b}_1 \mid ELSE \sim N\ \left(\tilde{\boldsymbol{b}}_1,\ \boldsymbol{F}_1\right)$. By defining $\boldsymbol{\eta}^2 = \boldsymbol{X}\ \boldsymbol{\beta}^*$ $+\boldsymbol{Z}_1\boldsymbol{b}_1$ in a similar way, we arrive at the full conditional of $\boldsymbol{b}_2$ as $\boldsymbol{b}_2 \mid ELSE \sim N\left(\tilde{\boldsymbol{b}}_2,\ \boldsymbol{F}_2\right)$, where $\boldsymbol{F}_2 = \left(\sigma_{b_2}^{-2}\boldsymbol{G}_2^{-1} + \boldsymbol{Z}_2^T\boldsymbol{D}_\omega\boldsymbol{Z}_2\right)^{-1}$, $\widehat{\boldsymbol{b}}_2 = \boldsymbol{F}_2\left(\boldsymbol{Z}_2^T\boldsymbol{\kappa} - \boldsymbol{Z}_2^T\boldsymbol{D}_\omega\boldsymbol{\eta}^2\right)$.

The full conditional distribution of $\sigma^2_{b_h}$ is

$$f\left(\sigma^2_{b_h}|ELSE\right) \sim \chi^{-2}\left(\tilde{\nu}_b = \nu_{b_h} + n_{b_h}, \tilde{S}_b = \left(\boldsymbol{b}_h^T \boldsymbol{G}_h^{-1} \boldsymbol{b}_h + \nu_{b_h} S_{b_h}\right)/\nu_{b_h} + n_{b_h}\right) \tag{4.18}$$

The conditional distribution of $\sigma^2_{\beta*}$ is

$$\begin{aligned} f(\sigma^2_{\beta*}|ELSE) & \\ \sim \chi^{-2}\left(\tilde{\nu}_{\beta*}\right. & = \nu_{\beta*} + I, \tilde{S}_\beta \\ & = [(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \nu_{\beta*} S_{\beta*}]/\nu_{\beta*} + I) \end{aligned} \tag{4.19}$$

Taking advantage of the fact that the NB distribution can also be generated using a Poisson representation as pointed out by Quenouille (1949) as $Y = \sum_{l=1}^{L} u_l$, where $u_l \sim Log(\pi)$ and is independent of $L \sim Pois(-r \log(1-\pi))$, where Log and $Pois$ denote logarithmic and Poisson distributions, respectively. Then we infer a latent count L for each $Y \sim NB(\mu, r)$ conditional on Y and r. Therefore, following Zhou et al. (2012), we obtain the full conditional of r by alternating

$$f(r|ELSE) \sim G\left(a_0 - \sum_{i=1}^{I} \sum_{j=1}^{J} \sum\nolimits_{t=1}^{n_{ij}} \log\left(1-\pi_{ij}\right), \frac{1}{b_0 + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{n_{ij}} L_{ijt}}\right) \tag{4.20}$$

$$f\left(L_{ijt}|ELSE\right) \sim CRT\left(y_{ijt}, r\right) \tag{4.21}$$

where $CRT(y_{ijt}, r)$ denotes a Chinese restaurant table (CRT) random count variable that can be generated as $L_{ijt} = \Sigma_{l=1}^{y_{ijt}} d_l$, where $d_l \sim Bernoulli\left(\frac{r}{l-1+r}\right)$ and $\pi_{ij} = \frac{\exp\left(\eta^*_{ij}\right)}{1+\exp\left(\eta^*_{ij}\right)}$.

4.2.5 Gibbs Sampler for Model BNBR

The Gibbs sampler for the latent parameters of **model BNBR** with $G \times E$ can be implemented by sampling repeatedly from the following loop:

Step 1. Sample ω_{ijt} values from the Pólya-Gamma distribution in (4.16).
Step 2. Sample $L_{ijt} \sim CRT(y_{ijt}, r)$ from (4.21)
Step 3. Sample the scale parameter (r) from the gamma distribution in (4.20).
Step 4. Sample the location effects ($\boldsymbol{\beta}^*$) from the normal distribution in (4.15).

Step 5. Sample the random effects ($\boldsymbol{b}_h$) with $h = 1\,,2$, from the normal distribution in (4.17).

Step 6. Sample the variance effect ($\sigma^2_{b_h}$) with $h = 1\,,2$, from the scaled inverted χ^2 distribution in (4.18).

Step 7. Sample the variance effect ($\sigma^2_{\beta*}$) from the scaled inverted χ^2 distribution in (4.19).

Step 8. Return to step 1 or terminate when chain length is adequate to meet convergence diagnostics.

4.2.5.1 Simulation Examples

We performed a simulation study under linear predictor (Equation 4.1) with two scenarios (S1 and S2) to show the performance of the proposed Gibbs sampler for ordinal categorical and count phenotypes that takes into account $G \times E$. Scenario 1 has three environments ($I = 3$), 20 genotypes ($J = 20$), $\boldsymbol{G}_1 = \boldsymbol{I}_{20}$, $\boldsymbol{G}_2 = \boldsymbol{I}_3 \bigotimes \boldsymbol{G}_1$, and $\sigma^2_{b_1} = \sigma^2_{b_2} = 0.5$, with four different numbers of replicates of each genotype in each environment, $n_{ij} = 5\,,10\,,20$, and 40. Scenario 2 is equal to scenario 1, except that $\boldsymbol{G}_1 = 0.7\boldsymbol{I}_{20} + 0.3\boldsymbol{J}_{20}$, where $\boldsymbol{J}_{20}$ is a square matrix of ones of order 20 × 20; this second scenario was done with the intention of mimicking the correlation between lines of real data available in genomic selection. We computed 20,000 MCMC samples. Bayes estimates were computed using 10,000 samples because the first 10,000 were discarded as burn-in, and we performed 50 replications for each scenario in both models. Next we provide the details of the simulation under each model.

4.2.6 Model BLOR

Under the **model BLOR**, we simulated data from the following liability:

$$l_{ijt} = \boldsymbol{x}_i^T \boldsymbol{\beta} + b_{1j} + b_{2ij} + \varepsilon_{ijt}$$

since $i = 1\,,2\,,3$, $j = 1\,,\,\ldots\,,20$, and $t = 1\,,\,\ldots\,,n_{ij}$, $\boldsymbol{\beta}^{\boldsymbol{T}} = [-6, 5, 7]$, and the vectors $\boldsymbol{x}_i^T = [x_{i1}, x_{i2}, x_{i3}]$, where $x_{i1}\,, x_{i2}$, and x_{i3} are indicator variables that take the value of 1 if the observed environment i is 1, 2, or 3, respectively, and 0 otherwise. The threshold parameters used were γ_1=-3, $\gamma_2 = -1$, $\gamma_3 = 1$, and $\gamma_4 = 3$. The error terms (ε_{ijt}) were obtained from an $L(0, 1)$. Then the response variable was generated as

$$y_{ij} = \begin{cases} 1 & if - \infty < l_{ij} < \gamma_1, \\ 2 & if \quad \gamma_1 < l_{ij} < \gamma_2, \\ & \vdots \\ 5 & if \quad \gamma_4 < l_{ij} < \infty \end{cases}$$

The priors used were not informative for $f\left(\boldsymbol{\beta}\middle|\sigma_\beta^2\right) \sim N\left(\boldsymbol{\beta}_0^T = [0,0,0], \boldsymbol{I}_3 \times 10000\right)$, $f(\boldsymbol{b}_1|\sigma_{b_1}^2) \sim N(\mathbf{0}, \boldsymbol{G}_1\sigma_{b_1}^2)$, a normal prior distribution for $f(\boldsymbol{b}_2|\sigma_{b_2}^2) \sim N(\mathbf{0}, \boldsymbol{G}_2\sigma_{b_2}^2)$, with $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ as defined above for scenarios 1 and 2**,** and for the hyper-parameters for thresholds, we used $\gamma_{min} = -4$ and $\gamma_{max} = 4$. Results of this simulation study are given in Table 4.1.

4.2.7 Model BNBR

Also in this model, the priors used for both scenarios (S1 and S2) in the simulation study were not informative for all parameters: for $f\left(\boldsymbol{\beta}^*\middle|\sigma_{\beta*}^2\right) \sim N\left(\boldsymbol{\beta}_0^T = [0,0,0], \boldsymbol{I}_3 \times 10000\right)$, for $f(r) \sim G(0.001, 1/0.001)$, for $\sigma_{b_1}^2$ and $\sigma_{b_2}^2$ a $\sim\chi^{-2}(0.50002, 4.0002)$, while for $f(\boldsymbol{b}_1|\sigma_{b_1}^2) \sim N_J(\mathbf{0}, \boldsymbol{G}_1\sigma_{b_1}^2)$ and $f(\boldsymbol{b}_2|\sigma_{b_2}^2) \sim N_{IJ}(\mathbf{0}, \boldsymbol{G}_2\sigma_{b_2}^2)$, with $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ as defined above for scenarios S1 and S2. We report average estimates obtained by using the proposed Gibbs sampler along with standard deviations (SD), which are given in Table 4.2.

4.2.7.1 Model Implementation

The Gibbs samplers described above (for **model BLOR** and **model BNBR**) were implemented using the R-software (R Core Team 2015). For the implementation of the proposed models, we used MCMC with the Gibbs sampler algorithm (Gelfand and Smith 1990). We performed a total of 60,000 iterations, and 30,000 samples were used for inference since 30,000 were used as burn-in. Thinning of the chains was not applied following the suggestions of Geyer (1992), MacEachern and Berliner (1994), and Link and Eaton (Link and Eaton 2012). For the implementation of **model BLOR** for the real data sets, we used the following hyper-parameters $\nu_\beta = 3$, $S_\beta = 0.001$, $\boldsymbol{\beta}_0^T = [0,0,0]$, $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_3 \times 10000$, $\gamma_{\min} = -1000$, and $\gamma_{\max} = 1000$ for threshold parameters; these hyper-parameters lead weakly informative priors. For **model BNBR**, the priors used were $f\left(\boldsymbol{\beta}^*|\sigma_{\beta*}^2\right) \sim N_p(\boldsymbol{\beta}_0 = \mathbf{0}_3^T, \boldsymbol{I}_3 \times 10{,}000)$; $f(\boldsymbol{b}_1|\sigma_{b1}^2) \sim N_J(\mathbf{0}_{nb1}^T, \boldsymbol{G}_1\sigma_{b1}^2)$, where $\boldsymbol{G}_1$ is the GRM, that is, the covariance matrix of lines; $f\left(\sigma_{b1}^2\right) \sim \chi^{-2}(\nu_{b1} = 3, S_{b1} = 0.001)$; $f\left(\boldsymbol{b}_2|\sigma_{b2}^2\right) \sim N_{IJ}(\mathbf{0}_{nb2}^T, \boldsymbol{G}_2\sigma_{b2}^2)$, $\boldsymbol{G}_2$ is the covariance matrix that belong to the

Table 4.1 Average values (mean) and standard deviation (SD) of *model BLOR* with four sample sizes (n_{ij})

			$n_{ij} = 5$		$n_{ij} = 10$		$n_{ij} = 20$		$n_{ij} = 40$	
S	Parameter	True	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	β_0	−0.5	−0.235	1.633	−0.352	0.841	−0.526	0.387	−0.531	0.293
	β_1	1	1.033	0.324	0.952	0.298	1.00	0.287	1.00	0.239
	β_2	−0.5	−0.576	0.262	−0.529	0.305	−0.492	0.273	−0.478	0.273
	γ_1	−3	−2.858	1.567	−2.933	0.823	−3.032	0.337	−3.069	0.137
	γ_2	−1	−0.806	1.583	−0.911	0.796	−1.034	0.324	−1.052	0.145
	γ_3	1	1.255	1.579	1.109	0.781	0.959	0.316	0.954	0.147
	γ_4	3	3.251	1.607	3.145	0.779	2.990	0.344	2.962	0.177
	$\sigma_{b_1}^2$	0.5	0.587	0.203	0.601	0.202	0.548	0.137	0.568	0.155
	$\sigma_{b_2}^2$	0.5	0.542	0.182	0.528	0.162	0.506	0.138	0.5437	0.133
2	β_0	−0.5	−0.433	1.302	−0.506	0.893	−0.417	0.619	−0.625	0.585
	β_1	1	1.222	0.350	1.089	0.272	0.956	0.222	0.968	0.224
	β_2	−0.5	−0.459	0.346	−0.511	0.261	−0.501	0.239	−0.499	0.186
	γ_1	−3	−2.959	1.245	−3.089	0.744	−2.919	0.334	−3.047	0.212
	γ_2	−1	−0.880	1.242	−1.044	0.728	−0.8963	0.326	−1.035	0.205
	γ_3	1	1.2422	1.259	1.000	0.753	1.111	0.335	0.978	0.209
	γ_4	3	3.424	1.299	3.081	0.793	3.135	0.366	2.964	0.247
	$\sigma_{b_1}^2$	5	0.577	0.157	0.52	0.194	0.519	0.218	0.519	0.184
	$\sigma_{b_2}^2$	0.5	0.510	0.149	0.433	0.105	0.407	0.103	0.416	0.105

S denotes scenario

Table 4.2 Average values (mean) and standard deviation (SD) of *model BNBR* with four sample sizes (n_{ij})

			$n_{ij}=5$		$n_{ij}=10$		$n_{ij}=20$		$n_{ij}=40$	
S	Parameter	True	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	β_0	1.5	1.484	0.357	1.488	0.269	1.542	0.233	1.549	0.213
	β_1	−1	−0.981	0.256	−0.994	0.247	−1.075	0.250	−1.016	0.190
	β_2	1	0.997	0.270	0.985	0.223	0.994	0.268	0.949	0.223
	r	5	5.079	0.916	5.078	0.519	5.017	0.471	5.027	0.330
	$\sigma^2_{b_1}$	0.5	0.542	0.196	0.594	0.176	0.582	0.180	0.590	0.216
	$\sigma^2_{b_2}$	0.5	0.503	0.134	0.524	0.136	0.531	0.110	0.512	0.114
2	β_0	1.5	1.4808	0.5009	1.4596	0.5041	1.5611	0.6108	1.4723	0.4979
	β_1	−1	−1.0631	0.2348	−0.9975	0.2040	−1.008	0.2226	−1.025	0.1908
	β_2	1	0.9504	0.2356	1.0294	0.2167	0.9925	0.1954	0.9685	0.2018
	r	5	5.1030	0.8060	4.9901	0.5928	5.0367	0.3485	5.0275	0.2033
	$\sigma^2_{b_1}$	0.5	0.5422	0.1827	0.5650	0.2199	0.5785	0.1872	0.5296	0.1837
	$\sigma^2_{b_2}$	0.5	0.4987	0.1155	0.5084	0.1423	0.5302	0.1301	0.5123	0.1047

S denotes scenario

$G \times E$ term; $f\left(\sigma_{b2}^2\right) \sim \chi^{-2}(\nu_{b2} = 3, S_{b2} = 0.001)$; and $f(r) \sim G(a_0 = 0.01, 1/(b_0 = 0.01))$. These hyper-parameters produces weakly informative priors.

4.2.7.2 Measures of Predictive Performance

Scoring Rules

Here we present the scoring rules (scoring functions) for assessing prediction accuracy for ordinal categorical and count data. A scoring rule provides a summary measure for evaluating probabilistic predictions by assigning a numerical score based on the predictive distribution on the value or event that materializes (Garthwaite et al. 2005). Assuming that with both prediction models (for ordinal and count) we return a predictive distribution P for each observed outcome (y) in the data set, then we can use a scoring function to give a reward of $s(P, y)$ if the kth event occurs. We write $s(P, Q)$ for the expected value of $s(P, .)$ under Q. When a proper scoring rule is implemented, the highest expected reward is obtained by reporting the true probability distribution (Czado et al. 2009). Proper scoring rules always encourage the forecaster to be honest and maximize the expected reward. Usually the mean score is reported, which can be expressed as

$$S = \frac{1}{n} \sum_{k=1}^{n} s\left(P^{(k)}, y^{(k)}\right),$$

where $y^{(k)}$ and $P^{(k)}$ denote the kth observed outcome and the kth predictive distribution.

Proper Scoring Rules

A scoring rule is strictly proper if it is uniquely optimized by true probabilities. Suppose that our best prediction with our model is the predictive distribution Q. Let us assume that our prediction model has no incentive to predict any $P \neq Q$ and is encouraged to quote its true belief, $P = Q$, if

$$s(Q, Q) \leq s(P, Q),$$

with equality if, and only if, $P = Q$. A scoring rule with this characteristic is said to be *strictly proper* if $s(Q, Q) \leq s(P, Q)$ for all P and Q. Propriety is an essential property of a scoring rule that encourages coherent and honest predictions. Strict propriety ensures that both calibration and sharpness are being addressed (Czado et al. 2009), understanding sharpness as the concentration of the predictive distribution, and the shorter the sharper the predictions and the sharper the better, subject to calibration.

Examples of Proper Scoring Rules

The *logarithmic score* is defined as

$$logs(P, y) = -\log p_y$$

where p_y refers to the probability mass function at the observed outcomes (Czado et al. 2009). This is the only proper scoring rule that depends on the predictive distribution P through its probability mass (p_y). The quadratic score or Brier score is defined as

$$\text{qs}(P, y) = -2p_y + \parallel p \parallel^2, \tag{4.22}$$

where $\parallel p \parallel^2 = \sum_{k=1}^{\infty} p_k^2$, which can frequently be computed analytically for the Poisson and the negative binomial distribution. The *spherical score* is defined as

$$\text{sphs}(P, y) = -\frac{p_y}{\parallel p \parallel}$$

The ranked probability score (Czado et al. 2009) was originally proposed for ranked categorical data. It is defined as

$$\text{rps}(P, y) = \sum_{l=1}^{\infty} \{P_l - 1(y \le l)\}^2,$$

In practice it is not easy to choose a scoring rule, unless there is a unique and clearly defined underlying decision problem. However, it is always preferable to use a proper scoring rule instead of a classical measures of predictive performance that is not proper, such as the mean absolute error, mean square error of prediction, and mean squared Pearson residuals. However, in many situations, probabilistic predictions have multiple simultaneous uses, and it may be appropriate to use a variety of tools and scores, to take advantage of their different emphases and strengths (Czado et al. 2009). For example, Wecker (1989) used the Brier score (quadratic score) in the assessment of time series predictions of counts. Montesinos-López et al. (2015a, 2015b) also used the Brier score for assessing prediction accuracy for ordinal data in the context of genomic-enabled prediction.

4.2.7.3 Assessing Prediction Accuracy

Following Burgueño et al. (2012), we implemented the cross-validation scheme (CV2). Cross-validations mimic real situations that a breeder might face. This cross-validation scheme (CV2) mimics a situation where lines were evaluated in some environments but missing in others. In this case, information from relatives is used, and prediction assessment can benefit from borrowing information between lines within an environment and between lines across environments.

For this cross-validation, we used tenfold cross-validations; each time, ninefolds were used for training and onefold for testing. The training set was used to fit the model, and the validation set was used to evaluate the prediction accuracy of the proposed models. However, only the information of 10% of the lines in one environment was missing. Among the variety of methods for assessing prediction accuracy for ordinal categorical and count data, we used the Brier score for the GLS data set and the Spearman correlation (Cor) and mean square error of prediction (MSEP) for the FHB data set. These two criteria were used to obtain comparable predictions between the proposed models for counts (BNBR and Poisson) and models for normal and lognormal data. However, we need to point out that proper scoring rules (as the Brier score) should be preferred since classical measure of predictive performance as the MSEP and Cor are not proper and are not the best option for ordinal or counts data. For the GLS data sets with ordinal categorical data, the Brier score (Brier 1950) was computed as

$$BS = n^{-1} \sum_{k=1}^{n} \sum_{c=1}^{C} \left(\widehat{\pi}_{kc} - d_{kc} \right)^2 \tag{4.23}$$

where BS denotes the Brier score, $\widehat{\pi}_k$ denotes the predictive distribution derived from the estimated model for observation k, and d_{kc} takes a value of 1 if the ordinal categorical response observed for individual k falls into category c; otherwise, $d_{kc} = 0$. The range of BS in Eq. (4.23) for ordinal data is between 0 and 2. For this reason, we divided BS/2 to get the Brier score bounded between 0 and 1; lower scores imply better predictions. For count data, the Brier score was computed using Eq. (4.22) but computing p_y and $\|p\|^2$ depending on the model used, e.g., with the Poisson probability mass function if the **Pois model** was used for fitting the data, or the negative binomial probability mass function if the **BNBR model** was used. Here, the Brier score can be any real value, and the lower the value, the better the model.

The Brier score rule uses all the information contained in the predictive distribution, not just a small portion like the hit rate or the log likelihood score. Therefore, it is a reasonable choice for comparing ordinal categorical and count regression models, although there are other scoring rules that also have good properties.

4.3 Results

In the following sections, we investigate the performance of the proposed BLOR and BNBR models through a simulation study and with real data.

4.3.1 *Simulated Data Sets*

The only purpose of the simulations performed in this section is to show that the proposed methods work well in terms of parameter estimation.

4.3.1.1 Model BLOR

In Table 4.1, we report average estimates obtained by all methods, along with standard deviations (SD) for both scenarios (S1 and S2) under study. Table 4.1 shows that the bias in the estimation of the parameters is a little larger in S1 compared to S2 (which takes into account the GRM). Also, parameter β_1 is the parameter with larger bias (underestimated) when the sample size is $n_{ij} = 5$ and 10. Both variances ($\sigma^2_{b_1}$, $\sigma^2_{b_2}$) are overestimated under scenario 1, but only $\sigma^2_{b_1}$ is overestimated under scenario 2. From Table 4.1, it is clear that as the sample size increases, the average biases and SD decrease in all cases. This confirmed the consistent properties of all the estimates.

4.3.1.2 Model BNBR

Table 4.2 gives the results of the simulation study under both scenarios (S1 and S2). Again, the bias in the estimation of the parameters is a little larger in S1 compared to S2. Table 4.2 shows that parameter β_0 is the parameter with larger bias (underestimated). Both variances ($\sigma^2_{b_1}$, $\sigma^2_{b_2}$) are overestimated under scenario 1, but only $\sigma^2_{b_1}$ is overestimated under scenario 2. Also, with sample size $n_{ij} = 5$, the parameter r shows the larger SD; however, for larger sample sizes ($n_{ij} = 20, 40$), the SD are considerably reduced. In general, there is no large reduction in SD when the sample size increases from 5 to 10, 20, and 40, the exception being the estimation of r under both scenarios and the estimation of β_0 under scenario 1, where there is a large reduction of SD when the sample size increases. Even though the estimations under both models are not perfect, the proposed Gibbs samplers for ordinal and count data that take into account $G \times E$ do a good job for estimating the parameters since the estimates are close to the true values and with a SD of reasonable size. However, in both models, a more in-depth simulation study is required to ensure that these findings are valid for all possible scenarios.

4.3.2 *Real Data Sets*

Using the real data sets, we compared four scenarios for each model (Table 4.3). The table shows that scenarios S1 and S2 do not take into account interaction effects in the linear predictor, only main effects. Also, scenarios S1 and S3 do not use

Table 4.3 Scenarios proposed to fit the real data set with both models

	Main effects			Interaction effects	
Scenario	E	L	G	EL	EG
S1	X	X			
S2	X		X		
S3	X	X		X	
S4	X		X		X

E denotes environments, L lines, G lines with markers, EL interaction effect between environment and line, and EG interaction effect between environments and lines with markers

Table 4.4 Posterior average values (mean) and standard deviation (SD) for the GLS data set with *model BLOR* for each scenario given in Table 4.3

	S1		S2		S3		S4	
Parameter	Mean	SD	Mean	SD	Mean	SD	Mean	SD
β_1	−0.422	0.063	−0.217	0.069	−0.324	0.108	−0.415	0.093
β_2	0.167	0.056	0.376	0.064	0.524	0.099	0.342	0.089
β_3	−0.254	0.070	−0.047	0.075	−0.042	0.111	−0.273	0.094
γ_1	−1.625	0.053	−1.430	0.061	−2.084	0.076	−2.200	0.070
γ_2	−0.434	0.041	−0.227	0.056	−0.278	0.057	−0.447	0.039
γ_3	0.546	0.041	0.751	0.064	1.105	0.061	0.905	0.050
γ_4	1.355	0.047	1.566	0.070	2.182	0.072	1.964	0.070
$\sigma^2_{b_1}$	0.205	0.030	0.200	0.032	0.097	0.075	0.321	0.072
$\sigma^2_{b_2}$					1.488	0.144	1.525	0.164
postMeanLogLik	−3489		−3482		−2729		−2769	

marker information. These four scenarios are studied with the goal of investigating the gain in model fit and prediction ability taking into account the interaction effect and using the marker information available.

4.3.3 *Model BLOR for GLS Data Set*

Table 4.4 depicts the posterior mean and posterior standard deviation of the parameter estimates with the **model BLOR** using the GLS data set. The parameter estimates and posterior mean log likelihood (postMeanLogLik) of S1 and S2 are alike, as are the parameter estimates and posterior mean log likelihood (postMeanLogLik) of S3 and S4. The postMeanLogLik favors models S3 and S4.

In Table 4.5, we see that in Colombia, the best model in terms of prediction accuracy using the Brier score was S4, while the worst was S1. In Harare, the best model was S3 and the worst was S1. In Mexico, the best model was S2 followed by S4, while the worst model was S3. It is important to point out that prediction accuracy were best in Mexico and worst in Colombia. Because the differences between scenarios in each country are not large, it is not easy to discriminate

Table 4.5 Brier scores (mean and standard deviation (SD)) from **model BLOR** evaluated for validation samples for each scenario given in Table 4.3 for the GLS data set. Lower scores indicate better predictions

	Colombia		Harare		Mexico	
Scenario	Mean	SD	Mean	SD	Mean	SD
S1	0.417	0.020	0.388	0.016	0.360	0.044
S2	0.407	0.015	0.382	0.019	0.353	0.039
S3	0.404	0.023	0.363	0.015	0.367	0.035
S4	0.399	0.021	0.383	0.025	0.359	0.038

Table 4.6 Estimated fixed effects, variance components, and deviance information criteria (DIC) for *models BNBR and Pois* for the GLS data set

	S1		S2		S3		S4	
Parameter	Mean	SD	Mean	SD	Mean	SD	Mean	SD
	Model BNBR							
β_1^*	−4.002	0.071	−4.007	0.062	−4.001	0.055	−4.003	0.045
β_2^*	−3.828	0.074	−3.833	0.064	−3.822	0.057	−3.812	0.05
β_3^*	−3.961	0.075	−3.966	0.066	−3.957	0.059	−3.954	0.053
r	143.656	9.771	144.249	8.745	141.124	7.339	139.77	6.627
$\sigma^2_{b_1}$	0.032	0.004	0.034	0.005	0.033	0.005	0.037	0.005
$\sigma^2_{b_2}$	–	–	–	–	0.034	0.004	0.0368	0.005
DIC	8516.966	(3)	8484.151	(2)	8564.722	(4)	8462.392	(1)
	Model Pois							
β_1^*	−5.952	0.026	−5.95	0.022	−5.968	0.029	−5.982	0.024
β_2^*	−5.772	0.019	−5.772	0.015	−5.785	0.022	−5.781	0.016
β_3^*	−5.907	0.03	−5.909	0.029	−5.924	0.034	−5.925	0.029
r	1000	_	1000	–	1000	–	1000	–
$\sigma^2_{b_1}$	0.033	0.004	0.035	0.005	0.034	0.005	0.037	0.005
$\sigma^2_{b_2}$	–	–	–	–	0.034	0.004	0.037	0.004
DIC	8488.566	(3)	8457.023	(2)	8533.377	(4)	8427.161	(1)

() denotes the ranking of the four scenarios with the DIC for each model

between models; even though in two of the three locations, scenario S4 was identified as the best model for prediction.

4.3.4 Model BNBR for GLS Data Set

The posterior means, posterior SD of the scalar parameters, and deviance information criteria (DIC) for **model BNBR** and **model Pois** are given in Table 4.6. The table shows that the posterior means of the beta regression coefficients (β_1^*, β_2^*, and

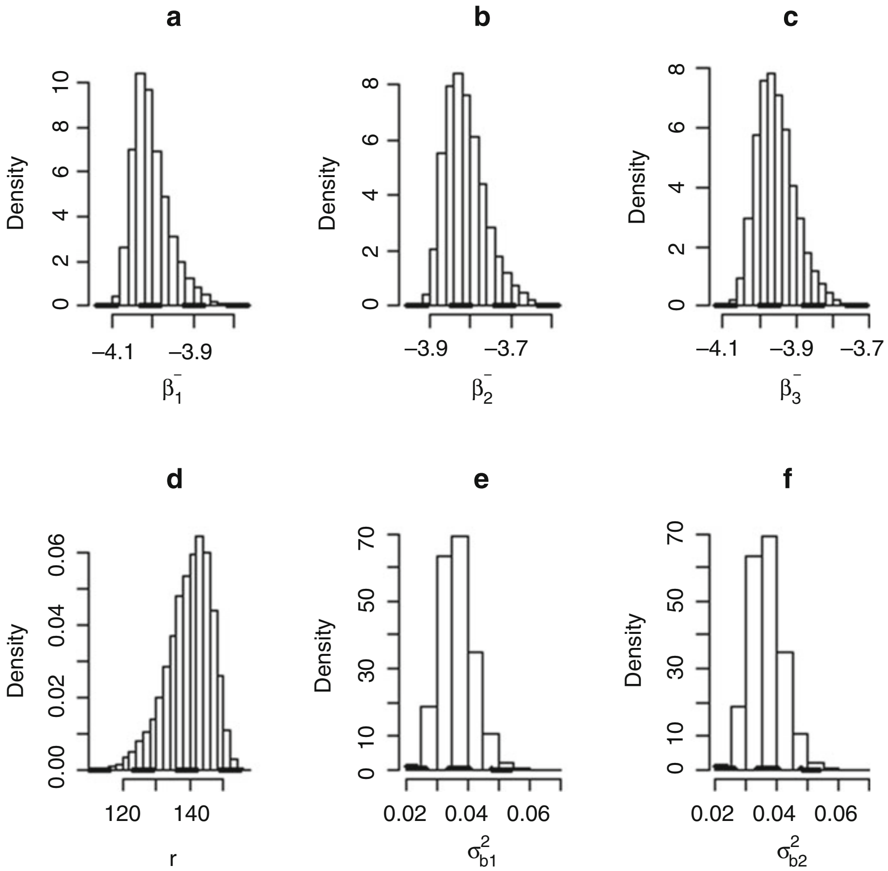


Fig. 4.2 Histogram representation of posterior distributions of scalar parameters for scenario S4 and **model BNBR** fitted with the whole data set GLS for (**a**) β_1^*, (**b**) β_2^*, (**c**) β_3^*, (**d**)r, (**e**) σ_{b1}^2, and (**f**) σ_{b2}^2 with priors superimposed as dashed lines at the bottom

β_3^*), variance components and scale parameter (r) were similar for the four proposed scenarios for **model BNBR** with the exception of scenarios S3 and S4, where the parameter r was lower, while the posterior SD of scenarios S3 and S4 were lower than those of scenarios S1 and S2. With regard to the DIC in **model BNBR,** the scenarios rank as follows: rank 1 for scenario S4, rank 2 for scenario S2, rank 3 for scenario S1, and rank 4 for scenario S3. Figure 4.2 shows a histogram representation of the posterior distributions for scalar parameters, and in all plots, the priors for each parameter in **model BNBR** are not informative.

Table 4.6 also shows that the posterior means and posterior SD of the beta regression coefficients(β_1^*, β_2^*, and β_3^*) and variance components for **model Pois** are similar between the four proposed scenarios. In terms of DIC, the scenarios rank as follows: rank 1 for scenario S4, rank 2 for scenario S2, rank 3 for scenario S1,

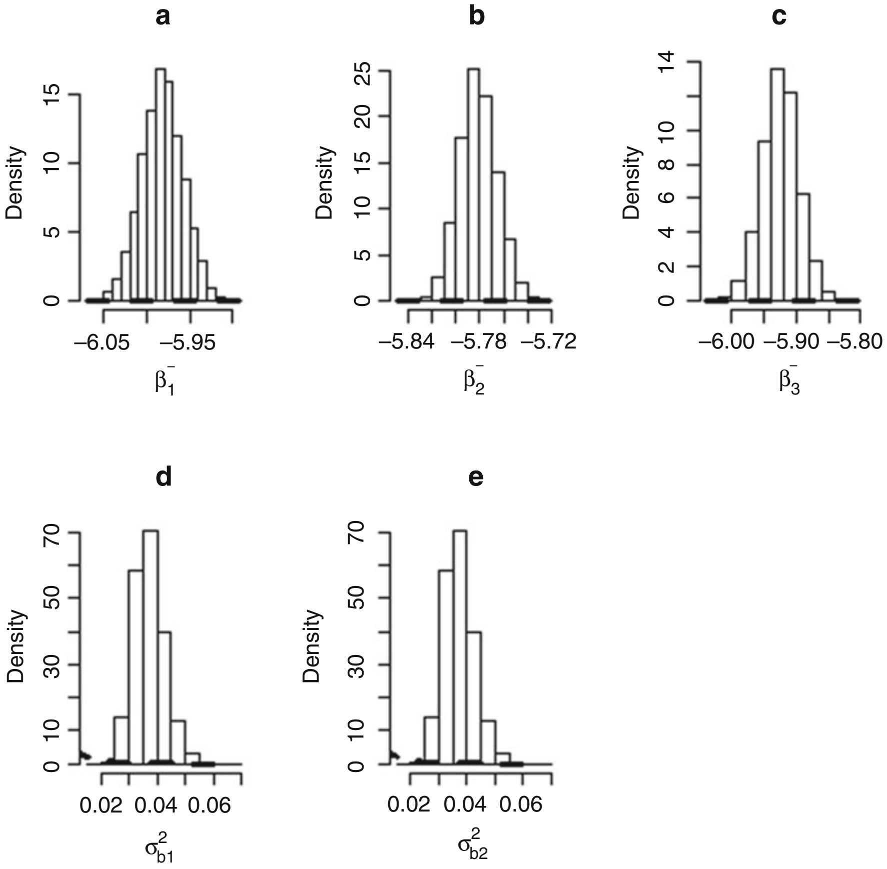


Fig. 4.3 Histogram representation of posterior distributions of scalar parameters for scenario 4 and **model Pois** fitted with the whole data set GLS for (**a**) β_1^*, (**b**) β_2^*, (**c**) β_3^*, (**d**) σ^2_{b1}, and (**e**) σ^2_{b2} with priors superimposed as dashed lines at the bottom

and rank 4 for scenario S3. Figure 4.3 also shows that the priors for each scalar parameter in **model Pois** are not informative.

Table 4.6 indicates that **model Pois** fits this real data set best, since comparing the DIC for these two models there is a clear superiority of **model Pois** over **model BNBR**. This means that **model Pois** is enough for this data set.

In Table 4.7 we present the mean and SD of the Brier scores resulting from the tenfold cross-validation performed. The Brier scores given in Table 4.7 were calculated using the testing set. According to the Brier scores, in **model BNBR,** the best model for prediction was S3 in the three environments. Under **model Pois,** the scenario with the best prediction accuracy was scenario S3 for Harare and Mexico, but scenario S2 in Colombia. It is not clear which model is the best (**model Pois** or **model BNBR)** since the Brier scores are very similar for both models. This may be due to the fact that the data are not really count, because we used the GLS data set that has the response variable as categorical ordinal.

Table 4.7 Brier scores (mean and standard deviation (SD)) from the *BNBR* and *Pois models* evaluated for validation samples for each scenario given in Table 4.3 for the GLS data set

	Colombia		Harare		Mexico	
Model	Mean	SD	Mean	SD	Mean	SD
	BNBR model					
S1	−0.224	0.013	−0.232	0.013	−0.219	0.023
S2	−0.226	0.010	−0.229	0.012	−0.219	0.020
S3	−**0.233**	0.011	−**0.242**	0.011	−**0.233**	0.014
S4	−0.221	0.008	−0.225	0.011	−0.219	0.017
	Pois model					
S1	−0.228	0.019	−0.216	0.027	−0.264	0.028
S2	−**0.231**	0.018	−0.218	0.020	−0.248	0.028
S3	−0.194	0.174	−**0.228**	0.036	−**0.281**	0.025
S4	−0.208	0.015	−0.218	0.016	−0.234	0.025

Lower scores indicate better predictions

4.3.5 Model BNBR for FHB Data Set

These results were taken from the paper of Montesinos-López et al. (2016). For this data set (FHB data) was implemented with the following linear predictor: $\eta_{ijk}=E_i+R(E)_{ik}+g_j+gE_{ij}$ with $i=1,\ldots,I;j=1,2,\ldots,J,k=1,\ldots,K,$ $t=1,2,\ldots,n_{ijk}$, where E_i represents environment i, $R(E)_{ik}$ represents the effect of block k within environment i, g_j is the marker effect of genotype j, and gE_{ij} is the interaction between markers and the environment. This predictor is very similar to that given in Eq. (4.1) with the term $R(E)_{ik}$ added. We assume that y_{ijkt} represents the count response for the tth replication of the jth line in the kth block in the ith environment. Four models were implemented using the linear predictor given above: model BNBR, model Poisson, normal, and lognormal (LN). Also for each model, the four scenarios given in Table 4.3 were studied. These models were implemented under a Bayesian approach (these Bayesian models were implemented through Gibbs sampler since all full conditionals were derived analytically, for the BNBR and Poisson model the Pólya-Gamma augmentation approach explained above was used, see Montesinos-López et al. Montesinos-López et al. 2015a, 2015b, 2015c and Montesinos-López et al. 2016 for details of all derivations). The Gibbs samplers for the four models were implemented in R Core Team (2015) using 60,000 iterations with a burn-in of 30,000, so that 30,000 samples were used for inference. The prediction accuracy of the proposed models were evaluated with the Spearman correlation (Cor) and the mean square error of prediction (MSEP), both calculated using the observed and predicted response variables of the validation set resulting of a tenfold cross-validation implemented for the four models and four scenarios. In this example we used the Spearman correlation and the MSEP to compare the results with models normal and lognormal.

In Table 4.8 we can see that the ranking of scenarios for the BNBR model were as follows: in Batan 2012, 1 for S4, 2 for S3, 3 for S1, and 4 for S2. In Batan 2014, the ranking was 1 for S4, 2 for S3, and 3 for S1 and S2. In Ecuador 2014, the ranking was 1 for S3, 2 for S2, 3 for S1, and 4 for S4. With the MSEP, the ranking for **model BNBR** in Batan 2012 was 1 for S3, 2 for S4, and 3 for S1 and S2. In Batan 2014, the ranking was 1 for S2, 2 for S1, 3 for S3, and 4 for S4. In Ecuador 2014, the ranking in terms of MSEP was 1 for S3, 2 for S2, 3 for S4, and 4 for S1. Under **model Pois,** the ranking of the four scenarios in each locality was exactly the same as the ranking reported for **model BNBR**. For **model normal** in terms of the Spearman correlation, S1 was the best in prediction accuracy in Batan 2012, while S4 was the worst in all three locations. In terms of MSEP, the best scenario was S3 in Batan 2012 and Ecuador 2014, and the worst was S4 in Batan 2014 and Ecuador 2014. For **model LN** in terms of the Spearman correlation, the best scenarios were scenarios S1, S2, and S3, and the worst was S4 in Batan 2012. In Batan 2014, the best scenario was S1, then scenario S3, and the worst was scenario S4. In Ecuador 2014, the best scenario was scenario S1 and S3, then S2 and S4. In terms of MSEP for Batan 2012, the best scenario was S3, then S1 and S2, and the worst was S4. In Batan 2014, the best scenario was S1, then S2, and the worst was scenario S4. Finally, in Ecuador 2014, the best scenario was S3, then S2, and the worst was scenario S1.

Table 4.9 gives the average of the ranks of the two posterior predictive checks (Cor and MSEP) that were used. Since we are comparing four scenarios for each model, the values of the ranks range from 1 to 4, and the lower the values, the better the scenario. For ties we assigned the average of the ranges that would have been assigned had there been no ties. Table 4.9 shows that the best scenarios were scenarios S3 and S4 under **models BNBR** and **Pois** in Batan 2012. In Batan 2014, under **models BNBR** and **Pois,** the best scenario was S2, while in Ecuador 2014, the best scenario was S3**.** Under **model normal**, the best scenario was S1 in Batan 2014, S1 and S3 in Ecuador 2014, while in Batan 2012, the best scenarios were S2 and S3. Finally, under **model LN,** the best scenario was S3 in Ecuador 2014, S3 in Batan 2012, and S1 in Batan 2014. Then according with results of Tables 4.8 and 4.9, the best models in terms of prediction accuracy are **models BNBR** and **Pois**, since they had better predictions in the validation set based on both the posterior predictive checks (Cor and MSEP) implemented. Also, we observed that in **models BNBR** and **Pois,** taking into account $G \times E$ considerably increased the prediction accuracy, which was expected since there is enough scientific evidence that including $G \times E$ interaction improves prediction accuracy. However, to use these models correctly, it is important to first understand the types of data we have before deciding on the modeling approach to be used.

Table 4.8 Estimated posterior predictive checks with cross-validation for *models BNBR*, *Pois*, *normal*, and *LN* for the FHB data set

		Batan 2012		Batan 2014		Ecuador 2014	
Scenario		Cor	MSEP	Cor	MSEP	Cor	MSEP
		Model BNBR					
S1	Mean	0.43 (3)	0.98 (3.5)	0.43 (3.5)	1.39 (2)	0.18 (3)	11.733 (4)
	SD	0.33	0.72	0.33	1.35	0.40	9.471
S2	Mean	0.42 (4)	0.98 (3.5)	0.43 (3.5)	1.38 (1)	0.20 (2)	11.222 (2)
	SD	0.33	0.72	0.33	1.36	0.37	8.614
S3	Mean	0.54 (2)	0.49 (1)	0.52 (2)	1.48 (3)	0.22 (1)	8.645 (1)
	SD	0.28	0.38	0.29	2.32	0.39	5.688
S4	Mean	0.56 (1)	0.61 (2)	0.56 (1)	1.85 (4)	0.12 (4)	11.343 (3)
	SD	0.24	0.44	0.22	2.68	0.41	8.154
		Model Pois					
S1	Mean	0.43 (3)	0.98 (3.5)	0.43 (3.5)	1.39 (2)	0.18 (3)	11.733 (4)
	SD	0.33	0.72	0.33	1.35	0.40	9.471
S2	Mean	0.42 (4)	0.98 (3.5)	0.43 (3.5)	1.38 (1)	0.20 (2)	11.222 (2)
	SD	0.33	0.72	0.33	1.36	0.37	8.614
S3	Mean	0.54 (2)	0.48 (1)	0.52 (2)	1.48 (3)	0.22 (1)	8.645 (1)
	SD	0.28	0.38	0.29	2.32	0.39	5.688
S4	Mean	0.56 (1)	0.61 (2)	0.56 (1)	1.85 (4)	0.12 (4)	11.343 (3)
	SD	0.24	0.44	0.22	2.68	0.41	8.154
			Model normal				
S1	Mean	0.36(1)	1.10 (4)	0.37 (1.5)	1.79 (1)	0.15 (1.5)	7.425 (2)
	SD	0.28	0.88	0.39	1.70	0.32	4.151
S2	Mean	0.34 (2)	0.99 (2)	0.33 (3)	2.01 (3)	0.07 (3)	7.454 (3)
	SD	0.33	0.65	0.44	2.46	0.33	4.339
S3	Mean	0.33 (3)	0.81 (1)	0.37 (1.5)	1.96 (2)	0.15 (1.5)	7.318 (1)
	SD	0.30	0.46	0.40	2.99	0.29	4.159
S4	Mean	0.27 (4)	1.03 (3)	0.24 (4)	2.37 (4)	0.04 (4)	8.482 (4)
	SD	0.34	0.73	0.45	3.42	0.24	4.326
		Model LN					
S1	Mean	0.51 (2)	0.66 (2.5)	0.46 (1)	1.60 (1)	0.15 (1.5)	8.10 (4)
	SD	0.21	0.42	0.31	2.35	0.38	5.11
S2	Mean	0.51 (2)	0.66 (2.5)	0.43 (3.5)	1.78 (2)	0.09 (3.5)	7.82 (2)
	SD	0.22	0.39	0.35	2.82	0.46	5.31
S3	Mean	0.51 (2)	0.64 (1)	0.45 (2)	1.871 (3)	0.15 (1.5)	7.76 (1)
	SD	0.21	0.45	0.31	3.16	0.37	5.21
S4	Mean	0.43 (4)	0.72 (4)	0.43 (3.5)	1.95 (4)	0.09 (3.5)	8.04(3)
	SD	0.25	0.42	0.33	3.15	0.41	5.18

The numbers in () denote the ranking of the four scenarios set for each posterior predictive check (extracted from Montesinos-López et al. 2016)

Table 4.9 Rank averages for the four scenarios resulting from the tenfold cross-validation implemented for the FHB data set

Scenarios	Batan 2012	Batan 2014	Ecuador 2014	Batan 2012	Batan 2014	Ecuador 2014
Model BNBR				**Model normal**		
S1	3.25	2.75	3.5	2.5	1.25	1.75
S2	3.75	2.25	2	2	3	3
S3	1.5	2.5	1	2	1.75	1.75
S4	1.5	2.5	3.5	3.5	4	4
Model Pois				**Model LN**		
S1	3.25	2.75	3.5	2.25	1	2.75
S2	3.75	2.25	2	2.25	2.75	2.75
S3	1.5	2.5	1	1.5	2.5	1.25
S4	1.5	2.5	3.5	4	3.75	3.25

Each average was obtained as the mean of the rankings given in Table 4.8 for the two posterior predictive checks (Cor and MSEP) in each scenario (extracted from Montesinos-López et al. 2016)

4.4 Conclusions

Generalized linear mixed models (GLMMs) are considered to be one of the major methodological developments of the second half of the last century. The main factor contributing to their wide applicability over the last 30 years or so is their flexibility, because they can be applied to different types of data (Berridge and Crouchley 2011), including continuous interval/scale, categorical (including binary and ordinal) data, count data, beta data, and others. Each member of the GLMMs family is appropriate for a specific type of data (Berridge and Crouchley 2011). However, GLMMs for non-normal data are scarce in the context of genomic-enabled prediction, since most of the models developed so far are linear mixed models (mixed models for Gaussian data). For these reasons, we believe that developing specific methods for categorical ordinal and count data for genomic-enabled prediction can help to improve the selection of candidate genotypes early when the phenotypes are ordinal and counts. Because using transformation to approximate the ordinal data and counts to normality or assuming that these types of data are normally distributed frequently produces poor parameter estimates and lower power, parameter interpretation is more difficult when transformation is used (Stroup 2015). However, in genomic selection, phenotypic data (dependent variable) are currently not taken into account before deciding on the modeling approach to be used, mainly due to the lack of genomic-enabled prediction models for non-normal phenotypes. Although our proposed Bayesian regression models are only for categorical ordinal and count data, they help fill this lack of genomic-enabled prediction models for non-normal data.

This chapter presents recent advances in models for whole genome prediction for ordinal categorical and count data. These models were built using the Pólya-Gamma data augmentation approach of Scott and Pillow (2013), which produces a

Gibbs sampler with full conditional distributions similar to that of the Bayesian probit ordinal regression (BPOR) model of Albert and Chib (1993). The proposed Bayesian logistic ordinal regression (model BLOR) is reduced to the BPOR model of Albert and Chib (1993) when the sampled values, ω_{ijt}, from the Pólya-Gamma distribution in Eq. (4.6) are set to 1. This is an advantage because researchers can perform an exact logistic or probit ordinal regression with the proposed model without having to do approximations to perform a logistic ordinal regression. The performance of the proposed model BLOR without interaction was compared to the BPOR model using the approximation ($logit(u) = (1.75)\Phi^{-1}(u)$) in a small simulation study and with real data sets using a 4- and 5-point ordinal scale by Montesinos-Lopez et al. (Montesinos-López et al. 2015b). They found that the estimation of parameters using the approximation $logit(u) = (1.75)\Phi^{-1}(u)$ produces a considerable amount of bias and can give rise to wrong conclusions in association studies. However, they observed no differences between the two models in terms of prediction ability with two real data sets. For this reason, the proposed BLOR model is a viable alternative for analyzing ordinal data since it is more robust for dealing with outlying data. This is because the logistic distribution has heavier tails and provides regression coefficients that are more interpretable due to their connection to odds ratios (Zucknick and Richardson 2014). This last advantage does not make sense when $p \gg n$,since the main driving force in Bayesian models in the case of $p \gg n$ is the prior and not the data (Gianola 2013). Even with this restriction, the proposed BLOR model unifies logistic and probit ordinal regression under a Bayesian framework and is a useful alternative for genomic-enabled prediction of ordinal categorical trials where available data sets have a larger number of parameters to estimate than observations.

The proposed Bayesian method for count data describes the work done by Montesinos-López et al. (2016), which extended the work of Montesinos-López et al. (2015c) to incorporate the G × E term. Modeling G × E for categorical ordinal and count data in the context of genomic-enabled prediction plays a central role in plant breeding for the selection of candidate genotypes that present high adaptation to a wide range of environmental conditions including local conditions. Also, incorporating G × E helps to predict yet-to-be observed phenotypes when the relative performance of genotypes varies across environments. To the best of our knowledge, this is the first work on genomic-enabled prediction that uses the NB and Poisson distributions with G × E.

It should also be noted that to use these models correctly, it is important to first understand the types of data being analyzed before deciding on the modeling approach to be employed. If the phenotypic data are normally distributed, the linear mixed models for genomic-enabled prediction developed so far for Gaussian phenotypes should be used. If the phenotypic data are binary or categorical ordinal, the methods proposed by Montesinos-López et al. (2015a, 2015b) and their extensions given in this chapter with the logit link should be preferred. If the phenotypic data are counts (number of panicles per plant, number of seeds per panicle, weed count per plot, etc.) and the counts are small, the models developed in this study (BNBR and Pois models) and those proposed by Montesinos-López et al. (2015c,

2016) are the best option, since they have more advantages over the conventional linear mixed models with Gaussian response, as was observed when we applied them to the real data set. However, Montesinos-López et al. (2015c, 2016) also found that when the count response is log transformed, the prediction accuracies are better than when using the counts as if they were normally distributed.

Finally, it is important to extend the proposed methods of Montesinos-López et al. (2015a, 2015b, 2015c, 2016) developed under the work of Scott and Pillow (2013) for ordered categorical responses and count data for multiple traits. Our methods are elegant, easy to implement, and produce a unified Gibbs sampler framework useful for both types of phenotypic responses. This is important because, of all the computational intensive methods for fitting complex multilevel models, the Gibbs sampler is the most popular due to its simplicity and ability to effectively generate samples from high-dimensional probability distributions (Park and van Dyk 2009). For this reason, we believe these methods are appealing alternatives for plant and animal researchers. Both models can be easily extended to take into account epistatic effects for the joint modeling of multiple traits, which, as is well documented, can increase the prediction accuracy of the models.

Acknowledgments We would like to thank all researchers in CIMMYT's Global Maize Program (GMP) and Global Wheat Program (GWP), as well as the national program researchers who generated the data used in this and other studies.

Appendix A: Derivation of Full Conditional Distributions for Model BLOR

Liabilities and ω_{ijt}. The fully conditional posterior distribution of liability l_{ijt} is

$$P(\boldsymbol{l}|ELSE) \propto P(\boldsymbol{l}|\,\boldsymbol{\beta},\boldsymbol{b})P(\boldsymbol{y}|\boldsymbol{l},\boldsymbol{\gamma}\ )$$

$$\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}} f(l_{ijt})\sum_{c=1}^{C} I(y_{ijt}=c)I(\gamma_{c-1}<l_{ijt}<\gamma_c)$$

$$\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}} \frac{\exp(-l_{ijt}+\boldsymbol{x}_i^T\boldsymbol{\beta}+b_{1j}+b_{2ij})}{[1+\exp(-l_{ijt}+\boldsymbol{x}_i^T\boldsymbol{\beta}+b_{1j}+b_{2ij})]^2}\sum_{c=1}^{C} I(y_{ijt}=c)I(\gamma_{c-1}<l_{ijt}<\gamma_c)$$

$$\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}} 2^{-2}\int_0^\infty \exp\left[-\frac{\omega_{ijt}(-l_{iljt}+\boldsymbol{x}_i^T\boldsymbol{\beta}+b_{1j}+b_{2ij})^2}{2}\right]P(\omega_{ijt};b=2,d=0)$$

$$\times d\omega_{ijt}\sum_{c=1}^{C} I(y_{ijt}=c)I(\gamma_{c-1}<l_{ijt}<\gamma_c)$$

The last inequality was obtained using a technique called the Pólya-Gamma method (Scott and Pillow 2013), which is useful when working with logistic likelihoods, and has the form

$$\frac{(e^{\psi})^{a}}{(1+e^{\psi})^{b}}=2^{-b}e^{\kappa\psi}\int_{0}^{\infty}e^{-\frac{\omega\psi^{2}}{2}}P(\omega;b,0)d\omega$$

where $\kappa = a - b/2$ and $P(\omega; b, d = 0)$ denotes the density of the random variable $\omega \sim PG(b, d = 0)$, where $PG(b, d)$ denotes a Pólya-Gamma distribution l_{ijt} with parameters b and d and density

$$P(\omega;b,d)=\left\{cosh^{b}\left(\frac{d}{2}\right)\right\}$$
$$\frac{2^{b-1}}{\Gamma(b)}\sum_{n=0}^{\infty}(-1)^{n}\frac{\Gamma(n+b)(2n+b)}{\Gamma(n+1)\sqrt{2\pi\omega^{3}}}\exp\left(-\frac{(2n+b)^{2}}{8\omega}-\frac{d^{2}}{2}\omega\right),$$

where cosh denotes the hyperbolic cosine.

