Aus dem Institut für
Pflanzenbau und Pflanzenzüchtung II
der Justus-Liebig-Universität Gießen
Professur für Biometrie und Populationsgenetik
Prof. Dr. Matthias Frisch

# Genomic Prediction of Crossing Partners on Basis of the Expected Mean and Variance of their Derived Lines

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften
im Fachbereich
Agrarwissenschaften, Ökotrophologie und Umweltmanagement
Justus-Liebig-Universität Gießen

von
Tanja Osthushenrich
aus Kamen

Gießen, 26. November 2019

# Contents



[1] Osthushenrich, T, Frisch, M & Herzog, E. 2017. Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. *PLoS ONE*, **12**(12), e0188839.

[2] Osthushenrich, T, Frisch, M, Zenke-Philippi, C, Jaiser, H, Spiller, M, Cselényi, L, Krumnacker, K, Boxberger, S, Kopahnke, D, Habekuß, A, Ordon, F & Herzog, E. 2018. Prediction of means and variances of crosses with genome-wide marker effects in barley. *Front Plant Sci* , **9**:1899.

# Abbreviations

| | |
|---|---|
| DH | doubled haploid |
| GP | genomic prediction |
| GV | genotypic value |
| LD | linkage disequilibrium |
| MPV | mid-parent value |
| OHV | optimal haploid value |
| QTL | quantitative trait loci |
| RIL | recombinant inbred line |
| S | superior progeny value |
| SSD | single seed descent |

# Chapter 1

# General Introduction

## Line and hybrid breeding programs

Plant breeders aim to improve desired characteristics of their breeding material. The breeding process is characterized by alternate phases of selection of promising candidates and recombination of plants for new variation. The main steps are: (1) Development of initial genetic variability; (2) selection of promising parents for new crosses; (3) test and branch off promising candidates as products for the market. My work is related to the second step, and has a focus on cross prediction and prediction of genetic segregation variance ($\hat{\sigma}_g^2$) in line and hybrid breeding programs.

In the breeding process, every plant species is treated differently according to their characteristics and the available breeding techniques (Becker, 2011). In commercially relevant crops, plant breeders strive for homozygous populations, which can be used as line varieties or as parental lines for hybrid breeding. Both breeding schemes involve a continuous improvement of the breeding pool, which is time-consuming, expensive and requires a careful handling of resources. Lines from the present pool are crossed to derive new improved inbred lines for the subsequent cycles. A breeder desires an optimal response to selection while maintaining the diversity of the breeding material. The response to selection is measured as the difference in mean of

the new improved breeding pool and the generation before selection. A large variation of new inbred lines improves the chance to find superior lines at the top fraction of the distribution. Therefore, a breeder aims to select superior crosses, which have a high predicted mean ($\hat{\mu}$) and a large $\hat{\sigma}_g^2$ for relevant traits. Planning of crosses based on their $\hat{\mu}$ was used for a long time, but despite all efforts, $\hat{\sigma}_g^2$ is difficult to predict.

## Development of inbred lines

The development of superior inbred lines as line varieties or as hybrid components are equally important for both breeding schemes and have many similarities. Main differences can be found in the evaluation of the lines and the development of the initial breeding pools. Line varieties are tested with line *per se* performance, while for hybrid components, the performance is evaluated in test crosses. Apart from the differences, the breeding process for inbred line development follows the scheme of advanced cycle breeding (Yu & Bernardo, 2004).

Continuous breeding cycles aim to improve the homozygous lines in the current breeding pool (Fig. 1.1, 1-4). A new cycle starts with multiple crosses among lines in the base population (1). All progenies of a cross form a breeding population with a full-sib structure. The first uniform heterozygous filial generation ($F_1$) within the breeding populations are selfed (2). New improved inbreds can be developed by two common methods, single seed descent (SSD) and doubled haploid (DH) (3; Fig. 1.2) (Becker, 2011). SSD can be applied in the first segregating generation ($F_2$). In each generation one seed per selfed plant is randomly selected and grown. This cycle is repeated until the progenies are nearly homozygous in later generations ($F_6$, 96.8 %/$F_7$, 98.4 %). DH technology produces homozygous lines without continuous selfing and shortens the breeding process. Male or female gametes from the cross between the $F_1$ plants are treated with colchicine and the $F_2$

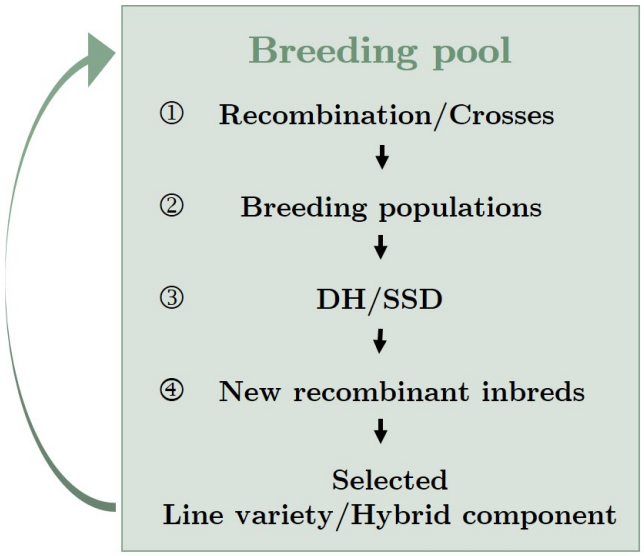


**Figure 1.1.** Scheme of advanced cycle breeding. New improved inbreds are derived from multiple crosses in the breeding pool (Yu & Bernardo, 2004).

haploid set of chromosomes is doubled. Today, DH protocols for many species are available (Maluszynski *et al.*, 2003). After selection within and among the breeding populations, improved lines can be used as parents for the following cycles (4).

Preliminary investigations have shown, that the variance of the parental genome contribution to inbred lines developed by the SSD or DH from bi-parental crosses is different. The variation around the expected mean value depends on expected number of crossovers during inbred line development with the applied breeding scheme. DH lines are developed after only one meiosis, whereas SSD lines have a meiosis in each selfing generation. Therefore, the variances of the parental genome contribution is greater for DH than for SSD lines (Fig. 1.2). Frisch & Melchinger (2007) proposed a theoretical concept for the expected gametic linkage disequilibrium (LD) between two loci at two map positions for different mating schemes. This important groundwork is necessary for the deterministic equations derived for the prediction of $\hat{\sigma}_g^2$ for new inbred lines in chapter 2.

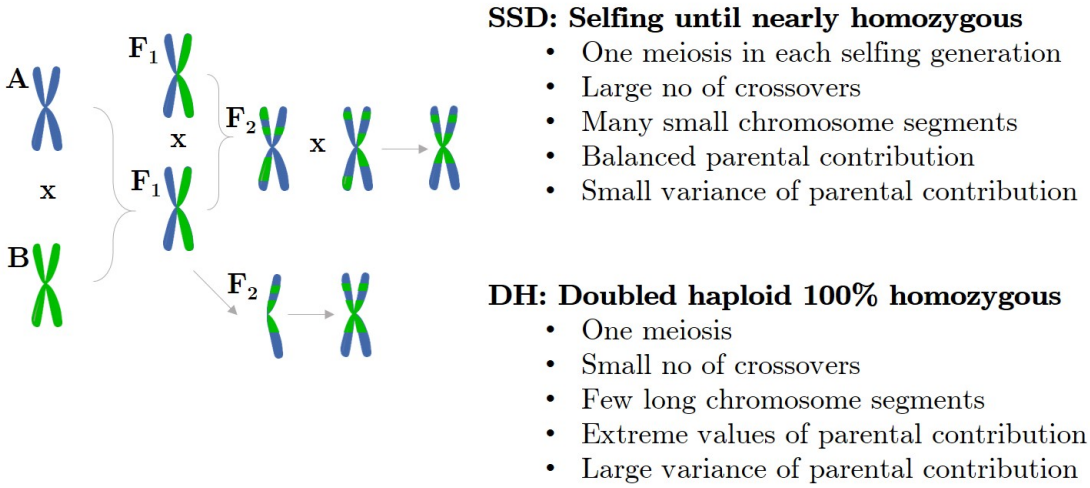


**Figure 1.2.** Inbred line development with two major systems. Continuous selfing with SSD or DH technology. Comparison of the breeding schemes with regard to the parental genome contribution (Frisch & Melchinger, 2007).

## Strategies to select promising crossing parents and find superior breeding populations

The critical step for any breeder is the selection of crossing parents to derive improved inbreds for subsequent breeding cycles. This stage is the basis for all success in response to selection. A superior breeding population is characterized by a high $\mu$ and $\sigma_g^2$ for the trait of interest. In such a breeding population, it is more likely to obtain superior inbred lines and gain a higher selection response (Fig. 1.3). A breeding pool with $N$ lines has $N * (N-1)/2$ cross combinations. This quadratic function does already indicate the tremendous increase in number of cross options with even minor increase in numbers of parental lines. The evaluation of all possible breeding populations in field trails is not feasible due to limited resources like field capacity, budget and time. Therefore, a prior estimate of the expected cross performance is a very helpful tool to plan crosses and allocate the resources.

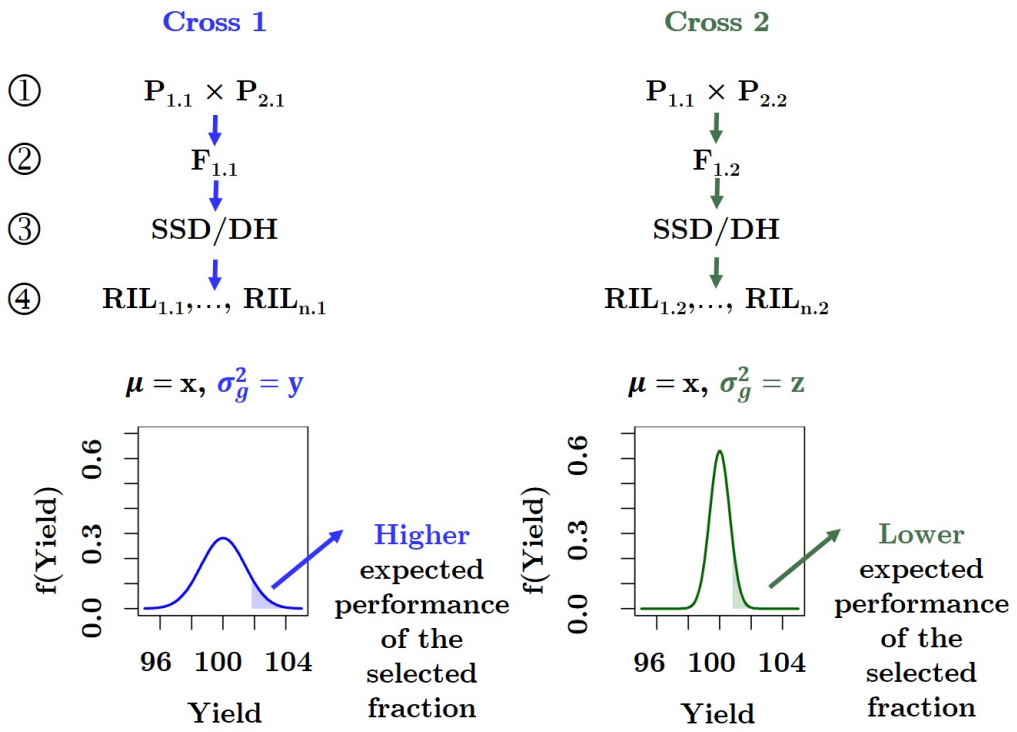


**Figure 1.3.** Distribution of homozygous full-sib progenies for a relevant trait e.g. yield from two biparental populations of two representative crosses within the breeding pool. The scheme of advanced cycle breeding is displayed in four steps. The biparental populations of recombinant inbred lines (RIL) have the same parameters for $\mu$, $h$ and the top 20 % of the RIL are selected. In comparison to cross one (blue), cross two (green) has a smaller $\sigma_g^2$. With otherwise equal framework conditions, the expectation value for the selected fraction of cross one is located further to the right hand side. Cross one has a higher expected performance and it is more likely to find a desirable RIL at the right hand side of the distribution. A preferred choice of crosses with large $\sigma_g^2$ and a high $\mu$ can be rewarded with a larger selection response per cycle.

This problem was addressed by different concepts which evaluate the cross performance. Aside from the concept of usefulness (U) (Schnell & Utz, 1975) and the related superior progeny value (S) (Zhong & Jannink, 2007), different strategies were proposed to guide mating decisions. The U of a cross is determined by the expected population mean of RILs of a trait $\mu$ plus the product of standardized selection intensity $i$, the population standard deviation $\sigma_g$ and the square root of heritability $h$ (Schnell & Utz, 1975). Since molecular markers gained ground in plant breeding, Zhong & Jannink (2007) modified the previous equation. Their S is based on genetically estimated parameters (Zhong & Jannink, 2007). Therefore, they omitted $h$ from the product term and simplified the formula. Today, genotypic values (GV) predicted by genome-wide prediction (GP) models, are used as a mating tool. These predictors for mean performance are good location parameters. Unfortunately, they provide no information about the potential to select high performing progenies of the top fraction (Lado *et al.*, 2017; Lehermeier *et al.*, 2017). Therefore, Cole & VanRaden (2011) predicted the minimum and maximum progeny of a cross in context of animal breeding, analogically Daetwyler *et al.* (2015) introduced the optimal haploid value (OHV) for plants. OHV is a measure for the best homozygous line available after a cross. OHV does not take into account recombination frequency and $i$. Therefore, Han *et al.* (2017) developed the predicted cross value for an optimal introgression of alleles from a donor. Their approach calculates the probability, that a cross produces the ideal genotype. The prerequisite for a practical application, is the identification of desirable alleles, which might be difficult in practical use. Akdemir & Sánchez (2016) proposed the genomic mating as a new tool for breeders. In comparison to GP which focuses on best performance of parents before mating, genomic mating includes information on complementary parents mated to reach higher selection gains in longer term. Genomic mating includes a "risk measure" similar to U to take advantage of within cross variances.

Apart from the above ideas, U and S are straightforward and have been repeatedly discussed in literature. The crosses can be ranked according to

their performance. The application of both theoretical ideas were a research subject for many years. The parameter $\mu$ can be predicted by the average phenotypic performance of the two parents, the mid-parent value (MPV) or by marker data with GVs (Utz *et al.*, 2001). MPVs have been used for a long time in breeding history (Lupton, 1961; Busch *et al.*, 1974; Melchinger *et al.*, 1998; Souza & Sorrells, 1991a; Kotzamanidis *et al.*, 2008; Utz *et al.*, 2001). From previous cycles, the phenotypic performance data of the parental lines is known and can be directly used. Such lines can have very similar $\mu$, but the expectation value for the selected fraction can be different depending on $\sigma_g^2$. Therefore, $\hat{\sigma}_g^2$ and other parameters balance the power for a breeder's decision and influence the selection response. Nevertheless, finding good predictors for $\hat{\sigma}_g^2$ was unsuccessful, even though the problem has been addressed with many different approaches.

## Approaches to predict $\hat{\sigma}_g^2$

The application of U and the S was mainly hindered by the lack of good predictors for $\hat{\sigma}_g^2$. Early quantitative genetics methods were sophisticated and resource-intensive. Variance component estimates (Falconer & Mackay, 1996; Lynch & Walsh, 1998) or approaches like generation mean analysis (Falconer & Mackay, 1996) were accompanied with extra costs for phenotypic evaluations (Bernardo, 2010). Furthermore, these methods are not applicable due to limited time frames in an actual breeding process. Therefore, the plant breeding community seek fast and easy to use prediction tools. A good predictor of $\hat{\sigma}_g^2$ is trait specific and aims to model the mosaic of all possible parental allele variants. A series of different approaches delivered inconsistent and partially context specific positive correlations between $\sigma_g^2$ and the used predictor (Hung *et al.*, 2012; Lian *et al.*, 2015; Tiede *et al.*, 2015a). Phenotypic distances in barley, maize, oat, wheat (Souza & Sorrells, 1991a; Melchinger *et al.*, 1998; Utz *et al.*, 2001; Kuczyńska *et al.*, 2007) and genetic distances based on the coefficient of parentage with pedigree data (Wright,

1922; Kempthorne, 1969) in oat (Cowen & Frey, 1985; Cowen & Frey, 1987; Souza & Sorrells, 1991a; Moser & Lee, 1994), soybean (Helms *et al.*, 1997; Kisha *et al.*, 1997; Manjarrez-Sandoval *et al.*, 1997) and wheat (Bhatt, 1970; Bhatt, 1973; Burkhamer *et al.*, 1998; Bohn *et al.*, 1999) were inconsistent and too weak for practical use. Similarly, ecogeographic diversity (Bhatt, 1973) or multivariate analysis to cluster and classify the parental lines with different characteristics in wheat (Bhatt, 1970; Bhatt, 1973), oat (Souza & Sorrells, 1991a; Souza & Sorrells, 1991b; Souza & Sorrells, 1991c; Moser & Lee, 1994), faba bean (Gumber *et al.*, 1999) or barley (Kuczyńska *et al.*, 2007) did not solve the problem. With the advent of molecular marker systems, genetic distances based on marker data (Nei, 1972) between two inbred lines were used to predict $\hat{\sigma}_g^2$. Just as the previous approaches, distances with restriction fragment length polymorphism, amplified fragment length polymorphism, simple sequence repeats, randomly amplified polymorphic DNA or single nucleotide polymorphism markers were not useful in various crops and model plants (Moser & Lee, 1994; Helms *et al.*, 1997; Kisha *et al.*, 1997; Manjarrez-Sandoval *et al.*, 1997; Burkhamer *et al.*, 1998; Melchinger *et al.*, 1998; Bohn *et al.*, 1999; Gumber *et al.*, 1999; Kuczyńska *et al.*, 2007; Brachi *et al.*, 2010). A series of experiments for different crops and model plants have only shown an insufficient predictive ability. Distance measures are specific for a cross. For practical use, a breeder desires predictors, which capture the individual dynamic of loci influencing $\hat{\sigma}_g^2$ for the trait of interest.

