Aus dem Institut für

Pflanzenzüchtung, Saatgutforschung und Populationsgenetik

der Universität Hohenheim

Fachgebiet Angewandte Genetik und Pflanzenzüchtung

Prof. Dr. Albrecht E. Melchinger

Factors influencing the Accuracy of Genomic Prediction in Plant Breeding

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften

vorgelegt
der Fakultät Agrarwissenschaften

von
Master of Science
Pascal Schopp
aus Merdingen

Stuttgart-Hohenheim
2017

Die vorliegende Arbeit wurde am 05.10.2017 von der Fakultät Agrarwissenschaften als “Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften (Dr. sc. agr.)” angenommen.

Tag der mündlichen Prüfung: 12.03.2018

1. Prodekan: Prof. Dr. Stefan Böttinger

Berichterstatter, 1. Prüfer: Prof. Dr. Albrecht E. Melchinger

Mitberichterstatter, 2. Prüfer: Prof. Dr. Chris-Carolin Schön

3. Prüfer: apl. Prof. Dr. Tobias Würschum

Contents

[1] Schopp, P., D. Müller, F. Technow and A.E. Melchinger, 2017. Genetics 205:1–14

[2] Schopp, P., C. Riedelsheimer, H.F. Utz, C.-C. Schön and A.E. Melchinger, 2015. Theor Appl Genet. 128:2189–2201

[3] Schopp, P., D. Müller, Y.C. J. Wientjes and A.E. Melchinger, 2017. G3 7:3571–3586

Abbreviations

BPF	Biparental family
DH	Doubled-haploid
(G)BLUP	(Genomic) best linear unbiased prediction
GCA	General combining ability
GP	Genomic prediction
GS	Genomic selection
IBD	Identity-by-descent
IBS	Identity-by-state
LD	Linkage disequilibrium
MAS	Marker-assisted selection
PA	Prediction accuracy
QTL	Quantitative trait locus/loci
TP	Testcross performance

1. General Introduction

The genetic improvement of complex traits by means of artificial selection is the primary objective in the breeding of most agriculturally relevant crop and animal species. In breeding programs targeting complex traits, a prerequisite for an efficient identification of superior genotypes is to accurately estimate their breeding (or genetic) values (Falconer and Mackay 1996). In cattle breeding, breeding values of bulls have traditionally been assessed by means of extensive progeny evaluations (Schaeffer 2006). By comparison, plant breeders typically evaluate the genetic potential of selection candidates by replicated testing of lines or clones in multi-environmental field trials (Smith *et al.* 2005). This approach is motivated by the prominent role of genotype by environment interactions, which impose limits on heritability and, thus, selection gain (Van Eeuwijk 2006). Since phenotyping of a large number of selection candidates in multiple environments is both time- and cost-consuming, a particular focus of research on plant breeding has been the development of more efficient selection strategies (Hallauer *et al.* 2010).

With the advances in genotyping technologies, genome-wide molecular markers have become increasingly available and affordable within the last two decades (Bernardo 2008). In the context of quantitative genetics, molecular markers opened new gateways for dissecting the genetic architecture of complex traits (Paterson 1998). Specifically, molecular mapping promised to identify the causal polymorphisms in the DNA sequence affecting the phenotype – so-called quantitative trait loci (QTL) – and employ them for marker-assisted selection (MAS) (Lande and Thompson 1990). Here, one attempts to localize unobserved QTL via statistical associations with genetically linked molecular markers, *i.e.,* linkage disequilibrium (LD). However, estimating each of the small genetic effects for the often hundreds of QTL underlying complex traits (Holland 2007) is infeasible with the sample sizes available in most species. To overcome this problem, molecular mapping approaches introduce thresholds to restrictively select the loci having a statistically significant effect on the target trait. This necessity comes at the expense of often only minor proportions of the total genetic variance being explained by markers (Myles *et al.* 2009), a phenomenon that has been termed 'missing heritability' (Maher 2008). The issue becomes particularly evident when considering the limited success in the dissection of complex traits in exceptionally large plant populations (Buckler *et al.* 2009) or even in molecular mapping studies on humans, which can be regarded as the upper benchmark for statistical power owing to data sets including tens of thousands of individuals (Altshuler *et al.* 2008). Thus, researchers came to realize that the hope put in MAS as an appropriate tool for the breeding of complex traits must be tempered; the only exception being traits controlled by few large-effect QTL (Xu and Crouch 2008). Consequently, there was an urgent need for alternative approaches to marker-driven selection of complex traits – such as grain yield in crops or fertility rate in cattle – to further promote the breakthrough of molecular markers in breeding.

Genomic prediction and selection

To overcome the drawbacks of MAS, Meuwissen *et al.* (2001) presented in their landmark paper an novel approach, which they termed genomic selection (GS). The fundamental idea behind GS is to renounce significance testing of marker effects and substitute it with a simultaneous estimation of genome-wide marker effects. Subsequently, marker effects associated with an individuals' genotype are summed up to obtain a so-called genomic estimated breeding value. Thus, emphasis is shifted from direct selection on putative QTL to selection on breeding values. This reduces missing heritability (Yang *et al.* 2010) and therefore facilitates capturing significantly larger amounts of the total genetic variance than MAS (Calus *et al.* 2008). Like for MAS (as opposed to GS), estimation of marker effects in GS requires a so-called training set of individuals for which both phenotypic and genotypic data are available to calibrate a prediction equation. In the literature, this estimation step and the associated methodology has been termed genome-wide, genome-enabled or simply genomic prediction (GP), to differentiate it from GS, *i.e.,* from the actual selection of candidates based on their genomic estimated breeding values.

Initially, the focus was primarily on developing statistical models for GP. These models have the common feature that they have to cope with the "small *n,* large *p"* problem (de Los Campos *et al.* 2013), which emerges if the number of predictor variables (*e.g.,* hundreds to thousands of molecular markers) strongly exceeds the number of (phenotypic) observations. To handle this problem, various types of models have been applied to or were newly developed for GP (de Los Campos *et al.* 2013; Gianola 2013). Besides fundamental methodical differences, these models are distinguished mainly in the assumed distributions of QTL effects underlying complex traits. Basic penalized regression models such as ridge-regression BLUP (Whittaker *et al.* 2000; Meuwissen *et al.* 2001) and a computationally efficient re-parameterization thereof (genomic BLUP; Habier *et al.* 2007; VanRaden 2008) assume that all markers capture an equally small proportion of the total genetic variance. This assumption is justified by one of quantitative genetics' most fundamental principles: the infinitesimal model

(Fisher 1918). Among the multitude of alternative models, most differ from these basic models in that they allow for variable selection, *i.e.,* for differentially weighing markers according to the genetic variance they explain (de Los Campos *et al.* 2013; Wimmer *et al.* 2013). Examples of this class include much-noticed Bayesian penalized regression models (Meuwissen *et al.* 2001; Gianola 2013), frequentist penalized regression models such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996; Usai *et al.* 2009), as well as a variety of semi- and non-parametric models such as reproducing kernel Hilbert spaces and neural networks (Gianola *et al.* 2006; Perez-Rodriguez *et al.* 2013). Despite sensible theoretical justification and promising results in simulation studies, there is scarce empirical evidence for the superiority of these more sophisticated models (Heslot *et al.* 2012; Daetwyler *et al.* 2013). This is particularly true for the prediction of highly complex traits in structured populations exhibiting high levels of LD (Wimmer *et al.* 2013), which represents the central application of GP in commercial plant breeding. In recognition of this and the widespread implementation of the BLUP-based approaches in practice, this thesis is based exclusively on the use genomic BLUP (GBLUP) and discusses the potential of alternative models only in few relevant cases.

Sources of quantitative-genetic information affecting the accuracy of genomic prediction

As in every regression problem, the estimation step in GP crucially depends on the sample size available for model training and on the signal-to-noise ratio of observations, reflected by the heritability of the phenotypes (Daetwyler *et al.* 2008, 2010). In addition, the targeted predictor variables are not observed in GP, because the QTL underlying complex traits are largely unknown (Zhong *et al.* 2009; Daetwyler *et al.* 2010). As long as GP relies on genotyping arrays containing markers in the order of several thousands, causal mutations are mostly not contained in the observed genotypic data (Pérez-Enciso *et al.* 2015). Therefore, GP rests on the fundamental assumption that each causal locus is in linkage disequilibrium (LD) with at least one molecular marker (Meuwissen *et al.* 2001). Just as in molecular mapping, the strength of LD determines the statistical power with which true QTL effects are captured by marker effects (Lande and Thompson 1990).

Originally, it was assumed that ancestral LD is the only 'source of information' contributing to the accuracy of GP (Meuwissen *et al.* 2001). Ancestral LD between QTL and markers refers to the LD measured in a defined base population and depends on ancestral effective population size (Sved 1971; Hill 1981), as well as on the allele frequencies at QTL and markers (Heslot *et al.* 2013) and marker density (Solberg *et al.* 2008). It was shown that prediction accuracy (*i.e.,* the correlation between true and estimated breeding or genetic values) generally increases with higher marker density, as well as with decreasing effective ancestral population size as a consequence of the higher level of LD due to genetic drift (Muir 2007). However, it took until the study of Habier *et al.* (2007) to realize that ancestral LD is not the only information source contributing to prediction accuracy (PA). Using simulation, they demonstrated that even under linkage equilibrium between QTL and markers (achieved by artificial localization on different chromosomes), markers provide substantial PA given that training and predicted individuals are related by pedigree. They concluded that PA increases with the pedigree relationships between training and predicted individuals, because they are implicitly captured in GP. This finding was later verified theoretically by Gianola *et al.* (2009) and empirically by Habier *et al.* (2010), Clark *et al.*

(2012) and others. While in principle, pedigree relationships contribute to PA of all GP models, the coherence becomes most obvious if (i) it is assumed that all markers contribute equally to the genetic variance (ridge-regression BLUP) and (ii) the model is re-parameterized as a natural extension of the traditional pedigree BLUP (Henderson 1984). This leads to GBLUP (Habier *et al.* 2007; VanRaden 2008), which replaces the numerator relationship matrix of pedigree BLUP with a marker-derived genomic relationship matrix. The fundamental point made by Habier *et al.* (2007) was that if QTL and markers are in linkage equilibrium, the genomic relationship matrix is approximately proportional to the numerator relationship matrix and therefore provides similar accuracy of estimated breeding values as pedigree BLUP.

In actuality, however, 'actual relationships' between individuals at QTL vary around their expected value (as expressed by their pedigree relationship) due to Mendelian sampling (Hill and Weir 2011). For instance, full sibs are expected to share 50% of their alleles, while the true value might range from around 20 to 80%. This variance in actual relationships is not reflected by the expected relationship and, hence, not utilized in pedigree BLUP. It is, therefore, the capability of markers to trace Mendelian sampling deviations that determines the superiority of GP over traditional pedigree BLUP (Hayes *et al.* 2009; Goddard *et al.* 2011).

Besides ancestral LD, it was suspected that actual relationships are captured if QTL and markers are inherited non-independently due to linkage on the same chromosomes, called co-segregation (Hayes *et al.* 2009; Goddard *et al.* 2011). Again, Habier *et al.* (2013) were the first to demonstrate this explicitly through well-designed simulations. They concluded that there are three sources of information affecting the accuracy of GP: pedigree relationships at QTL captured by markers, ancestral

LD between QTL and markers and co-segregation between QTL and markers. While ancestral LD is relatively stable across generations and contributes information to all related or unrelated descendants of the ancestral (or base) population (Habier *et al.* 2007), co-segregation information specifically refers to QTL and marker alleles originating from the same parental gametes (Hayes *et al.* 2009; Habier *et al.* 2013). Therefore, co-segregation information is conditional on the pedigree between training and predicted individuals. The pedigree structure is a function of the mating design, which itself is characterized by the number of parents sampled from the base population, as well as the type and number of generations of intermating. Hence, conclusions on the relevance of the three information sources in GP obtained from animal or plant breeding populations of large recent effective population size (Habier *et al.* 2007, 2013; Wientjes *et al.* 2013) are context-specific and not expected to be generalizable.

In plant breeding, few parental genotypes are often taken from a germplasm (corresponding to the base or ancestral population) and subsequently intermated to generate so-called synthetic populations (Bradshaw 2016). Synthetics took on a pioneering role in quantitative-genetic research, particularly in the development and success of recurrent selection approaches (Hallauer 1992). Further examples of synthetics include multi-parent advanced generation intercross (MAGIC) populations (Cavanagh *et al.* 2008), which were initially proposed for QTL mapping (Dell'Acqua *et al.* 2015; Wei and Xu 2015), but also gained momentum in breeding applications (Huang *et al.* 2012; Bandillo *et al.* 2013). Synthetics have also been suggested as suitable base material for recurrent GS (Windhausen *et al.* 2012; Massman *et al.* 2013; Gorjanc *et al.* 2016).