Then the joint posterior distribution of l_{ijt} and ω_{ijt} is equal to

$$P(\boldsymbol{l},\boldsymbol{\omega}|ELSE)\propto\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}}2^{-2}\exp\left[-\frac{\omega_{ij}(-l_{ijt}+\boldsymbol{x}_{i}^{T}\boldsymbol{\beta}+b_{1j}+b_{2ij})^{2}}{2}\right]P(\omega_{ijt};2,0)$$
$$\times\sum_{c=1}^{C}I(y_{ijt}=c)I(\gamma_{c-1}<l_{ijt}<\gamma_{c})$$

Therefore, the fully conditional posterior distribution of liability l_{ijt} is a truncated normal distribution and its density is

$$f(l_{ijt}|ELSE)$$
$$=\frac{\phi\left(\sqrt{\omega_{ijt}}(l_{ijt}-\boldsymbol{x}_{i}^{T}\boldsymbol{\beta}-b_{1j}-b_{2ij})\right)}{\Phi(\sqrt{\omega_{ijt}}(\gamma_{c}-\boldsymbol{x}_{i}^{T}\boldsymbol{\beta}-b_{1j}-b_{2ij}))-\Phi\left(\sqrt{\omega_{ijt}}(\gamma_{c-1}-\boldsymbol{x}_{i}^{T}\boldsymbol{\beta}-\boldsymbol{b}_{1\boldsymbol{j}}-\boldsymbol{b}_{2\boldsymbol{ij}})\right)}$$

For simplicity, *ELSE* is the data and the parameters, except for the one in question. ϕ and Φ are the density and distribution function of a standard normal random variable and the fully conditional posterior distribution l_{ijt} of ω_{ijt} is

$$f(\omega_{ijt}|ELSE)\propto 2^{-2}\exp\left[-\frac{\omega_{ijt}(-l_{ijt}+\boldsymbol{x}_{i}^{T}\boldsymbol{\beta}+b_{1j}+b_{2ij})^{2}}{2}\right]P(\omega_{ijt};2,0)$$
$$\propto\exp\left[-\frac{\omega_{ijt}(-l_{ijt}+\boldsymbol{x}_{i}^{T}\boldsymbol{\beta}+b_{1j}+b_{2ij})^{2}}{2}\right]P(\omega_{ijt};2,0)$$

From here and from Eq. (4.5) of Polson et al. (2013), we get that

$$f\left(\omega_{ijt}|ELSE\right) \sim PG\left(2, -l_{ijt} + \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta} + b_{1j} + b_{2ij}\right)$$

Regression Coefficients ($\boldsymbol{\beta}$) First note that the fully conditional posterior of $\boldsymbol{l}\,,\boldsymbol{\beta}\,,\boldsymbol{\omega}$ is

$$P(\boldsymbol{l},\boldsymbol{\beta},\boldsymbol{\omega}|ELSE) \propto P(\boldsymbol{l}|\,\boldsymbol{\beta},\boldsymbol{b}_1,\boldsymbol{b}_2)P(\boldsymbol{y}|\boldsymbol{l},\boldsymbol{\gamma}\;)P(\boldsymbol{\omega})P\left(\boldsymbol{\beta}\middle|\boldsymbol{\sigma}_{\boldsymbol{\beta}}^2\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(-\boldsymbol{l} + \boldsymbol{X\beta} + \sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h\right)^T\boldsymbol{D}_\omega\left(-\boldsymbol{l} + \boldsymbol{X\beta} + \sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h\right)\right)P(\boldsymbol{\omega})P\left(\boldsymbol{\beta}\middle|\boldsymbol{\sigma}_{\boldsymbol{\beta}}^2\right)$$

where $P(\boldsymbol{\omega}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}}P\left(\omega_{ijt};2,0\right)$. Then, the full conditional posterior distribution of $\boldsymbol{\beta}$ is

$$P(\boldsymbol{\beta}|ELSE)$$

$$\times\propto\exp\left(-\frac{1}{2}\left(-\boldsymbol{l}+\boldsymbol{X\beta}+\sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h\right)^T\boldsymbol{D}_\omega\left(-\boldsymbol{l}+\boldsymbol{X\beta}+\sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h\right)\right.$$

$$-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\sigma}_{\boldsymbol{\beta}}^{-2}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}_0))$$

$$\propto\exp\left(-\frac{1}{2}[\boldsymbol{\beta}^T(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}+\boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{X})\boldsymbol{\beta}-2\left(\boldsymbol{\Sigma}_0^{-1}\sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{\beta}_0-\boldsymbol{X}^T\boldsymbol{D}_\omega(\sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h)+\boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{l}\right)^T\boldsymbol{\beta}]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\right)^T\tilde{\boldsymbol{\Sigma}}_0^{-1}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\right)\right]\right)$$

where $\tilde{\boldsymbol{\Sigma}}_0 = (\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \boldsymbol{X}^T\boldsymbol{D}_{\boldsymbol{\omega}}\boldsymbol{X})^{-1}$, $\tilde{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\Sigma}}_0\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \boldsymbol{X}^T\boldsymbol{D}_\omega(\sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h) + \boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{l}\right)$. It is important to point out that if we use a prior for $\boldsymbol{\beta}\propto$ Constant (improper uniform distribution), then in $\tilde{\boldsymbol{\Sigma}}_0$ and $\tilde{\boldsymbol{\beta}}_0$ we need to make **0** the term $\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}$. Finally, the full conditional) posterior of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\boldsymbol{ELSE} \sim N_I\left(\tilde{\boldsymbol{\beta}}_0,\tilde{\boldsymbol{\Sigma}}_0\right)$$

Polygenic effects ($\boldsymbol{b}_h$) Now the full conditional posterior of $\boldsymbol{b}_h$ is given as

$$\begin{aligned} & L(\boldsymbol{b}_h|ELSE) \\ & \propto \exp\left(-\frac{1}{2}(-\boldsymbol{l}+\boldsymbol{X\beta}+\sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h)^T\boldsymbol{D}_\omega(-\boldsymbol{l}+\boldsymbol{X\beta}+\sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h)\right)P(\boldsymbol{b}_{\boldsymbol{h}}|\sigma_{b_h}^2) \\ & \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{b}_h^T\left(\sigma_b^{-2}\boldsymbol{G}^{-1}+\boldsymbol{Z}_h^T\boldsymbol{D}_\omega\boldsymbol{Z}_h\right)\boldsymbol{b}_h-2\left(\boldsymbol{Z}_h^T\boldsymbol{D}_\omega\boldsymbol{l}-\boldsymbol{Z}_h^T\boldsymbol{D}_\omega\boldsymbol{X\beta}\right)^T\boldsymbol{b}_h\right]\right\} \\ & \qquad\qquad \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{b}_h-\tilde{\boldsymbol{b}}_h\right)^T\boldsymbol{F}_h^{-1}\left(\boldsymbol{b}_h-\tilde{\boldsymbol{b}}_h\right)\right\} \end{aligned}$$

This implies that the full conditional posterior of $\boldsymbol{b}_h$ is

$$f(\boldsymbol{b}_h|ELSE) \sim N\left(\widetilde{\boldsymbol{b}}_{\boldsymbol{h}}=\boldsymbol{F}_{\boldsymbol{h}}(\boldsymbol{Z}_h^T\boldsymbol{D}_\omega\boldsymbol{l}-\boldsymbol{Z}_h^T\boldsymbol{D}_\omega\boldsymbol{\eta}^h),\boldsymbol{F}_{\boldsymbol{h}}=(\sigma_{b_h}^{-2}\boldsymbol{G}_h^{-1}+\boldsymbol{Z}_h^T\boldsymbol{D}_\omega\boldsymbol{Z}_h^T)^{-1}\right)$$

with $h=1\,,2$, $\boldsymbol{\eta}^1=\boldsymbol{X\beta}+\boldsymbol{Z}_2\boldsymbol{b}_2$ and $\boldsymbol{\eta}^2=\boldsymbol{X\beta}+\boldsymbol{Z}_1\boldsymbol{b}_1$.

Variance of polygenic effects $(\sigma_{b_h}^2)$. Next, the conditional distribution of $\sigma_{b_h}^2$ is obtained. If $\sigma_{b_h}^2 \sim \chi^{-2}(\nu_h, S_h)(shape\ \ and\ scale)$, then

$$P(\sigma_{b_h}^2|ELSE) \propto \frac{1}{(\sigma_{b_h}^2)^{\frac{\nu_h+n_h}{2}+1}}\exp\left(-\frac{\boldsymbol{b}_h^T\boldsymbol{G}_h^{-1}\boldsymbol{b}_{\boldsymbol{h}h}+\nu_hS_h}{2\sigma_{b_h}^2}\right)$$

This is the kernel of the scaled inverted χ^2 distribution; therefore, the full conditional posterior is

$$f(\sigma_{b_h}^2|ELSE) \sim \chi^{-2}\left(\widetilde{\nu}_{\boldsymbol{h}}=\nu_h+n_h, \widetilde{S}_{\boldsymbol{b}}=(\boldsymbol{b}_h^T\boldsymbol{G}_h^{-1}\boldsymbol{b}_h+\nu_hS_h)/\nu_b+n_h)\right.$$

Threshold effects ($\boldsymbol{\gamma}$) The density of the full conditional posterior distribution of the cth threshold, γ_c, is

$$\begin{aligned} P(\boldsymbol{\gamma}|\boldsymbol{ELSE}) &\propto P(\boldsymbol{y}|\boldsymbol{l},\boldsymbol{\gamma})P(\boldsymbol{\gamma}) \\ &\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}}\sum_{c=1}^{C}I(y_{ijt}=c)I(\gamma_{c-1}<l_{ijt}<\gamma_c)I(\boldsymbol{\gamma}\in\boldsymbol{T}) \end{aligned} \tag{4.A.1}$$

If Eq. (4.A.1) is seen as a function of γ_c, it is evident that the value of γ_c must be larger than all the $l_{ijt}|y_{ijt}=c$ and smaller than all the $l_{ijt}|y_{ijt}=c+1$. Hence, as a function of γ_c, Eq. (4.A.1) leads to the uniform density

$$P(\gamma_c|ELSE) = \frac{1}{\min\left(l_{ijt}\middle|y_{ijt} = c+1\right) - \max\left(l_{ijt}\middle|y_{ijt} = c\right)} I(\boldsymbol{\gamma} \in \boldsymbol{T}) \tag{4.A.2}$$

Equation (4.A.2) corresponds to a uniform distribution on the interval [min {min $(l_{ijt}|y_{ijt} = c+1), \gamma_{c+1}, \gamma_{max}$ }, max{max$(l_{ijt}|y_{ijt} = c), \gamma_{c-1}, \gamma_{min}$}] (Albert and Chib 1993; Sorensen et al. 1995).

Variance of location effects (σ_β^2) If we give $\sigma_\beta^2 \sim \chi^{-2}(\nu_\beta, S_\beta)(shape\ \ and\ scale)$, then

$$P\left(\sigma_\beta^2|ELSE\right) \propto P\left(\sigma_\beta^2\right)P\left(\boldsymbol{\beta}|\ \sigma_\beta^2\right) = \frac{1}{\left(\sigma_\beta^2\right)^{\frac{\nu_\beta}{2}+1}} \exp\left(-\frac{\nu_\beta S_\beta}{2\sigma_\beta^2}\right) P\left(\boldsymbol{\beta}|\ \sigma_\beta^2\right)$$

$$\propto \frac{1}{\left(\sigma_\beta^2\right)^{\frac{\nu_\beta+I}{2}+1}} \exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \nu_\beta S_\beta}{2\sigma_\beta^2}\right)$$

This is the kernel of the scaled inverted χ^2 distribution; therefore, the full conditional posterior is

$$\sigma_\beta^2 \mid ELSE \sim \chi^{-2}\left(\tilde{\nu}_\beta = \nu_\beta + I, \tilde{S}_\beta = \left[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T \Sigma_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \nu_\beta S_\beta\right]/\nu_\beta + I\right)$$

Appendix B: Derivation of Full Conditional Distributions for Model BNBR

Full conditional for $\boldsymbol{\beta}^*$

$$f(\boldsymbol{\beta}^*|ELSE) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}} \Pr\left(Y_{ijt} = y_{ijt}|\boldsymbol{x}_i^T, r, \omega_{ijt}, b_{1j}, b_{2ij}\right) f(\boldsymbol{\beta}^*)$$

$$\propto \exp\left(\boldsymbol{\kappa}^T \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\kappa}^T \sum_{h=1}^{2} \boldsymbol{Z}_h \boldsymbol{b}_h - \frac{1}{2}\left(\boldsymbol{X}\boldsymbol{\beta}^* + \sum_{h=1}^{2} \boldsymbol{Z}_h \boldsymbol{b}_h\right)^T \right.$$

$$\left. \boldsymbol{D}_\omega \left(\boldsymbol{X}\boldsymbol{\beta}^* + \sum_{h=1}^{2} \boldsymbol{Z}_h \boldsymbol{b}_h\right) - \frac{1}{2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_0^{-1} \sigma_\beta^{-2} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\beta}^{*T}\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}+\boldsymbol{X}^T\boldsymbol{D}_\omega\boldsymbol{X}\right)\boldsymbol{\beta}^* - 2\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \boldsymbol{X}^T\boldsymbol{D}_\omega\right.\right.\right.$$

$$\left.\left.\sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h+\boldsymbol{X}^T\boldsymbol{\kappa})^T\boldsymbol{\beta}^*\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\left(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}_0\right)^T\tilde{\boldsymbol{\Sigma}}_0^{-1}\left(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}_0\right)\right]\right) \propto N\left(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\Sigma}}_0\right)$$

where $\tilde{\boldsymbol{\Sigma}}_0 = \left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \boldsymbol{X}^T\boldsymbol{D}_{\boldsymbol{\omega}}\boldsymbol{X}\right)^{-1}$, $\widetilde{\boldsymbol{\beta}}_0 = \widetilde{\boldsymbol{\Sigma}}_0(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \boldsymbol{X}^T\boldsymbol{D}_\omega\sum_{h=1}^{2}\boldsymbol{Z}_h\boldsymbol{b}_h + \boldsymbol{X}^T\boldsymbol{\kappa})$.

Full conditional for ω_{ijt}

$$f(\omega_{ijt}|ELSE) \propto \exp\left[-\frac{\omega_{ijt}(\boldsymbol{x}_i^T\boldsymbol{\beta}^* + b_{1j} + b_{2ij})^2}{2}\right]f(\omega_{ijt}; y_{ijt} + r, 0)$$

$$\propto \exp\left[-\frac{\omega_{ijt}(\boldsymbol{x}_i^T\boldsymbol{\beta}^* + b_{1j} + b_{2ij})^2}{2}\right]f(\omega_{ijt}; y_{ijt} + r, 0)$$

$$\propto PG(y_{ijt} + r, \boldsymbol{x}_i^T\boldsymbol{\beta}^* + b_{1j} + b_{2ij})$$

Full conditional for $\boldsymbol{b}_1$

Defining $\boldsymbol{\eta}^1 = \boldsymbol{X}\,\boldsymbol{\beta}^* + \boldsymbol{Z}_2\boldsymbol{b}_2$, the conditional distribution of $\boldsymbol{b}_1$ is given as

$$f(\boldsymbol{b}_1|ELSE) \propto \exp\left(\boldsymbol{\kappa}^T\boldsymbol{Z}_1\boldsymbol{b}_1 - \frac{1}{2}\left(\boldsymbol{Z}_1\boldsymbol{b}_1 + \boldsymbol{\eta}^1\right)^T\boldsymbol{D}_\omega\left(\boldsymbol{Z}_1\boldsymbol{b}_1 + \boldsymbol{\eta}^1\right)\right)f\left(\boldsymbol{b}_1|\sigma_{b_1}^2\right)$$

$$\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{b}_1^T\left(\sigma_{b_1}^{-2}\boldsymbol{G}_1^{-1} + \boldsymbol{Z}_1^T\boldsymbol{D}_\omega\boldsymbol{Z}_1\right)\boldsymbol{u} - 2\left(\boldsymbol{Z}_1^T\boldsymbol{\kappa} - \boldsymbol{Z}_1^T\boldsymbol{D}_\omega\boldsymbol{\eta}^1\right)^T\boldsymbol{b}_1\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}(\boldsymbol{b}_1 - \widetilde{\boldsymbol{b}}_1)^T\boldsymbol{F}_1^{-1}(\boldsymbol{b}_1 - \widetilde{\boldsymbol{b}}_1)\right\} \sim N(\widetilde{\boldsymbol{b}}_1, \boldsymbol{F}_1)$$

where $\boldsymbol{F}_1 = \left(\sigma_{b_1}^{-2}\boldsymbol{G}_1^{-1} + \boldsymbol{Z}_1^T\boldsymbol{D}_\omega\boldsymbol{Z}_1\right)^{-1}$ and $\tilde{\boldsymbol{b}}_1 = \boldsymbol{F}_1\left(\boldsymbol{Z}_1^T\boldsymbol{\kappa} - \boldsymbol{Z}_1^T\boldsymbol{D}_\omega\boldsymbol{\eta}^1\right)$.

Full conditional for $\sigma_{b_h}^2$

$$f\left(\sigma_{b_h}^2|ELSE\right) \propto \frac{1}{\left(\sigma_{b_h}^2\right)^{\frac{\nu_{b_h}+n_{b_h}}{2}+1}}\exp\left(-\frac{\boldsymbol{b}_h^T\boldsymbol{G}_h^{-1}\boldsymbol{b}_h + \nu_{b_h}S_{b_h}}{2\sigma_{b_h}^2}\right)$$

$$\propto \chi^{-2}\left(\tilde{\nu}_b = \nu_{b_h} + n_{b_h}, \tilde{S}_b = \left(\boldsymbol{b}_h^T \boldsymbol{G}_h^{-1} \boldsymbol{b}_h + \nu_{b_h} S_{b_h}\right) / \nu_{b_h} + n_{b_h}\right)$$

with $n_{b_1} = J$ and $n_{b_2} = IJ$.

Full conditional for $\sigma^2_{\beta*}$

$$f(\sigma^2_{\beta*}|ELSE) \propto \frac{1}{(\sigma^2_{\beta*})^{\frac{\nu_{\beta*}+I}{2}+1}} \exp\left(-\frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \nu_{\beta*} S_{\beta*}}{2\sigma^2_{\beta*}}\right)$$

$$\propto \chi^{-2}\left(\widetilde{\nu}_{\beta*} = \nu_{\beta*} + I, \widetilde{S}_\beta = [(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \nu_{\beta*} S_{\beta*}] / \nu_{\beta*} + I\right)$$

Full conditional for r

To make the inference of r, we first place a gamma prior on it as $r \sim G(a_0, 1/b_0)$. Then we infer a latent count L for each $Y \sim NB(\mu, r)$ conditional on Y and r. Since $L \sim Pois(-r\log(1-\pi))$, by construction we can use the Gamma-Poisson conjugacy to update r. Therefore,

$$f(r|ELSE) \propto f(r) \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{t}^{n_{ij}} f\left(y_{ijt}|L_{ijt}\right) f\left(L_{ijt}\right)$$

$$\propto r^{a_0-1} \exp(-rb_0) \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{t}^{n_{ij}} \left(-rlog\left(1-\pi_{ij}\right)\right)^{L_{ijt}} \exp\left(r\log\left(1-\pi_{ij}\right)\right)$$

$$\propto r^{a_0 + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{n_{ij}} L_{ijt} - 1} \exp[-(b_0 - \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{n_{ij}} \log(1-\pi_{ij}) r)]$$

$$\propto G\left(a_0 - \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{n_{ij}} \log\left(1-\pi_{ij}\right), \frac{1}{b_0 + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{n_{ij}} L_{ijt}}\right) \quad (4.A.5)$$

According to Zhou et al. (2012), the conditional posterior distribution of L_{ijt} is a Chinese restaurant table (CRT) count random variable. That is, $L_{ijt} \sim CRT(y_{ijt}, r)$ and we can sample it as $L_{ijt} = \Sigma_{l=1}^{y_{ijt}} d_l$, where $d_l \sim Bernoulli\left(\frac{r}{l-1+r}\right)$.

References

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88(422):669–679

Berridge DM, Crouchley R (2011) Multivariate generalized linear mixed models using R. CRC Press, Boca Raton

Bartlett MS (1947) The use of transformations. Biometrics 3(1):39–52

Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78:1–3

Burgueño J, de los Campos GDL, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52:707–719

Casellas J, Caja G, Ferret A, Piedrafita J (2007) Analysis of litter size and days to lambing in the Ripollesa ewe. I. comparison of models with linear and threshold approaches . J Anim Sci 85:618–624

Cavanagh, C.R., Chao, S., Wang, S. et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proceedings of the National Academy of Sciences. 110(20):8057–8062

Crossa J, Pérez-Rodríguez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokosho C (2011) Genomic selection and prediction in plant breeding. Journal of Crop Improvement 25 (3):239–261

Czado C, Gneiting T, Held L (2009) Predictive model assessment for count data. Biometrics 65 (4):1254–1261

de los Campos, G., and Perez-Rodriguez, P. (2013). BGLR: Bayesian generalized linear regression. R package version. http://R-Forge.R-project.org/projects/bglr/

de Maturana EL, Gianola D, Rosa GJM, Weigel KA (2009) Predictive ability of models for calving difficulty in US Holsteins. J Anim Breed Genet 126:179–188

Garthwaite PH, Kadane JB, O'Hagan A (2005) Statistical methods for eliciting probability distributions. J Am Stat Assoc 100(470):680–701

Gelfand AE, Smith AF (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85(410):398–409

Geyer CJ (1992) Practical Markov chain Monte Carlo. Stat Sci 7(4):473–483

Gianola D (1980) A method of sire evaluation for dichotomies. J of Anim Sci 51(6):1266–1271

Gianola D (1982) Theory and analysis of threshold characters. J Anim Sci 54(5):1079–1096

Gianola D, Foulley JL (1983) Sire evaluation for ordered categorical data with a threshold model. Genet Sel Evol 15(2):1–23

Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194:573–596

González-Camacho JM, de los Campos G, Pérez-Rodríguez P, Gianola D, Cairns JE, Mahuku G, Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125(4):759–771

González-Recio O, Forni S (2011) Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet Sel Evol 43:7

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67

Kizilkaya K, Tait RG, Garrick DJ, Fernando RL, Reecy JM (2011) Whole genome analysis of infectious bovine keratoconjunctivitis in Angus cattle using Bayesian threshold models. BMC Proc 5:S22

Kizilkaya K, Fernando RL, Garrick DJ (2014) Reduction in accuracy of genomic prediction for ordered categorical data compared to continuous observations. Genet Sel Evol 46(1):37. doi: 10.1186/1297-9686-46-37

Link WA, Eaton MJ (2012) On thinning of chains in MCMC. Methods Ecol Evol 3(1):112–115

MacEachern SN, Berliner LM (1994) Subsampling the Gibbs sampler. Am Stat 48(3):188–190

McCulloch CE, Searle SR (2001) Generalized, linear, and mixed models (1st ed.). Chichester: Wiley. ISBN 0-471-19364-X.

Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, de los Campos G, Eskridge KM, Crossa J (2015a) Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. G3: Genes| Genomes| Genetics 5(1):291–300

Montesinos-López OA, Montesinos-López A, Crossa J, Burgueño J, Eskridge K (2015b) Genomic-enabled prediction of ordinal data with Bayesian logistic ordinal regression. G3: Genes|Genomes|Genetics 5(10):2113–2126. http://doi.org/10.1534/g3.115.021154

Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Eskridge K, He X, Juliana P, Crossa J (2015c) Genomic prediction models for count data. J Agric Biol Environ Stat 20 (2):533–554

Montesinos-López A, Montesinos-López OA, Crossa J, Burgueño J, Eskridge K, Falconi-Castillo-E, He X, Singh P, Cichy K (2016) Genomic Bayesian prediction model for count data with genotype × environment interaction. G3: Genes|Genomes|Genetics 6(5):1165–1177

Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc A 135:370–384. doi:10.2307/2344614

O'Hara RB, Kotze DJ (2010) Do not log-transform count data. Methods Ecol Evol 1(2):118–122

Park T, van Dyk DA (2009) Partially collapsed Gibbs samplers: illustrations and applications. J Comput Graph Stat 18(2):283–305

Polson NG, Scott JG, Windle J (2013) Bayesian inference for logistic models using Pólya–gamma latent variables. J Am Stat Assoc 108:1339–1349

Quenouille MH (1949) A relation between the logarithmic, Poisson, and negative binomial series. Biometrics 5:162–164

Ramirez-Valverde R, Misztal I, Bertrand J, K. (2001) Comparison of threshold vs linear and animal vs sire models for predicting direct and maternal genetic effects on calving difficulty in beef cattle. J Anim Sci 79:333–338

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

Schurink A, Wolc A, Ducro B, Frankena K, Garrick D, Dekkers J, van Arendonk J (2012) Genome-wide association study of insect bite hypersensitivity in two horse populations in the Netherlands. Genet Sel Evol 44(1):31

Scott J, Pillow JW (2013) Fully Bayesian inference for neural models with negative-binomial spiking. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25. Cornell University, New York, pp 1898–1906

Sorensen DA, Andersen S, Gianola D, Korsgaard I (1995) Bayesian inference in threshold models using Gibbs sampling. Genet Sel Evol 27(3):229–249

Stroup WW (2012) Generalized linear mixed models: modern concepts, methods and applications. CRC Press, Boca Raton

Stroup WW (2015) Rethinking the analysis of non-normal data in plant and soil science. Agron J 107(2):811–827

Teerapabolarn K, Jaioun K (2014) An improved Poisson approximation for the negative binomial distribution. Appl Math Sci 8(89):4441–4445

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91 (11):4414–4423

Vazquez AI, Weigel KA, Gianola D, Bates DM, Perez-Cabal MA et al (2009) Poisson versus threshold models for genetic analysis of clinical mastitis in US Holsteins. J Dairy Sci 92:5239–5247

Varona L, Misztal I, Bertrand J, K. (1999) Threshold-linear versus linear-linear analysis of birth weight and calving ease using an animal model. Ii. Comparison of models. J Anim Sci 77:2003–2007

Villanueva B, Fernandez J, Garcia-Cortes LA, Varona L, Daetwyler HD, Toro MA (2011) Accuracy of genome-wide evaluation for disease resistance in aquaculture breeding programs. J Anim Sci 89:3433–3442

Wang CL, Ding XD, Wang JY, Liu JF, Fu WX, Zhang Z, Jin ZJ, Zhang Q (2013) Bayesian methods for estimating GEBVs of threshold traits. Heredity 110(3):213–219

Wecker WE (1989) Assessing the accuracy of time series model forecasts of count observations. J Bus Econ Stat 7(4):418–419

Wright S (1934) An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics 19:506–536

Yang W, Tempelman RJ (2012) A Bayesian antedependence model for whole genome prediction. Genetics 190(4):1491–1501

Zucknick, M., and Richardson, S. (2014). MCMC algorithms for Bayesian variable selection in the logistic regression model for large-scale genomic applications. Technical Report. http://arxiv.org/abs/1402.2713.

Zhou M, Li L, Dunson D, Carin L (2012) Lognormal and gamma mixed negative binomial regression. In machine learning: proceedings of the international conference on machine learning. vol. 2012. p 1343. NIH Public Access.

Chapter 5
Genomic Selection for Small Grain Improvement

Jessica E. Rutkoski, Jared Crain, Jesse Poland, and Mark E. Sorrells

Abbreviations

BB	BayesB
BRR	Bayesian ridge regression
CIMMYT	International Maize and Wheat Improvement Center
DArT	Diversity Array Technology
DHs	Doubled haploids
DON	Deoxynivalenol
ECs	Environmental covariates
FHB	*Fusarium* head blight
GBS	Genotyping by sequencing
GEBV	Genomic estimated breeding value
GS	Genomic selection
GxE	Genotype-by-environment interaction
h^2	Heritability
HTP	High-throughput phenotyping
LD	Linkage disequilibrium
MAS	Marker-assisted selection
MEs	Mega-environments

J.E. Rutkoski
International Programs, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY 14853, USA

J. Crain • J. Poland
Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA

M.E. Sorrells (✉)
Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York 14853, USA
e-mail: mes12@cornell.edu

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7_5

MET	Multi-environment trials
MxE	Marker-by-environment interaction
PS	Phenotypic selection
QTL	Quantitative trait loci, RR-BLUP, ridge-regression best linear unbiased prediction
SNP	Single nucleotide polymorphism
TPE	Target population of environments

5.1 Introduction

Small grains include sorghum (*Sorghum bicolor*), wheat (*Tritcum aestivum* L.), oats (*Avena sativa* L.), barley (*Hordeum vulgare* L.), rye (*Secale cereale* L.), rice (*Oryza sativa*), millet (*Pennisetum glaucum*), and triticale (*Triticosecale*) but not the pseudo-cereals such as buckwheat (*Fagopyrum esculentum*) and quinoa (*Chenopodium quinoa*). Most of the small grain cereals are self-pollinated with the exception of rye. Consequently, this chapter will focus on genomic selection (GS) research conducted on wheat, oats, barley, and rice. GS in hybrid cereals will be covered in Chapter 7.

Animal breeders initiated GS research, in part, because of the high cost of phenotyping and the inability to replicate individual genotypes. Meuwissen et al. (2001) conducted foundational work in GS development by simultaneously estimating all genetic marker effects. Their simulation results showed up to a 0.84 correlation between estimated breeding values obtained through GS and the true breeding value. Based on these results, they proposed that GS could have significant impact in plant and animal breeding programs by using dense markers to predict performance of individuals that did not have phenotypic records. The genetic gain per unit time achieved by a breeding program can be summarized by the breeder's equation:

$$G = \frac{ir\sigma A}{Y} \tag{5.1}$$

where G is the gain per year, i is selection intensity, r is selection accuracy, σ_A is the square root of narrow-sense heritability, and Y is time in years to complete a cycle of selection (Falconer and Mackay 1996). By combining GS with methods to shorten the breeding cycle, significant gains should be achieved (Meuwissen et al. 2001), with gains proportional to the reduction in breeding cycle time. Plant breeders lagged behind in the use of mixed models and pedigrees for predicting breeding value because phenotyping was relatively less expensive and genotypes using inbred lines could be replicated to increase heritability. However, that has changed rapidly as plant breeders have begun to incorporate GS into their breeding programs. Although it is not considered a small grain, among the earliest publications on GS in crops was a simulation study by Bernardo and Yu (2007) using maize as an example. Using a population derived from a biparental cross of maize inbreds as a training population, they found that the predicted increase in selection gain from ridge-regression best linear unbiased prediction (RR-BLUP) was 18% greater than that from marker-assisted recurrent selection for a highly heritable trait ($h^2 = 0.8$) and 43% for a trait with low heritability ($h^2 = 0.2$). In 2009, Heffner et al. (2009) published a review and interpretation paper on GS highlighting the potential as well as the challenges of applying GS in an applied breeding program. They highlighted the importance of estimating allele effects rather than genotype effects using many unreplicated lines. They also noted that genotype-by-environment interaction (GxE) is likely to be much more problematic for plant breeders than for animal breeders. These were among the earliest publications on GS in plants, and publications that followed shortly thereafter were also simulations. The section that follows reviews the main body of literature on GS in small grains.

5.2 Overview of GS Research in Small Grains

At least 40 GS studies have been published in small grains to date (Table 5.1). Twenty-nine of them were conducted in bread wheat, five in barley, two in oat and rye, and one in durum wheat (*Triticum turgidum* L. spp. *durum*), perennial ryegrass (*Lolium perenne* L.), and intermediate wheatgrass (*Thinopyrum intermedium*). Across all studies Diversity Array Technology (DArT) was the most frequently used marker platform followed by genotyping by sequencing (GBS) and single nucleotide polymorphism (SNP). Taken together, these studies indicate that GS could be successfully applied in cereals breeding to increase rates of genetic gain.

The first few GS studies in small grains were published between 2009 and 2011 (Crossa et al. 2010; Heffner et al. 2011a; de los Campos et al. 2009). Both de los campos et al. (2009) and Cross et al. (Crossa et al. 2010) used data from the International Maize and Wheat Improvement Center (CIMMYT) wheat breeding program. A 13–42% increase in correlation between predicted values and observed values was reported by de los Campos et al. (2009) using Bayesian LASSO compared to prediction models using pedigree alone. Crossa et al. (2010) used CIMMYT wheat breeding data from the international yield trials as well as CIMMYT maize data to evaluate parametric and semi-parametric prediction models based on pedigree and/or genomic relationship. Both de los Campos et al. (2009) and Crossa et al. (2010) concluded that models that used genomic markers were superior to those that used only pedigree relationships. Heffner et al. (Heffner et al. 2011a) used data from a soft white wheat breeding program for multiple traits to compare prediction models, GS to marker-assisted selection (MAS) and phenotypic selection (PS) accuracies, and to look at the impact of training population size and number of markers on the GS accuracy. The authors concluded that based on their results, GS could increase the rate of genetic gain per unit time and cost if applied in a wheat breeding program. Many other studies of GS in cereals were published thereafter, and here we highlight a few key studies.

Table 5.1 Genomic selection studies in small grains

						Topics covered						
Reference	Small grain species	Traits[a]	Population size	Number of markers and platform[b]	Prediction models[c]	GS model comparison	New prediction model development	Effect of training population on accuracy	Other factors affecting accuracy	Coping with GxE	GS vs. MAS	Realized genetic gain
Arruda et al. (2015)	*Triticum aestivum* L.	Six Fusarium head blight resistance traits	273	5054 GBS	RR-BLUP, LASSO, elastic net	x			x			
Asoro et al. (2011)	*Avena sativa* L.	Beta-glucan concentration, days to heading, groat percent, plant height, grain yield	446	1005 DArT	RR-BLUP, Bayes $C\pi$	x		x	x			
Asoro et al. (2013)	*Avena sativa* L.	Beta-glucan concentration	446 and 482	866 and 675 DArT	G-BLUP						x	x
Burgueño et al. (2012)	*Triticum aestivum* L.	Grain yield	599	1279 DArT	G-BLUP, and G-BLUP modeling GxE	x	x			x		
Crossa et al. (2010)	*Triticum aestivum* L.	Grain yield	599	1279 DArT	RKHS, Bayesian LASSO, G-BLUP	x						
Crossa et al. (2016a), b)	*Triticum turgidum* L. spp. durum	Grain yield, grain volume weight, 1000-kernel weight, days to heading	388	7594 SNP	M×E RR-BLUP	x	x			x		
Daetwyler et al. (2014)	*Triticum aestivum* L.	Resistance to leaf rust, stem rust, and yellow rust based on disease severity scores	206	5568 SNP	G-BLUP, BayesR	x						

Dawson et al. (2013)	*Triticum aestivum* L.	Grain yield	622	34,843 GBS	G-BLUP, G-BLUP with environment cluster specific variances, G-BLUP with environment clusters covariances modeled as factor analytic	x				x		
Endelman (2011)	*Triticum aestivum* L.	Grain yield	599	1279 DArT	G-BLUP, BLUP with Gaussian and exponential kernels	x						
Fè et al. (2015)	*Lolium perenne* L.	Days to heading	1846	1,005,509 GBS	G-BLUP						x	
He et al. (2016)	*Triticum aestivum* L.	Grain yield	2325	12,642 SNP	RR-BLUP, Bayes $C\pi$, RKHS, EG-BLUP	x			x			
Heffner et al. (2011a)	*Triticum aestivum* L.	Days to heading, plant height, lodging tolerance, preharvest sprouting tolerance, grain yield, flour yield, and seven other soft wheat quality traits	374	1158 DArT	Multiple linear regression, RR-BLUP, BayesA, BayesB, Bayes $C\pi$	x			x		x	
Heffner et al. (2011b)	*Triticum aestivum* L.	Preharvest sprouting tolerance, flour yield, and seven other soft wheat quality traits	209 and 174	399 multiple platforms and 574 DArT	Stepwise regression multiple linear regression, RR-BLUP, Bayes $C\pi$	x			x		x	

(continued)

Table 5.1 (continued)

						Topics covered						
Reference	Small grain species	Traits[a]	Population size	Number of markers and platform[b]	Prediction models[c]	GS model comparison	New prediction model development	Effect of training population on accuracy	Other factors affecting accuracy	Coping with GxE	GS vs. MAS	Realized genetic gain
Heslot et al. (2014)	*Triticum aestivum* L.	Grain yield	2437	1287	Factorial regression and an extension of factorial regression with soft rules	x	x			x		
Heslot et al. (2013a)	*Hordeum vulgare* L.	Grain yield	996	335 SNP	Bayesian LASSO			x		x		
Heslot et al. (2013b)	*Triticum aestivum* L.	Grain yield, plant height, heading date, and preharvest sprouting tolerance	365	1544 DArT and 38,412 GBS	RR-BLUP				x			
Heslot et al. (2012)	*Hordeum vulgare* L. and *Triticum aestivum* L.	Grain yield, beta-glucan concentration, heading date, plant height, thousand kernel weight	761, 911, 599, 374, and 551	338 SNP, 2146 SNP, 1279 DArT, 1158 DArT, and 319 SNP	Bayes $C\pi$, Bayesian LASSO, BRR, E-Bayes, NNET, RF, RKHS, RR-BLUP, SVR, wBSR	x						
Isidro et al. (2015)	*Triticum aestivum* L.	Grain yield, test weight, lodging tolerance, days to heading, and plant height	1127	38,893 GBS	RR-BLUP			x				
Jarquín et al. (2014)	*Triticum aestivum* L.	Grain yield	139	2395 SNP	A reaction norm model (extension of G-BLUP)	x	x		x			

Jiang et al. (2015)	*Triticum aestivum* L.	Fusarium head blight resistance index based on disease severity and incidence scores	372	782 SNP and SSR combined	Multiple linear regression, RR-BLUP; RKHS; Bayes $C\pi$	x			x		x	
Lado et al. (2016)	*Triticum aestivum* L.	Grain yield	1477	81,999 GBS	G-BLUP and G-BLUP modeling covariances among environments	x	x	x		x		
Lopez-Cruz et al. (2015)	*Triticum aestivum* L.	Grain yield	693, 670, and 807	15,744, 15,744, and 14,217 GBS	BRR and BRR modeling M×E	x	x			x		
Lorenz et al. (2012)	*Hordeum vulgare* L	Fusarium head blight resistance based on severity and DON concentration	691	3072 SNP	RR-BLUP, Bayes $C\pi$, Bayesian LASSO	x		x	x			
Mirdita et al. (2015)	*Triticum aestivum* L.	Fusarium head blight resistance based on severity and *Septoria tritici* blotch resistance based on severity scores	2325	12,642 SNP	RKHS, RR-BLUP	x					x	
Ornella et al. (2012)	*Triticum aestivum* L.	Resistance to yellow rust and stem rust, based on severity scores	90–180	1400 DArT including non-polymorphic markers	Bayesian LASSO, RR, SVR	x						
Perez-Rodriguez et al. (2013)	*Triticum aestivum* L.	Days to heading, grain yield	306	1717 DArT	Bayesian LASSO, BayesA,	x						

(continued)

Table 5.1 (continued)

						Topics covered						
Reference	Small grain species	Traits[a]	Population size	Number of markers and platform[b]	Prediction models[c]	GS model comparison	New prediction model development	Effect of training population on accuracy	Other factors affecting accuracy	Coping with GxE	GS vs. MAS	Realized genetic gain
					BayesB, RKHS, BRR, NNET							
Poland et al. (2012)	*Triticum aestivum* L.	Grain yield, thousand kernel weight, days to heading	254	41,371 GBS	G-BLUP				x			
Rutkoski et al. (2014)	*Triticum aestivum* L.	Resistance to stem rust	365	4040 GBS	Multiple linear regression, G-BLUP, Bayesian LASSO, Bayes $C\pi$	x	x				x	
Rutkoski et al. (2015a)	*Triticum aestivum* L.	Resistance to stem rust	626, 626, 1163, and 1150	20,882, 20,882, 18,653 and 18,653 GBS	G-BLUP and BRR							x
Rutkoski et al. (2015b)	*Triticum aestivum* L.	Resistance to stem rust	365 and 503	17,168 GBS	G-BLUP			x				
Rutkoski et al. (2012)	*Triticum aestivum* L.	Five *Fusarium* head blight resistance traits and days to heading	170	2402 DArT and 38 SSR	RR-BLUP, Bayesian LASSO, RKHS, RF, multiple linear regression	x			x		x	
Sallam et al. (2015)	*Hordeum vulgare* L	*Fusarium* head blight resistance based on severity and DON concentration, grain yield, plant height	647	1536 SSR	RR-BLUP, RKHS, Bayes $C\pi$	x		x				

Schmidt et al. (2016)	*Hordeum vulgare* L	Twelve malting quality traits	65–424, and 102	4095 and 4359 SNP	RR-BLUP				x			
Schulthess et al. (2016)	*Secale cereale* L.	Grain yield, protein content	201 and 219	394 and 584 DArT	Multi-trait RR-BLUP	x			x			
Storlie and Charmet (2013)	*Triticum aestivum* L.	Grain yield	318	1279 DArT	RR-BLUP, G-BLUP, BRR, Bayesian LASSO				x			
Thavamanikumar et al. (2015)	*Triticum aestivum* L.	Time to young microspore, spike grain number	165 and 159	17,328 and 17,293 SNP	BayesB, Bayesian LASSO, RR-BLUP, PLS, and SPLS	x			x			
Wang et al. (2014)	*Secale cereale* L.	Grain yield, plant height, starch content, total pentosan content	220 and 220	1048	RR-BLUP and a regression model using identified QTL			x	x		x	
Ward et al. (2015)	*Triticum aestivum* L.	Days to heading, plant height, grain yield, and 21 quality traits	151	603 DArT and 655 metabolites	RR-BLUP and differentially penalized regression	x	x					
Zhang et al. (2016)	*Thinopyrum intermedium* (host) Barkworth & D. R. Dewey	Heading score, plant height, head weight, threshability, seed weight, biomass	1126	3883	RR-BLUP, G-BLUP with a Gaussian kernel, BayesA, BayesB, Bayes $C\pi$, Bayesian LASSO, BRR, RKHS, and RF	x		x				
Zhao et al. (2014)	*Triticum aestivum* L.	Days to heading, plant height	1739	1280 SNP and 3 function SNP markers	RR-BLUP, Bayes $C\pi$, RR-BLUP with differentially weighted predictors, stepwise	x	x				x	

(continued)

Table 5.1 (continued)

						Topics covered						
Reference	Small grain species	Traits[a]	Population size	Number of markers and platform[b]	Prediction models[c]	GS model comparison	New prediction model development	Effect of training population on accuracy	Other factors affecting accuracy	Coping with GxE	GS vs. MAS	Realized genetic gain
					regression with functional markers							

[a] *Fusarium* head blight resistance is caused by *Fusarium graminearum* Schwabe, leaf rust is caused by *Puccinia triticina*, stem rust is caused by *Puccinia graminis*, yellow rust is caused by *Puccinia striiformis*, *DON* deoxynivalenol, *Septoria tritici* blotch is caused by *Zymoseptoria tritici*

[b] *GBS* genotyping by sequencing, *DArT* Diversity Array Technology, *SNP* single nucleotide polymorphism

[c] *RR-BLUP* ridge-regression best liner unbiased prediction (Meuwissen et al. 2001; Whittaker et al. 2000), *LASSO* least absolute shrinkage and selection operator (Tibshirani 1996), elastic net (Zou and Hastie 2005), Bayesian LASSO (Park et al. 2008), Bayes $C\pi$ (Habier et al. 2011), *G-BLUP* genomic best linear unbiased prediction (Habier et al. 2010), *RKHS* reproducing kernel Hilbert space regression (Gianola and Van Kaam 2008), $M \times E$ marker-by-environment interaction, BayesR (Erbe et al. 2012), EG-BLUP (Jiang and Reif 2015), BayesA (Meuwissen et al. 2001), BayesB (Meuwissen et al. 2001), *BRR* Bayesian ridge regression (Pérez et al. 2010), *E-Bayes* empirical Bayes (Xu 2007), *RF* random forest regression (Breiman 2001), *SVR* support vector regression (Smola and Schölkopf 2004), *NNET* neural network (Ripley 1996), *wBSR* weighted Bayesian shrinkage regression (Hayashi and Iwata 2010), *PLS* partial least squares (Geladi and Kowalski 1986), *SPLS* sparse partial least squares (Chun and Keleş 2010)

In 2011, Asoro et al. published the first GS study in oat using data from an oat breeding program for multiple traits to evaluate GS accuracies and to assess factors affecting accuracy. This study was also the first GS study in small grains that examined how characteristics of the training population affected GS accuracy. The authors found that including older lines in the training set either increased or had no impact on accuracy, suggesting that an appropriate model training population can be constructed by accumulating breeding program data over time. The first study in small grains that aimed to increase prediction accuracy by modeling genotype-by-environment interaction was by Burgueño et al. (Burgueño et al. 2012). This study used the same data as in Crossa et al. (2010) and evaluated whether modeling covariance between environments could improve prediction accuracy. The results of this study indicated that modeling covariance between environments could increase accuracy when predicting performance within specific environments when the lines were observed in some environments but not others. This highlights that phenotypes of the selection candidates observed in one environment can be used to improve genomic prediction of selection candidates' breeding values in other environments.

In 2012, the first study in small grains using GBS for genomic selection was published by Poland et al. (2012). This study found that GBS led to greater GS accuracies compared to current DArT markers. A later study by Heslot et al. (2013b) followed up on this observation and found that GBS leads to higher GS accuracies compared to DArT because GBS produces a larger number of non-redundant, evenly distributed markers. Although many researchers are concerned that the relatively large amount of missing data commonly observed in GBS datasets may impede downstream analyses, Rutkoski et al. (2013) found that the amount of missing data commonly observed in wheat GBS datasets has very little impact on GS prediction accuracies.

The first report of realized gain from a GS experiment in small grains was published by Asoro et al. in Asoro et al. 2013. Working in oats, the authors compared GS, MAS, and phenotypic selection for beta-glucan concentration. In both the GS and MAS schemes, phenotypic data on the selection candidates were used in addition to marker data. The authors found that more superior individuals originated from the populations developed using GS or MAS, demonstrating the value of markers for improving selection. Rutkoski et al. (2015a) published the first realized gain from GS experiment in wheat. This study compared GS with PS for breeding for quantitative adult plant resistance to stem rust. Unlike in Asoro et al. (2013), in the GS scheme, phenotypic data were not available on the selection candidates prior to selection. The authors found that GS and PS lead to equal rates of genetic gain per unit time, but GS led to a faster loss of genetic variance because with GS two cycles were completed, while with PS only one cycle was completed. The lack of improvement in genetic gain per unit time from GS over PS in this case was due to the relatively low prediction accuracy in the first cycle of selection when the model training population is not closely related to the selection candidates. The University of Minnesota barley breeding program has implemented GS since 2010 for *Fusarium* head blight (FHB) resistance as well as for yield, winter hardiness, and malting quality resulting in reduced time and labor for phenotyping (Bernardo 2016). In evaluating dynamic germplasm from the barley breeding program, Sallam et al. (2015) found prediction accuracies that ranged from 0.03 to 0.99 for traits including plant height, yield, FHB resistance, and deoxynivalenol (DON) concentration.

5.3 Factors Affecting GS Prediction Accuracies

5.3.1 Theoretical Considerations of GS Accuracies

Utilizing foundational genetic theory, irrespective of species, we can predict GS accuracy based on the heritability of the trait, the number of independent chromosome segments, and the number of individuals in the training population (Daetwyler et al. 2010). For the simplest GS examples, this assumes (1) there is perfect linkage between markers and quantitative trait loci (QTL), (2) the model training and selection candidate individuals are sampled from the same population, and (3) the trait of interest is conferred by a large number of additive loci. A large body of work has extended these concepts to understand how GS performs in real-world applications when germplasm and phenotypic data diverge from the ideal conditions. For example, when markers are not in complete linkage disequilibrium (LD) with QTL, increasing the number of markers so that more markers are in LD with QTL leads to higher accuracies (Heffner et al. 2011b; Muir 2007). If model training and selection candidates are sampled from different populations, the level of relationship between the two populations is another factor affecting accuracy, with increasing levels of relationship leading to higher accuracies (Pszczola et al. 2012). Lastly, when the trait of interest is conferred by few loci, the choice of prediction model will affect accuracy.

5.3.2 Research to Increase GS Accuracy

Many of the research studies of GS in small grains have examined factors affecting prediction accuracy. In small grains, at least 29 studies have looked at the effect of prediction model on the GS accuracy, at least nine studies have examined how the relationship between the training population and the selection candidate population affects accuracy, and at least 16 studies have looked at other factors affecting accuracy including training population size and number of markers.

5.3.3 Effect of GS Model

The vast majority of studies looking at the effect of prediction model on accuracy report either no or only a small effect of the prediction model on the accuracy (Heslot et al. 2012; Sallam et al. 2015). Among studies that have compared additive and nonadditive GS models, Mirdita et al. (2015), He et al. (2016), Endelman et al. (2011), Perez-Rodriguez et al. (2013), and Rutkoski et al. (2012) observed that nonadditive GS models led to higher accuracies compared to additive models. In contrast, Ornella et al. (2012) reported that in biparental populations, nonlinear models led to lower accuracies compared to linear models.

5.3.4 Relationship of Training and Validation Population

Studies in small grains that have examined the effect of the relationship between the individuals in the training and validation population on the GS accuracy have observed that having a closer relationship between training and validation populations leads to higher GS accuracies (Asoro et al. 2011; Lorenz et al. 2012; Rutkoski et al. 2015b; Wang et al. 2014; Zhang et al. 2016). Interestingly, Rutkoski et al. (2015b) found that including older model training data after updating the model training population could lead to decreased prediction accuracy but only if the older model training data had a low heritability. On the other hand, Lorenz et al. (2012) found that including individuals from a different subpopulation in the model training set neither increased nor decreased accuracy. Asoro et al. (2011) reported that including older individuals in the model training population either slightly increased or did not affect the accuracy. In a study by Zhang et al. (2016) about GS in intermediate wheatgrass, the authors reported that accuracies increased with increasing numbers of families and genotypes per family in the training set, but after 30 families and six genotypes per family, the increase in accuracy was very small.

The potential to select subsets of the model training population of a given size that lead to the highest possible accuracy, referred to as training population optimization, has been investigated for wheat by Isidro et al. (2015) and Rutkoski et al. (2015b). Both studies found that training population optimization was better than random sampling for selecting subsets from a population for model training. This could be useful for selecting which set of lines to phenotype for prediction model updating to ensure the highest accuracy given the resources available. Overall, studies in small grains generally confirm that GS model training populations should be related to the selection candidates and should be frequently updated to maintain accuracy. While most envision updating the GS model with phenotypic data generated on the lines selected in the breeding program because of their superior breeding values, it may be wise to use training population optimization to select a set of lines to phenotype specifically for GS model updating. This would include lines that would ordinarily get discarded in the breeding program.

More research is required in this area to determine if the benefits of selecting lines to phenotype specifically for model training outweigh the costs.

5.3.5 Effect of Number of Markers and Individuals

Studies in small grains that have examined the effect of the number of markers and number of individuals in the training population on the GS accuracy have observed similar trends where accuracy increases linearly with marker number and training population size until a plateau is reached, and a plateau in accuracy is reached sooner with marker number than with training population size (Arruda et al. 2015; Heffner et al. 2011a, b; Lorenz et al. 2012). Heffner et al. (2011a) observed that for predicting within biparental populations, increasing training population size from 24 to 96 leads to a linear increase in accuracy, while increasing marker number beyond 256 did not improve prediction accuracy. For predicting within a population of advanced breeding lines, Heffner et al. (2011b) found that accuracy increased linearly with TP size from 96 to 288 and very gradually when the number of markers was increased beyond 384. Lorenz et al. (2012) reported that in a population of barley breeding lines from different breeding programs, accuracy reached a plateau at a population size of 200 and that marker number could be decreased to 384 without losing accuracy. Other work in barley by Sallam et al. (2015) found that DON concentration level predictions plateaued at a TP of 75, while grain yield prediction did not plateau based on the TP size, suggesting the TP size may be trait specific. The occurrence of plateaus in accuracy at relatively low numbers of markers and population sizes is a reflection of the low rate of LD decay with physical distance, or in other words, a low number of independent chromosome segments in the small grain breeding populations are used for GS studies.

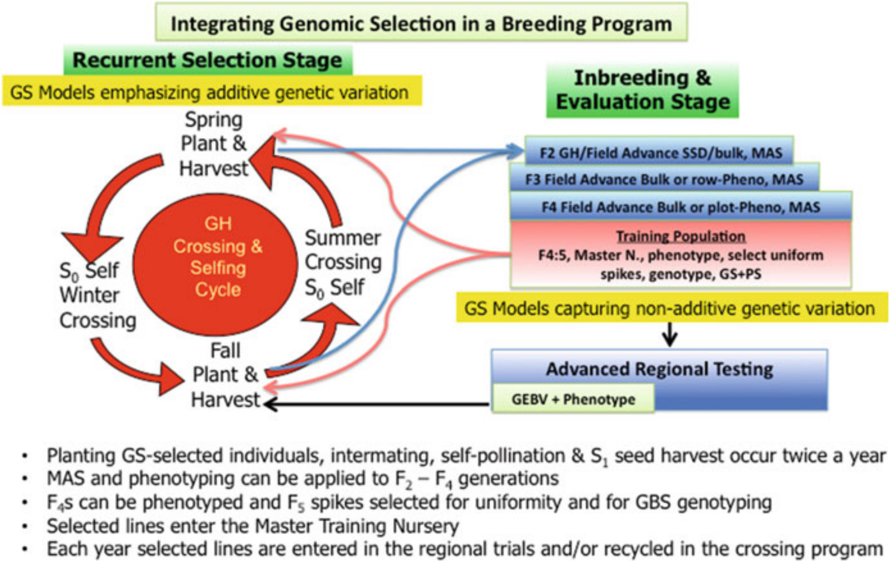
In a population consisting of advanced wheat breeding lines from breeding programs across the Midwestern and Eastern United States, Arruda et al. (2015) found that for most traits, a plateau in accuracy was reached when the training population was larger than 192 lines, and, depending on the trait, a plateau in accuracy occurred when the number of markers was greater than 1500 or 3000. In a breeding scenario where older breeding lines are used to predict newer breeding lines, larger training population sizes and more markers may be required compared to what cross validation studies would indicate because generations of recombination have taken place leading to faster rates of LD decay among the combined model training-selection candidate population. This was observed in a study by Rutkoski et al. (2015b) where the authors examined the increase in accuracy with increasing training population size for two training sets, one distantly related and one closely related to the selection candidates. They found that accuracy increased linearly with training population size in the more distantly related training set, while accuracy reached a plateau at 292 individuals with the more closely related training set. Research on training population size and number of markers will ultimately need to be conducted using forward validation within individual breeding programs.

5.4 Breeding Methods for Variety Development in Cereals

While any implementation of GS involves identifying high-performing individuals based on a genomic prediction model, the methods that different breeding programs could use may look quite different depending upon the budget and resources available to the breeding program, the relative cost of genotyping vs. phenotyping, the generation time that can be achieved, and the heritability of the traits of interest. An important feature of GS is that it is complementary to MAS. Marker-assisted selection is most effective for simply inherited, high heritability traits, whereas GS is relatively more effective for low heritability traits conferred by many QTL. Both methods could be readily incorporated into a molecular breeding program and used in concert to enable breeders to select for simply inherited traits with MAS and for quantitative traits using GS. Applications such as spiked GBS (Rife et al. 2015) which combine whole-genome profiling along with known marker assays could allow breeding programs to efficiently and affordably integrate both GS and MAS.

5.4.1 Timing of GS Application Within Breeding Programs

Implementation of GS in cereal crops can be imposed in early or late generations (Fig. 5.1). In early generation implementation, GS is applied to the selection candidates directly after crossing prior to any selfing or after one generation of selfing. Implementing GS in this way leads to a greater reduction in the breeding cycle duration because the two or more growing seasons that would normally be needed for selfing are eliminated (Heffner et al. 2010; Hickey et al. 2014). Selected individuals are cycled back into the crossing block as parents. This rapid cycling (Fig. 5.1) program allows individuals selected based on their genomic estimated breeding value (GEBV) to be planted, cross- or self-pollinated, and harvested two or more times a year for many cereal species. This would ordinarily not be possible because many important traits must be evaluated on fixed lines using relatively large quantities of seed. One consideration for implementing GS in this way is that model updating will need to come from inbred lines derived from the selection candidates in earlier cycles and potentially several cycles of GS selection could be conducted for every single cycle of training population updating that is completed. Once early generation lines are selected as parents in the rapid cycling program, they can begin the inbreeding phase where culling based on MAS and PS can be applied at any generation until the F_4 or F_5 generations as preferred. At that stage, the candidates should be whole-genome genotyped and phenotyped to train the model to predict GEBVs in the breeding population. The genome-wide marker data and phenotypic data can also be fit in a model that captures nonadditive genetic effects to improve selection accuracy on the lines per se for promotion as varieties. Phenotyping and genotyping all selection candidates have been shown to be more favorable for improving rates of genetic gain compared to only phenotyping a


Integrating Genomic Selection in a Breeding Program
Recurrent Selection Stage
GS Models emphasizing additive genetic variation
Spring Plant & Harvest
GH Crossing & Selfing Cycle
S_0 Self Winter Crossing
Summer Crossing S_0 Self
Fall Plant & Harvest
Inbreeding & Evaluation Stage
F2 GH/Field Advance SSD/bulk, MAS
F3 Field Advance Bulk or row-Pheno, MAS
F4 Field Advance Bulk or plot-Pheno, MAS
Training Population
F4:5, Master N., phenotype, select uniform spikes, genotype, GS+PS
GS Models capturing non-additive genetic variation
Advanced Regional Testing
GEBV + Phenotype

- Planting GS-selected individuals, intermating, self-pollination & S_1 seed harvest occur twice a year
- MAS and phenotyping can be applied to $F_2 - F_4$ generations
- F_4s can be phenotyped and F_5 spikes selected for uniformity and for GBS genotyping
- Selected lines enter the Master Training Nursery
- Each year selected lines are entered in the regional trials and/or recycled in the crossing program

Fig. 5.1 Integration of GS in a pure line breeding program. In the rapid cycling phase, GS is used to enhance gain per unit time. In the inbreeding phase, MAS and PS can be imposed until the F4 or F5 generation, and then whole-genome genotyping is used to select individuals that enter the training population or are recycled in the crossing program. Each phase is conducted simultaneously, and the GS models are updated annually

subset of the genotyped selection candidates (Endelman et al. 2014). GS can also be applied only among lines, without a rapid cycle program. Compared to a rapid cycling approach, applying GS only among lines would enable higher selection accuracies but would not reduce the breeding cycle duration as dramatically. Different GS breeding schemes should be evaluated for each specific situation to understand the trade-offs in cycle time and selection accuracy and to optimize gain from GS. Deterministic and stochastic simulations can be useful tools for this purpose.