To overcome the limitations of distance measures, quantitative trait analysis (QTL) were used to derive trait specific predictors for $\hat{\sigma}_g^2$. Molecular biological methods like linkage and association mapping are designed to estimate trait specific QTL effects and reveal the genetic trait architecture. This idea originates from van Berloo & Stam (1998), who suggested to use QTL analysis marker information to inter cross pairs of RIL to accumulate advantageous alleles. Their idea was transferred to model trait specific $\hat{\sigma}_g^2$ for unlinked (Bernardo *et al.*, 2006) and linked loci (Zhong & Jannink, 2007). QTL analysis is subject to uncertainty and often reveals artifacts for marker effects that are not reproducible in other experiments. Therefore, Zhong

& Jannink (2007) investigated genome-wide marker effects for their predictions. More recently, simulation studies (Endelman, 2011; Iwata *et al.*, 2013; Bernardo, 2014; Lian *et al.*, 2015; Mohammadi *et al.*, 2015; Tiede *et al.*, 2015a; Lado *et al.*, 2017; Yao *et al.*, 2018; Adeyemo & Bernardo, 2019), analytical approaches for plant (Zhong & Jannink, 2007; Lehermeier *et al.*, 2017) and animal breeding (Cole & VanRaden, 2011; Segelke *et al.*, 2014; Bonk *et al.*, 2016; Wittenburg *et al.*, 2016) based on marker effects estimated with GP approach (Meuwissen *et al.*, 2001) were used to predict the potential of a breeding population. Genetic marker effects associated with a trait are estimated in a training set where phenotypic and genotypic data is available. All marker effects, regardless of a significant threshold, can be used to predict crosses with or without considering LD in a cross.

# Objectives

The aim of my Ph.D. project was to investigate the prediction of $\hat{\mu}$ and $\hat{\sigma}_g/\hat{\sigma}_g^2$ of a cross with a new, fast and user-friendly analytical tool which uses genome-wide marker effects. In particular, my objectives were to:

(1) Present analytical formulas for the expected genotypic $\hat{\mu}$ and $\hat{\sigma}_g^2$ of the lines derived from a cross of two parental lines. The prediction is based on marker effects that are estimated with genomic prediction approaches such as ridge-regression best linear unbiased prediction (RR-BLUP) and accounts for the expected gametic LD between two loci (Chapter 2).

(2) Extend the formulas for breeding populations of two major systems (DH, SSD) used by plant breeders and geneticists to derive lines from a cross after various generations of inter-mating (Chapter 2).

(3) Illustrate the use of the formulas with published data of the nested association mapping population in maize. The results are compared to a previously published simulation approach regarding computation time. Further, predicted and observed parameters are correlated (Chapter 2).

(4) Demonstrate the practical applicability of the analytical approach for cross prediction based on genome-wide marker effects in a real-life barley data set from an ongoing resistance breeding project. Rank and select superior crosses with the S for the next breeding cycle (Chapter 3).

# Chapter 2

# Genomic Selection of Crossing Partners on Basis of the Expected Mean and Variance of Their Derived Lines[1]

[1]Osthushenrich, T, Frisch, M & Herzog, E. 2017. Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. *PLoS ONE*, **12**(12), e0188839.

PLOS ONE

RESEARCH ARTICLE

# Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines

**Tanja Osthushenrich, Matthias Frisch, Eva Herzog***

Institute of Agronomy and Plant Breeding II, Justus Liebig University, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany

* eva.herzog@uni-giessen.de





**Citation:** Osthushenrich T, Frisch M, Herzog E (2017) Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. PLoS ONE 12(12): e0188839. https://doi.org/10.1371/journal.pone.0188839

**Editor:** Aimin Zhang, Institute of Genetics and Developmental Biology Chinese Academy of Sciences, CHINA

**Received:** June 7, 2017

**Accepted:** November 14, 2017

**Published:** December 4, 2017



**Data Availability Statement:** The data used in this study are publically available third-party data. We downloaded the data from the Panzea database which can be accessed via http://www.panzea.org/data The direct URL for downloading the phenotypes of the 282 lines of the association panel, the phenotypes of the NAM RILs, the marker genotypes of the NAM RILs as well as the genetic map used in the present study is http://cbsusrv04.tc.cornell.edu/users/panzea/filegateway.aspx?category=Phenotypes The data used in our manuscript is contained in the zip-folder

## Abstract

In a line or a hybrid breeding program superior lines are selected from a breeding pool as parental lines for the next breeding cycle. From a cross of two parental lines, new lines are derived by single-seed descent (SSD) or doubled haploid (DH) technology. However, not all possible crosses between the parental lines can be carried out due to limited resources. Our objectives were to present formulas to characterize a cross by the mean and variance of the genotypic values of the lines derived from the cross, and to apply the formulas to predict means and variances of flowering time traits in recombinant inbred line families of a publicly available data set in maize. We derived formulas which are based on the expected linkage disequilibrium (LD) between two loci and which can be used for arbitrary mating systems. Results were worked out for SSD and DH lines derived from a cross after an arbitrary number of intermating generations. The means and variances were highly correlated with results obtained by the simulation software PopVar. Compared with these simulations, computation time for our closed formulas was about ten times faster. The means and variances for flowering time traits observed in the recombinant inbred line families of the investigated data set showed correlations of around 0.9 for the means and of 0.46 and 0.65 for the standard deviations with the estimated values. We conclude that our results provide a framework that can be exploited to increase the efficiency of hybrid and line breeding programs by extending genomic selection approaches to the selection of crossing partners.

## Introduction

In each cycle of a line or a hybrid breeding program, lines are selected which serve as the parents of the crosses from which the base population of the next breeding cycle is derived. However, not all possible crosses between the superior lines of a cycle can be made and evaluated due to limited resources. The decision which parental lines to cross is therefore an essential factor that determines the selection gain in a breeding program.

The usefulness of a cross [1] is defined as $U = \mu + i\sigma_g h$, where $\mu$ is the expectation and $\sigma_g$ the standard deviation of the genetic values of the lines derived from the cross, $i$ is the selection

Buckler_etal_2009_Science_flowering_time_data-090807.zip <http://de.iplantcollaborative.org/dl/d/0506883B-CF86-43EA-9BCC-E6682410A932/Buckler_etal_2009_Science_flowering_time_data-090807.zip> The direct URL for downloading the marker genotypes of the 282 lines of the association panel is http://cbsusrv04.tc.cornell.edu/users/panzea/download.aspx?filegroupid=7 <http://cbsusrv04.tc.cornell.edu/users/panzea/filegateway.aspx?category=Genotypes> The data used in our manuscript is contained in the file SNP55K_maize282_AGP3_20190904.hmp.txt.gz <http://de.iplantcollaborative.org/dl/d/76992989-F12A-4F48-A3E3-C60139852E9F/SNP55K_maize282_AGP3_20190904.hmp.txt.gz>.

**Funding:** The author(s) received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

intensity [2], and $h$ the square root of the heritability. The potential of this concept was recognized early, its practical application, however, was hindered by the difficulty of obtaining good estimates for the standard deviation $\sigma_g$. The prediction of the segregation variance $\sigma_g^2$ has therefore been a research subject for many years. A recent review of studies using genetic distances, pedigrees or QTL estimates for this purpose was presented by [3].

Bernardo et al. [4] suggested to estimate the variance $\sigma_g^2$ from QTL effect estimates assuming unlinked loci. This concept was extended to linked loci by Zhong and Jannink [5] who defined, by omitting the square root of the heritability $h$ from the equation of the usefulness, the superior progeny value as $s = \mu + i\sigma_g$. Their approach was developed for recombinant inbred lines derived by single-seed descent (SSD) and uses additive genetic effects estimated by QTL mapping or genome-wide prediction. Several studies investigated with computer simulations the prediction of the genetic variance within a cross [3, 6–8], but fast and versatile analytical solutions for predicting the variance $\sigma_g^2$ for arbitrary mating schemes have to our knowledge not yet been developed.

Our objectives were to (1) present an analytical derivation of $\sigma_g^2$ that is based on the expected linkage disequilibrium (LD) between two loci, and that can be used for arbitrary mating systems, (2) provide formulas for the genetic variance $\sigma_g^2$ in populations of doubled haploid (DH) and SSD lines derived from a cross after $t$ generations of intermating, and (3) illustrate the use of the formulas with published data of the nested association mapping (NAM) population in maize [9].

## Materials and methods

### Derivation of $\sigma_g^2$ based on LD

To derive the superior progeny value $s = \mu + i\sigma_g$ of the cross of two homozygous lines, we define the random variable $Z$ describing the genotypic values of a population of homozygous lines derived from the cross. The expectation of $Z$ is

$$\mu = \mathrm{E}(Z) = \beta_0 + \sum_c \sum_j \mathrm{E}(Z_j), \tag{1}$$

and its variance is

$$\sigma_g^2 = \mathrm{var}(Z) = \sum_c \sum_{j,k} \mathrm{cov}(Z_j, Z_k) \tag{2}$$

where the summation index $c$ sums over chromosomes, $j$ sums over the loci on a chromosome, and $j$, $k$ sums over all locus pairs on a chromosome. $\beta_0$ is the intercept in an additive genetic model and the $Z_j$ are random variables that describe the genetic effect of the allele at locus $j$. $Z_j$ can either be two times the additive effect of the maternal allele, which we denote by $g_j$, or two times the effect of the paternal allele denoted by $h_j$. The event space of $Z_j$ is $\omega_j \in \Omega_j = \{g_j, h_j\}$. The probability that the random variable $Z_j$ takes the value $g_j$ or $h_j$ is

$$P(Z_j = g_j) = P(Z_j = h_j) = 1/2 \tag{3}$$

and, hence, the expectation of $Z_j$ is

$$\begin{aligned} \mathrm{E}(Z_j) &= P(Z_j = g_j)g_j + P(Z_j = h_j)h_j \\ &= \frac{1}{2}(g_j + h_j). \end{aligned} \tag{4}$$

The effects of the maternal and paternal alleles at locus $k$ are $g_k$ and $h_k$. For deriving the covariance between the genotypic values at the two linked loci $j$ and $k$

$$\mathrm{cov}(Z_j, Z_k) = \mathrm{E}(Z_j Z_k) - \mathrm{E}(Z_j)\mathrm{E}(Z_k), \tag{5}$$

we need the expectation of the random variable $Z_jZ_k$ with event space $\omega_{j,k} \in \Omega_{j,k} = \{g_jg_k, g_jh_k, h_jg_k, h_jh_k\}$. To determine the probability of the four events in $\Omega_{j,k}$, we define the conditional probability

$$q_{jk} = P(Z_k = g_k \mid Z_j = g_j) \tag{6}$$

that $Z_k$ takes the value $g_k$ under the condition that $Z_j$ takes the value $g_j$, i.e., the probability that the locus $k$ carries the maternal gamete under the condition that locus $j$ carries the maternal gamete. Using $q_{jk}$, we have

$$\begin{aligned} P(Z_jZ_k = g_jg_k) &= P(Z_j = g_j)P(Z_k = g_k \mid Z_j = g_j) \\ &= \tfrac{1}{2}q_{jk} \end{aligned} \tag{7}$$

and for reasons of symmetry

$$P(Z_jZ_k = g_jh_k) = \tfrac{1}{2}(1 - q_{jk}) \tag{8}$$

$$P(Z_jZ_k = h_jg_k) = \tfrac{1}{2}(1 - q_{jk}) \tag{9}$$

$$P(Z_jZ_k = h_jh_k) = \tfrac{1}{2}q_{jk} \tag{10}$$

which can be used to determine the expectation

$$\begin{aligned} \mathrm{E}(Z_jZ_k) &= \sum_{\omega_{j,k}} P(Z_jZ_k = \omega_{j,k})\omega_{j,k} \\ &= \tfrac{1}{2}q_{jk}(g_jg_k + h_jh_k) + \tfrac{1}{2}(1 - q_{jk})(g_jh_k + h_jg_k). \end{aligned} \tag{11}$$

If Eqs (4) and (11) are inserted into Eq (5), it results

$$\mathrm{cov}(Z_j, Z_k) = \tfrac{1}{2}q_{jk}(g_jg_k + h_jh_k) + \tfrac{1}{2}(1 - q_{jk})(g_jh_k + h_jg_k) - \tfrac{1}{2}(g_j + h_j)\tfrac{1}{2}(g_k + h_k) \tag{12}$$

and, expanding the brackets,

$$\mathrm{cov}(Z_j, Z_k) = (\tfrac{1}{2}q_{jk} - \tfrac{1}{4})(g_jg_k + h_jh_k - g_jh_k - h_jg_k). \tag{13}$$

From the definition of the conditional probability $P(A|B) = P(A, B)/P(B)$ it follows that

$$q_{jk} = P(Z_k = g_k, Z_j = g_j)/P(Z_j = g_j). \tag{14}$$

Using the definition of the LD coefficient

$$D_{jk} = P(Z_k = g_k, Z_j = g_j) - P(Z_j = g_j)P(Z_k = g_k) \tag{15}$$

and that $P(Z_j = g_j) = P(Z_k = g_k) = 1/2$, we can write $q_{jk}$ as a function of the expected LD between the two loci $j$ and $k$ as

$$q_{jk} = \tfrac{1}{2} + 2D_{jk}. \tag{16}$$

## Derivations for SSD and DH lines

Eqs 1–16 can be used to determine the superior progeny value $s$ in terms of the expected LD for two linked loci. It can be used for arbitrary mating systems that were used to derive the populations of homozygous lines from the initial cross. Prerequisite is that the expected LD coefficient $D_{jk}$ and, hence, $q_{jk}$ for the mating system is known or can be derived.

Two major systems used by plant breeders and geneticists to derive inbred lines from a cross are DH lines and recombinant inbred lines developed by repeated selfing with SSD. DH lines are derived from the $F_1$ ($F_1$-DH), or after $t$ generations of random intermating of the $F_1$, which we denote by $(F_1)^t$-DH. SSD lines are derived from the $F_2$, or after $t$ generations of random intermating of the $F_1$, which we denote by $(F_2)^t$-SSD. We present results for $q_{jk}$ for the $(F_1)^t$-DH and $(F_2)^t$-SSD mating systems following the approach of [10].

As an $F_1$-DH population consists of gametes generated by an $F_1$, the probability that such a gamete carries the alleles of one parental line at the loci $j$ and $k$ is

$$P(Z_k = g_k, Z_j = g_j) = \frac{1 - r_{jk}}{2}, \tag{17}$$

where $r_{jk}$ is the recombination frequency between $j$ and $k$. Hence, according to Eq 15, the corresponding LD coefficient is

$$D_{jk} = \frac{1 - 2r_{jk}}{4}. \tag{18}$$

After $t$ generations of random mating, the LD coefficient $D_{jk}$ of the $F_1$ population is reduced by the factor $(1 - r_{jk})^t$ [2]. It follows for $(F_1)^t$-DH lines

$$D_{jk} = \frac{1 - 2r_{jk}}{4}(1 - r_{jk})^t \tag{19}$$

and

$$q_{jk} = \frac{1}{2} + \frac{1 - 2r_{jk}}{2}(1 - r_{jk})^t. \tag{20}$$

The LD coefficient in SSD lines derived from a population in Hardy-Weinberg-Equilibrium is [11]:

$$D_{jk} = \frac{D'_{jk}}{1 + 2r_{jk}}, \tag{21}$$

where $D'_{jk}$ is the LD coefficient in the initial population. The LD coefficient in an $(F_2)^t$ population is the same as in a population of $(F_1)^t$-DH lines (Eq 19), therefore, for $(F_2)^t$-SSD lines

$$D_{jk} = \frac{1}{1 + 2r_{jk}} \frac{1 - 2r_{jk}}{4}(1 - r_{jk})^t \tag{22}$$

and

$$q_{jk} = \frac{1}{2} + \frac{1 - 2r_{jk}}{2 - 4r_{jk}}(1 - r_{jk})^t \tag{23}$$

The recombination frequency $r_{uv}$ can be derived from an arbitrary mapping function. For example, using Haldane's mapping function [12], the recombination frequency between the

map positions that correspond to $Z_j$ and $Z_k$ is

$$r_{jk} = (1 - e^{-2|x_j - x_k|})/2, \quad (24)$$

where $x_j$ and $x_k$ are the map positions of the two loci in Morgan units.

For the estimation of the expectation and the standard deviation of a cross with the presented formulas, the parameter $\beta_0$ and the effects $g_j$ and $h_j$ are predicted with a genome-wide prediction model. Further, a linkage map of the markers used for the estimation of the genetic effects is required for calculation of the recombination frequencies.

## Data set for illustration

To illustrate the application of the formulas and to compare our results to results of simulations with PopVar we used publicly available data that was originally generated for the investigation of flowering time in maize [9, 13–15]. The genotypic and phenotypic data used in the present study were downloaded from www.panzea.org. The data comprised two data sets: an association panel of 282 diverse maize inbred lines and the maize NAM population. The NAM population consists of 25 families of 200 recombinant inbred lines that were derived from crosses of the inbred B73 with 25 lines from the association panel. The 25 parental lines were selected from the association panel to represent its diversity [14, 15]. The field experiments for the association panel and the NAM population were conducted in 2006 and 2007 in eight environments within the US. The phenotypic data were best linear unbiased predictors (BLUPs) for each line from these field experiments. Briefly, field spatial correction was applied within each environment. BLUPs for each line were then predicted using a combined mixed model across environments. A detailed description of the field design and the statistical analysis was presented in [15]. We used the data for the traits days to silking (DTS, female flowering) and days to anthesis (DTA, male flowering). We pre-processed the marker data for the available 1100 SNPs. Only polymorphic markers with a maximum of two alleles, less than 10% missing values and a gene diversity of at least 0.1 were used for estimating the genetic effects. Some individuals from the original 200 progeny per cross and a few lines from the association panel were also discarded as they had more than 10% missing marker data. After cleaning, we performed the final analyses for the present study with 258 diverse lines of the association panel as training set and 4641 recombinant inbred lines from the NAM population as validation set. Each NAM family consisted of 183-200 recombinant inbred lines. For both sets, 325 high-quality SNP markers were available. For calculating the recombination frequencies between marker loci, we used the published linkage map based on the NAM population [14]. The linkage map covered a total genome length of 1400 cM, resulting in an average marker density of one marker every 4.3 cM.

## Prediction of $\mu$ and $\sigma_g^2$

For calculating marker effects, the association panel was used as training set. Marker effects were calculated with ridge-regression best linear unbiased prediction (RR-BLUP) [16]. The marker effects $u_i$ were thus estimated by solving the mixed-model equation

$$\begin{pmatrix} 1'1 & 1'Z \\ Z'1 & Z'Z + \lambda I \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{u} \end{pmatrix} = \begin{pmatrix} 1'y \\ Z'y \end{pmatrix} \quad (25)$$

Employing these genetic effects and the marker genotypes of the parental lines, we estimated $\mu$ and $\sigma_g^2$ for the families of the NAM population with Eqs 1 and 2. The estimated means and

variances were compared with the observed means and variances for DTS and DTA in the NAM families. Effect estimation and estimation of genetic means and variances were implemented in the C programming language.

