Aus dem Institut für
Pflanzenzüchtung, Saatgutforschung und Populationsgenetik
der Universität Hohenheim
Fachgebiet Angewandte Genetik und Pflanzenzüchtung
Prof. Dr. A.E. Melchinger

Comparison of 'omics' technologies for hybrid prediction

Matthias Westhues

Dissertation

zur Erlangung des Grades eines Doktors

der Agrarwissenschaften

vorgelegt
der Fakultät Agrarwissenschaften

Stuttgart-Hohenheim
2019

Die vorliegende Arbeit wurde am 27.09.2019 von der Fakultät Agrarwissenschaften der Universität Hohenheim als "Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften (Dr. sc. agr)" angenommen.

Tag der mündlichen Prüfung: 08.11.2019

1. Prodekan:	Prof. Dr. T. Streck
Berichterstatter, 1. Prüfer:	Prof. Dr. A.E. Melchinger
Mitberichterstatter, 2. Prüfer:	Prof. Dr. M. Frisch
3. Prüfer:	Prof. Dr. J. Bennewitz

Contents

Contents iii

1 General Introduction 2

2 Omics-based hybrid prediction in maize 10

3 Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize 12

4 Efficient Genetic Value Prediction Using Incomplete Omics Data 14

5 General Discussion 16

6 Summary 29

7 Zusammenfassung 31

8 Acknowledgments 51

9 Curriculum Vitae 53

10 Erklärung 53

Abbreviations

BLUP	Best linear unbiased prediction
GBLUP	Genomic best linear unbiased prediction
GWAS	Genome-wide Association Study
LD	Linkage disequilibrium
MAF	Minor allele frequency
MAS	Marker-assisted selection
OLS	ordinary least squares
QTL	Quantitative trait locus
RKHS	Reproducing kernel hilbert spaces
RNA	ribonucleic acid
TRN	training set
TST	test/validation set

1. General Introduction

Hybrid breeding, which capitalizes on exploiting heterosis in crosses between genetically divergent parents, is an important driver of sustainable agricultural intensification needed for meeting increased demands for food and fiber (Duvick, 2005). With the emergence of the doubled-haploid (DH) technology (Wedzony et al., 2009), hybrid breeding has received further attention. The DH technology allows the production of diploid plants, which have two identical copies of the same chromosome, in a single generation. Assuming two heterotic groups, which are genetically distinct clusters of maternal and paternal germplasm (Melchinger and Gumber, 1998), and further assuming that a medium-sized breeding program can generate about 1,000 DH lines per heterotic group and year, a breeder would hypothetically have to select from $n^2 = 10^6$ hybrid candidates in each generation. Typically, 90% of DH lines are being produced anew in each season implying that $0.9^2 * 10^6 = 810.000$ putative hybrids would not share a single parent of hybrid progeny with phenotypic records, thereby exacerbating the selection even more. The great challenge now is to find ways to reduce the effort necessary to select the most promising of all possible candidates, which cannot be assessed exhaustively in multi-environment field trials.

Accelerating Hybrid Prediction

Early attempts to limit the number of genotypes, that need to be assessed for estimating the performance of putative hybrids, was to examine the *per se* performance of their parent lines. Due to dominance effects (Schrag et al., 2006; Smith, 1986) this approach failed, however, giving way to the evaluation of *general combining ability* (GCA), which is based on crosses between putative parent lines with few tester lines from the opposite heterotic group (Schrag et al., 2006; Sprague and Tatum, 1942). The latter approach is still practiced in many modern breeding programs, but it is time-consuming (Bernardo, 2010), thereby reducing potential selection gain (Falconer and Mackay, 1996). A promising way to address the huge number of candidates is to look into auxiliary information, data collected on the parents instead of the hybrid candidates themselves, thereby reducing the number of genotypes from n^2 to just $2n$. Since its inception, the goal of hybrid breeding was to pick the best hybrid candidate itself (Shull, 1908) but, due to the previous lack of suitable predictor data as well as absent powerful computational resources and software, this idea has been revisited only recently (Technow et al., 2014).