This thesis addresses the question of how the number of parents affects the accuracy of GP and the information sources contributing to it, by treating synthetics as a conceptual framework to generate a continuous gradient of numbers of parents employed for population development. This proceeding enabled us to demonstrate the role of the three information sources in GP across a wide range of scenarios relevant to breeding. We use these insights to discuss the implications for GP of related and unrelated material in a broad context, beyond that of synthetics in the traditional sense.

Genomic prediction in hybrid breeding

Hybrid cultivars are usually obtained by crossing to genetically distant parental inbred lines from opposite germplasm, called heterotic pools (Melchinger and Gumber 1998). In hybrid breeding, initial variation for selection within these pools is typically created by producing multiple biparental crosses between superior inbred lines from previous cycles (Hallauer 1990; Mikel and Dudley 2006). The already large number of inbred lines derived from biparental crosses that has been evaluated every season has further increased since the introduction of the doubled-haploid (DH) technology (Prigge and Melchinger 2012). To handle this quantity, breeders use multi-stage selection (Cochran 1951; Utz 1969) to divide up the search for the best individuals into multiple consecutive steps. First, lines are selected based on their *per se* performance for traits mainly related to seed production. Subsequently, the central step in hybrid breeding is to assess the lines' testcross performance (TP) with tester genotypes (so-called 'testers') from the opposite heterotic group (Mihaljevic *et al.* 2005). At this point, the number of selection candidates is still large, so that TP can only be evaluated in a moderate number of test environments and with only one to few testers. This imposes restrictions on the estimation of general combining ability

(GCA), and, therefore, on heritability and selection gain. GP of TP in early multi-stage selection promises to counteract these challenges by facilitating (i) increased selection intensities through evaluation of more selection candidates and (ii) higher heritabilities by phenotyping fewer candidates more precisely (Technow *et al.* 2013). For these reasons, GP of early-stage TP has been considered among the most promising applications in hybrid breeding, which was empirically backed by encouraging PAs within individual breeding stages (Albrecht *et al.* 2011; Zhao *et al.* 2012). However, extensions to multi-stage selection remained unclear, particularly because advanced stages typically include different or further testers and/or test environments. The study of Windhausen *et al.* (2012) indicated declining PA for GP across testers and test environments and emphasized the need for further research on this issue. Albrecht *et al.* (2014) investigated the PA when GP was performed across genetic groups, testers, environments and years, but results were not clear-cut due to confounding effects among these factors. Specifically, they were not able to determine to which extent PA decreased because of different testers in the training and prediction set, because different groups of individuals were paired with different testers. Although the imbalance of their data set was representative of what is regularly found in practice, theoretical insights into the influence of changing testers and test environments on PA are required to optimize multi-stage selection programs integrating GP. Given a breeding stage comprising a finite number of candidates, breeders need to know how PA evaluated within the same stage (*e.g.,* assessed by cross-validation) relates to the phenotypic performance of selected individuals in subsequent stages, because genotype by tester or genotype by environment interactions may lead to rank changes among genotypes from one stage to another. Moreover, stringent truncation selection is usually practiced after the first testcross evaluation, but its effect on the PA observed in subsequent stages has not been addressed hitherto.

Genomic prediction in biparental families

Since the GP methodology has been developed largely in an animal breeding context, research on GP in plant breeding primarily targeted issues related to the specific population structures (Guo *et al.* 2014). In commercial breeding programs, populations are typically stratified into multiple subpopulations, most commonly into related and unrelated biparental families (BPF) of inbred lines (Albrecht *et al.* 2011; Technow and Totir 2015). The expected differences among BPFs can be reliably predicted from the mean performance of their respective parents (Melchinger 1987). Thus, the central application of GP in commercial plant breeding is the identification of superior lines within BPFs (Riedelsheimer *et al.* 2013). Accordingly, it is most meaningful to also evaluate PA within BPFs (as opposed to across families) to exclude the effect of different population means for valid comparisons among populations and methods (Windhausen *et al.* 2012; Guo *et al.* 2014). In agreement with the fundamental insights of earlier studies (Habier *et al.* 2007; Daetwyler *et al.* 2010), PA crucially depends on the pedigree relationship between the predicted BPF and the BPF(s) used for model training (Riedelsheimer *et al.* 2013; Lehermeier *et al.* 2014), as well as on the sample size, marker density, and trait heritability (Jannink *et al.* 2010; Wimmer *et al.* 2013). Given the typical stratification of breeding populations, favorable preconditions for GP can be created by calibrating prediction models either with individuals from the same BPFs or from connected BPFs (*e.g.,* half-sibs sharing one common parent with the predicted BPF) from the same and/or previous breeding cycles

(Riedelsheimer *et al.* 2013; Jacobson *et al.* 2014; Lehermeier *et al.* 2014; Lian *et al.* 2014). Despite these straightforward guidelines for designing suitable training sets, the afore-cited studies found considerable variation in PA when predicting across different combinations of BPFs, as well as for different traits. Although uncertainty about PA hampers the confidence in GP as a novel tool in practice – where breeders relied for decades exclusively on phenotypic selection – no study has systematically investigated the reasons for this phenomenon.

For efficient allocation of resources and design of training sets in GP, it would be highly desirable to forecast PA prior to phenotyping and direct assessment via cross-validation. Various deterministic formulae have been derived for this task (VanRaden 2008; Goddard 2009; Daetwyler *et al.* 2010), and were applied rather successfully also within populations in plant breeding (Lorenz 2013; Riedelsheimer *et al.* 2013; Riedelsheimer and Melchinger 2013; Lian *et al.* 2014; He *et al.* 2016). However, PA within BPFs is naturally expected to be relatively high and depends primarily on the sample size and heritability of the phenotyped fraction of individuals (Riedelsheimer and Melchinger 2013). Thus, forecasting PA within BPFs has at most a general informational value regarding the allocation of resources to single BPFs (number vs. accuracy of phenotypic observations). For breeders, it would be much more relevant to forecast PA across BPFs, *e.g.*, for evaluating the prospects of GP based on related and unrelated BPFs phenotyped in preceding cycles, henceforth referred to as 'early prediction' (Jacobson *et al.* 2014; Auinger *et al.* 2016). Here, accurate forecasts of PA could assist breeders in differentially allocating resources to new crosses (number of produced and phenotyped lines), depending on the PA they can expect based on *a priori* existing data. Despite numerous empirical investigations on across-family GP (Riedelsheimer *et al.* 2013;

Jacobson *et al.* 2014; Lehermeier *et al.* 2014; Albrecht *et al.* 2014), deterministic formulae have so far received little attention in these scenarios. One reason for this might be the theoretical hurdles in the adoption of existing formulae for across-population GP derived for outbred individuals in animal breeding (Hayes *et al.* 2009; Wientjes *et al.* 2015, 2016) to BPFs of inbred lines in plant breeding. In this thesis, we present modifications of existing deterministic formulae and demonstrate their usefulness in simulations, using as a starting point the simple case of GP within and across single BPFs. The strong variation in PA that can be encountered in these prediction scenarios is carefully examined and discussed with respect to the use of deterministic formulae.

Objectives

This thesis sought to examine how various factors specific to plant breeding populations affect the accuracy of genomic prediction. In particular, the objectives were to

i. Assess by simulation how prediction accuracy and the contributions of the three information sources in synthetic populations depend on the number of parents, relatedness between training and prediction set and the level of ancestral disequilibrium;

ii. Develop a theoretical framework based on selection index theory to forecast prediction accuracy of testcross performance of inbred lines across different testers and test environments, and validate the formulae in an empirical data set of maize;

iii. Analyze the mean and variance in prediction accuracy within and across biparental families of inbred lines and provide extensions to existing deterministic formulae to forecast prediction accuracy;

iv. Discuss the implications of the results in a broad context, regarding relevant applications in commercial plant breeding and beyond.

Literature cited

Albrecht, T., H.-J. Auinger, V. Wimmer, J. O. Ogutu, C. Knaak *et al.*, 2014 Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor. Appl. Genet. 127: 1375–1386.

Albrecht, T., V. Wimmer, H. Auinger, M. Erbe, C. Knaak *et al.*, 2011 Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123: 339–350.

Altshuler, D., M. Daly, and E. Lander, 2008 Genetic Mapping in Human Disease. Science 322: 881–888.

Auinger, H.-J., M. Schönleben, C. Lehermeier, M. Schmidt, V. Korzun *et al.*, 2016 Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.). Theor. Appl. Genet.

Bandillo, N., C. Raghavan, and P. Muyco, 2013 Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6: 1–15.

Bernardo, R., 2008 Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. Crop Sci. 48: 1649–1664.

Bradshaw, J. E., 2016 *Plant Breeding: Past, Present and Future.* Springer International Publishing.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown *et al.*, 2009 The genetic architecture of maize flowering time. Science 325: 714–718.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics 178: 553–561.

Cavanagh, C., M. Morell, I. Mackay, and W. Powell, 2008

From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr. Opin. Plant Biol. 11: 215–221.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. J. van der Werf, 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.

Cochran, W. G., 1951 Improvement by means of selection. Proc. 2nd Berkeley Symp Math Stat Prob 449–470.

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

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams, 2010 The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021–1031.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams, 2008 Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3: e3395.

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

Dell'Acqua, M., D. M. Gatti, G. Pea, F. Cattonaro, F. Coppens *et al.*, 2015 Genetic properties of the MAGIC maize population : a new platform for high definition QTL mapping in Zea mays. Genome Biol. 1–23.

Falconer, D. F., and T. S. C. Mackay, 1996 *Introduction to Quantitative Genetics* (1996 Longman, Ed.). Pearson, Essex.

Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc.

Edinb. 52: 399–433.

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

Gianola, D., R. L. Fernando, and A. Stella, 2006 Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. Genetics 173: 1761–1776.

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

Goddard, M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245–257.

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen, 2011 Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128: 409–421.

Gorjanc, G., J. Jenko, S. J. Hearne, and J. M. Hickey, 2016 Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17: 30.

Guo, Z., D. Tucker, C. Basten, G. Harish, E. Ersoz *et al.*, 2014 The impact of population structure on genomic prediction in stratified populations. Theor. Appl. Genet. 127: 749–762.

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

Habier, D., R. L. Fernando, and D. J. Garrick, 2013 Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics 194: 597–607.

Habier, D., J. Tetens, F. Seefried, P. Lichtner, and G. Thaller, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42: 5.

Hallauer, A. R., 1990 Methods used in developing maize inbreds. Maydica 35: 1–16.

Hallauer, A. R., 1992 Recurrent selection in maize, in Plant Breeding Reviews Vol. 9. Wiley & Sons Ltd.

Hallauer, A. R., M. J. Carena, and J. M. Filho, 2010 *Quantitative genetics in maize breeding.* Springer New York.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard, 2009 Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41: 51.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. Cambridge 91: 47–60.

He, S., A. W. Schulthess, V. Mirdita, Y. Zhao, V. Korzun *et al.*, 2016 Genomic selection in a commercial winter wheat population. Theor. Appl. Genet. 129: 641–651.

Henderson, C., 1984 *Applications of linear models in animal breeding.* University of Guelph, ON.

Heslot, N., J. Rutkoski, J. Poland, J. L. Jannink, and M. E. Sorrells, 2013 Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. PLoS One 8: e74612.

Heslot, N., H.-P. Yang, M. E. Sorrells, and J.-L. Jannink, 2012 Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. 52: 146–160.

Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. 38: 209–216.

Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. Cambridge 93: 47–64.

Holland, J. B., 2007 Genetic architecture of complex traits in plants. Curr. Opin. Plant Biol. 10: 156–161.

Huang, B. E., A. W. George, K. L. Forrest, A. Kilian, M. J. Hayden *et al.*, 2012 A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol. J. 10: 826–839.

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo, 2014 General combining ability model for genomewide selection in a biparental cross. Crop Sci. 54: 895–905.

Jannink, J.-L., A. J. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Briefings Funct. genomics proteomics 9: 166–177.

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

Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan *et al.*, 2014 Usefulness of multi-parental populations of maize (Zea mays L.) for genome-based prediction. Genetics 198: 3–16.

Lian, L., A. Jacobson, S. Zhong, and R. Bernardo, 2014 Genomewide prediction accuracy within 969 maize biparental populations. Crop Sci. 54: 1514–1522.

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

Maher, B., 2008 Personal genomes: The case of the missing heritability. Nature 456: 18–21.

Massman, J. M., H.-J. G. Jung, and R. Bernardo, 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.