### Comparison with simulations

In addition to the estimates obtained with Eqs 1 and 2, we estimated $\mu$ and $\sigma_g^2$ for DTS and DTA in the NAM families with the simulation software PopVar [7]. We used the same data sets as input for the simulations as for the calculations with the formulas. As PopVar estimates the marker effects and $\mu$ and $\sigma_g^2$ in one analysis step, it was not possible to use exactly the same marker effects in the simulations as were used for the analytical approach. We used the implemented RR-BLUP routine of PopVar without the default cross-validation option it offers in order to obtain marker effects as similar as possible to the ones used in the formulas. For each NAM family, we simulated 200 progeny with 25 replications. Computing time required for simulations with PopVar and for estimating the means and variances with Eqs 1 and 2 was assessed using a Linux system with Intel x5670 processors. The calculations were carried out single-threaded.

## Results

The means and variances estimated with Eqs 1 and 2 showed correlations between 0.98 and 1 with the average of the 25 simulated estimates obtained from PopVar (Fig 1). Estimating $\mu$ and $\sigma_g^2$ with Eqs 1 and 2 took 3.3 s and 3.5 s for DTA and DTS, respectively, and obtaining the simulated parameters with PopVar took 46.2 s and 45.5 s, respectively. When the estimates from Eqs 1 and 2 were compared to the observed parameters from the NAM population, the correlations between the observed and estimated means of the crosses for DTA and DTS were 0.90 and 0.91, and the correlations between the observed and estimated standard deviations were 0.46 for DTA and 0.65 for DTS (Fig 2). The estimated standard deviations tended to overestimate the observed standard deviations by factors 1.5-3.

## Discussion

### Differentiation to previous approaches

Zhong and Jannink [5] suggested to assess the value of a cross by its superior progeny value $s = \mu + i\sigma_g$, where $\mu$ is the expectation of the genotypic values of the recombinant inbred lines derived by SSD from the cross and $\sigma_g^2$ is the variance of the genotypic values. $\mu$ and $\sigma_g$ were defined in terms of $\alpha_i$, which denotes half of the difference between the two homozygous QTL effects at a locus. This parameterization has the advantage that it can be directly applied for effect estimates from QTL mapping in bi-parental populations. In our notation, the absolute value of the $\alpha_i$ can be expressed as $|g_u - h_u|/2$. Our Eq 2 corresponds to Eq 3 of [5], and is equivalent if employed with an LD coefficient $D_{uv}$ for $F_2$-SSD lines. Our notation has the advantage that the allele effect estimates from genome-wide prediction approaches can be directly used without re-parameterization to $\alpha_i$, and that the formulas also can be applied when effects for multiple alleles were estimated. This is important for applications in which several SNPs are combined to haplotypes.

The analogy between [5] and our approach ends after the authors of [5] presented their Eq 3. For their further derivations, they consider a population of recombinant inbred lines derived from the cross of two inbred lines and they investigate the question which pair of recombinant inbred lines from this population should be crossed to obtain the maximum superior progeny value $s$. The situation we investigate is different. Our approach is not restricted to selection of

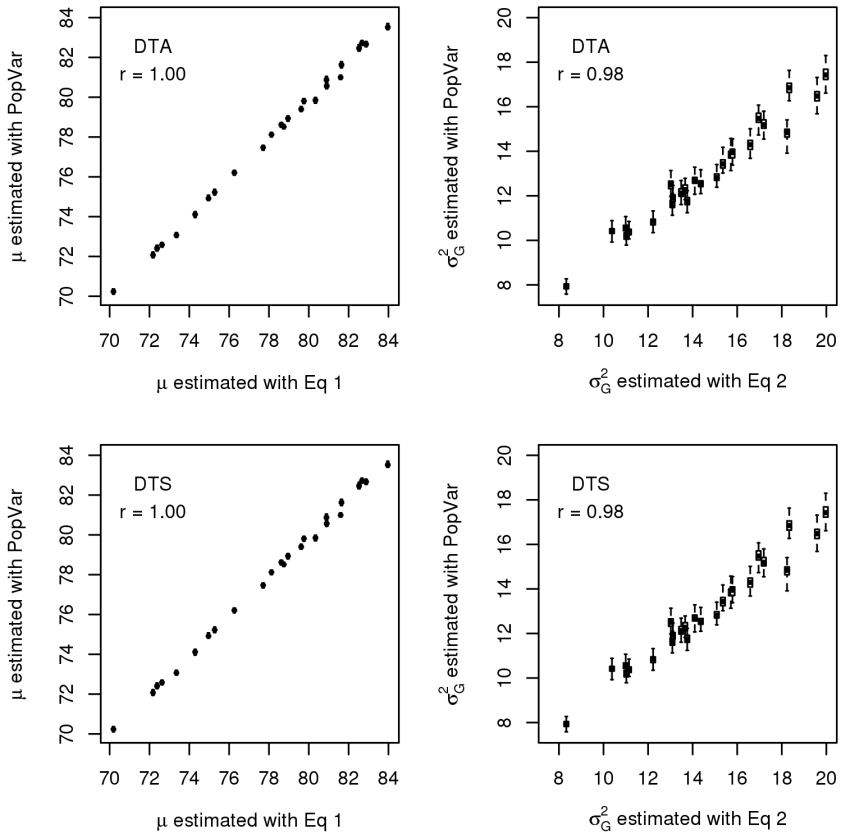


**Fig 1. Correlations between the means (left) and variances (right) estimated with Eqs 1 and 2 and the software PopVar.** Correlations were calculated for the traits DTA (top) and DTS (bottom) in the maize NAM population.

https://doi.org/10.1371/journal.pone.0188839.g001

crossing partners from a set of recombinant inbred lines which were derived from a bi-parental cross. Rather, our approach is targeted at choosing the optimal crossing partners from a set of lines that can be selected from multiple families of a certain cycle of a breeding program to be used as crossing parents for the inbred lines of the subsequent cycle.

As an extension to the approach of [5], our formulas can be flexibly adapted to arbitrary mating systems. Specifically, we presented derivations for DH lines which are used for line development in many breeding programs. Our formulas are also applicable in situations in which several generations of intermating might be required, e.g. if newly introduced diversity needs to be recombined, or if rare transgressive segregants are desired. Moreover, they can be used with arbitrary mapping functions. We therefore believe that our approach has the potential to make the concept of superior progeny value more versatile and applicable in a wider range of crossing scenarios.

In summary, in comparison to the approach of [5], we provide an alternative derivation for the genetic variance of a crossing population that can be used for arbitrary mating systems for which the LD coefficient $D_{uv}$ can be determined. As an advancement, we provide extensions for DH lines and for an arbitrary number of intermatings before deriving inbred lines by either

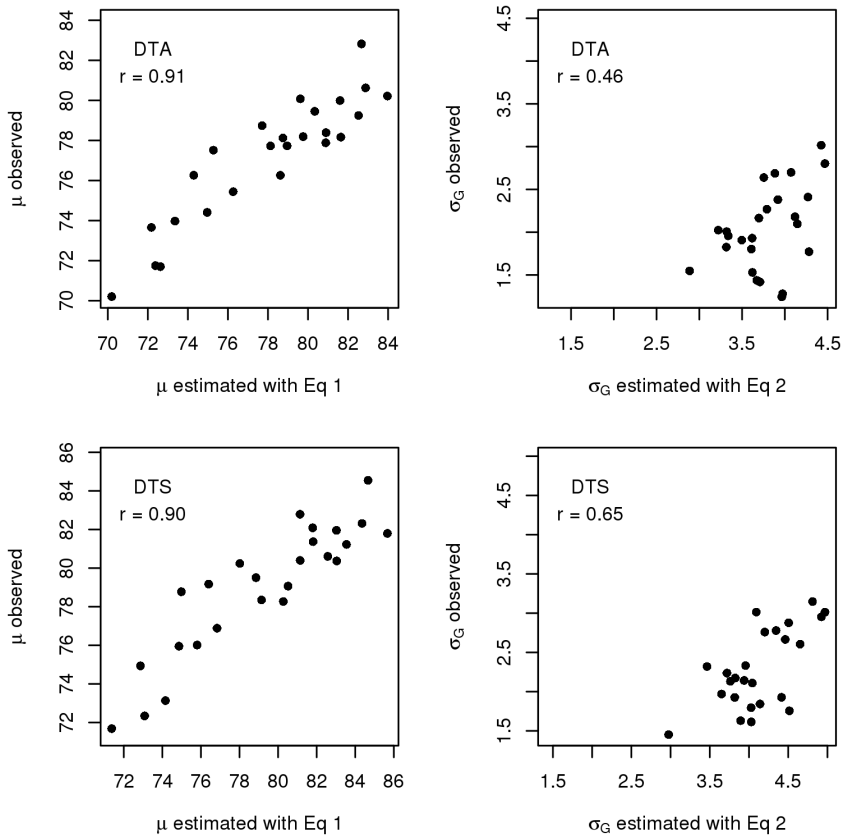


**Fig 2. Correlations between the means (left) and standard deviations (right) estimated with Eqs 1 and 2 and the observed parameters from the maize NAM population.** The observed parameters were estimated from the observed phenotypic values of the NAM population for the traits DTA (top) and DTS (bottom).

https://doi.org/10.1371/journal.pone.0188839.g002

DH or the SSD. Further, we provide an extension from recombinant inbred lines originating from a single bi-parental cross to sets of inbred lines derived from multiple families from a breeding pool with different parents. We suggest to apply the formulas to select crossing partners from such sets of inbred lines as parents for the subsequent breeding cycle.

## Comparison with PopVar simulations

Bernardo [6] and a series of subsequent investigations [3, 7, 8] suggested to use computer simulations with marker effects estimated by genome-wide prediction models to predict the genetic variance $\sigma_g^2$ within a cross. Genetic simulations are modeling the recombination along chromosomes, the distribution of the recombined chromatids to gametes, and the union of gametes. The mathematical model using Haldane's mapping function, which was used in our derivations, is equivalent to the simulation of recombination with a count-location process, using a Poisson distribution for the number of crossovers on a chromosome, and a uniform distribution of the location of crossovers [17].

The software PopVar [3] uses a $\chi^2$ distribution for the number of crossovers on a chromosome and a uniform distribution for the location of the crossovers. This difference in the underlying models is expected to contribute to the differences in the genetic variances predicted with the two approaches in our example data set for traits DTA and DTS (Fig 1). Moreover, due to the implementation of PopVar and the resulting user options, it was not possible to use exactly the same marker effects for estimating the means and variances with PopVar and the derived formulas. However, given that the same mathematically equivalent ridge-regression model was used in both approaches, we expect the differences in marker effects to be negligible. Despite of some small numerical differences in the estimated variances, the two models yielded to a large extent similar results, with correlations of 1 for the mean, and 0.98 for the genetic variances (Fig 1). This could be expected, as both our formulas and PopVar rely on similar assumptions.

However, the major advantage of analytical approaches compared to simulations is speed. In our example data set, only 325 SNP markers were employed, and the analytical approach was about ten times faster as the simulations. It can be expected that the time advantage will be considerably greater with data sets that comprise a greater number of markers.

We conclude that our formulas provide analytical results for the means and variances that are highly correlated to the simulation results, but can be computed more quickly.

## Prediction of genetic variances of the flowering time data set

The accuracy of predicting genetic means and variances obtained when applying our formulas to the flowering time data set is limited by the accuracy of the genetic effect estimates.

In addition to homoscedastic ridge regression, we estimated the marker effects also by using heteroscedastic models [18], but no substantial differences in the results were observed (results not shown).

In data sets from breeding programs, a high marker density is not necessarily required to obtain high accuracies of genomic prediction [19, 20]. This is often attributed to high levels of LD and relatedness in breeding pools. In contrast, here we used a diversity panel consisting of maize lines of different origin as training set. Due to the lower expected LD in such a data set, the number 325 markers seems low for genome-wide prediction.

The number of genotypes in the training set seems also low. In contrast to genetically narrow material from a breeding pool, the number of alleles in the diversity panel is expected to be greater. Hence, the number of replications for each allele in the training set is lower in a diversity set than in material from a breeding program.

Despite of these described properties of investigated data set, the correlations between the estimated and observed means of the crosses for DTA and DTS were 0.90 and 0.91, and the correlations of the estimated standard deviations were 0.46 and 0.65 (Fig 2). We think that these correlations are sufficiently high to create a ranking of crosses based on their superior progeny values *s*. With this ranking, a superior fraction of crosses could be identified and further evaluated to determine the crosses with the highest performance in field trials.

We expect that in breeding pools with longer LD stretches and less allelic variation in combination with a higher marker density, a more precise effect estimation can contribute to greater correlations between estimated and observed segregation variances. We plan further investigations with data sets from breeding programs in sugar beet.

## Application in breeding programs

Due to the comparatively low computation time and versatility with respect to the method of marker effect estimation as well as mating system, the presented formulas can be applied in a

wide range of breeding programs. Breeders working in smaller breeding programs for commercially less important crops might consider it a constraint that a linkage map for the markers used for genomic prediction must be available for calculating recombination frequencies. However, the investment might be worthwhile, as a more precise prediction of superior crosses will not only increase selection gain, but also allow for a more efficient allocation of resources.

A possible application scheme for our formulas is outlined as follows. In a breeding program the marker genotypes as well as the performance data for lines selected as crossing parents for the next cycle are routinely available. These data can be used to estimate genetic effects for the markers. Areas of application include prediction of line *per se* performance in line breeding programs, as well as prediction of testcross performance in hybrid breeding programs. On basis of the marker effect estimates from genome-wide prediction models, the means and standard deviations for each cross are predicted.

The relative superiority of the crosses would depend on the ratio of means and standard deviations, as was also pointed out by [5]. In plant breeding programs, it is very common to recombine best-by-best rather than constantly introducing novel variation from genetic resources with poor agronomic properties and adaptation. This implies that the major proportion of the crosses can be expected to have similar cross means with a low variance of the means. Consequently, the genetic variance gains importance as a decision criterion which crosses to make. This holds even more true if we assume that most elite lines are fixed for the same superior alleles, and that a negative covariance might exist between $\mu$ and $\sigma_g^2$ [5]. In this case, maintaining genetic variance in the breeding pool is a constant challenge in order to guarantee selection gain in the longer term.

The estimation of the genetic variance could also provide a guideline for the allocation of resources to specific crosses. Consider again the superior progeny value $s = \mu + i\sigma_g$. If we compare two crosses of elite lines with the same expectation $\mu$, the difference in selection gain solely depends on the segregation variance $\sigma_g^2$ if the same number of progeny is generated and the same selection intensity $i$ is applied. However, as the relation between $\sigma_g^2$ and $i$ is a multiplicative one, a cross with a moderately higher genetic variance can result in considerably larger selection gain if the selection intensity is increased. Thus, it makes more sense to invest resources and generate more progeny in crosses with higher segregation variance.

In this context, some consideration should be given to the fact that while the magnitude of correlations between estimated and observed means and variances seems reasonable and useful, the estimated standard deviations tended to systematically overestimate the observed values (Fig 2). This might reduce the efficiency of a breeding program. If the systematic upward bias should in general be so large that it considerably changes the relative magnitude of the segregation variance in comparison to the mean, breeders might invest too many resources in terms of family size without any return on investment. It is possible that the upward bias is in part due the fact that the formulas give us the expected value for an infinite population size, while we compared them to observed values from finite populations which might not realize the full potential of segregation variance. However, this is not likely, as our results were close to the simulation results which were also based on finite population sizes. Further possibilities might be overestimation of the actual recombination frequencies by the mapping function, or the choice of the genome-wide prediction model and the resulting marker effect shrinkage [7].

We still think that it is more efficient to plan the size of the single families within breeding programs based on estimates of the segregation variance than to simply create many small families with the same number of progeny, which is common practice in many plant breeding programs. Open research questions in the field of breeding applications therefore comprise breeding designs that include an optimum family size and selection intensity based on

estimated segregation variance. For this goal, the presented approach provides a fast and easy-to-use basis.