Most of the GS research in cereal crops has been conducted using inbred lines. In a deterministic simulation study by Heffner et al. (2010), GS among inbred lines was found to increase gain from selection per unit time twofold compared to MAS among inbred lines for known QTL. Examining how to optimize preliminary yield trials, Endelman et al. (2014) found that a 5% increase over phenotypic selection could be achieved by using GS in inbred lines.

Other authors have reported evaluations of various GS schemes (Longin et al. 2015). The study by Longin et al. (2015) evaluated GS using GS alone, GS followed by one or two rounds of PS, and the comparison to PS only. At an estimated GS accuracy of 0.3, the authors found that GS followed by one round of PS produced the highest genetic gain; however, the rankings of methods were dependent upon GS accuracy. For example, if GS accuracy could exceed 0.65 using GS only with no

PS provided the highest genetic gain, but under more realistic conditions (GS accuracy 0.3), combinations of GS and PS were more productive. The use of doubled haploid lines would preclude the use of MAS and PS during the inbreeding phase of line development (Fig. 5.1).

5.4.2 Genomic Selection for Germplasm Improvement Through Introgression of Alleles

While the previous GS strategies have focused mainly on reducing the length of the breeding cycle or increasing the selection accuracy to increase the rate of genetic gains, GS could also be used for introgression of exotic alleles into elite germplasm. Crossa et al. (2016a) evaluated GS in landraces populations including over 2000 Iranian and 8000 Mexican lines. They found good GS prediction accuracies even with GxE and population structure. They proposed that GS could be used to predict genotype performance of all genotyped accessions within a germplasm collection and then phenotyping could be conducted on the most promising lines, followed by introgressing the selected exotic alleles into elite germplasm. Bernardo (2009) simulated the effect of introgressing exotic alleles into adapted maize germplasm. His results showed that GS could be used to rapidly incorporate exotic alleles into elite germplasm. Additionally, this work provided some guidance on the number of cycles of GS that could be used with exotic alleles and where to apply GS in the introgression program. He found that the best starting material for GS was an F_2 cross between exotic and adapted germplasm rather than a BC_1 or BC_2. In general the rate of genetic progress declined after 7–8 cycles of GS, but assuming three cycles of GS could be completed per year resulted in an equivalent time frame of two rounds of PS (2 years per cycle). The 7–8 cycles of GS resulted in 1.25–2.4 times the rate of gain compared to the two cycles of PS depending on the number of favorable alleles in the exotic germplasm and a trait heritability of 0.8. Another simulation study in maize suggested that GS should be applied in exotic-by-exotic crosses and used to increase the frequency of favorable alleles (Gorjanc et al. 2016). This was because using GS with exotic-by-elite crosses quickly resulted in reconstructing elite material. In practice, applying GS among exotic-by-exotic crosses could be problematic due to poor adaptation. For example, if the exotic individuals are photoperiod sensitive in the environments of interest, it will not be possible to gather meaningful grain yield data for prediction modeling. Importantly, Bernardo (2009) found that the quality (accuracy) of the phenotypic data and the quantity (number) of phenotyped individuals were crucial for ensuring success of GS in introgressing exotic germplasm. Thus, training population size and composition needs to be carefully considered. Bernardo (2009) suggested that if trait heritability was low (h^2=0.2), then more field testing (replication) across multiple environments should be used to increase the entry-mean heritability (accuracy). Along with replication to improve the entry-mean heritability, Bernardo (2009) suggested that training populations should be larger, 288 compared to 144 in adapted-by-adapted crosses. Additionally, to achieve sufficiently high prediction accuracy, GS for introgression of exotic germplasm would need to be within biparental populations, which have slower rates of LD decay and fewer segregating chromosome segments compared to populations derived from multiple families.

5.4.3 Combining Genomics and Phenomics for Increased Precision

As reported by Endelman et al. (2014) and Lorenz (2013), phenotypic data on the selection candidates per se can be used in GS models to increase selection accuracy and gain from selection. With decreasing genotyping costs, phenotypes are quickly becoming the most valuable asset to breeding programs. Field-based phenomics or high-throughput phenotyping (HTP) is an active area of research that is working to provide image and sensor data for traits that are correlated with the phenotypes of interest (Cobb et al. 2013; White et al. 2012) which could be useful for prediction modeling. Using a variety of proximal sensors, researchers have mapped QTL for biomass growth in triticale (Busemeyer et al. 2013) as well as assessing differences in crop response to well-watered and drought conditions in cotton (Andrade-Sanchez et al. 2014). The ability to generate large volumes of data quickly and at multiple time points possibly before grain yield testing has led to efforts to combine phenotypic data within the GS model. There are several methods by which both GS and HTP could be integrated into a breeding program. Low-cost HTP could be used to evaluate lines in early generations to both provide data to train GS models and make screening decisions when thousands of lines need to be evaluated (Araus and Cairns 2014). Along with providing information about the crop, HTP could allow breeding programs to increase population size of material evaluated while increasing selection intensity resulting in higher genetic gain (Crain and Reynolds 2016).

Additionally, HTP and GS combinations could also be used for evaluating more advanced lines. Using spectral indices and canopy temperature, Rutkoski et al. (2016) and Crain et al. (unpublished) reported higher GS prediction accuracies by including HTP data in the GS model. Rutkoski et al. (2016) found up to a 70% increase in prediction accuracy for grain yield by including HTP traits of canopy temperature and vegetation indices. By including canopy temperature and spectral reflectance (Crain et al. unpublished) found an average of a 12% increase in GS model accuracy compared to GS models utilizing only marker information. By using multiple traits (phenotypes) from HTP data, the goal is to enhance model prediction accuracy. Previous work with multiple traits has shown that prediction accuracy of a low heritability trait can be greatly increased when a second or multiple traits that have higher heritability are added to the model (Jia and Jannink 2012). They found that with a low heritability trait ($h^2 = 0.1$), prediction accuracy went from 0.49 to 0.64 assuming a genetic correlation of 0.1. As the genetic correlation increased, the prediction accuracy further increased. Along with increasing prediction accuracy, they also found that prediction accuracies could be increased when there was missing information. While these are some of the earliest efforts to incorporate phenotypic data, HTP is offering many possibilities to further increase the genetic gains from GS. The utilization of HTP along with GS has the potential to provide scientists with rich datasets that adequately reflect real-world conditions. For example, Montesinos-López et al. (2016) utilized the same data presented by Rutkoski et al. (2016) to develop multi-trait, multi-environment models. Employing these types of complex models that account for the genotypic relationship between multiple traits and environmental factors (multiple environments) should allow breeders to get a more complete picture of how genetics are expressed in different conditions allowing for more accurate selection decisions to be made.

5.4.4 Additional Breeding Program Considerations

Along with increasing the accuracy of phenotyping, the design of the breeding program can be modified to take full advantage of GS. With whole-genome genotyping, testing more lines at different locations, rather than all lines at one location or fewer lines at all locations, Endelman et al. (2014) found that prediction accuracies were increased. This potentially surprising result comes from the fact that in GS, the alleles are under selection rather than an individual genotype. Thus, in field trials the goal is replication of alleles, but the replication of alleles is not limited to specific genotypes (Lorenz et al. 2011).

The adept use of data and training populations has allowed GS to predict new genotype performance in multiple environments as well as hybrid performance. Lado et al. (2016) found that using related genotype information at multiple locations could be used to predict new genotype performance with high (0.5) accuracy. Zhao et al. (2015) used GS to predict performance of hybrid lines to develop heterotic groups. While this application was for hybrid breeding, the ability to predict hybrid (cross) performance without testing the phenotype could be applicable for the pure line breeder as well. Both of these examples along with others (e.g., Endelman et al. 2014; Hickey et al. 2014) highlight the importance of training population design (size and quality) and genotyping (number of markers) in making accurate predictions.

GS can be incorporated into a breeding program in several different facets. Breeders could use GS to provide more information about the selections they are making or to make all selection decisions. In practice, GS will probably be applied somewhere in between these two extremes whether it be to increase favorable alleles, rapidly cycle germplasm, or introgress exotic alleles. The success that GS has will be dependent on the resourcefulness of the researcher and his/her ability to fully utilize molecular, phenotypic, and environmental information in an efficient way based on careful assessment of different breeding schemes.

5.5 GS for Analyzing and Predicting GxE

Plant breeders are concerned about GxE because it amplifies the phenotypic variation without contributing to the additive genetic variation, thereby reducing heritability across environments. The challenges of GxE have been with plant breeders since the beginning. However, whole-genome genotyping and GS present some opportunities to the modern plant breeder that are already changing our strategies for dealing with GxE, for both genotype selection and multiple environment trial evaluation methods.

5.5.1 Target Population of Environments

Fundamental to the concept of GxE is the definition and sampling of a target population of environments (TPE). The TPE represents the biotic and abiotic factors that released varieties are likely to encounter during their production. Because genotypes respond differently to environmental factors, in multi-environment trials (MET), varieties perform differently resulting in relative differences in performance as well as rank changes. Rank changes or crossover interactions complicate selection because one variety is not the best for all environments. The challenge is to adequately sample environments, especially over years where climate can be more variable. Generally, the composition of the TPE is unknown. Years with insufficient precipitation may be more common than those with adequate precipitation, and a streak of dry years could bias the evaluation of experimental lines. One approach to managing environmental variability is to classify environments based on historical data and then weight the trials by their expected frequency of occurrence in the TPE (Podlich et al. 1999). This approach effectively adjusts for the negative effects of unrepresentative environments on selection resulting in increased gain from selection.

5.5.2 Application of GS Models to GxE

GxE was recognized as a major issue for applying GS in plants (Crossa et al. 2010; Heffner et al. 2009). Crossa (2012) reviewed GxE and marker-by-environment interaction studies and discussed models for assessing marker effects, QTL, and marker effect by environment interactions and for studying the pattern of covariability of marker effects across environments. He made the case for evaluating different models from different areas of statistical research to better understand genetic effects and their interaction with environment. Studies that have evaluated GS predictions that model genotype-by-environment interaction report an increase in accuracy from modeling the interaction rather than ignoring it (Burgueño et al.

2012; Heslot et al. 2014; Jarquín et al. 2014; Lado et al. 2016; Lopez-Cruz et al. 2015). Within GS accounting for GxE remains an active area of research with numerous methods proposed for utilizing GxE including marker-by-environment interactions (Heslot et al. 2013a), using genetic correlations between TPE (Burgueño et al. 2012), and using environmental covariates (Heslot et al. 2014).

5.5.3 GxE by Modeling Marker Replication and Interaction

By design, plant breeding MET data are unbalanced, thus limiting the kinds of analyses that can be performed on the data because all genotypes are not represented in all environments. The transforming principle first employed by Heslot et al. (2013a) is that even though all genotypes are not represented in all environments, all marker effects *are* represented in all environments. This allows one to measure the relative similarities among environments based on marker effects as well as similarities based on prediction of marker effects. This approach is especially useful in cases where a factor analytic model fails convergence. Similarities based on marker effects can be determined using Euclidean distances and visualized using cluster analysis (Fig. 5.2, Heslot et al. 2013a). Using multi-environment yield trial data from a commercial barley breeding program, the authors found that outlier environments were readily identified. Additionally, the breeder field notes corroborated the results; however, for this dataset, grouping environments based on similarity of marker effects did not increase prediction accuracy. Likewise, prediction of marker effects in other environments can be calculated with a simple correlation analysis and visualized in a heat map. Although the patterns were not as clear as the marker effects, grouping environments based on average reciprocal prediction accuracies increased prediction accuracy for yield across environments (Heslot et al. 2013a).

In a second experiment designed to optimize the composition of the training population, Heslot et al. (2013a) used the average predictive ability of each environment for predicting performance of lines in the other environments in the same dataset. The environments were then ranked from least predictive to most predictive, and starting with the least predictive, one environment was removed at a time, and then the model was retrained on the remaining environments to determine if prediction accuracy improved (Fig. 5.3, Heslot et al. 2013a). The environments that were removed were placed in an unpredictive set, and the prediction accuracy of that set was calculated. When the prediction accuracy of the predictive set dropped and/or the unpredictive set increased, the remaining environments were considered to be the optimal set, and, in this study, accuracy was increased from 0.54 to 0.61. Out of the 58 environments, 18 unpredictive environments were removed. Interestingly, some outlier environments were included, and only one barley line was excluded in the optimal set of environments.

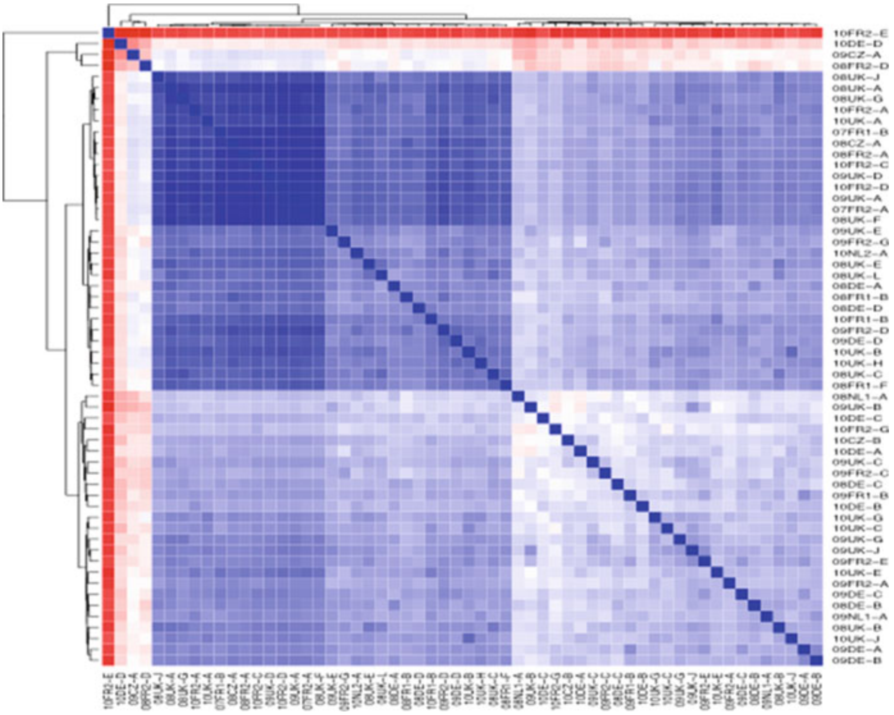


Fig. 5.2 Heat map showing the similarity of environments based on Euclidian distances computed using marker effects. Environment comparisons with red shading are more dissimilar, and those environments with *blue* shading are more similar (Fig. 5.3 from Heslot et al. 2013a)

Another benefit from whole-genome genotyping is that by modeling MxE, we can begin to understand what genetic regions have main (stable) effects and which ones interact with the environment. Lopez-Cruz (2015) modeled MxE using regression of wheat phenotypes on markers or using covariance structures (a genomic best linear unbiased prediction-type model) to estimate main (stable) effects and environment-specific (interaction) effects. Environments that were correlated exhibited low MxE, and those that were not correlated showed high MxE. Modeling MxE in general improved prediction accuracy over models that did not take MxE into account. The MxE model limits the ability to interpret patterns of GxE that are not positive leading the authors to conclude that the model is best suited for the joint analysis of positively correlated environments (Lopez-Cruz et al. 2015).

Model development to represent GxE is a research area that is continually advancing. To further extend research on MxE, Crossa et al. (2016b) explored the use of priors that produce shrinkage and variable selection including Bayesian ridge regression (BRR) and BayesB (BB) in durum wheat. They evaluated the genomic prediction accuracy of MxE models within and across environments. The MxE model minimized the model residual variance and improved data-fitting gain for

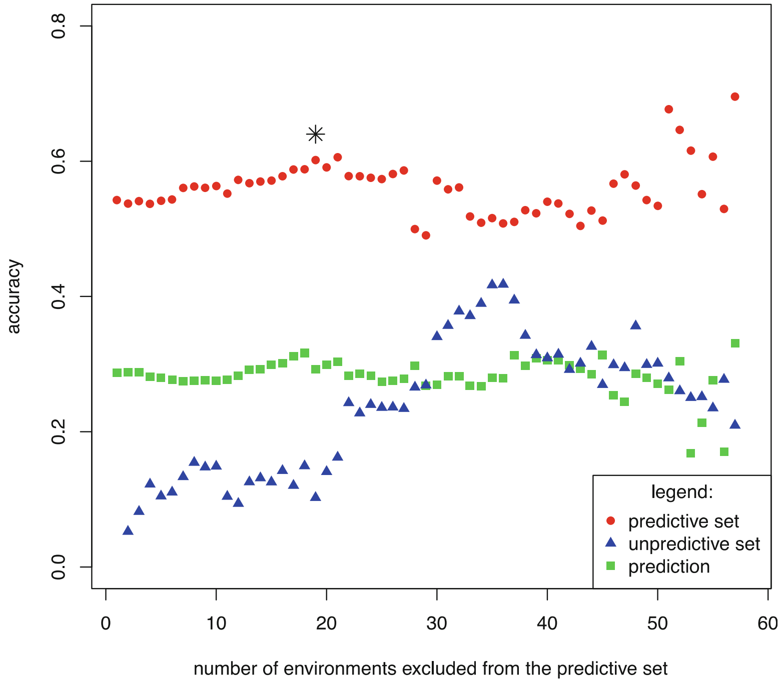


Fig. 5.3 Optimization of the training population. The *blue dots* are cross-validated accuracies for the selected training population (predictive set), and red triangles are prediction accuracies for the environments removed from the training population (unpredictive set). *Green squares* are the prediction accuracies for a validation set observed in 2011 (Fig. 5 from Heslot et al. 2013a)

more simply inherited traits compared to more complex traits such as grain yield and test weight. The MxE model identified markers for the major genes for heading date including *Ppd-A1*, *Ppd-B1*, and Ta*FT-A* on chromosomes 2A, 2B, and 7A, and their effects were stable across environments. For grain yield, several additional chromosome regions with large marker effects were identified in all chromosome groups. Another example of modeling GxE was given by Cuevas et al. (2016) in which nonlinear Gaussian kernels were used to model MxE. Because this model allowed for small, complex MxE interactions, they were able to capture up to 60% greater predictions compared to models using a single environment.

5.5.4 GxE by Treating Environments as Multiple Traits

Genotype-by-environment interactions can be analyzed by treating different environments as multiple traits and considering variety performance in different environments as correlated traits (Falconer and Mackay 1996) or, in the case of strong

GxE, a lack of correlation among environments. Consequently, genetic correlations among environments in the TPE can be used to increase the prediction accuracy across environments in the target region (Burgueño et al. 2012). In one of the earlier papers involving GxE, Burgueño et al. (2011) compared linear mixed models and factor analytic models for their predictive ability. When GxE was important, modeling GxE using the factor analytic model improved prediction accuracy; otherwise when GxE was not significant, most models gave relatively high prediction accuracies. Burgueño et al. (2012) examined wheat MET using GS multi-environment (multi-trait) models and evaluated their predictive accuracy with and without pedigree and marker information. In their cross validation, they predicted either the performance of untested genotypes or the performance of genotypes that had been evaluated in only some environments. Models that included both markers and pedigrees were superior to those that included either alone. Additionally, prediction accuracies were higher for predicting the performance of genotypes in untested environments than for predicting untested genotypes. They concluded that prediction accuracy could be improved using multi-environment GS models.

Modeling GxE when genotypic and environmental data are highly dimensional can present computational problems. More recent papers have focused on marker-by-environment interaction (MxE) effects. Jarquin et al. (2014) proposed a variance components approach where they used covariance functions to model high-dimensional interactions between markers and environmental covariates (ECs) for wheat and maize. In principle, it should be possible to model GxE by regressing phenotypes on markers and ECs and partitioning the GxE. In the reaction norm model, genetic and environmental gradients are described using a linear regression on genetic markers and on ECs. They used 68 ECs related to different crop developmental stages and compared interactions with main effects. Prediction models that included the interaction terms were 17–34% more accurate than models based only on main effects.

Lado et al. (2016) also used the correlations among environments to design sets of environments having low GxE to try and better predict genotype performance in untested environments. They used mixed models to generate the variance–covariance matrix across environments in a large, highly unbalanced, historical dataset from a wheat breeding program to obtain predictions within or across different sets of environments. They grouped environments into three mega-environments (MEs) based on a genotype-by-GxE biplot. The best predictions were within years across locations or within MEs for a given year or location. They concluded that borrowing information from environments using a variance–covariance matrix was useful for predicting new genotypes prior to phenotyping. Cuevas et al. (2017) presented results using a Bayesian genomic kernel model to account for the correlation between environments. This model accounting for GxE always showed superiority to models that only assessed one environment.

5.5.5 Dissecting GxE Using Environmental Covariates and Crop Models

As described above, GxE can be taken into account using multiplicative mixed models such as the factor analytic structure to model the covariance between environments responsible for GxE. However, those approaches have numerical limitations because of the highly unbalanced nature of multi-environment plant breeding datasets. Also, because they are based on observed covariance among environments, they are only explanatory of past performance rather than predictive of future performance. Another interesting approach to better understand and predict GxE involved the integration of ECs into the genomic selection framework to predict GxE deviations for unobserved environments (Heslot et al. 2014). Including environmental covariates in the analysis presented some of the same issues encountered using GS methods such as a high number of covariates, each explaining a small amount of the total variance while being highly correlated with each other.

Heslot et al. (2014) modeled genome-wide markers and their differential response to the environment to better understand the genetic architecture of GxE. Using more than 2000 winter wheat lines grown in 44 environments over 6 years in France, daily weather data (AGRI4CAST), and a wheat crop model known as SirusQuality (Martre et al. 2006), they first synchronized the developmental stages of the crop with the climatic conditions during those stages. Stress covariates (climatic variables at a specific developmental stage) were derived by developmental stage by using knowledge about the sensitivity of specific growth stages to abiotic stresses. The stress covariates were then used as independent variables in statistical genetic models for effect estimation and prediction. The factorial regression model was extended to the genomic selection context, and for each marker, they fit a main effect and a sensitivity to each of the stress covariates. A machine-learning algorithm was used to capture the interactions between markers and stress covariates as well as nonlinear effects. Genotype performance was predicted as a main effect plus a GxE deviation.

To deal with the high dimension of n markers by n covariate predictors, they assessed the variance of marker effects across environments and eliminated those explaining little or no variation. The photoperiod sensitivity gene *Ppd-D1* had the highest variance but alone did not capture a significant part of the GxE variance. The optimal model based on cross validation used 250 markers plus the nonlinear soft rule fit component. The most important stress covariate was the sum of the average daily temperature between meiosis and flowering. The second most important was drought in the early spring measured by “total number of dry days to 350 degree days” and the “sum of precipitation and evapotranspiration potential.” Heat stresses before flowering and during early grain fill were also important covariates.

A factor analytic model predicted a GxE response for any genotype in any environment, even if an environment had no phenotypic data for that genotype.

Euclidean distances could be calculated for all environments based on the predicted level of genetic correlation between environments, and a cluster analysis can be used to reveal the structure of the TPE. By including the GxE component, they were able to increase prediction accuracy for genotype performance in unobserved environments by 11.1% on average and the variability in prediction accuracy decreased by 10.8%. In contrast to the approaches that used covariances among environments to improve prediction accuracy, the use of carefully selected stress covariates allowed for prediction in unobserved environments rather than a retrospective view of GxE. This approach provides important information to the breeder because it offers a mechanism to leverage agronomy and physiology knowledge, reduce dimensionality and nonlinearity, use existing breeding data, and interpret results to identify specific environmental stresses.

Although their research involved maize, it is important to consider the application of the crop growth model and whole-genome prediction proposed by Technow et al. (2015) and empirically evaluated by Cooper et al. (2016). Cooper et al. (2016) were able to predict drought tolerance of a set of doubled haploids (DHs) (from the same cross) in maize hybrids using five measures of crop growth in a model that incorporated whole-genome prediction and an algorithm based on approximate Bayesian computation. As expected, prediction accuracies were high when predicting entries in the same environment (0.53–0.82) but generally low when predicting test DH entries in a new environment (0.22–0.38). If this approach is able to deliver a means for predicting genotype performance in the environments and management practices of the TPE, the breeder could use predicted performance of genotypes for important environment types of the TPE instead of using predictions based only on performance across environments.

5.6 Cost Benefit Analysis of GS

The breeder’s equation (Eq. 1) provides the information needed to optimize selection strategies for different GS applications. While manipulation of any of the variables in the formula can drive the rate of genetic gain, the majority of GS work has focused on shortening the length of time per cycle as the increase in genetic gain would increase proportionally. Any strategy that increases genetic gain could be used in a breeding program; however, only strategies that allow GS to surpass the rate of gain achieved by PS or where GS is sufficiently cheaper than PS will find applications in breeding programs.

Heffner et al. (2010) addressed this question by comparing the cost and genetic gain per unit time for conventional MAS and GS. Their results indicated that GS could achieve greater genetic gain than MAS on a per-year basis, even when GEBV accuracies are low. They predicted that given a prediction accuracy of 0.5, the expected annual gain from GS would exceed that of MAS by up to threefold for a high-intensity maize breeding program and up to twofold for a low-intensity winter wheat breeding program. The advantage realized by GS was almost entirely due to the shorter breeding cycle. The case for decreasing breeding cycle time has also been shown for introgressing wild alleles where Bernardo (2009) showed that the rate of genetic gain using 7–8 cycles of GS was higher than two cycles of phenotypic test-cross selection in maize.

Genomic selection may also find use where the trait of interest is challenging to measure. For example, FHB susceptibility and mycotoxin DON levels require laborious and expensive phenotyping. Using GS to predict FHB susceptibility and DON levels, Lorenz et al. (2012) found high prediction accuracy 0.72 and 0.68 for FHB and DON, respectively. These accuracies were equal to phenotypic selection and were estimated to cost only 25% as much as PS. Rutkoski et al. (2012) also reported high prediction accuracies (>0.62) for DON when markers were used in combination with phenotypes.

Early work in wheat quality by Heffner et al. (2011a) found that low marker density (256 markers) and small training population sizes (96 genotypes) in biparental crosses resulted in 0.66 ratio between the GS prediction and PS. Assuming that two GS cycles could be completed per year, they estimated that GS could provide more gains than PS for all nine of the wheat quality traits studied at 1/3rd the cost of PS. Further work in wheat by Battenfield et al. (2016) demonstrated that application of GS for predicting milling and baking quality traits in wheat had the potential to substantially outperform PS. The prediction accuracies for complex and expensive phenotypes in the CIMMYT bread wheat breeding program such as mixing time and loaf volume in the quality lab were moderate ranging from 0.32 (grain hardness) to 0.62 (mixing time). However, even with these moderate levels of prediction accuracy, the lower cost and higher throughput of genotyping relative to phenotyping for milling and baking gave substantial advantage to GS. Based on the current implementation of ten times more samples being genotyped in the breeding program than the capacity to phenotype for quality, they calculated a 1.4–2.7 higher rate of genetic gain for GS over PS for the quality traits.

Heslot et al. (2015) made a case for consideration of the problem of optimal resource allocation to obtain maximum genetic gains. One example of optimizing resources (Endelman et al. 2014) evaluated the optimal design of the preliminary yield trial. Using biparental barley and maize populations, they found that up to a 5% increase in genetic gain could be achieved if genotyping was 25% of the cost of one yield plot unit assuming a breeding program with 250 yield plot units per family. While GS prediction accuracy shows positive genetic gain, more cost benefit studies are needed (Heslot et al. 2015).

5.7 Summary

Breeding programs are dynamic entities, and consideration of the cost at each stage is required to optimize gain. For example, there will be questions about what germplasm to use, correlations among traits, trade-offs between family size and number of families, balance between phenotypic and GS or MAS at a constant budget, relationship between the training or mapping population, and the breeding germplasm. Some of the efficiencies that can be realized with GS and whole-genome genotyping include increased gain by genotyping individuals that are also phenotyped, more efficient experimental designs (sparse testing), reduced nursery sizes, and selection for costly traits or traits that are not expressed in each season. While researchers have worked to provide answers to many of these questions (e.g., Endelman et al. 2014; Heslot et al. 2015; Hickey et al. 2014), there are still many uncharted courses for how GS could play within individual breeding programs that are driven by different goals, environments, and policies. GS has considerable potential for improving quantitative traits, and new approaches for implementation of GS will continue to evolve in applied breeding programs.

References

Andrade-Sanchez P, Gore MA, Heun JT, Thorp KR, Carmo-Silva AE, French AN et al (2014) Development and evaluation of a field-based high-throughput phenotyping platform. Funct Plant Biol 41(1):68–79

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19(1):52–61

Arruda MP, Brown PJ, Lipka AE, Krill AM, Thurber C, Kolb FL (2015) Genomic selection for predicting fusarium head blight resistance in a wheat breeding program. Plant Genome 8 (3):1–12

Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J-L (2011) Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Genome J 4(2):132

Asoro FG, Newell MA, Beavis WD, Scott MP, Tinker NA, Jannink JL (2013) Genomic, marker-assisted, and pedigree-BLUP selection methods for β-glucan concentration in elite oat. Crop Sci 53(5):1894–1906

Battenfield SD, Guzmán C, Gaynor RC, Singh RP, Peña RJ, Dreisigacker S et al (2016) Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. Plant Genome 9(2):1–12

Bernardo R (2009) Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Sci 49(2):419

Bernardo R (2016) Bandwagons I, too, have known. Theor. Appl. Genet. Springer, Berlin Heidelberg. 129(12):2323–2332.

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47(3):1082–1090

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Burgueño J, Crossa J, Cotes JM, Vicente FS, Das B (2011) Prediction assessment of linear mixed models for multienvironment trials. Crop Sci 51(3):944–954

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. Crop Sci 52(2):707–719

Busemeyer L, Ruckelshausen A, Möller K, Melchinger AE, Alheit KV, Maurer HP et al (2013) Precision phenotyping of biomass accumulation in triticale reveals temporal genetic patterns of regulation. Sci Rep 3:2442

Chun H, Keleş S (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Ser B Stat Methodol 72(1):3–25

Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theor Appl Genet 867–887.

Cooper M, Technow F, Messina C, Gho C, Totir LR (2016) Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial. Crop Sci 56:1–16

Crain JL, Reynolds MP, Poland JA (2016) Utilizing high-throughput phenotypic data for improved phenotypic selection of stress adaptive traits in wheat. Crop Sci

Crossa J (2012) From genotype x environment interaction to gene x environment interaction. Curr Genomics 13(3):225–244

Crossa J, Jarquín D, Franco J, Pérez-Rodríguez P, Burgueño J, Saint-Pierre C et al (2016a) Genomic prediction of gene bank wheat landraces. G3 Genes|Genomes|Genetics 6 (7):1819–1834

Crossa J, de los Campos G, Maccaferri M, Tuberosa R, Burgueño J, Pérez-Rodríguez P (2016b) Extending the marker X environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Sci 56(5):2193–2209

Crossa J, De Los CG, Pérez P, Gianola D, Burgueño J, Araus JL et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2):713–724

Cuevas J, Crossa J, Montesinos-Lopez O, Burgueno J, Perez-Rodriguez P, de los Campos G (2017) Bayesian genomic prediction with genotype × environment interaction kernel models. G3 Genes|Genomes|Genetics 7(1):41–53

Cuevas J, Crossa J, Soberanis V, Pérez-Elizalde S, Pérez-Rodríguez P, de los Campos G et al (2016) Genomic prediction of genotype x environment interaction kernel regression models. Plant Genome 9(3):1–20

Daetwyler HD, Bansal UK, Bariana HS, Hayden MJ, Hayes BJ (2014) Genomic prediction for rust resistance in diverse wheat landraces. Theor Appl Genet 127(8):1795–1803

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185(3):1021–1031

Dawson J, Endelman J, Heslot N (2013) The use of unbalanced historical data for genomic selection in an international wheat breeding program. F Crop Res 154:12–22

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

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4(3):250–255

Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, Sorrells ME et al (2014) Optimal design of preliminary yield trials with genome-wide markers. Crop Sci 54(1):48–59

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM et al (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95(7):4114–4129

Falconer DS, Mackay TFC (1996) Quantitative Genetics, 4th edn. Pearson, New York

Fè D, Cericola F, Byrne S, Lenk I, Ashraf BH, Pedersen MG et al (2015) Genomic dissection and prediction of heading date in perennial ryegrass. BMC Genomics 16(1):921

Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 185:1–17

Gianola D, Van Kaam JBCHM (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178(4):2289–2303

Gorjanc G, Jenko J, Hearne SJ, Hickey JM (2016) Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17(1):30

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186

Habier D, Tetens J, Seefried F-R, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42(1):5

Hayashi T, Iwata H (2010) EM algorithm for Bayesian estimation of genomic breeding values. BMC Genet 11:3

He S, Schulthess AW, Mirdita V, Zhao Y, Korzun V, Bothe R et al (2016) Genomic selection in a commercial winter wheat population. Theor Appl Genet 129(3):641–651

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

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

Heffner EL, Lorenz AJ, Jannink JL, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci 50(5):1681–1690

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49(1):1–12

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127(2):463–480

Heslot N, Jannink J-L, Sorrells ME (2015) Perspectives for genomic selection applications and research in plants. Crop Sci 55(1):1–12

Heslot N, Jannink JL, Sorrells ME (2013a) Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Sci 53(3):921–933

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

Heslot N, Yang H-PP, Sorrells MEMEME, Jannink J-LL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52(1):146

Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM et al (2014) Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci 54(4):1476–1488

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

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127(3):595–607

Jia Y, Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192(4):1513–1522

Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. Genetics 201(2):759–768

Jiang Y, Zhao Y, Rodemann B, Plieske J, Kollers S, Korzun V et al (2015) Potential and limits to unravel the genetic architecture and predict the variation of Fusarium head blight resistance in European winter wheat (Triticum aestivum L.) Heredity (Edinb) 114:318–326

Lado B, González Barrios P, Quincke M, Silva P, Gutiérrez L (2016) Modeling genotype x environment interaction for genomic selection with unbalanced data from a wheat breeding program. Crop Sci 56(April):1–15

Longin CFH, Mi X, Würschum T (2015) Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. Theor Appl Genet Springer Berlin Heidelberg 128(7):1297–1306

Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink J-LJ et al (2015) Increased prediction accuracy in wheat breeding trials using a marker x environment interaction genomic selection model. G3 Genes|Genomes|Genetics 5(4):569–582

Lorenz AJ (2013 Mar) 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

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H et al (2011) Genomic selection in plant breeding. Adv Agron 110:77–122

Lorenz AJ, Smith KP, Jannink JL (2012) Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci 52(4):1609–1621

Martre P, Jamieson PD, Semenov MA, Zyskowski RF, Porter JR, Triboi E (2006) Modelling protein content and composition in relation to crop nitrogen dynamics for wheat. Eur J Agron 25(2):138–154

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

Mirdita V, He S, Zhao Y, Korzun V, Bothe R, Ebmeyer E et al (2015) Potential and limits of whole genome prediction of resistance to Fusarium head blight and Septoria tritici blotch in a vast Central European elite winter wheat population. Theor Appl Genet 128(12):2471–2481

Montesinos-López OA, Montesinos-López A, Crossa J, Toledo F, Pérez-Hernández O, Eskridge KM et al (2016) A genomic bayesian multi-trait and multi-environment model. G3 Genes|Genomes|Genetics 6(9):2725–2744

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

Ornella L, Singh S, Perez P, Burgueño J, Singh R, Tapia E et al (2012) Genomic prediction of genetic values for resistance to wheat rusts. Plant Genome J 5(3):136–148

Park T, Casella G, Ark TP, Asella GC (2008) The Bayesian lasso. J Am Stat Assoc 103 (482):681–686

Perez-Rodriguez P, Gianola D, Gonzalez-Camacho JM, Crossa J, Manes Y, Dreisigacker S (2013) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 Genes|Genomes|Genetics 2(12):1595–1605

Pérez P, de los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Genome 3(2):106–116

Podlich DW, Cooper M, Basford KE (1999) Computer simulation of a selection strategy to accommodate genotype-environment interactions in a wheat recurrent selection programme. Plant Breed 118(1):17–28

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y et al (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome J 5(3):103

Pszczola M, Strabel T, Mulder HA, Calus MPL (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci 95 (1):389–400

Rife TW, Wu S, Bowden R, Poland JA (2015) Spiked GBS: a unified, open platform for single marker genotyping and whole-genome profiling. BMC Genomics 16(1):1–7

Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press. Cambridge, UK

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

Rutkoski J, Poland J, Mondal S, Autrique E, González Párez L, Crossa JJ et al (2016) Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3 Genes|Genomes|Genetics 6 (9):2799–2808

Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL et al (2015a) Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. Plant Genome 8(2):1–10

Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL et al (2015b) Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. Plant Genome 8(1):1–10

Rutkoski JE, Poland JA, Singh RP, Huerta-espino J, Barbier H, Rouse MN et al (2014) Genomic selection for quantitative adult plant stem rust resistance in wheat. Plant Genome 7(3):1–10

Rutkoski JE, Poland J, Jannink JL, Sorrells ME, Breeding P, York N (2013) Imputation of unordered markers and the impact on genomic selection accuracy. G3 Genes|Genomes|Genetics 3(3):427–439

Sallam AH, Endelman JB, Jannink J-L, Smith KP (2015) Assessing genomic selection prediction accuracy in a dynamic barley breeding population. Plant Genome 8(1):1–15

Schmidt M, Kollers S, Maasberg-Prelle A, Großer J, Schinkel B, Tomerius A et al (2016) Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. Theor Appl Genet 129(2):203–213

Schulthess AW, Wang Y, Miedaner T, Wilde P, Reif JC, Zhao Y (2016) Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. Theor Appl Genet 129(2):273–287

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222

Storlie E, Charmet G (2013) Genomic selection accuracy using historical data generated in a wheat breeding program. Plant Genome 6(1):1–9

Technow F, Messina CD, Totir LR, Cooper M (2015) Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PLoS One 10(6):e0130855

Thavamanikumar S, Dolferus R, Thumma BR (2015) Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. G3 Genes|Genomes|Genetics 5(October):1991–1998

Tibshirani R (1996) Regression selection and shrinkage via the lasso. J R Stat Soc B 128:267–288

Wang Y, Mette MF, Miedaner T, Gottwald M, Wilde P, Reif JC et al (2014) The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. BMC Genomics 15(1):556

Ward J, Rakszegi M, Bedő Z, Shewry PR, Mackay I (2015) Differentially penalized regression to predict agronomic traits from metabolites and markers in wheat. BMC Genet 16(1):1–7

White J, Andrade-Sanchez P, Gore MA, Bronson KF, Coffelt TA, Conley MM et al (2012) Field-based phenomics for plant genetics research. F Crop Res 133:101–112

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75(2):249–252

Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63(2):513–521

Zhang X, Sallam A, Gao L, Kantarski T, Poland J, DeHaan LR et al (2016) Establishment and optimization of genomic selection to accelerate the domestication and improvement of intermediate wheatgrass. Plant Genome 9(1):1–18

Zhao Y, Li Z, Liu G, Jiang Y, Maurer HP, Würschum T et al (2015) Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc Natl Acad Sci U S A 112 (51):15624–15629

Zhao Y, Mette MF, Gowda M, Longin CFH, Reif JC (2014) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity (Edinb) 112(6):638–645

Zou H, Hastie T (2005) Regularization and variable selection via the elastic-net. J R Stat Soc 67 (2):301–320

Chapter 6
Current Status and Prospects of Genomic Selection in Legumes

Ankit Jain, Manish Roorkiwal, Manish K. Pandey, and Rajeev K. Varshney

6.1 Introduction

Availability of proper nutrition is of extreme importance as malnutrition at an early age may lead to reduced physical and mental development and limits the capacity to learn. UN World Food Program has reported that more than 900 million people in the world do not get nutritious food to eat. Global population has been growing at a fast pace, and feeding the ever increasing population with nutritious food is becoming more difficult day by day. This will continue until there is significant genetic gain by increasing crop productivity with enhanced nutrition. Although significant efforts have been focussing on enhancing the crop production to feed the world, still there are famines occurring in several parts of the world (http://www.latimes.com/world/africa/la-fg-southsudan-famine-20170220-story.html). Considering this alarming situation, the United Nations and other affiliated organizations have a challenge to eradicate hunger and malnutrition to ensure food and nutrition security by responding to nutritional needs, addressing emerging threats and meeting the zero hunger challenge. To overcome this devastating situation of malnutrition, legumes are expected to play significant role, and there is a dire need to enhance the productivity of these legumes.

Legumes have been cultivated since early civilizations and have been the major source of nutrition for humans and animals (Power 1987; Graham and Vance 2003; Varshney et al. 2013a; Rubiales and Mikic 2015; Pandey et al. 2016). Legumes have been recognized as most valuable food to meet the dietary requirements of undernourished or underserved global populations (Rebello et al. 2014). Research has shown that replacement of energy dense foods with legumes offers various

A. Jain • M. Roorkiwal (✉) • M.K. Pandey • R.K. Varshney
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru,
Telangana, India
e-mail: m.roorkiwal@cgiar.org

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7_6

health benefits (Tarawali and Ogunbile 1995). In addition, legumes have the ability to fix atmospheric nitrogen, which is vital for improving the soil nutritional profile, thereby reducing the requirement for nitrogen fertilizers enabling legumes more suited for crop rotation programs.

Legumes are among the important crop commodities and have high demand being a major supplement of protein, but the productivity is low compared with the increasing demand resulting from several biotic (Rubiales and Mikic 2015) and abiotic stresses (Araújo et al. 2015). The productivity trends for these legumes in the last five decades suggest very little improvement leading to low productivity in most of the legumes compared with cereal crops (FAOSTAT 2014). Nevertheless, several efforts made in these years identified the genetic variations for various traits of interest in these legumes to enhance the crop productivity. So far, limited success could be achieved with the application of conventional breeding approaches for enhancing the crop productivity by overcoming key constraints. It is time to adopt modern and new technologies for enhancing the rate of genetic gain, so that improved varieties can be developed faster and more precisely equipped with essential traits to face the climate and other stress factors.

A paradigm shift is required in approaches and breeding methodologies to develop superior varieties for the future. In this context, deployment of genomics tools and technologies has shown great potential in understanding the complex genetics and breeding problems. It has been realized that genomics-assisted breeding (GAB), with integration of conventional breeding is the key to overcome conventional breeding limitations (Varshney et al. 2013a). Further in the case of legumes, a journey from a status of orphan crops with a dearth of genomic resources a decade ago, to current well-enriched genomic resource crop status, opened the possibility of deployment of GAB for these crops. Additionally, recent advent of the next-generation sequencing (NGS) technologies had brought down the sequencing and genotyping cost significantly. As a result, draft genomes have become available for several legume crops including model legumes, i.e., *Medicago truncatula* (Young et al. 2011), *Lotus japonicus* (Sato et al. 2008) and crops such as *Glycine max* (Soybean) (Schmutz et al. 2010), *Cajanus cajan* (Pigeonpea) (Varshney et al. 2012), *Cicer arietinum* (Chickpea) (Varshney et al. 2013b; Jain et al. 2013); *Lupinus angustifolius* (Lupin) (Yang et al. 2013), *Vigna radiata* (Mung bean) (Kang et al. 2014) and *Arachis duranensis* and *A. ipaensis* (progenitors of cultivated groundnut) (Bertioli et al. 2016; Chen et al. 2016). Genome sequencing efforts followed by large scale re-sequencing efforts in each crop led to availability of millions of structural variations leading to availability of large numbers of genetic markers (see Varshney et al. 2013a; Bohra et al. 2014; Pandey et al. 2016).

Availability of large scale genome-wide genetic markers led to establishment of several high-throughput genotyping platforms, offering precise, rapid and cost-effective solutions to genotyping of large populations. For instance, informative single nucleotide polymorphisms (SNPs) with high genome density are being chosen and used to design assays/platforms for legumes such as in *Vigna unguiculata* (Egbadzor et al. 2014; Huynh et al. 2013; Lucas et al. 2013, Muñoz-Amatriaín et al. 2016), *Pisum sativum* (Deulvot et al. 2010; Bordat et al. 2011; Tayeh et al. 2015), *Lens culinaris* (Sharpe et al. 2013; Kaur et al. 2014a), *Vicia faba*

(Kaur et al. 2014b), soybean (Lee et al. 2015; Wang et al. 2016), chickpea (Gujaria et al. 2011; Hiremath et al. 2011; Roorkiwal et al. 2014), pigeonpea (Saxena et al. 2012) and groundnut (Pandey et al. 2017). Other alternative SNP detection systems like competitive allele-specific PCR (KASPar) (Cottage et al. 2012; Hiremath et al. 2012; Kumar et al. 2012; Saxena et al. 2012; Xu et al. 2012; Fedoruk 2013; Khera et al. 2013; Sharpe et al. 2013), custom-designed Illumina VeraCode assay (Deulvot et al. 2010; Roorkiwal et al. 2013, Duarte et al. 2014) have also been employed for various applications. The development and deployment of different genotyping platforms provide cost effective and precise genotyping solution to many legume crops leading to enhanced rate of progress in legume genomics. NGS-based genotyping by sequencing (GBS) allows simultaneous marker discovery as well as genotyping of the populations even in the absence of a reference genome (Davey et al. 2011). Among legumes, the GBS approach has been successfully used in lentil (Ates et al. 2016) and chickpea (Deokar et al. 2014; Jaganathan et al. 2015; Verma et al. 2015) for genome-wide SNP discovery and genetic mapping. Further, whole genome re-sequencing (WGRS) and restriction site-associated DNA (RAD) sequencing approaches have also been used to capture the variations in the genome and to understand diversity prevailing in the germplasm (see Varshney et al. 2013b).

GAB aims at to accelerate crop improvement by establishing and exploiting the relationships between genotype and phenotype. Of the three GAB approaches, marker-assisted backcrossing (MABC), marker-assisted recurrent selection (MARS) and genomic selection (GS), MABC has been deployed in most of the crops and proved to be an effective approach for development of improved varieties and lines in many legume crop plants (see Pandey et al. 2016). MABC uses markers linked to agronomical important traits and mainly aims at introgression of a limited number of alleles from one genetic background (donor) to other (recipient) (Hospital 2005). Further, the improved varieties developed as a result of MABC contain one or a few alleles at major gene/QTLs from the donor genotype, keeping intact the rest of the genome from recurrent parent (see Varshney et al. 2013a). For instance, one "*QTL-hotspot*" region having QTLs for several drought tolerance-related root traits was introgressed into JG11, a desi chickpea cultivar from the drought tolerant line ICC4958 (Varshney et al. 2013c). Similarly introgression lines developed using MABC for fusarium wilt (FW) and ascochyta blight (AB) resistance in the background of C214 have shown enhanced resistance for FW and AB (Varshney et al. 2014). In the case of groundnut, MABC has been exploited to introgress major QTLs for leaf rust resistance from GPBD 4, a leaf rust resistant cultivar into ICGV 91114, JL 24 and TAG 24 cultivars (Varshney et al. 2014). MABC along with MAS was further deployed in enhancing the oil quality by increasing oleic acid in three different groundnut varieties, viz. ICGV 06110, ICGV 06142 and ICGV 06420 (Janila et al. 2016). In the case of pea, Aphanomyces root rot resistance QTLs (Lavaud et al. 2015) and frost tolerance QTLs (Hascoët et al. 2014) were introgressed using MABC into different agronomically important genetic backgrounds. Likewise in soybean, MABC was deployed successfully to improve resistance to a defoliating insect (Zhu et al. 2007), bacterial leaf pustule resistance (Kim et al. 2008) and to reducing a kunitz trypsin inhibitor (Kumar et al. 2015).

In order to address the limitations of MABC approach for improving multiple complex traits, MARS has been proposed for combining major and minor QTLs in several crops. In the case of MARS, the de novo QTL identification is carried out in a breeding population derived from the crosses of superior varieties followed by crossing genotypes with superior alleles for pyramiding targeted QTLs into one or more genetic backgrounds (Bernardo and Charcosset 2006). However, the MARS approach was not effective for increasing yield in chickpea (Pandey et al. 2016). MARS was suggested a method for improvement of drought tolerance in groundnut, however more than 100 main and epistatic effect QTLs were reported because handling these small effect QTLs through MABC was not possible (Gautami et al. 2012).

GS utilizes phenotypic as well as genome-wide marker data to predict the genomic-estimated breeding values (GEBV) for selecting the superior lines. In brief, two populations, training population and testing population (sometimes, it is part of training population, hence known as validation set as well) are used. Training population is the one with comprehensive phenotypic data under different environmental conditions, that is, different locations/seasons/treatments. Genome-wide genotypic and phenotypic data for the training population are used to train different statistical GS models. The training population can be subdivided into five to ten groups, and then, cross validation is used to evaluate the GS models and prediction accuracy. Trained models, are used to calculate GEBV of a testing or selection candidate population that has been genotyped but not phenotyped. The predicted GEBVs are used to select superior lines from the population. One of the advantages associated with GS is that it reduces the selection cycle length by eliminating the phenotyping that is required for multiple rounds of selection hence reducing time and cost, leading to genetic gain.

Genomic prediction is a key to success in GS breeding, and it depends on high-throughput and high-density genotyping along with accurate, multilocation phenotyping data. Availability of ample genomic resources and affordable high-density and high-throughput genotyping in several legumes will facilitate deployment of GS in legumes. This chapter briefly describes the critical factors determining the success of genomic selection and summarises the ongoing efforts to deploy genomic selection in legumes and further the existing possibilities by integrating available genomic resources to harness the full potential of modern breeding approaches.

6.2 Critical Factors in Deployment of Genomic Selection

High-precision prediction accuracies are the most critical point that determines the success of any GS breeding program. Multiple simulation and empirical studies involving estimation of prediction accuracies rely on multiple factors *viz.* number and type of markers (Chen and Sullivan 2003; Poland and Rife 2012), population structure (Nakaya and Isobe 2012; Spindel et al. 2015), training population size (Daetwyler et al. 2008), heritability and architecture of target traits (Zhong et al. 2009; Zhang et al. 2014, 2016) and the relationship between training population and selection candidates.

Numerous GS models have been proposed to address the diverse requirements for achieving satisfactory prediction accuracies. Some of the routinely used GS models include Random Regression Best Linear Unbiased Predictor (RR-BLUP; Meuwissen et al. 2001; Liu et al. 2008; Zhang et al. 2010), Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996; de los Campos et al. 2009a), semiparametric strategies (Kinship GAUSS), Bayesian approach viz. Bayesian Ridge Regression, Bayesian LASSO (de los Campos et al. 2009b; Legarra et al. 2011), Bayes A (Meuwissen et al. 2001), Bayes B (Meuwissen et al. 2001) and Bayes Cπ (Habier et al. 2011) and machine learning Random Forest Regression (RFR) (Breiman, 2001), and Support Vector Regression (SVR) (Drucker et al. 1997). Various comparative accounts have been drawn to assess the performances of these GS models among different organisms (Moser et al. 2009, Heslot et al. 2012, Resende et al. 2012a, b). Selection of an appropriate GS model varies from case to case, and hence, multiple models should be considered in any GS study.

Size of training population is another important factor that has significant impact on prediction accuracies. Bernardo and Yu (2007) suggested that a minimum size of the training population to be 100–150 genotypes to obtain the optimum prediction accuracy. In the case of genetically diverse populations, larger training populations are required to attain better prediction accuracies (Mujibi et al. 2011). Genetic relatedness of the individuals in the training and selection populations is known to affect the accuracies of GS studies (Asoro et al. 2011). Among cattle, GEBVs estimated within breed were found to be more accurate than the ones estimated across breeds (Hayes et al. 2009). Price et al. (2010) and Guo et al. (2014) demonstrated significant reduction in prediction accuracies in structured populations.

Application of genome-wide markers results in better prediction accuracies (Meuwissen et al. 2001; Calus and Veerkamp 2007). Higher marker density has been demonstrated to produce higher genomic prediction accuracy (Zhong et al. 2009; Asoro et al. 2011; Heffner et al. 2011; Poland et al. 2012; Heslot et al. 2013). Low marker densities in some cases result in lower prediction accuracies, that could be explained as lower probability of LD between markers and QTLs, because of the smaller fraction of variation (Solberg et al. 2008). Hickey et al. (2014) reported that a small number of markers (200–500) and phenotypes (1000) are required in a closely related biparental population to achieve effective prediction accuracies, whereas for a population that is unrelated to the selection candidates, a much larger number of markers and phenotypes are required for the same prediction accuracy. A large mixed training population set with higher marker density is recommendable to achieve high prediction accuracies rather than using multiple training populations representing one germplasm group (Asoro et al. 2011). In another study, De Roos et al. (2009) suggested that a high marker density is required if training and selection populations are highly divergent.

High-throughput genotyping platforms such as DArT, SNP array and GBS are being used based on different needs. GBS has been deployed in almost all the crops in the initial genetic analysis as it provides a low cost option to plant species where there is no reference genome (Poland et al. 2012). A comparison made by Poland et al. (2012) using GBS for de novo genotyping of testing populations in case of the wheat (*Triticum aestivum* L.) genome showed higher prediction accuracies of 0.3–0.5 in comparison to established marker platforms.

Enhancing the marker numbers while imputing the missing marker data has been reported to improve in prediction accuracies. For instance, Poland et al. (2012) showed an improvement of prediction accuracies with the genotyping data set consisting of 35,000 SNPs with up to 80% missing data points, over the prediction accuracies estimated from 2000 DArT markers with missing data points up to 2%. In various studies including maize, wheat, barley and forest trees, a positive relationship between the trait heritability and prediction accuracies has been observed (Lorenzana and Bernardo 2009; Albrecht et al. 2011; Heffner et al. 2009, 2011; Grattapaglia et al. 2011; Guo et al. 2012; Combs and Bernardo 2013). In another study, Zhang et al. (2014) established higher prediction accuracies for less complex traits. Most of the results discussed here form the basis of ongoing efforts in legume genomic selection and serve as the guidelines for strategizing the future efforts. GS efforts in different legumes have been described below in detail.

6.3 Soybean (*Glycine max*)

Deployment of GS among legumes first started with improving yield and agronomic traits in soybean. A set of 301 elite breeding lines was genotyped with GBS and phenotyped for grain yield at multiple locations (Table 6.1) (Jarquín et al. 2014). By keeping a randomly selected set of 50 accessions for a validation population, a positive relationship was observed between the size of training population and prediction accuracy, which began to plateau at a training population size of 100; however, it continued to increase until the maximum available size. The study included the evaluation of three different imputation methods to impute the missing data for soybean. However, not many differences were obtained using these imputation methods. Although, random forest imputation produced the highest accuracies, no significant differences were observed. A high prediction accuracy (0.64) reflected high potential of GS for yield in soybean (Table 6.1) (Jarquín et al. 2014).

