Institute of Plant Breeding, Seed Science and Population Genetics

Applied Genetics and Plant Breeding

Prof. Dr. Albrecht E. Melchinger

University of Hohenheim

Genomic Selection in Synthetic Populations

Dissertation

submitted in fulfillment of the requirements for the degree

“Doktor der Agrarwissenschaften”

(Dr. sc. agr. / *Ph. D.* in Agricultural Sciences)

to the

Faculty of Agricultural Sciences

Dominik Müller

from Karlsbad-Langensteinbach

Stuttgart-Hohenheim

June 2017

This thesis was accepted as a doctoral dissertation in fulfillment of the requirements for the degree “Doktor der Agrarwissenschaften” (Dr. sc. agr. / Ph. D. in Agricultural Sciences) by the Faculty of Agricultural Sciences at the University of Hohenheim, on July 5th, 2017.

Day of oral examination: March 26th, 2018

Examination Committee:

Vice-Dean / Head of Committee:	Prof. Dr. T. Streck
1. examiner and reviewer:	Prof. Dr. A.E. Melchinger
2. examiner and reviewer:	Prof. Dr. C.C. Schön
3. examiner:	apl. Prof. Dr. T. Würschum

Contents

1 **General Introduction** 1

2 **Shrinkage estimation of the genomic relationship matrix can improve prediction accuracy in the training set** [1] 17

3 **Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness and ancestral linkage disequilibrium** [2] 19

4 **Persistency of Prediction Accuracy and Genetic Gain in Synthetic Populations under Recurrent Genomic Selection** [3] 21

5 **General Discussion** 23

6 **Summary** 34

7 **Zusammenfassung** 36

[1] Müller, D., Technow, F., and Melchinger, A. E. (2015). Shrinkage estimation of the genomic relationship matrix can improve genomic estimated breeding values in the training set. Theoretical and Applied Genetics, 128(4):693–703.

[2] Schopp, P., Müller, D., Technow, F. and Melchinger, A. E. (2017) Accuracy of Genomic Prediction in Synthetic Populations Depending on the Number of Parents, Relatedness, and Ancestral Linkage Disequilibrium. Genetics, 205(1):441-454.

[3] Müller, D., Schopp, P. and Melchinger, A. E. (2017) Persistency of Prediction Accuracy and Genetic Gain in Synthetic Populations under Recurrent Genomic Selection. G3: Genes, Genomes, Genetics, 7(3):801-811.

Symbols

Acronyms / Abbreviations

N_e	effective population size
N_p	number of parents
GEBV	genomic estimated breeding value
GRM	genomic relationship matrix
GS	genomic selection
LD	linkage disequilibrium
PA	prediction accuracy
QTL	quantitative trait locus
SNP	single nucleotide polymorphism
SOI	sources of information
SOV	sources of variation
TS	training set

Chapter 1

General Introduction

Genomic Selection

Genomic selection (GS) is an emerging tool in the genomic era that has first been conceived by Meuwissen et al. (2001). GS can be considered as an advancement of marker-assisted selection (MAS). In MAS, the idea is to identify and validate a small number of significant marker - trait associations, which can then be employed to obtain a higher efficiency compared to phenotypic selection alone (Lande and Thompson 1990). However, one major limitation of MAS is that the selected marker loci usually explain only a small portion of the available additive genetic variance (Yang et al. 2010). GS represents a paradigm shift, because here the goal is no longer to find significant associations, but to use all markers in a statistical model and simultaneously estimate their effects.

After its inception, GS rapidly developed into a very active area of research in animal breeding. Nowadays, GS is a standard procedure for dairy cattle and is practiced for dozens of traits in different breeds based on officially released genomic estimated breeding values (GEBVs). As of April 2016, there were already more than 1 million genotypes available for genomic evaluations in the U.S., and local producers have accepted GEBVs as reliable accounts of a bull's performance, such that nowadays over half of the matings involve young bulls without progeny records (Wiggans et al. 2017). There has been accumulating evidence that GS can also lead to major improvements in plant breeding by increasing genetic gain and reducing costs for phenotyping (Bernardo and Yu 2007; Crossa et al. 2011; Heffner et al. 2009, 2010, 2011; Lorenzana and Bernardo 2009). However, because plant breeding is mainly conducted by large international companies with closed germplasm pools, there is little public information available on the extent that GS is already implemented in practical breeding. Nevertheless, it can be safely assumed that all big companies are already routinely making use of GS in their selection programs to provide their breeders with additional information for taking their decisions.

The principle of GS is straightforward. First, a so-called training set (TS) of phenotyped individuals has to be established. These individuals are also genotyped with molecular markers, mostly single nucleotide polymorphisms (SNPs), and a statistical regression model is used to associate phenotypes with genotypic data. The model can then be used to derive predictions of breeding values for any genotyped individual. Based on these predictions, individuals can be selected for further breeding at a stage where phenotypic data are not yet available, *e.g.*, taking tissue samples from grains or seedlings. As GS rests on the potential of genome-wide molecular markers to capture QTL effects, markers are a key component in genomic breeding strategies (Lin et al. 2014), and it was a fortunate coincidence that the development of GS concurred with the availability of the SNP technology, offering high-throughput and inexpensive genomic data (Muir 2007).

As GS was first proposed and initially developed for the purpose of dairy cattle breeding, the available concepts cannot necessarily be directly applied to situations in plant breeding. This is due to fundamental differences between breeding of dairy cattle and crops (*cf.* Jonas and De Koning 2013). In plant breeding, effective population sizes (N_e) are usually much smaller as compared to animal breeding (Lin et al. 2014), either because of crops being naturally autogamous or because breeding populations are highly structured. A special type of population structure that is restricted to plant breeding are synthetic populations, where a limited number of parental individuals is intermated to form a new population with relatively small N_e. Up to now, there is no research available assessing the properties of GS in synthetic populations. Training data in dairy cattle is usually composed of historical data that has been accumulated over many years. On the other hand, individuals of the TS are themselves candidates for selection in many applications in plant breeding. Thus, it is reasonable to compare genomic predictions of such individuals with their phenotypic data and to assess opportunities for improving prediction accuracy (PA) within the TS.

Sources of Quantitative Genetic Information

When GS was proposed by Meuwissen et al. (2001), it was initially thought that the information provided by molecular markers stems only from ancestral LD with causative QTL. However, Gianola et al. (2009) and Habier et al. (2007) demonstrated that molecular markers are also able to explain additive genetic relationships between individuals that are traditionally assessed using pedigree records. Hence, it was shown that GS also implicitly utilizes pedigree-information. A third source of information was described in Habier et al. (2013) and was called co-segregation information. Co-segregation information is likely the most non-intuitive form of information. It is only present if individuals are related by their pedigree, *i.e.*, share common ancestors, but it is distinct from information from additive genetic relationships. Co-segregation information is grounded on the fact that alleles at linked marker loci and QTL co-segregate in the process of meiosis, and hence alleles at these loci can appear to be tightly associated within segregating families, even if they are unassociated at the germplasm level. It is important to recognize that the term *sources of information* (SOI) in the strict sense always refers to information that is explained from molecular markers about QTL. Later, I will discuss the *sources of variation*, *i.e.*, factors that give rise to variation in genetic relationships at QTL; where the latter determines the potential PA that can be achieved in GS.

Prediction of Genomic Values

A large number of parametric and non-parametric as well as frequentist and Bayesian models for GS have been developed during the last decade (*cf.* Campos et al. 2013). In the following, I describe the two most commonly used models, random-regression best linear unbiased prediction (RR-BLUP) and genomic best linear unbiased prediction (GBLUP), that were also employed in this thesis. A comprehensive study by Heslot et al. (2012) compared numerous models across different crops and traits, and no single model could be identified that uniformly outperformed the others. However, RR-BLUP and GBLUP appear to rely to a greater extent on additive genetic relationships and co-segregation, such that RR-BLUP/GBLUP is likely more suitable to situations where there are many QTL with small effects. On the other hand, Bayesian methods tend to better use ancestral LD between marker loci and QTL, such that they can better fit situations with few major QTL (Habier et al. 2007; Lin et al. 2014). These results from simulation studies were not always confirmed in empirical studies, where differences between models were mostly small and GBLUP performed well across a wide range of cases (Campos et al. 2013).