Melchinger, A. E., 1987 Expectation of means and variances of testcrosses produced from from F2 and backcross individuals and their selfed progenies. Heredity (Edinb).

59: 105–115.

Melchinger, A. E., and R. K. Gumber, 1998 Overview of heterosis and heterotic groups in agronomic crops. Concepts Breed. heterosis Crop plants 29–44.

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

Mihaljevic, R., C.-C. Schön, H. F. Utz, and A. E. Melchinger, 2005 Correlations and QTL correspondence between line per se and testcross performance for agronomic traits in four populations of European maize. Crop Sci. 45: 114–122.

Mikel, M. A., and J. W. Dudley, 2006 Evolution of North American dent corn from public to proprietary germplasm. Crop Sci. 46: 1193–1205.

Muir, W. M., 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.

Myles, S., J. Peiffer, P. J. Brown, E. S. Ersoz, Z. Zhang *et al.*, 2009 Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21: 2194–2202.

Paterson, A. H., 1998 *Molecular Dissection of Complex Traits*. Taylor & Francis Inc.

Pérez-Enciso, M., J. C. Rincón, and A. Legarra, 2015 Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 47: 43.

Perez-Rodriguez, P., D. Gianola, J. M. Gonzalez-Camacho, J. Crossa, Y. Manes *et al.*, 2013 Comparison Between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat. G3 2: 1595–1605.

Prigge, V., and A. E. Melchinger, 2012 Production of Haploids and Doubled Haploids in Maize, pp. 161–172 in *Plant Cell*

Culture Protocols, Humana Press, Totowa, NJ.

Riedelsheimer, C., J. B. Endelman, M. Stange, M. E. Sorrells, J. L. Jannink *et al.*, 2013 Genomic predictability of interconnected biparental maize populations. Genetics 194: 493–503.

Riedelsheimer, C., and A. E. Melchinger, 2013 Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor. Appl. Genet. 126: 2835–2848.

Schaeffer, L. R., 2006 Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123: 218–223.

Smith, A. B., B. R. Cullis, and R. Thompson, 2005 The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J. Agric. Sci. 143: 449.

Solberg, T. R., A. K. Sonesson, J. a Woolliams, and T. H. E. Meuwissen, 2008 Genomic selection using different marker types and densities. J. Anim. Sci. 86: 2447–2454.

Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2: 125–141.

Technow, F., A. Bürger, and A. E. Melchinger, 2013 Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3 3: 197–203.

Technow, F., and L. R. Totir, 2015 Using Bayesian Multilevel Whole Genome Regression Models for Partial Pooling of Training Sets in Genomic Prediction. G3 5: 1603–1612.

Tibshirani, R., 1996 Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B 58: 267–288.

Usai, M. G., M. E. Goddard, and B. J. Hayes, 2009 LASSO with cross-validation for genomic selection. Genet. Res. (Camb). 91: 427–436.

Utz, H. F., 1969 *Mehrstufenselektion in der Pflanzenzüchtung*.

Verlag Eugen Ulmer Stuttgart.

Van Eeuwijk, F., 2006 *Genotype by Environment Interaction — Basics and Beyond*. Plant Breeding: The Arnel R. Hallauer International Symposium.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Wei, J., and S. Xu, 2015 A Random Model Approach to QTL Mapping in Multi-parent Advanced Generation Inter-cross (MAGIC) Populations. Genetics 202: 471–486.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genet. Res. 75: 249–252.

Wientjes, Y. C. J., P. Bijma, R. F. Veerkamp, and M. P. L. Calus, 2016 An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments. Genetics 202: 799–823.

Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten *et al.*, 2015 Empirical and deterministic accuracies of across-population genomic prediction. Genet. Sel. Evol. 47: 5.

Wientjes, Y. C. J., R. F. Veerkamp, and M. P. L. Calus, 2013 The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics 193: 621–631.

Wimmer, V., C. Lehermeier, T. Albrecht, H.-J. Auinger, Y. Wang *et al.*, 2013 Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection. Genetics 195: 573–587.

Windhausen, V. S., G. N. Atlin, J. M. Hickey, J. Crossa, J.-L. Jannink *et al.*, 2012 Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and Environments. G3 2: 1427–1436.

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

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–569.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H. P. Maurer *et al.*, 2012 Accuracy of genomic selection in European maize elite breeding populations. Theor. Appl. Genet. 124: 769–776.

Zhong, S., J. C. M. Dekkers, R. L. Fernando, and J.-L. Jannink, 2009 Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics 182: 355–364.

2. Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium

Schopp, P.[1a], D. Müller[1a], F. Technow[1] and A. E. Melchinger[1]

[1]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim,70593 Stuttgart, Germany

[a]These authors contributed equally

Genetics 205:1–14 (2017)

The original publication is available at

http://www.genetics.org

Abstract

Synthetics play an important role in quantitative genetic research and plant breeding, but few studies have investigated the application of genomic prediction (GP) to these populations. Synthetics are generated by intermating a small number of parents (N_P) and thereby possess unique genetic properties, which make them especially suited for systematic investigations of factors contributing to the accuracy of GP. We generated synthetics in silico from N_P = 2 to 32 maize (*Zea mays* L.) lines taken from an ancestral population with either short- or long-range linkage disequilibrium (LD). In eight scenarios differing in relatedness of the training and prediction sets and in the types of data used to calculate the relationship matrix (QTL, SNPs, tag markers, and pedigree), we investigated the prediction accuracy (PA) of Genomic best linear unbiased prediction (GBLUP) and analyzed contributions from pedigree relationships captured by SNP markers, as well as from cosegregation and ancestral LD between QTL and SNPs. The effects of training set size N_{TS} and marker density were also studied. Sampling few parents ($2 \leq N_P < 8$) generates substantial sample LD that carries over into synthetics through cosegregation of alleles at linked loci. For fixed N_{TS}, NP influences PA most strongly. If the training and prediction set are related, using $N_P < 8$ parents yields high PA regardless of ancestral LD because SNPs capture pedigree relationships and Mendelian sampling through cosegregation. As N_P increases, ancestral LD contributes more information, while other factors contribute less due to lower frequencies of closely related individuals. For unrelated prediction sets, only ancestral LD contributes information and accuracies were poor and highly variable for $N_P = 4$ due to large sample LD. For large NP, achieving moderate accuracy requires large N_{TS}, long-range ancestral LD, and high marker density. Our approach for analyzing PA in synthetics provides new insights into the prospects of GP for many types of source populations encountered in plant breeding.

3. Forecasting the accuracy of genomic prediction with different selection targets in the training and prediction set as well as truncation selection

Schopp, P.[1], C. Riedelsheimer[1], H. F. Utz[1], C.-C. Schön[2] and A. E. Melchinger[1]

[1]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim,70593 Stuttgart, Germany

[2]Plant Breeding, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Liesel-Beckmann-Straße 2, 85354 Freising, Germany

Theoretical and Applied Genetics 128:2189–2201 (2015)

The original publication is available at

http://link.springer.com

Abstract

Genomic prediction of testcross performance (TP) was found to be a promising selection tool in hybrid breeding as long as the same tester and environments are used in the training and prediction set. In practice, however, selection targets often change in terms of testers, target environments or traits leading to a reduced predictive ability. Hence, it would be desirable to estimate for given training data the expected decline in the predictive ability of genomic prediction under such settings by deterministic formulas that require only quantitative genetic parameters available from the breeding program. Here, we derived formulas for forecasting the predictive ability under different selection targets in the training and prediction set and applied these to predict the TP of lines based on line per se or testcross evaluations. On the basis of two experiments with maize, we validated our approach in four scenarios characterized by different selection targets. Forecasted and empirically observed predictive abilities obtained by cross-validation generally agreed well, with deviations between −0.06 and 0.01 only. Applying the prediction model to a different tester and/or year reduced the predictive ability by not more than 18 %. Accounting additionally for truncation selection in our formulas indicated a substantial reduction in predictive ability in the prediction set, amounting, e.g., to 53 % for a selected fraction $\alpha = 10$ %. In conclusion, our deterministic formulas enable forecasting the predictive abilities of new selection targets with sufficient precision and could be used to calculate parameters required for optimizing the allocation of resources in multi-stage genomic selection.

4. Genomic prediction within and across biparental families: means and variances in prediction accuracy and usefulness of deterministic equations

Schopp, P.[1], D. Müller[1], Y. C. J. Wientjes[2] and A. E. Melchinger[1]

[1]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim,70593 Stuttgart, Germany

[2]Animal Breeding and Genomics, Wageningen University and Research, 6700 AH, The Netherlands

G3 7:3571–3586 (2017)

The original publication is available at

http://www.g3journal.org

Abstract

A major application of genomic prediction (GP) in plant breeding is the identification of superior inbred lines within families derived from biparental crosses. When models for various traits were trained within related or unrelated biparental families (BPFs), experimental studies found substantial variation in prediction accuracy (PA), but little is known about the underlying factors. We used SNP marker genotypes of inbred lines from either elite germplasm or landraces of maize (*Zea mays* L.) as parents to generate in silico 300 BPFs of doubled-haploid lines. We analyzed PA within each BPF for 50

simulated polygenic traits, using genomic best linear unbiased prediction (GBLUP) models trained with individuals from either full-sib (FSF), half-sib (HSF), or unrelated families (URF) for various sizes (N_{train}) of the training set and different heritabilities (Embedded Image In addition, we modified two deterministic equations for forecasting PA to account for inbreeding and genetic variance unexplained by the training set. Averaged across traits, PA was high within FSF (0.41–0.97) with large variation only for $N_{train} < 50$ and $h^2 < 0.6$. For HSF and URF, PA was on average ~40–60% lower and varied substantially among different combinations of BPFs used for model training and prediction as well as different traits. As exemplified by HSF results, PA of across-family GP can be very low if causal variants not segregating in the training set account for a sizeable proportion of the genetic variance among predicted individuals. Deterministic equations accurately forecast the PA expected over many traits, yet cannot capture trait-specific deviations. We conclude that model training within BPFs generally yields stable PA, whereas a high level of uncertainty is encountered in across-family GP. Our study shows the extent of variation in PA that must be at least reckoned with in practice and offers a starting point for the design of training sets composed of multiple BPFs.

5. General Discussion

Synthetics: a conceptual framework to disentangle the information sources contributing to prediction accuracy in plant breeding

A central interest of research on GP has been to unravel the sources of quantitative-genetic information (herein briefly called ‘information sources’) contributing to PA (Habier *et al.* 2007, 2013; Zhong *et al.* 2009; de los Campos *et al.* 2013; Wientjes *et al.* 2013). The principle of the three information sources apply in general to all GP models (Habier *et al.* 2013), but it was shown that the relative importance of information sources can differ among models. While BLUP-based approaches focus more on capturing pedigree relationships, variable selection models like BayesB emphasize more on ancestral LD, *i.e.,* on exploiting tight historical QTL-marker associations spanning short genetic map distances (Habier *et al.* 2007; Wimmer *et al.* 2013).

Using GBLUP, Habier *et al.* (2013) showed that in both animal and plant breeding schemes, the majority of predictive information is contributed by capturing pedigree relationships and ancestral LD. Co-segregation, on the other hand, contributed only to a smaller extent. Although their analyses covered important scenarios such as half-sibs designs in cattle breeding (MacLeod *et al.* 2010) and multiple biparental families in maize breeding (Albrecht *et al.* 2011), the transferability of these findings to other breeding scenarios remained questionable. In particular, it was unclear whether their results generalize to populations derived from small numbers of parents.

In Schopp *et al.* (2017a), we addressed this question by treating synthetics as a conceptual framework to analyze how the number of parents affects PA and the three information sources contributing to it. Our results show that capturing available pedigree relationships in GP is in nearly all scenarios of utmost relevance, regardless of the number of parents (except for GP within biparental families). The presence of close pedigree relationships between training and precited individuals resulted in much higher PAs compared with scenarios in which individuals were unrelated, in agreement with previous studies (Habier *et al.* 2007, 2010, 2013; Clark *et al.* 2012; de los Campos *et al.* 2013; Wientjes *et al.* 2013). However, the key finding of our study was the continuous shift from co-segregation towards ancestral LD information as the number of parents was successively increased. We provided an explanation for why this is the case by attributing the relevance of co-segregation for small numbers of parents to the large amount of (what we termed) 'sample LD'. Sample LD was defined as the LD created by sampling few parents from the ancestral population, which together with ancestral LD forms the LD that can be measured in the target population subjected to GP. From a population genetics' perspective, sample LD corresponds to LD generated by a strong bottleneck in recent effective population size (Hill 1981). We showed that sample LD is a specific property of the pedigree and thereby only informative for GP of progenitors of a given set of intermated parents, *e.g.,* for individuals belonging to the same synthetic as those used for model training. By comparison, only information from ancestral LD can be harnessed to predict individuals that are unrelated to the training set (Habier *et al.* 2013), which necessitates devoting increasing resources to sample size and marker density (Schopp *et al.* 2017a).