Further, exploiting the GAB, genotyping data for 31,045 SNPs on 309 soybean germplasm accessions were used to estimate the prediction accuracy for seed weight (SW) (Zhang et al. 2016). Five-fold cross validation (CV) was applied by randomly assigning 20% of the association panel as validation set and remaining

Table 6.1 Summary of key genomic selection studies in some legume crops

Legume crop	Population Size	Marker type	Traits	GS models	Reference
Soybean	301	GBS	1. Grain yield	A standard Genomic best linear unbiased prediction (G-BLUP) model including only additive effects, and an extended version of the G-BLUP model including additive-by-additive effects.	Jarquín et al. (2014)
Alfalfa	190	GBS (10,000 SNPs)	1. Single harvest biomass 2. Total biomass	Random Regression Best Linear Unbiased Predictor (RR-BLUP)	Li et al. (2015)
	278 (adapted to two different environment)	GBS	1. Dry matter yield	Support vector regression using linear and Gaussian kernel, RR-BLUP, random Forest regression and Bayes A, Bayes B and Bayesian lasso,	Annicchiarico et al. (2015)
Pea	372	331 SNP	1. Date of flowering 2. Number of seeds per plant 3. Thousand seed weight	LASSO (least absolute shrinkage ans selection operator), PLS (partial least squares), SPLS (sparse partial least squares), Bayes A, Bayes B and G-BLUP	Burstin et al. (2015)
	339	9824 SNPs (GenoPea 13.2 K SNP Array)	1. Date of flowering 2. Number of seeds per plant 3. Thousand seed weight	Kernel partial least squares regression (kPLSR), LASSO, G-BLUP, Bayes A, and Bayes B	Tayeh et al. (2015)
Chickpea	320	3000 DArT markers	1. Seed yield 2. 100 seed weight 3. Days to 50% flowering 4. Days to maturity	RR-BLUP, kinship GAUSS, Bayes Cπ, Bayes B, Bayesian LASSO, Random Forest (RF)	Roorkiwal et al. (2016)
Groundnut	188	2356 DArT markers	1. Days to flowering 2. Seed weight 3. Pod yield	RR-BLUP, kinship GAUSS, Bayes Cπ, Bayes B, Baysian LASSO and RF	Pandey et al. (2014b, 2015)

80% as the training set. Based on the number of SNPs used and the size of training population, the prediction accuracies were found to vary between 0.75 and 0.87. Like other studies (Asoro et al. 2011; Jarquin et al. 2014), on size of the training population, smaller populations resulted in lower prediction accuracies. Another observation was the prediction accuracy using all 2000 SNPs was found to be same, even reducing it to 500 SNPs. Higher prediction accuracies were observed compared to Jarquín et al. (2014) with same number of markers, similar population size, and broad sense heritability of traits, pointing towards the impact of genetic architecture of traits in populations under investigation.

6.4 Alfalfa (*Medicago sativa*)

Alfalfa is a perennial legume with a long breeding cycle, which limits crop improvement efforts. Selection cycle duration can be reduced by deploying GS for complex traits such as yield by using GS for predicting the breeding values (Li et al. 2015). Prediction accuracies were obtained using phenotyping data for yield traits during two selection cycles from three locations and using genotyping data for ~10,000 SNPs (Li et al. 2015). Varying levels of missing values from the marker data set were used for GS modelling using random forest method for missing values imputation. Validation of genomic prediction models was performed by cross validation, in which randomly selected 90% genotypes were used as training population and 10% was used for testing/validation. Marker data sets with more missing values resulted in a large number of markers and resulted in increased prediction accuracies. Prediction accuracies were validated for both the generation viz. cycle 0 and cycle 1. In individual generation analysis, prediction accuracies validated within locations were found to be much higher than prediction accuracies across the locations, possibility due to G × E interaction for biomass yield. Prediction accuracies of 0.43–0.66 for total biomass yield in a synthetic alfalfa breeding population showed the underlying potential of further application of GS in other complex traits (Li et al. 2015) (Table 6.1).

In total, 278 elite genotypes adapted to two different environments with a different genetic base were genotyped using GBS and phenotyped for dry matter yield of their densely planted half-sib progenies in separate environments (Annicchiarico et al. 2015). Prediction accuracies were higher using joint SNP calling in comparison to separate SNP calling for the two data sets. Random forest was used for missing marker imputation. A comparison of prediction accuracies within and across populations was performed with the same set of markers, and it was observed that within-population prediction accuracies were higher than across-population prediction accuracies, probably due to a high level of intra-population variation. Results indicated a greater than three-fold higher prediction for yield gain per unit time though GS in comparison to conventional selection (Annicchiarico et al. 2015) (Table 6.1).

6.5 Pea (*Pisum sativum*)

In the case of pea, SNP markers were used to predict the phenotypes using different statistical methods (Burstin et al. 2015). Phenotyping data for two seasons and genotyping data generated with 331 SNPs on >350 accessions representing various cultivars, diverse wild types, landraces, etc. were used to estimate the prediction accuracies (Table 6.1). To minimize the impact of population structure leading to spurious associations, authors used the approach recommended by Johnson et al. (2007). Thousand seed weight (TSW) was predicted better than the beginning of flowering (BegFlo) and number of seeds per plant (NSeed). During the same year, they reported deployment of a high-density genotyping platform for GS (Tayeh et al. 2015). Similarly, genotyping data from the GenoPea 13.2 K SNP Array on a collection of 339 accessions along with the phenotyping data for TSW, BegFlo and NSeed were used for estimating genomic prediction values using five different statistical methods (Tayeh et al. 2015). To estimate the impact of the training population size over the prediction accuracies, different sizes of training populations were selected randomly with multiple repetitions; however, the test set was fixed with 99 accessions. Similarly, to assess the effect of marker density on prediction accuracies, evenly distributed SNP subsets were selected for estimation. Of five models considered in the study, four showed equivalent performance, whereas performance of LASSO was less than others. Another highlight of the study was that no significant differences were observed whether or not the markers with low minor allele frequency (MAF) were included. The effect of a reduction in the size of the training population was reduction in accuracy of the prediction models (Q^2). In addition, reducing the marker density but retaining only a single marker per unique map position did not affect prediction accuracy. However, a further reduction in the number of markers led to reduced Q^2. Q^2 values obtained in Tayeh et al. (2015) were found to be higher than in Burstin et al. (2015).

6.6 Chickpea (*Cicer arietinum*)

In case of chickpea, there is only one report coming from ICRISAT about deploying GS breeding and conducting initial studies of standardizing different GS models (Roorkiwal et al. 2016). In this context, a training population containing 320 elite chickpea breeding lines consisting of desi and kabuli seed types, from the International Chickpea Screening Nursery (ICSN), was genotyped using the DArTseq platform. This platform generated 3000 polymorphic markers. Phenotyping data were generated for yield and yield-related traits *viz.* seed yield (SY), 100 seed weight (SDW), days to 50% flowering (DF) and days to maturity (DM), at two different locations during two different crop seasons for two different treatments, that is, rainfed and irrigated conditions. Six different statistical models were used to calculate prediction accuracies and perform five-fold cross validation to estimate the prediction accuracies by randomly selecting 80% of the lines for the training population and the remaining 20% as the testing population (Roorkiwal et al. 2016). A large variation in prediction accuracies were observed among the traits undertaken in the study, but overall performance of the models were found to be similar for every trait. The effect of G × E interaction was observed in the prediction accuracies of individual traits. For instance, the best prediction accuracy was observed for SDW (trait least affected by G x E interaction and treatments, etc.); however, prediction accuracies were lower for SY trait, which is known to be affected by G × E. The impact of missing marker data and MAF on prediction accuracies was assessed for 100 seed weight, using nine different combinations of missing marker data and MAF (including markers in combination with 0%, $\leq$10% and $\leq$30% missing data, and 0%, $\geq$5% and $\geq$10% MAF). The results showed that the random forest model at 0% missing marker data and $\geq$5% MAF combination had the best prediction accuracy, whereas the Bayes B model with 0% missing marker data and $\geq$10% MAF produced lowest accuracies. This study also assessed the impact of population structure on GEBV prediction accuracy. Desi and kabuli seed types were undertaken as separate groups and also grouped together to calculate prediction accuracies. The results reflected a higher prediction accuracy using the complete set in comparison to different seed types considered separately, which might be attributed to a larger population size (Roorkiwal et al. 2016) (Table 6.1).

6.7 Groundnut (*Arachis hypogaea*)

In case of groundnut, ICRISAT has taken some initiatives towards deploying GS breeding and conducting initial studies of standardizing different GS models (Pandey et al. 2016).While undertaking deployment of GS in groundnut, the focus of the study was to assess the impact of associated markers on prediction accuracies for three important traits viz. days to flowering (DF), seed weight (SW) and pod yield (PY) with different heritabilities (Pandey et al. 2014a, b; Pandey et al. 2015). Six seasons of phenotyping data for these traits and genotyping of the reference set with 2356 DArT markers were used for GS analysis (Table 6.1). When comparing the prediction accuracy for total and associated markers, the impact of population size and two different approaches were used to estimate the prediction accuracies. In the first approach, the whole population set was considered as a training population, and a part of the training population was considered as validation set to calculate the prediction accuracies. However, in another approach, the whole population was fractioned into five random smaller sets, of which one set was used to train the GS model, hence acted as training population, and the rest four were used as validation sets. Associated markers were compared with using all markers and the associated marker set showed higher prediction accuracies. However in a second approach where randomly selected smaller sets were used to genotype the training population, prediction accuracies obtained with associated markers were less predictive than all genome-wide markers. Overall, only marginal differences were observed between the prediction accuracies estimated using total genome-wide markers by both the approaches. As expected, the traits with higher heritability showed higher prediction accuracies in comparison to those with lower heritability. A positive relation between the heritability and prediction accuracies was observed, supporting similar observations in maize, wheat, barley, etc. (Lorenzana and Bernardo 2009; Albrecht et al. 2011; Heffner et al. 2011; Guo et al. 2012; Combs and Bernardo 2013). So far, the lack of a high-throughput genotyping platform to generate high-density genotyping data has been the major obstacle in deploying the GS breeding in groundnut. However, the availability of genome sequences of a diploid progenitor species and 58 K Axiom_*Arachis* SNP (Pandey et al. 2017) array during 2016 will further boost the deployment of GS breeding in groundnut.

6.8 Conclusions

The majority of legume crops lacked the attention of researchers for generating genomic resources for a longer time compared with cereal crops. Nevertheless, the speedy development in NGS technologies and assembly methodologies made generating genomic resources affordable and technically sound over the time. The legume crops have made much progress from poor resource to highly enriched genomic resourced crops. This has provided many opportunities to implement advanced genomic-assisted breeding. GS breeding has demonstrated its great value to the ongoing conventional breeding programs of cattle and in some plant species. This approach is gaining attention from other crop breeders including legumes as it promises greater genetic gain by improving complex traits in less time with more precision. Seeing the benefits achieved in the maize and wheat breeding programs, legume crops are now looking forward to deploying GS breeding to address its some of the most complex problems that are the key obstacles in achieving higher productivity. Selected studies conducted so far in legumes have suggested the possibility of achieving high prediction accuracies. These preliminary studies also indicated the potential role of GS in developing superior varieties with enhanced genetic gain and ability to overcome various stresses, hence ensuring food security with higher productivity. Currently, the majority of the legume crops are in the process of deploying GS in their breeding program; however, it will take a few years for GS to become routine similar to other major crop breeding programs.

References

Albrecht T, Wimmer V, Auinger H, Erbe M, Knaak C et al (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339–350

Annicchiarico P, Nazzicari N, Li X, Wei Y, Pecetti L et al (2015) Accuracy of genomic selection for alfalfa biomass yield in different reference populations. BMC Genomics 16:1020

Araújo SS, Beebe S, Crespi M, Delbreil B, González EM et al (2015) Abiotic stress responses in Legume crops: strategies used to cope with environmental challenges. Crit Rev Plant Sci 34:237–280

Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J (2011) Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. The Plant Genome 4:132–144

Ates D, Sever T, Aldemir S, Yagmur B, Temel HY et al (2016) Identification QTLs Controlling Genes for Se Uptake in Lentil Seeds. PLOS ONE 11(4): e0154054

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

Bernardo R, Yu J (2007) Prospects for genome-wide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD et al (2016) The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. Nat Genet 48:438–446

Bohra A, Pandey MK, Jha UC, Singh B, Singh IP et al (2014) Genomics-assisted breeding in four major pulse crops of developing countries: present status and prospects. Theor Appl Genet 127:1263–1291

Bordat A, Savois V, Nicolas M, Salse J, Chauveau A et al (2011) Translational genomics in legumes allowed placing in silico 5460 unigenes on the pea functional map and identified candidate genes in Pisum sativum L. Genes Genome Genet 1:93–103

Breiman L (2001) Random forests. Mach Learn 45:5–32. doi:10.1023/A:1010933404324

Burstin J, Salloignon P, Chabert-Martinello M, Magnin-Robert JB, Siol M et al (2015) Genetic diversity and trait genomic prediction in a pea diversity panel. BMC Genomics 16:105

Calus MPL, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J Ani Breed Genet 124:362–368

Chen X, Sullivan PF (2003) Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. Pharmacogenomics J 3:77–96

Chen X, Li H, Pandey MK, Yang Q, Wang X et al (2016) Draft genome of the peanut A-genome progenitor (Arachis duranensis) provides insights into geocarpy, oil biosynthesis, and allergens. Proc Nat Acad Sci 113:6785–6790

Combs E, Bernardo R (2013) Accuracy of genomewide selection for different traits with constant population size, heritability and number of markers. The Plant Genome 6:1

Cottage A, Gostkiewicz K, Thomas JE, Borrows R, Torres AM et al (2012) Heterozygosity and diversity analysis using mapped SNPs in a faba bean inbreeding programme. Mol Breed 30:1799–1809

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:e3395

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM et al (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

de los Campos G, Gianola D, GJM R (2009a) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci 87:1883–1887

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A et al (2009b) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385

de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic breeding values across multiple populations. Genetics 183:1545–1553. doi:10.1534/genetics.109.104935

Deokar AA, Ramsay L, Sharpe AG, Diapari M, Sindhu A et al (2014) Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. BMC Genomics 15:708

Deulvot C, Charrel H, Marty A, Jacquin F, Donnadieu C et al (2010) Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. BMC Genomics 11:468

Drucker H, Burges CJC, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. Adv Neural Info Process Syst 9:155–161

Duarte J, Rivière N, Baranger A et al (2014) Transcriptome sequencing for high throughput SNP development and genetic mapping in pea. BMC Genomics 15:126

Egbadzor KF, Ofori K, Yeboah M, Aboagye LM, Opoku-Agyeman MO et al (2014) Diversity in 113 cowpea [Vigna unguiculata (L) Walp] accessions assessed with 458 SNP markers. Springer Plus 3:541

Fedoruk M (2013) Linkage and association mapping of seed size and shape in lentil. Thesis (Masters of Science), University of Saskatchewan, Saskatoon

Gautami B, Pandey MK, Vadez V, Nigam SN, Ratnakumar P et al (2012) Quantitative trait locus analysis and construction of consensus genetic map for drought tolerance traits based on three recombinant inbred line populations in cultivated groundnut (Arachis hypogaea L.) Mol Breed 30:757–772

Graham PH, Vance CP (2003) Legume crops: importance and constraints to greater use. Plant Physiol 131:872–877

Grattapaglia D, Resende MDV, Resende MR, Sansaloni CP, Petroli CD et al (2011) Genomic selection for growth traits in eucalyptus: accuracy within and across breeding populations. BMC Proc 5:O16. doi:10.1186/1753-6561-5-S7-O16

Gujaria N, Kumar A, Dauthal P, Dubey A, Hiremath P et al (2011) Development and use of genic molecular markers (GMMs) for construction of a transcript map of chickpea (Cicer arietinum L.) Theor Appl Genet 122:1577–1589

Guo Z, Tucker D, Lu J, Kishore V, Gay G (2012) Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theor Appl Genet 124:261–275

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E et al (2014) The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet 127:749–762

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186

Hascoët E, Jaminon O, Devaux C, Blassiau C, Bahrman N, Bochard A-M et al (2014) Towards fine mapping of frost tolerance QTLs in pea, in 2nd PeaMUST Annual Meeting (Dijon)

Hayes B, Bowman P, Chamberlain A, Goddard M (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443

Heffner EL, Me S, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

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

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160

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

Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R et al (2014) Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci 54:1476–1488

Hiremath PJ, Farmer A, Cannon SB, Woodward J, Kudapa H et al (2011) Large-scale transcriptome analysis in chickpea (Cicer arietinum L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. Plant Biotechnol J 9:922–931. doi:10.1111/j.1467-7652.2011.00625.x

Hiremath PJ, Kumar A, Penmetsa RV, Farmer A, Schlueter JA et al (2012) Large-scale development of cost-effective SNP marker assays for diversity assessment and genetic mapping in chickpea and comparative mapping in legumes. Plant Biotechnol J 10:716–732

Hospital F (2005) Selection in backcross programmes. Philos Trans Roy Soc Lond B Biol Sci 360:1503–1511

Huynh BL, Close TJ, Roberts PA, Hu Z, Wanamaker S et al (2013) Gene pools and the genetic architecture of domesticated cowpea. The Plant Genome 6:3

Jaganathan D, Thudi M, Kale S, Azam S, Roorkiwal M et al (2015) Genotyping-by-sequencing based intra-specific genetic map refines a "QTL-hotspot" region for drought tolerance in chickpea. Mol Gen Genomics 290:559–571

Jain M, Misra G, Patel RK, Priya P, Jhanwar S et al (2013) A draft genome sequence of the pulse crop chickpea (Cicer arietinum L.) Plant J 74:715–729. doi:10.1111/tpj.12173

Janila P, Pandey MK, Shasidhar Y, Variath MT, Sriswathi M et al (2016) Molecular breeding for introgression of fatty acid desaturase mutant alleles (*ahFAD2A* and *ahFAD2B*) enhances oil quality in high and low oil containing peanut genotypes. Plant Sci 242:203–213

Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J et al (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics 15:740

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics 8:118–127. doi:10.1093/biostatistics/kxj037

Kang YJ, Kim SK, Kim MY, Lestari P, Kim KH et al (2014) Genome sequence of mungbean and insights into evolution within Vigna species. Nat Commun 5:5443

Kaur S, Cogan NO, Stephens A, Noy D, Butsch M et al (2014a) EST-SNP discovery and dense genetic mapping in lentil (Lens culinaris Medik.) enable candidate gene selection for boron tolerance. Theor Appl Genet 127:703–713

Kaur S, Kimber RBE, Cogan NOI, Materne M, Forster JW et al (2014b) SNP discovery and high-density genetic mapping in faba bean (Vicia faba L.) permits identification of QTLs for ascochyta blight resistance. Plant Sci 217–218:47–55

Khera P, Upadhyaya HD, Pandey MK, Roorkiwal M, Sriswathi M et al (2013) SNP-based genetic diversity in the reference set of peanut (Arachis spp.) by developing and applying cost-effective KASPar genotyping assays. Plant Genome 6:1–11

Kim KH, Kim MY, Van K, Moon JK, Kim DH et al (2008) Marker-assisted foreground and background selection of near isogenic lines for bacterial leaf pustule resistant gene in soybean. J Crop Sci Biotechnol 11:263–268

Kumar S, Banks TW, Cloutier S (2012) SNP discovery through next-generation sequencing and its applications. Int J Plant Genomics 15

Kumar V, Rani A, Rawal R, Mourya V (2015) Marker assisted accelerated introgression of null allele of kunitz trypsin inhibitor in soybean. Breed Sci 65:447–452

Lavaud C, Lesne A, Piriou C, Le Roy G, Boutet G et al (2015) Validation of QTL for resistance to Aphanomyces euteiches in different pea genetic backgrounds using near-isogenic lines. Theor Appl Genet 128:2273–2288

Lee YG, Jeong N, Kim JH, Lee K, Kim KH et al (2015) Development, validation and genetic analysis of a large soybean SNP genotyping array. Plant J 81:625–636

Legarra A, Robert-Granie P, Croiseau G, Guillaume F, Fritz S (2011) Improved LASSO for genomic selection. Genet Res 93:77–87

Li X, Wei Y, Acharya A, Hansen JL, Crawford JL et al (2015) Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. Plant Genome 8

Liu XQ, Rong JY, Liu XY (2008) Best linear unbiased prediction for linear combinations in general mixed linear models. J Multivariate Analysis 99:1503–1517

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161

Lucas MR, Ehlers JD, Huynh BL, Diop NN, Roberts PA et al (2013) Markers for breeding heat-tolerant cowpea. Mol Breed 31:529–536

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol 41:56

Mujibi FD, Nkrumah JD, Durunna ON, Stothard P, Mah J et al (2011) Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. J Anim Sci 89:3353–3361

Muñoz-Amatriaín M, Mirebrahim H, Xu P, Wanamaker SI, Luo M et al (2016) Genome resources for climate-resilient cowpea, an essential crop for food security. Plant J. doi:https://doi.org/10.1101/059261

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

Pandey MK, Rathore A, Das RR, Khera P, Upadhyaya HD et al (2014a) Selection of appropriate genomic selection model in an unstructured germplasm set of peanut (Arachis hypogaea L.). 6th international Food Legumes Research conference & 7th international conference on Legume Genetics and Genomics on 7–11 July 2014, Saskatoon

Pandey MK, Upadhyaya HD, Rathore A, Vadez V, Sheshshayee MS et al (2014b) Genome-wide association studies for 50 agronomic traits in peanut using the 'reference set' comprising 300 genotypes from 48 countries of semi-arid tropics of the world. PLoS One 9:e113326

Pandey MK, Agarwal G, Rathore A, Janila P, Upadhyaya HD, et al. (2015). Development of high density 60K "Axiom_Arachis" SNP Chip and optimization of genomic selection model for enhancing breeding efficiency in peanut. Proceedings of 8th international conference of the Peanut Research Community on "Advances in Arachis through Genomics and Biotechnology", Brisbane, 5–9 Nov 2015

Pandey MK, Roorkiwal M, Singh VK, Ramalingam A, Kudapa H et al (2016) Emerging genomic tools for legume breeding: current status and future prospects. Front Plant Sci 7

Pandey MK, Agarwal G, Kale SM, Clevenger J, Nayak SN et al (2017) Development and evaluation of a high density genotyping 'Axiom_Arachis' array with 58K SNPs for accelerating genetics and breeding in groundnut. Nat Sci Rep 7:40577. doi:10.1038/srep40577

Poland J, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92–102

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S et al (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5:103–113

Power JF (1987) Legume crops: their potential role in agricultural production. Am J Alt Agri 2:69–73

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463

Rebello CJ, Greenway FL, Finley JW (2014) A review of the nutritional value of legumes and their effects on obesity and its related co-morbidities. Obesity Rev 15:392–407

Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA et al (2012a) Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 194:116–128

Resende MFR, Munoz P, Resende MDV, Garrick DJ, Fernando RL et al (2012b) Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda l.) Genetics 190:1503–1510

Roorkiwal M, Rathore A, Das RR, Singh MK, Jain A, Srinivasan S, Gaur PM, Chellapilla B, Tripathi S, Li Y, Hickey JM, Lorenz A, Sutton T, Crossa J, Jannink J-L, Varshney RK (2016) Genome-Enabled prediction models for yield related traits in chickpea. Front Plant Sci 7

Roorkiwal M, Sawargaonkar SL, Chitikineni A, Thudi M, Saxena RK et al (2013) Single nucleotide polymorphism genotyping for breeding and genetics applications in chickpea and pigeonpea using the BeadXpress platform. Plant Genome 6

Roorkiwal M, Von Wettberg EJ, Upadhyaya HD, Warschefsky E, Rathore A et al (2014) Exploring germplasm diversity to understand the domestication process in Cicer spp. using SNP and DArT markers. PLoS One 9(7):e102016

Rubiales D, Mikic A (2015) Introduction: legumes in sustainable agriculture. Crit Rev Plant Sci 34:2–3

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T et al (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res 15:227–239

Saxena RK, Penmetsa RV, Upadhyaya HD, Kumar A, Carrasquilla-Garcia N et al (2012) Large-scale development of cost-effective single-nucleotide polymorphism marker assays for genetic mapping in pigeonpea and comparative mapping in legumes. DNA Res 19:449–461

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Sharpe AG, Ramsay L, Sanderson LA, Fedoruk MJ, Clarke WE et al (2013) Ancient orphan crop joins modern era: gene-based SNP discovery and mapping in lentil. BMC Genomics 14:192

Solberg TR, Sonesson AK, Woolliams JA (2008) Genomic selection using different marker types and densities. J Anim Sci 86(10):2447–2454

Spindel J, Begum H, Akdemir D, Virk P, Collard B et al (2015) Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet 11:e1005350

Tarawali G, Ogunbile OA (1995) Legumes for sustainable food production in semi-arid savannahs. ILEIA Newslett 11(4):18–23

Tayeh N, Aluome C, Falque M, Jacquin F, Klein A et al (2015) Development of two major resources for pea genomics: the GenoPea 13.2 K SNP Array and a high-density, high-resolution consensus genetic map. Plant J 84:1257–1273

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Series B 58:267–288

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK et al (2012) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89

Varshney RK, Mohan SM, Gaur PM, Gangarao NVPR, Pandey MK et al (2013a) Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. Biotechnol Adv 31:1120–1134

Varshney RK, Song C, Saxena RK, Azam S, Yu S et al (2013b) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat Biotechnol 31:240–246. doi:10.1038/nbt.2491

Varshney RK, Gaur PM, Chamarthi SK, Krishnamurthy L, Tripathi S et al (2013c) Fast-track introgression of "*QTL-Hotspot*" for root traits and other drought tolerance traits in JG 11, an elite and leading variety of chickpea. Plant Genome 6:3. doi:10.3835/plantgenome2013.07.0022

Varshney RK, Mohan SM, Gaur PM, Chamarthi SK, Singh VK et al (2014) Marker-assisted backcrossing to introgress resistance to Fusarium wilt (FW) race 1 and Ascochyta blight (AB) in C 214, an elite cultivar of chickpea. Plant Genome. doi:10.3835/plantgenome2013.10.0035

Verma S, Gupta S, Bandhiwal N, Kumar T, Bharadwaj C et al (2015) High-density linkage map construction and mapping of seed trait QTLs in chickpea (Cicer arietinum L.) using genotyping-by-sequencing (GBS). Sci Rep 5:17512

Wang J, Chu S, Zhang H, Zhu Y, Cheng H et al (2016) Development and application of a novel genome-wide SNP array reveals domestication history in soybean. Sci Rep 6

Xu P, Wu XH, Wang BG, Luo J, Liu YH et al (2012) Genome wide linkage disequilibrium in Chinese asparagus bean (Vigna unguiculata ssp. sesquipedialis) germplasm: implications for domestication history and genome wide association studies. Heredity 109:34–40

Yang H, Tao Y, Zheng Z, Zhang Q, Zhou G et al (2013) Draft genome sequence, and a sequence-defined genetic linkage map of the legume crop species Lupinus angustifolius L. PLoS One 8:e64799

Young ND, Debellé F, Oldroyd GE, Geurts R, Cannon SB et al (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature 480:520–524

Zhang Z, Liu J, Ding X, Bijma P, de Koning D-J et al (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5:e12648. doi:10.1371/journal.pone.0012648

Zhang Z, Ober U, Erbe M, Zhang H, Gao N et al (2014) Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS One 9: e93017

Zhang J, Song Q, Cregan PB, Jiang GL (2016) Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). Theor Appl Genet 129:117–130

Zhong S, Dekkers JC, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355–364

Zhu S, Walker DR, Warrington CV, Parrott WA, All JN et al (2007) Registration of four soybean germplasm lines containing defoliating insect resistance QTLs from PI 229358 introgressed into 'Benning'. J Plant Reg 1:162–163

Chapter 7
Genomic Selection in Hybrid Breeding

Albert Wilhelm Schulthess, Yusheng Zhao, and Jochen C. Reif

Abbreviations

BLUP	Best linear unbiased prediction
e-Bayes	Empirical Bayes method
GCA	General combining ability
GS	Genomic selection
LD	Linkage disequilibrium
MAS	Marker assisted selection
PS	Phenotypic selection
RE	Relative efficiency
REML	Restricted maximum likelihood
RKHS	Reproducing kernel Hilbert space
RR-BLUP	Ridge regression best linear unbiased prediction
RRS	Recurrent reciprocal selection
SCA	Specific combining ability
SNP	Single nucleotide polymorphism
W-BLUP	Weighted best linear unbiased prediction

7.1 Introduction

A "hybrid variety" will be understood as the offspring of a controlled cross of two or more different (inbred or not) genotypes (Becker 2011). The ultimate goal of hybrid breeding is the exploitation of the phenomenon known as "heterosis"

A.W. Schulthess • Y. Zhao • J.C. Reif (✉)
Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466, Gatersleben, Germany
e-mail: schulthess@ipk-gatersleben.de; zhao@ipk-gatersleben.de; reif@ipk-gatersleben.de

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7_7

(Whitford et al. 2013), in which the performance of hybrids is superior to the mean of its parents (Bernardo 2010; Falconer and Mackay 1996). However, there are additional reasons why hybrid breeding is preferred to line breeding (Longin et al. 2012):

(i) Hybrids have greater yield stability, which is a major advantage for agriculture in marginal environments.
(ii) Heterozygosity allows the potential combination of dominant major genes in the hybrid genotype.
(iii) Hybrids offer a built-in plant variety protection system by means of inbreeding depression when growing farmed-saved seeds.

In the last years, the plant breeding community started to look at genomic selection (GS) as a promising tool to reduce the costs and to accelerate plant breeding programs (Desta and Ortiz 2014; Jannink et al. 2010; Zhao et al. 2014b). The main objective of this chapter is to explain the basic concepts of hybrid breeding and to integrate them with methods of GS. Thereby, it is expected that these concepts and methods allow the reader to understand the philosophy underlying GS in hybrid breeding.

7.2 Basic Concepts Relevant to Hybrid Breeding

Even though heterozygosity does not necessarily imply the occurrence of dominance, heterozygosity is fully required for its existence (Bernardo 2010). This contrasts with the situation observed for fully inbred genotypes used in line breeding, in which heterozygosity is practically residual (Bos and Caligari 2008) and, as a result, dominance effects are expected to be negligible or absent within this particular system. Therefore, dominance will be considered as a particular feature of hybrid breeding and will receive special attention in the following sections. Furthermore, heterosis is of special interest to hybrid breeding, and thus, basic concepts related to this topic will be presented. Finally, the breeding concepts of combining ability and heterotic groups will be introduced to understand the philosophy of hybrid breeding.

7.2.1 Dominance

We will first consider a single-locus model to explain the concept of dominance (Fig. 7.1). Given two homozygous but contrasting genotypes for locus A, coded as A_1A_1 and A_2A_2, the genotypic values, i.e., the effect of the genotype on the phenotype, of them and their cross A_1A_2 can be denoted as: $\mathrm{MP} - a$, $\mathrm{MP} + a$, and $\mathrm{MP} + d$, respectively, where MP corresponds to the average of the phenotypes of parents with genotypes A_1A_1 and A_2A_2. The value a stands for the additive effect at

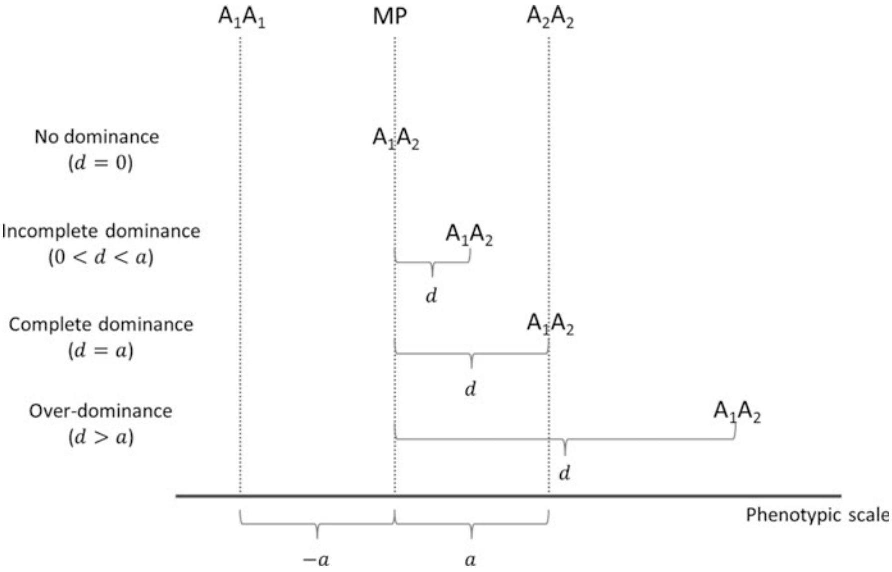


Fig. 7.1 Schematic representation of the dominance effect (d) at locus A according to different levels of dominance: no dominance ($d = 0$), incomplete dominance ($0 < d < a$), complete dominance ($d = a$) and overdominance ($d > a$). A_1A_1 and A_2A_2 represent two homozygous but contrasting parents at locus A, whereas A_1A_2 corresponds to the offspring of the cross between them. MP stands for the mid-parent value between parents A_1A_1 and A_2A_2, and a is the difference between parent A_1A_1 or A_2A_2 and the MP value at the phenotypic scale

locus A and is defined as half the difference between the genotypic values of A_1A_1 and A_2A_2. When dominance is present at locus A, the value d will differ from zero, being $0 < d < a$, $d = a$ or $d > a$, in the cases of incomplete dominance, complete dominance and overdominance, correspondingly (Falconer and Mackay 1996).

Nevertheless, when crosses are performed, the unit of inheritance corresponds to an allele and not to the genotype itself; therefore, an allelic value should be defined. The average effect or value of an allele is defined as the mean of individuals (expressed as a deviation from the population mean) that inherited that particular allele provided that the second allele was inherited at random (Falconer and Mackay 1996). In consequence, the genotypic value (g_{12}) of the cross $A_1A_1 \times A_2A_2$ is denoted as

$$g_{12} = \mu + \alpha_1 + \alpha_2 + \delta_{12}, \qquad (7.1)$$

where μ refers to the mean of population in Hardy-Weinberg equilibrium; α_1 and α_2 are the allele effects of the genes inherited from parents A_1A_1 and A_2A_2, respectively; and δ_{12} represents a residual value, which cannot be explained by the average allelic effects. This residual term δ_{12} will be denoted as the dominance deviation (Falconer and Mackay 1996).

7.2.2 Heterosis

Quantitative geneticists use the term heterosis to make reference to the superiority of a hybrid over the mean of its parents, a term known as mid-parent heterosis. In addition, plant breeders tend to use the term better-parent heterosis, in which the economic advantage of a hybrid is defined as the superior performance over both of its parents (Becker 2011). Furthermore, in crops such as wheat or barley, where line breeding historically played a major role compared with hybrid breeding, plant breeders often use the concept of commercial heterosis, in which the superiority of the hybrids is judged based on comparison(s) with the best available inbred line(s) (Longin et al. 2012). Nevertheless, in this chapter, the term "heterosis" will be understood, if not explicitly stated otherwise, as midparent heterosis.

The genetic causes of heterosis are still a topic of debate in quantitative genetics. However, there are two main hypotheses that have been proposed to explain this phenomenon (Bernardo 2010; Whitford et al. 2013):

(i) Dominance hypothesis:

Hybrids are expected to be superior to their homozygous parents because of the masking of unfavorable recessive alleles in the heterozygous genotype. Then, at a single locus level, the heterozygote would be expected to not exceed the genotypic value of the better parent ($0 < d \leq a$ in Fig. 7.1). Nonetheless, it is very likely that neither of the homozygous parents carry all positive alleles at all loci for a particular polygenic trait, implying that the hybrid could be superior to both parents (Bruce 1910; Collins 1921; Jones 1917; Keeble and Pellew 1910).

(ii) Over-dominance hypothesis:

Hybrids are expected to be superior to their parents because of the inherent superiority of the heterozygous genotype over both homozygotes ($d > a$ in Fig. 7.1). In consequence, a single locus would be enough to explain heterosis (Crow 1948; East 1936; Hull 1945).

Last but not least, even in the absence or with very low degrees of dominance effects, a situation which could be expected for autogamous species, heterosis could be present for a particular trait (Bernardo 2010; Whitford et al. 2013). In this case, epistasis would play an important role in heterosis (Richey 1942; Schnell and Cockerham 1992). It is likely that a mixture of the above-mentioned mechanisms ultimately underlies heterosis (Whitford et al. 2013).

7.2.3 Combining Abilities

Once it is known that heterosis guarantees the successful development of commercial hybrid genotypes for a particular species, plant breeders are in general no more interested in heterosis itself, but rather in the performance of hybrid genotypes.

In this sense, hybrid breeders' efforts and resources will be completely allocated to the development of hybrids with superior performance, independently of whether the superior performance is due to heterosis between the parents or because the parents have a high per se performance (Bernardo 2010).

Hybrid crop studies have shown for complex traits such as grain yield that hybrid performance cannot be predicted with high accuracy using the per se performance of the parents (for review, see Becker 2011 and Hallauer et al. 2010). Therefore, plant breeders normally evaluate and select good parents based on their performance as parents of hybrids, which is often referred to as the combining ability of the parent lines (Hallauer et al. 2010). Given a pool of parent lines to select, the hybrid performances of their crosses are evaluated. Then, the mean hybrid performance of a particular cross $F_i \times M_j$ ($\mu_{F_i \times M_j}$) can be expressed using the combining abilities of parent lines as follows (Bernardo 2010; Falconer and Mackay 1996):

$$\mu_{F_i \times M_j} = \mu + \mathrm{GCA}_{F_i} + \mathrm{GCA}_{M_j} + \mathrm{SCA}_{F_i \times M_j}, \tag{7.2}$$

where μ is the mean of all hybrids, whereas GCA and SCA correspond to the general and specific combining abilities of parents, respectively. The GCA_{F_i} and GCA_{M_j} are the mean values expressed as deviations from μ of all F_1 hybrids having F_i or M_j as one of the parents, correspondingly. However, even ignoring nongenetic sources of error, there would be a remaining proportion of variability for the hybrid performance, which could not be explained by the GCA of both parents. This last term, referred to as $\mathrm{SCA}_{F_i \times M_j}$, measures the interaction between parents F_i and M_j which cannot be accounted by the main effects of their GCAs. The combining abilities are assumed as independent from each other. Hence, assuming no error variance, the total variance between hybrids can be decomposed as $\sigma^2_{\mathrm{GCA}_F} + \sigma^2_{\mathrm{GCA}_M} + \sigma^2_{\mathrm{SCA}}$, if parent lines conform groups by sex or factors (denoted as F for female or M for male) or alternatively as $2\sigma^2_{\mathrm{GCA}} + \sigma^2_{\mathrm{SCA}}$, if they do not configure any kind of groups (Falconer and Mackay 1996).

GCA is often interpreted as the influence of additive effects, whereas SCA as an indication of genes having dominance and epistatic effects (Hallauer et al. 2010). Moreover, provided that both parents are completely homozygous inbred lines and assuming the absence of epistatic effects, σ^2_{GCA} and σ^2_{SCA} are equal to the variances due to α (σ^2_{A}) and δ effects (σ^2_{D}), respectively (Wricke and Weber 1986).

7.2.4 Heterotic Groups and Patterns

A heterotic group is defined as a group of genotypes that display similar combining ability and heterotic response when crossed with genotypes from other genetically distinct germplasm groups. In addition, a pair of specific heterotic groups with a high hybrid performance in their cross will conform a heterotic pattern (Melchinger and Gumber 1998). It has been suggested that using genetically divergent populations is relevant for the establishment of heterotic groups and patterns. Genetically diverse heterotic groups do not only allow the maximum exploitation of heterosis and hybrid performance but also lead to a lower $\sigma_{\mathrm{D}}^{2}/\sigma_{\mathrm{A}}^{2}$ ratio or, equivalently, to a lower $\sigma_{\mathrm{SCA}}^{2}/\sigma_{\mathrm{GCA}}^{2}$ proportion. The latter implies that hybrid performance could be predicted using Eq. (7.2) only relying on GCAs of parents (Reif et al. 2007). Furthermore, because combining ability can be exploited in a recurrent fashion, breeding efforts within a hybrid breeding program can then be allocated to the selection of the best parents for each heterotic group within a specific and previously identified heterotic pattern mainly based on their GCA (Hallauer et al. 2010). This breeding method is known as recurrent reciprocal selection (RRS). Briefly, provided that at least two heterotic groups or pools are available, genotypes to be tested of one pool are testcrossed with a small number of random sampled genotypes that belong to the opposite pool. Therefore, a small number of crosses for each tested genotype are generated. Later, all seeds pertaining to crosses of a particular tested parent are harvested and bulked together as a single progeny. Thus, each parent will be represented by its corresponding progeny in field trials during the next season. Then, the best parent lines of each pool are recognized and selected based on the performance of each progeny. Subsequently, selected parents are inter-mated within each pool and serve as base material for their respective heterotic group during the next breeding cycle. This whole process is repeated in parallel for each considered pool (Comstock et al. 1949). The RRS method has been widely and successfully used for hybrid breeding in crops such as maize, where heterotic groups and patterns were empirically developed by observing which crosses produced superior hybrid performance and which do not (Tracy and Chandler 2006).

7.3 GS for Hybrid Genotypes

7.3.1 Cross-Validated Prediction Accuracy of GS

Studies on GS often mix the concepts of prediction ability (or predictability) and prediction accuracy. Nevertheless, prediction ability is expressed as the correlation between genomic predictions and observed phenotypes, whereas prediction accuracy is generally defined as the prediction ability divided by h (the square root of the heritability h^2) for the trait being predicted (Lorenzana and Bernardo 2009; Riedelsheimer et al. 2012; Zhao et al. 2012b). In this sense, the prediction accuracy value is interpreted as prediction ability for a trait with heritability equal to 1, and, in consequence, prediction accuracies are expected to be higher than prediction ability values. Moreover, because prediction accuracies provide an estimate of the genotypic correlation, they are more relevant for the estimation of the effects of indirect selection by means of GS (Zhao et al. 2013b). However and for simplicity, both terms will be indistinctly used in the present chapter, unless the contrary is stated.

Predictabilities of GS are overestimated when the same genotypes used for the estimation of marker effects are considered for prediction (Krchov et al. 2015; Xu et al. 2014), even if the observed phenotypes used for the computation of predictabilities are different from those considered for marker effects estimation (Krchov et al. 2015). Such approaches to compute predictability do not properly mimic the situation faced in practice by plant breeders: genomic models will be trained with an estimation set of genotypes for which phenotypic and genomic data are available and predictions will be obtained for a group of genotyped, but not phenotyped, selection candidates, which are in principle independent from the estimation set. Therefore, validation of the predictability is crucial to show the actual potential of GS in plant breeding and can be efficiently achieved by means of cross-validations (Hjorth 1994). In k-fold cross-validation, for instance, the population with available genomic and phenotypic data is divided in k subgroups of similar size. Then, the first $k-1$ subgroups are used to predict the effects of markers, and the genotypes included in the k^{th} subgroup are predicted and compared with their observed values. This process can be iteratively repeated to obtain robust estimates for the cross-validated prediction accuracy of GS.

7.3.1.1 Relatedness Plays a Major Role in Determining the Cross-Validated Prediction Accuracy of GS in Hybrid Breeding

From simulation and experimental plant data studies, it is well known that relatedness between estimation and validation sets influences the prediction ability of GS (Gowda et al. 2014; Habier et al. 2007; Meuwissen 2009; Meuwissen et al. 2001; Mirdita et al. 2015; Technow et al. 2012, 2014; Zhao et al. 2013b, 2015). In this sense, the more related these two sets are, the higher would be the predictability. In addition, relatedness will be present at different levels of a plant breeding program, and, in consequence, this should be taken into account at the moment of performing GS and interpreting predictability levels. The current section aims to illustrate this point in detail.

Relatedness and Its Implications on the Predictabilities of GS Within Hybrid Breeding Programs

Different levels of relatedness can be found within the plant material used in a hybrid breeding program. This is schematically represented in Fig. 7.2 by considering factorial single crosses between biparental populations of pools A and B. First, elite genotypes A_1 to A_4 (belonging to pool A) and B_1 to B_4 (coming from pool B) will be crossed within each pool for the generation of segregating inbred populations. In Fig. 7.2, the inbred progenies are represented by genotypes a* to t* and a to t for pools A and B, correspondingly. Nowadays, inbreds can be obtained in a singlegeneration by means of doubled haploid techniques (Becker 2011). Thus, eight different populations of full-sib lines are generated, with four

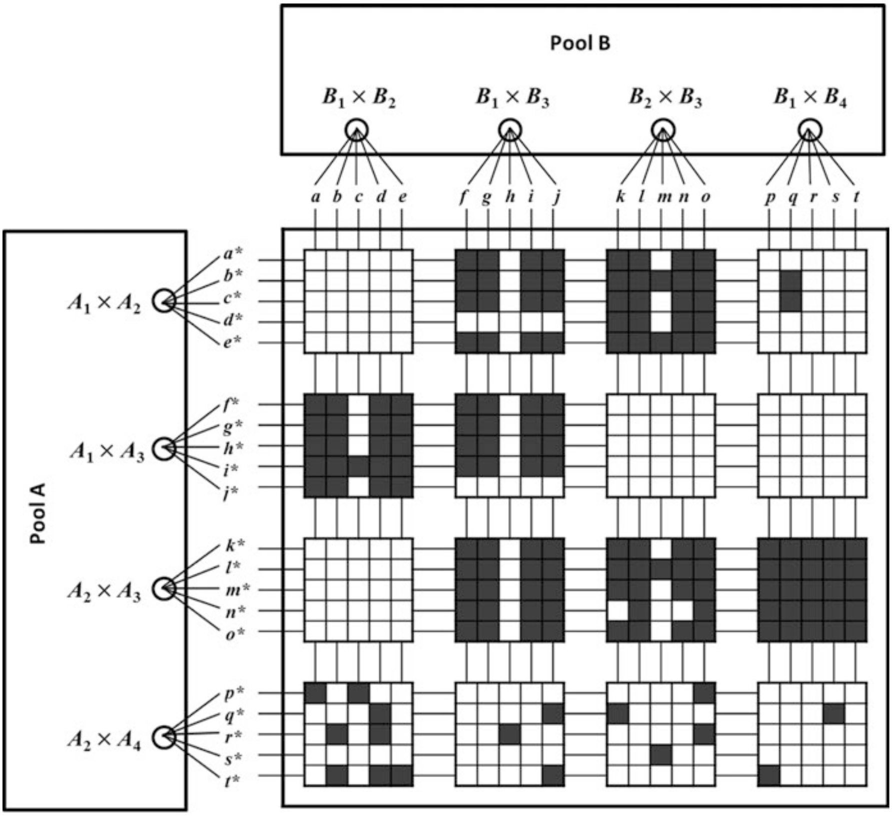


Fig. 7.2 Illustration of a hybrid breeding program using a factorial cross-design between biparental populations of two pools (namely pools A and B). Genotypes A_1 to A_4 and B_1 to B_4 represent different elite lines belonging to pools A and B, respectively. Circles correspond to the different families generated by crossing elite lines within each pool (denoted by the × symbol), whereas genotypes a^* to t^* and a to t are the corresponding progenies from these crosses. Progenies connected to the same node (family) are assumed as full-siblings. The squares in the center of the figure represent the 400 possible hybrids obtained between inbreds a^* to t^* and a to t of pools A and B, correspondingly. The empty white squares denote hybrids with available phenotypic data, whereas the shaded ones indicate missing hybrids

populations for each pool. Subsequently, the generated lines are crossed in a factorial way with genotypes of the opposite pool to evaluate their performance as parents. For simplicity, only the two extreme levels of relatedness will be considered as examples. At one extreme, the most related individuals correspond to hybrids derived from crossing genotypes of a biparental population with the same genotype of the opposite pool. For instance, this situation is well represented by hybrids $d^* \times b$ and $d^* \times c$ in Fig. 7.2. At the other extreme, there are less related hybrids like $h^* \times n$ and $q^* \times s$, in which the genotypes being crossed do not share any of the elite lines involved in the generation of biparental populations. Nevertheless, according to Sect. 7.2.4, in RRS, the best parent lines recognized within each pool will be subsequently used as founder material of the upcoming segregating inbred populations during the next breeding cycle. Presumably, it is possible that in some occasions, two or more of these good parent lines belong to the same biparental cross (for example, genotypes *p* and *q* in Fig. 7.2), which implies that the elite pool available to serve as parent lines of the different segregating populations would not always include completely unrelated individuals (Bernardo 1994). Moreover, it is also anticipated in Fig. 7.2 that the phenotypes of some hybrids will be missed (represented as shaded squares). This could be because of some evaluation plots that were missed during the crop season, an insufficient number of seeds available for field testing, a limited budget for the plant breeding program that ultimately limited the number of hybrids tested, among other reasons. Furthermore, this unbalanced scenario provides a very good opportunity to perform GS for the individuals without phenotypic records. For example, implementing GS to perform within-population prediction could allow plant breeders to partially testcross a particular segregating population and then to predict the untested individuals with a model whose marker effects were estimated using the tested population fraction (Krchov and Bernardo 2015; Windhausen et al. 2012). This particular breeding scenario can be found in an illustrative manner in Fig. 7.2 by crossing individuals from the $A_1 \times A_2$ population with the tester *m* of the opposite pool, being three of the five possible hybrids available for marker effects estimation. Later, performances of the two remaining hybrids, i.e., $b^{*}\times m$ and $e^{*}\times m$, could be predicted by GS. Accordingly, within-population prediction schemes have been applied in GS studies to obtain cross-validated prediction accuracies for testcross performance of biparental populations in crops like rye (Wang et al. 2014), sugar beet (Würschum et al. 2013) and maize (Albrecht et al. 2011; Krchov and Bernardo 2015; Lorenzana and Bernardo 2009; Zhao et al. 2012a). Prediction within families is a closed system and corresponds to the most favorable scenario for GS, which results in the maximum attainable GS predictabilities (Crossa et al. 2013). This is mainly because of the combination of high levels of relatedness between estimation and prediction sets plus the long-range haplotype blocks within families leading to high linkage disequilibrium (LD) between markers and the loci with true effects on traits (Albrecht et al. 2011) in addition to the absence of population structure expected for this situation (Crossa et al. 2013). However, this approach has two main limitations:

(i) If the estimated marker effects are used for prediction of selection candidates that are less related to the estimation population, there is a potential risk of drop in predictability (Albrecht et al. 2014; Habier et al. 2007; Meuwissen 2009; Meuwissen et al. 2001; Wang et al. 2014).

(ii) Predictability levels in within-population prediction could be constrained by a limited number of genotypes used for marker effect estimation (Albrecht et al. 2014; Lehermeier et al. 2015; Meuwissen 2009). In this sense, increasing the size of the estimation set is expected to create more recombination events that allow an increased resolution for marker effects estimation, ultimately leading to a model with superior predictability for progenies, which are several generations away from the estimation set (Lorenz 2013).

Therefore, it has been proposed that a way to obtain robust marker effect estimates would be to combine the data from different biparental populations in a single and comprehensive estimation set followed by within-population prediction (Albrecht et al. 2014; Lehermeier et al. 2015; Wang et al. 2014; Zhao et al. 2012a). For instance, all the phenotypic data available in Fig. 7.2 could be used to estimate marker effects, and the remaining missing hybrids would be predicted by GS. Nevertheless, in some occasions, it seems that marker × population interactions (when marker effects are not the same in all populations) could negatively impact the predictabilities obtained by approaches like this one (Albrecht et al. 2011, 2014; Lehermeier et al. 2015; Zhao et al. 2012a). In general, neither of the following GS approaches could improve predictabilities compared with a model, in which marker effects are simply estimated across testcross populations: including a general population effect, excluding markers with significant marker × population interaction (Zhao et al. 2012a), or modeling population-specific marker effects considering a variance–covariance structure between populations (Lehermeier et al. 2015). Presumably, the different levels of relatedness expected in breeding programs allow keeping acceptable within-population predictability levels when marker effects are assumed constant across populations, and all these populations constitute a big combined estimation set. In consequence, this last GS approach could be a good choice for robust marker effects estimation due to its simplicity.

Cross-Validation Methods Considering Different Levels of Relatedness in Factorial Crosses

In complete factorial mating designs, $a \times b$ combinations are possible, with a and b being the number of lines belonging to pools A and B, respectively (Fig. 7.3, based on schemes from Schrag et al. 2009). A basic scheme to perform cross-validation in factorial crosses is the leave-one-out method (Fig. 7.3a). In this cross-validation scheme, $a \times b - 1$ hybrid genotypes are used as the estimation set to predict the remaining $(a \times b)^{\text{th}}$ genotype. Then, after obtaining genomic predictions for all the $a \times b$ hybrids in an iterative manner, these values are compared with the observed ones (Jacobson et al. 2014). The concept behind this method is the prediction of a small number of unintentionally missing hybrids that in practice failed (Schrag et al. 2009). Predictions of testcross performance in maize (Jacobson et al. 2014) and of sunflower hybrid performance (Reif et al. 2013) correspond to some examples in which this cross-validation method has been applied. In practice, however, a large number of early candidate lines of each pool will be only tested as parents with the best lines of the opposite group because the evaluation of an extremely large number of hybrids from the $a \times b$ combination becomes unfeasible (Schrag et al. 2009). This situation is better represented by the L-shaped cross-validation scheme (Fig. 7.3b), which has been used as the T2 validation sets in simulated and experimental data for testcrosses in maize (Technow et al. 2012, 2014) and experimental data for factorial crosses of diversity panels in wheat (Gowda et al. 2014; Mirdita et al. 2015; Zhao et al. 2015).