In the following, we present the RR-BLUP and GBLUP models. Let x_{ik} be a genotypic score variable for the ith individual at the kth locus. x_{ik} is informative about the number of alleles of a given type present at the respective locus. If the minor allele of a bi-allelic locus is coded with 1 and the major allele with 0, an individual homozygous for the minor allele (the major allele) would receive a genotypic score of $x_{ik} = 2$ ($x_{ik} = 0$) and a heterozygous individual $x_{ik} = 1$. A simple linear model, associating phenotypes with genotypes, is written as

$$y_i = \mu + \sum_k x_{ik}\alpha_k + \varepsilon_i, \tag{1.1}$$

where μ is the general intercept, y_i is the phenotypic value of the ith individual, α_k is the regression coefficient associated with the kth locus, and ε_i is the model residual. Letting

p_k be the minor allele frequency at the kth locus, then using a transformed predictor $x'_{ik} = x_{ik} - 2p_k$ leads to regression coefficients α_k from model 1.1 that can be interpreted as average effects of allele substitutions (Lynch and Walsh 1998), illustrating how this model relates to the estimation of breeding values.

The assumptions about the distribution of α_k are crucial for defining a variety of different models for GS, a summary of which can be found in Campos et al. (2013). For RR-BLUP, originally proposed in Meuwissen et al. (2001), the assumptions are $\alpha_k \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2_\alpha\right)$ and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2_\varepsilon\right)$, where σ^2_α is the variance of the substitution effects and σ^2_ε is the residual variance. This model is referred to as "random-regression BLUP" (Habier et al. 2007) because the regression coefficients α_k are assumed to be random effects of a linear mixed-model. Best linear unbiased predictions (BLUPs) are obtained by solving the mixed-model equations (Henderson 1984). The general intercept μ is assumed to be fixed. This model was shown to be equivalent to GBLUP (Goddard 2009; Habier et al. 2007; Strandén and Garrick 2009). In GBLUP, the pedigree-based numerator relationship matrix (A) of the classical animal model (Henderson 1973) that describes the additive genetic relationships is replaced by a marker-derived genomic relationship matrix (GRM), denoted by $\mathbf{G}$. As in the animal model, the genetic value of each individual and the relationships between all individuals are explicitly included in the model (Goddard et al. 2009). The genomic relationship coefficient g_{ij} between two individuals i and j is usually computed as

$$g_{ij} = \sum_k c(x_{ik} - 2p_k)(x_{jk} - 2p_k), \tag{1.2}$$

where $c = \left(2\sum_k p_k(1-p_k)\right)^{-1}$ (Habier et al. 2007; VanRaden 2008). Multiplication by c makes G similar to A (VanRaden 2007). The GBLUP model equivalent to model 1.1 is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Za} + \boldsymbol{\varepsilon}, \tag{1.3}$$

where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{a}$ is a vector of breeding values with $\mathbf{a} \sim \mathscr{N}\left(0, \sigma_a^2 \mathbf{G}\right)$, $\mathbf{Z}$ is an incidence matrix that associates breeding values with phenotypes, and $\boldsymbol{\varepsilon}$ is a vector of model residuals with $\varepsilon_i \overset{iid}{\sim} \mathscr{N}\left(0, \sigma_\varepsilon^2\right)$. The connection between the RR-BLUP model and the GBLUP model is made when using $\sigma_\alpha^2 = c\sigma_a^2$ in RR-BLUP, and conversely breeding values can be predicted in RR-BLUP by simply summing up effect estimates over all loci.

BLUP estimates of $\mathbf{a}$ in equation 1.3 can be obtained by solving the classical mixed-model equations (Lynch and Walsh 1998)

$$\begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{1} & \mathbf{Z}^T\mathbf{Z} + \frac{\sigma_\varepsilon^2}{\sigma_a^2}\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix}. \tag{1.4}$$

Using the GBLUP formulation is often times advantageous compared to RR-BLUP, because instead of one effect per locus, one has to estimate only one breeding value per indivdiual (Goddard 2009), and the number of individuals is commonly much smaller than the number of markers.

Prediction Accuracy in the Training Set

Research on GS has so far been focused on investigating PA and its influencing factors when the targets of prediction are individuals without phenotypic data. However, the phenotyped training individuals that are used to calibrate the statistical model are themselves possible selection candidates in many practical applications in plant breeding, and the TS may constitute a considerable fraction of the total number of candidates. An important example of such a case is the evaluation of inbred lines for testcross-performance in hybrid breeding. Here, lines are crossed with a tester from the opposite heterotic pool, which is commonly an elite inbred line unrelated to the candidates with promising potential of producing superior hybrids (Melchinger 1987). In commercial hybrid breeding programs, a large number of candidate lines are produced and need to be evaluated for testcross performance in each season (Albrecht et al. 2011). Genomic prediction offers the possibility to save valuable resources by evaluating only a part of the candidates (the TS) by their testcrosses in the field and predicting the performance of the remaining ones.

In the case of the GBLUP model, predictive information between the training and the prediction set[1] comes from the estimation of genomic relationships at QTL, and there is no obstacle to exploit this information for GS of TS individuals as well. However, molecular markers do not necessarily fully describe relationships between individuals as they are present at the level of the underlying QTL of a trait. A reason for this can be insufficient marker coverage of the genome, but also that in reality, QTL can be multi-allelic and can follow a different allele frequency spectrum compared to molecular markers, both of which is expected to reduce linkage disequilibrium between QTL and markers (Jannink 2010). Hence, genomic relationships are inherently associated with some level of noise. Endelman and Jannink (2012) showed that shrinking of the GRM toward a less complex target matrix has the potential to improve PA in the TS. Despite its practical importance, GS within the TS has so far received only very little attention in the literature, and there is a need for research to possibilities for improving PA, which will be addressed in one part of this thesis.

Synthetic Populations

Synthetic populations, commonly denoted as synthetics, are population varieties that are produced either by controlled matings or open pollination from a limited number of selected parental components, with ensuing cross-pollination of the F_1 progeny for one or several generations (Becker 2011; Falconer and Mackay 1996). Synthetic varieties are common in forage crops, where most grasses and many legumes can be easily propagated vegetatively, so that clones are often used as parental components (Becker 2011). For

[1] With "prediction set", I refer to the individuals on which no phenotypic information is available and that are the original target of genomic evaluations, *i.e.*, that can only be assessed by their GEBVs.

example, alfalfa (*Medicago sativa*) is one of the most widespread forage legume and its varieties are synthetics from a variable number of parents (N_p), which can be clones, half-sib or full-sib families (Flajoulot et al. 2005). Already in 1919, Hayes and Garber (1919) suggested the use of synthetic varieties of maize. Later, Sprague and Jenkins (1943) pointed out the value of synthetics as a reservoir to harbor desired gene combinations. In this respect, a particularly prominent example of a synthetic in maize breeding is the "Iowa Stiff Stalk Synthetic", which was developed from 16 inbred lines in the 1930s and has since then been subjected to two long-term recurrent selection programs (Hallauer 2008). From this synthetic, numerous very successful elite inbred lines such as B73 have been derived (Hagdorn et al. 2003; Sprague and Jenkins 1943). These inbred lines have contributed a large proportion of today's commercial maize germplasm (Mikel 2006), illustrating the importance of the this synthetic for modern maize breeding. Besides their usage for practical breeding, synthetics have played an important role in quantitative genetic research on gene action in complex heterotic traits and the comparison of selection methods (*cf.* Hallauer et al. 2010).