## Author Contributions

**Conceptualization:** Matthias Frisch, Eva Herzog.

**Formal analysis:** Tanja Osthushenrich.

**Investigation:** Tanja Osthushenrich.

**Methodology:** Tanja Osthushenrich, Matthias Frisch.

**Software:** Tanja Osthushenrich.

**Validation:** Tanja Osthushenrich.

**Visualization:** Tanja Osthushenrich.

**Writing – original draft:** Tanja Osthushenrich, Eva Herzog.

**Writing – review & editing:** Eva Herzog.

## References


1. Schnell F, Utz H. $F_1$-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern. In: Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter. Bundesversuchsanstalt für alpenländische Landwirtschaft Gumpenstein; 1975. p. 243–248.
2. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. Uk, Longman: Harlow; 1996.
3. Mohammadi M, Tiede T, Smith KP. PopVar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. Crop Science. 2015; 55:2068–2077. https://doi.org/10.2135/cropsci2015.01.0030
4. Bernardo R, Moreau L, Charcosset A. Number and fitness of selected individuals in marker-assisted and phenotypic recurrent selection. Crop Science. 2006; 46:1972–1980. https://doi.org/10.2135/cropsci2006.01-0057
5. Zhong S, Jannink JL. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. Genetics. 2007; 177:567–576. https://doi.org/10.1534/genetics.107.075358 PMID: 17660556
6. Bernardo R. Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. Crop Science. 2014; 54:2586–2595. https://doi.org/10.2135/cropsci2014.01.0088
7. Tiede T, Mohammadi M, Smith K. PopVar: genomic breeding tools: genetic variance prediction and cross-validation. R package version. 2015;1.2.
8. Lian L, Jacobson A, Zhong S, Bernardo R. Prediction of genetic variance in biparental maize populations: Genomewide marker effects versus mean genetic variance in prior populations. Crop Science. 2015; 55:1181–1188. https://doi.org/10.2135/cropsci2014.10.0729
9. Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, et al. Maize association population: A high-resolution platform for quantitative trait locus dissection. The Plant Journal. 2005; 44:1054–1064. https://doi.org/10.1111/j.1365-313X.2005.02591.x PMID: 16359397
10. Frisch M, Melchinger AE. Variance of the parental genome contribution to inbred lines derived from biparental crosses. Genetics. 2007; 176:477–488. https://doi.org/10.1534/genetics.106.065433 PMID: 17409089
11. Cockerham CC, Weir BS. Descent measures for two loci with some applications. Theoretical population biology. 1973; 4:300–330. https://doi.org/10.1016/0040-5809(73)90013-0 PMID: 4747657
12. Haldane J. The combination of linkage values and the calculation of distances between the loci of linked factors. Journal of Genetics. 1919; 8:299–309.
13. Yu J, Holland JB, Mcmullen MD, Buckler ES. Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics. 2008; 178:539–551. https://doi.org/10.1534/genetics.107.074245 PMID: 18202393
14. McMullen MD, Kresovich S, Sanchez Villeda H, Bradbury P. Genetic Properties of the Maize Nested Association Mapping Population. Science. 2009; 325:737–740. https://doi.org/10.1126/science.1174320 PMID: 19661427

15. Buckler ES, Holland JB, Bradbury Peter. The Genetic Architecture of Maize Flowering Time. Science. 2009; 325:714–718. https://doi.org/10.1126/science.1174276 PMID: 19661422

16. Meuwissen T, Hayes B, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157:1819–1829. PMID: 11290733

17. Maurer HP, Melchinger AE, Frisch M. Population genetic simulation and data analysis with Plabsoft. Euphytica. 2008; 161:133–139. https://doi.org/10.1007/s10681-007-9493-4

18. Hofheinz N, Frisch M Heteroscedastic Ridge Regression Approaches for Genome-Wide Prediction With a Focus on Computational Efficiency and Accurate Effect Estimation. G3: Genes, Genomes, Genetics. 2014; 4:539–546. https://doi.org/10.1534/g3.113.010025

19. Hofheinz N, Borchardt D, Weissleder K, Frisch M. Genome-based prediction of test cross performance in two subsequent breeding cycles. Theoretical and Applied Genetics. 2012; 125:1639–1645. https://doi.org/10.1007/s00122-012-1940-5 PMID: 22814724

20. Zenke-Philippi C, Thiemann A, Seifert F, Schrag T, Melchinger A, Scholten S, et al. Prediction of hybrid performance in maize with a ridge regression model employed to DNA markers and mRNA transcription profiles. BMC Genomics. 2016; 17:262. https://doi.org/10.1186/s12864-016-2580-y PMID: 27025377

# Chapter 3

# Prediction of Means and Variances of Crosses With Genome-Wide Marker Effects in Barley[1]

[1]Osthushenrich, T, Frisch, M, Zenke-Philippi, C, Jaiser, H, Spiller, M, Cselényi, L, Krumnacker, K, Boxberger, S, Kopahnke, D, Habekuß, A, Ordon, F & Herzog, E. 2018. Prediction of means and variances of crosses with genome-wide marker effects in barley. *Front Plant Sci* , **9**:1899.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 21 December 2018
doi: 10.3389/fpls.2018.01899

# Prediction of Means and Variances of Crosses With Genome-Wide Marker Effects in Barley

*Tanja Osthushenrich[1], Matthias Frisch[1], Carola Zenke-Philippi[1], Heidi Jaiser[2], Monika Spiller[3], László Cselényi[4], Kerstin Krumnacker[5], Susanna Boxberger[6], Doris Kopahnke[7], Antje Habekuß[7], Frank Ordon[7] and Eva Herzog[1]**

[1] Institute of Agronomy and Plant Breeding II, Justus Liebig University, Gießen, Germany, [2] Saatzucht Josef Breun GmbH & Co. KG, Herzogenaurach, Germany, [3] Syngenta Seeds GmbH, Bad Salzuflen, Germany, [4] W. von Borries-Eckendorf GmbH & Co. KG, Leopoldshöhe, Germany, [5] Limagrain GmbH, Edemissen, Germany, [6] Ackermann Saatzucht GmbH & Co. KG, Irlbach, Germany, [7] Institute for Resistance Research and Stress Tolerance, Julius Kühn-Institute, Quedlinburg, Germany

**Background:** The expected genetic variance is an important criterion for the selection of crossing partners which will produce superior combinations of genotypes in their progeny. The advent of molecular markers has opened up new vistas for obtaining precise predictors for the genetic variance of a cross, but fast prediction methods that allow plant breeders to select crossing partners based on already available data from their breeding programs without complicated calculations or simulation of breeding populations are still lacking. The main objective of the present study was to demonstrate the practical applicability of an analytical approach for the selection of superior cross combinations with experimental data from a barley breeding program. We used genome-wide marker effects to predict the yield means and genetic variances of 14 DH families resulting from crosses of four donor lines with five registered elite varieties with the genotypic information of the parental lines. For the validation of the predicted parameters, the analytical approach was extended by the masking variance as a major component of phenotypic variance. The predicted parameters were used to fit normal distribution curves of the phenotypic values and to conduct an Anderson-Darling goodness-of-fit test for the observed phenotypic data of the 14 DH families from the field trial.

**Results:** There was no evidence that the observed phenotypic values deviated from the predicted phenotypic normal distributions in 13 out of 14 crosses. The correlations between the observed and the predicted means and the observed and predicted variances were $r = 0.95$ and $r = 0.34$, respectively. After removing two crosses with downward outliers in the phenotypic data, the correlation between the observed and predicted variances increased to $r = 0.76$. A ranking of the 14 crosses based on the sum of predicted mean and genetic variance identified the 50% best crosses from the field trial correctly.

**Conclusions:** We conclude that the prediction accuracy of the presented approach is sufficiently high to identify superior crosses even with limited phenotypic data. We therefore expect that the analytical approach based on genome-wide marker effects is applicable in a wide range of breeding programs.

**Keywords: cross prediction, genomic prediction, variance prediction, segregation variance, genetic variance**

OPEN ACCESS

***Edited by:***
*Chengdao Li,*
*Murdoch University, Australia*

***Reviewed by:***
*Thomas Lubberstedt,*
*Iowa State University, United States*
*Martin O. Bohn,*
*University of Illinois at Urbana-Champaign, United States*

****Correspondence:***
*Eva Herzog*
*eva.herzog@agrar.uni-giessen.de*

***Specialty section:***
*This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science*

***Received:*** *30 April 2018*
***Accepted:*** *07 December 2018*
***Published:*** *21 December 2018*

***Citation:***
*Osthushenrich T, Frisch M, Zenke-Philippi C, Jaiser H, Spiller M, Cselényi L, Krumnacker K, Boxberger S, Kopahnke D, Habekuß A, Ordon F and Herzog E (2018) Prediction of Means and Variances of Crosses With Genome-Wide Marker Effects in Barley. Front. Plant Sci. 9:1899. doi: 10.3389/fpls.2018.01899*

## INTRODUCTION

Selection gain in breeding programs relies on the selection of suitable crossing partners which will result in derived lines with superior performance. The best cross is not necessarily the cross with the greatest mean performance, but the cross of which the best lines show the highest performance (Zhong and Jannink, 2007). Looking at the criteria which have been suggested to evaluate the potential of a certain cross to generate high-performing progeny, such as the usefulness criterion $U = \mu + i\sigma_g h$ (Schnell and Utz, 1975) or the superior progeny value $s = \mu + i\sigma_g$ (Zhong and Jannink, 2007), it becomes clear that the expected genetic variance within a cross is the key factor for identifying the best crosses. Nevertheless, strategies for identifying superior crosses in applied breeding programs have so far mostly relied on pedigree information, mid-parent performance and phenotypic evaluation (Lado et al., 2017). The main reason why the selection of crosses on the basis of their progeny variance has not yet been widely implemented in plant breeding programs was that before the advent of molecular markers there were only limited possibilities of obtaining sufficiently precise predictors for these genetic variances.

In the era of high-throughput genotyping and genomic selection, recent research has focused on obtaining predictors for the genetic variance from genome-wide marker estimates by either simulations (Bernardo, 2014; Lian et al., 2015; Mohammadi et al., 2015) or analytical approaches (Zhong and Jannink, 2007; Bonk et al., 2016; Lehermeier et al., 2017). Versatile analytical methods that allow plant breeders to make a fast selection of superior crossing partners based on already available genotypic and phenotypic data from their breeding programs without the need of reparametrization of estimated marker effects, complicated calculations, or simulation of breeding populations promise to improve the efficiency of breeding programs. In a previous study, we have presented an analytical approach for the prediction of the means and genetic variances of crosses based on maker effects estimated by methods of genomic selection that works for arbitrary mapping functions and mating systems (Osthushenrich et al., 2017). First promising results of cross prediction with analytical approaches were published for simulated populations or multi-parental mapping populations (Bonk et al., 2016; Lado et al., 2017; Lehermeier et al., 2017; Osthushenrich et al., 2017). However, as the design of mapping populations deviates from the design of typical breeding populations, the practical applicability of the analytical approaches in plant breeding populations remains to be demonstrated. To our knowledge, no studies are available which investigate the application of analytical approaches for cross prediction for agronomically important complex quantitative traits with data from actual breeding populations.

The aims of the present study were to apply the analytical formulas for prediction of the means and variances of crosses by Osthushenrich et al. (2017) to a data set from a resistance breeding project in barley, and to investigate the model fit for yield in 14 families of doubled haploid (DH) lines derived from crosses of four pre-breeding lines and five registered commercial elite varieties. Our objective was to investigate the practical relevance and applicability of our analytical approach for the identification of superior cross combinations in plant breeding programs.

**Abbreviations:** DH, doubled haploid.

## MATERIALS AND METHODS

### Genetic Material

For a resistance breeding project the registered six-row barley varieties Jenny (JEN, Saatzucht Breun), KWS Meridian (MER, KWS Saat SE), Otto (OTT, W. von Borries-Eckendorf), Etincel (ETI, Secobra), and Quadriga (QUA, Secobra) were crossed with the resistance donor lines BAZ 2L101 (101), BAZ 2L146 (146), DH 33 (D33), DH 37 (D37) developed by the Julius Kühn Institute and the registered variety Antonella (ANT, Nordsaat). The resistance donor lines carried resistances to either barley yellow dwarf virus (BYDV; 101, 146), net blotch (*Pyrenophora teres* f. *teres*; D33, D37), or were a registered variety (ANT) carrying resistance to net blotch, powdery mildew (*Blumeria graminis*) and scald (*Rhynchosporium commune*). By crossing each registered elite variety with each donor line, respectively, a 5 × 5 factorial cross was attempted. However, not all crosses were successful and yielded viable offspring (**Table 1**). Different numbers of $F_1$-DH lines were produced from each successful cross, resulting in 250 $F_1$-DH lines in total (**Table 1**). The genetic relationship between parental lines and the emerging DH lines are displayed in a principal coordinate analysis in **Figure 1**.

### Field Data

An augmented design with five blocks was used to evaluate all genotypes for yield in one year at five locations in Germany with one replication per location. The field experiment was carried out in Adenstedt (State Niedersachsen, Region Südhannover), Harzhof (State Schleswig-Holstein, Region Ost-Holstein), Irlbach (State Bayern, Region Niederbayern), Lenglern (State Niedersachsen, Region Südniedersachsen), Morgenrot (State Sachsen Anhalt, Region Östliches Harzvorland). The parental lines were used as checks and were replicated five times.

**TABLE 1 |** Size n of the families of DH lines derived from the crosses of Parent 1 x Parent 2.

| | Parent 1 | | | | |
|---|---|---|---|---|---|
| Parent 2 | 101 | 146 | ANT | D33 | D37 |
| | n | | | | |
| ETI | 14 | 13 | 18 | 20 | 0 |
| JEN | 12 | 7 | 0 | 0 | 0 |
| MER | 19 | 10 | 16 | 22 | 0 |
| OTT | 4 | 10 | 4 | 0 | 13 |
| QUA | 1 | 10 | 12 | 8 | 37 |

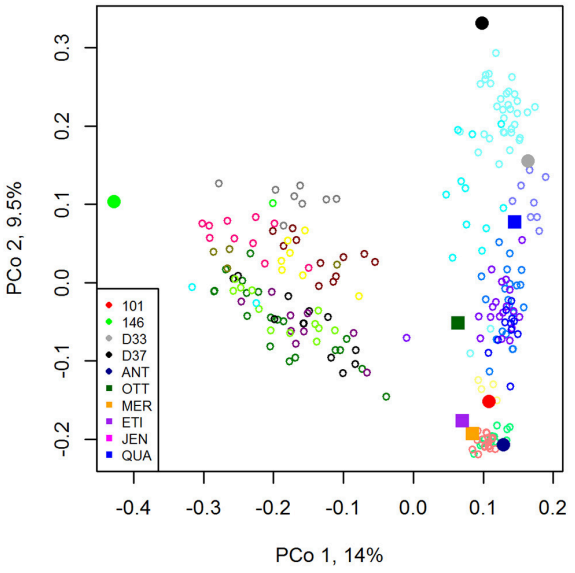


**FIGURE 1 |** Principal coordinates analysis plot with nine parental lines and 239 emerging DH lines. Parental lines: filled symbols. Circles: resistance donors, Squares: elite lines. DH lines: bordered circles with 19 different colors representing the crosses with viable offspring.

The field data were analyzed with the mixed linear model

$$Yield \sim \mu + Genotypes + Location + Location : Blocks + Error$$

where the common mean $\mu$ and genotypes were treated as fixed factors, whereas blocks, locations, and heteroscedastic model errors were assumed as random. The resulting adjusted entry means for yield for each DH line were used in further calculations.

## Genotypic Data

All 250 resulting DH lines and the ten parental lines were genotyped with the 50 k iSelect Chip (Trait Genetics, Gatersleben). All SNP markers with more than two recorded alleles, more than 10% missing values and a gene diversity of <10% were excluded from the analysis, as well as all individuals with more than 15% missing marker information. As a result, 9,597 SNP markers and 259 genotypes (249 DH lines and 10 parental lines) remained for the analysis.

## Genomic Prediction of Marker Effects

For the prediction of marker effects, we used ridge-regression best linear unbiased prediction (Meuwissen et al., 2001). As training set for the prediction of marker effects we used the complete genotypic and the phenotypic data of the 249 DH lines from the 5 × 5 factorial which remained after data cleaning.

## Prediction of Cross Parameters $\hat{\mu}_g$ and $\hat{\sigma}_g^2$

For the prediction of the expectation $\hat{\mu}_g$ and the genetic variance $\hat{\sigma}_g^2$ of the crosses we used the analytical approach of Osthushenrich et al. (2017) and the marker effects estimated with RR-BLUP. The required recombination frequencies were derived from a published linkage map (Bayer et al., 2017). We used the genotypes of the ten parental lines to predict $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ of the resulting DH lines of the validation set.

## Validation Set

For validating the prediction of $\hat{\mu}_g$ and $\hat{\sigma}_g^2$, we compared the predictions from the formulas with the observed phenotypic values $\bar{x}$ and $s_p^2$ from the field trial. As validation set, we used the 200 DH lines resulting from the following 14 crosses: 146ETI, 146JEN, 146MER, 146OTT, 146QUA, ANTETI, ANTMER, ANTOTT, ANTQUA, D33ETI, D33MER, D33QUA, D37OTT, D37QUA. The remaining crosses did not result in viable offspring. For line 101, the resulting DH lines from all five crosses had to be excluded from the validation set, as the genotype of the parental line 101 did not match the genotype of the resulting DH lines, meaning that a problem with seed identification of the parental line had at some point occurred during the project. The final validation set thus comprised an unbalanced 5 × 4 factorial of 14 families of 200 DH lines in total (**Table 1**).

### Comparison of Predicted $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ and Observed Parameters $\bar{x}$ and $s_p^2$

For comparing the predicted and the observed values from the field trial, we used the yield data of the validation set (**Table 1**). As the variance of the phenotypic data is defined as $\sigma_p^2 = \sigma_g^2 + \sigma_m^2$, the approach of Osthushenrich et al. (2017) was extended by an estimate of the distribution of the phenotypic data by adding an estimate $s_m^2$ of the masking variance $\sigma_m^2$ to the predicted variance $\hat{\sigma}_g^2$. For this purpose, the masking variance $s_m^2$ was estimated as the square of the average standard error of the adjusted treatment mean of the mixed models analysis of the field trial (Piepho and Möhring, 2007).

Due to the balanced design of the field trial, the estimated masking variance $s_m^2$ resulted in the same value of 33.41 dt²/ha² for all 14 crosses. An Anderson Darling goodness-of-fit test (Anderson and Darling, 1954) was carried out to test the null hypothesis that the observed yield values of the 14 DH families are a sample from a normal distribution $\mathcal{N}\left(\hat{\mu}_g, \hat{\sigma}_g^2 + s_m^2\right)$.

## Ranking of Crosses

To validate the identification of superior cross combinations with the analytical approach of Osthushenrich et al. (2017), we created a ranking of crosses based on the criterion $\hat{\mu}_g + \hat{\sigma}_g$. This predicted ranking of the crosses was compared to the ranking of crosses based on the best-performing DH line from each cross.