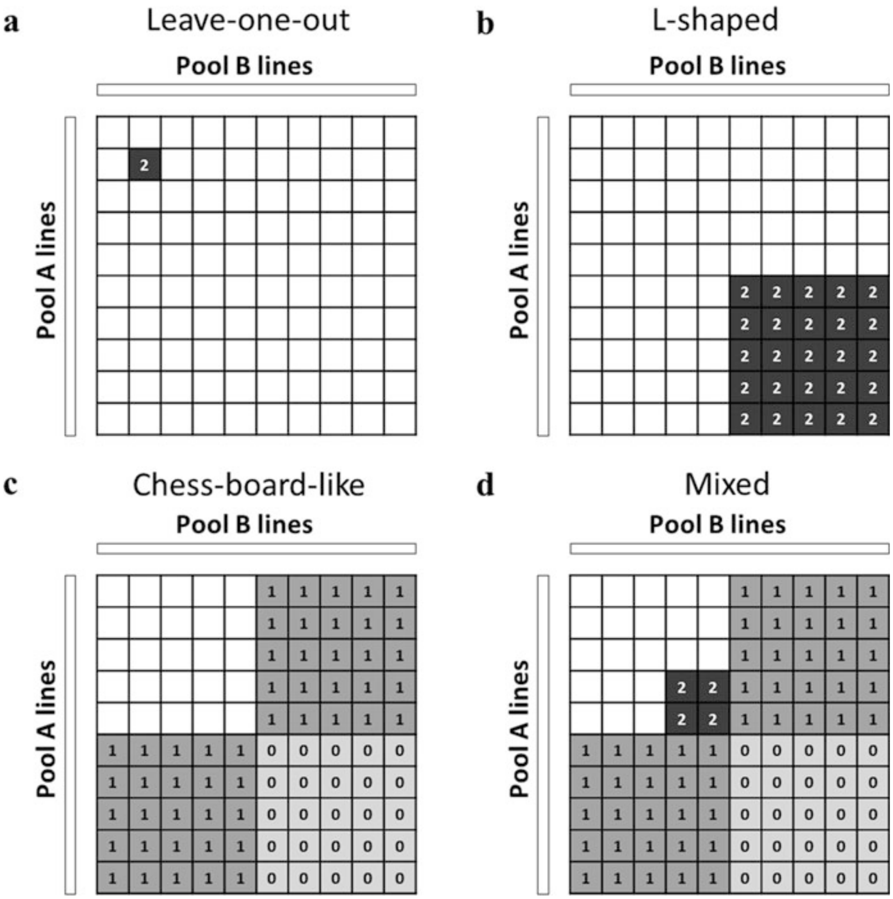


Fig. 7.3 Methods of cross-validation for genomic selection (GS) in factorial crosses: (**a**) Leave-one-out, (**b**) L-shaped, (**c**) Chess-board-like and (**d**) Mixed scheme. In all cases, two pools of parent lines (namely pool A and B) were considered. Each square corresponds to a different hybrid between pools A and B. The empty white squares represent hybrids with available phenotypic data (estimation set), whereas the shaded ones correspond to different hybrids being predicted by GS (validation sets). The degree of shading denotes the level of expected relatedness between estimation and validation sets (the darker the shading, the higher the relatedness), according to the number of parent lines shared between the hybrids of the estimation and validation sets. Numbers 0, 1 and 2 indicate none, one or two parents in common between estimation and validation sets, respectively (based on schemes presented by Schrag et al. 2009)

In the leave-one-out and L-shaped schemes, both parents of the predicted hybrids in the validation set were already evaluated as parents for other hybrids in the estimation set (hybrids denoted with the number 2 in Fig. 7.3a, b). Nevertheless, although estimation and validation sets are related through common parents in these situations, this does not mean that these cross-validation schemes imply overoptimistic outcomes and should be avoided. This is mainly because both methods were designed for scenarios in which predictions rely mainly on relatedness and plant breeders want to profit from it. In contrast, when the objective is the introduction of new parent lines into the breeding program, the level of relatedness between the estimation set and the predicted selection candidates is expected to decrease. In this situation, only one or none of the parents (hybrids represented by the numbers 1 and 0 in Fig. 7.3c, correspondingly) will be shared between the estimation set and the predicted hybrids (Schrag et al. 2009). Therefore, LD should play a more important role than relatedness for the predictability of GS at this point (Habier et al. 2007). The chess-board-like scheme mimics this scenario (Fig. 7.3c) and corresponds to the T1 and T0 validation sets used in simulated and experimental data for factorial crosses in maize (Technow et al. 2012, 2014) and experimental data for factorial crosses in wheat (Gowda et al. 2014; Mirdita et al. 2015; Zhao et al. 2015). In addition, a cross-validation scheme with no shared parents between estimation and validation sets has been used for GS in wheat (Miedaner et al. 2013; Zhao et al. 2013a, b, 2014a). Last but not least, in reality, plant breeders would use the same estimation set for all the above-mentioned prediction scenarios; in consequence, a mixed cross-validation scheme (Fig. 7.3d) will be expected in hybrid breeding programs (Gowda et al. 2014; Mirdita et al. 2015; Technow et al. 2012, 2014; Zhao et al. 2015). Interestingly, prediction accuracy levels for hybrid grain yield performance of the T2 validation set were similar in maize (Technow et al. 2014) and wheat (Zhao et al. 2015), but even though prediction accuracies decreased when shifting from T2 to T0 validation sets in both species, this decay in prediction accuracy was much more pronounced in wheat than in maize. One possible explanation for this phenomenon is that Zhao et al. (2015) based their conclusion on a data set concerning factorial crosses of a diversity panel in wheat, whereas Technow et al. (2014) relied on factorial crosses of maize lines belonging to a RRS program. As it was already mentioned in section "Relatedness and Its Implications on the Predictabilities of GS Within Hybrid Breeding Programs", lines belonging to a particular pool can be closely related in a RRS program, implying that there could be some degree of residual relatedness between the estimation and the T0 validation sets in this particular breeding scheme.

7.3.1.2 Discrepancies Between Test and Target Environments Are Expected to Impact the Cross-Validated Prediction Accuracy of GS

Genotype × environment interaction is expected to negatively impact predictabilities of GS when the target environments for the selection candidates differ from the environments considered to test the genotypes used in the estimation set (Albrecht et al. 2014; Krchov et al. 2015; Schulz-Streeck et al. 2013; Wang et al. 2014; Windhausen et al. 2012). Furthermore, even though the target locations could be exactly the same between the estimation set and the selection candidates, the genotype × year interaction can still have a potential negative impact on predictabilities (Krchov et al. 2015; Wang et al. 2014). Moreover, different agronomical practices can also derive in different environments. For instance, phenotypic data coming from field trials with high levels of nitrogen fertilization could be used as estimation set for genomic predictions of selection candidates targeted to marginal environments agriculture. In consequence, across-environment cross-validation schemes have been applied to mimic this situation in studies on GS for testcross performance in maize (Albrecht et al. 2014; Krchov et al. 2015; Schulz-Streeck et al. 2013; Windhausen et al. 2012; Zhang et al. 2015) and rye (Wang et al. 2014). Nevertheless, although these approaches are expected to give more realistic predictability estimates than cross-validating with the same group of environments used for the estimation set, most studies on GS for hybrid crops have ignored this issue (Krchov et al. 2015). Hopefully, future studies would consider across-environment cross-validation approaches more often, leading to potentially lower but more realistic predictability values for GS.

7.3.2 Accommodating Dominance Effects Within the GS Model

Because dominance is a particular feature of hybrid genotypes, the accommodation of dominance effects within the GS models will receive special attention in this section.

7.3.2.1 Model Based on Marker Effects

A general model for GS including the dominance component is defined as follows (Zhao et al. 2013b):

$$\mathbf{Y} = 1_{\mathbf{n}}\mu + \mathbf{Z_A a} + \mathbf{Z_D d} + \mathbf{e}, \tag{7.3}$$

where $\mathbf{Y}$ is the n-length vector of phenotypic values pertaining to a particular trait, $1_{\mathbf{n}}$ corresponds to a n-length vector of ones, μ stands for the general mean, whereas $\mathbf{Z_A}$ and $\mathbf{Z_D}$ are $n \times m$ design matrices for additive and dominance effects of m bi-allelic markers, respectively. The elements of $\mathbf{Z_A}$ are coded as 0, 1, 2 according to the homozygous (first allele), heterozygous and homozygous (second allele) states, whereas the elements of $\mathbf{Z_D}$ are 0, 1 for the homozygous and heterozygous states at the i^{th} locus, correspondingly. The m-length vectors $\mathbf{a} = (a_1, a_2, \ldots a_m)^T$ and $\mathbf{d} = (d_1, d_2, \ldots d_m)^T$ contain the elements a_i and d_i, which denote the additive and dominance effects for the i^{th} marker, respectively, whereas $\mathbf{e} = (e_1, e_2, \ldots e_n)^T$ is a vector of length n and e_j is the residual for the j^{th} genotype (Zhao et al. 2013b). Equation (7.3) could be modified and alternatively expressed in terms of combining abilities (Piepho 2009).

Ridge Regression Best Linear Unbiased Prediction

An efficient way to obtain the estimate of μ ($\widehat{\mu}$) along with the predictions for **a** ($\hat{a}$) and **d** $(\hat{d})$ of Eq. (7.3) is by means of Ridge Regression-Best Linear Unbiased Prediction (RR-BLUP) (Whittaker et al. 2000). In this method, it is assumed that a_i and d_i marker effects follow the normal distributions $N(0, \sigma_a^2)$ and $N(0, \sigma_d^2)$, correspondingly, being σ_a^2 and σ_d^2 the constant variances of additive and dominance effects, respectively (Zhao et al. 2013b). A normal distribution is assumed for the residuals $e_j \sim N(0, \sigma_e^2)$, where σ_e^2 is the residual variance of Eq. (7.3). Then, the solution of the mixed-model equations (Henderson 1984), allowing the obtainment of $\widehat{\mu}$, $\hat{a}$, and $\hat{d}$, corresponds to:

$$\begin{bmatrix} \widehat{\mu} \\ \hat{a} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} n & 1_n^T \mathbf{Z_A} & 1_n^T \mathbf{Z_D} \\ \mathbf{Z_A^T} 1_n & \mathbf{Z_A^T Z_A} + \lambda_A \mathbf{I_m} & \mathbf{Z_A^T Z_D} \\ \mathbf{Z_D^T} 1_n & \mathbf{Z_D^T Z_A} & \mathbf{Z_D^T Z_D} + \lambda_D \mathbf{I_m} \end{bmatrix}^{-1} \begin{bmatrix} 1_n^T \mathbf{Y} \\ \mathbf{Z_A^T Y} \\ \mathbf{Z_D^T Y} \end{bmatrix}, \qquad (7.4)$$

where $\mathbf{I_m}$ stands for an identity matrix of size m and the shrinkage parameters λ_A along with λ_D are accordingly defined as the ratios $\lambda_A = \sigma_e^2/\sigma_a^2$ and $\lambda_D = \sigma_e^2/\sigma_d^2$ (Meuwissen et al. 2001; Zhao et al. 2013b). The λ terms (λ_A and λ_D) prevent over fitting the model and thus allow the estimation of effects for all markers (Piepho 2009).

Bayesian Approaches

Briefly, Bayesian approaches provide a description of how existing knowledge is modified by experience. The central concept within *Bayesian learning* is to combine what is already known about the statistical ensemble before the data are observed—such knowledge is represented in terms of prior probability distributions—with the information coming from the data. As a result, a posterior distribution is obtained, from which inferences are made using standard probability calculus techniques, and the outcomes are interpreted probabilistically (Sorensen and Gianola 2002). In GS, Bayesian statistics are mainly used to relax some of the assumptions used within the genomic prediction models (Jannink et al. 2010). In the pioneering study of Meuwissen et al. (2001), two Bayesian methods were introduced, namely BayesA and BayesB. The linear model at the level of the data is equal to that used for the RR-BLUP approach in Eq. (7.3), excepting for the assumptions made for σ_a^2 and σ_d^2 (Meuwissen et al. 2001; Zhao et al. 2013b).

BayesA

In RR-BLUP, σ_a^2 and σ_d^2 are assumed as common variances for all loci effects, being this assumption not necessarily realistic for all genetic architectures. BayesA

(Meuwissen et al. 2001; Zhao et al. 2013b) gives solution to this problem because each of the i loci has its own additive ($\sigma^2_{a_i}$) and dominance ($\sigma^2_{d_i}$) variances. Then, the prior probability distributions for $\sigma^2_{g_i}$ in BayesA (denoting g the **a** or **d** effects in Eq. (7.3), irrespectively) correspond to a scaled inverted chi-square distribution in the way $\chi^{-2}\left(v_g, S^2_g\right)$, where v_g and S^2_g are the degrees of freedom and the scale parameter associated to g, respectively. However, these prior probability distributions will lead to posteriors which cannot be directly used for estimation, because they would be conditioned to the unknown value of g (Meuwissen et al. 2001). In consequence, samples are obtained from full-conditional posterior distributions using methods like Gibbs sampling. The full-conditional posterior for g corresponds to a normal distribution $N\left(Z^T_{G_i}(Z_{G_i}g_i + e)/\tilde{\theta}_i, \sigma^2_e/\tilde{\theta}_i\right)$, being $\tilde{\theta}_i = Z^T_{G_i}Z_{G_i} + \sigma^2_e/\sigma^2_g$, whereas Z_{G_i} denotes the i^{th} column of $\mathbf{Z_A}$ or $\mathbf{Z_D}$ from Eq. (7.3), irrespectively. For $\sigma^2_{g_i}$, the full-conditional posterior is a scaled inverted chi-square distribution denoted as $\left(g^2_i + v_g S^2_g\right)\chi^{-2}_{v_g+1}$ (Zhao et al. 2013b).

BayesB

There could be many loci that do not contribute to the variation on traits with less-complex genetic architectures (loci with $\sigma^2_{g_i} = 0$); however, this is not considered by BayesA (Meuwissen et al. 2001). RR-BLUP and BayesA always fit all markers in the GS model, even if they truly have zero effects on the trait under study. Although these markers without true effects are expected to have small predicted effects, they would add noise to the genomic predictions (Habier et al. 2011). In contrast, BayesB considers a proportion π of markers whose $\sigma^2_{g_i} = 0$ and $(1 - \pi)$ with $\sigma^2_{g_i} > 0$. Nonetheless, this new consideration makes the usage of Gibbs sampling unfeasible; hence, the Metropolis-Hastings algorithm has been recommended for sampling (for a detailed explanation refer to Meuwissen et al. 2001). In addition, in Sect. 7.3.1.1, it was already mentioned that relatedness between estimation and validation set influences the prediction ability of GS, and, interestingly, a simulation study showed that BayesB is less impacted by the genetic relatedness among individuals than RR-BLUP because the former model uses better the information due to LD (Habier et al. 2007).

BayesCπ

BayesA and BayesB treat π as known, with $\pi = 0$ for BayesA and an arbitrary π value within the range 0 and 1 for BayesB (Habier et al. 2011; Meuwissen et al. 2001), which is in contradiction with the concept of *Bayesian learning* (Habier et al. 2011; Sorensen and Gianola 2002). To give solution to these drawbacks, Habier et al. (2011) proposed a new Bayesian approach called BayesCπ, which treats a common σ^2_g and π as unknown with a scaled inverted chi-square and a uniform (0,1) distribution as prior probability distributions, respectively. An extension for

BayesCπ accommodating dominance effects is described in detail by Zhao et al. (2013b). Based on results using simulated and animal breeding data, Habier et al. (2011) recommended BayesCπ for routine applications because of its relatively short computational time among other advantages.

7.3.2.2 Model Based on Genotypes and Relationship Matrices

Provided that normally the number of loci (or molecular markers) surpasses the number of genotypes, models based on marker effects are expected to be less computationally efficient than a model based on genotypic effects (Hayes et al. 2009). By far, the most applied genotype-based method of GS is the one proposed by VanRaden (2007, 2008). In the literature, this method is often referenced as GBLUP (Guo et al. 2014; Habier et al. 2011; Su et al. 2012) because of its similarities with the original best linear unbiased prediction (BLUP) method for breeding value prediction using pedigree records information presented by Henderson (1984). Considering only additive effects and that the sum $\mathbf{Z}_{\mathrm{A}}\mathbf{a}$ from Eq. (7.3) corresponds to the breeding value of individuals (VanRaden 2008), Hayes et al. (2009) demonstrated that GBLUP is equivalent to the mixed models methods based on markers effects (like RR-BLUP) when the matrix of relationships among genotypes is calculated from marker profiles. Consequently, there is no reduction in the prediction accuracy of breeding values by shifting to GBLUP. The extension of GBLUP for the inclusion of dominance effects is originally defined in the classical work of Henderson (1985), and it has been lately implemented using marker-estimated relationship matrices (Da et al. 2014; Su et al. 2012). Hereafter, we will refer to this method as DGBLUP. In Henderson's nomenclature, the linear model underlying DGBLUP looks like Eq. (7.3), but in this case, $\mathbf{Z}_{\mathrm{A}}$ and $\mathbf{Z}_{\mathrm{D}}$ are the design matrices for the **a** and **d** vectors of additive and dominance genetic effects of n genotypes, correspondingly. In addition, **a** and **d** vectors follow normal distributions $N\left(0, \mathbf{A}*\sigma_{\mathrm{A}}^2\right)$ and $N(0, \mathbf{D}*\sigma_{\mathrm{D}}^2)$, respectively, where σ_{A}^2 and σ_{D}^2 are now the total additive and dominance variances, correspondingly. Regarding **A** and **D**, they now correspond to the additive (VanRaden 2007, 2008) and dominance (Da et al. 2014; Su et al. 2012) marker-estimated relationship matrices, respectively. Then, the mixed models equations (Henderson 1985) for the DGBLUP are

$$\begin{bmatrix}\widehat{\mu}\\ \hat{a}\\ \hat{d}\end{bmatrix} = \begin{bmatrix} n & \mathbf{1}_{\boldsymbol{n}}^{\boldsymbol{T}}\mathbf{Z}_{\mathrm{A}} & \mathbf{1}_{\boldsymbol{n}}^{\boldsymbol{T}}\mathbf{Z}_{\mathrm{D}} \\ \mathbf{Z}_{\mathrm{A}}^{\boldsymbol{T}}\mathbf{1}_{\boldsymbol{n}} & \mathbf{Z}_{\mathrm{A}}^{\boldsymbol{T}}\mathbf{Z}_{\mathrm{A}} + \mathbf{A}^{-1}\sigma_e^2/\sigma_{\mathrm{A}}^2 & \mathbf{Z}_{\mathrm{A}}^{\boldsymbol{T}}\mathbf{Z}_{\mathrm{D}} \\ \mathbf{Z}_{\mathrm{D}}^{\boldsymbol{T}}\mathbf{1}_{\boldsymbol{n}} & \mathbf{Z}_{\mathrm{D}}^{\boldsymbol{T}}\mathbf{Z}_{\mathrm{A}} & \mathbf{Z}_{\mathrm{D}}^{\boldsymbol{T}}\mathbf{Z}_{\mathrm{D}} + \mathbf{D}^{-1}\sigma_e^2/\sigma_{\mathrm{D}}^2 \end{bmatrix}^{-1} \begin{bmatrix}\mathbf{1}_{\boldsymbol{n}}^{\boldsymbol{T}}\mathbf{Y}\\ \mathbf{Z}_{\mathrm{A}}^{\boldsymbol{T}}\mathbf{Y}\\ \mathbf{Z}_{\mathrm{D}}^{\boldsymbol{T}}\mathbf{Y}\end{bmatrix}. \tag{7.5}$$

Subsequently, the variance component estimates along with the predictions of genetic effects in Eq. (7.5) are simultaneously computed by the restricted maximum likelihood (REML) algorithm. Moreover, based on relationship matrices, the mixed models for hybrid prediction can be modified to accommodate random terms such as the GCA effects of parents and their corresponding SCAs. For instance, Bernardo (1994, 1996) used a relationship matrix termed **S** for the SCA component, which was expressed as the direct product between the relationship matrices pertaining to the GCAs of two heterotic groups (Stuber and Cockerham 1966). In recent years, the GCA plus SCA model using marker-estimated relationship matrices has been implemented in the context of hybrid genomic prediction (Massman et al. 2013; Piepho 2009).

7.3.2.3 Classical Mixed Models or Bayesian Approaches?

There are no GS methods that are suitable for all genetic architectures and/or breeding schemes. Therefore, the superior performance in terms of predictability of different GS approaches relies always on the context of their applications. Interestingly, most studies on GS for hybrid genotypes have relied on classical mixed model predictions by means of GBLUP (Albrecht et al. 2011, 2014; Guo et al. 2013; Massman et al. 2013; Riedelsheimer et al. 2012; Technow et al. 2014; Zhao et al. 2015) and RR-BLUP (Gowda et al. 2014; Guo et al. 2013; Hofheinz et al. 2012; Jacobson et al. 2014; Lorenzana and Bernardo 2009; Massman et al. 2013; Miedaner et al. 2013; Mirdita et al. 2015; Riedelsheimer et al. 2012; Windhausen et al. 2012; Würschum et al. 2013; Zhao et al. 2012a, b, 2013a, b, 2014a). In contrast, lesser studies have applied Bayesian approaches for hybrid performance prediction (Lorenzana and Bernardo 2009; Miedaner et al. 2013; Mirdita et al. 2015; Technow et al. 2014; Zhao et al. 2013a,b, 2014a, 2015). One reason for these observations could be that the understanding and implementation of classical mixed model methods is much more straightforward than for Bayes approaches, which is also facilitated by the large number of user-friendly REML and BLUP packages available (Guo et al. 2014). However, as it was already mentioned in section "BayesB", Bayesian methods like BayesB, which assign effects equal to zero to a proportion π of markers, are expected to be less impacted by the relatedness between estimation and validation sets than methods conferring nonzero effects to all markers available, like RR-BLUP. This particular issue is not trivial if one takes into consideration that the information from genetic relationships is halved with each additional generation and that LD information is more persistent through time (Habier et al. 2007). Nevertheless, in situations in which pedigree relatedness can be efficiently exploited by plant breeders, RR-BLUP could be valuable for predicting hybrid performance (Zhao et al. 2013b, see Sect. 7.3.1.1 for a detailed explanation). Moreover, in practice, the joint evidence of studies on hybrid performance prediction has been inconclusive about the superior predictability of Bayesian over classical mixed models approaches (Lorenzana and Bernardo 2009; Miedaner et al. 2013; Mirdita et al. 2015; Technow et al. 2014; Zhao et al. 2013a, b, 2014a, 2015). For instance, although it was expected that predictions obtained by means of BayesCπ outperformed RR-BLUP predictions for the medium-complexity trait *Fusarium* head blight resistance in hybrid wheat, both methods performed equally (Mirdita et al. 2015). The joint evidence suggests, in consequence, that both groups of methods are in practice relatively equivalent, and also that GS methods must be ultimately selected based on their implementability and understandability, which makes methods like GBLUP and RR-BLUP the preferred ones.

7.3.2.4 Benefits of Modeling Dominance Effects

Studies based on simulated data have shown that the higher the relative importance of dominance versus additive effects is, the more beneficial (in terms of predictability) the inclusion of dominance over additive effects within the GS models would be (Guo et al. 2013; Nishio and Satoh 2014; Technow et al. 2012; Zhao et al. 2013b). Nevertheless, evaluating the benefits of including, in general, any effect within the GS model using empirical data is challenging because often it is not exactly known which of the following situations is being confronted when no benefits are observed: a) some assumptions of the GS methods that include the particular effect evaluated are disrupted; thus, the methods cannot accurately capture the true effect; b) the GS methods can accurately model the effect under evaluation, but the influence of the true effect is extremely low and c) the GS methods cannot accurately capture the effect evaluated, and the influence of the true effect is also negligible. A study on genomic predictions for grain yield in a population of 1604 wheat hybrids found some predictability improvements by using DGBLUP over GBLUP (Zhao et al. 2015), but these benefits were not observed by means of RR-BLUP and Bayesian approaches in a population compromising 90 wheat hybrids (Zhao et al. 2013b). In addition, genomic predictions considering additive plus dominance effects were not superior to predicting frost tolerance exclusively by additive effects in hybrid wheat, presumably because of the low contribution of dominance compared with additive effects for this trait (Zhao et al. 2013a). However, Guo et al. (2013) observed using experimental data for different traits in an F_1 maize population that, in general, the benefits of including dominance over additive effects were more pronounced when the differences between broad-sense and narrow-sense heritabilities for the traits were higher. They expressed broad-sense heritabilities as the ratio of the additive plus dominance variances estimates to the total phenotypic variance, whereas in the narrow-sense heritability, only the additive variance was considered as the numerator. Therefore, their findings suggest that the more different were these two values, the higher was the importance of the dominance variance and, in consequence, the higher the benefits from including dominance over additive effects. Moreover, in general, accommodating dominance over additive effects has been also beneficial in GS for plant height and heading date in a hybrid population of wheat (Zhao et al. 2014a). In conclusion, joint evidence of simulated and experimental data studies points out that modeling dominance over additive effects is beneficial when dominance effects have an important contribution to the total genetic variation (Bernardo 1994; Gowda et al. 2013; Guo et al. 2013; Nishio and Satoh 2014; Reif et al. 2013; Technow et al. 2012; Zhao et al. 2013a, b, 2014a, 2015).

Estimation of genetic parameters such as variance components could give some insights for the relative contribution of dominance effects to total genetic variance, which also highlights the importance of phenotypic analyses as a decision tool before performing GS.

7.3.3 Beyond the Modeling of Dominance

7.3.3.1 Accommodating Epistasis

Using simulated data, Guo et al. (2013) showed that epistasis can bias the predictions achieved by GS models based solely on additive and dominance main effects. In the past, models including epistasis were presumably avoided because of their high computational burden, especially if a large number of markers was available (Jiang and Reif 2015). Nevertheless, this limitation has also encouraged scientists to search for more efficient GS methods accommodating epistasis. In principle, GBLUP can be extended for the inclusion of any order of epistatic interactions by approximating the epistatic genomic relationship matrix of the interaction effects with the Hadamard product operation (denoted as #) between the relationship matrices of main effects (Henderson 1984; Jiang and Reif 2015; Su et al. 2012). For instance, additive × additive, additive × dominance and additive × additive × dominance interactions are represented as $\mathbf{A}\,\#\,\mathbf{A}$, $\mathbf{A}\,\#\,\mathbf{D}$ and $\mathbf{A}\,\#\,\mathbf{A}\,\#\,\mathbf{D}$, respectively. Hereafter, this method will be called EGBLUP (Jiang and Reif 2015). Another approach for GS considering epistatic interactions corresponds to the semi-parametric reproducing kernel Hilbert space (RKHS) regression method (Gianola et al. 2006; Gianola and van Kaam 2008). Recently, it has been demonstrated that both RKHS and EGBLUP considering epistasis are similar approaches and, as a result, reach comparable levels of predictability (Jiang and Reif 2015). Solely taking into account the additive effects and their interactions, the RKHS method is at first sight similar to the additive effects GBLUP (VanRaden 2007, 2008), but a **K** matrix is used instead of the original **A** matrix. The **K** matrix is a $n \times n$ kernel matrix whose entries are functions of marker profiles of pairs of genotypes in the way $\mathbf{K} = (k(x_i, x_j))$, where k ( ) represents a particular function (e.g., the Gaussian kernel function), whereas x_i and x_j are the rows of the marker profile matrix pertaining to genotypes i and j, correspondingly (Jiang and Reif 2015). Furthermore, Bayesian approaches have been also to accommodate epistatic interactions, like the empirical Bayes (e-Bayes) method, in which marker additive main effects and second-order epistatic interactions are calculated based on estimates of true marker variances (Lorenzana and Bernardo 2009). However, only few studies have used GS models taking into account epistatic interactions for hybrid performance prediction. Lorenzana and Bernardo (2009) observed that ignoring epistatic interactions within the GS model leads to higher predictabilities than accommodating epistasis by means of e-Bayes for different traits in maize testcross populations. In addition, two recent crop plant studies considered additive, dominance and their second-level interactions in GS for hybrid performance by means of EGBLUP (Xu et al. 2014; Zhao et al. 2015). Xu et al. (2014) concluded that in general, adding dominance plus epistasis over the main additive component did not help to improve genomic predictions for hybrid performance in rice, whereas Zhao et al. (2015) only observed some benefits from including the dominance over the additive component, but there was no further accuracy improvement from including epistasis in hybrid wheat. Moreover, both authors partly attributed this issue to the fact that the **A** and **D** matrices are correlated with the relationship matrices of epistatic effects; hence, the two former matrices already capture much of the variation for hybrid prediction. Nonetheless, Xu et al. (2014) observed by means of simulated data that for large estimation sets, there is a benefit in prediction by including the epistatic effects within the model, reflecting the need of large population sizes to accurately take advantage of epistasis prediction in GS. In consequence, more studies on hybrid performance prediction are needed to explore the benefits and limitations of GS approaches, which accommodate epistatic effects.

7.3.3.2 Other GS Approaches

W-BLUP Method

Recently, a new GS method, named weighted best linear unbiased prediction (W-BLUP), was designed to properly incorporate the information of previously known functional markers (Zhao et al. 2014a). Alternatively, Bernardo (2014) suggested modeling known functional markers as fixed effects. In the study of Zhao et al. (2014a), it was observed that the predictability values for heading date obtained by means of marker assisted selection (MAS) using the functional marker *Ppd-D1* were higher than by performing GS based on 1280 single nucleotide polymorphism (SNP) markers in a hybrid wheat population. Nevertheless, when both types of information were combined using W-BLUP, predictability values surpassed the ones obtained by MAS or GS alone. In consequence, W-BLUP holds the promise to bridge the gap between MAS and GS when known functional markers are available.

Multiple-Trait GS

Simulation studies have shown that prediction accuracies for a trait with relatively low heritability can be improved when a genetically correlated trait with higher heritability is included within a multiple-trait GS model (Guo et al. 2014; Hayashi and Iwata 2013; Jia and Jannink 2012). However, plant studies exploiting these benefits are scarce, and, to the best of our knowledge, there are only a couple of studies evaluating the advantages and limitations of these methods for hybrid prediction. In a study concerning two testcross populations of rye (Schulthess et al. 2016), grain protein content predictions markedly benefited from the availability of grain yield information in one of the two testcross populations. These benefits were even more pronounced when a few phenotypic records were available for the predicted target trait, but many more phenotypic records were at hand for the indicator trait during the estimation of marker effects. In addition, Lehermeier et al. (2015) performed genomic predictions in ten different testcross populations of maize by means of a multiple-trait GS approach called MG-GBLUP, which considered each testcross population as if it were a different trait. Nevertheless, MG-GBLUP was, in general, less than or equivalently accurate to estimating marker effects by means of a much simpler GS model that assumes the same marker effect for all testcross populations. In the future, more studies are needed to evaluate in detail the routine implementation of multiple-trait GS approaches for hybrid breeding.

Metabolomic Prediction

In the omics era, metabolomics corresponds to the systematic study of metabolite profiles pertaining to a particular process at the organism, tissue or cell level. Metabolomics provides a tool for measuring biochemical activity directly by monitoring the substrates and products transformed during metabolism (Patti et al. 2012). Nowadays, the availability of massive and automated analytical platforms has facilitated the routine generation of this high dimensional data (Patti et al. 2012; Ward et al. 2015). The weak correspondence between the information of genetic and metabolic profiles obtained from leaves of maize (Riedelsheimer et al. 2012) and wheat (Zhao et al. 2015) suggests that both information sources content connected but, at the same time, different biological information (Riedelsheimer et al. 2012). In this sense, it is expected that metabolite profiles condense genetic and environmental influences together (Feher et al. 2014). The basic model of metabolomic prediction is similar to Eq. (7.3) but omitting the dominance term and respecifying the design matrix for **a,** corresponding this last term now to the vector of metabolite effects. Accordingly, the new **Z** matrix for metabolite effects contains the normalized metabolite levels instead of the $-1, 0, 1$ nomenclature originally used for additive effects of bi-allelic markers (Riedelsheimer et al. 2012). However, it should be noted that metabolite profiles from parental lines interact in a very complex way to determine the metabolite profiles of hybrids. For example, a particular metabolite found at low levels in parents A, B and C can be at high and low levels in the hybrids A × B and B × C, respectively. This could be a consequence of variation in a second metabolite at the average parental levels of crosses A × B and B × C, implying that metabolite levels cannot be regarded as independent from each other. In addition, the influence of dominance between parental metabolite profiles can make the situation even more complex (Feher et al. 2014). Riedelsheimer et al. (2012) performed metabolomic prediction for different traits in maize testcrosses by using 130 leaf metabolite profiles and reported prediction accuracy levels that were only slightly lower than those achieved based on highly dense genomic profiles of 38,019 SNP markers.

Nevertheless, the results of this comparison should be carefully interpreted, because accuracies of metabolomic prediction were previously normalized by the repeatability of metabolite profiles, leading to an overestimation of the predictabilities achieved by metabolomic prediction. Lately, Zhao et al. (2015) found that predictions for grain yield based on 34 leaf-extracted metabolomic profiles reached substantially inferior prediction accuracies compared with GS based on 17,372 SNP markers in hybrid wheat. Furthermore, Riedelsheimer et al. (2012) and Zhao et al. (2015) observed no benefits in terms of prediction accuracies from combining genomic and metabolomic information into a single prediction model. In the future, further studies on prediction by means of models that simultaneously integrate genomic, transcriptomic, proteomic, and metabolomic information would be needed; thus, helping to understand how these different layers of biological information interact to shape the complex phenotypes of hybrid plants.

Considering Marker × Environment Interactions

It was already mentioned in Sect. 7.3.1.2 that genotype × environment interactions have the potential to negatively impact the predictability of GS. Considering this issue and by means of single-stage RR-BLUP approaches, Schulz-Streeck et al. (2013) partitioned the additive marker effects of Eq. (7.3) into main and marker × environment interaction effects for testcross performance prediction in maize. In a first attempt, the authors predicted untested genotypes in tested environments (i.e., environments already included within the estimation set), and they observed that GS predictabilities improved when shifting from a model with only additive effects to a model including main additive plus marker × environment interaction effects. Moreover, similar results were in general obtained by Zhang et al. (2015) for testcross performance prediction of grain yield considering well-watered and water-stressed environments, although the advantages of accommodating genotype × environment interactions within the GS model were less pronounced for days to anthesis and plant height prediction. Nevertheless, Schulz-Streeck et al. (2013) also found that these benefits disappeared when predictions were performed for untested genotypes in untested environments, highlighting the importance of across-environment cross-validation schemes to evaluate the prospects of GS in a more realistic manner. More studies should be conducted in this research area to elucidate if GS models including marker × environment interactions or approaches that dissect these interactions by considering environmental covariates (Heslot et al. 2014) have promising applications in hybrid crop species.

7.3.4 Implementation of GS in Hybrid Breeding

So far, the reader has probably realized that most studies on GS for hybrid performance in plants have mainly focused on the factors influencing the predictability of GS. For sure, this vast amount of knowledge has helped researchers and the plant breeding community in how to carefully interpret predictability values according to different biological and breeding scenarios and also in how to improve it when possible. Lately, studies have started to extensively discuss the implementation of GS in plant breeding. Because prediction within families corresponds to the most favorable scenario for the implementation of GS (Crossa et al. 2013), implementation studies have mainly focused on plant breeding programs based on biparental populations (Endelman et al. 2014; Krchov and Bernardo 2015; Longin et al. 2015; Lorenz 2013; Riedelsheimer and Melchinger 2013). In general, the authors concluded that GS should be considered as a tool to assist plant breeding in a similar way as MAS; hence, it is not intended to completely replace phenotypic selection (PS). Therefore, as already stated in section "Relatedness and Its Implications on the Predictabilities of GS Within Hybrid Breeding Programs", the main idea would be to partially testcross a particular biparental population and then to predict the remaining individuals using a model previously trained with the tested population fraction (Krchov and Bernardo 2015; Windhausen et al. 2012). Here lies the paradigm shift of GS because the purpose of phenotypic evaluations turns from exclusively guiding PS toward additionally calibrating statistical models for GS (Lorenz 2013). Subsequently, selection is supposed to be simultaneously performed in, both, estimation and prediction sets (Endelman et al. 2014; Krchov and Bernardo 2015; Riedelsheimer and Melchinger 2013). Certainly, one general question that arises in studies on GS implementation is: Given a limited budget, what is the optimal allocation of resources between estimation and prediction sets that maximizes the ratio of selection gains per unit of time between GS and pure PS? In principle, the essence of this problem can be represented by means of a mathematical model, in which a particular objective function, which is subjected to some constraints, aims to be maximized based on the optimization of some decision variables according to certain parameters. Moreover, this dissected representation can potentially clarify the interrelationships between the different components of the problem, thus, facilitating its comprehension and analysis (Hillier and Lieberman 2001). In fact, GS implementation has been already described as a nonlinear optimization problem (Endelman et al. 2014; Riedelsheimer and Melchinger 2013). Relying on the available literature along with our own concepts and criteria about the topic, the current section of this chapter aims to introduce the problem of GS implementation for a testcrossed biparental population pertaining to a hybrid breeding program.

7.3.4.1 Towards a Successful Implementation of GS in Hybrid Breeding

Objective Function: The Relative Efficiency of GS over Pure PS

The objective function is the particular value aimed to be maximized (or minimized) throughout the optimization process and should be one of the first model components to be identified or defined. This value measures the performance expressed as a mathematical function of the decision variables (Hillier and Lieberman 2001). For instance, plant breeders rely on the selection gain (ΔG) to compare and measure the performance of different selection methods. In principle, ΔG is the difference of the mean genetic value between the offspring of the selected fraction and the whole population before selection. This value can be predicted as:

$$\Delta G = ih\sigma_G, \tag{7.6}$$

being i the selection intensity, whereas h and σ_G denote the square roots of the heritability (h^2) and the genetic variance of the population (σ_G^2), correspondingly. In truncation selection (one-tail selection), the i term is a function of the proportion of individuals being selected according to a particular threshold located away from the mean phenotype of the original population before selection (Falconer and Mackay 1996). Moreover, h^2 is expressed as follows:

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{G\times E}^2}{\text{Nr. Env}} + \frac{\sigma_e^2}{\text{Nr. Env}\times\text{Nr. Rep}}}, \tag{7.7}$$

where $\sigma_{G\times E}^2$ and σ_e^2 correspond to the genotype × environment interaction and residual variance components, respectively, whereas Nr. Env and Nr. Rep are the number of environments and replicates used in balanced field tests, correspondingly (Holland et al. 2003; Piepho and Möhring 2007). Then, h is interpreted as the accuracy of PS (Endelman et al. 2014; Lorenz 2013; Riedelsheimer and Melchinger 2013), whereas the accuracy of GS can be mechanistically decomposed as:

$$\text{GS accuracy} = \sqrt{\frac{\lambda h^2}{\lambda h^2 + 1}}, \tag{7.8}$$

being $\lambda = \frac{N_E}{M_e}$, where N_E is the number of genotyped and phenotyped individuals in the estimation set and M_e is the effective number of loci (Daetwyler et al. 2008). Nevertheless, Eq. (7.8) is presented here only for explanatory purposes because the relationship between N_E and GS accuracy should be ultimately determined in an empirical manner based on previous available genomic and phenotypic data from plant breeding institutions (Endelman et al. 2014; Krchov and Bernardo 2015). On top of this, considering that selection would be simultaneously performed within, both, estimation and prediction sets, the overall ΔG_{GS} of plant breeding programs based on GS should be a function of the ΔG achieved by GS for the genomic predicted fraction in addition to the ΔG within the estimation set (Endelman et al. 2014; Krchov and Bernardo 2015; Riedelsheimer and Melchinger 2013). In this sense, because genomic predictions can also be obtained for the genotypes conforming the estimation set at no extra costs, these predictions could also be combined with the phenotypic data in a molecular selection index (Endelman et al. 2014; Lorenz 2013; Riedelsheimer and Melchinger 2013) as originally proposed for MAS more than two decades ago by Lande and Thompson (1990). Therefore, accuracies of selection within the estimation and prediction sets will be differentially denoted as r_E and r_P, correspondingly. Consequently and according to Eq. (7.6), ΔG_{GS} can be expressed as:

$$\Delta G_{GS} = \left(\left(\frac{S_E}{S} \right) i_E r_E + \left(\frac{(S - S_E)}{S} \right) i_P r_P \right) \sigma_G, \tag{7.9}$$

where S and S_E are the number of genotypes selected from the whole population and the estimation set, respectively, whereas i_E and i_P are the selection intensities within the estimation and prediction sets, correspondingly (Endelman et al. 2014; Riedelsheimer and Melchinger 2013). A similar formulation to Eq. (7.9) was presented by Krchov and Bernardo (2015). Ultimately, the relative efficiency (RE) of GS compared with pure PS should be maximized. This metric is defined as the ratio between ΔG_{GS} and the previously maximized ΔG for pure PS given the same assumptions or conditions. In this sense, ΔG_{GS} should be compared with the best ΔG attainable by means of pure PS in the same biparental population. Thus, the breakeven point for a successful GS implementation is when the RE of GS reaches unity because values above this threshold reflect a better competitiveness for GS over pure PS (Endelman et al. 2014; Krchov and Bernardo 2015; Riedelsheimer and Melchinger 2013). Last but not least, because GS has the potential to accelerate plant breeding programs by reaching more selection cycles than pure PS within the same amount of time, more important than the direct comparison between ΔG_{GS} and ΔG of pure PS is the contrast between their ΔG per unit of time (Longin et al. 2015). Examples of GS implementation with more than one stage of selection can be found elsewhere (Endelman et al. 2014; Longin et al. 2015; Lorenz 2013); however and for simplicity, we will approach the optimization problem by solely considering one-stage selection.

The Main Constraint: The Budget

The constraints are any restrictions on the values that the decision variables can take, and they are mathematically expressed by means of inequalities or equations (Hillier and Lieberman 2001). In other words, restrictions basically shape the space of all possible (optimal and suboptimal) solutions for the optimization problem.

In consequence, the basic constraint for any plant breeding strategy would be that its costs do not exceed the available budget. In the context of GS, this is well represented by the following inequation:

$$N_E C_E + (N - N_E) C_P \leq \text{Budget}, \tag{7.10}$$

where C_E and C_P correspond the costs of estimation and prediction sets, respectively, whereas N is the total number of individuals in the biparental population before selection. Costs and budget can be conveniently expressed in plot equivalents or yield plot units, that is, the cost of phenotyping one yield plot (Endelman et al. 2014; Riedelsheimer and Melchinger 2013). Moreover, costs can be further decomposed as follows: C_P corresponds to the costs of producing a line (C_L) plus the costs of genotyping (C_G), whereas $C_E = C_L + C_G + C_F$, being C_F the cost of the field trials using a number of plots determined by the Nr. Env × Nr. Rep combination. In this sense, Eq. (7.10) represents the trade-off existent in GS between the number of plots used for field evaluations (phenotyping intensity) and N_E (Lorenz 2013). In parallel, the constraint for pure PS is reduced to: $N(C_L + C_F) \leq$ Budget, reflecting that the trade-off between the number of individuals and phenotyping intensity is also present in pure PS (Endelman et al. 2014). Similar formulations of costs can be found in studies on GS implementation (Endelman et al. 2014; Krchov and Bernardo 2015; Longin et al. 2015; Lorenz 2013; Riedelsheimer and Melchinger 2013).

So far, studies on GS implementation have considered, either explicitly or indirectly, that the quantity of seed pertaining to the F_1 of the testcross will be enough to perform sufficient field trials. However, the production of F_1 seeds is a well-known constraint in hybrid breeding, especially in naturally self-pollinated species (Longin et al. 2012; Whitford et al. 2013). Plant breeding institutions give solution to this problem, for instance, by increasing the planting area for the parent line(s), which in turn results in higher costs (Longin et al. 2012). Nevertheless, GS could allow the prediction of testcross performance for parent lines, which do not produce enough F_1 seeds for testing in field trials. Being this an advantage for GS over PS, future studies on GS implementation should consider a seed quantity constraint to explore this potential benefit of GS.

Decision Variables

Decision variables are a quantitative representation of the decision to be made. They conform the group of all quantities that will be changed (optimized) along the space of solutions (determined by the constraints) during the optimization process and, ultimately, leading to the maximization (or minimization) of the objective function at their optimum values (Hillier and Lieberman 2001). According to the objective function and constraint presented, the decision variables for the optimal allocation of resources in GS correspond to:

(i) N_E: provided N was previously optimized for pure PS, optimizing N_E would elucidate which proportion of the population should be, both, phenotyped and genotyped, or only genotyped to maximize the RE of GS considering a particular budget and other assumptions.

(ii) i_E and i_P: given N and N_E along with preknown properties of the probability distributions for the testcross performances of estimation and prediction sets, the optimization of i_E and i_P would determine the fraction of top-ranking genotypes that should be selected within each respective set (Burrows 1975; Falconer and Mackay 1996). Ultimately, this optimized fraction would set the optimum values for S_E and (S-S_E), indicating the number of individuals that should be selected from estimation and prediction sets, respectively, and allowing the maximization of the RE of GS according to the particular budget and other assumptions.

(iii) Nr. Env and Nr. Rep: the optimization of the Nr. Env × Nr. Rep combination will indicate the phenotyping intensity within the estimation set, which is necessary to maximize the RE of GS taking on count the particular budget limitation and other assumptions.

Parameters

Parameters are the constants (coefficients and right-hand sides) present in the constraints as well as in the objective function (Hillier and Lieberman 2001) and, in contrast to the decision variables, they are assumed as fixed known values. The main parameters considered in the present formulation of GS implementation problem are the budget and costs (C_L, C_G and C_F). However, because of the uncertainty associated to the actual values of parameters, assigning the proper quantities to them is a very delicate task. In consequence, it would be important to investigate how the solution for the optimization problem would change if other possible values were assigned to the parameters. Accordingly, sensitive parameters will correspond to all those constants whose value cannot be modified without changing the optimal solution of the problem. Thus, special care should be allocated to the assignment of values for those particular parameters (Hillier and Lieberman 2001). For instance, even though it has been shown that there is a great amount of flexibility in choosing the optimal N_E and phenotyping intensity levels that maximize ΔG_{GS} (Lorenz 2013; Riedelsheimer and Melchinger 2013), this flexibility for the optimal solution strongly relies on the available budget because smaller budgets will restrict the set of possible solutions that maximize ΔG_{GS} in one-stage selection (Riedelsheimer and Melchinger 2013). In addition, the budget and C_G will determine whether or not GS can compete with pure PS (RE $\geq$ 1), with higher budgets (Krchov and Bernardo 2015; Riedelsheimer and Melchinger 2013) and lower C_G (Endelman et al. 2014; Krchov and Bernardo 2015; Riedelsheimer and Melchinger 2013) having in general positive effects on the RE of GS. Furthermore, the influence of C_G on RE of GS becomes more important with smaller budgets (Riedelsheimer and Melchinger 2013). Therefore, studies on GS implementation often speculate about a future decline of C_G, implying a more favorable scenario for GS implementation (Endelman et al. 2014; Krchov and Bernardo 2015; Riedelsheimer and Melchinger 2013).

Last but not least, M_e, σ^2_G, $\sigma^2_{G\times E}$ and σ^2_e are estimated values according to a particular population, test environments and target trait. Like any estimator, they would rely on the sample used for estimation, and hence, their values would be also subjected to uncertainty. Nevertheless, the impacts of changes in these parameters on the decision-making process in GS have been less studied. Briefly, high values of M_e would reflect the polygenic nature of a particular trait (Falconer and Mackay 1996). Assuming different M_e values for grain yield testcross performance in maize, Riedelsheimer and Melchinger (2013) observed that M_e has a negative influence on the RE of GS, being this trend more pronounced when low budgets are available. Additionally, these authors also studied the influence of the importance of $\sigma^2_{G\times E}$ and σ^2_e relative to σ^2_G ($\sigma^2_{G\times E} : \sigma^2_G$ and $\sigma^2_e : \sigma^2_G$ ratios, respectively) on the solutions for the optimal allocation of resources in GS. Under the assumption of a high budget, and that C_G is less than the cost of phenotyping one plot, Riedelsheimer and Melchinger (2013) showed that the RE of GS stays nearly constant at ~1.3 regardless of the assumed value for $\sigma^2_{G\times E} : \sigma^2_G$ and also, that increasing the $\sigma^2_e : \sigma^2_G$ ratio has a positive effect on the RE of GS. The authors attributed their observations to the differences in reallocation of resources between GS and pure PS that concomitantly occurred when varying $\sigma^2_{G\times E} : \sigma^2_G$ and $\sigma^2_e : \sigma^2_G$. Nonetheless, future studies should evaluate if these last observations hold truth for a less favorable scenario with more severe budget limitations and current genotyping costs.

7.3.4.2 Model Recalibration After a Successful GS Implementation in Early Breeding Stages: A Proposal

It was already highlighted in section "The Main Constraint: The Budget" that the decisions of plant breeders are always restricted by a limited budget; hence, they will always confront a trade-off between phenotyping intensity and the number of individuals being tested in field trials. Moreover, from section "The Main Constraint: The Budget", it is also concluded that even ignoring budget limitations, the limited quantity of seed belonging to the F_1 of a testcross would be an additional constraint for the phenotyping intensity of the testcross performance, especially during early breeding stages. In addition, it becomes clear from Eq. (7.7) that a restriction in phenotyping intensity is expected to decrease the potentially achievable h^2, and, according to Eq. (7.8), this limitation in h^2 will constrain the GS accuracy at any given N_E. Nevertheless, as a plant breeding program proceeds, the phenotyping intensity will be increased for the individuals being selected. Thus, incorporating this high-quality phenotypic data into the estimation set has the potential to improve prediction accuracies for future lines from (or very close related to) the biparental population in which GS was originally implemented. Consequently, new marker effects will be obtained by using updated estimation sets, allowing the recalibration of GS models. At the same time, new phenotypic data will be available for those genotypes whose early selection relied only on genomic predictions and this new information could be used to increase the size of the estimation set, which in turn would further improve GS accuracy according to Eq. (7.8). However, there are two important points that should be taken on count before proceeding with the GS recalibration:

First, there is an intrinsic issue of selection in plant breeding which has being ignored within the present proposal for GS recalibration: plant breeders want to increase the frequency of favorable alleles within the selected set (Bernardo 2010), leading to a decrease in σ_G^2 and to the misrepresentation of alleles with negative effects for a particular trait in the selected fraction. In this sense, updating the estimation set with the new information would mostly increase the phenotyping intensity for favorable alleles and, at the same time, would increase the frequency of favorable alleles within the estimation set. The impacts of PS within the estimation set on GS accuracy were studied by Zhao et al. (2012b) using grain yield testcross performance data of maize. They observed that using one-tail selection combined with high selection intensities within the estimation set led to a substantial decrease in GS accuracy compared with unselected estimation sets of the same size. In addition, they found that unselected estimation sets with low phenotyping intensities reached higher GS accuracies than estimation sets with much higher phenotyping intensity but subjected to one-tail selection and high selection intensity. Interestingly, their results also showed that estimation sets subjected to bidirectional selection (two-tails selection) in combination with high selection intensities reached superior GS accuracies than unselected estimation sets of the same size. Moreover, they concluded that just a small proportion of low performing genotypes (10–15% of the total fraction selected from both tails) would be enough for this purpose. Nevertheless, updating the estimation set by means of this last strategy implies a paradigm shift of selection within plant breeding programs and, consequently, should be further analyzed under the eye of the "economics" of GS for its wide acceptance by the plant breeding community.

Second, a simulation study showed that GS predictability does not always coincide with the accuracies at the individual level (Clark et al. 2012), implying that even though GS could reach high predictability or accuracy levels (according to the definitions in Sect. 7.3.1), some genotypes within the predicted set would be very accurately predicted and others would not. Prediction accuracy at the genotype level is often termed as "reliability," and it has been extensively used in the field of animal science (Clark et al. 2012; Mrode 2005; VanRaden 2008) and later in the context of GS in hybrid crops (Akdemir et al. 2015; Rincent et al. 2012). Furthermore, the simulation study of Clark et al. (2012) also showed that the reliability is highly associated to the maximum level of relatedness between estimation set and the particular genotype being predicted. In other words, highly reliable predictions are expected for genotypes, which were very well represented by a few or even by a single extremely closely related genotype(s) within the estimation set. Additionally, the reliability criterion has been used to identify individuals that are best suited for the conformation of the estimation set within diverse populations of maize (Akdemir et al. 2015; Rincent et al. 2012). In this sense, both studies in maize demonstrated that, in general, estimation sets that maximize the average reliability of the prediction set lead to higher GS predictabilities or accuracies than estimation sets constructed by random sampling. Nonetheless, the impacts of using the reliability of predicted genotypes as threshold criteria for updating the information within the estimation set of biparental populations have not been evaluated so far; therefore, studies will be needed to elucidate if reliabilities are also useful to further improve or maintain accuracies in the context of GS recalibration.

7.3.4.3 Final Words for the Implementation of GS in Hybrid Breeding

GS implementation in hybrid breeding is challenging because of the number of variables and imponderables involved in the optimal allocation of resources between estimation and prediction sets that influence the RE of GS under a given budget constraint. The current section was intended to give readers the basic conceptual framework of this problem considering biparental populations. It is anticipated that the problem formulation presented here is a simplification of the real picture; hence, readers are encouraged to make use of available decision support software (e.g., Endelman et al. 2014; Riedelsheimer and Melchinger 2013) to get further insights into how decision variables and parameters influence the RE of GS. Furthermore, readers with intermediate to advanced programming/planning skills are invited to elaborate their own models to find an optimal solution for the GS implementation problem. Preferably, the GS implementation should be planned and evaluated in an integrated manner and from a multidisciplinary point of view, considering together the skills and knowledge of plant breeders, scientists, technicians, process engineers and managers. In the future, once several plant breeding institutions and companies have already implemented GS, people should start to analyze successful and unfortunate cases of study to gain the empirical knowledge that is necessary to bridge the gap between theory and practice for the GS implementation problem in hybrid breeding.