Genomic Selection in Synthetics

Genomic selection has been considered and investigated in detail in numerous types of structured plant breeding populations (*e.g.* Albrecht et al. 2011; Lorenzana and Bernardo 2009), but a systematic analysis of the prospects and properties of GS in synthetics is still missing so far. This is especially relevant in situations where synthetics represent actual variety types, as it is the case for many perennial forage crops and amenity grasses because of their superiority compared to open-pollinated varieties (Bradshaw 2016). Examples are alfalfa (*Medicago sativa*, Becker 2011) and perennial ryegrass (*Lolium perenne*, Miedaner 2010). GS has been proposed for breeding of alfalfa by recurrent selection (Veronesi et al. 2010), and recently first empirical results on GS in alfalfa became available. These report prediction accuracies of 0.32 and 0.35 for biomass yield, indicating over three-fold greater expected genetic gain per unit time compared to phenotypic selection (Annicchiarico et al. 2015), and in some cases moderate (0.3 – 0.4) accuracies for forage quality traits were observed (Biazzi et al. 2017).

Moreover, the investigation of GS in synthetics also provides a nearly ideal framework for studying the factors influencing the contributions of the three SOI to PA. This is because synthetics are constructed from a defined number of parental components, in the case of maize fully homozygous inbred lines. Varying the number of parental components allows for seamlessly modeling a plethora of breeding scenarios ranging between the common case of a bi-parental family to the case of a population variety. Different numbers of parents will give rise to different levels of pedigree-relationships between members of the synthetic, as well as a different relative importance of Mendelian sampling, which enables studying the importance of the associated SOI for PA under full control of all factors. Furthermore, parental inbred lines are themselves derived from source germplasm, which may exhibit a variable level of LD. Ancestral LD between marker loci and QTL has been shown to contribute information to PA, but it is yet unclear how ancestral LD affects PA in the derived synthetic, given that LD needs to pass the funnel of parental lines. These questions are addressed in the second part of this thesis.

Recurrent Genomic Selection in Synthetics

Recurrent selection has played a central role in plant breeding. Essentially, plant breeding itself can be broadly considered as a process of recurrent, *i.e.*, repetitive, selection. For maize (*Zea mays*), Hallauer and Carena (2012) differentiate between two types of cyclical selection: The first focuses on the development of improved inbred lines for hybrid production, where selection mostly takes place within F_2 families produced from recycled elite lines. The second is recurrent selection, where in general the goal is to enhance genetically broad-based populations by gradually increasing the frequency of favorable alleles. However, both methods can be coupled, in that superior lines for hybrid breeding can be produced as spin-offs during recurrent selection (Hallauer 1992). In this case, a population improved by recurrent selection has the purpose to provide a high frequency of superior individuals from which inbred lines are derived (Penny et al. 1963).

Different recurrent selection methods (*e.g.*, mass, half-sib, full-sib, S_1) have been applied to a large array of autogamous and allogamous crop species, including crop-specific modifications, and to many traits of varying complexity. Interpopulation recurrent selection has been used on crops that exhibit a large level of heterosis that is utilized to develop high-performance hybrids. A comprehensive discussion of recurrent selection can be found in Hallauer (1985) and Hallauer (1992).

Synthetic populations have been used as source material for recurrent selection (*e.g.*, the “Iowa Stiff Stalk Synthetic”) and recently, they have been proposed as a particularly suitable source material for the application of GS to recurrent selection (Gorjanc et al. 2016; Windhausen et al. 2012). Here, the virtue of GS would be that an established prediction equation could be repeatedly applied and possibly updated across multiple cycles for selecting candidates. If this is combined with the use of off-season nurseries on the other hemishpere, this would allow for completing up to two selection cycles per year, and hence promises to increase response to selection while reducing the costs for phenotyping (Bernardo and Yu 2007).

Although the usefulness of GS across multiple selection cycles has been investigated in numerous simulation studies (*e.g.* Bastiaansen et al. 2012; Jannink 2010; Liu et al. 2015; Muir 2007; Sonesson and Meuwissen 2009; Yabe et al. 2013, 2016), these studies generally considered large effective population sizes ($N_e \geq 100$), whereas in synthetics N_e is small due to the limited number of parental components. So far, no study has systematically investigated the influence of N_p on the importance of the SOI for the persistency of PA and genetic gain in recurrent GS in synthetics, which will be addressed in the third part of this thesis.

Objectives

The encompassing topic of the present thesis was the investigation of the properties of GS in scenarios specific to plant breeding. The first part is concerned with finding possibilities to improve PA of GS in the TS used for calibration. In the second part, the focus is on GS in synthetic populations and a dissection of the SOI contributing to PA in a single cycle. The third part analyzes PA and genetic gain across multiple cycles of recurrent selection. In particular, the objectives of this thesis were to

1a) investigate whether shrinkage estimation of the genomic relationship matrix can improve PA in the TS,

1b) compare newly developed with existing shrinkage approaches under different scenarios regarding population structure and marker density,

2a) examine how PA in synthetics depends on the N_p and ancestral LD,

2b) assess the importance of the three SOI for prediction accuracy and how they are influenced by TS size and marker density,

2c) analyze the relationship of observed LD between QTL and markers among the ancestral population, parents, and the synthetics generated from them,

3a) analyze PA and genetic gain in recurrent GS in synthetics, depending on N_p, ancestral LD, and the number of recombination generations,

3b) assess the relative importance of the three SOI in recurrent GS, considering also TS size and marker density.

References

Albrecht, Theresa et al. (2011). “Genome-based prediction of testcross values in maize.” *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 123.2, pp. 339–50.

Annicchiarico, Paolo et al. (2015). “Accuracy of genomic selection for alfalfa biomass yield in different reference populations”. *BMC Genomics* 16.1, p. 1020.

Bastiaansen, John W M et al. (2012). “Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures.” *Genetics, selection, evolution : GSE* 44.3.

Becker, Heiko (2011). *Pflanzenzüchtung*. Uni-Taschenbücher basics M. UTB GmbH, p. 368.

Bernardo, Rex and Jianming Yu (2007). “Prospects for genomewide selection for quantitative traits in maize”. *Crop Science* 47.3, pp. 1082–1090.

Biazzi, Elisa et al. (2017). “Genome-wide association mapping and genomic selection for alfalfa (Medicago sativa) forage quality traits”. *PLoS ONE* 12.1, pp. 1–17.

Bradshaw, John E. (2016). *Plant Breeding: Past, Present and Future*. 1st ed. Springer International Publishing, p. 693.

Campos, Gustavo de los et al. (2013). “Whole-genome regression and prediction methods applied to plant and animal breeding.” *Genetics* 193.2, pp. 327–45.

Crossa, José et al. (2011). “Genomic Selection and Prediction in Plant Breeding”. *Journal of Crop Improvement* 25.3, pp. 239–261.

Endelman, Jeffrey B. and Jean-Luc Jannink (2012). “Shrinkage estimation of the realized relationship matrix.” *G3* 2.11, pp. 1405–1413.

Falconer, Douglas S. and Trudy F. C. Mackay (1996). *Introduction to Quantitative Genetics*. 4th ed. San Francisco: Benjamin Cummings, p. 480.

Flajoulot, Sandrine et al. (2005). “Genetic diversity among alfalfa (Medicago sativa) cultivars coming from a breeding program, using SSR markers”. *Theoretical and Applied Genetics* 111.7, pp. 1420–1429.

Gianola, Daniel et al. (2009). “Additive genetic variability and the Bayesian alphabet.” *Genetics* 183.1, pp. 347–363.

Goddard, Michael E. (2009). “Genomic selection: prediction of accuracy and maximisation of long term response.” *Genetica* 136.2, pp. 245–257.

Goddard, Michael E. et al. (2009). “Estimating Effects and Making Predictions from Genome-Wide Marker Data”. *Statistical Science* 24.4, pp. 517–529.

Gorjanc, Gregor et al. (2016). “Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations”. *BMC Genomics* 17.1, p. 30.