## Software

The statistical analysis of the field data was conducted in R version 3.4.2 (R Core Team, 2017). The estimation of marker effects as well as the prediction of the means and genetic variances of the crosses was conducted in R version 3.4.2 with the software package SelectionTools, which is freely available for download

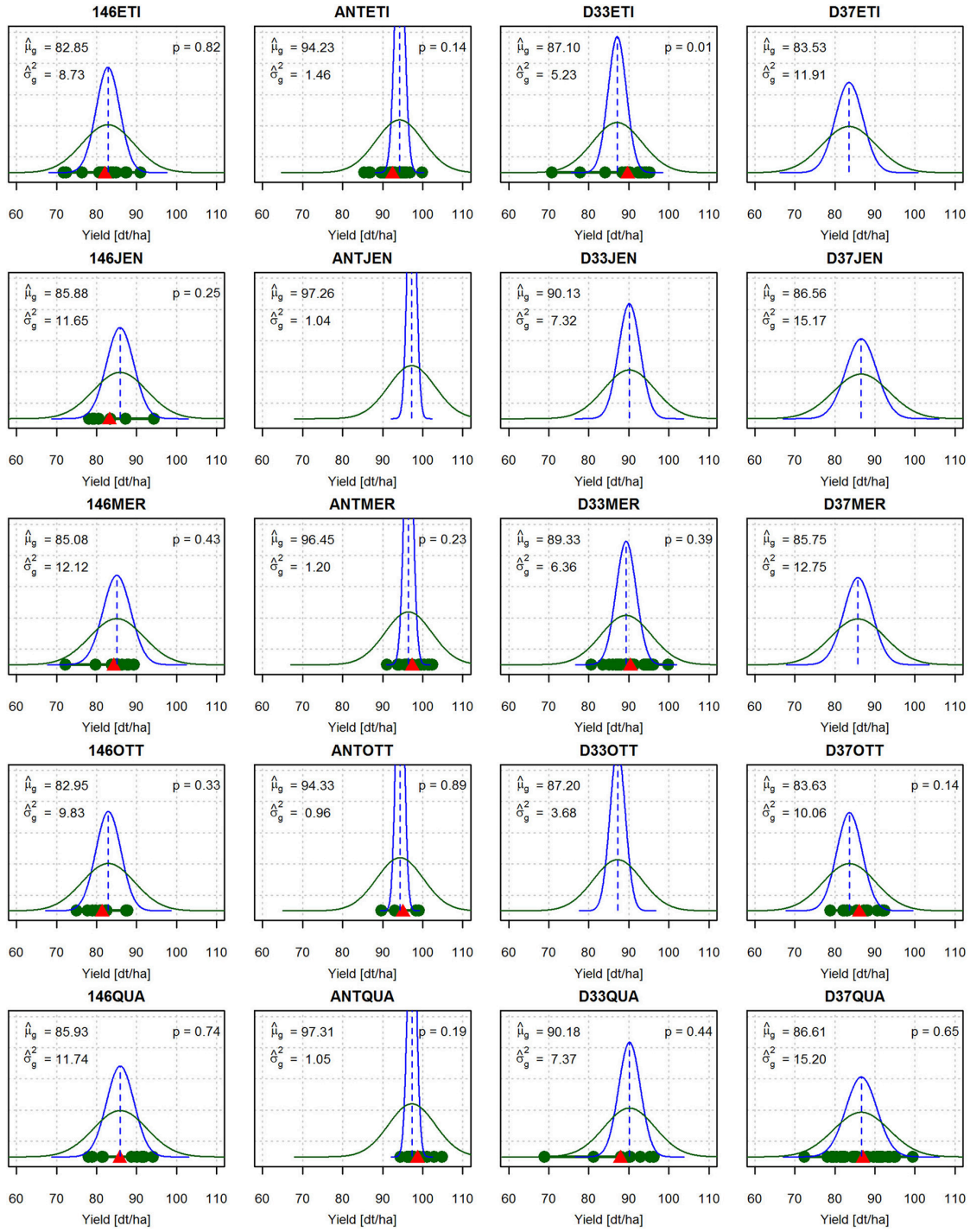


**FIGURE 2 |** Marker-based predictions of the genetic means $\hat{\mu}_g$ and variances $\hat{\sigma}^2_g$ of the DH lines derived from all crosses of the complete factorial. The density of a normal distribution with the predicted genetic parameters $\mathcal{N}\left(\hat{\mu}_g, \hat{\sigma}^2_g\right)$ is depicted in blue. The density of a normal distribution with the predicted phenotypic

*(Continued)*

**FIGURE 2 |** parameters $\mathcal{N}\left(\hat{\mu}_g, \hat{\sigma}_g^2 + s_m^2\right)$ is depicted in green, where $s_m^2$ is the estimated masking variance obtained as the square of the standard error of the adjusted phenotypic means of the field trial. For the crosses for which field data are available, the adjusted treatment means are marked with green dots and the respective family means $\bar{x}$ with red triangles. $p$ is the $p$-value of the Anderson-Darling goodness-of-fit test for the null hypothesis that the observed adjusted treatment means are a sample of a normal distribution $\mathcal{N}\left(\hat{\mu}_g, \hat{\sigma}_g^2 + s_m^2\right)$.

from our homepage[1]. A code and output example is available in **Figure 5**.

## RESULTS

The observed mean yield performance $\hat{\mu}_g$ of the crosses ranged from 82.85 dt/ha (146ETI) to 97.31 dt/ha (ANTQUA) (**Figure 2**). The genetic variances $\hat{\sigma}_g^2$ ranged from 0.96 dt²/ha² (ANTOTT) to 15.20 dt²/ha² (D37QUA). The differences between the predicted yield means $\hat{\mu}_g$ and the genetic variances $\hat{\sigma}_g^2$ were larger between crosses of the same elite variety with different donor lines (columns of **Figure 2**) than between crosses of the same donor line with different elite varieties (rows of **Figure 2**). For example, the crosses of the elite variety QUA with four donor lines showed a comparatively large variation of $\hat{\mu}_g$ and ranged between 85.93 dt/ha and 97.31 dt/ha (last row of **Figure 2**). The genetic variance $\hat{\sigma}_g^2$ also showed a comparatively large variation and ranged between 1.05 dt²/ha² and 15.20 dt²/ha². In contrast, for the five crosses with donor line 146, $\hat{\mu}_g$ for yield ranged only between 82.85 dt/ha and 85.93 dt/ha, and $\hat{\sigma}_g^2$ ranged only between 8.73 dt²/ha² and 12.12 dt²/ha² (first column of **Figure 2**). Crosses with donor line ANT, which is a highly resistant elite variety, displayed the overall highest values of $\hat{\mu}_g$ and the lowest values of $\hat{\sigma}_g^2$ (second column of **Figure 2**).

The crosses D33ETI and D33QUA showed downward outliers which resulted in high observed phenotypic variances $s_p^2$ of 36.57 dt²/ha² and 80.64 dt²/ha² (data not shown, but outliers visible in **Figure 2**). The phenotypic variances of the other twelve crosses with viable offspring ranged between 9.38 and 36.46 dt²/ha² (data not shown). The estimate of the masking variance based on the average standard error from the field data was $s_m^2 = 33.41$ dt²/ha² and thus was higher than the observed phenotypic variances for ten out of 14 crosses (data not shown).

The Anderson-Darling goodness-of-fit test indicated that there is no evidence to reject the null hypothesis that the observed yield values (**Figure 2**, green dots) are sampled from a normal distribution $\mathcal{N}\left(\hat{\mu}_g, \hat{\sigma}_g^2 + s_m^2\right)$ (green curves) in 13 out of 14 crosses. The exception was cross D33ETI which featured downward outliers and a left-skewed sample distribution ($p = 0.01$).

The correlation between the observed yield means $\bar{x}$ (**Figure 2**, red triangles) and the predicted yield means $\hat{\mu}_g$ was $r = 0.95$ (data not shown). The correlation between the observed phenotypic variance $s_p^2$ and the predicted genetic variance $\hat{\sigma}_g^2$ was $r = 0.34$ for all 14 crosses (data not shown). However, when the two crosses D33ETI and D33QUA with downward outliers were removed, this correlation increased to $r = 0.76$ (data not shown).

A comparison of the ranking of crosses based on the observed yield data of the best resulting DH line from each cross with the ranking of the crosses based on the criterion $\hat{\mu}_g + \hat{\sigma}_g$ which relied on the predicted parameters showed that the prediction accuracy was sufficient to correctly identify the 50% best crosses (**Figure 3**).

A negative covariance existed between $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ for all crosses (**Figure 4**). However, when the five potential crossing partners were regarded separately for each donor line, the covariances between $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ were positive.

## DISCUSSION

Despite the recent large interest in methods of cross prediction and the selection of promising crossing partners based on marker data in the plant breeding community (Bernardo, 2014; Lian et al., 2015; Mohammadi et al., 2015; Bonk et al., 2016; Han et al., 2017; Lado et al., 2017; Lehermeier et al., 2017), the application of the published analytical approaches was either demonstrated with simulated data sets or in mapping populations which are not comparable in their structure to typical breeding populations. No studies are available to our knowledge in which the applicability of analytical approaches for marker data was tested for relevant traits such as yield in plant breeding data sets. In the present study, we tested if the formulas for variance prediction presented in Osthushenrich et al. (2017) show sufficient precision for the identification of the most promising crossing partners in an ongoing resistance breeding project in barley.

The data set in use in this investigation was not specifically designed for a rigorous validation of the formulas of Osthushenrich et al. (2017). For such a validation study, several parameters would need a different experimental design. We outline these parameters here to show the limits of the present evaluation.

The present study uses a set of intercrossed lines as a training set, and we evaluate the genetic variance in the same data set. Consequently, the results presented here cannot be regarded as an independent validation. Instead, we are rather investigating the fit of the model to the data. If the model does not fit the data in such an analysis, the conclusion can be drawn that the model is not suitable to explain the data. If the model is able to explain the data, however, a considerable overfitting of the model might still be present, because genomic prediction is an $p > n$ problem where the number of independent variates ($p$, markers) is greater than the number of observations ($n$, lines).

[1]www.uni-giessen.de/population-genetics Homepage of the Department of Biometry and Population Genetics, Institute for Agronomy and Plant Breeding II, Justus Liebig University Giessen. www.uni-giessen.de/population-genetics Accessed 21 February 2018.1514.

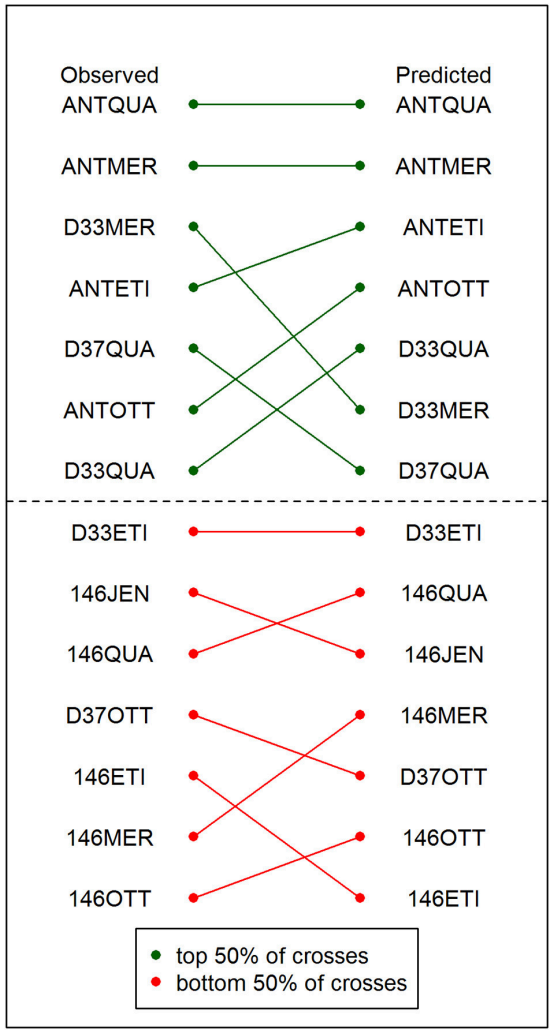


**FIGURE 3 |** Comparison of the observed and the predicted ranking of crosses. The observed ranking is based on the yields of the best DH lines resulting from each cross in the field trial (**left**). The predicted ranking of crosses is based on the criterion $\hat{\mu}_g + \hat{\sigma}_g$ from the predicted cross parameters (**right**). The 50% top-ranked crosses are depicted in green. The 50% bottom-ranked crosses are depicted in red. The green and red lines indicate how the position of the crosses has changed between the observed and the predicted ranking.

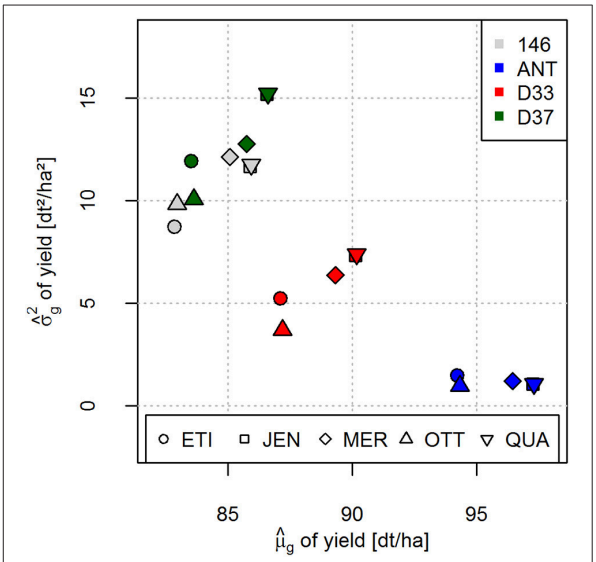


**FIGURE 4 |** Predicted genetic variances $\hat{\sigma}^2_g$ plotted against the predicted cross means $\hat{\mu}_g$ for yield for the 20 crosses of four donor and five registered elite varieties. Crosses with the different donor lines are distinguished by different colors (right corner legend). Crosses with different elite varieties are distinguished by different symbols (bottom legend). Specific crosses can be identified by the combination of color and symbol.

This potential overfitting was not quantified by the analyses we present here.

We are using only small numbers of lines per cross. The estimates of the phenotypic variances within each cross are therefore not estimated with a high precision, but instead they have large standard errors and large confidence intervals. In an experiment designed to validate the formulas for variance prediction, larger family sizes would be desirable.

Due to their large standard errors, we decided not to further decompose the per-cross variances into genetic variance and within-cross residual variance. Such an analysis would have the advantage of being able to compare genetic within-cross variances, and in addition would be able to model cross-specific residuals. Nevertheless, the estimation errors of genetic variances are large, even for experiments that were designed specifically for that purpose, and in the present data set we consider the precision of per-cross estimates of genetic variances as too low for drawing valid conclusions. For this reason, we decided to present only the phenotypic per-cross variances, and to compare those with the masking variance estimated across all crosses. This enables an explorative comparison of the magnitudes of the variance components. In a purposely designed experiment, the estimation of per-cross genetic variances and their comparison with the predicted genetic variance would provide not only an explorative comparison but rather would allow more stringent hypothesis testing.

The field trial in our experiment consisted of five replications for each genotype, this resulted in a limited precision of the phenotypic data. As a consequence, the masking variance in our experiment still amounts to considerable size. In a validation experiment carrying out replicated trials in more than five locations and more than one year would result in a smaller masking variance. Ideally, the design of the validation experiment should result in a masking variance that is smaller than the within-family variance. This would allow an effective within-family selection. Further, it would be desirable if the validation experiment was of a size that allowed heteroscedastic error variances for locations or even for the location:cross combinations.

```
library("SelectionTools")                          # attaches package to search list
st.read.marker.data("markerdata.mpo")              # reads in the marker data
st.read.map("linkagemap.map")                      # reads in the linkage map
st.read.performance.data("phenotypicdata.dta")     # reads in the phenotypic data
gs.esteff.rr("BLUP")                 # genome-wide prediction of marker effects
gs.cross.eval.mu()                   # predicts cross means
gs.cross.eval.va(pop.type = "DH")    # predicts the genetic variances
gs.cross.eval.mi()                # predicts the minimum haplotype values
gs.cross.eval.ma()                # predicts the maximum haplotype values
gs.cross.eval.es(alpha = 0.25)    # superior progeny value for selected fraction alpha
gs.cross.eval.gd()                # genetic distances of the crossing partners
results <- gs.cross.info()   # sorts and saves the results of the cross prediction
head(results)                # output: list of all pairwise crosses and parameters
   P1No  P2No  P1Name  P2Name gd       mu       mi       ma        va       es
1     1     2  RGS001  RGS002 0.257409 83.36176 60.53005 104.74666 4.438759 86.03978
2     1     3  RGS001  RGS003 0.233721 81.25195 60.26730 102.59095 2.288458 83.17484
3     1     4  RGS001  RGS004 0.269231 80.34271 58.75304 103.28010 2.659816 82.41575
4     1     5  RGS001  RGS005 0.251289 83.39370 60.87191 104.29224 3.488243 85.76772
5     1     6  RGS001  RGS006 0.000435 78.19133 62.90081  97.39307 0.003951 78.27122
6     1     7  RGS001  RGS007 0.243835 81.98775 60.69462 103.17051 2.868892 84.14072

...
```

**FIGURE 5 |** Demonstration of R Code used for cross prediction with package SelectionTools.

A further issue that is not addressed with our experimental setup is the question of whether random genetic drift or selection during the DH process might have an effect on the estimated variances, this might also be addressed in a validation experiment.

Our motivation to use the present data set in spite of its limitations and in spite of the fact that it was not specifically designed for validation of formulas for variance prediction was, that it actually originates from a practical breeding program. Our argumentation is that the results presented here have a high transferability to applied breeding programs, whereas the results of a pure validation study would have only a limited transferability due to differences in the experimental setup.

The prediction of the yield means $\hat{\mu}_g$ and genetic variances $\hat{\sigma}_g^2$ of the 14 crosses of five registered elite varieties and four resistance donors for which phenotypic data was available yielded overall plausible results (**Figure 2**). For example, for the crosses of the elite variety QUA with four donor lines (last row of **Figure 2**), $\hat{\mu}_g$ for yield ranged between 85.93 dt/ha and 97.31 dt/ha and $\hat{\sigma}_g^2$ ranged between 1.05 dt²/ha² and 15.20 dt²/ha². For the five crosses with donor line 146 (first column of **Figure 2**), $\hat{\mu}_g$ for yield ranged only between 82.85 dt/ha and 85.93 dt/ha and $\hat{\sigma}_g^2$ ranged between 8.73 dt²/ha² and 12.12 dt²/ha². Differences between the crosses in $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ were thus more influenced by donor lines (columns of **Figure 2**) than by the elite varieties (rows of **Figure 2**), indicating that the elite varieties contributed little to the genetic variance $\hat{\sigma}_g^2$ of the crosses and had similar mean performance $\hat{\mu}_g$. This is also illustrated by the fact that all crosses of elite varieties with donor line ANT, which is also a highly resistant elite variety, had a comparatively high $\hat{\mu}_g$ and a low $\hat{\sigma}_g^2$ compared to the other crosses. These findings are reflected in the varying spread of the blue normal distribution curves in **Figure 2** with $\mathcal{N}\left(\hat{\mu}_g, \hat{\sigma}_g^2\right)$ for the different crosses. It is also confirmed by the corresponding values for the observed yield means $\bar{x}$ (red triangles) and the observed variances $s_p^2$ from the field trial (data not shown).

While a direct comparison of $\hat{\mu}_g$ and $\bar{x}$ from the field trial is straightforward and yielded a correlation of $r = 0.95$ (data not shown), a direct comparison of $\hat{\sigma}_g^2$ predicted from genetic marker effects with the estimated phenotypic variance $s_p^2$ from the field trials is problematic and less straightforward.