References

Akdemir D, Sanchez JI, Jannink JL (2015) Optimization of genomic selection training populations with a genetic algorithm. Genet Sel Evol 47:38

Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzunova M, Simianer H, Schön CC (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339–350

Albrecht T, Auinger HJ, Wimmer V, Ogutu JO, Knaak C, Ouzunova M, Piepho HP, Schön CC (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor Appl Genet 127:1375–1386

Becker H (2011) Pflanzenzüchtung (in German). Auflagennr. 2. Verlag Eugen Ulmer, Stuttgart

Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34:20–25

Bernardo R (1996) Best linear unbiased prediction of maize single-cross performance. Crop Sci 36:50–56

Bernardo R (2010) Breeding for quantitative traits in plants. Stemma Press, Woodbury

Bernardo R (2014) Genomewide selection when major genes are known. Crop Sci 54:68–75

Bos I, Caligari P (2008) Selection methods in plant breeding, 2nd edn. Springer, Dordrecht

Bruce AB (1910) The Mendelian theory of heredity and the augmentation of vigor. Science 32:627–628

Burrows PM (1975) Expected selection differentials for directional selection. Biometrics 28:1091–1100

Clark SA, Hickey JM, Daetwyler HD, Van der Werf JHJ (2012) The importance of information on relatives for the prediction of genomic breeding values and implications for the makeup of reference populations in livestock breeding schemes. Genet Sel Evol 44:4

Collins GN (1921) Dominance and vigor of first generation hybrids. Am Nat 55:116–133

Comstock RE, Robinson HF, Harvey PH (1949) A breeding procedure designed to make maximum use of both general and specific combining ability. Agron J 41:360–367

Crossa J, Pérez-Rodríguez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J et al (2013) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112:48–60

Crow JF (1948) Alternative hypotheses of hybrid vigor. Genetics 33:477–487

Da Y, Wang C, Wang S, Hu G (2014) Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. PLoS One 9: e87666

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:e3395

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

East EM (1936) Heterosis. Genetics 21:375–397

Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, Sorrells ME, Jannink JL (2014) Optimal design of preliminary yield trials with genome-wide markers. Crop Sci 54:48–59

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Ronald Press Company, New York

Feher K, Lisec J, Römisch-Margl L, Selbig J, Gierl A, Piepho HP, Nikiloski Z, Willmitzer L (2014) Deducing hybrid performance from parental metabolic profiles of young primary roots of maize by using a multivariate Diallel approach. PLoS One 9:e85435

Gianola D, van Kaam JB (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776

Gowda M, Zhao Y, Maurer HP, Weissmann EA, Würschum T, Reif JC (2013) Best linear unbiased prediction of triticale hybrid performance. Euphytica 191:223–230

Gowda M, Zhao Y, Würschum T, Longin CFH, Miedaner T, Ebmeyer E, Schachschneider R, Kazman E, Schacht J, Martinant JP, Mette MF, Reif JC (2014) Relatedness severely impacts accuracy of marker-assisted selection for disease resistance in hybrid wheat. Heredity 112:552–561

Guo T, Li H, Yan J, Tang J, Li J, Zhang Z, Zhang L, Wang J (2013) Performance prediction of F1 hybrids between recombinant inbred lines derived from two elite maize inbred lines. Theor Appl Genet 126:189–201

Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G (2014) Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet 15:30

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397

Habier D, Fernando R, Kizilkaya K, Garrick D (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinf 12:186

Hallauer AR, Carena MJ, Miranda Filho JB (2010) Quantitative genetics in maize breeding. Iowa State University Press, Ames

Hayashi T, Iwata H (2013) A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinf 14:34

Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91:47–60

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph

Henderson CR (1985) Best linear unbiased prediction of non-additive genetic merits. J Anim Sci 60:111–117

Heslot N, Akdemir D, Sorrells M, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127:463–480

Hillier FS, Lieberman GJ (2001) Introduction to operations research, 7nd edn. McGraw Hill, New York

Hjorth JSU (1994) Computer intensive statistical methods. Validation model selection and bootstrap. Chapman & Hall, London

Hofheinz N, Borchardt D, Weissleder K, Frisch M (2012) Genome-based prediction of test cross performance in two subsequent breeding cycles. Theor Appl Genet 125:1639–1645

Holland JB, Nyquist WE, Cervantes-Martińex CT (2003) Estimating and interpreting heritability for plant breeding: an update. In: Janick J (ed) Plant breeding reviews, vol 22. Wiley, New York, pp 9–112

Hull FH (1945) Recurrent selection for specific combining ability in corn. J Am Soc Agron 37:134–145

Jacobson A, Lian L, Zhong S, Bernardo R (2014) General combining ability model for genomewide selection in a biparental cross. Crop Sci 54:895–905

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–177

Jia Y, Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513–1522

Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. Genetics 201:759–768

Jones DF (1917) Dominance of linked factors as a means of accounting for heterosis. Genetics 2:466–479

Keeble F, Pellew C (1910) The mode of inheritance of stature and of time of flowering in peas (Pisum sativum). J Genet 1:47–56

Krchov LM, Bernardo R (2015) Relative efficiency of genomewide selection for testcross performance of doubled haploid lines in a maize breeding program. Crop Sci 55:2091–2099

Krchov LM, Gordillo GA, Bernardo R (2015) Multienvironment validation of the effectiveness of phenotypic and genomewide selection within biparental maize populations. Crop Sci 55:1068–1075

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

Lehermeier C, Schön CC, de los Campos G (2015) Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. Genetics. doi:10.1534/genetics.115.177394

Longin CFH, Mühleisen J, Maurer HP, Zhang H, Gowda M, Reif JC (2012) Hybrid breeding in autogamous cereals. Theor Appl Genet 125:1087–1096

Longin CFH, Mi X, Würschum T (2015) Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. Theor Appl Genet 128:1297–1306

Lorenz AJ (2013) Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. G3 3:481–491

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161

Massman JM, Gordillo A, Lorenzana RE, Bernardo R (2013) Genomewide predictions from maize single-cross data. Theor Appl Genet 126:13–22

Melchinger AE, Gumber RK (1998) Overview of heterosis and heterotic groups in agronomic crops. In: Lamkey KR, Staub JE (eds) Concepts and breeding of heterosis in crop plants. ASACSSA-SSSA Publication, Madison, pp 29–44

Meuwissen THE (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol 41:35

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Miedaner T, Zhao Y, Gowda M, Longin CFH, Korzun V, Ebmeyer E, Kazman E, Reif JC (2013) Genetic architecture of resistance to Septoria Tritici blotch in European wheat. BMC Genomics 14:858

Mirdita V, Liu G, Zhao Miedaner T, Longin CFH, Gowda M, Mette MF, Reif JC (2015) Genetic architecture is more complex for resistance to Septoria Tritici blotch than to Fusarium head blight in central European winter wheat. BMC Genet 16:430

Mrode RA (2005) Linear models for the prediction of animal breeding values, 2nd edn. CABI Publishing, Wallingford

Nishio M, Satoh M (2014) Including dominance effects in the genomic BLUP method for genomic evaluation. PLoS One 9:e85792

Patti GJ, Yanes O, Siuzdak G (2012) Innovation: metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 13:263–269

Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49:1165–1176

Piepho HP, Möhring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881–1888

Reif JC, Gumpert F, Fischer S, Melchinger AE (2007) Impact of genetic divergence on additive and dominance variance in hybrid populations. Genetics 176:1931–1934

Reif JC, Zhao YS, Würschum T, Gowda M, Hahn V (2013) Genomic prediction of sunflower hybrid performance. Plant Breed 132:107–114

Richey FD (1942) Mock-dominance and hybrid vigor. Science 96:280–281

Riedelsheimer C, Melchinger AE (2013) Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor Appl Genet 126:2835–2848

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44:217–220

Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodriguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize in breds (*Zea mays* L.) Genetics 192:715–728

Schnell FW, Cockerham CC (1992) Multiplicative vs. Arbitrary gene action in heterosis. Genetics 131:461–469

Schrag TA, Frisch M, Dhillon BS, Melchinger AE (2009) Marker-based prediction of hybrid performance in maize single-crosses involving doubled haploids. Maydica 54:353–362

Schulthess AW, Wang Y, Miedaner T, Wilde T, Reif JC, Zhao Y (2016) Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. Theor Appl Genet 129:273–287

Schulz-Streeck T, Ogutu JO, Gordillo A, Karaman Z, Knaak C, Piepho HP (2013) Genomic selection allowing for marker-by-environment interaction. Plant Breed 132:532–538

Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York

Stuber CW, Cockerham CC (1966) Gene effects and variances in hybrid populations. Genetics 54:1279–1286

Su G, Christensen OF, Ostersen T et al (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One 7:e45293

Technow F, Riedelsheimer C, Ta S, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125:1181–1194

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197:1343–1355

Tracy WF, Chandler MA (2006) The historical and biological basis of the concept of heterotic patterns in corn belt dent maize. In: Lamkey KR, Lee M (eds) Plant breeding: the Arnel R Hallauer international symposium. Blackwell Publishing, Ames, pp 219–233

VanRaden PM (2007) Genomic measures of relationship and inbreeding. Interbull Bull 37:33–36

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Wang Y, Mette MF, Miedaner T, Gottwald M, Wilde P, Reif JC, Zhao Y (2014) The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. BMC Genomics 15:556

Ward J, Rakszegi M, Bedo Z, Shewry P, Mackay I (2015) Differentially penalized regression to predict agronomic traits from metabolites and markers in wheat. BMC Genet 16:19

Whitford R, Fleury D, Reif JC, Garcia M, Okada T, Korzun V, Langridge P (2013) Hybrid breeding in wheat: technologies to improve hybrid wheat seed production. J Exp Bot 64:5411–5428

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink JL, Sorrels ME, Raman B, Cairns JE, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, Melchinger AE (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 2:1427–1436

Wricke G, Weber WE (1986) Quantitative genetics and selection in plant breeding. Gruyter, Berlin

Würschum T, Reif JC, Kraft T, Janssen G, Zhao Y (2013) Genomic selection in sugar beet breeding populations. BMC Genet 14:85

Xu S, Zhu D, Zhang Q (2014) Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc Natl Acad Sci U S A 111:12456–12461

Zhang X, Pérez-Rodríguez P, Semagn K, Beyene Y, Babu R, López-Cruz MA, San Vicente F, Olsen M, Buckler E, Jannink JL, Prasanna BM, Crossa J (2015) Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity 114:291–299

Zhao Y, Gowda M, Liu W, Würschum T, Maurer HP, Longin CFH, Ranc N, Reif JC (2012a) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124:769–776

Zhao Y, Gowda M, Longin CFH, Würschum T, Ranc N, Reif JC (2012b) Impact of selective genotyping in the training population on accuracy and bias of genomic selection. Theor Appl Genet 125:707–713

Zhao Y, Gowda M, Würschum T, Longin CFH, Korzun V, Kollers S, Schachschneider R, Zeng J, Fernando R, Dubcovsky J (2013a) Dissecting the genetic architecture of frost tolerance in Central European winter wheat. J Exp Bot 64:4453–4460

Zhao Y, Zeng J, Fernando R, Reif JC (2013b) Genomic prediction of hybrid wheat performance. Crop Sci 53:802–810

Zhao Y, Mette MF, Gowda M, Longin CFH, Reif JC (2014a) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity 112:638–645

Zhao Y, Mette MF, Reif JC (2014b) Genomic selection in hybrid breeding. Plant Breed. doi:10.1111/pbr.12231

Zhao Y, Li Z, Liu G, Jiang Y, Maurer HP, Würschum T, Mock HP, Matros A, Ebmeyer E, Schachschneider R, Kazman E, Schacht J, Gowda M, Longin CFH, Reif JC (2015) Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc Natl Acad Sci U S A. doi:10.1073/pnas.1514547112

Chapter 8
Opportunities and Challenges to Implementing Genomic Selection in Clonally Propagated Crops

Dorcus C. Gemenet and Awais Khan

8.1 Clonally Propagated Crops

Clonal, or vegetative, crops are asexually propagated, that is, successive mitoses of specialized plant tissues develop into a new clonal population from a single mother plant (Bisognin 2011). Asexual propagation is used in all important root and tuber crops, many forage crops, almost 75% of perennial fruit trees, wooden ornamentals, many cut flowers, pot plants, and forest trees (Miller and Gross 2011; Denis and Bouvet 2013; Grunenberg et al. 2009), and presents a number of advantages. It can lead to increased levels of heterozygosity, fix favorable combinations of important traits, eliminate undesirable crosses and the resulting deleterious effects, and allow easy identification and propagation of favorable mutations. It is also an efficient method for maintenance, conservation, and in vitro and ex vitro propagation of cultivars with no viable seeds (Bisognin 2011). Despite these benefits, breeding of clonally propagated crops also includes several challenges. Most clonal fruit and forest trees have long juvenile phases, extensive outcrossing, widespread hybridization, limited population structure, multiple origins, and ongoing crop–wild gene flow, and have suffered from mild domestication bottlenecks due to clonal propagation (Miller and Gross 2011). Many clonally propagated crops are polyploid, which enables them to adapt to rapidly changing environment by maintaining increased heterozygosity, thus reducing inbreeding depression (Griffin et al. 2011). The high natural heterozygosity means that these crops are not amenable

D.C. Gemenet
International Potato Centre (CIP), Apartado 1558, Lima 12, Peru

A. Khan (✉)
Plant Pathology and Plant-Microbe Biology Section, School of Integrative Plant Science, Cornell University, New York State Agricultural Experiment Station, Geneva, NY 14456, USA
e-mail: awais.khan@cornell.edu

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7_8

to self-pollination due to high inbreeding depression. These effects lead to decreased second- and third-order favorable interactions and a reduced frequency of trigenic and tetragenic loci interactions, as well as the possibility to accumulate masked deleterious recessive alleles (Bradshaw 1994). Overdominance and epistatic interactions at a given locus of a heterozygous genotype can mask deleterious effects, but they can emerge after selfing. Genetic studies in polyploid crops are often further complicated by the presence of different ploidy levels. Cultivated potato (*Solanum* sp.) could be diploid ($2n = 2x$), triploid, ($2n = 3x$), tetraploid, ($2n = 4x$), or pentaploid ($2n = 5x$) (Spooner et al. 2010); cultivated sweet potato (*Ipomoea batatas*) is hexaploid ($2n = 6x$) (Jones 1965); and the ploidy level of sugarcane is yet unknown (Gouy et al. 2013), whereas *Musa species* are triploid ($2n = 3x$) (Simmonds 1962). In these crops, genetic analysis is complicated by the presence of multiple alleles at a given locus, mixed inheritance patterns, association between ploidy and mating system variation, among others (Dufresne et al. 2014). The presence of several segregating alleles at each locus of highly heterozygous clonally propagated crops make their breeding challenging. For example, in potato, an autotetraploid, four alleles per loci could be segregating, meaning that crossing two tetra-allelic potatoes could result in 32 genetically distinct genotypes with different levels of trait expression, different from the original parents. Therefore, potato breeders need to evaluate large populations to find at least one genotype with the desirable allelic combinations (Jansky et al. 2016). Expanding to the six different alleles possible for autohexaploid sweet potato, the number of genotypes to be screened will be far too large to find desirable trait combinations from multiple different loci. In this chapter, we discuss conventional breeding methods and their challenges, the potential of genomic selection (GS), challenges for its implementation with some examples, and outlook of GS as a population improvement strategy in clonal crops.

8.2 Breeding Strategies for Clonal Crops

According to Simmonds (1979), breeding clonal crops requires a crossing step to provide sexual seeds, as a break from the normal clonal propagation, and to create genetic variation that can be exploited during selection in subsequent cycles, before reverting to clonal selection. This crossing step involves two heterozygous parents to produce clonally propagated hybrids and is a distinctive feature in these crops. The resulting hybrids are therefore heterozygous and heterogeneous, and can display all forms of genetic effects (additive, dominance, overdominance, and epistatic; Ceballos et al. 2015). The cross is followed by phenotypic mass selection or recurrent selection. These conventional approaches are both time and resource consuming, as they involve crossing in one generation, planting of true seed plants in another generation, and observation of clones from selected true seed plants over several generations and different environments, to evaluate genotype-by-environment (G x E) interactions (Gruneberg et al. 2009). Another challenge in breeding, selection, and conducting yield trials for clonally propagated crops is producing enough planting material. Due to their heterozygosity and self-incompatibility, selfing may result in inbreeding depression, and crossing to make seed can take a long time. Therefore, vegetative propagules are produced and then used for field and selection trials. However, each plant can only make a fixed number of vegetative propagules, and multiplication can thus be slow and costly. For example, it takes about 45 years from a cross to have enough planting material for replicated multienvironment field trials in cassava (Ceballos et al. 2015), whereas in tree fruits the breeding cycle may extend to a dozen years or more (van Nocker and Gardiner 2014). Another challenge, in several clonally propagated crops, is maintaining disease-free ("clean" "good *seed* quality," or low disease) stocks (clones) during the clonal-increase phase; otherwise, trait evaluations could be significantly affected. Genome-based selection, especially at the seedling stage of true seed plants, can be an important approach towards expediting the breeding process by shortening the lengthy selection cycle, removing the need of several subsequent clonal selection cycles and reducing the time required for multiplication (Myles 2013). Resources are also saved by maintaining fewer genotypes for phenotypic evaluation. Genome-based selection can be achieved through using either one of two methods: Firstly, conventional marker-assisted selection (MAS) using diagnostic markers linked with a few qualitative and quantitative trait loci (QTL) with large effects; secondly, GS using genomic estimated breeding values (GEBVs) predicted by high-density genome-wide molecular markers to select superior progeny (Meuwissen 2007) or combining both (Spindel et al. 2015). Conventional selection (phenotypic selection, PS) is usually based on multiple traits. In both PS and GS, a multistep selection and/or indexes could be developed to filter and select clones with the best combinations of traits of interest (Fig. 8.1).

8.3 Key Features of Genomic Selection

Genomic selection (GS) is a method of selection proposed by Meuwissen et al. (2001) using genome-wide genotypic data to predict the phenotypic performance of a genotype by estimating its breeding value or total genetic value, referred to as GEBV. In the GS process, statistical models are used to estimate the relationship between phenotypes and genotypes in a subset of the population normally referred to as a training population. The models developed are then tested with a validation set, which is a subset of the population phenotyped in the same environment(s) as the training set, a process called cross-validation. Validated models are then applied to a breeding population with only genotypic data to determine GEBV of the genotypes, and finally, elite individuals with desirable trait combinations are selected based on these GEBVs only (Nakaya and Isobe 2012). Several statistical models have been proposed for GS, each with different assumptions on marker effects and the relationship between the markers. However, for a given predictive model to perform well for a given trait, it has to follow the continuum of the genetic

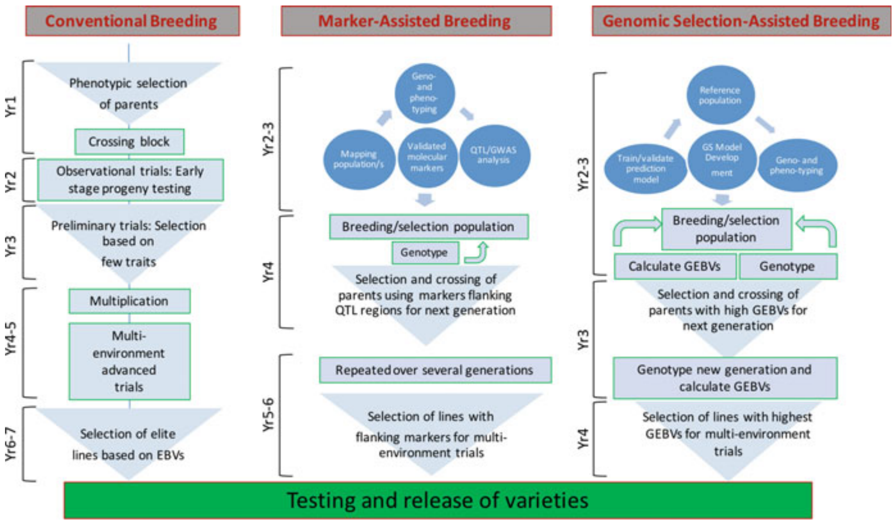

Conventional Breeding
Marker-Assisted Breeding
Genomic Selection-Assisted Breeding
Yr1
Phenotypic selection of parents
Crossing block
Yr2
Observational trials: Early stage progeny testing
Yr3
Preliminary trials: Selection based on few traits
Yr4-5
Multiplication
Multi-environment advanced trials
Yr6-7
Selection of elite lines based on EBVs
Yr2-3
Geno- and pheno-typing
Mapping population/s
Validated molecular markers
QTL/GWAS analysis
Yr4
Breeding/selection population
Genotype
Selection and crossing of parents using markers flanking QTL regions for next generation
Yr5-6
Repeated over several generations
Selection of lines with flanking markers for multi-environment trials
Yr2-3
Reference population
Train/validate prediction model
GS Model Develop ment
Geno- and pheno-typing
Breeding/selection population
Calculate GEBVs
Genotype
Yr3
Selection and crossing of parents with high GEBVs for next generation
Genotype new generation and calculate GEBVs
Yr4
Selection of lines with highest GEBVs for multi-environment trials
Testing and release of varieties

Fig. 8.1 (**a**) A conventional breeding scheme for clonally propagated crop includes crossing, selection, and yield trials. (**b**) A genomic-assisted scheme for clonally propagated crop includes crossing, selection, and yield trials

architecture of that trait (Poland and Rutkoski 2016). For quantitative traits, mixed models such as genomic best linear unbiased prediction (G-BLUP) and ridge-regression best linear unbiased prediction (RR-BLUP) are widely used, as they mimic the conventional best linear unbiased prediction (BLUP) normally used in phenotypic selection (PS). Several Bayesian models work better for traits that fall in-between quantitative and qualitative, that is, they are regulated by a few major genes. These include BayesA, BayesB (Meuwissen et al. 2001), BayesCπ (Habier et al. 2011), and Bayessian LASSO (least absolute shrinkage and selection operator; Park and Casella 2008). However, all these methods assume additive genetic effects, which is not the case for clonally propagated crops. For clonal crops, these methods would be adequate if GS was only applied to select for new parents for the crossing step, in which case, additive genetic variation would be important. However, for variety development in clonal crops, both additive and nonadditive genetic effects are important. Methods that model both additive and nonadditive effects would then be required if GS was to be successfully deployed for variety development in these crops (Azevedo et al. 2015). For these crops, models such as reproducing kernel Hilbert space (RKHS; Gianola and van Kaam 2008) and random forest (RF; Breiman 2001), which have been shown to capture both additive and nonadditive effects, could be applied. The predictive ability for each model is estimated as the correlation between the observed breeding values and the predicted breeding values from GS, whereas the prediction accuracy is calculated by dividing the predictive ability by the mean heritability of the validation set (Poland and Rutkoski 2016). As such, this prediction accuracy is affected by several factors.

These include the size of the training set: the larger the training set, the better the accuracy of prediction; the heritability of the trait: the higher the heritability of the trait, the better the accuracy; relationship between training and breeding populations: the closer the relationship, the better the accuracy; linkage disequilibrium (LD) between the marker and the trait: the higher the LD, the better the prediction accuracy. The LD between markers and QTLs (traits) also determines the number of markers required to perform GS: plants with rapid LD decay, like outcrossing species, require more markers, whereas less markers may be required for inbreeding species (Spindel et al. 2015; Nakaya and Isobe 2012). Other factors reported to affect prediction accuracy include population structure (pedigree), environment, data redundancy, epistasis, type of cross-validation (i.e. fivefold, tenfold, Jacknife), GS prediction models and accuracy calculation approach. GS requires a clear definition of the breeding scenario in which selection will be implemented and a detailed analysis of the population structure. Larger training sets that are closely related to the target breeding and selection populations give higher prediction accuracy. GS studies in maize have shown that a major portion of the prediction accuracy estimated using prediction models developed with unrelated populations comes from population structure and is affected by environment. Higher prediction accuracy can be achieved by also modeling GE (Genotype × Environment) and borrowing information from related environments (Crossa et al. 2014; Windhausen et al. 2012).

8.4 Challenges to Implementing Genomic Selection in Clonal Crops

8.4.1 Modeling of Genetic Effects and Heritability

The unique features of the population and quantitative-genetic parameters of clonally propagated crops may pose challenges to the adoption of GS models currently developed for seed-propagated crops. Most of the proposed models in GS mainly model additive effects and assume dominance and epistatic effects as part of the residual. This holds true for seed-propagated crops, as it is not possible to transfer nonadditive genetic effects to the next generation sexually, rather new nonadditive combinations are formed during each sexual recombination cycle. However, for clonally propagated crops, dominance and epistatic effects play an important role beside additive effects and need special consideration. This is because the whole set of alleles, together with their interactions, are passed to the next generation through clonal propagation. Ceballos et al. (2015) demonstrated the presence of additive, dominance, and epistatic effects in cassava whose magnitudes differed for individual traits. Because gene action is locus and trait specific, the currently available GS models will give different prediction accuracy for different traits in clonally propagated crops. In sugarcane, Gouy et al. (2013) reported similar predictive accuracy for several GS models on a single trait but significantly different predictive accuracy for each model on different traits. This means that additive, dominance, and epistatic genetic effects must be properly analyzed for every trait to select the best model for each. Most often, specialized populations with specific mating designs are needed to estimate the extent of these gene actions, and so far, these populations have not been properly defined for most traits of importance in clonally propagated crops. Furthermore, estimation of narrow sense heritability, one of the key factors affecting prediction accuracy in GS (Nakaya and Isobe 2012), is mainly based on additive genetic variation, which holds true for seed-propagated crops, but not for clonal crops. This then calls for the use of additive, dominance, and epistatic genetic effects in calculating broad sense heritability estimates used in GS for clonally propagated crops. Munoz et al. (2014) used both pedigree-based and marker-based information to model additive, dominance, and first-order epistatic interaction effects in the tree species *Pinus taeda*. They concluded that prediction accuracy of GEBV improved by including additive and nonadditive effects to the predictive models. Wolfe et al. (2016) used both additive only and additive plus nonadditive effect models to show that including nonadditive effects in the model improved prediction accuracy. On the other hand, including large effect QTL as fixed effects in additive-only model improved prediction accuracy for cassava mosaic disease resistance. Prediction accuracies ranged from 0.53 to 0.58 with different models, indicating that GS would be useful for selecting cassava mosaic disease resistance.

8.4.2 Linkage Disequilibrium between Markers and Quantitative Trait Loci

Linkage disequilibrium (LD), the nonrandom association of alleles at different loci, is another factor affecting prediction accuracy in GS (Nakaya and Isobe 2012). As opposed to linkage, which refers to the physical connection of loci on a chromosome and are inherited together, LD refers to the correlation among alleles in the whole population (Flint-Garcia et al. 2003). LD breaks down both by recombination (intrachromosomal LD) and independent assortment (interchromosomal LD), as well as by other factors that affect the Hardy–Weinberg equilibrium (Flint-Garcia et al. 2003). Estimation of LD in clonally propagated crops is only possible during the true seed plant stage resulting from meiotic events at the crossing step. The heterozygosity and heterogeneous nature of most clonal species ensures a large breakdown of LD at this crossing step. Because GS assumes that the marker density used is large enough that all genes are in LD with some of the markers (Meuwissen 2007), this implies that several markers are required to ensure higher prediction accuracy of the GEBV in clonally propagated crops relative to seed-propagated in which selfing may be possible. There has been little systematic evaluation of the extent of LD in most clonally propagated crops; in most cases, the effective number of markers required for efficient GS is unknown. Even for those crops where LD has been studied a little, there are contrasting findings regarding LD depending on the marker systems used. In potato, Simko et al. (2006), using SNPs within 100 bp derived from bacterial artificial chromosome (BAC) ends, reported LD extending to 10 cM possibly due to the shorter physical distance of the SNPs, whereas D'hoop et al. (2010) reported LD extending to 5 cM using over 3000 AFLP markers, whereas Stich et al. (2013) reported LD decay after 275 bp using genome-wide SNPs from the SOLCAP array. In sugarcane, Jannoo et al. (1999) reported LD extending to 10 cM, whereas Raboin et al. (2008) reported LD ranging from 0–30 cM. In case of cassava, LD analysis based on SNPs from GBS showed decay between 10–50 kb (Wolfe et al. 2016). In banana (Musa sp), Sardos et al. (2016) showed that LD extended to 10–100 kb. In apple, Kumar et al. (2013) reported LD persisting to approximately 1 cM. For crops like sweet potato and yam, efforts are still underway at the International Potato Center (CIP) and the International Institute of Tropical Agriculture (IITA). Findings from the above studies, with the exception of the findings by Stich et al. (2013), indicate that LD persists longer for polyploid clonal crops compared with diploid clonal crops. This can be attributed to the bottleneck in breeding polyploid crops using only few parents and the confounding effects of polyploidy on marker identification. Despite this, LD in clonal crops in general persists longer compared with outcrossing seed-propagated crops. In maize, Yan et al. (2009) showed LD decay within 1–10 kb on a global maize collection. This aspect can be attributed to clonal propagation that ensures reduced meiotic events. It is important to precisely estimate LD for most of these crops if GS were to be successful as this will enhance determining effective population sizes and genotyping densities that have great impact on the accuracy of genomic prediction (Grattapaglia and Resende 2011).

8.4.3 Genetic Architecture of Traits and Size of Training Population

The genetic architecture of a trait of interest affects prediction accuracy of GS (Nakaya and Isobe 2012). Genetic architecture of a trait is a complex of factors, including the number of genes controlling the trait and their genomic location, the effects of substituting alleles of these genes and the heritability of a trait (Poland and Rutkoski 2016). Many clonally propagated crops are self-incompatible and polyploid, resulting in multiple alleles and dosages at a given locus (Slater et al. 2016). Therefore, the allele combinations responsible for a given trait are numerous and mostly unknown. To apply GS in a breeding program, the training population used to develop prediction models should be large enough to capture representative combinations of alleles for traits of interest in the breeding population (Jannink et al. 2010). This is important because selection reduces additive genetic variation and reduces genetic gains in subsequent generations, whereas development of commercial varieties requires simultaneous improvement of many quantitative traits (Poland and Rutkoski 2016). For clonal crops, there has been little genetic gain for complex traits like yield compared with cereals, possibly due to using only a few parents and reduced meiotic combinations due to clonal propagation (Slater et al. 2016). So, to avoid bottlenecks from breeding, the effective population sizes may need to increase with increasing ploidy levels. Effective population size refers to the number of individuals who contribute offspring to the next generation while meeting the Hardy–Weinberg equilibrium. However, increasing effective population size also leads to more rapid decay of LD and reduced prediction accuracy (Grattapaglia and Resende 2011). No systematic analysis of the effective population sizes required for accurate estimation of GEBV has been studied in most clonally propagated crops. Initial estimation of effective population size for tetraploid potato did indicate the need for 79 initial parents (Slater et al. 2016). Furthermore, traits controlled by several small effect loci across the genome with complex genetic architecture are more responsive to genotype-by-environment interaction (G x E). Ly et al. (2013), using 17 traits in cassava, showed that prediction accuracy reduced between 0.01 to 0.18 for different locations compared with the training set, indicating a strong relationship between prediction accuracy and G x E. Resende et al. (2012) also showed reducing prediction accuracies with geographical distance between the model development sites and validation sites in loblolly pine (*Pinus taeda*). Because cassava and loblolly pine are diploid clonally propagated crops, it could be speculated that prediction accuracy for higher ploidy levels will be further reduced for traits showing strong G × E interactions. Therefore, the initial investment in development of GS models may be higher for clonally propagated crops compared with seed-propagated crops because of the need to evaluate larger numbers of training sets across several target environments, where environment here refers to both sites and seasons.

8.4.4 Number of Generations Following Training Model

One of the main attractions to implementing GS for crop improvement programs is the potential of lower costs and shorter generation intervals, arising from the ability to predict GEBVs with high accuracy, early in the breeding cycle, over several generations, without phenotyping selection populations. Prediction accuracy of GS depends on LD between markers and QTL, and is expected to decline in the generations following the population initially used for developing the training model for the estimation of GEBVs (Habier et al. 2007; Nakaya and Isobe 2012). The composition and genetic distance of individuals in a training population also affects prediction accuracy and differs for traits and the heritability of traits (Weng et al. 2016a, b). This is expected to be much more important in seed-propagated crops, where new combinations of nonadditive genetic effects are formed during each sexually reproductive cycle (Grattapaglia and Resende 2011). This may be less of a problem in clonally propagated crops, especially if the training population is sufficiently related to the breeding population because there is only one recombination step (the crossing step) per selection cycle, which can recur, and fixed genetic effects are fully passed on to the next generation through clones. Ly et al. (2013) did not find significant reduction in prediction accuracy due to relatedness or lack of relatedness between training and prediction sets in cassava. So, although initial investment for developing GS models may be higher for clonally propagated crops due to the higher marker density required to cover for a relatively large number of effective population size, the cost may be evened out by application of the same models for longer than for seed-propagated crops.

8.5 Examples of Genomic Selection in Clonal Crops

Most of the GS carried out in clonally propagated crops thus far has been proof-of-concept, that is, developing GS models on training sets, estimating prediction accuracies, and testing models in validation sets. Extension and implementation into practical breeding and selection programs has not been achieved fully, especially in the public sector. Here, we enumerate a few examples where GS has been tested in clonally propagated crops. de Oliveira et al. (2012) used the random regression-best linear unbiased prediction model (RR-BLUP) to estimate prediction accuracies in shoot weight (SW), fresh root weight (FRW), dry matter content (DMC), and starch yield (SY) in cassava. They showed that using only informative markers associated with a trait results in higher prediction accuracies for the respective traits. Prediction accuracies ranging from 0.67 to 0.83 were reported for SW, FRW, DMC, and SY. Given that prediction accuracies are always less than 1, phenotypic selection (PS) is always more efficient at selection that GS. However, GS becomes advantageous when considering the shortening of generation cycle involved in PS. de Oliveira et al. (2012) estimated genetic gains of GS versus PS for the above traits and concluded that reducing the generation cycle by half with GS would increase genetic gains by 39.4%, 56.9%, and 73.96% for DMC, FRY, and SW, respectively. In sugarcane, another complex, polyploid, clonally propagated crop, Gouy et al. (2013) tested four statistical models, Bayesian LASSO, ridge regression, reproducing kernel Hilbert space, and partial least square regression, showing correlations ranging from 0.11 to 0.62 between phenotypes and genotypes during cross validation, depending on the trait. Equal accuracy was seen for all models within a trait but with marked differences between traits. They concluded that their marker density (1499 dominant markers) may not have been large enough to cover the large sugarcane genome and capture the whole haplotype diversity, and suggested using multi-allelic markers to improve prediction. In oil palm (*Elaeis guineensis*), prediction accuracy ranged from 0.41 to 0.94, depending on the trait and the relationship of the population to the training set, but was not affected by the statistical method used (Cros et al. 2015). This was attributed to the small size of the training population and markers used, as well as to the complex genetic architecture for different traits. In potato, Habyarimana et al. (2014) used three GS models, Bayesian LASSO, genomic best linear unbiased prediction (G-BLUP), and reproducing kernel Hilbert space (RKHS), to evaluate prediction accuracy for yield, yield components, and quality traits. They reported prediction accuracy of $r > 0.60$ for several traits including carotenoids, tuber dry matter, and total yield. Meanwhile, in apple, prediction accuracy for ten traits (median of 0.19 and a maximum of 0.5) was strongly affected by distribution of traits and their heritability (Muranty et al. 2015). Traits with high heritability and normal phenotypic distribution showed response to selection. Furthermore, Kumar et al. (2012) used an 8 K Illumina Infinium chip to demonstrate the potential of GS for fruit quality traits at seedling stage, reducing the generation interval for the apple fruit trees with long juvenile phases. They compared RR-BLUP and Bayesian LASSO methods to show prediction accuracies ranging from 0.7 to 0.9 according to the trait analyzed but with little difference between the prediction models. They concluded that GS could accelerate the breeding process for fruit quality by making selections prior to the lengthy fruit quality phenotyping. Resende et al. (2012) showed prediction accuracies differing for different methods depending on genetic architecture of the traits in loblolly pine (*Pinus taeda*). Bayesian methods outperformed RR-BLUP for oligogenic traits because RR-BLUP assumes equal contribution of all markers and can overparameterize by fitting a large number of markers to a trait that is controlled by a few major genes. In other clonally propagated crops like horticultural and forest trees, sweet potato, yam, and banana, efforts are still underway to develop and implement GS models and to estimate prediction accuracy for traits of importance.

8.6 Outlook for Implementing Genomic Selection in Clonal Crop Breeding Programs

Genomic selection has great potential to expedite the breeding process in clonally propagated crops by shortening their long breeding cycle. However, the examples above show relatively moderate prediction accuracies for GS models in different crops, indicating room for improvement and refinement. Therefore, before the full potential for GS can be exploited for clonal crops, the challenges associated with population structure, architecture of traits, polyploidy, rapid LD decay, and heterozygosity have to be addressed in the GS models (Dufresne et al. 2014). Distinguishing between paralogous copies and the presence of high copy numbers of repetitive elements is difficult and poses a challenge for full genome annotations in polyploid crops (Leitch and Leitch 2008). The majority of conventional SNP-genotyping platforms and analytical tools have been developed for diploid crops and are often not suited to clonal crops. High heterozygosity and multiple alleles per locus in many of the clonally propagated crops is a challenge in developing pipelines for generating the high-density markers necessary for

GS. Further analytical solutions and pipelines are therefore required to allow these platforms to incorporate partial heterozygosity and allele dosage determination. Voorrips et al. (2011) developed models that estimate partial heterozygosity in tetraploid potato. Their methods efficiently assign bi-allelic marker scores in tetraploid species. Such a method could also be applied to higher ploidies if the data are of high quality, as more closely spaced peaks are expected at higher ploidies, which makes efficient assignment to classes a challenge if the data have too much noise. Serang et al. (2012) developed an algorithm that can estimate SNP-allele frequencies in individuals with multiple ploidy. They tested the methods on potato and sugarcane data and found that the methods identified the correct ploidies for all potato genotypes, whereas a few differences were observed in sugarcane, in agreement with the unknown ploidy levels of sugarcane genotypes. These studies are in the right direction but should be validated further in other polyploid species to allow genetic study of clonally propagated crops to benefit from next generation sequencing techniques. Once genotype calling and phasing can be done properly to allow development of high-density markers, efforts should be put into developing QTL, association mapping, and GS models that account for all quantitative-genetic (additive, dominant, and epistatic effects) parameters. Then, proper analysis of such parameters should be performed in available breeding populations to allow advancement from proof-of-concept status to applied breeding status.

References

Azevedo CF, de Resende MDV, e Silva FF, Viana JMS, Valente MSF, Resende MFR, Muñoz P (2015) Ridge, Lasso and Bayesian additive-dominance genomic models. BMC Genet 16(1):1

Bisognin DA (2011) Breeding vegetatively propagated horticultural crops. Crop Breed Appl Biotechnol S1:35–43, Brazilian Society of Plant Breeding

Bradshaw JE (1994) Quantitative genetics theory for tetrasomic inheritance. In: Bradshaw JE, Mackay GR (eds) Potato Genetics. CABI, Cambridge, pp 71–99

Breiman L (2001) Random forests. Mach Learn 45:5–32

Ceballos H, Kawuki RS, Gracen VE, Yencho GC, Hershey CH (2015) Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. Theor Appl Genet 128:1647–1667

Cros D, Denis M, Sánchez L, Cochard B, Flori A, Durand-Gasselin T, Nouy B, Omoré A, Pomiès V, Riou V, Suryana E, Bouvet J-M (2015) Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.) Theor Appl Genet 128:397–410

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112(1):48–60

Denis M, Bouvet JM (2013). Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genetics & Genomes, 9(1):37–51

D’hoop BB, Paulo MJ, Kowitwanich K, Sengers M, Visser RGF, van Eck HJ, van Eeuwijk FA (2010) Population structure and linkage disequilibrium unravelled in tetraploid potato. Theor Appl Genet 121:1151–1170 de Oliveira EJ, Resende MDV, Santos VDS, Ferreira CF, Oliveira GAF, da Silva MS, de Oliveira LA, Carlos Ivan Aguilar-Vildoso CI (2012) Genome-wide selection in cassava. Euphytica 187:263–276

Dufresne F, Stift M, Vergilino R, Marble BK (2014) Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Mol Ecol 23(1):40–69

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Gianola D, Van Kaam JB (2008) Reproducing Kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303

Gouy M, Rousselle Y, Bastianelli D, Lecomte P, Bonnal L, Roques D, Efile JC, Rocher S, Daugrois J, Toubi L, Nabeneza S, Hervouet C, Telismart H, Denis M, Thong Chane A, Glaszmann JC, Hoarau JY, Nibouche S, Costet L (2013) Experimental assessment of the accuracy of genomic selection in sugarcane. Theor Appl Genet 126:2575–2586

Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. Tree Genet Genomes 7:241–255

Griffin PC, Robin C, Hoffmann AA (2011) A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. BMC biology, 9(1):19

Grüneberg W, Mwanga R, Andrade M, Espinoza J (2009) Breeding clonally propagated crops. In FAO, selection methods: chapter 13, part 5 ftp://ftp.fao.org/docrep/fao/012/i1070e/i1070e04.pdf

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinform 12:186

Habyarimana E, Parisi B, Onofri C, Govoni F, Mandolino G Efficiency of genomic selection for yield, merceological and nutritional quality traits in hundred thirty-nine cultivates potato genotype. EAPR 2014 Brussels - 19th triennial conference, 6 to 11 July 2014, conference paper, doi:10.13140/2.1.1966.9286

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9(2):166–177

Jannoo N, Grivet L, Dookun A, D'Hont A, Glaszmann JC (1999) Linkage desequilibrium among modern sugarcane cultivars. Theor Appl Genet 99:1053–1060

Jansky SH, Charkowski AO, Douches DS, Gusmini G, Richael C, Bethke PC, Spooner DM, Novy RG, De Jong H, De Jong WS, Bamberg JB (2016) Reinventing potato as a diploid inbred line-based crop. Crop Sci 56:1–11

Jones A (1965) Cytological observations and fertility measurements of sweetpotato [*Ipomoea batatas* (L.) Lam]. Proc Am Soc Horticult Sci 86:527–537

Kumar S, Chagné D, Bink MCAM, Volz RK, Whitworth C, Carlisle C (2012) Genomic selection for fruit quality traits in apple. PloS One 7(5):e36674

Kumar S, Garrick DJ, Bink MCAM, Whitworth C, Chagné D, Richard K, Volz RK (2013) Novel genomic approaches unravel genetic architecture of complex traits in apple. BMC Genomics 2013(14):393

Leitch A, Leitch I (2008) Genomic plasticity and the diversity of polyploid plants. Science 320:481–483

Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, Gauch HG Jr, Okechukwu R, Dixon AGO, Kulakow P, Jannink J-L (2013) Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: a study in cassava. Crop Sci 53:1312–1325

Meuwissen T (2007) Genomic selection: marker-assisted selection on a genome-wide scale. J Anim Breed Genet 124(6):321–322

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Miller AJ, Gross BL (2011) From forest to field: perennial fruit crop domestication. Am J Bot 98 (9):1389–1414

Muñoz PR, Resende MFR, Gezan SA, Resende MDV, de los Campos G, Kirst M, Huber D, Peter GF (2014) Unraveling additive from non-additive effects using genomic relationship matrices. Genetics 198:1759–1768

Muranty H, Troggio M, Sadok IB, Rifai MA, Auwerkerken A, Banchi E, Velasco R, Stevanato P, van de Weg WE, Di Guardo M, Kumar S, Laurens F, Bink MCAM (2015) Accuracy and responses of genomic selection on key traits in apple breeding. Horticult Res 2:15060. doi:10.1038/hortres.2015.60

Myles S (2013) Improving fruit and wine: what does genomics have to offer? Trends Genet 29 (4):190–196

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

Park T, Casella G (2008) The Bayesian lasso. J Am Stat Assoc 103:681–686

Poland J, Rutkoski J (2016) Advances and challenges in genomic selection for disease resistance. Annu Rev Phytopathol 54:79–98

Raboin L-M, Pauquet J, Butterfield M, D'Hont A, Glaszmann J-C (2008) Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. Theor Appl Genet 116:701–714

Resende MFR, Muñoz P Jr, Resende MDV, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.) Genetics 190:1503–1510

Sardos J, Rouard M, Hueber Y, Cenci A, Hyma KE, van den Houwe I et al (2016) A genome-wide association study on the seedless phenotype in banana (Musa spp.) reveals the potential of a selected panel to detect candidate genes in a vegetatively propagated crop. PLoS One 11(5): e0154448

Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS One 7:e30906

Simko I, Haynes KG, Jones RW (2006) Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 173:2237–2245

Simmonds NW (1962) The Evolution of the Bananas. Tropical Sciences Series. Longman, London, pp. 170

Simmonds NW (1979) Principles of crop improvement. Longman, New York

Slater AT, Cogan NOI, Forster JW, Hayes BJ, Daetwyler HD (2016) Improving genetic gain with genomic selection in autotetraploid potato. Plant Genome 9. doi:10.3835/plantgenome2016.02.0021

Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E et al (2015) Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet 11(2):e1004982

Spooner DM, Gavrilenko T, Jansky SH, Ovchinnikova A, Krylova E, Knapp S, Simon R (2010) Ecogeography of ploidy variation in cultivated potato (Solanum sect. Petota). Am J Bot 97 (12):2049–2060

Stich B, Urbany C, Hoffmann P, Gebhardt C (2013) Population structure and linkage disequilibrium in diploid and tetraploid potato revealed by genome-wide high density genotyping using the SolCAP SNP array. Plant Breed 132:718–724

van Nocker S, Gardiner SE (2014) Breeding better cultivars, faster: applications of new technologies for the rapid deployment of superior horticultural tree crops. Horticult Res 1:14022. doi:10.1038/hortres.2014.22

Voorrips R, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi allelic marker data using mixture models. BMC Bioinformatics 12:172

Weng Z, Wolc A, Shen X, Fernando RL, Dekkers JC, Arango J et al (2016a) Effects of number of training generations on genomic prediction for various traits in a layer chicken population. Genet Sel Evol 48(1):1

Weng Z, Wolc A, Shen X, Fernando RL, Dekkers JCM, Arango J, Settar P, Fulton JE, O'Sullivan NP, Garrick DJ (2016b) Effects of number of training generations on genomic prediction for various traits in a layer chicken population. Genet Sel Evol 48:22

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink J-L, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, Melchinger AE (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 Genes Genomes Genet 2:1427–1436

Wolfe MD, Rabbi IY, Egesi C, Hamblin M, Kawuki R, Kulakow P, Lozano R, Del Carpio DP, Ramu P, Jannink J-L (2016) Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. Plant Genome 9. doi:10.3835/plantgenome2015.11.0118

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD et al (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS One 4(12):e8451. doi:10.1371/journal.pone.0008451

Chapter 9
Status and Perspectives of Genomic Selection in Forest Tree Breeding

Dario Grattapaglia

9.1 Introduction: From Trait Dissection to Genomic Prediction in Forest Trees

Advanced tree breeding involves a large number of steps around the basic concept of increasing the frequency of favorable alleles for a number of traits simultaneously in the target population. Recurrent cycles of selection, mating, and testing are used to develop genetically improved seeds or elite clonal stocks by maximizing genetic gain per unit time at the lowest possible cost (Namkoong et al. 1988; White et al. 2007). Trees have long life cycles and become reproductively active only after several years. The progress and success of tree breeding programs are therefore strongly dependent on the time needed to complete a breeding generation. This may last several years to decades depending on the biology of the species, the age at which phenotypes can be accurately measured, and the deployment plan of improved material, whether seeds or clones. Additionally, the uncertainties associated with conducting decade-long breeding programs can be high. Breeding investments are made several years before the eventual utilization of genetically improved material, making it susceptible to changes in the economic objectives of the forest products, market demands, and management policies.

The time challenges faced by tree breeders have historically led to substantial efforts to understand juvenile-mature correlations for late-expressing traits (Namkoong et al. 1988), devise ways to accelerate recombination by artificial flower induction (Greenwood et al. 1991; Hasan and Reid 1995), and practice early selection on juvenile traits (Williams 1988). In the early 1990s, when DNA marker technologies became more accessible, marker-assisted selection (MAS) was

D. Grattapaglia (✉)
EMBRAPA Genetic Resources and Biotechnology – EPqB, 70770-910 Brasilia, DF, Brazil

Universidade Católica de Brasília- SGAN, 916 modulo B, Brasília, DF 70790-160, Brazil
e-mail: dario.grattapaglia@embrapa.br

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7_9

immediately seen as a powerful tool to overcome some of the challenges. MAS could shorten breeding cycles by early selection for late-expressing traits such as wood properties. Furthermore, MAS could also be applied to increase selection intensity, reduce the effort of field-testing, and possibly improve selection precision for low-heritability traits such as volume growth (Grattapaglia et al. 1992; Neale and Williams 1991; Williams and Neale 1992). Nevertheless, the potential of MAS for forest trees was immediately questioned based on the state of linkage equilibrium that the typically large, random mating population with recent domestication history would be found and the concerns regarding stability of QTLs across the highly variable genetic backgrounds of breeding populations and environments (Strauss et al. 1992).

Despite those early, well-grounded arguments, linkage mapping and QTL detection experiments in species of pines, spruces, poplars, and eucalypts were carried out based on the implicit assumption that it would be possible to map and estimate the effects of all the relevant genes for traits such as growth and wood quality during the life of the tree, in every population and environment. A considerable number of studies describing QTLs and gene-trait associations in forest trees were reported (reviewed in (Grattapaglia et al. 2009, 2012; Harfouche et al. 2012; Neale and Kremer 2011). Mirroring what was the canonical approach to QTL mapping in the major crops and model systems, QTL mapping in forest trees was carried out using single biparental populations of relatively limited size. The difference was that two-generation pedigrees were used because any cross between heterozygous parents would provide a segregating F1 population under a pseudo-testcross (Grattapaglia and Sederoff 1994). Several "major effect" QTLs were mapped in early studies. However, later multifamily experiments conducted with larger populations revealed many more QTLs with smaller effects and largely inconsistent across backgrounds and environments (Dillen et al. 2008; Freeman et al. 2013; Gion et al. 2011; Novaes et al. 2009; Rae et al. 2008; Thumma et al. 2010). With the exception of a few QTLs of moderately large effect mapped for disease resistance (Junghans et al. 2003; Stirling et al. 2001; Wilcox et al. 1996), and candidate-gene associations for phenological traits (Ingvarsson et al. 2008), results generally showed that QTL and association mapping do not explain sufficient genetic variation to lead to any effective implementation of MAS for complex traits in forest trees. On hindsight, it is perplexing to consider how far removed from the reality of forest tree breeding were those biparental mapping populations and the QTL mapping data derived from them.

The ineffectiveness in dissecting complex traits, i.e., determining the position, variation, and magnitudes of allelic effects at QTLs underpinning quantitative traits and the consequent failure to implement MAS, has not been exclusive to forest trees. With the exception of a few simple qualitative or monogenic traits in crops (Bernardo 2008), major genes in fruit trees (Arus et al. 2012), and recessive genetic defects in domestic animals (Charlier et al. 2008), this has been the general conclusion for the vast majority of species undergoing selective breeding. This fact has caused a major paradigm shift in animal and plant molecular breeding in the last 10 years. The field has now moved from trying to a priori discover, validate, and use marker-trait associations to dealing with the aggregate of the whole-genome effect, much like quantitative genetics always did. This revolution was only possible following the development of large numbers of markers, mostly SNPs, together with cost-effective platforms to query them genome-wide and new statistical methods to deal with large datasets accounting for the large numbers of markers (p) and relatively small number of individuals (n) problem. The approach termed "genomic selection" (GS) estimates all marker effects simultaneously, retaining all of them as predictors of performance and precluding the prior search for significant marker-trait associations but focusing exclusively on prediction efficiency (Goddard and Hayes 2009; Meuwissen et al. 2001).

Genomic selection has become a theme of considerable interest in the tree genetics and breeding community worldwide in the last few years since the first perspectives based on simulations (Grattapaglia and Resende 2011; Iwata et al. 2011) and experimental results (Resende et al. 2012a, b) were reported. In this chapter, an update is provided of an earlier review on this topic (Grattapaglia 2014). However, a more comprehensive discussion of the main factors (theoretical and practical) relevant to GS in tree breeding that has emerged from experimental studies in the last few years is provided. This discussion is preceded by a concise explanation of the basic insights of GS and its perspectives and challenges in tree breeding. An updated compilation of all published experimental GS studies in forest trees follows, highlighting their main contributions to our current understanding of this new approach for tree breeding. The conclusion finally summarizes the main lessons learned so far in an attempt to provide a nine-point tentative roadmap for implementing GS in a tree breeding program.

9.2 Genomic Selection: Reviewing the Basic Principles

Nejati-Javaremi et al. (1997) were probably the first ones to show that "total allelic" relationship estimated from marker data would be a powerful alternative to the pedigree-derived additive genetic relationship to derive best linear unbiased prediction estimates (EBV) of breeding values using mixed model equations. Haley and Visscher (1998) proposed the idea of "total genomic selection," that is, that by genotyping at the genome-wide scale with sufficient marker density and low cost, it would be possible to assure that markers will be in complete association with any trait locus and, therefore, capture the most genomic effects underlying complex traits. However, it was the groundbreaking paper by Meuwissen et al. (2001) that anticipated that selection on genetic values predicted from markers could considerably increase the rate of genetic gain in animal and plant breeding programs. They also outlined the statistical approach and potential caveats to estimate the genetic value of unphenotyped individuals based exclusively on phenotype and genotype data of a reference ancestral population using "genomic selection" (GS), a term surprisingly not used in the main text but only in the running title of that paper.

Both MAS and GS start by establishing associations between discrete marker genotypes and continuously distributed phenotypes in relevant populations. However, they are fundamentally different following this initial step. MAS typically targets the discovery of marker-trait associations in one or a few biparental populations or association mapping panels using rigorous significance tests, with the later goal of using such marker-trait associations for selection. GS instead uses a dense genome-wide panel of markers whose effects on the phenotype are estimated simultaneously in a large and representative "training" population of individuals without applying rigorous significance tests. All or a subset of markers are retained as forecasters of phenotypes in prediction models to be later applied to "selection candidates" for which only genotypes are collected. Thus, in GS a marker effect does not need to exceed a stringent significance threshold to be used in the subsequent breeding phase, and the effects of the marker alleles are estimated in a much larger and more representative population rather than within one or a few mapping families. The "training" population involves at least several hundreds to a few thousand individuals representative of the target breeding population, which are genotyped for the marker panel and phenotyped for all traits of interest. The prediction models developed for each trait are cross-validated in a "validation" population, a randomly sampled subset of individuals of the same reference population that did not participate in the estimation of marker effects. Once a prediction model is shown to provide satisfactory accuracy, i.e., correlation between the observed and predicted breeding values following cross-validation, it can be used in the breeding phase to calculate the genomic estimated breeding values (GEBV) or total genomic estimated genotypic values (GEGV) (when nonadditive effects are also included in the model) of the selection candidates. Put simply, a GEBV is calculated by multiplying the genotypes at all markers by their effect estimated by, for example, random regression best linear unbiased prediction (RR-BLUP) or any other statistical method that adequately avoids model over-fitting by marker-specific shrinkage of regression coefficients (Crossa et al. 2010; Lorenz et al. 2011). There is also a second approach to use genotypic data in GS. Marker genotypes are used to estimate a genomic relationship matrix between individuals with genotypes and phenotypes of the training population and the yet-to-be phenotyped selection candidates for which only genotypes are available. This genomic relationship matrix can then be used to estimate a variance/covariance matrix between the genetic values in a mixed model generally called G-BLUP that stands for genomic BLUP. It has been shown that RR-BLUP and G-BLUP are statistically equivalent under theoretical conditions that are generally met in practice (Habier et al. 2007).

GS exploits both the linkage disequilibrium (LD) between the dense marker data and all QTL effects associated and the genetic relationship between the training population and the prospective selection candidates. By avoiding prior marker selection and estimating marker effects in a large and representative population, GS potentially captures all genetic variance for the trait explained by the large numbers of small effects that QTL or association genetics-based MAS does not capture. Genomic selection also known as genome-wide selection (GWS) has now become the paradigm for marker-assisted selection (MAS) of complex traits in plants and animals. GS is the standard molecular breeding technology in dairy cattle and increasingly been adopted in other animal species such as swine and broiler (Van Eenennaam et al. 2014). The most extraordinary example continues to be in dairy cattle where progeny testing of young bulls has been replaced by GS, resulting in rapid improvements across multiple traits. By 2011 over 40% of the market share of tested bulls across several countries was composed of bulls without milking daughters, exclusively selected based on GS (Pryce and Daetwyler 2012).

While GS was rapidly being adopted in animal breeding, it also became a topic of interest in plants, starting with the influential papers by Bernardo (2008) and Bernardo and Yu (2007), soon followed by others that discussed the potential of genomic prediction in crops (Heffner et al. 2009; Jannink et al. 2009) and forest tree improvement (Grattapaglia et al. 2009). An exponential growth of published studies about GS in all major cultivated plant species has taken place in the last 5 years. Following the early enthusiasm and prospects fueled by several simulation-based studies validated by experimental results, we have now moved to a phase where several detailed and careful considerations are necessary (Heslot et al. 2015; Jonas and de Koning 2015). These include the strategic breeding and tactical logistics and resource allocation aspects of implementing GS, the issues related to the optimal planning of training populations and phenotyping efforts associated to them, the marker platforms to be used, and a thorough cost-benefit analysis of the entire process.

9.3 Perspectives of Genomic Selection in Tree Breeding

The objective of selective breeding is to accelerate the rate of genetic improvement or selection response per unit time. As noted above, the time factor is extremely relevant to tree breeding due to the long generation times typically necessary to complete a full breeding cycle. To go back to the classic breeder's equation is therefore useful to understand how GS can have a tremendous impact on the rate of genetic gain. In the equation ($\Delta G = ir\sigma_A/L$), i is the selection intensity (the proportion of trees that are selected to become parents of the next generation); r is the accuracy of selection, i.e., the correlation between the estimated breeding value (EBV) and the true breeding value; σ_A is the additive genetic standard deviation of the trait of interest, i.e., the genetic variation available in the population for selection; and L is the generation interval or time needed to achieve the genetic gain. GS can directly increase the rate of genetic gain of a tree breeding program by increasing the selection intensity (i), because many more young seedlings can be genotyped and their phenotypes predicted by GS than the number of seedlings typically planted and managed in field trials. This is particularly relevant for traits expensive to measure or expressed late in the life of the tree. However, the largest impact of GS on the rate of genetic gain will result from radically reducing the generation interval (L). Phenotypes of the selection candidates can be predicted at ultra-early ages, for example, when the seedlings are a few weeks old, still in the nursery, instead of waiting half of the breeding cycle (usually 4–20 years or more, depending on the species) before having access to their phenotypes, especially those expressed late in the life of the tree. The accuracy of selection (r) evidently is also a main driver of the genetic gain. In standard breeding this accuracy is provided by the square root of the heritability, i.e., the proportion of the phenotypic variance explained by genetic components. In GS the accuracy of selection is estimated by the correlation between the genomic estimated breeding value (GEBV) and the true breeding value. When the breeding objective is to select individuals to be deployed as clones, this correlation needs to involve not only the additive effects but also the nonadditive component, such that r measures the correlation between the genomic estimated genotypic value (GEGV) and the true genotypic value. Finally, without genetic variation (σ_A) for the target trait, no progress will happen.

The implementation of a genomic selection program for tree breeding encompasses essentially two stages (Fig. 9.1). The first one involves the definition of a "training population" of individuals that are genotyped and phenotyped to develop and cross-validate predictive models to be later used in the second stage, where GS is actually put in practice. A training population is usually sampled from an existing progeny trial derived from inter-mating a group of elite parents that were established as the population to undergo breeding for the subsequent generations. Usually this group of elite parents will have an effective population size (N_e) between 30 and 100 and a census number (N) that will be in that same range or slightly larger, taking into account any cryptic relatedness that exists between the individuals. The training population will have at least 1,000–2,000 individuals. However, the more individuals are genotyped and phenotyped, the better will the marker effects be estimated and more robust will become the predictive model.

In the second stage, GS will be effectively employed on the selection candidates, typically an array of full of half-sib families derived from intercrossing either the original elite parents or elite individuals selected in the progeny trial used as training population. These selection candidates are genotyped and have their breeding values (GEBV) and/or genotypic values (GEGV) estimated using the predictive model developed earlier. Top ranked seedlings for GEBV are subject to early flower induction and inter-mated to create the next generation of breeding. Top ranked seedlings for GEGV are clonally propagated and tested in verification clonal trials where elite clones are eventually selected for operational plantation. Additionally, a random subset of the already genotyped selection candidates could be planted in experimental design and phenotyped at the target age to provide genotype and trait data for GS model updating as generations of GS advance, mitigating the erosion of marker-QTL LD and decay of relationships, and maintaining accuracy of GS predictions over generations.