Habier, David, Rohan L. Fernando, and Jack C. M. Dekkers (2007). “The impact of genetic relationship information on genome-assisted breeding values.” *Genetics* 177.4, pp. 2389–2397.

Habier, David, Rohan L. Fernando, and Dorian J. Garrick (2013). “Genomic BLUP decoded: a look into the black box of genomic prediction.” *Genetics* 194.3, pp. 597–607.

Hagdorn, Sandra et al. (2003). “Molecular genetic diversity among progenitors and derived elite lines of BSSS and BSCB1 maize populations”. *Crop Science* 43.2, pp. 474–482.

Hallauer, Arnel R. (1985). “Compendium of recurrent selection methods and their application”. *CRC Critical Reviews in Plant Sciences* 3.1, pp. 1–33.

– (1992). “Recurrent Selection in Maize”. *Plant Breeding Reviews* 9.2778, pp. 115–179.

– (2008). “Corn Breeding”. *Iowa State Research Farm Progress Reports* Paper 549.

Hallauer, Arnel R. and Marcelo J. Carena (2012). “Recurrent selection methods to improve germplasm in maize”. *Maydica* 57.3-4, pp. 266–283.

Hallauer, Arnel R., Marcelo J. Carena, and J.B. Miranda Filho (2010). *Quantitative genetics in maize breeding.* 6th ed. Springer.

Hayes, H. K. and R. J. Garber (1919). “Synthetic production of high protein corn in relation to breeding”. *Society* 15.178, pp. 309–318.

Heffner, Elliot L., Mark E. Sorrells, and Jean-Luc Jannink (2009). “Genomic Selection for Crop Improvement”. *Crop Science* 49.1, p. 1.

Heffner, Elliot L. et al. (2010). “Plant breeding with Genomic selection: Gain per unit time and cost”. *Crop Science* 50.5, pp. 1681–1690.

Heffner, Elliot L. et al. (2011). “Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations”. *Crop Science* 51.6, p. 2597.

Henderson, Charles R. (1973). “Sire Evaluation and Genetic Trends”. *Journal of Animal Science*, pp. 10–41.

– (1984). *Applications of linear models in animal breeding.* University of Guelph.

Heslot, Nicolas et al. (2012). “Genomic Selection in Plant Breeding: A Comparison of Models”. *Crop Science* 52.1, p. 146.

Jannink, Jean-Luc (2010). “Dynamics of long-term genomic selection.” *Genetics, selection, evolution : GSE* 42.35.

Jonas, Elisabeth and Dirk-Jan De Koning (2013). “Does genomic selection have a future in plant breeding?” *Trends in Biotechnology* 31.9, pp. 497–504.

Lande, Russell and Robin Thompson (1990). “Efficiency of marker-assisted selection in the improvement of quantitative traits.” *Genetics* 124.3, pp. 743–56.

Lin, Zibei, Ben J. Hayes, and Hans D. Daetwyler (2014). “Genomic selection in crops, trees and forages: A review”. *Crop and Pasture Science* 65.11, pp. 1177–1191.

Liu, Huiming et al. (2015). “Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs”. *Genetics, selection, evolution : GSE* 47.1, p. 19.

Lorenzana, Robenzon E. and Rex Bernardo (2009). “Accuracy of genotypic value predictions for marker-based selection in biparental plant populations.” *Theoretical and applied genetics* 120.1, pp. 151–61.

Lynch, Michael and Bruce Walsh (1998). *Genetics and Analysis of Quantitative Traits*. 1st ed. Sunderland: Sinauer Associates, p. 980.

Melchinger, Albrecht E. (1987). "Expectation of means and variances of testcrosses produced from from F2 and backcross individuals and their selfed progenies". *Heredity* 59.1, pp. 105–115.

Meuwissen, Theo H. E., Ben J. Hayes, and Michael E. Goddard (2001). "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps". *Genetics* 157.4, pp. 1819–1829.

Miedaner, Thomas (2010). *Grundlagen der Pflanzenzüchtung*. DLG-Verlag, p. 261.

Mikel, Mark A. (2006). "Availability and analysis of proprietary dent corn inbred lines with expired U.S. plant variety protection". *Crop Science* 46.6, pp. 2555–2560.

Muir, William M. (2007). "Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters". *Journal of Animal Breeding and Genetics* 124.6, pp. 342–355.

Penny, L. H. et al. (1963). "Recurrent Selection". *Statistical Genetics and Plant Breeding: A Symposium and Workshop*. Ed. by Warren D. Hanson and Harold F. Robinson. National Academy of Sciences-National Research Council, pp. 352–367.

Sonesson, Anna K. and Theo H. E. Meuwissen (2009). "Testing strategies for genomic selection in aquaculture breeding programs." *Genetics, selection, evolution : GSE* 41, p. 37.

Sprague, George F. and Merle T. Jenkins (1943). "A Comparison of Synthetic Varieties, Multiple Crosses, and Double Crosses in Corn". *Agronomy Journal* 35.2, p. 137.

Strandén, I. and Dorian J. Garrick (2009). "Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit." *Journal of dairy science* 92.6, pp. 2971–2975.

VanRaden, Paul M. (2007). "Genomic Measures of Relationship and Inbreeding". *Interbull Annual Meeting Proceedings* 37, pp. 33–36.

VanRaden, Paul M. (2008). “Efficient methods to compute genomic predictions.” *Journal of dairy science* 91.11, pp. 4414–4423.

Veronesi, Fabio, E. Charles Brummer, and Christian Huyghe (2010). “Alfalfa”. *Fodder Crops and Amenity Grasses Band 5 von Handbook of Plant Breeding*. Ed. by Beat Boller, Ulrich K. Posselt, and Fabio Veronesi. Springer New York. Chap. 17, p. 524.

Wiggans, George R. et al. (2017). “Genomic Selection in Dairy Cattle: The USDA Experience”. *Annual Review of Animal Biosciences* 5.1, pp. 309–327.

Windhausen, Vanessa S. et al. (2012). “Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments.” *G3* 2.11, pp. 1427–1436.

Yabe, Shiori, Ryo Ohsawa, and Hiroyoshi Iwata (2013). “Potential of genomic selection for mass selection breeding in annual allogamous crops”. *Crop Science* 53.1, pp. 95–105.

Yabe, Shiori et al. (2016). “Island-Model Genomic Selection for Long-Term Genetic Improvement of Autogamous Crops”. *Plos One* 11.4.

Yang, Jian et al. (2010). “Common SNPs explain a large proportion of the heritability for human height.” *Nature genetics* 42.7, pp. 565–569.

Chapter 2

Shrinkage estimation of the genomic relationship matrix can improve prediction accuracy in the training set

Dominik Müller[1], Frank Technow[1], Albrecht E. Melchinger[1]

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

Theoretical and Applied Genetics (2015) 128(4):693-703

The original publication is available at https://link.springer.com/article/10.1007/s00122-015-2464-6 (DOI: 10.1007/s00122-015-2464-6).

Abstract

In genomic prediction in plant breeding, the training set constitutes a large fraction of the total number of genotypes assayed and is itself subject to selection. The objective of our study was to investigate whether genomic estimated breeding values (GEBVs) of individuals in the training set can be enhanced by shrinkage estimation of the genomic relationship matrix. We simulated two different population types: a diversity panel of unrelated individuals and a biparental family of doubled haploid lines. For different training set sizes (50, 100, 200), number of markers (50, 100, 200, 500, 2,500) and heritabilities (0.25, 0.5, 0.75), shrinkage coefficients were computed by four different methods. Two of these methods are novel and based on measures of LD, the other two were previously described in the literature, one of which was extended by us. Our results showed that shrinkage estimation of the genomic relationship matrix can significantly improve the reliability of the GEBVs of training set individuals, especially for a low number of markers. We demonstrate that the number of markers is the primary determinant of the optimum shrinkage coefficient maximizing the reliability and we recommend methods eligible for routine usage in practical applications.