Factorizing actual and genomic relationships illustrates the role of identity-by-descent and identity-by-state relationships

To improve the interpretability of the three information sources, Schopp *et al.* (2017a) additionally presented the concept in a formal way. For this purpose, we used $q_{ij} = f_{ij} + t_{ij} + e_{ij}$ to factorize actual relationships at QTL (q_{ij}) between two individuals i and j into their expected identity-by-descent (IBD) relationship (*i.e.,* their pedigree relationship f_{ij}), the deviation ($t_{ij} = m_{ij} - f_{ij}$) between actual IBD relationship (m_{ij}) and expected IBD relationship (f_{ij}) due to Mendelian sampling, and their 'residual' genetic relationship (e_{ij}), which captures the genetic similarity (*i.e.,* identity-by-state (IBS) relationship) among their nominally unrelated parents in the ancestral population. A central prerequisite of GP, and explicitly of GBLUP, is to estimate actual relationships at QTP q_{ij} by marker-derived genomic relationships g_{ij} (Goddard *et al.* 2011; de los Campos *et al.* 2013; Habier *et al.* 2013). Factorizing genomic relationships g_{ij} in the same way yields $g_{ij} = f_{ij} + t_{ij}^{*} + e_{ij}^{*}$, where the expected IBD f_{ij} is identical to that assumed for q_{ij}, t_{ij}^{*} denotes the deviations between the actual and expected IBD relationship, which is expected to be captured by markers via co-segregation of QTL and markers, and e_{ij}^{*} denotes the residual genetic similarity (IBS) among individuals, which is expected to be captured by markers via ancestral LD between QTL and markers. It shall be emphasized that $f_{ij} + t_{ij}$ is only different from zero if individuals are related by pedigree with respect to an unrelated base population, corresponding in our terminology to the ancestral population. While f_{ij} can be calculated from pedigree records, the ('true') actual IBD relationships (*i.e.,* $m_{ij} = f_{ij} + t_{ij}$) can be obtained from simulations using so called *tag* markers (*see* Schopp *et al.* 2017a *for details*), which mark (tag) the parental origin of QTL alleles.

The same can be done for markers to obtain $f_{ij} + t_{ij}^{*}$. However, methods exist to estimate IBD relationships from marker data for practical applications (Meuwissen *et al.* 2011; Luan *et al.* 2012).

In additional simulations, we demonstrated the strength of linear relationships among q_{ij} and g_{ij} (expressed as their R^2 values), as well as among their subfactors, as a function of the number of parents. The results of these simulations are shown in Figure 1 and extend the results presented in Schopp *et al.* (2017a). We found that the capability of g_{ij} to explain variation in q_{ij} in synthetics peaks for $N_P = 4$ parents (Figure 1A). The factorization suggests that this is mainly a result of the strong variation in actual IBD relationships ($f_{ij} + t_{ij}$) in genetically narrow populations, which explains a large proportion of the variance in actual (IBS) relationships (q_{ij}). These actual IBD relationships can be efficiently captured by markers ($f_{ij} + t_{ij}^{*}$), even under low marker density (*e.g.,* 0.25 tag markers/cM, Figure 1A). The difference in R^2 between f_{ij} and $f_{ij} + t_{ij}$ with q_{ij}, respectively, underlines unambiguously the central role of deviations between actual IBD and expected IBD relationships in explaining variation in actual (IBS) relationships at QTL (q_{ij}). Increasing the number of parents resulted in declining importance of IBD information (f_{ij} and $f_{ij} + t_{ij}$, Figure 1A) and increasing importance of genetic similarity among parents (IBS) in the ancestral population (e_{ij}). This is attributable to a lower frequency of closely related individuals in the target population when deriving it from a large number of parents (*e.g.,* $N_P \geq 16$), which can only partially be captured by ancestral IBS information, *i.e.,* ancestral LD between QTL and markers (e_{ij}^{*}). Capturing IBS instead of only IBD information by markers (g_{ij} vs. $f_{ij} + t_{ij}^{*}$) can increase R^2 substantially, but only if marker density is high (Figure 1A). Here, it must be noted that in our example, the strong contribution of ancestral

IBS information to IBS information in the target population is a consequence of (i) the assumed long-range ancestral LD that caused considerable variation in IBS relationships among parents and (ii) the definition of a recent base population. This shows that the importance of IBD must be judged relative to that definition, and that capturing ancestral IBS by markers is particularly important if the ancestral population corresponds to a recent generation, *e.g.,* the parents of a synthetic. As pedigree records are often not available for several past generations in plant breeding, IBS information takes on an important role in GP compared with previous examples from animal breeding (Luan *et al.* 2012).

Comparing the associations (R^2) among subfactors of q_{ij} and g_{ij} shows that (i) $f_{ij} + t_{ij}$ is explained almost entirely by $f_{ij} + t_{ij}^*$ (Figure 1B), underlining that IBD relationships can be efficiently tracked by markers, even under low marker density, (ii) e_{ij}^* can only explain moderate amounts of e_{ij} due to incomplete ancestral LD between QTL and markers, even under high marker density and (iii) the terms $f_{ij} + t_{ij}$ and $f_{ij} + t_{ij}^*$ are uncorrelated with e_{ij} and e_{ij}^*, respectively, which shows that IBD and ancestral IBS contribute independently to the approximation of actual (IBS) relationships by genomic relationships in the target population.

This factorization experiment illustrates that the three information sources act together in a non-additive way to explain variation in actual relationships at QTL by marker-derived genomic relationships. Since actual relationships are in reality unknown, knowledge on how markers enable approximating the true relationship matrix in GBLUP (and equivalently parameterized also by all other GP models) are crucial for understanding and evaluating the merit of GP in many different scenarios.

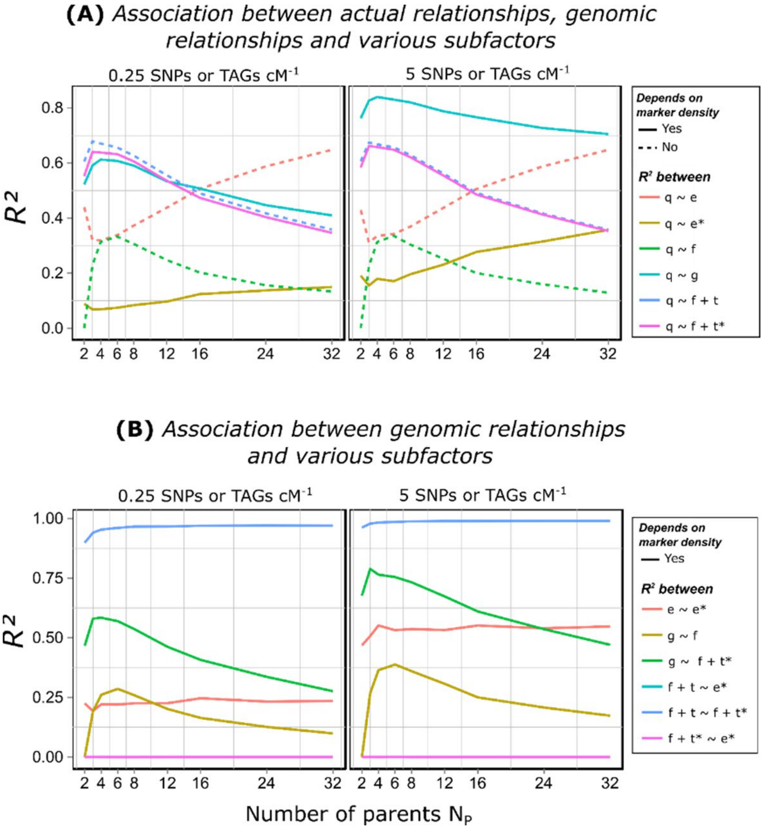

Figure 1 Coefficients of determination (R^2) among actual (q_{ij}) and genomic relationships (g_{ij}) and various subfactors thereof (*see text for detailed description of the factorization*) in synthetic populations derived from different number of parents (N_P). Synthetics were simulated as described in detail in Schopp *et al.* (2017a) and parents were sampled from an ancestral population with long-range linkage disequilibrium. The line type indicates whether an association depended on marker density (solid) or not (dashed).

The role of the three information sources in plant breeding: possible extensions of the standard GBLUP model to multiple cycles or populations

The findings of Schopp *et al.* (2017a) have important consequences for the implementation of GP into various practically relevant scenarios. In the narrow sense, our results suggest that about 6 to 8 inbred parents are advisable to develop synthetics for the purpose of recurrent GS (*cf.* Windhausen *et al.* 2012; Gorjanc *et al.* 2016). This number arises as a consequence of balancing high PA in the model training generation due to extensive use of co-segregation with good persistency across cycles due to ancestral LD (Müller *et al.* 2017; Schopp *et al.* 2017a). Together with decent marker density (> 2.5 SNPs/cM) and sample size (>250 individuals), a prediction equation might be reusable for two to three cycles without retraining (Massman *et al.* 2013; Müller *et al.* 2017). Besides the aspects related to PA and its persistency, the number of parents also determines the additive-genetic variance available for selection (Hill 1971). A follow-up study of our group (Müller *et al.* 2017) confirmed the recommended number of parents also in terms of cumulative genetic gain across several cycles of recurrent GS, which resulted from an optimal balance between initial PA, persistency of PA and additive-genetic variance. Altogether, our work promises high potential for the use of ‘rapid cycle’ GS (Windhausen *et al.* 2012), which was specifically proposed to increase genetic gains per unit time (Bernardo and Yu 2007).

The insights obtained in Schopp *et al.* (2017a) also extend to other relevant breeding scenarios, such as GP across multiple biparental families (BPF). While in GP within a single BPF only co-segregation is utilized, PA across multiple BPFs stems from all three information sources (Habier *et al.* 2013; Schopp *et al.* 2017a). However, in contrast to synthetics produced by random mating, the availability of a clear pedigree structure raises the question of whether it might be beneficial to model co-segregation, and thus IBD information, explicitly (Habier *et al.* 2013). In the standard linear mixed model framework, this can be done by separating IBD from IBS relationships (Luan *et al.* 2012; Vela-Avitúa *et al.* 2015). Alternatively, hierarchical Bayesian models that differentially weigh co-segregation (IBD) and ancestral LD information (IBS) were proposed (Technow and Totir 2015). However, these studies found little to no benefit from explicitly modelling IBD when comparing the PA with that of GBLUP, which implicitly integrates all three information sources simultaneously. As discussed above, the results of such model comparisons crucially depend on the depth of the pedigree (Habier *et al.* 2013), *i.e.,* on how many generations in the past the base population is defined. The further back, the more information is captured by co-segregation relative to ancestral LD, and the less relevant ancestral IBS information becomes in comparison with IBD information. Apart from that, explicit IBD models seem to be advantageous mainly if ancestral LD is weak (Luan *et al.* 2012), which can be caused by strong differences in the minor allele frequencies at QTL and markers, insufficient marker density, or large ancestral effective population size (Muir 2007; Solberg *et al.* 2008; Heslot *et al.* 2013). Further research must show whether there are plant breeding scenarios in which explicit modelling of co-segregation can be truly advantageous. One such scenario might be the above-mentioned use of recurrent GS, where differential weighing of IBD and IBS information could enhance the persistency of PA across repeated cycles of selection and recombination without retraining (Habier *et al.* 2013; Müller *et al.* 2017).

Similar to the considerations on explicit modelling of information sources, our results can provide guidelines for identifying suitable GP models in the narrow-sense (*e.g.* GBLUP vs. BayesB). As demonstrated by Wimmer *et al.* (2013), variable selection models like BayesB cannot utilize their theoretical potential if strong LD prevails in the target population. Hence, the large contributions of co-segregation and pedigree relationships to PA as found by Schopp *et al.* (2017a) in synthetics, together with the generally prevalent long-range LD in advanced breeding germplasm (Flint-Garcia *et al.* 2003; Van Inghelandt *et al.* 2011) underline the small benefit that can be expected from variable selection models in most plant breeding scenarios. Exceptions might be encountered if ancestral LD is weak (as discussed above), for example in largely unselected germplasm, such as landraces. However, efficient GP in such populations simultaneously requires high marker densities and large sample sizes (Schopp *et al.* 2017a). While the former will be increasingly feasible in the future (Pérez-Enciso *et al.* 2015), the latter may lie beyond of what is realistic in practical plant breeding.