The data set used in the present study comprises field data from only one year, a very limited number of locations and only one replication. In such a small data set, large standard errors are expected for the estimation of the phenotypic variance $s_p^2$, which result in large confidence intervals. A confidence interval for an observed variance $s^2$ of a normal distribution is defined as Bronshtein et al. (2003):

$$\frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2};\,n-1}} \leq \sigma^2 \leq \frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2};\,n-1}}$$

For example, $s_p^2$ of the 13 yield values in the field trial for cross 146ETI was 32.63 dt²/ha², resulting in a large 0.95 confidence interval of [16.78; 88.91]. From this we can deduce that the point estimator of the phenotypic variance has only limited accuracy. Moreover, marker-based predictions of $\hat{\sigma}_g^2$ are predictions of the genetic variance within a cross, whereas the variance of the true observed values in a field trial is $\sigma_p^2 = \sigma_g^2 + \sigma_m^2$, where $\sigma_g^2$ is the genetic variance and $\sigma_m^2$ is the masking variance due to environmental effects and inaccuracies of the field trial (Piepho and Möhring, 2007). In the present study, the $s_m^2$ estimated from the field trial was 33.41 dt²/ha², while the predicted genetic variances $\hat{\sigma}_g^2$ ranged between 0.96 dt²/ha² for the cross ANTOTT to 15.20 dt²/ha² for the cross D37QUA. Thus, $s_m^2$ was in all crosses about 2–30 times higher than $\hat{\sigma}_g^2$, and was consequently the major component of the phenotypic variance $\hat{\sigma}_p^2$.

To account for $\sigma_m^2$ in our comparison of predicted and observed variances $\hat{\sigma}_g^2$ and $s_p^2$, we fitted the green normal distribution curve $\mathcal{N}\left(\hat{\mu}_g, \hat{\sigma}_g^2 + s_m^2\right)$. We conducted an Anderson-Darling goodness-of-fit test to test the hypothesis that the phenotypic yield values of the DH lines from the field trial are samples from these normal distributions (Anderson and Darling, 1954). There was no evidence that this null hypothesis could be rejected for 13 out of 14 crosses (**Figure 2**). However, when looking at the absolute values of the observed phenotypic variances $s_p^2$ (data not shown) and the predicted phenotypic variances $\hat{\sigma}_p^2$, our prediction of $\hat{\sigma}_p^2 = \hat{\sigma}_g^2 + s_m^2$ tended to overestimate the observed variance $s_p^2$ of the phenotypic values.

This overestimation could be expected, as precise field trials to assess the yield are only carried out for a limited number of pre-selected individuals, while the analytical approach yields estimates for infinite unselected population sizes. Moreover, the crosses D33ETI and D33QUA featured downward outliers that might have inflated the average standard error for the adjusted treatment means and consequently the derived masking variance $s_m^2$. Under the assumption that the masking variance $\sigma_m^2$ is constant for all crosses, the correlation $r$ between the predicted genetic variance $\hat{\sigma}_g^2$ and the observed phenotypic variance $s_p^2$ gives an idea how valid the predictions for the evaluation of suitable crossing partners are. This correlation was $r = 0.34$ for all 14 crosses (data not shown). However, this was also mainly due to the crosses D33ETI and D33QUA, which each displayed outliers in the form of two very low yield values (**Figure 2**), resulting in high observed variances $s_p^2$ of the phenotypic values. Excluding these two crosses, the correlation increased to $r = 0.76$ (data not shown). From these findings, we draw two conclusions. First, low correlations between the predicted genetic variances $\hat{\sigma}_g^2$ and the observed phenotypic variances $s_p^2$ can be caused by outliers in the field trial which result in overestimated phenotypic variances. They do not necessarily mean that the prediction approach in itself is faulty or inaccurate. Rather, accurate field trials are of major importance not only for estimating marker effects and cross prediction, but also for the plausible validation of cross prediction. The evaluation of the accuracy of cross prediction should therefore comprise a careful monitoring of the field data. Estimates of the phenotypic variance $s_p^2$ from samples with outliers should be treated with caution. Second, the results shown in **Figure 2** indicate that our predictions of $\hat{\sigma}_g^2$ overall yielded reasonable results in light of the limitations of the available phenotypic data.

Despite the fact that the predicted genetic variances $\hat{\sigma}_g^2$ are difficult to validate with phenotypic data from breeding programs, they can still improve the efficiency of breeding programs with respect to long-term response to selection and efficient use of the limited plot number for field trials. Even for lower correlations between $\hat{\sigma}_g^2$ and $s_p^2$ it is reasonable to focus on crosses with high predicted genetic variance in order to maintain genetic diversity and long-term response to selection, given that reliable phenotypic and genotypic data is available for predicting marker effects.

More importantly, we argue that the main application of cross prediction in practical breeding programs is not so much to provide 100 percent accurate predictions of $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ but to allow the breeder to identify a certain fraction of promising crosses from the complete list of potential crosses in order to use the limited number of field plots efficiently. We compared the ranking of the crosses based on the criterion $\hat{\mu}_g + \hat{\sigma}_g$ to the ranking of the crosses based on the yield data of the best resulting DH line from each cross (**Figure 3**). In this comparison, all seven top-ranked crosses were identified correctly with the predicted parameters, allowing the breeder to efficiently narrow down the number of lines which have to be evaluated in costly field trials by 50% without reduction in selection gain.

It has been postulated that a negative covariance exists between $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ (Zhong and Jannink, 2007). This suggestion is very reasonable, as elite varieties which are fixed at many loci for superior alleles will result in crosses with high $\hat{\mu}_g$ and low $\hat{\sigma}_g^2$. This negative covariance is also observed in our data set if $\hat{\mu}_g$ is plotted against $\hat{\sigma}_g^2$ (**Figure 4**). For example, the ANT crosses can be considered as crosses between two elite varieties and consequently have a comparatively high $\hat{\mu}_g$ and low $\hat{\sigma}_g^2$ compared to the other crosses. In our data set, in line with the suggestions of Zhong and Jannink (2007), genetic variances $\hat{\sigma}_g^2$ were more influenced by donor lines (columns of **Figure 2**) than by the elite varieties (rows of **Figure 2**), indicating that the elite varieties contributed little to the genetic variances $\hat{\sigma}_g^2$ of the crosses. Crosses of elite varieties with donor lines 146, D33 and D37 which are pre-breeding lines with overall lower agronomic performance have lower $\hat{\mu}_g$ and higher $\hat{\sigma}_g^2$ in comparison to the ANT crosses.

Thus, we observed that the negative covariance between $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ of the crosses is mainly due to the different level of breeding intensity and selection that the donor lines have been subjected to (**Figure 4**). If the crosses of donor lines are regarded separately, as indicated by the different colors in **Figure 4**, a positive covariance existed between $\hat{\mu}_g$ and $\hat{\sigma}_g^2$. We therefore conclude that for many scenarios, for example if a specific donor line carrying desired resistance genes has to be used for trait introgression into the breeding pool, prediction of the genetic variance $\hat{\sigma}_g^2$ allows the breeder to identify the best crossing partner for this donor line from a set of different elite varieties. In addition, these predictions can also be used for improved resource allocation by investing more resources in terms of

number of progeny into crosses with higher predicted genetic variance $\hat{\sigma}_g^2$. We plan further investigations in this area.

In order to provide breeders with a fast and easy-to-use tool to implement the presented approach in their breeding pipelines, routines for data pre-processing, estimation of marker effects and cross prediction with the formulas of Osthushenrich et al. (2017) have been included in the software package SelectionTools. SelectionTools allows breeders to make use of the advantages of cross prediction in a convenient way without the need of comprehensive mathematical and programming skills. With standard data formats, the presented approach can be reproduced with only a few lines of R code (**Figure 5**).

## CONCLUSION

The analytical approach of Osthushenrich et al. (2017) yields plausible cross predictions which allow breeders to establish a ranking of potential crosses and identify a superior fraction of crosses for field evaluation. The approach is versatile and can be used for arbitrary mating systems. A major advantage of the presented approach is that it can be directly and easily used with marker effects from genome-wide prediction without time-consuming additional calculations or simulations. The prediction accuracy of means and variances is sufficiently high for practical application to derive meaningful predictions even with limited phenotypic data. We therefore expect that the formulas are applicable in a wide range of breeding programs.

## AVAILABILITY OF DATA AND MATERIAL

The datasets generated and/or analyzed during the current study are not publicly available due to the confidential genotypic data of the donor lines from an ongoing research project but are available from the corresponding author on reasonable request.

## AUTHOR CONTRIBUTIONS

HJ, MS, LC, KK, SB, AH and DK developed the genetic materials, conducted the field and greenhouse experiments. CZP and MF planned the field experiments and analyzed the field data. TO analyzed the genotypic data. TO and EH wrote the manuscript. MF, FO, and EH directed the project, contributed to the analyses and manuscript revisions. All authors proof-read the draft and approved the final manuscript.

## FUNDING

This work was funded by the German Federal Ministry of Food and Agriculture (BMEL), grant no. 2818203515. The funding body has no roles in the design of the study or collection, analysis, and interpretation of data or in writing the manuscript.

## REFERENCES

Anderson, T. W., and Darling, D. A. (1954). A test of goodness of fit. *J. Am. Stat. Assoc.* 49, 765–769.

Bayer, M. M., Rapazote-Flores, P., Ganal, M., Hedley, P. E., Macaulay, M., Plieske, J., et al. (2017). Development and evaluation of a barley 50k iSelect SNP array. *Front. Plant Sci.* 8:1792. doi: 10.3389/fpls.2017.01792

Bernardo, R. (2014). Genomewide selection of parental inbreds: classes of loci and virtual biparental populations. *Crop Sci.* 54, 2586–2595.doi: 10.2135/cropsci2014.01.0088

Bonk, S., Reichelt, M., Teuscher, F., Segelke, D., and Reinsch, N. (2016). Mendelian sampling covariability of marker effects and genetic values. *Genet. Sel. Evol.* 48:36. doi: 10.1186/s12711-016-0214-0

Bronshtein, I. N., Semendyayew, K. A., Musiol, G., and Muehlig, H. (2003). *Handbook of Mathematics*. 4th Edn. Berlin; New York, NY: Springer.

Han, Y., Cameron, J. N., Wang, L., and Beavis, W. D. (2017). The predicted cross value for genetic introgression of multiple alleles. *Genetics* 205, 1409–1423. doi: 10.1534/genetics.116.197095

Lado, B., Battenfield, S., Guzmán, C., Quincke, M., Singh, R. P., Dreisigacker, S., et al. (2017). Strategies for selecting crosses using genomic prediction in two wheat breeding programs. *Plant Genome* 10. doi: 10.3835/plantgenome2016.12.0128

Lehermeier, C., Teyssèdre, S., and Schön, C. C. (2017). Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. *Genetics* 207, 1651–1661. doi: 10.1534/genetics.117.300403

Lian, L., Jacobson, A., Zhong, S., and Bernardo, R. (2015). Prediction of genetic variance in biparental maize populations: genomewide marker effects versus mean genetic variance in prior populations. *Crop Sci.* 55, 1181–1188. doi: 10.2135/cropsci2014.10.0729

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

Mohammadi, M., Tiede, T., and Smith, K. P. (2015). PopVar: a genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. *Crop Sci.* 55, 2068–2077. doi: 10.2135/cropsci2015.01.0030

Osthushenrich, T., Frisch, M., and Herzog, E. (2017). Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. *PLoS ONE* 12:e0188839. doi: 10.1371/journal.pone.0188839

Piepho, H. P., and Möhring, J. (2007). Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* 177, 1881–1888. doi: 10.1534/genetics.107.074229

R Core Team (2017) *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna. Available online at: https://www.R-project.org/ (Accessed February 21, 2018).

Schnell, F., and Utz, H. (1975). *F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern*. Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter. Bundesversuchsanstalt für alpenländische Landwirtschaft Gumpenstein, 243–248.

Zhong, S., and Jannink, J. L. (2007). Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics* 177, 567–576. doi: 10.1534/genetics.107.075358

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Chapter 4

# General Discussion

## Strategies to predict $\mu$ and $\sigma_g^2$ in plant breeding

### Simulations

A series of authors used computer simulations to estimate cross parameters. The R-package rrBLUP named after the ridge-regression best linear unbiased prediction approach, provides a tool to simulate RILs of a cross in a breeding program which uses genome-wide marker effects assuming linkage equilibrium (Endelman, 2011). They calculate $\hat{\mu}$ and $\hat{\sigma}_g^2$ from a point cloud of predicted GVs for crosses among wheat lines but did not compare their results to observed data. Also Bernardo (2014) introduced and demonstrated a similar concept. In comparison, the virtual populations of progeny GVs consider LD structure in the population. A corresponding simulation tool is the ready-to-use R-package PopVar, which simulates GVs including LD information by using a genetic map for the prediction of $\hat{\sigma}_g^2$ (Mohammadi *et al.*, 2015; Tiede *et al.*, 2015a). Lado *et al.* (2017) compared simulation methodologies from R-package rrBLUP and PopVar to predict $\hat{\sigma}_g^2$ in two wheat breeding program assuming linkage equilibrium (Endelman, 2011) and LD (Mohammadi *et al.*, 2015). The first method does not require marker positions to simulate recombination points, while for PopVar a genetic map is necessary. Modeling $\hat{\sigma}_g^2$ to apply different selection criteria on their crosses had less

impact for quantitative traits than for the quality traits. They found no differences in selection using $\hat{\sigma}_g^2$ with linkage equilibrium and LD for all traits, even though simulations with LD had larger predicted $\hat{\sigma}_g^2$ values. They preferred the less computational demanding method with linkage equilibrium. Assuming linkage equilibrium sets the covariance between markers to zero. Based on their findings, Endelman's (2011) approach might be an alternative for smaller breeding programs, where the genetic map is not available. Markers in coupling phase have a positive covariance and in repulsion phase a negative one. Lado *et al.* (2017) showed a predominance of coupling phases on most chromosomes which explains the larger $\hat{\sigma}_g^2$.

Simulations have the advantage to predict trait correlations and more complicated models including e.g. epistasis (Yao *et al.*, 2018). At the same time simulations of progenies with an infinite population size have sampling errors, are computing resource intensive and therefore very time consuming for even minor data sets. New deterministic formulas deliver fast and highly correlated results with only small numerical differences (Osthushenrich *et al.*, 2017). Analytical approaches calculate all possible gametes with a finite population size. In result, compact formulas were developed for different breeding objectives.

## Analytical approaches

Analytical approaches aim to explain quantitative genetics relationships in a mathematical context and circumvent time consuming simulations. Closed formulas have a speed advantage (Osthushenrich *et al.*, 2017). Bernardo *et al.* (2006) calculated $\hat{\sigma}_g^2$ as the sum of $\hat{\sigma}_g^2$ across significant markers from QTL analysis. They assumed that all markers are on different chromosomes, and therefore, are in linkage equilibrium. Zhong & Jannink (2007) pointed out the importance of covariance structures between QTL markers, and developed analytical formulas to predict $\hat{\sigma}_g^2$ for linked loci. Their value of a cross depends on the performance of its best progeny that can be derived

by a cross among the RILs. Based on formulas of Zhong & Jannink (2007), Osthushenrich *et al.* (2017) extended the formulas for choosing the optimal crossing partners. The original term $\alpha$ (half of the difference between homozygous lines) was adapted, therefore marker effects from GP models can be directly used without re-parameterization. They also introduce formulas for different mating systems a breeder can use to develop new inbred lines with SSD and DH. The formulas were implemented in the R-package SelectionTools, which is straightforward for application. Analytical and simulation results from SelectionTools and PopVar showed a similar outcome. Lehermeier *et al.* (2017) derived similar formulas and compared different selection procedures like U, GVs or OHV (Daetwyler *et al.*, 2015). In all their cases, genetic gain based on U and predicted parameters from marker effects were better than from GVs or OHV. The core principal is similar to Osthushenrich *et al.* (2017) but they used Bayesian approaches without an applicable software tool.

## Strategies to predict $\hat{\mu}$ and $\hat{\sigma}_g^2$ in animal breeding

In animal breeding mating decisions are of great interest. An additional challenge is the predominant heterozygous population structure. The prediction of the gamete phases of non-inbreds is more complex and the derivation of diplotypes is accompanied with unknown bias (Bonk *et al.*, 2016). Cole & VanRaden (2011) predict the minimum and maximum Mendelian sampling $\hat{\sigma}_g^2$ similar to the OHV approach (Daetwyler *et al.*, 2015) with unlinked and fully linked additive marker effects. Segelke *et al.* (2014) and Wittenburg *et al.* (2016) used simulations of potential progenies, which consider the contribution of LD for genomic mating programs. Bonk *et al.* (2016) used gametes from parents with known diplotypes and recombination rates to predict Mendelian sampling covariances of GVs. In comparison to simulations, their approach avoids the Monte Carlo error typically for *in silico* methods. They used additive and dominance marker effects and found a good average

agreement between the exact values and the simulation results for a practical data set.

### Software and usability

For all strategies, the practical applicability is of great importance. Therefore, user-friendly software that can be easily integrated into the breeding pipeline is desirable. Good examples for implemented breeder's tools are published in open source R-packages, like rrBLUP (Endelman, 2011), PopVar (Tiede *et al.*, 2015a) and SelectionTools (Osthushenrich *et al.*, 2018). Especially for $\hat{\sigma}_g^2$ prediction approaches, SelectionTools has the advantage of a compact computational complexity, various genome-wide prediction models, options for different mating systems like SSD and DH and an example in a barley breeding program (Osthushenrich *et al.*, 2018). In comparison to simulation approaches, the computational efficiency of analytical approaches is better. A comparison of routines for $\hat{\sigma}_g^2$ prediction implemented in PopVar and SelectionTools, the closed formulas were about ten times faster with marginal numerical differences. Whenever possible, plant breeders prefer to work with homozygous populations and implement hybrid breeding pipelines. However, this approach is not possible for all plant species. Therefore, software packages like SelectionTools could be extended for heterozygous populations to enlarge it's fields of application. Once this is established, it can be applied as an universal tool in plant and animal breeding programs.

## Factors with impact on the accuracy of $\sigma_g^2$ prediction

Prediction of $\mu$ in biparental populations by genome-wide marker effects is robust (Bernardo, 2014; Mohammadi *et al.*, 2015; Tiede *et al.*, 2015b; Lado

*et al.*, 2017; Osthushenrich *et al.*, 2017). Modern $\hat{\sigma}_g^2$ prediction methods are based on marker effect estimates and predicted recombination events. In some cases $\hat{\sigma}_g^2$ prediction is more prone to errors from inaccurate marker effects (Zhong & Jannink, 2007). Both simulation and analytical approaches are affected by errors and bias that occur in upstream processes. Even though the main application and benefit of $\hat{\sigma}_g^2$ prediction is for breeding programs, a validation study requires special properties. Ideally, we are looking for large families of unselected progenies that were evaluated in field experiments with high power, to calculate cross specific $\hat{\sigma}_g^2$ compounds or precise $\sigma_p^2$. Data sets that meet the criteria for a validation study are hard to find. Therefore, the approaches were evaluated based on empirical data sets, which are not specially designed for a validation study (Mohammadi *et al.*, 2015; Bonk *et al.*, 2016; Lado *et al.*, 2017; Lehermeier *et al.*, 2017; Yao *et al.*, 2018; Allier *et al.*, 2019). Nevertheless, these may help to identify the crucial factors.