Chapter 3

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

Pascal Schopp[1*], Dominik Müller[1*], Frank Technow[1], Albrecht E. Melchinger[1]

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

[*] These authors contributed equally to this work.

Genetics (2017) 205(1):441-454

The original publication is available at http://www.genetics.org/content/early/2016/11/08/genetics.116.193243 (DOI: 10.1534/genetics.116.193243).

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, pedigree), we investigated the prediction accuracy of GBLUP and analyzed contributions from pedigree relationships captured by SNP markers as well as from co-segregation and ancestral LD between QTL and SNPs. The effects of training set size N_{TS} and marker density were also studied. Sampling few parents ($2 \leq N_P \leq 8$) generates substantial sample LD that carries over into synthetics through co-segregation of alleles at linked loci. For fixed N_{TS}, N_P influences prediction accuracy most strongly. If the training and prediction set are related, using $N_P < 8$ parents yields high prediction accuracy regardless of ancestral LD, because SNPs capture pedigree relationships and Mendelian sampling through co-segregation. As N_P increases, ancestral LD contributes more information, but 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 $N_P \leq 4$, due to large sample LD. For large N_P, achieving moderate prediction accuracy requires large N_{TS}, long-range ancestral LD and sufficient marker density. Our approach for analyzing prediction accuracy in synthetics provides new insights into the prospects of GP for many types of source populations encountered in plant breeding.

Chapter 4

Persistency of Prediction Accuracy and Genetic Gain in Synthetic Populations under Recurrent Genomic Selection

Dominik Müller[1*], Pascal Schopp[1*], Albrecht E. Melchinger[1]

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

[*] These authors contributed equally to this work.

G3: Genes, Genomes, Genetics 7(3):801–811

The original publication is available at http://www.g3journal.org/content/7/3/801 (DOI: 10.1534/g3.116.036582).

Abstract

Recurrent selection (RS) has been used in plant breeding to successively improve synthetic and other multiparental populations. Synthetics are generated from a limited number of parents (N_p), but little is known about how Np affects genomic selection (GS) in RS, especially the persistency of prediction accuracy ($r_{g,\hat{g}}$) and genetic gain. Synthetics were simulated by intermating $N_p = 2$ - 32 parent lines from an ancestral population with short- or long-range linkage disequilibrium (LD_A) and subjected to multiple cycles of GS. We determined $r_{g,\hat{g}}$ and genetic gain across 30 cycles for different training set (TS) sizes, marker densities, and generations of recombination before model training. Contributions to $r_{g,\hat{g}}$ and genetic gain from pedigree relationships, as well as from cosegregation and LD_A between QTL and markers, were analyzed via four scenarios differing in (i) the relatedness between TS and selection candidates and (ii) whether selection was based on markers or pedigree records. Persistency of $r_{g,\hat{g}}$ was high for small N_p, where predominantly cosegregation contributed to $r_{g,\hat{g}}$, but also for large N_p, where LD_A replaced cosegregation as the dominant information source. Together with increasing genetic variance, this compensation resulted in relatively constant long- and short-term genetic gain for increasing $N_p > 4$, given long-range LD_A in the ancestral population. Although our scenarios suggest that information from pedigree relationships contributed to $r_{g,\hat{g}}$ for only very few generations in GS, we expect a longer contribution than in pedigree BLUP, because capturing Mendelian sampling by markers reduces selective pressure on pedigree relationships. Larger TS size (N_{TS}) and higher marker density improved persistency of $r_{g,\hat{g}}$ and hence genetic gain, but additional recombinations could not increase genetic gain.

Chapter 5

General Discussion

The Three Sources of Variation and Information in Genomic Selection

In this part, the concept of the three *sources of variation* (SOV) is developed. The three SOV characterize processes that generate quantitative genetic information that can be exploited in GS. With quantitative genetic information, we specifically refer to the variance of q_{ij}, where q_{ij} is the genetic relationship coefficient at QTL between an individual i from the PS and j from the TS. This variation is the crucial factor determining the PA that can be potentially achieved in GS. As causal loci cannot be directly observed, molecular markers are used in order to capture this variation by estimating q_{ij}, and their power to do so is governed by the three SOI.

Two individuals that share a common ancestor (*i.e.*, that are related by pedigree) show a non-zero probability that two alleles randomly sampled from them (at an arbitrary locus) are identical copies of some founder allele (identical by descent, IBD) with respect to the base of the pedigree. This similarity is measured by the additive genetic relationship, which is equal to twice the coefficient of coancestry (Lynch and Walsh 1998; Malécot and Blaringhem 1948). In the following, we refer to this kind of similarity as the "expected IBD relationship", abbreviated by f_{ij} (Schopp et al. 2017). However, Mendelian sampling eventually leads to variation in the similarities between pairs of individuals that exhibit the same expected IBD relationship (Hill and Weir 2011). For instance, independent segregation of chromosomes during meiosis causes a variable proportion of shared genome between full-sibs with respect to their grandparents. Taking into account Mendelian sampling leads to another measure of similarity between (related) individuals, which we call the "actual IBD relationship" (τ_{ij}) (Schopp et al. 2017). Actual IBD relationships are similarly defined as expected ones, but they take into account the proportion of the genome of two individuals that is derived from the same founder genome at the base of the pedigree. In other words, the question "What is the probability that two randomly chosen alleles are IBD?" is answered by expected IBD relationships conditional on the pedigree records, but by actual IBD relationships conditional on the origin of alleles with respect to the base of the pedigree. The Mendelian sampling term can then be defined as the deviation of actual from expected IBD relationships, *i.e.*, $m_{ij} = \tau_{ij} - f_{ij}$. Both expected and actual IBD relationships are measures of genetic similarity that assume uniqueness of alleles at the base of the pedigree, *i.e.*, that all alleles present in the founder individuals are distinct. However, distinct founder alleles can be identical by state (IBS), and hence can be functionally equivalent if they are alleles of a causative locus. This induces similarity between individuals that can neither be explained by expected, nor by actual IBD relationships. This similarity at causative loci was called "actual IBS relationship", which is equivalent to genetic relationship (q_{ij}) (Schopp et al. 2017). The deviation of actual IBS relationship from actual IBD relationship is $\xi_{ij} = q_{ij} - \tau_{ij}$. Hence, actual IBS relationships at QTL can be factorized as

$$q_{ij} = f_{ij} + m_{ij} + \xi_{ij}. \tag{5.1}$$

It is important to note that while f_{ij} may be known from pedigree records, m_{ij} and ξ_{ij} can not be directly observed and can only be quantified in simulation studies. The contributions of $var(f_{ij})$, $var(m_{ij})$ and $var(\xi_{ij})$ to $var(q_{ij})$ will be the subject of later discussion.