Further examples related to our analyses include the augmentation of training data with unrelated genotypes as investigated by Technow *et al.* (2013). They found that PA could be increased by using augmented training sets spanning across heterotic groups, given good persistency of QTL-marker LD and linkage phases across these groups. Transferred to our framework, this strategy attempts to exploit ancestral LD information of an assumed common and distant founder population for the two heterotic groups. The approach becomes beneficial likely because the mandatory increase in sample size to efficiently utilize this more 'distant' ancestral LD – which is typically lower than the LD within groups (Technow *et al.* 2013) – is automatically assured by combining populations.

The variety of examples in plant breeding to which our findings can be applied are manifold and could only partially be elaborated in Schopp *et al.* (2017a) as well as in this discussion. Our work show that the relative importance of the three information sources is highly context-specific and not generalizable from the few specific examples presented in the literature (Wientjes *et al.* 2013; Habier *et al.* 2013). In conclusion, each prediction scenario must be evaluated individually in terms of appropriate modelling, as well as in terms of the required resources (sample size, marker density, etc.) with respect to the prevailing genetic preconditions. I believe that an enhanced understanding about the role of the three information sources in plant breeding further motivates the development of improved prediction models, mating schemes and training set designs for GP.

Implementing genomic prediction into multi-stage selection in hybrid breeding

Hybrid breeding aims at developing inbred lines showing superior testcross performance (TP) with testers from an opposite heterotic group (Sprague 1983). For efficient selection of lines with superior general combining ability (GCA) as well as specific combining ability, breeders employ the concept of multi-stage selection (Cochran 1951; Utz 1969). Here, an initially large number of candidate lines is first evaluated for their TP with one or few testers in a small sample of test environments. In subsequent stages, a reduced number of candidates is tested more extensively in evaluations typically including multiple testers and test environments. This allows also for a potential identification of superior inter-group combinations and release them as hybrid cultivars (Hallauer *et al.* 2010). In early stages, however, the estimation of GCA can be aggravated by the significance of genotype by environment (Van Eeuwijk 2006) and genotype by tester interactions (Longin *et al.* 2007). Therefore, balancing the allocation of resources among selection stages has been a main focus of plant breeding research for decades (Utz 1969; Schnell 1996; Longin *et al.* 2006; Mi *et al.* 2014). Despite being previously mentioned and experimentally addressed (Windhausen *et al.* 2012; Albrecht *et al.* 2014), the influence of using different testers and test environments in the course of multi-stage selection has not been examined theoretically in the context of GP. Hereafter, different testers and test environments, as well as TP and *per se* performance of lines will be referred to as different ‘selection targets’.

In Schopp *et al.* (2015), we provided a theoretical framework to deterministically forecast the predictive ability (*i.e.,* the correlation between phenotypic values and predicted genetic values) when training and prediction set represent different selection targets. Compared with GP within the same selection target (*cf.* Albrecht *et al.* 2011; Zhao *et al.* 2012), the predictive ability across selection targets is expected to decline in the presence of strong 'genotype by selection target interactions' (Windhausen *et al.* 2012; Albrecht *et al.* 2014). For instance, calibrating a prediction equation with data on TP of a set of lines based only on a single tester might hamper the transferability of predicted genetic values to another set of lines, if the goal is to predict their TP with a different tester or GCA in general. This is particularly the case if genotype by tester interactions explain a considerable amount of the phenotypic variance, *i.e.,* if the specific combining ability variance is large.

The derivation of predictive ability across selection targets in Schopp *et al.* (2015) was based on extensions of Dekkers' (2007) application of selection theory to marker-assisted selection. We demonstrated that under certain, well-justified assumptions, the predictive ability across selection targets is approximately equal to the product of the predictive ability evaluated within a selection target (*e.g.,* training and predicted individuals evaluated with the same tester) and the genetic correlation between the selection targets. The derivation for this formula rests on the assumption that the phenotypic performance of candidates evaluated for different selection targets can be considered as correlated traits (Falconer and Mackay 1996). While the factors influencing predictive ability within selection targets (*e.g.,* sample size, heritability, effective number of chromosome segments) are discussed elsewhere (*e.g.,* Daetwyler *et al.* 2010; Habier *et al.* 2013) as well as in other parts of this thesis, Schopp *et al.* (2015) formally showed that the transferability of a prediction equation to different testers or test environments can be described adequately by the genetic correlation between the original and the new selection target.

To demonstrate this relationship in practically relevant prediction scenarios, Schopp *et al.* (2015) derived genetic correlations between various selection targets (*e.g.*, different testers, different years as well as between line *per se* and testcross performance). Subsequently, we validated the formulae based on estimated genetic correlations and predictive abilities across the exemplarily considered selection targets obtained from a fully balanced dataset (all lines evaluated on both testers and in both years). Overall, we found good agreement between observed and forecasted predictive abilities, with deviations ranging only from -0.06 to 0.01.

Genetic correlation between selection targets – a central parameter to plant breeders

From a broader point of view, our formulae for forecasting predictive ability are an intuitive extension of what has long been used to evaluate the observed phenotypic performance for specific selection targets in the course of multi-stage selection. A breeder's goal is to evaluate a population of candidate lines under a fixed budget for an 'ultimate' selection target, which might be the candidates' 'true' general combining ability in a predefined target population of environments (Löffler *et al.* 2005). Due to limited testing capacity, the entirety of ultimate selection targets cannot be observed, and breeders must judge the available phenotypic evaluations in terms of their genetic correlation with the ultimate selection target. A practical example are genetic correlations between testers, which are commonly moderate to high (Bernardo 1991; Melchinger *et al.* 1998) owing to the minor role of specific combining ability in hybrids produced between lines from advanced heterotic groups (Reif *et al.* 2007; Technow *et al.* 2012). This knowledge has been used to justify selection on TP based on only one to few testers in early phases of multi-stage selection (Schnell 1996), because high genetic correlations among testers assure a certain transferability of observed TP to later stages, which potentially involve different or additional testers. Alternatively, breeders have been using hybrid testers to increase the effective number of included tester lines and improve GCA estimation (Longin *et al.* 2007). Another example are the at most medium genetic correlations between TP and line *per se* performance for complex traits such as grain yield (Melchinger *et al.* 1998; Mihaljevic *et al.* 2005). This strengthened the recognition that selection based on line *per se* performance is not efficient regarding the selection gain of the ultimate selection target. Integrating this thesis' findings in that reasoning suggests that similar recommendations as for phenotypic selection apply to GP across different selection targets. This is because the relationship between the observed and the ultimate selection target(s) can, in both cases, be sufficiently described by the genetic correlations between these targets (Schopp *et al.* 2015).

The prospects of genomic prediction in biparental families

Variation in PA imposes uncertainty and therefore a risk on the implementation of GP in practice. Previous studies showed that variation in PA can be substantial even within specific prediction scenarios, *e.g.,* when predicting a certain biparental family (BPF) based on training data obtained from the same or distinct BPFs (Riedelsheimer *et al.* 2013; Lehermeier *et al.* 2014; Lian *et al.* 2014). However, these studies were primarily concerned with the general influence of various factors (*e.g.,* sample size, heritability, pedigree relationship) on the expected PA. In Schopp *et al.* (2017b), we conducted the first extensive simulation study that systematically investigated the reasons determining the variation in PA in BPFs. Similar to most of the above cited studies, we considered for this purpose the simplistic case of GP within and across individual BPFs. Further, we analyzed two available deterministic formulae for forecasting PA and evaluated their usefulness in BPFs. These formulae were originally developed in the context of within-population GP in animal breeding (VanRaden 2008; Daetwyler *et al.* 2010; Wientjes *et al.* 2013) and were recently extended to across-population scenarios (Wientjes *et al.* 2015, 2016). Previous studies on plant breeding already used these formulae within populations (Lorenz 2013; Riedelsheimer *et al.* 2013; Riedelsheimer and Melchinger 2013; Lian *et al.* 2014, Akdemir *et al.* 2015; He *et al.* 2016), but ignored the fact that the original formulae do not account for inbreeding, despite being a central property of BPFs of inbred lines. Moreover, extensions to GP across BPFs have so far not been presented. Such extensions should account for family-specific polymorphisms, which can lead to genetic variance among predicted individuals that is not explained by the training set.

In Schopp *et al.* (2017b), we show that the variation in PA is generally negligible for GP within BPFs, except if sample size and/or heritability are small, as expected. By contrast, we found remarkable variation in PA for GP across BPFs, even under favorable prediction conditions, *i.e.,* large sample size, high heritability and connectedness of training and predicted individuals through a common parent. A causal analysis of GP across half-sib families showed that unexpectedly low PA can arise in the prediction of polygenic traits (1,000 QTL in our study), if there are many QTL that account for relatively large proportions of the genetic variance in the predicted BPF, which do not segregate in the BPF used for model training. We further show that deterministic formulae cannot account for these unobserved characteristics at QTL, because they merely utilize information from molecular markers and population parameters (VanRaden 2008; Daetwyler *et al.* 2010). This led to large deviations between observed and deterministic PAs for single traits. When PAs were averaged across traits, the trait-specific deviations due to different segregation QTL in the training and prediction set integrated out, which resulted in excellent agreement between observed and deterministic accuracies. We concluded that the PA expected across many traits can be forecasted precisely in the absence of phenotypic information for the predicted BPF.

Towards multifamily training sets

Our findings have important implications for commercial plant breeding programs, where the key role of GP lies in the identification of superior lines within BPFs, for which pedigree relationships cannot be exploited (Riedelsheimer *et al.* 2013; Schopp *et al.* 2017a). As discussed previously (Crossa *et al.* 2010; Lehermeier *et al.* 2014), GP within BPFs is a robust approach with the potential to provide high PA, but this requires phenotypic characterization of a sufficient number of training individuals (>50) from the targeted BPFs. Such numbers might be achievable in large BPFs of doubled-haploid lines in maize breeding, comprising often >100 individuals, but they are unrealistic in most other crops, such as small grain cereals (*e.g.,* wheat) (Heffner *et al.* 2011). For this and other reasons, alternative training set designs must be considered, and our study shows that among these, GP across single BPFs is certainly not a promising one. Although average PAs across traits were in a reasonable range when model training was based on half-sib or even unrelated families (~0.4 – 0.6), the variation among traits was in both cases tremendous and would be detrimental for practical applications. Hence, breeders should rely on multifamily training sets (Heffner *et al.* 2011; Albrecht *et al.* 2011; Jacobson *et al.* 2014; Lehermeier *et al.* 2014), which pool together available related and unrelated BPFs derived from the same germplasm. In addition to increasing sample sizes, these designs benefit particularly from an extensive use of germplasm-wide (ancestral) LD information (Habier *et al.* 2013; Technow and Totir 2015; Schopp *et al.* 2017a).

A further advantage of multifamily training sets is the possibility to employ them for predicting BPFs based on material evaluated in previous breeding cycles (*cf.* Jacobson *et al.* 2014; Auinger *et al.* 2016), *i.e.,* early prediction. This approach is expected to provide good PAs, because in multifamily training sets, the available ancestral LD information – which is known to persist well across cycles (Habier *et al.* 2013; Müller *et al.* 2017) – can be efficiently captured (Schopp *et al.* 2017a). Early prediction enables breeders to preselect individuals within and among BPFs prior to their phenotypic evaluation in time- and cost-consuming field trials (Marulanda *et al.* 2016). In my view, a promising strategy would be to combine preselection based on early predictions with subsequent updating of the prediction equation as soon as phenotypic data of closely related BPFs or even from the same BPF become available from the current breeding cycle. This would assure that recent IBD information (co-segregation) is optimally exploited in the updated predictions, based on which the final selection decision could be taken.

Additionally, multifamily training sets likely allow for a better agreement between observed and deterministic PAs than found for GP across individual BPFs. This can be expected based on the findings of Schopp *et al.* (2017b), who identified different QTL segregating across BPFs as the core problem causing variation in the observed PAs. By comparison, it is likely that the majority of QTL segregating in a predicted BPFs also segregate in at least some related BPFs included in a multifamily training set (Lehermeier *et al.* 2014). This is particularly the case if BPFs are derived from recycling a genetically narrow set of elite inbred lines as typically practiced in commercial breeding programs (Mikel and Dudley 2006).