Simulation-based and experimental reports outlined the promising prospects of GS to increase the efficiency of tree breeding programs (see below). In eucalypts and poplars, GS not only could eliminate the progeny trial but would also reduce the time and costs involved in the clonal testing phase by reducing the number of selected trees that are evaluated as clones in a preliminary, typically large-scale,

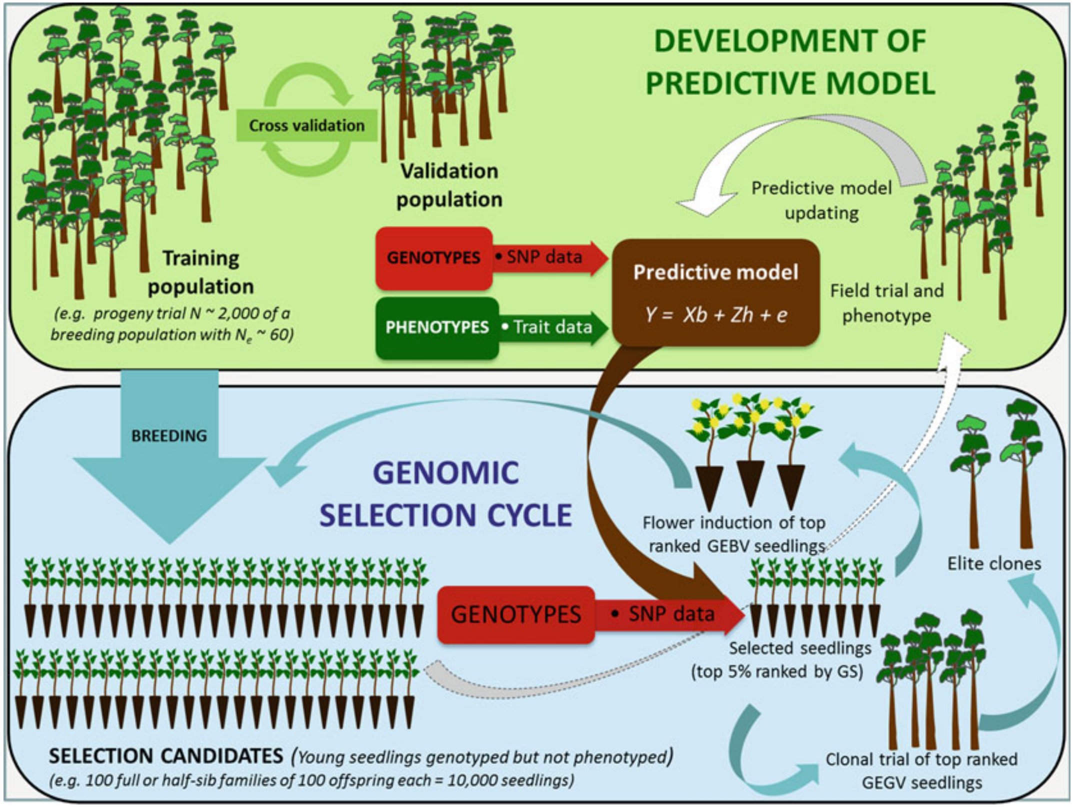

DEVELOPMENT OF PREDICTIVE MODEL
Training population
(e.g. progeny trial N ~ 2,000 of a breeding population with N_e ~ 60)
Cross validation
Validation population
GENOTYPES
• SNP data
PHENOTYPES
• Trait data
Predictive model
Y = Xb + Zh + e
Predictive model updating
Field trial and phenotype
BREEDING
GENOMIC SELECTION CYCLE
Flower induction of top ranked GEBV seedlings
Elite clones
GENOTYPES
• SNP data
Selected seedlings (top 5% ranked by GS)
Clonal trial of top ranked GEGV seedlings
SELECTION CANDIDATES (Young seedlings genotyped but not phenotyped)
(e.g. 100 full or half-sib families of 100 offspring each = 10,000 seedlings)

Fig. 9.1 The two stages of the development and implementation of a genomic selection program in tree breeding (Modified from Grattapaglia (2014))

clonal trial. In conifers, as pointed out by Resende et al. (2012b), GS combined to somatic embryogenesis (SE) could considerably boost the efficiency of current clonal propagation protocols by allowing preselection of zygotic embryos based on their GEBV and their immediate expansion into elite SE lines for the establishment of clonal trials or directly into commercial plantations. Besides the time gain, a less mentioned advantage of GS is related to the possibility of efficiently carrying out selection for several traits simultaneously in large numbers of individuals. It is virtually impossible for any breeding program to complete a rigorous assessment of all traits of interest in all trees in a progeny trial. Traits usually include wood volume, stem taper and straightness, physical and chemical wood properties, sprouting and rooting abilities, nutritional efficiency, and tolerance to pests, diseases, drought, and frost. In tropical eucalypts, for example, even in clonal trials, this is typically accomplished only in the very final stages of selection and for a very limited number of clones (20–50) that had been preselected for volume growth and wood density (Rezende et al. 2014). In traditional breeding, a sequential approach is typically used that combines different forms of selection indices and independent culling levels for estimating the ultimate value of candidates. In GS, because breeding values are predicted for each trait separately (i.e., a separate GEBV for each measured trait), selection indices can be used to combine data from all the traits under analysis into a single value for each candidate. The validity of this multiple-step approach rests on a property that the BLUP of any linear combination of traits is equal to that linear combination of the BLUP predicted values of the individual traits (White et al. 2007). Therefore, the net effect of GS would be a remarkable increase in selection intensity at the seeding stage for all traits simultaneously, considerably improving the overall efficiency of the breeding program. Additionally, applying GS to multiple traits could significantly increase the prediction accuracy for a low-heritability trait or for traits with a limited number of phenotypic records when a correlated high-heritability trait is available (Jia and Jannink 2012).

9.4 Genomic Selection: Experimental Results in Forest Trees

A compilation of all experimental GS studies in forest trees published to date is provided in a format that allows a quick perusal of their key attributes and performance of predictive abilities for different traits (Table 9.1). Reports of GS in forest trees have been unique in that they used considerably larger training population sizes and numbers of markers when compared to GS studies in crop plants. Experiments have typically mirrored the structure of true breeding populations and adopted designs that accounted at satisfaction with the theoretical expectations of higher diversity and the necessary relationship between training and validation sets. Another distinctive aspect has been the attempt to evaluate the

Table 9.1 Summary of the main features and results of the published experimental studies of genomic selection in forest trees

Species	Population structure	Population size	# and type of markers used	Trait (h^2)[a]	Predictive ability[b]	Reference
Eucalypts (*Eucalyptus grandis x E. urophylla hybrids*)	Progeny trial of 43 full-sib families from 11 elite hybrid parents	738	3,129 DArT fixed array	DBH (0.53)	0.54	Resende et al. (2012a)
				HG (0.42)	0.51	
				WSG (0.59)	0.60	
				PY (0.38)	0.54	
Eucalypts (*E. grandis, E. urophylla, E.* globulus, and their F1 hybrids)	Progeny trial of 232 full-sib families from 51 elite parents	920	3,564 DArT fixed array	DBH (0.56)	0.55	
				HG (0.48)	0.46	
				WSG (0.42)	0.42	
				PY (0.47)	0.38	
Loblolly pine *Pinus taeda*	Progeny trial of 61 full-sib families from 32 parents replicated in four environments; trees were clonally replicated	800	4,852 SNP Infinium chip	DBH (0.21–0.32[c])	0.31–0.37[c]	Resende et al. (2012b)
				HG (0.13–0.26[a])	0.26–0.34[c]	
Loblolly pine *Pinus taeda*	Progeny trial of 61 full-sib families from 32 elite parents; trees were clonally replicated	951	4,853 SNP Infinium chip	GR (0.22–0.35)	0.38–0.49	Resende et al. (2012c)
				DVP (0.07–0.45)	0.24–0.51	
				FRR (0.12–0.21)	0.23–0.34	
				WS (0.37)	0.39–0.43	
				LC (0.11)	0.17	
				LTW (0.17)	0.23–0.24	
				WSG (0.09)	0.20–0.22	
				SC (0.14)	0.25–0.26	
Loblolly pine *Pinus taeda*	13 full-sib families related by common parents	149	3,406 SNP Infinium chip	LC (0.76)[d]	0.66–0.76	Zapata-Valenzuela et al. (2012)
				CC (0.69)	0.61–0.83	
				HG (0.95)	0.47–0.52	
				VG (0.94)	0.30–0.56	
Loblolly pine *Pinus taeda*	13 full-sib families related by common parents	165	3,461 SNP Infinium chip	HG	0.37–0.74[e]	Zapata-Valenzuela et al. (2013)
Loblolly pine *Pinus taeda*	Progeny trial of 61 full-sib families from 32 elite parents	951	4,853 SNP Infinium chip	HG (0.32)	0.66–0.86[f]	Munoz et al. (2014)

(continued)

Table 9.1 (continued)

Species	Population structure	Population size	# and type of markers used	Trait (h^2)[a]	Predictive ability[b]	Reference
White spruce *Picea glauca*	214 open-pollinated families from 43 natural populations	1,694	6,385 SNPs Infinium chip	CP (0.24)	0.39; 0.20; 0.09[g]	(Beaulieu et al. 2014a)
				FC (0.33)	0.39; 0.18; 0.13	
				CW (0.57)	0.33; 0.17; 0.0	
				WSG (0.39)	0.37; 0.13; 0.07	
				MA (0.38)	0.38; 0.16; 0.0	
				WS (0.31)	0.38; 0.18; 0.03	
				RW (0.04)	0.33; 0.13; 0.0	
				SFS (0.48)	0.39; 0.16; 0.12	
				CRD (0.44)	0.44; 0.28; 0.13	
				CTD (0.26)	0.35; 0.15; 0.11	
				CWT (0.39)	0.41; 0.13; 0.13	
				HG (0.25)	0.36; 0.18; 0.09	

White spruce *Picea glauca*	Two unrelated breeding groups with 851 and 897 individuals in 27 and 32 full-sib families, respectively, in two environments	1,748	6,932 SNP Infinium chip	WSG (0.33–0.43)^{h}	0.79–0.06^{i} 0.75–0.66^{j} 0.59–0.10^{k} 0.63–0.53^{l}	Beaulieu et al. (2014b)
				MA (0.30–0.32)	0.77–0.03 0.71–0.61 0.36–0.04 0.52–0.41	
				HG (0.39–0.57)	0.58–0.0 0.68–0.34 0.33–0.0 0.55–0.22	
				DBH (0.32–0.39)	0.52–0.0 0.69–0.24 0.29–0.0 0.48–0.15	
Eucalypts (*Eucalyptus grandis x E. urophylla hybrids*)	Progeny trial of 45 full-sib families from 46 elite parents	1,000	29,090 SNPs from the EuCHIP60K fixed array	DBH (0.64)	0.44	Lima (2014)
				HG (0.54)	0.34	
				VG (0.63)	0.42	
				MAI (0.63)	0.42	
				CC (0.74)	0.52	
				HC (0.81)	0.55	
				SGR (0.93)	0.83	
				ILC (0.82)	0.64	
				SLC (0.84)	0.72	
				LC (0.84)	0.63	
				WSG (0.76)	0.63	
				MA (0.26)	0.17	
				FL (0.80)	0.49	
				FW (0.22)	0.14	
				FC (0.49)	0.33	

(continued)

Table 9.1 (continued)

Species	Population structure	Population size	# and type of markers used	Trait (h^2)[a]	Predictive ability[b]	Reference
Interior spruce *Picea glauca x Picea engelmannii*	25 open-pollinated families in three environments	1,126	8,868 to 62,198 SNPs by GbS with different imputation methods to account for 30% or 60% missing data	HG (0.43-0.98)[h]	0.17–0.63[m]	El-Dien et al. (2015)
				DBH(0.28-0.55)	0.01–0.77	
				VG (0.29-0.76)	0.01–0.73	
				WDV(0.31-0.78)	0.17–0.67	
				WDR(0.42-0.65)	0.14–0.64	
				WDX(0.48-0.59)	0.23–0.62	
				MOD(0.31-0.78)	0.16–0.67	
Interior spruce *Picea glauca x Picea engelmannii*	25 open-pollinated families in two environments	769	34,570 to 50,803 SNP by GbS depending imputation method	HG (0.43-0.98)	0.04–0.47	Ratcliffe et al. (2015)
Maritime pine *Pinus pinaster*	184 founders (G0) and 477 progeny individuals (G1) in 191 maternal half-sib families	661	2,500 SNPs Infinium chip	DBH (0.17)	0.38–0.47	Isik et al. (2016)
				HG (0.30)	0.45–0.47	
				SS (0.25)	0.48–0.55	

Maritime pine *Pinus pinaster*	46 founders (G0), 62 (G1) and 710 progeny individuals (G2) in 35 maternal half-sib families but with full pedigree inferred by SNP data	818	4,332 SNPs Infinium chip	DBH (0.17)	0.52–0.74	Bartholome et al. (2016)
				HG (0.30)	0.58–0.69	
				SS (0.25)	0.65–0.82	

[a] Trait heritability estimated using pedigree data; *trait legend*: *HG* height growth, *DBH* diameter at breast height, *WSG* wood-specific gravity, *PY* pulp yield, *LC* lignin content, *CC* cellulose content, *GR* growth (several traits), *DVP* developmental traits, *FRR* fusiform rust resistance, *WS* wood stiffness, *LTW* latewood %, *SC* sugar content, *VG* volume growth, *CP* cell population, *FC* fiber coarseness, *CW* crystallite width, *MA* microfibril angle, *RW* ring width, *SFS* specific fiber surface, *CRD* cell radial diameter, *CTD* cell tangential diameter, *CWT* cell wall thickness, *SS* stem sweep, *WDV* wood density by acoustic velocity, *WDR* wood density by resistance to drilling, *WDX* wood density by X-ray densitometry, *MOD* modulus of elasticity, *HC* hemicellulose content, *SGR* syringyl/guaiacyl lignin ratio, *ILC* insoluble lignin content, *SLC* soluble lignin content, *FL* fiber length, *FW* fiber width

[b] Predictive abilities reported correspond to the correlation between observed and expected breeding values; *GRM* genomic relationship matrix, *PA* predictive ability

[c] Heritability or PA variation across the four environments evaluated

[d] Heritabilities were reported as clone mean repeatabilities

[e] Variation in PA depending on the cross-validation method used

[f] PAs of the additive model and additive*dominance model

[g] PAs estimated by three different cross-validation schemes with decreasing relatedness between training and validation (see text for further details)

[h] Heritabilities across the different environments tested

[i] PAs from within-families cross-validation within and between the two breeding groups

[j] PAs from within-families cross-validation within and between the two environments

[k] PAs from between-families cross-validation within and between the two breeding groups

[l] PAs from between-families cross-validation within and between the two environments

[m] PAs reported are the lowest estimates by cross-validation across environments and the highest multisite using all individuals for training and validation

impact of different issues particularly relevant to tree breeding on the accuracy of prediction. These included the level of relationship between training and validation sets, the effect of genotype by environment (G*E), the influence of age-age correlations, and the performance of different analytical approaches that use variable underlying assumptions of trait architecture. Nevertheless, all studies until recently were not able to evaluate the actual performance of GS across generations, i.e., using training data of an ancestral generation to predict and validate phenotypes of progeny individuals. Cross-validation and estimation of predictive accuracy was carried out exclusively within the same generation. Recently, however, studies with *Pinus pinaster* that had access to three generations (G0, G1, and G2) showed encouraging results of intergeneration prediction both by mixing parents and progeny in the same training set (Isik et al. 2016) and later using only G0 and G1 individuals to predict in the G2 generation (Bartholome et al. 2016). Reported prediction accuracies of experimental studies have been generally very satisfactory, in line with the expectations from previous simulations (Grattapaglia and Resende 2011; Iwata et al. 2011) and results in crop plants and domestic animals.

Two experimental studies pioneered the field of genomic prediction in forest trees. A report in *Eucalyptus* involving two independent genetically unrelated breeding populations with contrasting effective population sizes assessed in completely different environments (Grattapaglia et al. 2011b; Resende et al. 2012a) and a second one involving a cloned set of loblolly pine full-sib families assessed across four different environments and two different ages (Resende et al. 2012b). Predictive abilities between 0.26 and 0.60, with an overall average of 0.44, were estimated by cross-validation for a range of growth and wood quality traits. These results approximated well to the accuracies predicted from deterministic (Grattapaglia and Resende 2011) and stochastic simulations (Iwata et al. 2011) for similar parameters of trait heritability, effective population size, and genotyping density. These experimental results suggested that potential gains of 50–200% in selection efficiency predicted by simulations could be achieved, if adequate prediction abilities would be kept across generations. These studies also showed that prediction accuracies strongly depend on the existence of genetic relationship between training and validation sets and are impacted by G*E and age-age correlations such that predictions will be effective when carried out in the same environment and same age as where and when the training data was collected. Studies in white spruce soon followed where the impact of G*E and the key significance of having genetic relationships between training and validation were thoroughly evaluated and corroborated (Beaulieu et al. 2014a, b). Recently published reports in eucalypts (Lima 2014), spruce (El-Dien et al. 2015; Ratcliffe et al. 2015), and maritime pine (Bartholome et al. 2016; Isik et al. 2016) provided additional promising results on the ability to predict complex traits in forest trees and confirmed what had been previously observed as far as the impact of relationship, G*E, and age. The main results and contributions of all these studies are detailed below, while discussing the main factors that affect the prospects of GS to forest tree breeding.

9.5 Factors Affecting the Success of Genomic Selection in Tree Breeding

The success of GS is dependent on a number of factors including both the fundamental aspects predicted from population and quantitative genetics theory and the more practical and logistics aspects of resource allocation and cost-benefit analysis. The following discussion tries to cover the main factors in these two realms reminding that they are in many ways interconnected and interdependent. The accuracy of a genomic prediction model, i.e., the correlation between the genomic estimated breeding value (GEBV) and the true breeding value, is undoubtedly the key factor that will have the major impact on the success of GS. Four fundamental factors from the theory of population and quantitative genetics are known to affect the accuracy of genomic prediction: (1) the effective population size (N_e) and genotyping density that in turn determines the extent of LD between markers and QTLs; (2) the size and composition of the training population, i.e., the number of individuals with phenotypes and genotypes from which the marker effects are estimated; (3) the heritability of the trait in question; and (4) the genetic architecture of the target trait, i.e., the distribution of QTL effects (number of loci and size effects) (Hayes et al. 2009a; Lin et al. 2014). An assessment of the impact of each one individually in the context of tree breeding was reported early on, providing some broadly useful guidelines for GS regardless of the target species, recombinant genome size, or breeding cycle length (Grattapaglia and Resende 2011).

When considering the practicalities of tree breeding, some of the most relevant factors that impact the prospects of GS include (1) the size, composition, and phenotyping effort devoted to the training population; (2) the genotyping platform employed and the resulting data quality, cost, turn-around time, and breeder's friendliness; (3) the extent of genotype by environment interaction; and (4) the long-term performance of genomic prediction models, including the need for model retraining and the potential effect on loss of diversity and increased inbreeding. With all these issues considered, an attempt is made here to answer a common question posed by tree breeders: what are the main issues that one should be aware of when considering the investment in a GS program?

9.5.1 Effective Population Size of the Tree Breeding Population

The main issue generally considered to determine the accuracy of GS is the extent of linkage disequilibrium, i.e., the nonrandom association between marker alleles and QTL alleles. This factor in turn directly depends on the effective population size (N_e) and genotyping density. The effective population size corresponds to the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration (Wright 1931). As the effective population size gets smaller, the effect of genetic drift gets stronger, and more LD is generated because it is unlikely that combinations of marker alleles and QTL alleles get sampled at a frequency that corresponds to the product of their individual frequencies. The resulting nonrandom association between alleles at marker loci and QTLs allows marker alleles to predict the allelic state of nearby QTL and thus to predict phenotypes. At equilibrium, the LD generated by random drift is balanced by recombination that takes place as breeding generations advance, causing it to dissipate, such that closer loci are expected to be in higher LD than more distant ones. Consequently, the relationship between N_e and LD affects the marker density needed to achieve and sustain adequate prediction accuracy of a GS model across generations. In other words, marker density needs to scale with the effective population size, and the level of LD between markers and QTL can be increased by reducing N_e (Grattapaglia 2014).

The discussion of the extent of LD in forest trees takes us back for a moment to the original criticism about the prospects of MAS in forest trees seen in those early days (Strauss et al. 1992). Given the preferentially outbred nature and large N_e of natural populations of trees, the claim was that the state of linkage equilibrium would therefore be such that prohibitively large training populations and marker density would be required to attain success of MAS. That original prediction was correct in that it would apply to very conservative breeding programs that aim not only at genetic gain but also at preserving diversity by managing breeding populations with very large $N_e > 300$. However, the reality of more advanced breeding programs where genetic gain is prioritized and alternative strategies are devised to maintain diversity (White et al. 2007) is not one of very large effective population sizes. Rather, small effective population sizes in the range of $N_e = 10$–100 are used to maximize gain. On a genome-wide basis, genetic drift is the main contributor to LD, and drift is generated by the breeder when a closed, selected breeding population is established. The effective population size influences the number of independently segregating chromosome segments expected in the population (M_e) which in turn will determine the necessary genotyping density to capture all the effects of the QTLs co-segregating with those segments. To understand this relationship, a common derivation proposed for M_e in populations is $M_e = 2N_eL$ (Hayes et al. 2009b) where L is the genome size in Morgans. Larger N_e and recombinationally larger genomes result in more independently segregating chromosome segments requiring more markers. On the other hand, the smaller the N_e, the closer the genetic relationship among individuals, the longer the independent chromosome segments, the smaller becomes M_e, and less markers are needed to reach a certain accuracy of GS. It is important to note that the key parameter here is not the physical genome length but rather the recombination size and number of chromosomes. The fact that conifer genomes are very large (~20–23 Gbp) does not matter here. Their recombination size in Morgans, ~15 M in *Pinus taeda* (Echt et al. 2011) to ~18 M in *Picea* (Pelgas et al. 2005) for 12 chromosomes, is not that different from the recombination size of *Eucalyptus*, ~13 M in 11 chromosomes (Brondani et al. 2006) with a much smaller physical genome (~0.65 Gb).

Calibrating the extent of LD by managing the effective population size, such that near-maximum genetic gain can be achieved in a long-term breeding program, is thus a key element when adopting GS. Theoretical studies and practical considerations regarding the appropriate size of a tree breeding population have shown that N_e between 20 and 50 will support selection with appreciable genetic gains for several generations (Namkoong et al. 1988; White et al. 2007). Although suitable for short-term genetic gains, such constrained N_e may be subject to larger deviations of actual versus predicted progress and may result in a faster buildup of relatedness. To remain on a conservative side in sustaining long-term gains, effective population sizes between $N_e = 40$ and 100 have been used, typically corresponding to a census number (i.e., the total number of selections retained in the breeding population in any given generation) around 200 individuals with some level of relatedness (White et al. 2007). As examples, the third breeding cycle of loblolly pine in the Southeastern USA has adopted a highly selected group of 40 selections to provide rapid gains (McKeand and Bridgwater 1998). In *Eucalyptus*, populations with N_e between 30 and 60 are typically used for each species in reciprocal recurrent selection strategies for hybrid breeding. Similar effective population sizes are also used in recurrent selection programs based on synthetic hybrid populations, an approach that exploits the variation derived from multiple species aiming at the selection of elite hybrid clones for deployment (Assis and de Resende 2011; Kerr et al. 2004). In conclusion, the effective population sizes currently used in most tree breeding programs largely fit within the perspectives of reaching high GS accuracies, provided that sufficient genotyping densities are used, so that the number of independently segregating chromosome segments is adequately tracked.

9.5.2 Genotyping Density and SNP Platforms for GS in Forest Trees

"Recent advances in molecular genetic techniques will make dense marker maps available and genotyping many individuals for these markers feasible." This far-seeing statement that introduces the seminal paper on genomic selection (Meuwissen et al. 2001) was written when SNP discovery by Sanger sequencing was still a prohibitively expensive endeavor for most species and SNP genotyping platforms were in their infancy. However, it clearly recognized the key role that the advent of faster and cheaper DNA marker genotyping would have for the new breeding method proposed in that article. In the last 15 years, a major revolution took place in the ability to discover large numbers of SNPs and develop new methods to assay DNA polymorphisms, starting in 2005 with the advent of next-generation sequencing technologies based on miniaturized and parallelized platforms. What makes genomic selection different from what breeders have done so far using the tools of quantitative genetics is the adoption of dense DNA marker data instead of relying solely on the expected pedigree relationships. Marker data allows one to build a genomic relationships matrix that precisely determines the realized kinship among individuals in a breeding population. This procedure not only allows correcting pedigree errors but, more importantly, captures the random Mendelian sampling term resulting from gamete formation, such that the realized genetic covariances are now based on the actual proportion of the genome that is IBD or IBS between any two individuals (VanRaden 2008). Genomic selection based on realized genomic relationships can produce more accurate predictions than the pedigree-based method precisely because it exploits the variation created by Mendelian segregation. It is therefore relevant to devote some time to discuss the advantages and limitations of the different marker technologies currently available for the application of GS.

Prior to the times of easy SNP discovery and genotyping, microsatellite markers were the workhorse of genetic analysis in forest trees, and they still are for many applications. Microsatellite genotyping has been adopted into breeding practice to resolve clonal identity, verify parentage, and reconstruct pedigrees as a way to reduce costs of controlled crosses (El-Kassaby and Lstiburek 2009; Grattapaglia et al. 2004; Lambeth et al. 2001). Usually between 10 and 20 microsatellites have been used providing abundant power to resolve parentage even with relatedness between alleged parents. In a recent study, it was shown that some 100 selected high-frequency SNPs are needed to match the power of 16 microsatellites for such applications in eucalypts (Telfer et al. 2015). However, typical microsatellite marker density is not sufficient to estimate the genome fraction shared by two individuals and to apply this information to genomic predictions. The genotyping density together with the effective population size showed by far the largest impact on the prospects of GS in forest tree breeding (Grattapaglia and Resende 2011). The upper bound benchmark accuracy of phenotypic BLUP selection, set at 0.68 in that study, can be reached at a relatively low marker density, around 2–3 markers/cM, as long as the effective population size is kept below $N_e = 60$. For an average genome of 1,500–2,000 cM, some ~5,000 SNPs would be necessary. For larger effective population sizes up to $N_e = 100$, however, 10 or up to 20 markers/cM would be necessary for keeping high accuracies of GS. Such a target genotyping density will require genotyping platforms to yield somewhere between 20,000 and 50,000 informative markers depending on the size of the recombining genome and the effective population size of the breeding population.

The impact of the genotyping density used in the practice of GS will become even more important as generations of selection advance. In the absence of selection, increasing marker density is beneficial to the persistence of GEBV prediction accuracy over generations (Solberg et al. 2009) because higher marker densities enable GEBV accuracy to persist over time due to a slower decay of LD among tightly linked marker and trait loci. However, directional selection following the initial training population is expected to result in a rapid decline of accuracy (Muir 2007). High-density genotyping was shown to be essential to sustain accuracy and keep selection effective for more generations in the presence of directional selection when a finite number of QTL loci are assumed rather than an infinitesimal model (Long et al. 2011). In such cases, selection, together with recombination, may change the pattern of LD between markers and QTLs. The new LD generated by selection can be unfavorable for GEBV prediction which was based on the original marker-QTL LD structure in the training population. Although the decrease of accuracy of GS over time can be mitigated by reestimating marker effects or varying the weight given to markers, the possibility of using higher genotyping densities is generally preferred.

The use of lower-density marker panels has been an interesting option to reduce genotyping cost (Habier et al. 2009). The training population is genotyped with a full set of markers, but selection candidates are genotyped with a smaller selected subset. The pattern of linkage disequilibrium among markers in the training population is used to predict genotypes for the missing markers in the candidates. This strategy has been successfully implemented in dairy cattle (Berry and Kearney 2011) and became a standard practice in operational GS (Boichard et al. 2012). It could become an important strategy for GS in forest trees as well. Although theory predicts that a lower marker density would make GS more susceptible to the decay of LD with recombination, if prediction accuracies are mainly driven by relationship, low-density marker panels would be perfectly suitable together with continuous model retraining strategies. However, in forest trees the much wider genetic diversity across breeding programs might be such that the shared use of a common high-density SNP panels instead of each program developing a custom low-density panel could be economically more advantageous at least in the initial stages of GS. As GS programs of each individual organization advance and larger numbers of samples are genotyped by each breeding program, low-density SNP panels might become the standard practice.

9.5.2.1 Fixed Content SNP Arrays

The easy availability of shared commercial SNP chips has been a major strength of specific communities in advancing genomic selection into operational use. The best example of a widely uses common SNP platform for GS is the 50K bovine chip (Matukumalli et al. 2009). Large-scale genome-wide SNP discovery projects started relatively recently for forest tree genera such as *Picea* (Pavy et al. 2006), *Eucalyptus* (Novaes et al. 2008), *Populus* (Geraldes et al. 2011), and *Pinus* (Eckert et al. 2010; Lepoittevin et al. 2010). These efforts resulted in the development of some fixed content low-density arrays with hundreds of SNPs for *Pinus* (Chancerel et al. 2011; Eckert et al. 2009), *Picea* (Pavy et al. 2008), and *Eucalyptus* (Grattapaglia et al. 2011c). Moderate-density Infinium arrays were reported for *Pinus taeda* with 7,216 SNPs (Eckert et al. 2010), *Picea* with 9,539 SNPs (Pavy et al. 2013), *Pinus pinaster* with 9,000 SNPs (Plomion et al. 2016), and higher-density array with 34,000 SNPs for *Populus* (Geraldes et al. 2013). Although the SNP contents for these arrays were published, their use was restricted to those that developed it. They did not become commercially available products that one could order from a vendor or buy service from a provider.

Recently, however, a high-density multispecies SNP chip for eucalypts was developed from whole-genome resequencing of a large sample of 240 trees of 12 species (Silva-Junior et al. 2015). The EuCHIP60K, the highest-density SNP platform so far for a forest tree, provides close to 60,000 polymorphic SNPs across all the most widely planted and bred *Eucalyptus* species worldwide, providing a 96% genome-wide coverage and a density of one SNP every 12–20 kb or ~20–30 markers/cM. More importantly, however, not only its content is open source, but it was deliberately developed as a commercial product. The EuCHIP60K was made possible by a sort of community crowd-funding effort where eucalypt-based forest companies mainly from Brazil agreed to genotype at least 960 trees of their breeding programs, such that the minimum number of 15,000 samples was reached to cover the upfront cost of chip fabrication. This SNP chip is fully available to every interested institution, public or private, at a very competitive price through GeneSeek (NE, USA), an agricultural genomics service provider.

Over 30,000 *Eucalyptus* trees have already been genotyped with the EuCHIP60K at the time of this writing, the vast majority of the data used to start eucalypt GS experiments and pilot programs in several forest-based companies across the world. The use of a common genotyping platform across breeding programs should become a very valuable asset for future research and utilization of genomic selection. Common high-quality SNP data will allow, for example, the development of large-scale meta-analyses of GS data opening possibilities to develop and test prediction models based on much larger training populations. Furthermore, genomic selection experiments carried out with this chip are now providing the necessary information for the development of lower-density SNP chips for specific applications.

9.5.2.2 Genotyping-by-Sequencing (GbS) Approaches

Due to the general lack of accessible SNP arrays for the majority of forest tree species, GbS methods have been a useful entry technology to develop SNP resources and, in some cases, carry out high-density genotyping of forest tree populations. GbS allows capturing SNP diversity of much larger numbers of samples and carrying out SNP discovery even in very large and complex genomes with no reference. Large numbers of SNPs were discovered using RAD sequencing in *Eucalyptus* (Grattapaglia et al. 2011a), while GbS was used to mine SNPs in *Picea glauca* (Chen et al. 2013) and to carry out a genome-wide association study in *Pinus contorta* (Parchman et al. 2012). Optimized GbS methods were recently reported for three *Pinus* species (Pan et al. 2015). Sequence capture-based genotyping has also been successfully applied for a more targeted complexity reduction SNP discovery and genotyping in *Populus trichocarpa* (Zhou and Holliday 2012) and *Pinus taeda* (Neves et al. 2014). The first study to use GbS to carry out a GS study in forest trees was reported for *Picea engelmannii* (El-Dien et al. 2015).

GbS allows simultaneous discovery and genotyping of large numbers of markers with essentially no upfront costs (Davey et al. 2011). These methods require a genome complexity reduction step targeting a portion of the genome for selective enrichment, carried out either by PCR, restriction enzyme digestion, or sequence capture, followed by high-throughput NGS to ensure high sequence coverage of the targeted reduced representations (Cronn et al. 2012). Dominant presence/absence variants (PAVs), derived from polymorphism in the restriction recognition sites, and codominant single-nucleotide polymorphisms (SNPs) within the sequence tags are detected after aligning the sequence reads with or without the aid of a reference genome sequence. Restriction enzyme-based genome complexity reduction was also the basic approach of the widely used Diversity Arrays Technology (DArT) genotyping method (Kilian et al. 2012). This technology was successfully used to develop a fixed 7,680 probe array for *Eucalyptus* that allowed the first reported experimental results of genomic selection in forest trees (Grattapaglia et al. 2011b). This highly validated hybridization-based platform was later converted to an NGS-based assay named DArT-Seq, significantly improving throughput and marker number for *Eucalyptus* (Sansaloni et al. 2011).

9.5.2.3 Fixed Content SNP Arrays Versus GbS for Genomic Selection in Forest Trees

GbS methods have been attractive for they provide large numbers of SNPs at a relatively lower cost per sample when compared to fixed SNP arrays. GbS does not require prior sequence information, and there is no need to assemble a minimum number of samples of several hundred or thousands to defray the cost of chip fabrication. One only pays for the samples genotyped. The downside of GbS, however, is that to keep sample costs down, the sequencing coverage is generally low and therefore highly variable across the sampled loci in the genome (Beissinger et al. 2013). Technical issues associated with DNA digestion, PCR amplification of libraries, and sequencing process itself add a considerable amount of variation in what genomic loci are sampled and at what sequence depth during sequencing. This fact results in large proportions of missing data, usually around 40% up to 80% depending on the depth of sequencing employed (Poland and Rife 2012). This problem is mitigated by SNP imputation in inbred species where reference haplotypes are easily determined by deep sequencing of founder lines and expected genotypes are homozygous. In outbred forest trees, however, genotype imputation is not straightforward as genomes are highly heterozygous, and multiple unrelated parents are used such that reference haplotypes are not easily determined.

The problem of missing data in GbS tends to become substantial when attempting to genotype complex and highly heterozygous genomes of forest trees due to much higher restriction-site variation across individuals causing presence/absence variants and the need of higher coverage to declare heterozygous genotypes with confidence. This in turn leads to genotype reproducibility issues when one attempts to genotype the same sample across independent experiments.

Genomic loci and SNPs contained into them may be sampled or not in the replicates, and provided that the genomic locus is sampled, the genotype declared could match or not between replicates. In a study with Poplar, out of 16 GBS replicates of the same exact *Populus trichocarpa* tree Nisqually-1, the genotype used for genome sequencing, only 27% of in silico predicted restriction sites were sampled. Across the 16 replicates, on average, 26% of the SNPs were detected in only one of the 16 replicates, and only 9.6% were detected in all 16 replicates. Still, this amounted to ~34,000 loci out of the 334,000 total loci sampled. It is expected, however, that with a larger number of samples, the proportion of SNPs genotyped across all samples with high call rates would drop considerably. Genotype mismatches between replicated samples were largely due to low read coverage and were about 2% after heavy filtering (Schilling et al. 2014). These results are in line what has been reported for larger sample sizes genotyped by GbS in *Pinus engelmannii* (El-Dien et al. 2015). Out of 1.2 million initially sampled SNPs and after filtering by allowing 30% or up to 60% of missing data, the number of useful SNPs dropped to 8,868 or 62,198 depending on the different imputation approaches used. An imputation accuracy of 0.77–0.82 indicates that some ~20% of the genotype data could still be incorrect. No mention was made of genotype reproducibility in that study or the impact of such inaccuracies on genomic predictions.

It is intuitive that the success of genomic selection is dependent on SNP data quality. One has to be able to repeatedly genotype the same set of SNPs across generations with which the prediction models were initially developed in the training population. It is therefore not clear at this point whether the current GbS methods will be able to provide such data quality or, conversely, what is the tolerable genotyping inaccuracy for successfully applying GS. Demonstration GS proof-of-concept experiments are probably okay when carried out using GbS. However, to implement a professional routine of long-term GS into an industrial breeding program, very high standards of data quality should be sought. Currently only fixed SNP array provides the gold standard of data reproducibility (>99%), both in terms of sampling the same SNP loci and declaring the same exact genotype for the same individual across different sample batches and laboratories. This is probably one of the reasons why fixed SNP arrays have been the only platform used so far in domestic animal breeding, animal model research, and human genomic medicine.

Additionally, fixed SNP array data are breeder friendly and easily manageable and stored without the bioinformatics burden associated with GbS data. The common criticism of ascertainment bias of fixed content chips, a potential problem for population genetic studies, does not represent an issue for GS (Heslot et al. 2013), reminding that any GBS method is equally subject to such bias due to the genome complexity reduction methods involved, the biases inherent to next-generation sequencing, and the filtering pipeline applied for data analysis. Despite the falling prices of sequencing, the cost advantage of GbS in relation to fixed SNP arrays has dropped substantially in recent times with more flexible chip fabrication formats and competition among the main SNP chips vendors. It is therefore likely that fixed arrays will also become the standard for GS in breeding of the major tree species. Besides the EuCHIP60K already fitting such requirements, similar "community" chips are currently in development for loblolly pine (F. Isik, 'Technical meeting on Pine SNP chip development' 9 September 2016). It is also true, however, that novel, targeted GbS methods based on amplicon sequencing, sequence capture, or padlock probes will continue to evolve in parallel to dropping prices and increased precision of NGS platforms and analytical software. Once quality and cost issues are carefully evaluated, the key feature in choosing a SNP platform will be the flexibility that a new method provides to move seamlessly from one SNP platform to another while querying the same SNP set.

9.5.2.4 Whole-Genome Sequence Data for GS

With the evolution of sequencing technologies, a discussion has taken place on the value of moving from sparse SNP data to whole genome sequence data for the practice of genomic selection. Notwithstanding the challenge of managing massive NGS datasets for large numbers of individuals, in theory, if sequence data were used instead of dense SNPs, accuracy should increase because rare causal alleles would be better captured in the predictive models, and these in turn could be more stable as generations advance. However, simulation studies have shown that whole-genome sequence data does not bring any advantage in accuracy when the effective population size is reduced and LD is longer range which is usually the case in breeding programs (MacLeod et al. 2014). Another study found no justification to move to whole-genome sequencing for genomic selection unless accurate prior estimates on the functionality of SNP data could be included in the model (Perez-Enciso et al. 2015). If all SNPs within causal genes were included in the prediction model, accuracy could increase by ~40%. However, this advantage would be quickly lost if incorrect or incomplete biological information regarding SNP function was used.

9.5.3 Training Population: Size, Composition, and Phenotyping

Assembling a large number of trees into a training population to accurately estimate SNP effects is generally not a problem in forest tree breeding, although phenotyping costs can be an issue especially for traits that are expensive to measure. Choice of a training population evidently will depend in large part on the breeding strategy adopted. Training populations are typically established by sampling several hundred or a few thousand trees in existing progeny trials at ages that will allow extensive high-quality phenotyping of all traits targeted by the breeding program. These trials are derived from the inter-mating (open pollinated or controlled) of a set of a few to several dozen elite parents representative of the target germplasm, encompassing an adequate effective population size to provide sustained gains for a few generations ahead. Combining training sets from different populations can be useful to boost accuracy when individual populations lack sufficient size, although considerable risks exist of lowering the performance of such multi-population prediction models because relatedness with the prospective selection candidates is reduced or eliminated.

How many individuals should be included in a training population for GS in forest tree breeding? With up to $N = 1,000$ individuals, the selection accuracy was shown to rapidly increase, reaching satisfactory levels. With 2,000 individuals an improvement of ~10% in the accuracy would be expected, and larger improvements can be achieved under conditions of lower-heritability traits, larger numbers of QTLs involved, and larger effective population sizes. After $N = 2,000$ simulations have shown that the accuracy tends to plateau irrespective of the effective population size and genotyping density (Grattapaglia and Resende 2011). However, if the QTL distribution violates the infinitesimal model assumption of equal size effect and common variance, not all of the genetic variance is explained, and the selection accuracy can be lower depending on the method used to calculate the GEBV (Coster et al. 2010). Using training sets around $N = 2,000$ might, therefore, be warranted to protect against such model violations or cases where several hundred QTLs control trait variation. Simulation studies mirroring a eucalypt breeding scheme showed a considerable improvement of genomic prediction accuracies when increasing the training population size by consolidating phenotypic and genotypic data of individuals from previous breeding cycles (Denis and Bouvet 2013). Furthermore, larger training populations mitigate the probability of losing rare favorable alleles from the breeding population as generations of selection advance, although some will inevitably be lost because they are in low LD with any marker. A higher marker density will also help in this respect, i.e., in preserving rarer alleles in the breeding populations, thus allowing better long-term gains from selection.

Phenotyping large training populations of forest trees can be challenging and expensive. To mitigate this problem, a common approach widely used in forest tree breeding is to use indirect phenotyping methods such as NIRS (near-infrared reflectance spectroscopy) or X-ray diffraction for high-throughput measurements of chemical and physical wood properties. Although data collected by such methods are generally precise, they might not be accurate to the actual whole tree value, but they still allow confident raking of trees, which is generally satisfactory for GS. These methods were employed in the first experimental assessments of GS in *Eucalyptus* (Resende et al. 2012a) and white spruce (Beaulieu et al. 2014a, b) and recently for the assessment of a large set of chemical and physical traits in a GS study in *Eucalyptus* (Lima 2014). With current drops in genotyping costs, while phenotyping costs remain constant or increase, considerations have been given to a reverse approach in defining training populations so that individuals to be phenotyped are chosen on the basis of their genotypes. For example, Rincent et al. (2012) proposed different metrics to maximize the reliability of genomic predictions by optimizing the composition of individuals in the training population based exclusively on their genotypic data. Different criteria based on the diversity or on the prediction error variance from G-BLUP prediction were proposed to select reference individuals.

9.5.3.1 Clonal Replication of the Training Population

A common question made for tree species amenable to cloning, such as eucalypts and poplars, is whether vegetatively propagating the individuals of a training population would benefit the accuracy of a predictive model. In principle, by clonally replicating individuals, trait heritability would be increased, with a likely positive impact of accuracy. However, the impact of heritability on accuracy of GS has been shown to be very modest when genotyping is dense (see below). To answer this question it is relevant to remember that a key feature of GS is that phenotyping of the training population is done to train a model, not to directly select individuals. Selection subsequently proceeds on the basis of genomic estimated breeding values (GEBVs) such that prediction based on allele effects is now the selection criterion and the allele becomes the unit of evaluation. Alleles are therefore the units that need to be replicated not individuals (Lorenz et al. 2011). Therefore, when establishing a training population under a fixed phenotyping budget, it is more beneficial to increase the number of individuals phenotyped than clonally propagating and phenotyping a smaller number of individuals. Evidently, however, when no budget restrictions exist as far as phenotyping, clonally propagating a large number of individuals ($N \geq 2000$) as a training population would be advantageous, probably increasing the prediction accuracies, especially for low-heritability traits. Additionally, clonally propagating a training population would allow replicating it in several different environments and thus implementing a strategy in which the same breeding population is used to breed improved genetic material for different environments instead of a more costly option of advancing different populations for different environments. Phenotypes collected on the same genotypes in each environment would be used to build different prediction models for each environment therefore optimizing budgets even in the presence of significant genotype x environment interaction.

9.5.3.2 Genetic Relationship Between Training and Selection Candidates

The importance of relationship as a driver of accuracy in GS was shown early on from simulation studies and underscored in all recent reviews on the perspectives GS in plant and domestic animals breeding (Heslot et al. 2015; Lin et al. 2014; Van Eenennaam et al. 2014). Individuals closely related to the training population are always expected to have an advantage in accuracy over distantly related individuals. The demonstration that RR-BLUP and G-BLUP are equivalent implicitly showed that no LD between markers and QTLs is required for GS to work. The accuracy of GEBV is nonzero even without LD. However, when SNPs are the QTL themselves or in LD with the QTLs, RR-BLUP will provide better accuracy than

G-BLUP (Habier et al. 2007). This expectation was corroborated experimentally in forest trees where models developed for one population had limited or no ability of predicting phenotypes in an unrelated one (Beaulieu et al. 2014a, b; Resende et al. 2012a). These results indicate that the relatively low marker density used in these experiments has not been able to capture LD between QTLs and markers, such that prediction models have relied essentially on relatedness and are in principle population specific. Using stochastic simulations of a typical eucalypt breeding program (Denis and Bouvet 2013) also showed a marked decrease of the prediction accuracy at a rate of 10–15% per breeding cycle as the relationship between the training and candidate populations decreased.

Increasing the genetic relationships between training and selection candidates effectively has the same consequence as reducing the effective population size such that the stronger the relationship, the higher is the accuracy. Furthermore, in maize biparental populations it was shown that it is better to increase the accuracy of prediction by increasing relatedness between training and validation populations, rather than by increasing the size of the training set with less relatedness to predicted individuals (Riedelsheimer et al. 2013). Increased relatedness reduces the number of independently segregating chromosome segments (M_e) therefore increasing the probability that chromosome segments identical by descent sampled in the training population are also found in the selection candidates. For the successful implementation of GS it is therefore crucial that the selection candidates are genetically related to the training population.

The issue of relationship is one that should also be carefully considered when cross validating prediction models. The individuals on whom the models will be applied are the selection candidates, but the accuracy of predicting their phenotypes cannot be estimated because their phenotypes are not available. The models are therefore tested by cross-validation, typically using a subsample of the training population. Because relatedness is an important component of prediction accuracy, the most important principle of selecting a testing population is that it should mirror the relationship of the selection candidates to the training population (Daetwyler et al. 2013). If the testing population is more or less related to the training population than the selection candidates, then the prediction accuracy will be over- or underestimated, respectively. In replicated cross-validation, the manner in which individuals are assigned to particular folds affects accuracy. Random assignment of individuals to training or testing sets is prone to inflate accuracies because of within-family components driving them. A more realistic approach is to randomly assign whole full- or half-sib families to training or testing sets to evaluate prediction accuracy across families or to design cross-validation schemes that use genomic relationship data to partition individuals into the various folds to minimize the relationships between training and testing populations (Saatchi et al. 2011).

Beaulieu et al. (2014a) carried out the most informative study so far to evaluate the impact of genetic relationship on the accuracy of genomic prediction in outbred forest trees. A training population of 1,694 trees representative of 214 open-pollinated families was phenotyped for 12 wood and growth traits and genotyped for 6,385 SNPs. Three cross-validation schemes were applied with decreasing relationship between training and validation sets. CV1 involved allowing half-sib relationships between the sets, CV2 was performed by eliminating all maternal relatedness by assigning entire families to folds of the training and validation sets, and CV3 was designed to control for any possible contribution of the pollen parent to relatedness, thus eliminating as much as possible any possibility of coancestry between training and validation sets. Confirming expectations, they found that predictive ability between remotely related individuals (CV2) was only slightly lower (5–20% depending on the trait) than that of those built for closely related individuals (CV1). When the possibility of coancestry between cross-validation sets was eliminated and confirmed by an average estimated kinship coefficient of zero, the prediction accuracy was considerably reduced but still clearly different from zero for several of the traits supporting the putative presence of historical LD between SNPs and trait loci, despite the relatively sparse SNP data. In a subsequent study (Beaulieu et al. 2014b), this time dealing with two totally unrelated breeding groups, good predictions were obtained within each breeding group. However a sharp drop of accuracies near zero was seen when training was carried out in one group and cross-validated in the other. SNP genotyping was low density (6,932 SNPs) for an estimated recombining genome of 2,100 cM, and SNPs on the chip were not evenly distributed across the genome but rather targeted a limited set of candidate genes. The ability to capture historical LD was therefore very limited, if any, further confirming the key role of relationship as the pivotal driver of accuracy in these genomic prediction experiments.

Understanding the drivers of prediction accuracy in GS is a relevant issue because it has a direct impact on the ability of GS models to predict phenotypes in future generation removed from training. The concept of GS, as originally outlined, was based on the understanding that LD alone would explain the predictive ability of a model (Meuwissen et al. 2001). Later, however, it became clear both from simulation and experimental studies that prediction accuracy was also affected by genetic relationships between training and validation sets captured by SNP data (Habier et al. 2007; Legarra et al. 2008). The relative contributions of LD, additive genetic relationship, and cosegregation to the accuracy of predictions were modeled under different scenarios and their persistence over generations assessed (Habier et al. 2013). Among the several results that those simulations revealed, it was shown that the correlation between GEBVs within families depends largely on additive genetic relationship, which is determined by the size of the training populations and the effective number of SNPs. This latter one was defined as the number of ideal SNPs that provides the same accuracy due to additive genetic relationships as the actual number of SNPs in the model. It decreases with the increasing range of LD and therefore with decreasing effective population size, explaining why the accuracy due to additive genetic relationships does not improve beyond a certain SNP density. The lack of improvement in accuracy with increasing number of SNPs or, conversely, the rapid attainment of a plateau of accuracy with relatively few hundred SNPs has been a common observation in almost all GS studies in forest trees to date (Beaulieu et al. 2014a, b; Resende et al. 2012a, b). Surprisingly, the same prediction accuracies were seen whether using SNPs selected based on highest estimated effects for any particular trait or chosen randomly (Beaulieu et al. 2014b) providing strong evidence in support of relationships as the main source of accuracy. While the LD component of predictive accuracy is expected to persist over generations without the need for retraining, the component due to additive genetic relationships is anticipated to decay rapidly with the successive generations of recombination.

9.5.3.3 Populations for GS in Hybrid Breeding

Breeding for interspecific hybrids is an established strategy in some of the main plantation forest tree species. Hybrids combine desirable traits from two or more species through complementation of additive gene action. The best documented example in trees is the *E. grandis x E. urophylla* hybrid that combines growth and fungal disease resistance, respectively, and displays a heterotic effect due to nonadditive gene action. Elite hybrid individuals are clonally deployed, frequently exhibiting greater phenotypic stability that allows extending plantation range to sites where one or both parental species have a suboptimal performance (Rezende et al. 2014). Exploiting hybrid breeding in eucalypts can be done in a single synthetic population where the original species are hybridized at the outset to form a single breeding population which is then advanced by conventional recurrent selection (Kerr et al. 2004). Alternatively, reciprocal recurrent selection between the two species can be adopted, and the breeding goal in the pure species is to optimize the performance of hybrid descendants that are deployed either as clones or hybrid seed varieties.

The question arises on what would be the population to train a model for hybrid breeding under such a reciprocal recurrent selection strategy. Would it be the hybrid population or the pure species populations? The maintenance of predictive ability of a GS model across different populations or species will essentially rely on the consistency of LD across them, which in turn depends on the recombination rate between marker and QTLs and the time since the two diverged. The less diverged the populations are and the higher the marker density, better performance of the predictive model is expected across populations. An analogous situation takes place in bovine breeding in which selection is carried out in pure breeds, but the aim is to improve crossbred performance. Results from simulation studies generally show that training on crossbred data provides good prediction accuracy for selecting purebred individuals for crossbred performance. The incorporation of dominance in the model and the use of high-marker densities are generally beneficial (de Roos et al. 2009; Ibanz-Escriche et al. 2009; Kizilkaya et al. 2010; Zeng et al. 2013). When crossbred data is not available, separate purebred training populations can be used either separately or combined depending on the correlation of LD phase between the pure lines (Esfandyari et al. 2015).

In trying to make a parallel between the bovine breeding scenario and the case of eucalypt hybrid breeding, a prediction model could be trained on a hybrid population, i.e., a hybrid progeny trial, and used to select individuals in the two pure species for their performance as parents of hybrids. It is important to note, however, that while the estimates of the age of the most recent common ancestor of domesticated cattle range from 200 to 300 KYA (Murray et al. 2010), the estimated divergence time between the *Eucalyptus* species used in hybrid breeding is much older at 2–5 MYA (Silva-Junior and Grattapaglia 2015) such that it is not clear at this point if such an approach would actually work given the much wider divergence. If such a strategy proves effective, however, the hybrid progeny trial would also serve to train a model to select hybrid candidates to be deployed as clones. In other words, a prediction model involving additive effects would in principle be developed to select for high GEBV individuals in each species separately to serve as parents of the subsequent generation. If nonadditive effects are relevant to the target traits, a separate model including also nonadditive components would better serve to select individuals based on their genomic estimated genotypic value (GEGV) for clonal deployment. A simulation study showed that a GS model including dominance effect outperforms an additive model only when the training population is large and updated by combination of data from previous breeding cycles (Denis and Bouvet 2013). Clearly, experimental examination of the potential approaches and feasibility of applying GS to reciprocal recurrent selection in hybrid eucalypt populations deserve further attention.

9.5.4 Trait Heritability and Genetic Architecture

Theory predicts that the number of QTLs underlying trait variation will have an important impact on the accuracy of GS. Fewer loci controlling larger fractions of the phenotypic variance are more easily captured relative to a more complex genetic architecture involving larger numbers of loci with smaller effects. As pointed out earlier, QTL mapping experiments in forest trees have revealed increasing numbers of QTLs controlling each trait as more and larger mapping populations were used. It is reasonable, therefore, to assume that quantitative traits are controlled by several tens to hundreds of QTLs. Simulations have shown that the reduction of GS accuracy with an increasing number of QTLs tends to be more pronounced at lower-marker densities or larger effective population sizes. Assuming a total of 200 QTLs, marker densities $\geq$5–10 markers/cM would be necessary assuming a simpler genetic architecture, while 20 markers/cM would be necessary with larger numbers of QTLs (Grattapaglia and Resende 2011).

Heritability on the other hand was shown to have a relatively minor impact on accuracy when the training population size is large enough so that marker effects are adequately estimated. GS accuracy is directly proportional to the product of the heritability and the ratio between the number of phenotypic records in the training population and the number of QTLs involved (Daetwyler et al. 2008). Therefore, by simulating a scenario with a rather modest training set for a tree breeding situation of $N \geq 1000$ individuals, a trait controlled by 100 QTL, and an effective population size $N_e = 60$, the GS accuracy increased only slightly, from 0.71 to 0.83, as the heritability went from 0.2 to 0.6 (Grattapaglia and Resende 2011).

Simulation studies for animal breeding scenarios also showed that a decrease in accuracy with decreasing heritability is readily compensated by using larger training sets (Meuwissen et al. 2001; Nielsen et al. 2009).

9.5.5 Data Analysis Approaches for GS in Forest Trees

Genomic prediction requires methods that are capable of handling cases where the number of marker variables (p) greatly exceeds the number of individuals (n) (the large p small n problem) while mitigating the risk of model over parameterization. Several analytical approaches have been proposed and used for prediction of genome-estimated breeding or genotypic values. Ideally, a genomic prediction method should provide high accuracy, limit over-fitting on the training population, and preferably capture marker-QTL LD rather than relatedness for higher long-term stability. A good method should be easy to implement, reliable across a wide range of traits and datasets, and computationally efficient (Heslot et al. 2012). Several thorough reviews are available regarding the features of the main prediction methods for GS (de los Campos et al. 2013; Heffner et al. 2009; Lorenz et al. 2011), guidelines to compare them (Daetwyler et al. 2013), and comparative benchmark assessments in animal (Moser et al. 2009), crops (Heslot et al. 2012), and forest trees (Resende et al. 2012c). The current methods basically differ with respect to the assumptions regarding the genetic architecture of the trait for which genomic predictions are sought.