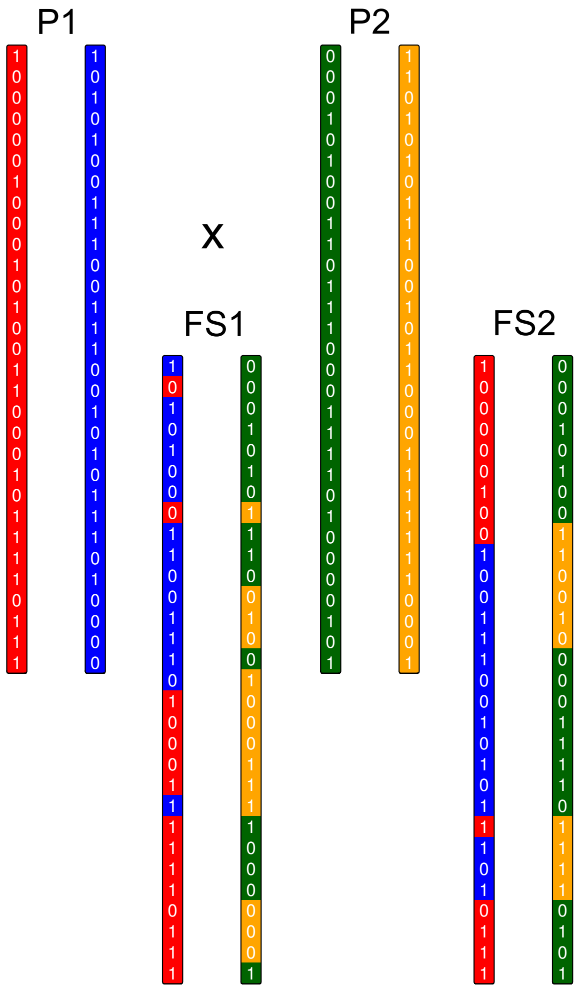

Figure 5.1 Illustration of the three sources of variation. Two parents (P1, P2) at the top produce two full-sibs (FS1, FS2). Assuming that both parents are non-inbred and form the root of the pedigree, expected IBD relationship between the sibs is equal to 0.5. The actual IBD relationship, calculated from parental origins of alleles, equals 0.633 , whereas the actual IBS relationship, calculated from states of alleles, equals 0.4 .

The concept of the three SOV is illustrated in Figure 5.1. Two non-inbred and heterozygous parents produce two full-sibs and form the root of the pedigree. Colors indicate the origin of alleles with respect to the chromosomes of the parents, whereas the numbers code for the actual state of the alleles. The expected IBD relationship (f_{ij}) between the sibs is equal to 0.5. Calculating the actual IBD relationship (τ_{ij}) from the parental origin of alleles yields a value of 0.633 , *i.e.*, knowledge of which allele is inherited from which parental chromosome suggests that the sibs are closer related to each other than expected, due to random Mendelian sampling. However, calculating the actual IBS relationship (q_{ij}) from allelic states according to formula 1.2 (assuming a frequency of 0.5 for all alleles) gives only 0.4 . Obviously, the extreme deviation of actual IBS from actual IBD relationship is an artifact and due to the small number of loci on only a single chromosome in this example. It is important to keep in mind that the sources of variation only describe what happens at the QTL.

The three SOI refer to what molecular markers can explain about $var(q_{ij})$, *i.e.*, how well marker-based genomic relationships estimate q_{ij}. Markers can capture expected IBD relationships f_{ij} (Habier et al. 2007) by segregating in the appropriate proportions as prescribed by the pedigree relationships between individuals (pedigree information). This does not require an association between markers and QTL and works best if markers segregate independently and have an intermediate allele frequency (Habier et al. 2013). Co-segregation and ancestral LD between markers and QTL both imply statistical associations between the alleles at these loci. In the case of co-segregation, the statistical associations, which we termed "sample LD", are created when a limited N_p gives rise to a new population, such as in the case of a synthetic. The smaller N_p, the larger the importance of this sample LD compared to ancestral LD in the parents' population. Co-segregation then propagates these associations from the set of parents to the resulting population. On the other hand, ancestral LD refers to associations that are the product from population genetic processes, such as genetic drift and selection, but is, eventually, also maintained by co-segregation of loci. Hence, what are association from co-segregation and what from ancestral LD is sometimes a matter of definition.

Prediction Accuracy in the Training Set

Our results demonstrate that shrinking the GRM towards a simpler target matrix has the potential to improve PA in the TS. However, the extent of the possible improvement strongly depends on (i) the type of population, (ii) the applied marker density and (iii) the heritability of the phenotypic data. In general, potential improvements in PA were much higher for a diversity panel of nominally unrelated inbred lines than for a bi-parental family. This can be explained by the fact that in bi-parental populations, there are extensive linkage blocks (Frisch and Melchinger 2007; Smith et al. 2008), and marker loci strongly co-segregate with QTL. Hence, a much smaller number of marker loci is necessary to accurately capture genetic relationships at QTL (q_{ij}). On the contrary, in the diversity panel, markers rely entirely on the usually much weaker ancestral LD to capture genetic relationships, such that a larger number of loci is necessary.

In a study of Endelman and Jannink (2012), it was found that shrinkage of the GRM could not improve PA in the PS, a finding that was corroborated by our results. Contrary to the TS, there is no phenotypic information available for the PS, and hence shrinking the GRM will not lead to an upweighting of the own phenotype of an individual, but merely of a downweighting of the information from TS individuals. In the scenarios that we considered, populations were unstructured in the sense that there was no variation in expected IBD relationships. Accordingly, the only SOI was co-segregation in the case of a bi-parental family and population-wide LD in the case of a diversity panel. Hence, because the only available SOI was downweighted, it is plausible that PA deteriorated. However, the situation might be different in structured populations. Assuming that expected IBD relationships are known and using these as target for shrinkage, it might be the case that shrinkage improves PA in the PS by upweighting reliable information from the pedigree (f_{ij}) in favor of a noisy estimate of the Mendelian sampling term and deviations thereof (m_{ij} and ξ_{ij}).

GS within bi-parental families is one of its most important applications in plant breeding. Our results indicate that even for low marker densities, there is practically no chance to improve PA in the TS, such that we can not recommend a shrinkage approach in such a situation, unless the marker density is actually extremely low and heritability is moderate to high ($0.5 - 0.75$). Another application of GS in plant breeding is for recurrent selection in broad-based populations, such as a synthetic produced from a large N_p. This type of population is comparable to the diversity panel used in our study. In this situation, individuals would likely be genotyped using low-density chips for saving cost, so that a shrinkage approach may be worthwhile.

Genomic Selection in Synthetic Populations

The number of parents N_p as well as ancestral LD crucially affects the distributions of f_{ij}, m_{ij} and ξ_{ij}, and hence influences the SOV as well as the SOI of PA in synthetics. It was shown that the closer f_{ij} between individuals, the larger $var(m_{ij})$ relative to these individuals, because the longer are the haplotypes that they inherit from their common ancestors (Goddard et al. 2011; Hill and Weir 2011). Increasing N_p decreases the proportion of close f_{ij} relationships (*i.e.*, full-sibs, half-sibs) between the TS and the PS. This leads to an overall reduction in $var(m_{ij})$, and hence, for a given size of the TS, to a lower absolute frequency of exceptionally close relatives (as measured by q_{ij}). Because cryptic relationships between the parents of a synthetic are correlated with ancestral LD, increasing ancestral LD leads to higher $var(\xi_{ij})$ in the synthetic. As a consequence, given that TS size is fixed, increasing N_p or reducing ancestral LD reduces $var(q_{ij})$ and therefore decreases the potential to obtain high PA. These observations apply to the case when the TS and PS are subsets of the same synthetic, *i.e.*, derived from the same set of parents. If this is not the case and they stem from two distinct synthetics with non-overlapping parents, only ξ_{ij} but neither f_{ij} nor m_{ij} contributes to q_{ij}, which strongly limits the possibility of reaching high PA.

In GS within a bi-parental family, *i.e.*, where both TS and PS come from the same family, it is only co-segregation that provides information about $var(m_{ij})$ and $var(\xi_{ij})$. This is because marker loci will (co-)segregate in approximately equal proportions as QTL, hence capturing $var(m_{ij})$, and because the allelic states at markers inform about the allelic states at QTL. Pedigree relationships (f_{ij}) do not contribute information, because they do not vary between the individuals. Also, ancestral LD does not contribute, because all the observed LD between markers and QTL is complete (*i.e.*, $r^2 = 1$), independent of the locations of loci on the genome. Our findings corroborate existing experience and former claims that the potential accuracy of GS is maximal within bi-parental families (Lehermeier et al. 2014; Riedelsheimer et al. 2013). We now understand that this is because of the large $var(q_{ij})$ due to the outstanding contribution of $var(m_{ij})$ and marker being able to successfully capture this variance due to co-segregation. Contrary to the situation depicted in Figure 5.1, $var(m_{ij})$ increases with inbreeding of the progeny and is maximal if they are fully inbred, *e.g.*, doubled haploids, a common situation in plant breeding.