Expected Relationship

Pedigree data, which can be collected at almost no cost, emerged as the first predictor to lift some of the burden of intensive phenotype testing. Pedigree data capture the expected relatedness between individuals and have been proven to provide valuable information in hybrid prediction models (Bernardo, 1994; Crossa et al., 2014). However, pedigree data suffer from some shortcomings: The choice of a founder generation is rather arbitrary and typically depends on the time at which a breeding program started collecting pedigree records. If the pedigree is extremely deep, another problem arises in that kinship coefficients converge to one, thereby requiring truncation of the pedigree at some point (Speed and Balding, 2015). Moreover, Mendelian sampling, representing the randomness in which genes are passed down from parents to their progeny, is not accounted for by pedigree information. Finally, pedigree-information may give inaccurate estimates of the gene substitution effect because its prediction models falsely assume that residuals are identically and independenty distributed (Duenk et al., 2017).

Dawn of Genomics

Between 1974 and 2005 several marker systems, such as restriction fragment length polymorphisms (RFLP), simple sequence repeats (SSR), amplified fragment length polymorphisms (AFLP), single nucleotide polymorphisms (SNP) and diversity arrays technology markers (DArT), were developed (see Bernardo (2008) for a review). Marker-assisted selection (MAS) was suggested for integrating these molecular markers with phenotype data to optimize selection indices and drive the improvement of quantitative characters (Lande and Thompson, 1990). While MAS successfully facilitated the manipulation of quantitative trait loci (QTL) with large effects (Zhong et al., 2006) it has failed to improve traits influenced by many QTL, each with small effects on the trait (Bernardo, 2008; Dekkers and Hospital, 2002). Moreover, MAS suffers from biased effect sizes of QTL estimated in different genetic backgrounds (Melchinger et al., 1998) and may require large mapping populations when effect sizes are small and/or allele frequencies depart from 0.5 (Mackay et al., 2009).

Realized Relationship and Algorithms

While the path outlined by Lande and Thompson (1990) is deemed a rather inefficient use of the wealth of molecular data (Meuwissen et al., 2001), Fernando and Grossman (1989) proposed to use all genomic data simultaneously by appying BLUP to a linear mixed model, thereby jointly incorporating effects of all genomic markers. Especially in plant breeding, the number of features of a predictor p typically exceeds the number of genotypes n, thereby rendering ordinary least squares (OLS) infeasible for such prediction scenarios because no degrees of freedom are left for estimating feature effects (de los Campos et al., 2010). Instead, researchers adopted methods that incorporate effect shrinkage, and thus, circumvent this shortcoming of OLS. Among the earliest, and still widely used (Mrode, 2014), methods was best linear unbiased prediction (BLUP) (Henderson, 1949) and its equivalent ridge regression (RR) (Fernando and Grossman, 1989; Meuwissen et al., 2001). In short, BLUP incorporates information on non-random environmental effects as well as appropriately weighted information on related individuals to adjust the phenotypic observations (Mrode, 2014). BLUP, respectively RR, shrinks all effects equally and is well-suited for the infinitesimal model (Fisher, 1918) where the number of features is large and where no single feature explains a large proportion of phenotypic variance (Mackay et al., 2009). Popular alternatives to BLUP are Bayesian estimation methods, which assign varying prior distributions to effects and are therefore particularly suited for situations where the number of features is small to moderate and where some effects have appreciably higher impacts on phenotypic variation than others (de los Campos et al., 2010; Daetwyler et al., 2010; Zhao et al., 2015). These models mitigate the problem of overshrinking large effects that may occur with RR (Heffner et al., 2009). Empirical studies, however, have usually shown that the difference between algorithms like RR-BLUP and BayesB, which shrinks many effects to zero and assigns stronger priors to non-zero effects than RR-BLUP or BayesA, is usually small (Hayes et al., 2009; Technow et al., 2014; VanRaden et al., 2009). A plethora of further algorithms for predicting breeding values, including RR-BLUP, Bayesian LASSO, elastic net, Bayes Cπ, empirical Bayes, reproducing kernel Hilbert space, weighted Bayesian shrinkage regression, support vector machines, random forests and neural networks, were explored on eight plant breeding datasets but no clear recommendation for a particular algorithm could be made (Heslot et al., 2012). Generally, the optimal algorithm-choice depends on the phenotype as well as the predictors involved (Xu et al., 2017) but most recent studies typically use either GBLUP, when epistasis is small or not of interest, or reproducing kernel hilbert space when epistasis should be accounted for (e.g. e Souza et al. (2017); Xu et al. (2017)).