Less optimism might be appropriate in scenarios where multifamily training sets are used to predict entirely unrelated BPFs, *e.g.,* if they are derived from exotic material for introgression purposes or from distinct heterotic groups. In a comparable scenario, Schopp *et al.* (2017a) found considerable variation in PA when predicting synthetics derived from two parents (corresponding to F_2-derived BPFs of inbred lines) with training data from unrelated multiparental synthetics. Following the arguments of previous studies (de Roos *et al.* 2008; Riedelsheimer *et al.* 2013), Schopp *et al.* (2017a) suspected that this was due to inconsistency of linkage phases across populations. However, the findings of Schopp *et al.* (2017b) suggest that QTL exclusively segregating in the predicted BPF might have been a driving factor for the variation in PA also in the case of unrelated synthetics. Thus, a large overlap of QTL as well as markers segregating in both the training and prediction set, together with good consistency of linkage phases seem to be the prerequisites for robust predictions across BPFs, which likely also enables to accurately forecast PA by deterministic formulae.

Extensions to deterministic formulae: potential and limits in more complex scenarios

Extensions to different selection targets

Schopp *et al.* (2015) pointed out that extensions of the formulae to forecast PA across different selection targets, in addition to those considered in the study, are straightforward. For instance, one could easily derive the decline in predictive ability when model training is based on a certain sample of test locations, which are different from or a subset of the set of actual target locations. In fact, we emphasized that despite the encouraging results, further research must confirm the usefulness of the formulae in a broader range of scenarios, particularly in such characterized by stronger genotype by selection target interactions. I believe that for this purpose, an application of the formulae to a range of diverse locations could serve as a valuable benchmark, because this represents a scenario highly relevant to plant breeders that often suffers from low genetic correlations between selection targets due to strong genotype by environment interactions (Löffler *et al.* 2005).

Validation under truncation selection

Since GS is still a relatively novel tool, breeders might see the need to validate GP-based selection decisions taken in a certain stage with phenotypic data obtained in subsequent stages. Schopp *et al.* (2015) used theoretical results from Cochran (1951) to show that in addition to the expected decline in predictive ability under changing selection targets, the correlation between observed and predicted phenotypic values is tremendously reduced by truncation selection. Stringent selection (*e.g.,* $\alpha = 10\%$) is typically applied in the first selection stage to confine the number of candidates for further testing. Therefore, it should not be surprising that the predictive abilities detected in such validations might be very low. In this case, one has to keep in mind that selection gain is function of selection accuracy, selection intensity and phenotypic variance in the population undergoing selection (Falconer and Mackay 1996). Consequently, the formulae of Cochran (1951) and Schopp *et al.* (2015) can be employed to perform reverse calculations, *i.e.,* determine the predictive ability prior to selection from that observed in a validation experiment after selection. Results from such calculations can provide further insights into the prospects of GS in multi-stage selection.

Estimating genetic correlations and the role of epistasis

In my view, further research on the refinement of the simplistic approach of Schopp *et al.* (2015) to estimate genetic correlations could be worthwhile. For instance, allowing for heterogeneous (genetic or error) variances among testers and environments might improve estimates of genetic correlations. Another possibility could be to model the covariance among selection targets explicitly, as opposed to implicitly based on phenotypic observations for all selection targets, as done by Schopp *et al.* (2015) and other previous studies (Burgueño *et al.* 2012; Lopez-Cruz *et al.* 2015). Explicit modelling could, for example, be based on information from marker-derived relationships among testers (similar to hybrid prediction, *cf.* Technow *et al.* 2014) or on environmental covariates (*cf.* Jarquín *et al.* 2014). In contrast to implicit modelling, explicit modelling of covariance structures allows for the prediction of different selection targets even if they are not observed (Malosetti *et al.* 2016). This principle has already been successfully applied to predict the performance of genotypes in entirely unobserved environments (Jarquín *et al.* 2014). From such predictions for specific selection targets (either observed or unobserved), one could approximate genetic correlations in a second step. A possible area of application are genetic correlations between observed and unobserved (future) years, which can never be estimated directly, but always contribute to genotype by selection target interaction variance. To overcome this problem, at least to some extent, one could derive estimates of genetic correlations among years either (i) implicitly from phenotypic observations of previous years or (ii) explicitly using historical records of weather data (*cf.* Tuberosa 2012; Heslot *et al.* 2013). Whether or not explicit modelling can be useful for estimating genetic correlations warrants further research.

Extensions of the deterministic formulae for forecasting the PA within and across BPFs presented by Schopp *et al.* (2017b) to multifamily training sets should be relatively straight forward. Our theoretical framework already accounted for population-specific allele frequencies, but further research should examine the influence of population-specific heritabilities and the consideration of incomplete QTL-marker LD on forecasted PAs (Lorenz and Smith 2015; Wientjes *et al.* 2016). In addition, it remains to be investigated to which degree incomplete genetic correlations among BPFs restrict the usefulness of deterministic formulae (Wientjes *et al.* 2015). In Schopp *et al.* (2015, 2017b), we assumed that genetic correlations among training and prediction set were equal to 1, which basically implies that QTL effects (and thus, estimated marker effects) are constant across populations (Lehermeier *et al.* 2015). This is expected to hold in the absence of epistasis in general for TP and also for the *per se* performance of completely homozygous lines (*i.e.,* no dominance) (Melchinger 1987). Nevertheless, empirical and theoretical evidence for both considerable and negligible contributions of epistasis have been reported in the context of hybrid breeding in maize for several traits (Lamkey *et al.* 1995; Hinze and Lamkey 2003; Mihaljevic *et al.* 2005; Melchinger *et al.* 2007). If epistasis would indeed explain a considerable amount of the genetic variance for a given trait, genetic correlations between BPFs might be smaller than 1 and, thus, forecasted PAs might become inaccurate. In such cases, deterministic formulae should account for genetic correlations (Wientjes *et al.* 2015), but these are difficult to obtain for BPFs, because either (i) phenotypes of the predicted BPFs are not yet observed (early prediction) or (ii) sample sizes are too small for an accurate estimation. Therefore, I suggest that prior to any practical implementation, the usefulness of deterministic formulae should be validated carefully based on experimental data.

Using software for optimizing breeding schemes

Recently, software has been developed for optimizing the allocation of resources in multi-stage selection (Mi *et al.* 2014). The underlying algorithms are based on deterministic calculation of cumulative selection gain (Cochran 1951; Utz 1969; Longin *et al.* 2006). The software has also been used to evaluate the merit of integrating GS into existing and newly proposed multi-stage selection schemes (Longin *et al.* 2015; Marulanda *et al.* 2016). However, to the best of my knowledge, it does not (yet) consider the simultaneous use of phenotypic and genomic selection within a certain selection stage. Instead, it treats GS as a tool for preselection of genotypes based on *a priori* available training data, *i.e.,* early prediction. Despite early prediction being a promising alternative to implement GS into existing breeding schemes (Auinger *et al.* 2016; Marulanda *et al.* 2016), the complementary use of GS within stages of phenotypic multi-stage selection certainly demands further attention (Riedelsheimer and Melchinger 2013). Incorporating our theoretical results into optimization procedures should therefore facilitate advances towards a more efficient integration of GS into multi-stage selection.

Further potential for optimization of breeding programs using GS lies in incorporating computer simulations to complement the existing breeding methodology (Daetwyler *et al.* 2013). An example shall be given for employing deterministic formulae to evaluate the prospects of early predictions: Both formulae as implemented in Schopp *et al.* (2017b) require genotypic data of both the training and the predicted individuals, so they can only be used as soon as a new cross is produced and genotyped. However, for strategic planning of early predictions, knowing the expected PA beforehand would be highly advantageous. For example, if the forecasted PA would be high, breeders could renounce phenotyping of this cross in the upcoming season and advance lines with superior predicted breeding values directly to the next selection stage (*e.g.,* the second testcross stage). Conversely, if the forecasted PA would be low, one could (i) change the strategy towards GP based on newly generated training data from the current cycle, which can additionally be complemented by partially phenotyping the target cross or (ii) rely entirely on phenotypic selection. Computer simulation can assist breeders in gaining access to the necessary information, *e.g.,* by generating *in silico* BPFs based on the marker profiles of the parents of a planned cross and linkage map information, using available software packages like ‘Meiosis’ (Müller and Broman 2017). If PA is supposed to be forecasted across breeding cycles, which in addition to new genetic material also involve different selection targets (*e.g.,* testers), the formulae could be further extended using the derivations of Schopp *et al.* (2015). A comparable simulation-driven approach was recently used by Mohammadi *et al.* (2015) to predict the genetic segregation variance within BPFs, which has been a major goal to plant breeders for decades. Beyond this, the possibilities for applying computer simulation are basically limitless (Daetwyler *et al.* 2013) and, in my view, will play an increasingly central role in the design, comparison and evaluation of entire breeding schemes.

Conclusions

The results of this thesis demonstrate the potential and limits of GP in populations typically encountered in plant breeding and give detailed insights into the behavior of prediction accuracy in various practically relevant scenarios. The main conclusions of this thesis are

i. The relative contributions of the three information sources to prediction accuracy are context-specific and differ strongly between genetically narrow and diverse breeding populations.

ii. Knowledge about the prevailing information sources affecting prediction accuracy can provide guidelines for choosing adequate prediction models, constructing suitable training sets and optimizing breeding schemes.

iii. Prediction accuracy is expected to decline if the training and prediction set constitute different selection targets.

iv. The reduction in prediction accuracy across selection targets can be described by the genetic correlation between these targets, which can be estimated from parameters routinely generated in breeding programs.

v. GP within biparental families is a promising approach that provides high and robust prediction accuracies, whereas GP across individual biparental families is prone to substantial variation in prediction accuracy leading to unexpected selection outcomes.

vi. Multiparental training set designs should be favored in practice for increasing sample sizes, efficiently exploiting both co-segregation and ancestral linkage disequilibrium information and allowing for transferability of prediction equations to different breeding cycles.

vii. Deterministic formulae for forecasting prediction accuracy require modifications when applied to biparental families of inbred lines, but can provide accurate estimates that together with computer simulations open new gateways in optimizing breeding schemes integrating genomic selection.

Literature cited

Albrecht, T., H.-J. Auinger, V. Wimmer, J. O. Ogutu, C. Knaak *et al.*, 2014 Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor. Appl. Genet. 127: 1375–1386.

Albrecht, T., V. Wimmer, H. Auinger, M. Erbe, C. Knaak *et al.*, 2011 Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123: 339–350.

Auinger, H.-J., M. Schönleben, C. Lehermeier, M. Schmidt, V. Korzun *et al.*, 2016 Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.). Theor. Appl. Genet.

Bernardo, R., 1991 Correlation between testcross performance of lines at early and late selfing generations. Theor. Appl. Genet. 82: 17–21.

Bernardo, R., and J. Yu, 2007 Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci. 47: 1082–1090.

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

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. J. van der Werf, 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.

Cochran, W. G., 1951 Improvement by means of selection. Proc. 2nd Berkeley Symp Math Stat Prob 449–470.

Crossa, J., G. D. L. Campos, P. Pérez, D. Gianola, J. Burgueño *et al.*, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–24.

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

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams, 2010 The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021–1031.

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen, 2013 Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genet. 9: 7.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard, 2008 Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179: 1503–1512.

Dekkers, J. C. M., 2007 Prediction of response to marker-assisted and genomic selection using selection index theory. Genetics 124: 331–341.

Falconer, D. F., and T. S. C. Mackay, 1996 *Introduction to Quantitative Genetics* (1996 Longman, Ed.). Pearson, Essex.

Flint-Garcia, S. A., J. M. Thornsberry, and E. S. Buckler, 2003 Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54: 357–74.

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen, 2011 Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128: 409–421.

Gorjanc, G., J. Jenko, S. J. Hearne, and J. M. Hickey, 2016 Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17: 30.

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

Habier, D., R. L. Fernando, and D. J. Garrick, 2013 Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics 194: 597–607.

Habier, D., J. Tetens, F. Seefried, P. Lichtner, and G. Thaller, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42: 5.

Hallauer, A. R., M. J. Carena, and J. M. Filho, 2010 *Quantitative genetics in maize breeding.* Springer New York.

Heffner, E. L., J. Jannink, and M. E. Sorrells, 2011 Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. Plant Genome 4: 65–75.

Heslot, N., D. Akdemir, M. E. Sorrells, and J. L. Jannink, 2013 Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor. Appl. Genet. 1–18.

Heslot, N., J. Rutkoski, J. Poland, J. L. Jannink, and M. E. Sorrells, 2013 Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. PLoS One 8: e74612.

Hill, R. R., 1971 Effect of the Number of Parents on the Mean and Variance of Synthetic Varieties 1. Crop Sci. 11: 283–286.

Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. 38: 209–216.

Hinze, L., and K. Lamkey, 2003 Absence of epistasis for grain yield in elite maize hybrids. Crop Sci. 37: 46–56.

Van Inghelandt, D., J. C. Reif, B. S. Dhillon, P. Flament, and A. E. Melchinger, 2011 Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. Theor. Appl. Genet. 123: 11–20.