When examining GS across multiple cycles of recurrent selection, we observed a strong drop of PA at the beginning of selection, especially after the first and the second cycle. This is most likely caused by a rapid decline in $var(f_{ij})$ across generations and the contributed information therefrom, as it was also hypothesized by others (Habier et al. 2007; Wolc et al. 2011a, 2016). This decline is exacerbated by the fact GS uses f_{ij} relationships, which causes inbreeding and genetically narrows down the population (Daetwyler et al. 2007; Quinton et al. 1992).

When considering genetic gain, it is necessary to not only take PA into account, but also the available additive genetic variance. Increasing N_P broadens the genetic basis of the synthetic and therefore increases the initial genetic variance available for selection. This explains why long-term genetic gain increases when increasing N_p from 2 to 6, in contrast to the results for single-cycle GS where PA clearly decreased. Increasing N_p beyond 6 only moderately increased the genetic variance and long-term genetic gain reached a plateau, but only if strong LD was present in the ancestral population. The plateau was reached although our results for single-cycle PA indicated that with increasing N_P, the important contribution of co-segregation (via $var(m_{ij})$) to PA vanishes in favor of ancestral LD, decreasing PA. This indicates that both SOI largely compensate for each other with respect to long-term genetic gain. As synthetics with less than about 6 parents are unrealistic in practice, we conclude that prolonged genetic progress by GS in synthetics can be obtained irrespective of N_p, but if N_p is large, substantial ancestral LD is required.

Although it has been recommended to utilize information from expected IBD relationships in GS (*e.g.* Wolc et al. 2011b), our findings indicate that this might impair long-term genetic gain by increasing the rate of inbreeding, a result well known in animal breeding

(Belonsky and Kennedy 1988). Hence, if the breeding objective is long-term genetic gain, deliberately avoiding the implicit use of pedigree relationships (f_{ij}) in GS might be desirable with respect to long-term genetic gain and would also allow for a less frequent necessity of re-training the GS model.

Conclusions

This thesis investigated prediction accuracy of GS in the TS and in synthetic populations across multiple cycles of recurrent selection. Our specific conclusions are:

- Shrinkage estimation of the GRM has the potential to increase PA in the TS, but only if populations are unstructured and marker density is low, because in such a situation information mainly comes from ancestral LD.
- In synthetic populations, N_p is the predominant factor that influences PA. Within bi-parental families ($N_p = 2$), the only SOI for PA is co-segregation between markers and QTL. When $N_p > 2$, pedigree information and ancestral LD also contribute; the importance of the latter becomes larger as N_p increases. High ancestral LD in the source germplasm is especially important for broad-based synthetics.
- The observable LD between markers and QTL in the synthetic is the result of a superposition of ancestral LD and "sample LD" generated during the bottleneck of sampling the parents.
- During GS across multiple cycles, information from pedigree relationships rapidly vanishes after a few cycles, whereas co-segregation and even more so ancestral LD are stable SOI across many cycles.
- The long-term genetic gain of GS in synthetics is only mildly affected if N_p lies within a realistic range. This is because information from co-segregation and ancestral LD compensated for each other, provided that there is sufficient LD in the ancestral population.

References

Belonsky, G. M. and B. W. Kennedy (1988). “Selection on individual phenotype and best linear unbiased predictor of breeding value in a closed swine herd.” *Journal of animal science* 66.5, pp. 1124–1131.

Daetwyler, Hans D. et al. (2007). “Inbreeding in genome-wide selection”. *Journal of Animal Breeding and Genetics* 124.6, pp. 369–376.

Endelman, Jeffrey B. and Jean-Luc Jannink (2012). “Shrinkage estimation of the realized relationship matrix.” *G3* 2.11, pp. 1405–1413.

Frisch, Matthias and Albrecht E. Melchinger (2007). “Variance of the parental genome contribution to inbred lines derived from biparental crosses.” *Genetics* 176.1, pp. 477–88.

Goddard, Michael E., Ben J. Hayes, and Theo H. E. Meuwissen (2011). “Using the genomic relationship matrix to predict the accuracy of genomic selection.” *Journal of animal breeding and genetics* 128.6, pp. 409–421.

Habier, David, Rohan L. Fernando, and Jack C. M. Dekkers (2007). “The impact of genetic relationship information on genome-assisted breeding values.” *Genetics* 177.4, pp. 2389–2397.

Habier, David, Rohan L. Fernando, and Dorian J. Garrick (2013). “Genomic BLUP decoded: a look into the black box of genomic prediction.” *Genetics* 194.3, pp. 597–607.

Hill, W. G. and B. S. Weir (2011). “Variation in actual relationship as a consequence of Mendelian sampling and linkage.” *Genetics research* 93.1, pp. 47–64.

Lehermeier, Christina et al. (2014). “Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction”. *Genetics* 198.1, pp. 3–16.

Lynch, Michael and Bruce Walsh (1998). *Genetics and Analysis of Quantitative Traits*. 1st ed. Sunderland: Sinauer Associates, p. 980.

Malécot, Gustave; and Louis Blaringhem (1948). *Les Mathématiques de l’Hérédité*. Paris: Masson.

Quinton, M., C. Smith, and Michael. E. Goddard (1992). “Comparison of selection methods at the same level of inbreeding.” *Journal of animal science* 70.4, pp. 1060–1067.

Riedelsheimer, Christian et al. (2013). “Genomic predictability of interconnected biparental maize populations.” *Genetics* 194.2, pp. 493–503.

Schopp, Pascal et al. (2017). “Accuracy of Genomic Prediction in Synthetic Populations Depending on the Number of Parents, Relatedness and Ancestral Linkage Disequilibrium”. *Genetics* 205.1, pp. 441–454.

Smith, J. S. C. et al. (2008). “Use of doubled haploids in maize breeding: implications for intellectual property protection and genetic diversity in hybrid crops”. *Molecular Breeding* 22.1, pp. 51–59.

Wolc, Anna et al. (2011a). “Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model.” *Genetics, selection, evolution : GSE* 43.5.

Wolc, Anna et al. (2011b). “Persistence of accuracy of genomic estimated breeding values over generations in layer chickens.” *Genetics, selection, evolution : GSE* 43.23.

Wolc, Anna et al. (2016). “Mixture models detect large effect QTL better than GBLUP and result in more accurate and persistent predictions”. *Journal of Animal Science and Biotechnology* 7.1.

Chapter 6

Summary

The foundation of genomic selection has been laid at the beginning of this century. Since then, it has developed into a very active field of research. Although it has originally been developed in dairy cattle breeding, it rapidly attracted the attention of the plant breeding community and has, by now (2017), developed into an integral component of the breeding armamentarium of international companies. Despite its practical success, there are numerous open questions that are highly important to plant breeders.

The recent development of large-scale and cost-efficient genotyping platforms was the prerequisite for the rise of genomic selection. Its functional principle is based on information shared between individuals. Genetic similarities between individuals are assessed by the use of genomic fingerprints. These similarities provide information beyond mere family relationships and allow for pooling information from phenotypic data. In practice, first a training set of phenotyped individuals has to be established and is then used to calibrate a statistical model. The model is then used to derive predictions of the genomic values for individuals lacking phenotypic information. Using these predictions can save time by accelerating the breeding program and cost by reducing resources spent for phenotyping.

A large body of literature has been devoted to investigate the accuracy of genomic selection for unphenotyped individuals. However, training individuals are themselves often times

selection candidates in plant breeding, and there is no conceptual obstacle to apply genomic selection to them, making use of information obtained via marker-based similarities. It is therefore also highly important to assess prediction accuracy and possibilities for its improvement in the training set. Our results demonstrated that it is possible to increase accuracy in the training set by shrinkage estimation of marker-based relationships to reduce the associated noise. The success of this approach depends on the marker density and the population structure. The potential is largest for broad-based populations and under a low marker density.