Limitations of genomic information

Genomic information has propelled the field of trait predictions in the past two decades and the technology has matured so much that sampling genomic information can nowadays be done at a fraction of the initial costs. With the great success in predicting complex traits using genotyping arrays, such as the

Illumina50K chip for maize, it was initially expected that the next technologial leap, whole-genome sequencing (WGS) would help to precisely pinpoint quantitative trait loci influencing the trait of interest. Genotyping arrays are known to suffer from ascertainment bias, meaning that SNPs with high minor allele frequencies (MAF) are overrepresented (Mathew et al., 2018; Pérez-Enciso et al., 2015; Wray et al., 2013; Yang et al., 2010), putting sufficient linkage disequilibrium (LD) between SNPs and QTL in jeopardy (Hayes et al., 2009; Yang et al., 2017). Whole-genome sequencing (WGS) and genotyping-by-sequencing data are expected to remedy such shortcomings of SNP-arrays but suffer from the so-called 'curse of dimensionality', which implies that an increasing number of features, relative to the number of samples, increases overfitting (Libbrecht and Noble, 2015). In a simulation study, Pérez-Enciso et al. (2015) observed that the expected gain in prediction accuracy from WGS-data compared to array-data is merely in the range of 4% to 8%. This result clearly demonstrates the tradeoff between the richer information content of sequence data and added noise in the predictions compared to genotyping arrays. In human genetics, the $p > n$ problem is currently being addressed, particularly through the formation of consortia accumulating data for studies with related research questions. Still, a study investigating genetic influences on human height using 13,558,738 markers with a sample size of 102,221 individuals could only realize an explained variance of about one-third of the trait heritability in prediction (Kim et al., 2017).

Integrating Physiological Epistasis

Epistasis, describing deviations from Mendelian genetics where the phenotype expression of a genotype at a locus is altered by (an)other loci in the genome (Sackton and Hartl, 2016), adds to the aforementioned issues with genomic information. A caveat of genomic data is its inability to account for physiological epistasis (Dalchau et al., 2011; Zhu et al., 2012) while statistical epistasis, which is defined as genetic variation at the population level (Sackton and Hartl, 2016), is typically of negligible magnitude (Guo et al., 2016; Hill et al., 2008;

Mackay, 2014; Vazquez et al., 2016). Even recent methodological developments, addressing the 'curse of dimensionality' (Jarquín et al., 2014; Martini et al., 2016; Vitezica et al., 2017), could not exploit statistical epistasis unless training and prediction set shared the same or closely related parents (Jiang and Reif, 2015).

Endophenotypes, such as messenger RNA, proteins and metabolites, constitute biological layers governing the phenotype manifestation of a genotype (Mackay et al., 2009). Given their intermediary position in the genotype-phenotype cascade, they offer to capture physiological epistasis, *i.e.*, interactions among and between any biological layer preceding the phenotype, including epistasis at the genome level (Sackton and Hartl, 2016). Contrary to the pedigree as well as the genome, endophenotypes are not static but increasingly in flux with growing proximity to the phenotype as exemplified by metabolites of the Calvin-Benson cycle (Sweetlove et al., 2016). Combining endophenotypes, as well as pedigree and genomic information, was suggested for compensating missing or unreliable information in any single data source (Ritchie et al., 2015) and is a focal point of this thesis.

Imputing Incomplete Predictors

Recent studies in maize inbred lines (Guo et al., 2016), rice hybrids (Dan et al., 2016; Xu et al., 2016) and maize hybrids (Zenke-Philippi et al., 2017) have shown that transcriptomic and metabolomic data, sampled at an early stage of plant development, could improve upon pedigree- and even genome-based predictions of some important and complex traits. Despite these promising results, novel "omics" predictors have not found widespread application, yet. One of the reasons for reservations among breeders regarding the use of "omics" predictors might be additional costs and logistical efforts for sampling these data compared to pedigree or SNP data. Borrowing a statistical framework from animal breeding, which allows for the combination of two predictors sampled on partially overlapping sets of genotypes, might incentivize breeders to spend resources on covering merely a subset of genotypes with "omics" predictors while imputing the latter for the remaining set of genotypes.