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo, 2014 General combining ability model for genomewide selection in a biparental cross. Crop Sci. 54: 895–905.

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt *et al.*, 2014 A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127: 595–607.

Lamkey, K. R., B. J. Schnicker, and A. E. Melchinger, 1995 Epistasis in an Elite Maize Hybrid and Choice of Generation for Inbred Line Development. Crop Sci. 35: 1272–1281.

Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan *et al.*, 2014 Usefulness of multi-parental populations of maize (Zea mays L.) for genome-based prediction. Genetics 198: 3–16.

Lehermeier, C., C.-C. Schön, and G. de los Campos, 2015 Assessment of Genetic Heterogeneity in Structured Plant Populations Using Multivariate Whole-Genome Regression Models. Genetics 201: 323–337.

Lian, L., A. Jacobson, S. Zhong, and R. Bernardo, 2014 Genomewide prediction accuracy within 969 maize biparental populations. Crop Sci. 54: 1514–1522.

Löffler, C. M., J. Wei, T. Fast, J. Gogerty, S. Langton *et al.*, 2005 Classification of maize environments using crop simulation and geographic information systems. Crop Sci. 45: 1708–1716.

Longin, C. F. H., X. Mi, and T. Würschum, 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.

Longin, C. F. H., H. F. Utz, A. E. Melchinger, and J. C. Reif, 2007 Hybrid maize breeding with doubled haploids: II. Optimum type and number of testers in two-stage selection for general combining ability. Theor. Appl. Genet. 114: 393–402.

Longin, C. F. H., H. F. Utz, J. C. Reif, W. Schipprack, and A. E. Melchinger, 2006 Hybrid maize breeding with doubled haploids: I. One-stage versus two-stage selection for testcross performance. Theor. Appl. Genet. 112: 903–912.

Lopez-Cruz, M., J. Crossa, D. Bonnett, S. Dreisigacker, J. Poland *et al.*, 2015 Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3 5: 569–582.

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

Lorenz, A. J., and K. P. Smith, 2015 Adding genetically distant individuals to training populations reduces genomic prediction accuracy in Barley. Crop Sci. 55: 2657–2667.

Luan, T., J. A. Woolliams, J. Ødegård, M. Dolezal, S. I. Román-Ponce *et al.*, 2012 The importance of identity-by-state information for the accuracy of genomic selection. Genet. Sel. Evol. 44: 28.

MacLeod, I. M., B. J. Hayes, K. W. Savin, A. J. Chamberlain, H. C. McPartlan *et al.*, 2010 Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. J. Anim. Breed. Genet. 127: 133–142.

Malosetti, M., D. Bustos-Korts, M. P. Boer, and F. A. Van Eeuwijk, 2016 Predicting responses in multiple environments: Issues in relation to genotype x environment interactions. Crop Sci. 56: 2210–2222.

Marulanda, J. J., X. Mi, A. E. Melchinger, J. L. Xu, T. Würschum *et al.*, 2016 Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. Theor. Appl. Genet. 129: 1901–1913.

Massman, J. M., H.-J. G. Jung, and R. Bernardo, 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.

Melchinger, A. E., 1987 Expectation of means and variances of testcrosses produced from from F2 and backcross individuals and their selfed progenies. Heredity (Edinb). 59: 105–115.

Melchinger, A. E., H. F. Utz, H. P. Piepho, Z. B. Zeng, and C.-C. Schön, 2007 The role of epistasis in the manifestation of heterosis: A systems-oriented approach. Genetics 177: 1815–1825.

Melchinger, A. E., H. Utz, and C.-C. Schön, 1998 Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in. Genetics 149: 383–403.

Meuwissen, T. H. E., T. Luan, and J. A. Woolliams, 2011 The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J. Anim. Breed. Genet. 128: 429–439.

Mi, X., H. F. Utz, F. Technow, and A. E. Melchinger, 2014 Optimizing Resource Allocation for Multistage Selection in Plant Breeding with R Package Selectiongain. Crop Sci. 54: 1413–1418.

Mihaljevic, R., C.-C. Schön, H. F. Utz, and A. E. Melchinger, 2005 Correlations and QTL correspondence between line per se and testcross performance for agronomic traits in four populations of European maize. Crop Sci. 45: 114–122.

Mihaljevic, R., H. F. Utz, and A. E. Melchinger, 2005 No Evidence for Epistasis in Hybrid and Per Se Performance of Elite European Flint Maize Inbreds from Generation Means and QTL Analyses. Crop Sci. 45: 2605–2613.

Mikel, M. A., and J. W. Dudley, 2006 Evolution of North American dent corn from public to proprietary germplasm. Crop Sci. 46: 1193–1205.

Mohammadi, M., T. Tiede, and K. P. Smith, 2015 Popvar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. Crop Sci. 55: 2068–2077.

Muir, W. M., 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.

Müller, D., and K. W. Broman, 2017 Meiosis: Simulation of Meiosis in Plant Breeding Research. R Packag. version 1.0.0.

Müller, D., P. Schopp, and A. E. Melchinger, 2017 Persistency of prediction accuracy and genetic gain in synthetic populations under recurrent genomic selection. G3 7: 801–811.

Pérez-Enciso, M., J. C. Rincón, and A. Legarra, 2015 Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 47: 43.

Reif, J. C., F.-M. Gumpert, S. Fischer, and A. E. Melchinger, 2007 Impact of interpopulation divergence on additive and dominance variance in hybrid populations. Genetics 176: 1931–1934.

Riedelsheimer, C., J. B. Endelman, M. Stange, M. E. Sorrells, J. L. Jannink *et al.*, 2013 Genomic predictability of interconnected biparental maize populations. Genetics 194: 493–503.

Riedelsheimer, C., and A. E. Melchinger, 2013 Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor. Appl. Genet. 126: 2835–2848.

Schnell, F., 1996 Über Zuchtplanung und die Entscheidungsspielräume des Pflanzenzüchters. Vor. über Pflanzenzüchtung 33: 227–244.

Schopp, P., D. Müller, F. Technow, and A. E. Melchinger, 2017a Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness and ancestral linkage disequilibrium. Genetics 205: 1–14.

Schopp, P., D. Müller, Y. C. J. Wientjes, and A. E. Melchinger, 2017b Genomic prediction within and across biparental families: means and variances in prediction accuracy and usefulness of deterministic equations. G3: *In review*

Schopp, P., C. Riedelsheimer, H. F. Utz, C.-C. Schön, and A. E. Melchinger, 2015 Forecasting the accuracy of genomic prediction with different selection targets in the training and prediction set as well as truncation selection. Theor. Appl. Genet. 128:.

Solberg, T. R., A. K. Sonesson, J. a Woolliams, and T. H. E. Meuwissen, 2008 Genomic selection using different marker types and densities. J. Anim. Sci. 86: 2447–2454.

Sprague, G. F., 1983 Heterosis in Maize: Theory and Practice, pp. 47–70 in *Heterosis: Reappraisal of Theory and Practice*, Springer Berlin Heidelberg.

Technow, F., A. Bürger, and A. E. Melchinger, 2013 Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3 3: 197–203.

Technow, F., C. Riedelsheimer, T. Schrag, and A. E. Melchinger, 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., T. Schrag, W. Schipprack, E. Bauer, H. Simianer *et al.*, 2014 Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. Genetics 197:1343-1355

Technow, F., and L. R. Totir, 2015 Using Bayesian Multilevel Whole Genome Regression Models for Partial Pooling of Training Sets in Genomic Prediction. G3 5: 1603–1612.

Tuberosa, R., 2012 Phenotyping for drought tolerance of crops in the genomics era. Front. Physiol. 3: 1–26.

Utz, H. F., 1969 *Mehrstufenselektion in der Pflanzenzüchtung*. Verlag Eugen Ulmer Stuttgart.

Van Eeuwijk, F., 2006 *Genotype by Environment Interaction — Basics and Beyond*. Plant Breeding: The Arnel R. Hallauer International Symposium.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Vela-Avitúa, S., T. H. Meuwissen, T. Luan, and J. Ødegård, 2015 Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. Genet. Sel. Evol. 47: 9.

Wientjes, Y. C. J., P. Bijma, R. F. Veerkamp, and M. P. L. Calus, 2016 An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments. Genetics 202: 799–823.

Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten *et al.*, 2015 Empirical and deterministic accuracies of across-population genomic prediction. Genet. Sel. Evol. 47: 5.

Wientjes, Y. C. J., R. F. Veerkamp, and M. P. L. Calus, 2013 The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics 193: 621–631.

Wimmer, V., C. Lehermeier, T. Albrecht, H.-J. Auinger, Y. Wang *et al.*, 2013 Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection. Genetics 195: 573–587.

Windhausen, V. S., G. N. Atlin, J. M. Hickey, J. Crossa, J.-L. Jannink *et al.*, 2012 Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and Environments. G3 2: 1427–1436.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H. P. Maurer *et al.*, 2012 Accuracy of genomic selection in European maize elite breeding populations. Theor. Appl. Genet. 124: 769–776.

Zhong, S., J. C. M. Dekkers, R. L. Fernando, and J.-L. Jannink, 2009 Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics 182: 355–364.

6. Summary

Genomic prediction (GP) is a novel statistical tool to estimate breeding values of selection candidates without the necessity to evaluate them phenotypically. The method calibrates a prediction model based on data of phenotyped individuals that were also genotyped with genome-wide molecular markers. The renunciation of an explicit identification of causal polymorphisms in the DNA sequence allows GP to explain significantly larger amounts of the genetic variance of complex traits than previous mapping-based approaches employed for marker-assisted selection. For these reasons, GP rapidly revolutionized dairy cattle breeding, where the method was originally developed and first implemented. By comparison, plant breeding is characterized by often intensively structured populations and more restricted resources routinely available for model calibration. This thesis addresses important issues related to these peculiarities to further promote an efficient integration of GP into plant breeding.

The accuracy of GP depends on three sources of quantitative-genetic information, which are harnessed by the prediction model to approximate the targeted characteristics at unknown causal polymorphisms by observable molecular markers. These information sources include pedigree relationships captured by markers, as well as co-segregation and ancestral linkage disequilibrium between causal polymorphisms and molecular markers. Gaining insights into the predominant information sources affecting prediction accuracy is crucial for choosing and developing suitable prediction models, as well as for optimizing calibration set designs and breeding strategies. However, previously reported results are context-specific and therefore not easily generalizable to all scenarios relevant in plant breeding.

A central step in hybrid breeding is to evaluate the selection candidates for their testcross performance with tester genotypes from an opposite heterotic group. This serves to assess their general heterotic performance for subsequent identification of superior, marketable hybrid combinations. For this purpose, breeders use multi-stage selection to successively narrow down an initially large number of selection candidates in several years of testing. Simultaneously, they increase the number of test environments and tester genotypes to mirror as representative as possible the targeted genetic and environmental characteristics. GP of testcross performance has been identified as an effective tool within selection stages, mainly because it allows for increasing the number of manageable individuals and, therefore, selection gain. The question remained how representative predicted genetic values are of subsequent selection stages, where tester genotypes and test environments may change in the course of multi-stage selection.

Identification of superior inbred lines within biparental families via GP is of central interest in commercial plant breeding. However, inconsistent prediction accuracies have been reported, which hampers the confidence in this new selection tool for practical application. Insights into the factors responsible for this variation and strategies for improvement are therefore urgently required. Moreover, deterministic formulae to forecast prediction accuracy prior to empirical evaluation were developed in animal breeding. Their use could greatly enhance the strategic planning of plant breeding programs integrating GP. However, the applicability of the formulae in biparental families of inbred lines requires further assessment.

Using computer simulation, we show that in synthetic populations generated from intermating a small number of parents, there is a shift of the predominant information sources harnessed in GP compared with previously examined scenarios characterized by larger effective population sizes. These findings have important implications in plant breeding for predicting both related and unrelated material, as well as for the allocation of resources and the design of more advanced breeding strategies.

Our results show that GP across different selection targets (*e.g.,* tester genotypes and years) results in lower prediction accuracies than within the same selection target. We demonstrate that the reduction in prediction accuracy across selection targets can be described by their genetic correlation. Validation of a novel deterministic formula to forecast prediction accuracy of altered selection targets resulted in good agreement with empirically observed values.

GP within biparental families was found to be a promising implementation strategy, whereas GP across individual biparental families is not recommended due to unstable prediction accuracies. Together with modified deterministic formulae to forecast prediction accuracy, the use of multifamily training populations for model calibration holds great promise to enhance the merit of GP in commercial plant breeding.

In conclusion, this thesis contributes important insights into the factors determining the accuracy of GP in plant breeding. Awareness of the idiosyncrasies of typical plant breeding populations is crucial for an increasingly efficient integration of GP into both existing and newly developed breeding schemes.