Genome-wide marker effects can be estimated with various models from different types of training populations, that differ in sample size, relatedness, marker density, trait heritability and architecture, evaluation of phenotypic data in different environments and LD between markers and QTL. All these factors can influence the $\hat{\sigma}_g^2$ prediction. The properties of training populations have a large impact on success of $\hat{\sigma}_g^2$ prediction. The best results were observed for marker effects estimated in training populations which have a relevant relationship structure to the prediction set (Riedelsheimer *et al.*, 2013; Bernardo, 2014; Isidro *et al.*, 2015; Sallam *et al.*, 2015). Apart from the general GP approach, this has a significant impact on the cross prediction (Mohammadi *et al.*, 2015; Yao *et al.*, 2018; Adeyemo & Bernardo, 2019). A low heritability, sparse markers or insufficient population size probably have a negative effect on the accuracy of $\hat{\sigma}_g^2$ estimation. Dense marker spacing increased accuracy of $\hat{\sigma}_g^2$ due to higher accuracy of marker effects (Zhong & Jannink, 2007). Higher trait heritability had a positive impact on the prediction accuracy of marker effects. That applies to initial GP studies (Wimmer *et al.*, 2013) and was observed for prediction of $\hat{\sigma}_g^2$ and U (Zhong & Jannink, 2007; Lian *et al.*, 2015; Tiede *et al.*, 2015b; Lehermeier *et al.*, 2017;

Osthushenrich *et al.*, 2017; Yao *et al.*, 2018).

A comparison of different genome-wide prediction models only showed minor influence on the prediction of $\hat{\sigma}_g^2$ (Lado *et al.*, 2017; Osthushenrich *et al.*, 2017; Yao *et al.*, 2018). Nevertheless, there may be superior models for various trait architectures, like in some studies in the original research context of GP (Zhong *et al.*, 2009; Daetwyler *et al.*, 2010; Hayes *et al.*, 2010; Riedelsheimer *et al.*, 2012; Wimmer *et al.*, 2013). The existing studies using plant material are mainly based on additive marker effects. Including epistasis (Jiang & Reif, 2015), dominance (Bonk *et al.*, 2016) or genotype environmental effects (Heslot *et al.*, 2014; Lopez-Cruz *et al.*, 2015) could improve the accuracy. This could be a subject for further research. For fully homozygous lines, dominance effects are negligible (Hill *et al.*, 2008; Mohammadi *et al.*, 2015), but for lines developed with SSD, a small rest heterozygosity is likely. For heterogeneous livestock populations dominance variance in a model is very reasonable (Bonk *et al.*, 2016). Modeling epistasis for large training sets has been challenging for the GP approach (Jiang & Reif, 2015).

Genetic maps are useful to model crossing over during meiosis. Unfortunately, the recombination rates are not family specific, but summarized in a consensus map. A study for the nested association mapping population has shown the diverse recombination pattern within the families (McMullen *et al.*, 2009). Further, the populations used for generating the map are not necessarily closely related to the material used for predictions. This source of error was not investigated at this point.

## Validation approaches

Only a few studies correlated predicted and observed progeny distributions. In spite of all efforts, a strict validation was limited by special features

of the collected data. Different data sets like breeding material or mapping populations were used to correlate the parameters. A real validation study most likely requires expensive experiments with large unselected family sizes, many crosses and precise field trails.

Breeding material has the advantage of genetically narrow material with longer LD stretches, less allelic variation and is fixed for superior alleles evaluated with relatively high marker density. Rather small and selected family sizes and a limited precision of phenotypic data can hamper a validation study. Observed $\sigma_p^2$, estimated from small selected families, have large standard errors and confidence intervals (Osthushenrich *et al.*, 2018). Mapping populations have very often wide crosses which are not typically found in the in breeding context. On the other hand, they can have family sizes of 200 RILs which are evaluated across multiple environments (McMullen *et al.*, 2009). Therefore, the observed $\sigma_p^2$ can have a smaller sampling error. The disadvantage of this population structure is the diverse parental material with lower expected LD and great number of alleles. For validation, the number of families and the lower number of common markers for both sets and the size of the training set could be problematic. A hybrid of both kind of data sets would balance the disadvantages, but is not feasible in either one or the other issue.

A majority of the validation studies with diverse data sets showed positive results. The correlations give reasonable results of the ranking of observed and predicted $\sigma^2$ (Segelke *et al.*, 2014; Lian *et al.*, 2015; Tiede *et al.*, 2015b; Osthushenrich *et al.*, 2017; Osthushenrich *et al.*, 2018; Neyhart & Smith, 2019). Lian *et al.* (2015) used empirical data of 85 related biparental maize breeding populations. They found significant correlations between predicted and observed parameters ($r = 0.18 - 0.52$), but preferred their proposed mean-variance-model. Estimates predicted with the mean-variance-model were less prone to bias. This concept uses phenotypic data from prior related populations and is not based on genome-wide marker effects. A validation study with 40 barley breeding populations found promising results

for fusarium head blight resistance ($r = 0.61$) and did not observe such large bias like Lian *et al.* (2015) (Tiede *et al.*, 2015b). In a subsequent study, they investigated hybrid traits in the training set of hybrid crop breeding programs (Beckett *et al.*, 2019). A prior knowledge of the most promising parental pairs may be beneficial to improve the genetic gain. More recently, a study which focuses on the selection of training populations was published. They investigated 284 diverse maize inbreds to predict eight crosses among non-related diverse inbreds to identify the top 10 % of the crosses with the U criterion. The virtual populations had an erratic effectiveness for predicting the $\hat{\sigma}_g^2$ (Adeyemo & Bernardo, 2019). Opposed to their previous study (Lian *et al.*, 2015), where they used related biparental crosses evaluated in various environments, they used a set of diverse inbreds evaluated in only one environment to predict distributions of eight crosses. While $\hat{\mu}_g$ was predicted reliably, the $\hat{\sigma}_g^2$ and U predictions were insufficient. They discuss four reasons that explain the poor association in their study. The first reason is the weak link of training and prediction set. Second, they found large confidence intervals of observed $\sigma_p^2$, even though they used $120 - 144$ progenies per family. The recombination rates are cross-specific and may not be pictured by the consensus map. Investigated $F_3$ plants may be confounded by dominance variance due to rest heterozygosity. With their specially designed experiment, they clearly demonstrated the demerits of the $\hat{\sigma}_g^2$ prediction. A breeder can include the sensitive issues whether the new approach might be useful in the breeding process. For such a data set, they suggested to use their mean-variance-model, but did not investigate the accuracy. A user has to consider carefully, weather the data set can be evaluated with GP-based approaches.

Certain studies observed an underestimated $\hat{\sigma}_g^2$ (Lian *et al.*, 2015; Lehermeier *et al.*, 2017), but Osthushenrich *et al.* (2017) observed an overestimation. Adeyemo & Bernardo (2019) observed a downward and upward bias for different traits. A possible reason for downward bias is the marker effect shrinkage, but this does not explain the upward bias (Cole & VanRaden, 2011; Tiede *et al.*, 2015b; Lehermeier *et al.*, 2017; Adeyemo & Bernardo,

2019). Lado *et al.* (2017) discussed the predominance of coupling phases, but still, this is a question for further research. Neyhart & Smith (2019) investigated a LD based simulation approach of $\hat{\sigma}_g^2$ for three relevant quantitative traits within a genomic selection based breeding program. Comparing observed and predicted values, they found a relative consistency of this bias across the family validation, which allows a selection of crosses. They concluded that accurate predictors of $\hat{\sigma}_g^2$ are feasible but reliable phenotypic data are critical for the implementation. This supports earlier results of Osthushenrich *et al.* (2018). Based on their analytical approach, they reported meaningful ranking of crosses in a breeding program, although they faced limited phenotypic data. Generally, $\hat{\sigma}_g^2$ prediction has to be based on a solid foundation of GP. Therefore, the rigors validation with an optimal data set remains an open research question.

A fair comparison of $\sigma_p^2$ and $\hat{\sigma}_g^2$ seems difficult for breeding populations. Since $\hat{\sigma}_g^2$ is derived for unselected progeny populations with infinite population size. Typically estimated $\hat{\sigma}_g^2$ is compared to observed $\sigma_p^2$. In small family sizes, large standard errors and confidence intervals, limited precision of phenotypic data, random genetic drift and selection in the DH generations can limit the accuracy. Specially designed experiments with large number of families and family sizes can improve the precision (Segelke *et al.*, 2014). In animal breeding context, they discussed a number of 150 offspring per sire as a compromise to provide a good data foundation. Adeyemo & Bernardo (2019) observed large confidence intervals for eight crosses and did not find meaningful nor significant correlations for family sizes of 120-144 progenies. Similar to Segelke *et al.* (2014), the number of crosses or matings are important to find significant correlations. Already one outlier can confound the ranking (Osthushenrich *et al.*, 2018). The observed values can be confounded by environmental errors and inaccuracies of the field trails. It would be therefore advisable to account for this major component of $\sigma_p^2$ (Osthushenrich *et al.*, 2018) or use observed $\sigma_g^2$ for the comparison. A further decomposition of genetic variance to estimate variance components for each family and the accuracy of $\sigma_g^2$ remains to be examined.

## Application in breeding process

The prerequisites for an integration into the breeding pipeline are a training set which includes the crossing parents, marker genotypes, performance data and linkage map of the relevant markers. These data sets are routinely available in larger breeding programs for new lines and hybrids. The functions in the R-package SelectionTools provide all parental combinations of the breeding material with only little effort and time. An overview of all crosses offers a good starting point for an evaluation based on objective formulas, subjective breeder's adjustment and other information resources. This provides the opportunity to select parents for the next cycle and allocate resources to the most potential crosses. Another application is the ranking and preselection of the top and bottom fraction, to have a more efficient selection step (Osthushenrich *et al.*, 2018). The ranking of crosses is subject of consistent bias over all families and therefore, is useful to select promising crosses. (Neyhart & Smith, 2019). Beckett *et al.* (2019) see the most valuable application in evaluation of untested but genotyped lines in the training set and elimination of crosses which appear promising by pedigree and MP-values but have a small $\hat{\sigma}_g^2$. A different context is the selection of parental lines for mapping populations to maximize the power of QTL detection (Hung *et al.*, 2012). A simultaneous selection of multiple traits, to circumvent undesirable correlations, is discussed as a new application based on genome-wide prediction. The application can support maintaining genetic diversity for the subsequent cycles and improve short and long term goals (Yao *et al.*, 2018). This idea was examined by different authors (Akdemir & Sánchez, 2016; Yao *et al.*, 2018; Allier *et al.*, 2019; Neyhart *et al.*, 2019).

## Relevance of $\hat{\mu}$ and $\hat{\sigma}_g^2$ for cross performance

Parameter $\mu$ can be predicted with very high accuracy. Both approaches, the MPV from phenotypic and genotypic data deliver good results. Zhong

& Jannink (2007) discussed the limited parameter space of $\hat{\sigma}_g^2$. It is generally accepted, that $\mu$ has the bigger influence on selection gain. It depends on the ratio of the variation in $\hat{\mu}$ and $\hat{\sigma}_g^2$ in the population where we select in. In general, $\hat{\sigma}_g^2$ gains importance for situations with similar cross $\mu$ and a difference in $\sigma_g^2$. Therefore, $\sigma_g^2$ can provide additional information for cross evaluation. Osthushenrich *et al.* (2018) found large $\hat{\sigma}_g^2$ for crosses among elite and donor lines and small $\hat{\sigma}_g^2$ for crosses among elite lines. Overall, they observed a strong negative correlation of $\hat{\sigma}_g^2$ and $\hat{\mu}$. But within crosses among elite and several resistance donor lines, they observed a positive correlation. This variation can be used for selection of the most promising cross among an elite line and a potential donor. The goal is to discriminate among common crosses in a breeding pool and identify a certain fraction of promising crosses (Tiede *et al.*, 2015b; Osthushenrich *et al.*, 2018). For Lado *et al.* (2017) $\hat{\mu}$ was the best approach to select the parents. However, they recommended to consider genetic diversity in the breeding process which could be more important for qualitative traits (Bernardo, 2014; Lado *et al.*, 2017; Yao *et al.*, 2018). Other studies (Lehermeier *et al.*, 2017; Yao *et al.*, 2018) found a higher genetic gain by selection with U than selection only with the MPV. Parameter $i$ had a strong influence on the selection with U, a large proportion of selected progenies reduced the weight of the $\hat{\sigma}_g^2$ in U application and the selection gain is more similar to a selection based on $\mu$ (Yao *et al.*, 2018). A problem is the erratic bias that was observed in different validation studies. However, findings of Neyhart & Smith (2019) are encouraging. They observed a consistent bias distributed across the families and found a plausible selection basis with their simulation approach. Nevertheless, a breeder might be misled and invest to many or less resources in a certain cross and this is a risk for the realized selection gain (Osthushenrich *et al.*, 2017).

# Conclusions

The modern methods for $\sigma_g^2$ prediction and application of U are good tools which can support the breeder decision, but are not a standalone criterion.

Other information resources like field trails, genetic background, favorable alleles or the breeder's experience should be involved in the process. A breeder has to evaluate the data basis to apply the tool in the breeding process. Good experience with the GP approach may presage a more positive assessment to predict $\hat{\sigma}_g^2$. A continued validation with relevant data sets are desirable to test the influence of selection criteria like U in comparison to other criteria on genetic gain and genetic diversity in short and longer term. But such validation experiments are most likely expensive. Besides basic features like a representative training set, they require large unselected family sizes, many crosses and precise field trails. Selection indices, different GP models or uncertainty due to unknown bias of the predictors leave room for investigation. Open source, fast and easy to apply R-packages like SelectionTools are a good foundation for new projects. They provide an initial picture of the cross options and can guide a breeder's decision. In comparison to historic approaches like distance measures, new approaches using simulations and closed formulas yielded mainly positive results.

# Chapter 5

# Summary

In plant breeding programs for line varieties and hybrid components, superior lines are selected from a breeding pool as parental lines for the next breeding cycle. However, not all possible crosses between the parental lines can be evaluated in the field due to limited resources. It is therefore more efficient to preselect the most promising parental combinations based on genotypic values. Possible progenies of a cross can be characterized by distribution with parameters mean $\mu$ and the standard deviation $\sigma$. For a cross with large $\mu$ and $\sigma$ it is more likely to find superior progeny. Recombination of elite breeding material often results in crosses with similar $\mu$, therefore $\sigma$ can help to discriminate among the crosses.

Predicting the genetic segregation variance $\sigma_g^2$ of crosses based on genomic data is currently of great interest in the animal and plant community. This is underlined by a range of recent publications in this field. The knowledge of expected $\hat{\mu}$ and $\hat{\sigma}_g^2$ is helpful to use selection criteria like the concept of usefulness or superior progeny value. Both selection strategies involve the distribution parameters $\mu$ and $\sigma_g$ to assess the cross value. Parameter $\mu$ can be predicted with simple methods, while predictors associated with $\sigma_g^2$ were hard to find and made the selection strategies difficult to employ.

Distance measures based on phenotypic, genotypic or pedigree data were insensitively studied but were not robustly associated with $\sigma_g^2$. The combination of phenotypic and genotypic data, lay the foundation for the recently

published methods. At present, simulation approaches and analytical formulas were published to estimate $\hat{\mu}$ and $\hat{\sigma}_g^2$ based on marker effects that are predicted with genome-wide prediction models. Therefore, new approaches are an extension of genomic prediction that can be obtained with less effort since the method is gaining ground as a tool in breeding programs where genotypic and phenotypic data is routinely available.

The first study presents a resource-efficient tool for breeders to select parental lines within a line or hybrid breeding program to distinguish between the most promising crosses that could be made. The estimation is based on marker effects that are predicted with genome-wide prediction models and accounts for the expected gametic disequilibrium between two loci. The derived formulas can be used for typical mating systems like single seed descent and doubled haploid lines, and also consider several generations of intermating before inbred line derivation. A published maize data set was tested and compared with the simulation approach PopVar. The analytical results for means and variances are highly correlated to simulation results. In times of big data management the formulas have a promising speed advantage. At that time, the prediction of $\hat{\mu}$ and $\hat{\sigma}_g^2$ and application of usefulness and superior progeny value has been tested with simulated data and mapping populations. However, 'breeders' data sets represent a major field of application for cross prediction.

In the second study, the practical applicability of an analytical approach for cross prediction based on genome-wide marker effects in a real-life barley data set from an ongoing resistance breeding project. The presented approach is fast and convenient to use, and sufficiently accurate to identify the 50 % best crosses from the field trial.

The new methods are promising for increasing response to selection of line and hybrid breeding programs by extending genomic prediction approaches. The application can support the selection of crossing partners, optimizing or reducing resource use for phenotyping and maintaining genetic diversity in breeding programs.

# Kapitel 6

# Zusammenfassung

In Pflanzenzuchtprogrammen für Liniensorten und Hybridkomponenten werden aus einem Zuchtpool überlegende Elternlinien für den nächsten Zuchtzyklus ausgewählt. Aufgrund begrenzter Ressourcen können nicht alle möglichen Kreuzungen im Feld evaluiert werden. Eine Vorauswahl der vielversprechendsten Kreuzungskombinationen basierend auf genotypischen Werten kann eine effizientere Nutzung der vorhandenen Ressourcen ermöglichen. Nachkommen einer Kreuzung werden durch den Mittelwert $\mu$ und Standardabweichung $\sigma$ charakterisiert. Eine Kreuzung mit großem $\mu$ und $\sigma^2$ bietet eine bessere Grundlage für die Selektion von vielversprechenden Nachkommen. Eine Rekombination von elitärem Zuchtmaterial führt oft zu Kreuzungen mit ähnlichem $\mu$, daher kann der Parameter $\sigma^2$ für die Bewertung der Kreuzung entscheidend sein.

Die Vorhersage genetischer Aufspaltungsvarianz $\sigma_g^2$ einer Kreuzung basierend auf genomischen Daten ist derzeit von großem Interesse in der Tier- und Pflanzenzüchtung. In jüngster Zeit wurden in diesem Gebiet einige Studien veröffentlicht. Selektionsstrategien wie beispielsweise das Konzept der 'usefulness' oder dem 'superior progeny value' verwenden die Parameter $\mu$ und $\sigma$, um ein Ranking der Kreuzungen zu ermöglichen. Der Parameter $\mu$ kann mit einfachen Methoden vorhergesagt werden, während $\sigma^2$ schwer zu schätzen ist. Daher war die Anwendung der Selektionsstrategien lange Zeit nicht in vollem Umfang nutzbar.