For the scope of this discussion, it is relevant to highlight the fact that across several reports in crops, trees, and domestic animals, the ridge regression best linear unbiased prediction (RR-BLUP) method using a mixed model has been very effective in providing the best compromise between computation time and prediction efficiency (Lorenz et al. 2011). RR-BLUP assumes that the trait is controlled by many loci of small effect, so that all marker effects are treated as random, normally distributed, and with a common variance. Results therefore suggest that most economically important quantitative traits adequately fit into the assumption of an infinitesimal model. In a loblolly pine study, for example, the performance of RR-BLUP and three Bayesian methods was only marginally different when compared across 17 traits with distinct heritabilities, with a small improvement using BayesA only for fusiform rust resistance where loci of relatively larger effect had been described (Resende et al. 2012c). Equivalent results were obtained for growth and wood traits in other forest trees showing no performance difference between RR-BLUP and Bayesian methods (Beaulieu et al. 2014b; Isik et al. 2016; Lima 2014; Ratcliffe et al. 2015). Considering the overall efficiency of RR-BLUP or the equivalent G-BLUP, a general recommendation has been made to use it as a starting point from which to explore additional alternative models (Heslot et al. 2012; Lorenz et al. 2011), although additional research in this area is warranted. Additional models would include Bayesian methods, when suspicion or prior information exists regarding the existence of loci of larger effect, or machine learning methods when nonadditive effects are known or presumed important.

9.5.5.1 Modeling Nonadditive Effects in Genomic Prediction

In several plant species, and particularly in some forest trees such as eucalypts, vegetative propagation of outstanding individuals is a key strategy for deploying elite genetic material. Clones maximize gains from selection by capturing additive and nonadditive effects. In forest trees, it is also common to observe that top parents may not be top clones and vice versa, suggesting considerable levels of nonadditive variation depending on the trait. A dominance to additive variance ratio close to 1.2 for growth was estimated in *E. grandis x E. urophylla* (Bouvet et al. 2009), while in *E. globulus* this ratio was 0.8 with indications that epistasis might be the main component of the nonadditive variance (Araujo et al. 2012). GS for tree breeding has therefore received increased attention in evaluating models including nonadditive effects. A simulation study directed to *Eucalyptus* breeding showed that a model including dominance effects performed better for clone selection only when dominance effects were preponderant (i.e., a dominance to additive variance ratio approaching 1.0) and heritability was >0.6 (Denis and Bouvet 2013).

Genomic data has also been successfully used to understand the relative importance of additive versus nonadditive variation and its implication in tree breeding. A number of studies have shown that the accuracy and stability of prediction models were improved by using marker-based instead of pedigree-based relationship matrices (Beaulieu et al. 2014a; Bouvet et al. 2016; Munoz et al. 2014; Zapata-Valenzuela et al. 2013). Besides correcting pedigree errors, marker-based matrices capture both the Mendelian segregation within full-sib families and genetic links through unknown common ancestors which are not available in the known pedigree. In *Pinus taeda* the use of a genomic relationships matrix yielded a better separation of additive and nonadditive components of the variance in height growth when compared to the pedigree-based model. Results provided evidence that additive pedigree-based models tend to inflate breeding values by capturing a large proportion of variance due to interaction terms. Additionally, it was shown that models including nonadditive relationship were more stable than traditional G-BLUP at predicting breeding values (Munoz et al. 2014). In hybrid eucalypts, using genome-wide information was also shown to improve the variance partition (Bouvet et al. 2016). At this point, however, no experimental data exist yet in forest trees regarding the ability of GS in predicting the total genotypic value of individual trees including additive and nonadditive effects, across generations. Research into this topic is one of the top priorities for forest tree species that are deployed as clonal varieties.

9.5.5.2 Genomic Prediction as a Ranking Problem

When judging the potential value of genomic prediction for selection, it is essential that the training and validation scheme adopted must reflect the way genomic prediction will be used in practice. The discussion on the feasibility of selecting individual trees for clonal propagation takes us to the recognition that until now, the predictive accuracy of a model has been typically assessed using the Pearson correlation between the observed trait values and the predicted trait values. GS has been essentially formulated as a regression problem. However, in tree breeding programs where clonal propagation is possible and clones are a genetic "dead end," i.e., they are not used back in breeding, it is common that the unit of selection and deployment is the individual tree. The breeder is simply interested in ranking individuals for their own merit from the best to the worst, without necessarily predicting their breeding value. This is particularly relevant to hybrid breeding strategies in eucalypts where individual trees are selected, ranked, tested, and eventually deployed as clones. When evaluating what individual tree ranking would GS reveal as compared to standard BLUP phenotypic selection, Lima (2014) reported a coincidence above 70% when selecting the top 30 trees out of the 1,000 of the training population by leave-one-out cross-validation and of 60% when tandem selection for volume growth and wood density was applied.

In a recent study, particularly relevant to tree breeding strategies that target individual tree selection for clonal propagation, Blondel et al. (2015) proposed to formulate GS as a ranking problem, showing that Pearson's correlation may correlate poorly with individual ranking accuracy. The approach also involves model estimation and candidate selection stages. However, instead of imposing that the model satisfies the equivalence of predicted with observed value, a score is assigned to each candidate, and the scores are used to rank the candidates. Machine learning methods were employed to rank individuals in six different datasets of both inbred and outbred plants. The approach showed a significantly higher efficiency to correctly rank individuals when compared to several standard regression methods. Clearly, this study opens a new avenue of GS research to develop methods that better fit the case of selecting top individuals for clonal propagation.

9.5.6 Genomic Prediction Accuracy Across Environments and Ages

G*E is essentially a lack of consistency in the relative performance of individuals when they are grown in different environments. Genotype by environment (G*E) interaction is a fact that all tree breeding programs commonly deal with. G*E can be of different levels, depending on the species, environmental variability and extent of the intended forest plantation sites, and type of planting material, whether families or clones, with clones typically being more interactive than families. Interactions can be more subtle when differences in performance are observed, but the relative ranking of tested individuals does not change across different environments (termed scale-effect interaction) or more severe types of interactions when rank changes are observed. Correct ranking of individual trees by their genetic value is a key component of the successful implementation of GS. Therefore, while the presence of scale-effect interactions should not represent a major limitation of a prediction model, rank changes are critical. When large rank change interactions are found, the GS strategy must account for this (Grattapaglia 2014).

Considerations and treatment of the interaction between genome predictions and environment will follow the same procedures used in dealing with standard G*E effects. Technically, there is nothing different between dealing with conventional G*E or genomic effect by environment interaction. The same consideration regarding the definition of breeding or management zones (i.e., the set of environments for which an improved variety is being developed), commonly applied to tree breeding programs, will apply to GS as well. Prediction models might be accurate across sites within the same breeding zone but probably not across breeding zones. However, the need to develop specific GS models for each breeding zone will largely depend upon the type of interaction observed, whether scale effect or rank change.

In forest trees multi-environmental G*E interaction not only is commonplace, but it is used to assess the performance of the same clones or families across different environmental conditions, to study genotype stability and to predict the performance of untested genotypes. Heffner et al. (2009) pointed out that GS opens the opportunity to evaluate the effect of particular genomic segments that are shared between lines across multiple environments. This information sharing should provide GS with stability of predictions even in the presence of G*E. This concept was put in practice by Burgueno et al. (2012) using a multi-environment dataset of wheat lines, showing that combining pedigree and marker data can yield substantial increases in prediction accuracy relative to traditional pedigree-based prediction and to single-environment pedigree and genomic prediction models. Multi-environment GS models enhanced predictive power in across-environment prediction, i.e., predicting the performance of genotypes that were evaluated in some environments but not in others.

The impact of environmental variation on the success of genomic predictions in forest trees has been evaluated, corroborating the expectations based on previous knowledge of G*E trends. Generally, all studies showed an important impact of environmental variation, but its magnitude and variation across traits followed what was already known from G*E studies, with growth traits showing higher interaction than wood properties. Resende et al. (2012b) clonally propagated and deployed the same set of 951 loblolly pine individuals in four locations on a north-south gradient along the Southeastern USA. Prediction models trained using the local phenotypes provided good predictions within site, but predictions got increasingly poorer as the geographical distance between training and testing sites increased along the latitudinal gradient. In white spruce, across-environment predictions were essentially the same as those within environment for wood traits but dropped for growth traits, confirming the contrasting behavior previously seen for these traits in typical G*E studies (Beaulieu et al. 2014b).

The accuracy of GS models in predicting the GEBV was assessed in interior spruce (*Picea glauca x P. engelmannii*) using a set of 1,126 38-year-old trees planted across three different sites originating from 25 open-pollinated families. Predictions of seven growth and wood traits were evaluated using four cross-validation scenarios: (1) training and validation within each individual site; (2) cross-site validation (all possible combinations); (3) within multisite, i.e., the three sites combined into a single training set; and (4) multisite training and validation in each individual site (El-Dien et al. 2015). Good accuracies were obtained when training and testing were carried out within each site despite the small training population available within each site, but, as expected, they dropped across sites. The estimated type-b genetic correlations between sites closely reflected the trend observed for the GS accuracy observed across sites. Prediction accuracies of a single multisite training model were higher for all seven traits when compared to the accuracies estimated in within-site validations, likely driven by the considerably larger training population used in this scenario with all 1,126 trees. Similarly, when the multisite model was validated on each separate site, accuracies were essentially as good as within site, suggesting that the positive effect of increasing the training population size counterbalanced the effect of environmental variation.

Another key aspect in forest tree breeding is the impact of age on the accuracy of predictions. Ideally, selection should be applied on trees at the same age when they are usually harvested. However, it is common for tree breeders to make selections at an earlier age in an attempt to accelerate a breeding program. The feasibility of such an approach will depend essentially on the magnitude of the age-age or juvenile-mature correlations which can be relatively high for wood quality traits but low for growth traits, although ample variation exists depending on species, environment, and ages considered (White et al. 2007). GS accuracy across ages was assessed in loblolly pine using diameter and height growth measurements obtained over multiple years (Resende et al. 2012b). As expected, given the weak juvenile-mature correlations typically observed in conifers (Namkoong et al. 1988), GS models trained on phenotypes measured at ages 1–2 years had unacceptable accuracy in predicting phenotypes at age 6 years. Equivalent results were reported in a recent study in *Picea engelmannii* using a series of repeated tree height measurements through ages 3–40 years on a population of 769 trees belonging to 25 open-pollinated families. Prediction accuracies varied substantially through time mirroring the spatial competition among trees. As expected, the behavior of genomic prediction accuracies across time was highly correlated with age-age genetic correlations and decreased substantially with increasing difference in age between the training and validation populations (Ratcliffe et al. 2015).

Results of the experimental studies reported to date on the impact of environment and age on the accuracy of genomic prediction in forest trees lead to a general conclusion. Existing data from traditional G*E or age-age correlation studies will inform with good precision what to expect from genomic prediction across environments and ages. As a rule, accurate predictions will require training models on traits measured at the same age and environment as the ones where predictions on selection candidates are planned. As age-age correlations between training and testing age improve and the magnitude and trend of the G*E interaction becomes inconsequential between training and testing sites, predictions will tend to be satisfactory, provided that genetic relationship between training and selection candidates is kept in the population.

There is an additional aspect to be considered about the prospects of GS across environments that was examined in the context of GS in crops (Heslot et al. 2015), but that will potentially be much more relevant and challenging in forest trees due to their longer life span. We have seen above that the appropriate tree phenotypes to train a prediction model should be collected at or very close to harvest age, which usually spans several years or decades, and preferably in the same target environment as the one where GS will be practiced on future selection candidates. The target environment of a forest tree is a consolidation of the action of several abiotic and biotic factors during the long and variable preceding life of the tree, including severe droughts, frosts, pest, and diseases attacks. Assuring that the target environment where phenotypes are collected for model training will be the same for the future selection candidates may therefore be much more critical (and challenging) for GS than in conventional phenotypic selection. In the latter, phenotypic data are used only to rank and select individuals on which phenotypes were measured. Thus, if a particular year of data is a misleading sample of the target environment due to some severe climate fluctuations during the life span of the tree, it will impact genetic gain for only that particular generation of selection. In GS, on the other hand, the unrepresentative data may affect genetic gain over a much longer period of time, as it will influence marker effect estimates that, in turn, will affect selection criteria going several generations forward. Periodical retraining with phenotypes collected in more recent generations of breeding might help mitigate this problem. Finally, despite the challenges of dealing with G*E, GS provides opportunities to integrate environmental covariates (e.g., climate data) to predict G*E deviations for unobserved environments. This approach can in turn allow prediction of individual stability, identification of important stresses, and understanding of the target environmental variation that is critical for breeding strategies (Heslot et al. 2015; Jarquin et al. 2014).

9.5.7 Performance of GS Across Generations

Proof-of-concept experiments in forest trees have been carried out by sampling training and validation sets within the same generation, usually the same progeny trial or different progeny trials, involving the same set of half- or full-sib families. Marker density was generally low, with a few thousand markers only, and accuracy was mostly driven by relatedness and not by marker-QTL LD. Not only experimental data is still lacking on simple two-generation cross-validation, but nothing is known about the performance of GS for long-term gain. The duly posed question by breeders is how accurate will the genomic predictions be on individuals several generations removed from the training population? As generations advance, recombination will erode both marker-QTL LD and links of relatedness between training and selection candidates reducing accuracy, while directional selection may change both the genetic architecture of the trait, via changes in allele frequencies, and the patterns of LD making them potentially unfavorable for GEBV prediction.

Recently, the first GS studies to evaluate GS accuracy by training and validating models using individuals of three generations G0, G1, and G2 were carried out in maritime pine (*Pinus pinaster*). In a first study, 184 individuals of the G0 parental and 477 of the G1 progeny generation were used (Isik et al. 2016). Mixed sets of parents and progeny were used for training and validation, resulting in good predictive abilities (0.43–0.49) for stem sweep, total height, and tree diameter. Recently, that population was expanded by including individuals of a G2 population. G2 individuals were preselected to include exclusively individuals that would limit the effective population size to $N_e = 25$, fully confirmed pedigree and highest BLUP for volume and stem straightness. Following simulations to select the best subsample of G2 individuals to maximize prediction accuracy, models trained on 46 G0 and 62 G1 individuals and validated on 710 G2 individuals showed high (0.70–0.85) predictive abilities despite the very small training population size, possibly a result boosted by the preselection of G2 individuals maximizing relatedness. Therefore, while promising results of GS have been reported in essentially all forest tree studies (Table 9.1), strictly speaking only one result of genomic prediction across generations is available so far, although several experiments are in the ground as we speak. Despite the inherent limitations of GS models validated exclusively within generation, they could still be quite useful in situations where the same crosses are repeated and prediction is applied on sibs of the original training set to increase selection intensity. This approach would be particularly useful to select top individuals to be deployed as clones by capturing additive and nonadditive effects, especially for late-expressing traits. However, when GS is applied to advance generations, selection candidates will rarely belong to the same population as the training set and may well be several generations removed from it.

Experimental studies assessing the performance of GS across multiple generations of breeding take some time to happen or rely on existing individuals of ancestral generations like the *Pinus pinaster* described above. However, several studies approached this issue by simulations. In the seminal study of Meuwissen et al. (2001), the decline of GS accuracy over generations was estimated at 5% per generation, getting smaller in later generations. Other studies under more complex models including the effect of directional selection and the structure and depth of the training population have been reported (Bastiaansen et al. 2012; Iwata et al. 2011; Jannink 2010; Long et al. 2011; Muir 2007; Sonesson and Meuwissen 2009). All these studies fundamentally converged to a similar recommendation: marker effects have to be reestimated frequently in order to maintain accuracy of predictions over generations. The issue of model updating was specifically assessed for a 60-year conifer tree breeding program by comparing the performance of GS with conventional phenotypic selection using stochastic simulations (Iwata et al. 2011). Results showed that GS outperformed phenotypic selection in the short term (30 years) but not in the long term (60 years). When the prediction model was updated, however, the genetic gain of GS was nearly twice that of phenotypic selection, even for low-heritability traits, with a greater advantage of GS as genotyping density increased. Two model updating strategies were tested. In a more conventional one, the prediction model generated in the initial cycle of selection is updated after three (or more) generations of GS by carrying out a progeny trial of already genotyped selection candidates, and their data is used to reestimate the marker effects. In a second strategy, in each cycle of GS, a subset of the genotyped selection candidates of that cycle is planted in a progeny trial. After a few years (depending on the species), phenotypes for that subset of trees become available and are used to update the prediction model. From that point on, every year the prediction model gets updated with the inclusion of phenotypic data of the extra subset of trees from previous generations. Because a set of trees from every cycle of GS is actually field grown, this second updating strategy allows continuous verification of the genetic progress of the GS program, although it involves greater costs of growing and measuring trees every generation and could theoretically increase the probability of unintended fixation of unfavorable alleles (Iwata et al. 2011). A significant advantage of model updating on GS accuracy by including phenotypic data from previous cycles was also shown by simulations in the context of *Eucalyptus* breeding (Denis and Bouvet 2013).

From the practical standpoint of a breeding program, continuously associating phenotypic data from previous cycles of GS and thus progressively updating prediction models and increasing the size and pedigree depth of the training population seem to be a very sensible and feasible approach to adopt. The cost of genotyping the subsets of selection candidates would have already been covered in the GS cycle, and growing and measuring a few hundred trees would not represent a significant cost while allowing for permanent monitoring of the realized performance of GS. Such a continuous retraining approach would allow the additive relationship component of predictive ability to be sustained across generations such that GS could be successfully practiced despite a limited ability to capture SNP-QTL LD due to the lower genotyping densities necessary to keep costs affordable in a breeding program.

9.5.8 Inbreeding and Maintenance of Genetic Diversity with GS

Finally, two additional issues have been raised regarding the performance of GS over the long term: inbreeding and loss of useful variation. GS could potentially result in a fast and unintended frequency increase of deleterious alleles causing inbreeding depression or fixation of unfavorable QTL alleles due to the progressive effect of drift with the restriction of effective population size. Daetwyler et al. (2007) showed that GS reduces the rate of inbreeding per generation when compared with sib and BLUP selection. High accuracies of estimated breeding values are achieved through better prediction of the Mendelian sampling term. This genomic-level resolution increases differentiation among sibs, allowing the breeder to better manage coancestry and to mitigate the rate of inbreeding even when selecting related individuals in breeding programs that are pushing for high genetic gains. Consistent with this expectation, the effect of nonrandom mating on the rate of inbreeding was found to be smaller for breeding schemes that adopt genome predictions when compared to conventional mating and selection designs (Nirea et al. 2012).

While GS is more efficient in reducing pedigree-based inbreeding when compared to BLUP by increasing emphasis on the individual rather than family information, pedigree inbreeding might not accurately reflect loss of genetic variation and the true level of inbreeding due to changes in allele frequencies and hitch-hiking. Liu et al. (2014) evaluated this issue using simulations, concluding that GS can have a greater impact than pedigree-based BLUP on the reduction of genetic diversity surrounding QTLs by a "hitch-hiking" effect because GS leads to a higher accuracy of selection on the QTL. Another reason might be that instead of directly selecting the QTL, selection acts on markers in LD with the QTL. This effect becomes more important when QTL effects are large, such that when implementing long-term genomic selection, genomic control of inbreeding is therefore essential to reduce the considerable hitch-hiking effects that are associated with genomic selection, regardless of the prediction model used.

The second issue regarding the impact of GS over time relates to the loss of favorable alleles with the faster successive cycles of breeding, potentially causing a progressive reduction of response to selection. Besides the loss of useful diversity, the hitch-hiking effect could also increase the frequency of linked deleterious alleles. Measures to mitigate this effect include using higher genotyping densities, periodical model updating, and verification of performance of a subset of selected trees along the GS cycles of breeding to monitor any possible reduction of vigor attributable to weakly or moderately deleterious mutations (Iwata et al. 2011). Additionally it has been shown that adopting weighed GS (Goddard 2009) together with using a larger training set (Jannink 2010) will help reducing the loss of low-frequency favorable alleles in the breeding population, although some will inevitably be lost due to low LD with any genotyped marker. In a simulation study, Jannink (2010) showed that placing additional weight on low-frequency favorable marker alleles allowed GS to increase their frequency earlier on, causing an initial increase in genetic variance. This procedure led to higher long-term gain while mitigating losses in short-term gain. Weighted GS also increased the maintenance of marker polymorphism, ensuring that QTL-marker linkage disequilibrium was higher than in conventional unweighted GS.

9.6 Conclusions and Perspectives

A number of recent experimental reports have now showed that the prospects of GS applied to forest tree breeding are very encouraging. To illustrate how one would envisage the operational flow of GS in tree breeding, Fig. 9.2 outlines the

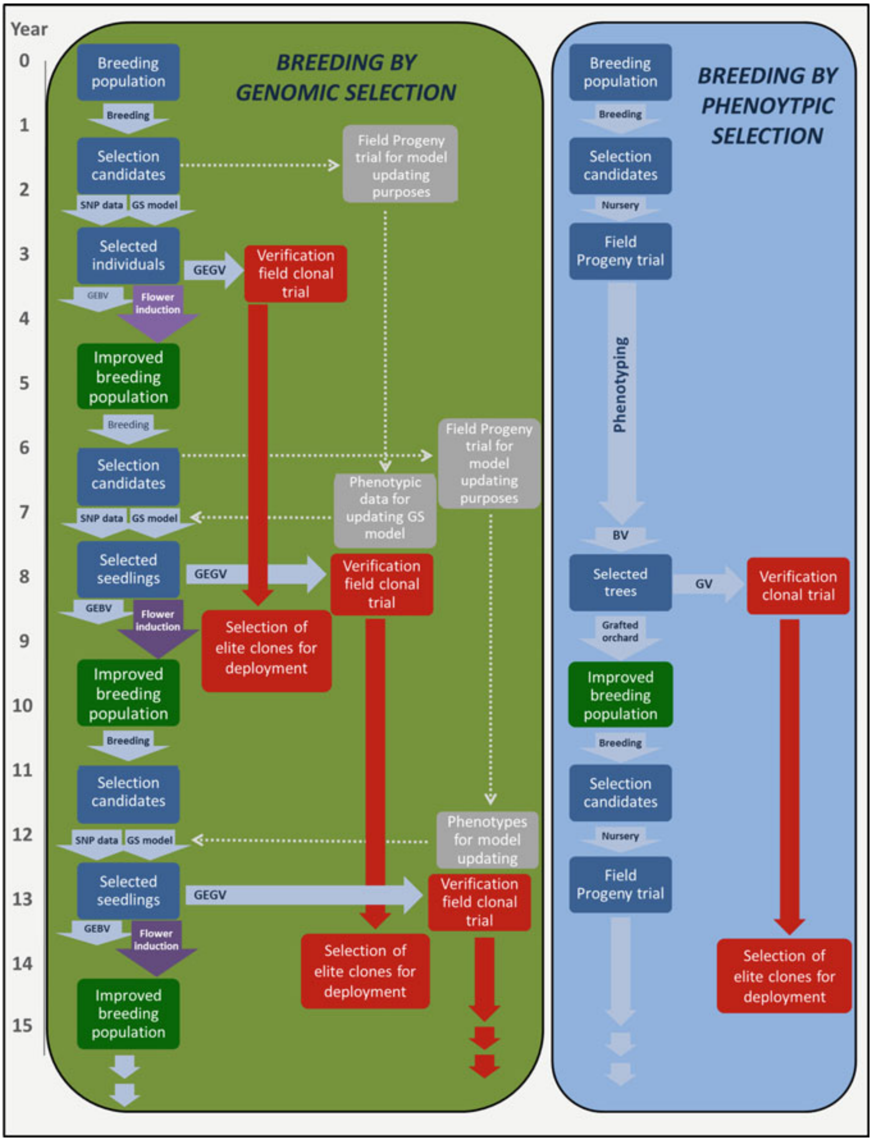

Year
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
BREEDING BY GENOMIC SELECTION
Breeding population
Breeding
Selection candidates
Field Progeny trial for model updating purposes
SNP data
GS model
Selected individuals
GEGV
Verification field clonal trial
GEBV
Flower induction
Improved breeding population
Breeding
Selection candidates
Field Progeny trial for model updating purposes
Phenotypic data for updating GS model
SNP data
GS model
Selected seedlings
GEGV
Verification field clonal trial
GEBV
Flower induction
Selection of elite clones for deployment
Improved breeding population
Breeding
Selection candidates
Phenotypes for model updating
SNP data
GS model
Selected seedlings
GEGV
Verification field clonal trial
GEBV
Flower induction
Selection of elite clones for deployment
Improved breeding population
BREEDING BY PHENOTYPIC SELECTION
Breeding population
Breeding
Selection candidates
Nursery
Field Progeny trial
Phenotyping
BV
Selected trees
GV
Verification clonal trial
Grafted orchard
Improved breeding population
Breeding
Selection candidates
Nursery
Field Progeny trial
Selection of elite clones for deployment

Fig. 9.2 Comparative timelines of genomic selection (GS) breeding and phenotypic selection (PS) breeding for tropical *Eucalyptus* (see text for details) (Modified from Grattapaglia (2014))

comparative timelines of breeding by GS and breeding by standard phenotypic selection for a recurrent selection strategy in tropical *Eucalyptus*. Both methods start at year zero with the same breeding population, and in the GS route, it is assumed that predictive models were previously developed. In a GS breeding cycle, following SNP genotyping and genomic prediction of all target traits (i.e., growth, form, wood properties, disease resistance, etc.), selection candidates can follow three possible nonexclusive routes:

1. Top ranked seedlings for GEBV (genomic estimated breeding value) are immediately routed to flower induction treatment and recombined to create the improved population (green boxes) completing the recurrent breeding cycle (green boxes);
2. Top ranked seedlings for GEGV (genomic estimated genotypic value) are cloned directly by mini-cutting methods and deployed in field verification clonal trials and ultimately submitted to a final selection for elite operational clones for plantation (red boxes);
3. A random subset of a few hundred selection candidates in each GS cycle can be planted in field trials to provide in due course additional phenotypic data to be added to the initial training dataset allowing continuous predictive model updating (gray boxes).

In the proposed scheme, GS is expected to eliminate the field progeny trial phase, accelerating the completion of a breeding cycle by allowing the selection of elite clones much faster. With GS, a cycle of recurrent selection in tropical *Eucalyptus* breeding, going from an original population to an improved population, will last 5 years, while in standard breeding it lasts at least 10 years. Two generations of elite clones can be developed by GS in 14 years, while standard phenotypic selection will only provide one generation in 15 years. Note that in standard breeding the verification clonal trial lasts 5–6 years to allow adequate phenotyping of wood properties traits. In GS, although accurate predictions of wood properties traits should obtained by GEBV, still this 5–6 years verification clonal trial is kept mostly to validate the general field performance and adaptability of the prospective clones. Depending on the performance of GS as the program proceeds, it might eventually be possible to preclude or shorten this final verification clonal trial, therefore further accelerating the deployment of new clones into the commercial forest.

The effective application of genomic prediction in a tree breeding program will vary on a case-by-case basis following a detailed cost-benefit analysis. GS might not be an option for small-scale breeding programs for tree species with a limited or niche market share, little prior genetic information on the species, and modest budgets. On the other hand, for aggressive breeding programs of the major tree species that support large industrial forest-based operations, it seems clear that time gains by eliminating progeny testing and streamlining clonal trials of young genomically ranked trees for multiple traits should be valuable. The adoption of GS might therefore become a competitive advantage in turning breeding generations quicker and thus deploying improved genetic stocks in the commercial forest at a faster rate. In concluding this chapter, it seems therefore useful to review the main lessons learned that have emerged so far from the experimental reports of genomic prediction in forest trees. They are summarized in a nine-point tentative roadmap that should assist tree breeders and managers when considering research or operational implementation of GS in their organizations:

1. *Starting Population for Training GS Models*. Leveraging existing progeny trials of the current breeding program, consisting of several tens of half- or full-sib families with relatively constrained effective population sizes, has been a successful approach to establish training populations. By mirroring actual tree breeding settings, the satisfactory prediction abilities estimated in essentially all studies and for all traits are good indications of the promising operational prospects of GS. Sampling preferably 2,000 individuals or at least 1,000 from such progeny trials for detailed phenotyping and SNP genotyping should be an effective way to establish a robust training population to start a GS program.
2. *Genetic Relationship Between Training Population and Selection Candidates*. For the successful implementation of GS, it is crucial that the selection candidates are genetically related to the training population. Studies that evaluated the impact of removing relatedness between training and validation sets have provided strong evidence in this respect. Higher genotyping densities and evaluation across multiple generations of breeding will now be needed to assess the relative importance of the decay of relatedness versus the SNP-QTL LD in maintaining satisfactory prediction abilities. Model updating strategies will likely be very important to counteract the expected decay of relationship and LD such that good prediction abilities might still be maintained with relatively sparse SNP genotyping densities.
3. *Genotyping Platform*. Studies in forest trees have shown satisfactory predictive abilities using relatively modest genotyping densities (2,500 ~ 10,000 SNPs) likely due to the leading role of relatedness as driver of accuracy. Higher marker densities should however be recommended to capture true LD and sustain long-term accuracies. There is ample room for improvement of SNP genotyping platforms in parallel with the development and experimental assessment of lower-density marker panels. While improved genotyping-by-sequencing (GbS) methods will likely surface in the near future, at this point fixed SNP array technologies unquestionably constitute the gold standard for data quality and breeder friendliness. Costs of such arrays have dropped significantly in recent years, although they still require upfront development costs which can be easily shared by interested organizations. This has been successfully done for species of *Eucalyptus* where a public SNP chip is available. Similar efforts are underway for species of *Pinus* such that high-standard public SNP genotyping platforms are today realistic targets for the mainstream plantation forest tree species.
4. *Genotype by Environment and Age Interactions*. Studies that evaluated the impact of G*E on the efficiency of GS showed that predictive abilities were reduced when models trained in one environment were validated in a different one, although the magnitude of such reduction varied across traits. Data from

traditional G*E or age-age correlation studies will inform with good precision what to expect from genomic prediction across environments and ages. As a general rule, accurate predictions will require training models on traits measured at the same age and environment as the ones where predictions on selection candidates are planned. As age-age correlations between training and testing age improve and the magnitude and trend of the G*E interaction becomes inconsequential between training and testing sites, predictions will tend to be satisfactory, provided that genetic relationships between training and selection candidates are kept in the population.

5. *Data Analysis*. A genomic prediction method should provide high accuracy, limit over-fitting on the training population, and capture marker-QTL LD besides relatedness for higher long-term stability. A good method should be easy to implement, reliable across a wide range of traits and datasets, and computationally efficient. Across several reports in crops, trees, and domestic animals, the RR-BLUP method and the G-BLUP equivalent have been effective in providing the best compromise between computation time and prediction efficiency. RR-BLUP assumes that the trait is controlled by many loci of small effect, therefore suggesting that most economically important quantitative traits in forest trees adequately fit into the assumption of an infinitesimal model. Several open-access softwares are available to implement this method, and training courses on their use are regularly offered by several institutions worldwide. Still, research on the subject is warranted to develop improved approaches including methods to efficiently incorporate nonadditive variation and individual ranking of trees for clonal selection.
6. *Logistics*. Logistic issues such as specific nursery infrastructure, sample collection and tracking system, large-scale DNA extraction and qualification, genotyping service providers, and data analysis pipelines are equally important modules for the successful implementation of a GS operation but beyond the scope of this chapter. Nevertheless, several of these components are either already routinely used in standard nursery operations of large forest-based companies or can be easily established in-house (e.g., DNA extraction lab) or accessed through specialized service providers in agricultural genomics.
7. *Cost-Benefit Analysis*. A detailed cost-benefit analysis of adopting GS using net present value methodologies is an absolutely necessary step before considering its implementation. The groundbreaking advance that GS caused in dairy cattle breeding is frequently used as an example of the economically successful use of this technology. It has been questioned whether it is an adequate benchmark for annual crops, although not so for forest trees where GS was considered to be potentially even more successful than in dairy cattle (Jonas and de Koning 2013). Still, while cattle and trees share the same challenge of long generation times, the logistics and cost of progeny testing a bull is substantially higher than progeny testing a tree, such that the cost of genotyping is easily justified and a remarkable gain in selection intensity has been possible. Current cost of

genotyping a sample even at USD ~40 is still expensive for many forest tree breeding programs and more so for those that run on very tight budgets. Assembling very large numbers of samples across breeding programs of several organizations on the same SNP genotyping array in long-term contracts is expected, however, to provide the necessary economy of scale to drive costs down in the near future.

8. *Prediction Across Generations*. Despite the encouraging estimates of predictive ability so far, it should be stressed once again that all studies but a recent one only evaluated the potential of GS within the same generation. In other words, training and validation sets were contemporary. Results of GS across independent generations of parents and progeny are limited so far and much less the performance of GS in generations farther removed from training. Moreover, the impact of recombination and selection across generations on prediction ability could not yet be assessed too, an issue that might become more relevant as generations of GS advance. There is a general urgency among research groups working in the area to provide additional experimental data on actual genomic selection across generations. Several experiments are underway especially in eucalypts, to compare the ranking of individual trees predicted at seedling stage based on genomic data with their realized ranking at rotation age for growth and wood quality traits.
9. *Changing Environment and Model Retraining*. Assuring that the target environment where phenotypes are collected for model training will be the same for the future selection candidates is a challenging issue for GS. In conventional breeding, phenotypic data are used only to rank and select individuals on which phenotypes were measured. Thus, if a particular year of data is a misleading sample of the target environment, it will impact genetic gain for only a short period of time. In GS, on the other hand, unrepresentative phenotypic data collected in the training population will affect genetic gain over a much longer period of time, affecting selection criteria going forward. Periodical retraining with phenotypes collected in more recent generations of breeding that were exposed to more recent environments should mitigate this problem.

GS is definitely a hot topic in tree breeding and a fast-moving area of research in several organizations worldwide, both public and private, working on the interface of genomics and quantitative genetics. While some of the fundamental genetic aspects discussed here are not likely to change much, or are valid under current technologies and circumstances, some others will almost unquestionably change in the future as new genotyping and sequencing technologies materialize and improved statistical approaches are developed. As GS adoption evolves and large experimental datasets are gathered across unrelated populations of tens of thousands of trees, the accumulation of genomic prediction data should also provide a powerful experimental framework, beyond QTL mapping and association genetics, toward the fundamental investigation of complex trait variation. The evolution of integrative approaches based on such large genotype and phenotype datasets should deliver important additional hints toward understanding the connections and interactions between the multitude of discrete genome-wide elements and the continuous phenotypic variation in complex traits. The full elucidation of such connections will nevertheless continue to be a very challenging endeavor due to the time and space dynamics of the effects of these genomic elements and the stochastic processes that thwart the expected one-to-one relationship between genotypes and phenotypes.

Acknowledgments This work was supported by CNPq grant 577047/2008-6, PRONEX-FAP-DF grant "NEXTREE" 2009/00106-8, EMBRAPA Macroprogram 2 grant 02.07.01.004, and a CNPq research fellowship to DG. Special thanks to all my students, collaborators, and colleagues worldwide working in genomic prediction and forest tree breeding with whom I have had the privilege to share and discuss several of the ideas presented in this chapter.

References

Araujo JA, Borralho NMG, Dehon G (2012) The importance and type of non-additive genetic effects for growth in Eucalyptus globulus. Tree Genet Genomes 8:327–337

Arus P, Verde I, Sosinski B, Zhebentyayeva T, Abbott AG (2012) The peach genome. Tree Genet Genomes 8:531–547

Assis TF, de Resende MDV (2011) Genetic improvement of forest tree species. Crop Breed Appl Biotechnol 11:44–49

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

Bastiaansen JWM, Coster A, Calus MPL, van Arendonk JAM, Bovenhuis H (2012) Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet Sel Evol 44:3

Beaulieu J, Doerksen T, Clement S, Mackay J, Bousquet J (2014a) Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity 113:343–352

Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J (2014b) Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC Genomics 15:1048

Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, Vaillancourt B, Buell CR, Kaeppler SM, de Leon N (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193:1073–1081

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649–1664

Bernardo R, Yu JM (2007) Prospects for genome wide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Berry DP, Kearney JF (2011) Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal 5:1162–1169

Blondel M, Onogi A, Iwata H, Ueda N (2015) A ranking approach to genomic selection. PLoS One 10:e0128570

Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, Gietzen KJ, Hayes BJ, Lawley CT, Sonstegard TS, Van Tassell CP, PM VR, Viaud-Martinez KA, Wiggans GR, Consortium BL (2012) Design of a bovine low-density SNP array optimized for imputation. PLoS One 7: e34130

Bouvet JM, Makouanzi G, Cros D, Vigneron P (2016) Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping: prediction accuracy implications. Heredity 116:146–157

Bouvet JM, Saya A, Vigneron P (2009) Trends in additive, dominance and environmental effects with age for growth traits in Eucalyptus hybrid populations. Euphytica 165:35–54

Brondani RP, Williams ER, Brondani C, Grattapaglia D (2006) A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers for the genus. BMC Plant Biol 6:20

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

Chancerel E, Lepoittevin C, Le Provost G, Lin YC, Jaramillo-Correa JP, Eckert AJ, Wegrzyn JL, Zelenika D, Boland A, Frigerio JM, Chaumeil P, Garnier-Gere P, Boury C, Grivet D, Gonzalez-Martinez SC, Rouze P, Van de Peer Y, Neale DB, Cervera MT, Kremer A, Plomion C (2011) Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. BMC Genomics 12:368

Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, Carta E, Dardano S, Dive M, Fasquelle C, Frennet JC, Hanset R, Hubin X, Jorgensen C, Karim L, Kent M, Harvey K, Pearce BR, Simon P, Tama N, Nie H, Vandeputte S, Lien S, Longeri M, Fredholm M, Harvey RJ, Georges M (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet 40:449–454

Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA (2013) Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genet Genomes 9:1537–1544

Coster A, Bastiaansen JWM, Calus MPL, van Arendonk JAM, Bovenhuis H (2010) Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet Sel Evol 42:9

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J (2012) Targeted enrichment strategies for next-generation plant biology. Am J Bot 99:291–311

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

Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365

Daetwyler HD, Villanueva B, Bijma P, Woolliams JA (2007) Inbreeding in genome-wide selection. J Anim Breed Genet 124:369–376

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:e3395

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345

de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. Genetics 183:1545–1553

Denis M, Bouvet JM (2013) Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genet Genomes 9:37–51

Dillen S, Storme V, Marron N, Bastien C, Neyrinck S, Steenackers M, Ceulemans R, Boerjan W (2008) Genomic regions involved in productivity of two interspecific poplar families in Europe. 1. Stem height, circumference and volume. Tree Genet Genomes 5:147–164

Echt CS, Saha S, Krutovsky KV, Wimalanathan K, Erpelding JE, Liang C, Nelson CD (2011) An annotated genetic map of loblolly pine based on microsatellite and cDNA markers. BMC Genet 12:17

Eckert AJ, Pande B, Ersoz ES, Wright MH, Rashbrook VK, Nicolet CM, Neale DB (2009) High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (Pinus taeda L.) Tree Genet Genomes 5:225–234

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, Gonzalez-Martinez SC, Neale DB (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* l., *Pinaceae*). Genetics 185:969–982

El-Dien OG, Ratcliffe B, Klapste J, Chen C, Porth I, El-Kassaby YA (2015) Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics 16:370

El-Kassaby YA, Lstiburek M (2009) Breeding without breeding. Genet Res 91:111–120

Esfandyari H, Sorensen AC, Bijma P (2015) Maximizing crossbred performance through purebred genomic selection. Genet Sel Evol 47:16

Freeman JS, Potts BM, Downes GM, Pilbeam D, Thavamanikumar S, Vaillancourt RE (2013) Stability of quantitative trait loci for growth and wood properties across multiple pedigrees and environments in Eucalyptus globulus. New Phytol 198:1121–1134

Geraldes A, Difazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N, Porth I, Mckown AD, Skyba O, Li E, Fujita M, Klapste J, Martin J, Schackwitz W, Pennacchio C, Rokhsar D, Friedmann MC, Wasteneys GO, Guy RD, El-Kassaby YA, Mansfield SD, Cronk QCB, Ehlting J, Douglas CJ, Tuskan GA (2013) A 34K SNP genotyping array for Populus trichocarpa: design, application to the study of natural populations and transferability to other Populus species. Mol Ecol Resour 13:306–323

Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao YJ, Tam A, Wang SC, Friedmann M, Birol I, Jones SJM, Cronk QCB, Douglas CJ (2011) SNP discovery in black cottonwood (Populus trichocarpa) by population transcriptome resequencing. Mol Ecol Resour 11:81–92

Gion JM, Carouche A, Deweer S, Bedon F, Pichavant F, Charpentier JP, Bailleres H, Rozenberg P, Carocha V, Ognouabi N, Verhaegen D, Grima-Pettenati J, Vigneron P, Plomion C (2011) Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: Eucalyptus. BMC Genomics 12:301

Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245–257

Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381–391

Grattapaglia D (2014) Breeding forest trees by genomic selection: current progress and the way forward. Chapter 26. In: Tuberosa R, Graner A, Frison E (eds) Advances in genomics of plant genetic resources. Springer, New York, pp 652–682

Grattapaglia D, Chaparro J, Wilcox P, Mccord S, Werner D, Amerson H, Mckeand S, Bridgwater F, Whetten R, O'malley D, Sederoff RR (1992) Mapping in woody plants with RAPD markers: applications to breeding in forestry and horticulture. Proceedings of the symposium "applications of RAPD technology to plant breeding". Crop Science Society of America, American Society of Horticultural Science, American Genetic Association, pp 37–40

Grattapaglia D, de Alencar S, Pappas G (2011a) Genome-wide genotyping and SNP discovery by ultra-deep Restriction-Associated DNA (RAD) tag sequencing of pooled samples of E. grandis and E. globulus. BMC Proc 5:P45

Grattapaglia D, Plomion C, Kirst M, Sederoff RR (2009) Genomics of growth traits in forest trees. Curr Opin Plant Biol 12:148–156

Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. Tree Genet Genomes 7:241–255

Grattapaglia D, Resende MDV, Resende M, Sansaloni C, Petroli C, Missiaggia A, Takahashi E, Zamprogno K, Kilian A (2011b) Genomic selection for growth traits in Eucalyptus: accuracy within and across breeding populations. BMC Proc 5:O16

Grattapaglia D, Ribeiro VJ, Rezende GD (2004) Retrospective selection of elite parent trees using paternity testing with microsatellite markers: an alternative short term breeding tactic for Eucalyptus. Theor Appl Genet 109:192–199

Grattapaglia D, Sederoff R (1994) Genetic-linkage maps of Eucalyptus-grandis and Eucalyptus-urophylla using a pseudo-testcross – mapping strategy and RAPD markers. Genetics 137:1121–1137

Grattapaglia D, Silva OB, Kirst M, de Lima BM, Faria DA, Pappas GJ (2011c) High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. BMC Plant Biol 11:65

Grattapaglia D, Vaillancourt R, Shepherd M, Thumma B, Foley W, Külheim C, Potts B, Myburg A (2012) Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. Tree Genet Genomes 3:463–508

Greenwood MS, Adams GW, Gillespie M (1991) Stimulation of flowering by grafted black spruce and white spruce – a comparative-study of the effects of gibberellin A4/7, cultural treatments, and environment. Can J For Res 21:395–400

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397

Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. Genetics 182(1):343–353

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194:597–607

Haley CS, Visscher PM (1998) Strategies to utilize marker-quantitative trait loci associations. J Dairy Sci 81:85–97

Harfouche A, Meilan R, Kirst M, Morgante M, Boerjan W, Sabatti M, Mugnozza GS (2012) Accelerating the domestication of forest trees in a changing world. Trends Plant Sci 17:64–72

Hasan O, Reid JB (1995) Reduction of generation time in Eucalyptus-globulus. Plant Growth Regul 17:53–60

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009a) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443

Hayes BJ, Visscher PM, Goddard ME (2009b) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91:47–60

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Heslot N, Jannink JL, Sorrells ME (2015) Perspectives for genomic selection applications and research in plants. Crop Sci 55:1–12

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

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160

Ibanz-Escriche N, Fernando RL, Toosi A, Dekkers JCM (2009) Genomic selection of purebreds for crossbred performance. Genet Sel Evol 41:12

Ingvarsson PK, Garcia MV, Luquez V, Hall D, Jansson S (2008) Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 Locus in European aspen (Populus tremula, Salicaceae). Genetics 178:2217–2226

Isik F, Bartholome J, Farjat A, Chancerel E, Raffin A, Sanchez L, Plomion C, Bouffier L (2016) Genomic selection in maritime pine. Plant Sci 242:108–119

Iwata H, Hayashi T, Tsumura Y (2011) Prospects for genomic selection in conifer breeding: a simulation study of *Cryptomeria japonica*. Tree Genet Genomes 7:747–758

Jannink JL (2010) Dynamics of long-term genomic selection. Genet Sel Evol 42:35

Jannink JL, Zhong SQ, Dekkers JCM, Fernando RL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355–364

Jarquin D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Perez P, Calus M, Burgueno J, de los Campos G (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127:595–607

Jia Y, Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513–1522

Jonas E, de Koning DJ (2013) Does genomic selection have a future in plant breeding? Trends Biotechnol 31:497–504

Jonas E, de Koning DJ (2015) Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. Front Genet 6:49

Junghans DT, Alfenas AC, Brommonschenkel SH, Oda S, Mello EJ, Grattapaglia D (2003) Resistance to rust (Puccinia psidii Winter) in eucalyptus: mode of inheritance and mapping of a major gene with RAPD markers. Theor Appl Genet 108:175–180

Kerr RJ, Dieters MJ, Tier B (2004) Simulation of the comparative gains from four different hybrid tree breeding strategies. Can J For Res 34:209–220

Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, Caig V, Heller-Uszynska K, Jaccoud D, Hopper C, Aschenbrenner-Kilian M, Evers M, Peng K, Cayla C, Hok P, Uszynski G (2012) Diversity arrays technology: a generic genome profiling technology on open platforms. Methods Mol Biol 888:67–89

Kizilkaya K, Fernando RL, Garrick DJ (2010) Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J Anim Sci 88:544–551

Lambeth C, Lee BC, O'Malley D, Wheeler NC (2001) Polymix breeding with parental analysis of progeny: an alternative to full-sib breeding and testing. Theor Appl Genet 103:930–943

Legarra A, Robert-Granie C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. Genetics 180:611–618

Lepoittevin C, Frigerio JM, Garnier-Gere P, Salin F, Cervera MT, Vornam B, Harvengt L, Plomion C (2010) In vitro vs in silico detected SNPs for the development of a genotyping array: what can we learn from a non-model species? PLoS One 5:e11034

Lima BM (2014) Bridging genomics and quantitative genetics of Eucalyptus: genome-wide prediction and genetic parameter estimation for growth and wood properties using high-density SNP data. Genetics Dep. University of São Paulo, Piracicaba, SP, Brazil, pp 93. Available in English at http://www.teses.usp.br/teses/disponiveis/11/11137/tde-25062014-25085814/pt-br.php

Lin Z, Hayes BJ, Daetwyler HD (2014) Genomic selection in crops, trees and forages: a review. Crop Pasture Sci 65:1177–1191

Liu HM, Sorensen AC, Meuwissen THE, Berg P (2014) Allele frequency changes due to hitch-hiking in genomic selection programs. Genet Sel Evol 46:8

Long N, Gianola D, Rosa GJM, Weigel KA (2011) Long-term impacts of genome-enabled selection. J Appl Genet 52:467–480

Lorenz AJ, Chao SM, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink JL (2011) Genomic selection in plant breeding: knowledge and prospects. Adv Agron 110:77–123

MacLeod IM, Hayes BJ, Goddard ME (2014) The Effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. Genetics 198 (4):1671–1684

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS, Van Tassell CP (2009) Development and characterization of a high density SNP genotyping assay for cattle. PLoS One 4:e5350

McKeand SE, Bridgwater FE (1998) A strategy for the third breeding cycle of loblolly pine in the Southeastern US. Silvae Genet 47:223–234

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol 41:56

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

Munoz PR, Resende MFR, Gezan SA, Resende MDV, de los Campos G, Kirst M, Huber D, Peter GF (2014) Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics 198:1759–1768

Murray C, Huerta-Sanchez E, Casey F, Bradley DG (2010) Cattle demographic history modelled from autosomal sequence variation. Philos T R Soc B 365:2531–2539

Namkoong G, Kang HC, Brouard JS (1988) Tree breeding: principles and strategies. Springer Verlag, New York

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nat Rev Genet 12:111–122

Neale DB, Williams CG (1991) Restriction-fragment-length-polymorphism mapping in conifers and applications to forest genetics and tree improvement. Can J For Res 21:545–554

Nejati-Javaremi A, Smith C, Gibson JP (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. J Anim Sci 75:1738–1745

Neves LG, Davis JM, Barbazuk WB, Kirst M (2014) A high-density gene map of loblolly pine (Pinus taeda L.) based on exome sequence capture genotyping. G3 Genes Genom Genet 4:29–37

Nielsen HM, Sonesson AK, Yazdi H, Meuwissen THE (2009) Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. Aquaculture 289:259–264

Nirea KG, Sonesson AK, Woolliams JA, Meuwissen THE (2012) Effect of non-random mating on genomic and BLUP selection schemes. Genet Sel Evol 44:11

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9:312

Novaes E, Osorio L, Drost DR, Miles BL, Boaventura-Novaes CRD, Benedict C, Dervinis C, Yu Q, Sykes R, Davis M, Martin TA, Peter GF, Kirst M (2009) Quantitative genetic analysis of biomass and wood chemistry of Populus under different nitrogen levels. New Phytol 182:878–890

Pan J, Wang BS, Pei ZY, Zhao W, Gao J, Mao JF, Wang XR (2015) Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Mol Ecol Resour 15:711–722

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol Ecol 21:2991–3005

Pavy N, Gagnon F, Rigault P, Blais S, Deschenes A, Boyle B, Pelgas B, Deslauriers M, Clement S, Lavigne P, Lamothe M, Cooke JEK, Jaramillo-Correa JP, Beaulieu J, Isabel N, Mackay J, Bousquet J (2013) Development of high-density SNP genotyping arrays for white spruce (Picea glauca) and transferability to subtropical and nordic congeners. Mol Ecol Resour 13:324–336

Pavy N, Parsons LS, Paule C, MacKay J, Bousquet J (2006) Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. BMC Genomics 7:174

Pavy N, Pelgas B, Beauseigle S, Blais S, Gagnon F, Gosselin I, Lamothe M, Isabel N, Bousquet J (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics 9:1–17

Pelgas B, Bousquet J, Beauseigle S, Isabel N (2005) A composite linkage map from two crosses for the species complex Picea mariana x Picea rubens and analysis of synteny with other Pinaceae. Theor Appl Genet 111:1466–1488

Perez-Enciso M, Rincon JC, Legarra A (2015) Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol 47:43

Plomion C, Bartholome J, Lesur I, Boury C, Rodriguez-Quilon I, Lagraulet H, Ehrenmann F, Bouffier L, Gion JM, Grivet D, de Miguel M, de Maria N, Cervera MT, Bagnoli F, Isik F, Vendramin GG, Gonzalez-Martinez SC (2016) High-density SNP assay development for genetic analysis in maritime pine (Pinus pinaster). Mol Ecol Resour 16:574–587

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92–102

Pryce JE, Daetwyler HD (2012) Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim Prod Sci 52:107–114

Rae A, Pinel M, Bastien C, Sabatti M, Street N, Tucker J, Dixon C, Marron N, Dillen S, Taylor G (2008) QTL for yield in bioenergy *Populus*: identifying G×E interactions from growth at three contrasting sites. Tree Genet Genomes 4:97–112

Ratcliffe B, El-Dien OG, Klapste J, Porth I, Chen C, Jaquish B, El-Kassaby YA (2015) A comparison of genomic selection models across time in interior spruce (Picea engelmannii x glauca) using unordered SNP imputation methods. Heredity 115:547–555

Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA, Pappas GJ, Kilian A, Grattapaglia D (2012a) Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 194:116–128

Resende MFR, Munoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV, Kirst M (2012b) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol 193:617–624

Resende MFR, Munoz P, Resende MDV, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012c) Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.) Genetics 190:1503–1510

Rezende GDSP, Resende MDV, Assis TF (2014) Eucalyptus breeding for clonal forestry. In: Fenning T (ed) Challenges and opportunities for the world's forests in the 21st century. Springer Science+Business Media, Dordrecht, pp 393–424

Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503

Rincent R, Laloe D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodriguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E, Schoen CC, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.) Genetics 192:715–728

Saatchi M, McClure MC, McKay SD, Rolf MM, Kim J, Decker JE, Taxis TM, Chapple RH, Ramey HR, Northcutt SL, Bauck S, Woodward B, Dekkers JCM, Fernando RL, Schnabel RD, Garrick DJ, Taylor JF (2011) Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet Sel Evol 43:40

Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, Kilian A (2011) Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. BMC Proc 5: P54

Schilling MP, Wolf PG, Duffy AM, Rai HS, Rowe CA, Richardson BA, Mock KE (2014) Genotyping-by-sequencing for populus population genomics: an assessment of genome sampling patterns and filtering approaches. PLoS One 9:95292

Silva-Junior OB, Faria DA, Grattapaglia D (2015) A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing 240 *Eucalyptus* tree genomes across 12 species. New Phytol 206:1527–1540

Silva-Junior OB, Grattapaglia D (2015) Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis. New Phytol 208:830–845

Solberg TR, Sonesson AK, Woolliams JA, Odegard J, Meuwissen THE (2009) Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. Genetics Selection Evolution 41 (1):53

Sonesson AK, Meuwissen THE (2009) Testing strategies for genomic selection in aquaculture breeding programs. Genet Sel Evol 41:37

Stirling B, Newcombe G, Vrebalov J, Bosdet I, Bradshaw HD (2001) Suppressed recombination around the MXC3 locus, a major gene for resistance to poplar leaf rust. Theor Appl Genet 103:1129–1137

Strauss SH, Lande R, Namkoong G (1992) Limitations of molecular-marker-aided selection in forest tree breeding. Can J For Res 22:1050–1061

Telfer EJ, Stovold GT, Li YJ, Silva OB, Grattapaglia DG, Dungey HS (2015) Parentage reconstruction in Eucalyptus nitens using SNPs and microsatellite markers: a comparative analysis of marker data power and robustness. PLoS One 10:e0130601

Thumma BR, Southerton SG, Bell JC, Owen JV, Henery ML, Moran GF (2010) Quantitative trait locus (QTL) analysis of wood quality traits in Eucalyptus nitens. Tree Genet Genomes 6:305–317

Van Eenennaam AL, Weigel KA, Young AE, Cleveland MA, Dekkers JCM (2014) Applied animal genomics: results from the field. Annu Rev Anim Biosci 2:105–139

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

White TL, Adams WT, Neale DB (2007) Forest genetics. CABI Publishing, Cambridge, MA, p 682

Wilcox PL, Amerson HV, Kuhlman EG, Liu BH, O'Malley DM, Sederoff RR (1996) Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. Proc Natl Acad Sci U S A 93:3859–3864

Williams CG (1988) Accelerated short-term genetic testing for loblolly-pine families. Can J For Res 18:1085–1089

Williams CG, Neale DB (1992) Conifer wood quality and marker-aided selection: a case-study. Can J For Res 22:1009–1017

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J, Neale D, McKeand S, Whetten R (2012) SNP markers trace familial linkages in a cloned population of *Pinus taeda* – prospects for genomic selection. Tree Genet Genomes 6:1307–1318

Zapata-Valenzuela J, Whetten RW, Neale D, McKeand S, Isik F (2013) Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. G3 Genes Genom Genet 3:909–916

Zeng J, Toosi A, Fernando RL, Dekkers JCM, Garrick DJ (2013) Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol 45:11

Zhou LC, Holliday JA (2012) Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. BMC Genomics 13:703

Index

R.K. Varshney et al. (eds.), *Genomic Selection for Crop Improvement*,
DOI 10.1007/978-3-319-63170-7