Synthetic populations are produced by intermating a small number of parental components, and they have played an important role in the history of plant breeding for improving germplasm pools through recurrent selection as well as for actual varieties and research on quantitative genetics. The properties of genomic selection have so far not been assessed in synthetics. Moreover, synthetics are an ideal population type to assess the relative importance of three factors by which markers provide information about the state of alleles at QTL, namely (i) pedigree relationships, (ii) co-segregation and (ii) LD in the source germplasm. Our results show that the number of parents is a crucial factor for prediction accuracy. For a very small number of parents, prediction accuracy in a single cycle is highest and mainly determined by co-segregation between markers and QTL, whereas prediction accuracy is reduced for a larger number of parents, where the main source of information is LD within the source germplasm of the parents. Across multiple selection cycles, information from pedigree relationships rapidly vanishes, while co-segregation and ancestral LD are a stable source of information. Long-term genetic gain of genomic selection in synthetics is relatively unaffected by the number of parents, because information from co-segregation and from ancestral LD compensate for each other. Altogether, our results provide an important contribution to a better understanding of the factors underlying genomic selection, and in which cases it works and what information contributes to prediction accuracy.

Chapter 7

Zusammenfassung

Die jüngste Entwicklung von großen, kosteneffizienten Genotypisierungsplattformen stellt eine Grundvoraussetzung für den Erfolg der genomischen Selektion dar. Das funktionale Prinzip beruht auf der Ausnutzung von Informationen zwischen Individuen. Vorhandene genetische Ähnlichkeiten werden durch den genomischen Fingerabdruck erfasst. Diese Ähnlichkeiten liefern Informationen, die über die reinen Verwandschaftsverhältnisse hinausgehen und erlauben die Ausnutzung phänotypischer Daten über Individuen hinweg. In der Praxis muss zunächst ein Kalibrierungsdatensatz mit phänotypisierten Individuen erstellt werden, der zur Schätzung eines statistischen Modells dient. Dieses Model wird hernach eingesetzt, um Vorhersagen über den genomischen Wert von Individuen ohne phänotypische Daten zu treffen. Die Verwendung dieser Vorhersagen kann Zeit einsparen, indem das Zuchtprogramm beschleunigt wird, aber auch durch eine Verringerung der zur Phänotypisierung eingesetzten Ressourcen Kosten senken.

Die Untersuchung der Vorhersagegenauigkeit genomischer Selektion innerhalb nicht phänotypisierter Individuen war bereits Gegenstand zahlreicher Forschungsarbeiten. Bei den Trainingsindividuen zur Kalibrierung des Modells handelt es sich in der Pflanzenzüchtung jedoch häufig ebenfalls um potentielle Selektionskandidaten und es existiert kein prinzipielles Hindernis, genomische Selektion ebenso auf diese anzuwenden und die

Information von markerbasierten Ähnlichkeiten auszunutzen. Daher ist es wichtig, die Vorhersagegenauigkeit sowie deren Verbesserungsmöglichkeiten im Trainingsdatensatz zu prüfen. Unsere Ergebnisse zeigen, dass es grundsätzlich möglich ist durch Schrumpfungsschätzung von markerbasierten Verwandschaften deren Störsignale zu vermindern und die Genauigkeit im Trainingsdatensatz zu steigern. Dabei hängt der Erfolg von der Markerdichte und der Populationstruktur ab. Das Potential ist am größten für breite Populationen bei einer geringen Markerdichte.

Synthetische Populationen werden durch Kreuzung einer geringen Anzahl an elterlichen Komponenten erzeugt und haben in der Geschichte der Pflanzenzüchtung eine wichtige Rolle gespielt. Dies betrifft sowohl die Verbesserung des Zuchtmaterials durch rekurrente Selektion, als auch die Erstellung von Sorten sowie die quantitativ-genetische Züchtungsforschung. Die Eigenschaften genomischer Selektion wurden bisher nicht in Synthetiks untersucht. Zudem handelt es sich bei Synthetiks um einen idealen Populationstyp, um die Bedeutung der drei Faktoren zu untersuchen, durch welche Marker Informationen über den Zustand an QTL liefern, nämlich (i) Verwandschaftsverhältnisse (ii) Kosegregation und (iii) Kopplungsphasenungleichgewicht (LD) im Zuchtmaterial. Unsere Ergebnisse zeigen, dass die Elternzahl einen entscheidenden Faktor für die Vorhersagegenauigkeit darstellt. Bei einer sehr geringen Elternzahl ist die Vorhersagegenauigkeit innerhalb eines Zyklus am größten und wird hauptsächlich durch Kosegregation zwischen Markern und QTL bestimmt. Ist die Elternzahl hingegen groß, so tritt als vornehmliche Informationsquelle LD im Ursprungsmaterial der Eltern hervor. Wird genomische Selektion über mehrere Zyklen hinweg praktiziert, so verschwindet die Information aus Verwandschaftsverhältnissen sehr schnell, wohingegen sich Kosegregation und LD als stabile Informationsquellen erweisen. Der langfristige Selektionserfolg genomischer Selektion in einem Synthetik ist nur in einem geringen Maße abhängig von der Elternzahl, da sich Informationen aus Kosegregation und LD gegenseitig aufwiegen. Insgesamt liefern unsere Ergebnisse einen wichtigen Beitrag für ein besseres Verständnis der Grundlagen der genomischen Selektion, in welchen Fällen sie Erfolg verspricht, und welche Informationen die Vorhersagegenauigkeit beeinflussen.

Acknowledgements

First, I would like to thank my academic supervisor Prof. Dr. A. E. Melchinger for providing me with the opportunity to become a Ph.D. candidate in his group and for his advice, guidance and continuous support during my work.

Many thanks to my all of my Ph.D. colleagues at the Institute of Plant Breeding, Seed Science and Population Genetics and the State Plant Breeding Institute, Hohenheim. I think the overall atmosphere was always very supportive and motivating. Especially, I want to acknowledge the Ph.D. students of my working group for the great time that we spend together, at the institute and in private life.

Special thanks goes to Pascal Schopp, with whom I entertained a very close and productive collaboration. He was always a supportive and responsive partner for open discussions and a fruitful exchange of ideas.

Also special thanks to Matthias Westhues. We were always on the same wavelength when it came to questions about tools, approaches and best practices of research. We supported and encouraged each other until the end of our doctorate.

Last but not least, I recognize the support of Helga Kösling, Beate Devezi-Savula, Margit Lieb, and Susanne Meyer in all organizational affairs.

Curriculum Vitae

Name:	Dominik Müller
Date and place of birth:	30.07.1989, Karlsruhe, Germany

School Education

1999 - 2005	Secondary school, Realschule Karlsbad, Karlsbad, Germany
2005 - 2008	High school degree (Abitur), Albert-Einstein-Schule, Ettlingen, Germany

University Education

2009 - 2012	B.Sc. Agricultural Sciences, University of Hohenheim, Germany
2012 - 2013	M.Sc. Crop Sciences, University of Hohenheim, Germany
2013 - 2017	Doctorate candidate in Applied Genetics and Plant Breeding (Prof. A. E. Melchinger), University of Hohenheim, Germany

Employment Record

2014 - 2017	Research Assistant, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, Germany

Dominik Müller

Stuttgart, 14.06.2017

Declaration

Eidesstattliche Versicherung gemäß § 8 Absatz 2 der Promotionsordnung der Universität Hohenheim zum Dr.sc.agr.

1. Bei der eingereichten Dissertation zum Thema *Genomic Selection in Synthetic Populations* handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Ich habe nicht die Hilfe einer kommerziellen Promotionsvermittlung oder -beratung in Anspruch genommen.
4. Die Bedeutung der eidesstattlichen Versicherung und der strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Die Richtigkeit der vorstehenden Erklärung bestätige ich. Ich versichere an Eides Statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Dominik Müller

Stuttgart, 14.06.2017