7. Zusammenfassung

Die genomische Leistungsvorhersage (GLV) ist ein neues statistisches Werkzeug zur Zuchtwertschätzung von Selektionskandidaten ohne die Notwendigkeit diese zuvor zu phänotypisieren. Die Methode kalibriert ein Vorhersagemodell auf Basis bereits phänotypisierter Individuen, welche zudem mit genomweiten molekularen Markern genotypisiert wurden. Der Verzicht auf die explizite Identifikation von kausalen Polymorphismen in der DNA-Sequenz ermöglicht der GLV signifikant größere Anteile der genetischen Varianz komplexer Merkmale zu erklären als frühere, kartierungsbasierte Ansätze zur markergestützten Selektion. Aus diesen Gründen revolutionierte die GLV bereits die Milchrinderzüchtung, in welcher die Methode ursprünglich entwickelt und auch erstmalig praktisch angewendet wurde. Im Vergleich hierzu zeichnet sich die Pflanzenzüchtung durch starke Populationsstruktur und stärker limitierte Ressourcen für den Zweck der Modellkalibration aus, welche in regelmäßigen Abständen zur Verfügung stehen. Die vorliegende Arbeit widmet sich wichtigen Fragen, die sich aus diesen Eigenschaften ergeben, um eine effizientere Integration der GLV in der Pflanzenzüchtung zu fördern.

Die Vorhersagegenauigkeit der GLV hängt von drei quantitativ-genetischen Informationsquellen ab, welche durch das Vorhersagemodell genutzt werden, um die Zielcharakteristika der unbekannten, kausalen Polymorphismen mit Hilfe beobachtbarer molekularer Marker anzunähern. Diese Informationsquellen umfassen die additiv-genetische Verwandtschaft, sowie die Ko-segregation und das den Vorfahren entstammende Kopplungsphasenungleichgewicht zwischen kausalen Polymorphismen und molekularen Markern. Tiefere Einblicke in die vorherrschenden Informationsquellen zu gewinnen, welche die Vorhersagegenauigkeit in

Pflanzenzüchtungspopulationen bestimmen, ist entscheidend, um angemessene Vorhersagemodelle zu ermitteln und zu entwickeln, sowie um die Zusammensetzung von Referenzpopulationen und Zuchtschemata zu optimieren. Die bislang berichteten Erkenntnisse sind jedoch kontext-spezifisch und daher nicht ohne Weiteres auf alle pflanzenzüchterisch relevanten Szenarien übertragbar.

Ein zentraler Schritt in der Hybridzüchtung ist die Evaluierung der Selektionskandidaten anhand ihrer Testkreuzungsleistung mit Testergenotypen einer entgegengesetzten heterotischen Gruppe. Dies dient dazu die generelle heterotische Leistungsfähigkeit zu bewerten, um im nachfolgenden Schritt überlegene, vermarktbare Hybridkombinationen zu identifizieren. Zu diesem Zweck verwenden Züchter die Mehrstufenselektion, um eine anfänglich große Anzahl Selektionskandidaten schrittweise in mehreren Testjahren einzuschränken. Hierbei wird gleichzeitig die Zahl der Testergenotypen und Testumwelten erhöht wird, um die anvisierten genetischen und umweltspezifischen Zieleigenschaften möglichst repräsentativ abzubilden. Die GLV der Testkreuzungsleistung wurde bereits als ein nützliches Werkzeug innerhalb einzelner Selektionsstufen identifiziert, besonders weil sie die Anzahl handhabbarer Individuen und dadurch den möglichen Selektionsgewinn erhöhen kann. Eine offene Frage ist jedoch, wie repräsentativ das innerhalb einer Stufe vorhergesagte genetische Leistungspotential für nachfolgende Selektionsstufen ist, in welchen sich sowohl Testergenotypen als auch Testumwelten im Zuge der Mehrstufenselektion ändern können.

Die Identifikation von überlegenen Inzuchtlinien innerhalb biparentaler Familien mittels der GLV ist von zentralem Interesse in der kommerziellen Pflanzenzüchtung. Es wurden jedoch bereits inkonsistente Vorhersagegenauigkeiten festgestellt, die das Vertrauen in dieses neue

Selektionswerkzeug hinsichtlich einer praktischen Anwendung beeinträchtigen. Weitere Erkenntnisse über die Faktoren, die diese Inkonsistenzen bedingen, sowie mögliche Verbesserungsstrategien, werden daher dringend benötigt. Darüber hinaus wurden im Rahmen der Tierzüchtung Formeln entwickelt, welche eine deterministische Prognose der Vorhersagegenauigkeit vor der eigentlichen empirischen Evaluierung liefern können. Deren Nutzen könnte die strategische Planung von Pflanzenzüchtungsprogrammen, welche die GLV integrieren, erheblich verbessern. Jedoch bedarf die Anwendbarkeit der Formeln in biparentalen Familien weiterer Beurteilung.

Unter Nutzung von Computersimulationen zeigen wir, dass in synthetischen Populationen, erzeugt mittels Durchkreuzung einer geringen Anzahl von Eltern, eine Verlagerung der in der GLV vornehmlich genutzten Informationsquellen hinsichtlich früherer untersuchter Szenarien größerer effektiver Populationsgrößen stattfindet. Diese Erkenntnisse haben in der Pflanzenzüchtung sowohl große Bedeutung für die Nutzung der GLV zur Vorhersage von verwandtem und unverwandtem Material, als auch für die Allokation von Ressourcen, sowie dem Design fortgeschrittener Zuchtstrategien.

Unsere Resultate zeigen, dass die GLV über Selektionsziele hinweg (z.B. über verschiedene Tester Genotypen oder über Jahre hinweg), verglichen mit der Anwendung innerhalb desselben Selektionsziels, zu einer reduzierten Vorhersagegenauigkeit führt. Wir zeigen auf, dass die Reduktion in der Vorhersagegenauigkeit über Selektionsziele hinweg mit Hilfe der genetischen Korrelation beschrieben werden kann. Die Validierung einer neu entwickelten deterministischen Formel zur Prognose der Vorhersagegenauigkeit geänderter Selektionsziele resultierte in guter Übereinstimmung mit den empirisch beobachteten Werten.

Die GLV innerhalb biparentaler Familien wurde als eine vielversprechende Anwendungsstrategie befunden, wohingegen die GLV über einzelne Familien hinweg aufgrund unsteter Vorhersagegenauigkeiten nicht empfohlen werden kann. Zusammen mit modifizierten deterministischen Formeln zur Prognose der Vorhersagegenauigkeit, scheint die Nutzung von Referenzpopulationen aus mehreren Familien zur Modellkalibration ein vielversprechender Ansatz zu sein, um den Nutzen der GLV in der kommerziellen Pflanzenzüchtung weiter zu erhöhen.

Zusammenfassend kann gesagt werden, dass diese Arbeit wichtige Erkenntnisse zu jenen Faktoren beiträgt, welche die Genauigkeit der GLV in für die Pflanzenzüchtung relevanten Szenarien beeinflussen. Das Bewusstsein über die Eigenheiten typischer Pflanzenzüchtungspopulationen ist hierbei essentiell, um eine zunehmend effiziente Integration der GLV sowohl in existente, als auch in neu entwickelte Zuchtschemata zu ermöglichen.

8. Acknowledgments

First of all, I would like to thank my academic supervisor and mentor Prof. Dr. Albrecht E. Melchinger for awakening my interest in plant breeding and for giving me the opportunity to do my PhD in his prestigious research group. I am especially grateful for his constant dedication, his deep enthusiasm for my research topics and his exceptional support during the transition towards my professional career in plant breeding.

Sincere thanks to Prof. Dr. Chris-Carolin Schön and Prof. Dr. Jörn Bennewitz for serving on my graduate committee. Special thanks also to Prof. Dr. Chris-Carolin Schön for her helpful comments and contributions to my publications.

Many thanks to my good colleague Dominik Müller for countless and inspiring discussions and constant mutual feedback.

I am grateful to Dr. Frank Technow for awakening my interest in genomic prediction, computer simulation and programming, and for giving me a unique perspective on scientific work.

I would also like to thank Prof. Dr. H.F. Utz and Dr. Tobias Schrag for their advice regarding statistical and software-related issues.

Many thanks to Mrs. Kösling and Mrs. Lieb for their constant help regarding organizational matters.

I am especially grateful for the financial support received from KWS Saat SE during my master studies. Special thanks also to Syngenta for partially funding my PhD and for the interesting discussions on practical breeding with Nicolas Ranc, Christoph Mainka and others.

Special thanks go to KWS Saat SE, particularly to Dr. Milena Ouzunova and Dr. Thomas Presterl for their flexibility and support during the final stages of my PhD.

Many thanks to all my colleagues and friends in the working group of plant breeding and applied genetics, Mathias Westhues, Juliane Böhm, Dominik Müller, Sen Han, Felix Sattler, Anna Bürger, Haixiao Hu, Moctar Kante, Willem Mollenar, Pedro Brauner, José Marulanda, Wilmar Leiser and all the other members of the institute of plant breeding, for creating a pleasant working atmosphere and for many hours of fun, both during work and after.

Last but not least, I want to thank my family for their constant and unconditional support and for encouraging me in difficult phases, whether personally or professionally. Without you, this would not have been possible. Thank you very much.

Curriculum vitae

NAME	Pascal Schopp
BIRTH	7 September 1987 in Breisach/Rhein
CURRENT POSITION	Research scientist, KWS Saat SE
SCHOOL	1995 – 1999, elementary school (Grund- und Hauptschule Merdingen)
	1999 – 2005, highschool (Martin-Schongauer Gymnasium Breisach)
	2005 – 2007, highschool (Heinrich-Heine Gymnasium Kaiserslautern) Abitur, March 2007
STAY ABROAD	2007, Work-and-travel Australia
UNIVERSITY EDUCTATION	2008 – 2011, Agricultural Biology, University of Hohenheim, Stuttgart Bachelor of Science August 2011
	2011 – 2013, Crop Sciences, University of Hohenheim, Stuttgart Master of Science August 2013
	2013 – 2017, Doctoral Student, Plant Breeding and Applied Genetics, University of Hohenheim, Stuttgart

…………………………...

Pascal Schopp

List of publications

Schopp, P., C. Riedelsheimer, H.F. Utz, C.-C. Schön and A.E. Melchinger, 2015. Forecasting the accuracy of genomic prediction with different selection targets in the training and prediction set as well as truncation selection. Theor Appl Genet. 128:2189–2201

Schopp, P., D. Müller, F. Technow and A.E. Melchinger, 2017. Accuracy of Genomic Prediction in Synthetic Populations Depending on the Number of Parents, Relatedness, and Ancestral Linkage Disequilibrium. Genetics 205:1–14

Schopp, P., D. Müller, Y.C.J. Wientjes and A.E. Melchinger, 2017. Genomic prediction within and across biparental families: means and variances in prediction accuracy and usefulness of deterministic equations. G3 7:3571–3586

Müller, D., **P. Schopp** and A.E. Melchinger, 2017. Persistency of prediction accuracy and genetic gain in synthetic populations under recurrent genomic selection. G3 7:801-811

Melchinger, A.E., **P. Schopp**, D. Müller, T.A. Schrag, E. Bauer, S. Unterseer, L. Homann, W. Schipprack, and C.-C. Schön, 2017. Libraries of doubled-haploid lines to safeguard our genetic resources. Genetics 207: 1611-1619

Müller D., **P. Schopp** and A. E. Melchinger, 2018. Selection on Expected Maximum Haploid Breeding Values Can Increase Genetic Gain in Recurrent Genomic Selection. G3 8:1173-1181

Brauner P. C., D. Müller, **P. Schopp**, J. Böhm, E. Bauer, C.-C. Schön and A. E. Melchinger, 2018. Genomic Prediction Within and Among Doubled-Haploid Libraries from Maize Landraces. Genetics Early Online https://doi.org/10.1534/genetics.118.301286

Erklärung

Hiermit erkläre ich an Eides statt, dass die vorliegende Arbeit von mir selbst verfasst wurde und lediglich unter Zuhilfenahme der angegebenen Quellen und Hilfsmittel angefertigt wurde. Wörtlich oder inhaltlich übernommene Stellen wurden als solche gekennzeichnet.

Die vorliegende Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Institution oder Prüfungsbehörde vorgelegt.

Insbesondere erkläre ich, dass ich nicht früher oder gleichzeitig einen Antrag auf Eröffnung eines Promotionsverfahrens unter Vorlage der hier eingereichten Dissertation gestellt habe.

Stuttgart-Hohenheim, Mai 2017

....
Pascal Schopp