Distanzmaße, die auf phänotypischen, genotypischen oder Stammbaumdaten basieren, wurden intensiv untersucht, sind aber nicht robust mit $\sigma_g^2$ assoziiert. Die Kombination phänotypischer und genotypischer Daten legen den Grundstein für die kürzlich veröffentlichten Methoden. Derzeit werden Simulationen und analytische Formeln basierend auf Markereffekten aus genomweiten Vorhersagemodellen verwendet, um geeignete Schätzer für $\hat{\mu}$ und $\hat{\sigma}_g^2$ zu finden. Die neuen Methoden sind eine Erweiterung der genomischen Vorhersage, die als wichtiges Instrument in Zuchtprogrammen Verwendung findet. Die neuen Methoden können daher mit wenig Aufwand integriert werden, weil genotypische und phänotypische Daten oft routinemäßig verfügbar sind.

In der ersten Studie wird ein ressourceneffizientes Werkzeug zur Auswahl der Elternlinien innerhalb eines Linien- oder Hybrid-Zuchtprogramms zur Unterscheidung der vielversprechendsten möglichen Kreuzungen vorgestellt. Die Schätzung basiert auf Markereffekten aus genomweiten Vorhersagemodellen. Die abgeleiteten Formeln schätzen die Streuungs- und Lageparameter basierend auf dem erwarteten Kopplungsphasen-Ungleichgewicht zwischen zwei Loci. Diese können für typische Zuchtschemen für Linienentwicklung wie Ein-Korn-Ramsche oder Doppelthaploide verwendet werden und berücksichtigen auch mehrere vorangegangene Zwischenkreuzungen bevor die Inzuchtlinien generiert werden. Die abgeleiteten Formeln wurden an einem veröffentlichten Mais Datensatz getestet und mit dem Simulationsansatz 'PopVar' verglichen. Die analytischen Ergebnisse sind stark mit den Simulationsergebnissen korreliert. In Zeiten großer Datenmengen haben die Formeln einen vielversprechenden Geschwindigkeitsvorteil. Zu diesem Zeitpunkt war die Vorhersage von $\hat{\mu}$ and $\hat{\sigma}_g^2$ der 'usefulness' nur mit simulierten Daten und Kartierungs-Populationen getestet. Das Hauptanwendungsgebiet der Kreuzvorhersage sind jedoch typische Datensätze aus Zuchtprogrammen.

In der zweiten Studie wurde die praktische Anwendbarkeit eines analytischen Ansatzes zur Kreuzvorhersage basierend auf genomweiten Markereffekten in einem aktuellen Gerste Datensatz aus einem laufenden Resistenzzuchtprojekt getestet. Der vorgestellte Ansatz ist benutzerfreundlich, schnell

und ausreichend genau, um die Besten 50 % der neuen Linien aus dem Feldversuch zu identifizieren.

Die neuen Methoden sind vielversprechend, um den Selektionserfolg in Zuchtprogrammen zu erhöhen, indem geeignete Kreuzungspartner gewählt werden. Hierdurch kann der Ressourcenverbrauch für Phänotypisierung optimiert oder reduziert, und gleichzeitig die genetische Vielfalt überwacht werden.

# Chapter 7

# References

Adeyemo, E, & Bernardo, R. 2019. Predicting genetic variance from genomewide marker effects estimated from a diverse panel of maize inbreds. *Crop Sci*, **59**, 583–590.

Akdemir, D, & Sánchez, JI. 2016. Efficient breeding by genomic mating. *Front Genet*, **7**, 210.

Allier, A, Moreau, L, Charcosset, A, Teyssèdre, S, & Lehermeier, C. 2019. Usefulness criterion and post-selection parental contributions in multi-parental crosses: Application to polygenic trait introgression. *G3 (Bethesda)*, **9**, 1469–1479.

Becker, H. 2011. *Pflanzenzüchtung*. Ulmer, Stuttgart.

Beckett, TJ, Rocheford, TR, & Mohammadi, M. 2019. Reimagining maize inbred potential: Identifying breeding crosses using genetic variance of simulated progeny. *Crop Sci*, **59**, 1457–1468.

Bernardo, R. 2010. *Breeding for quantitative traits in plants. 2nd ed.* Stemma Press, Woodbury, MN.

Bernardo, R. 2014. Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. *Crop Sci*, **55**, 2586–2595.

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

Bhatt, G. 1970. Multivariate analysis approach to selection of parents for hybridization aiming at yield improvement in self-pollinated crops. *Aust J Agric Res*, **21**, 1–7.

Bhatt, G. 1973. Comparison of various methods of selecting parents for hybridization in common bread wheat (*Triticum aestivum* L.). *Aust J Agric Res*, **24**, 457–464.

Bohn, M, Utz, HF, & Melchinger, AE. 1999. Genetic similarities among winter wheat cultivars determined on the basis of RFLPs, AFLPs, and SSRs and their use for predicting progeny variance. *Crop Sci*, **39**, 228–237.

Bonk, S, Reichelt, M, Teuscher, F, Segelke, D, & Reinsch, N. 2016. Mendelian sampling covariability of marker effects and genetic values. *Genet Sel Evol*, **48**, 36.

Brachi, B, Faure, N, Horton, M, Flahauw, E, Vazquez, A, Nordborg, M, Bergelson, J, Cuguen, J, & Roux, F. 2010. Linkage and association mapping of arabidopsis thaliana flowering time in nature. *PLoS Genet*, **6**, e1000940.

Burkhamer, RL, Lanning, SP, Martens, RJ, Martin, JM, & Talbert, LE. 1998. Predicting progeny variance from parental divergence in hard red spring wheat. *Crop Sci*, **38**, 243–248.

Busch, RH, Janke, JC, & Frohberg, RC. 1974. Evaluation of crosses among high and low yielding parents of spring wheat (*Triticum aestivum* L.) and bulk prediction of line performance. *Crop Sci*, **14**, 47–50.

Cole, JB, & VanRaden, PM. 2011. Use of haplotypes to estimate Mendelian sampling effects and selection limits. *J Anim Breed Genet*, **128**, 446–455.

Cowen, NM, & Frey, KJ. 1985. Relationship between genealogical distance and breeding behaviour in oats. *Euphytica*, **36**, 413–424.

Cowen, NM, & Frey, KJ. 1987. Relationships between three measures of genetic distance and breeding behaviour in oats ((*Avena sativa* L.). *Genome*, **29**, 97–106.

Daetwyler, HD, Pong-Wong, R, Villanueva, B, & Woolliams, JA. 2010. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, **185**, 1021–1031.

Daetwyler, HD, Hayden, MJ, Spangenberg, GC, & Hayes, BJ. 2015. Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics*, **200**, 1341–1348.

Endelman, JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, **4**, 250–255.

Falconer, DS, & Mackay, TFC. 1996. *Introduction to quantitative genetics*. Longman, England.

Frisch, M, & Melchinger, AE. 2007. Variance of the parental genome contribution to inbred lines derived from biparental crosses. *Genetics*, **176**, 477–488.

Gumber, RK, Schill, B, Link, W, von Kittlitz, E, & Melchinger, AE. 1999. Mean, genetic variance, and usefulness of selfing progenies from intra- and inter-pool crosses in faba beans (*Vicia faba* L.) and their prediction from parental parameters. *Theor Appl Genet*, **98**, 569–580.

Han, Y, Cameron, JN, Wang, L, & Beavis, WD. 2017. The predicted cross value for genetic introgression of multiple alleles. *Genetics*, **205**, 1409–1423.

Hayes, BJ, Pryce, J, Chamberlain, AJ, Bowman, PJ, & Goddard, ME. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in holstein cattle as contrasting model traits. *PLoS Genet*, **6**, e1001139.

Helms, T, Orf, J, Vallad, G, & McClean, P. 1997. Genetic variance, coefficient of parentage, and genetic distance of six soybean populations. *Theor Appl Genet*, **94**, 20–26.

Heslot, N, Akdemir, D, Sorrells, ME, & Jannink, J-L. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor Appl Genet*, **127**, 463–480.

Hill, WG, Goddard, ME, & Visscher, PM. 2008. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*, **4**, e1000008.

Hung, H-Y, Browne, C, Guill, K, Coles, N, Eller, M, Garcia, A, Lepak, N, Melia-Hancock, S, Oropeza-Rosas, M, Salvo, S, Upadyayula, N, Buckler, ES, Flint-Garcia, S, McMullen, MD, Rocheford, TR, & Holland, JB. 2012. The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity*, **108**, 490–499.

Isidro, J, Jannink, J-L, Akdemir, D, Poland, J, Heslot, N, & Sorrells, ME. 2015. Training set optimization under population structure in genomic selection. *Theor Appl Genet*, **128**, 145–158.

Iwata, H, Hayashi, T, Terakami, S, Takada, N, Saito, T, & Yamamoto, T. 2013. Genomic prediction of trait segregation in a progeny population: a case study of Japanese pear *(Pyrus pyrifolia)*. *BMC Genet*, **14**, 81.

Jiang, Y, & Reif, JC. 2015. Modeling epistasis in genomic selection. *Genetics*, **201**, 759–768.

Kempthorne, O. 1969. *An introduction to genetic statistics*. Iowa State University Press, Ames (Iowa).

Kisha, TJ, Sneller, CH, & Diers, BW. 1997. Relationship between genetic distance among parents and genetic variance in populations of soybean. *Crop Sci*, **37**, 1317–1325.

Kotzamanidis, ST, Lithourgidis, AS, Mavromatis, AG, Chasioti, DI, & Roupakias, DG. 2008. Prediction criteria of promising $F_3$ populations in durum wheat : A comparative study. *Field Crops Res*, **107**, 257–264.

Kuczyńska, A, Surma, M., Kaczmarek, Z., & Adamski, T. 2007. Relationship between phenotypic and genetic diversity of parental genotypes and the frequency of transgression effects in barley *(Hordeum vulgare L .)*. *Plant Breed*, **126**, 361–368.

Lado, B, Battenfield, S, Guzmán, C, Quincke, M, Singh, RP, Dreisigacker, S, Peña, RJ, Fritz, A, Silva, P, Poland, JA, & Gutiérrez, L. 2017. Strategies for selecting crosses using genomic prediction in two wheat breeding programs. *Plant Genome*, **10**, 1–12.

Lehermeier, C, Teyssèdre, S, & Schön, C-C. 2017. Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. *Genetics*, **207**, 1651–1661.

Lian, L, Jacobson, A, Zhong, S, & Bernardo, R. 2015. Prediction of genetic variance in biparental maize populations: Genomewide marker effects versus mean genetic variance in prior populations. *Crop Sci*, **55**, 1181–1188.

Lopez-Cruz, M, Crossa, J, Bonnett, D, Dreisigacker, S, Poland, J, Jannink, J-L, Singh, RP, Autrique, E, & de los Campos, G. 2015. Increased prediction accuracy in wheat breeding trials using a marker environment interaction genomic selection model. *G3 (Bethesda)*, **5**, 569–582.

Lupton, FGH. 1961. Studies in the breeding of self-pollinating cereals. 3. Further studies in cross prediction. *Euphytica*, **10**, 209–224.

Lynch, M, & Walsh, B. 1998. *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, MA.

Maluszynski, M, Kasha, KJ, Forster, BP, & Szarejko, I. 2003. *Doubled haploid production in crop plants. A manual*. Kluwer Academic Publishers, Dordrecht. Pages 309–335.

Manjarrez-Sandoval, P, Carter, TE, Webb, DM, & Burton, JW. 1997. RFLP genetic similarity estimates and coefficient of parentage as genetic variance predictors for soybean yield. *Crop Sci*, **37**, 698–703.

McMullen, MD, Kresovich, S, Villeda, HS, Bradbury, P, Li, H, Sun, Q, Flint-Garcia, S, Thornsberry, J, Acharya, C, Bottoms, C, Brown, P, Browne, C, Eller, M, Guill, K, Harjes, C, Kroon, D, Lepak, N, Mitchell, SE, Peterson, B, Pressoir, G, Romero, S, Rosas, MO, Salvo, S, Yates, H, Hanson, M, Jones, E, Smith, S, Glaubitz, JC, Goodman, M, Ware, D, Holland, JB, & Buckler, ES. 2009. Genetic properties of the maize nested association mapping population. *Science*, **325**, 737–740.

Melchinger, AE, Gumber, RK, Leipert, RB, Vuylsteke, M, & Kuiper, M. 1998. Prediction of testcross means and variances among $F_3$ progenies of $F_1$ crosses from testcross means and genetic distances of their parents in maize. *Theor Appl Genet*, **96**, 503–512.

Meuwissen, THE, Hayes, BJ, & Goddard, ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

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

Moser, H, & Lee, M. 1994. RFLP variation and genealogical distance, multivariate distance, heterosis, and genetic variance in oats. *Theor Appl Genet*, **87**, 947–956.

Nei, M. 1972. Genetic distance between populations. *Am Nat*, **106**, 283–292.

Neyhart, JL, & Smith, KP. 2019. Validating genomewide predictions of genetic variance in a contemporary breeding program. *Crop Sci*, **59**, 1062–1072.

Neyhart, JL, Lorenz, AJ, & Smith, KP. 2019. Multi-trait improvement by predicting genetic correlations in breeding crosses. *G3 (Bethesda)*, **9**(10), 3153–3165.

Osthushenrich, T, Frisch, M, & Herzog, E. 2017. Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. *PLoS ONE*, **12**, e0188839.

Osthushenrich, T, Frisch, M, Zenke-Philippi, C, Jaiser, H, Spiller, M, Cselényi, L, Krumnacker, Kerstin, Boxberger, Susanna, Kopahnke, D, Habekuß, A, Ordon, F, & Herzog, E. 2018. Prediction of means and variances of crosses with genome-wide marker effects in barley. *Front Plant Sci*, **9**, 1899.

Riedelsheimer, C, Technow, F, & Melchinger, AE. 2012. Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics*, **13**, 452.

Riedelsheimer, C, Endelman, JB, Stange, M, Sorrells, ME, Jannink, J-L, & Melchinger, AE. 2013. Genomic predictability of interconnected biparental maize populations. *Genetics*, **194**, 493–503.

Sallam, AH, Endelman, JB, Jannink, J-L, & Smith, KP. 2015. Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *Plant Genome*, **8**, 1–15.

Schnell, FW, & Utz, HF. 1975. $F_1$-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern. *Pages 243–248 of: Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter*. Bundesversuchsanstalt für alpenländische Landwirtschaft Gumpenstein.

Segelke, D, Reinhardt, F, Liu, Z, & Thaller, G. 2014. Prediction of expected genetic variation within groups of offspring for innovative mating schemes. *Genet Sel Evol*, **46**, 42.

Souza, E, & Sorrells, ME. 1991a. Prediction of progeny variation in oat from parental genetic relationships. *Theor Appl Genet*, **82**, 233–241.

Souza, E, & Sorrells, ME. 1991b. Relationships among 70 North American oat germplasms: I. Cluster analysis using quantitative characters. *Crop Sci*, **31**, 599–605.

Souza, E, & Sorrells, ME. 1991c. Relationships among 70 North American oat germplasms: II. Cluster analysis using qualitative characters. *Crop Sci*, **31**, 605–612.

Tiede, T, Mohammadi, M, & Smith, KP. 2015a. PopVar: genomic breeding tools: genetic variance prediction and cross-validation. *R package version*, **1.2**.

Tiede, T, Kumar, L, Mohammadi, M, & Smith, KP. 2015b. Predicting genetic variance in bi-parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. *Mol Breeding*, **35**, 199.

Utz, HF, Bohn, M, & Melchinger, AE. 2001. Predicting progeny means and variances of winter wheat crosses from phenotypic values of their parents. *Crop Sci*, **41**, 1470–1478.

van Berloo, R, & Stam, P. 1998. Marker-assisted selection in autogamous RIL populations : a simulation study. *Theor Appl Genet*, **96**, 147–154.

Wimmer, V, Lehermeier, C, Albrecht, T, Auinger, HJ, Wang, Y, & Schön, CC. 2013. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, **195**, 573–587.

Wittenburg, D, Teuscher, F, Klosa, J, & Reinsch, N. 2016. Covariance between genotypic effects and its use for genomic inference in half-sib families. *G3 (Bethesda)*, **6**, 2761–2772.

Wright, S. 1922. Coefficients of inbreeding and relationship. *Am Nat*, **56**, 330–338.

Yao, J, Zhao, D, Chen, X, Zhang, Y, & Wang, J. 2018. Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). *Crop J*, **6**, 353–365.

Yu, J, & Bernardo, R. 2004. Changes in genetic variance during advanced cycle breeding in maize. *Crop Sci*, **44**, 405–410.

Zhong, S, & Jannink, J-L. 2007. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics*, **177**, 567–576.

Zhong, S, Dekkers, JCM, Fernando, RL, & Jannink, J-L. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics*, **182**, 355–364.

# Acknowledgments

I am very grateful to my academic supervisor Prof. Dr. Matthias Frisch for the opportunity to do my Ph.D. project in the Institute of Planzenbau and Pflanzenzüchtung II, his advice, excellent suggestions and support during this thesis work.

Many thanks to my colleague and co-author Eva Herzog for her patience, guidance during my work, valuable help, great conversations and critical reading of the manuscript.

A special thank to to Prof. Dr. Rod Snowdon for being my second supervisor, serving on my graduate committee and dedicated advice.

Many thanks to all members of sugar beet breeding group at KWS for giving me insight in their breeding practice. They always provide guidance, fruitful discussions, expertise and their valuable data to my Ph.D. project. Especially I would like to acknowledge Dietrich Borchardt for his mentorship.

Sincere thanks for sharing information on ongoing research and for good collaboration in our joint project to all members of barley breeding program.

Many thanks to my office mate Daniel Krenzer for great conversations, discussions, comments, helpful advice, cooperative work and awesome journeys.

Dr. Birgit Samans for sharing her experience and mentorship, great conversations, discussions, comments, helpful advice and great times at the PAG.

I would like to thank all my colleagues, family and friends for their constant encouragement and support.

# Eidesstattliche Erklärung

Ich erkläre:

Ich habe die vorgelegte Dissertation selbständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe.

Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.

Tanja Osthushenrich

Gießen, 26. November 2019