The introduction of MAS, and later genomic selection, in animal breeding suddenly confronted researchers with a situation where a large number of "old" animals had only pedigree records whereas a much smaller number of "young" animals had both, pedigree and genomic records. First attempts at combining the two predictors for estimating breeding values of animals used a two step approach where, in a first step, conventional breeding values were estimated using just pedigree data and phenotypic values from related animals (VanRaden et al., 2009). In a second step, marker effects were estimated by regressing conventional breeding values onto the marker genotypes of "young" animals. These effects could then be used to impute breeding values of the "young" animals. A problem with this approach is that the error distribution of the conventional breeding values may be different from that of the imputed breeding values (Aguilar et al., 2010). Garrick et al. (2009) proposed to accomodate this error by deregressing the breeding values and weighting the residuals according to the prediction error variances. Independently from each other, Legarra et al. (2009) and Christensen and Lund (2010) developed a statistical framework in which both, complete and incomplete predictor information could be used simultaneously for infering breeding values. Their method was refined by transforming their BLUP-models into equivalent single step marker-effect-models, which implicitly model the structure of the imputation error and use the information of both predictors without sophisticated weighting of their coefficients (Fernando et al., 2014).

Objectives

The goal of this thesis' research was to compare different 'omics' technologies regarding their utility for hybrid prediction. In particular, the objectives were to

1. compare the performance of 'omics' or pedigree data as single predictors for the prediction of hybrid performance.

2. investigate the benefit of combining features, sampled at the level of parent lines, for predicting genetic values of maize hybrids using multi-environmental phenotypic data on complex agronomic traits.

3. explore the single-step prediction framework as a viable improvement over current methods for imputing 'omics' predictors.

4. transfer the single-step framework for breeding value prediction to the prediction of hybrids derived from pure-breeding inbred lines.

2. Omics-based hybrid prediction in maize

Matthias Westhues[1,*], Tobias A. Schrag[1,*], Claas Heuer[2,3], Georg Thaller[2], H. Friedrich Utz[1], Wolfgang Schipprack[1], Alexander Thiemann[4], Felix Seifert[4], Anita Ehret[2], Armin Schlereth[5], Mark Stitt[5], Zoran Nikoloski[5], Lothar Willmitzer[5], Chris C. Schön[6], Stefan Scholten[4], Albrecht E. Melchinger[1]

[*]These authors contributed equally

[1]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, D-70599 Stuttgart, Germany

[2]Institute of Animal Breeding and Husbandry, Christian-Albrechts-University Kiel, D-24098 Kiel, Germany

[3]Inguran, LLC dba STGenetics, 22575 SH6 South, Navasota, TX 77868

[4]Biocenter Klein Flottbeck, Developmental Biology and Biotechnology, University of Hamburg, D-22609 Hamburg, Germany

[5]Max-Planck Institute of Molecular Plant Physiology, D-14476 Potsdam, Germany

[6]Plant Breeding, Technische Universität München, D-85354 Freising, Germany

Theor. Appl. Genet. (2017) 130:1927-1939

The original publication is available at

https://doi.org/10.1007/s00122-017-2934-0

Abstract

Accurate prediction of traits with complex genetic architecture is crucial for selecting superior candidates in animal and plant breeding and for guiding decisions in personalized medicine. Whole-genome prediction has revolutionized these areas but has inherent limitations in incorporating intricate epistatic interactions. Downstream 'omics' data are expected to integrate interactions within and between different biological strata and provide the opportunity to improve trait prediction. Yet, predicting traits from parents to progeny has not been addressed by a combination of 'omics' data. Here, we evaluate several 'omics' predictors —genomic, transcriptomic and metabolomic data —measured on parent lines at early developmental stages and demonstrate that the integration of transcriptomic with genomic data leads to higher success rates in the correct prediction of untested hybrid combinations in maize. Despite the high predictive ability of genomic data, transcriptomic data alone outperformed them and other predictors for the most complex heterotic trait, dry matter yield. An eQTL analysis revealed that transcriptomic data integrate genomic information from both, adjacent and distant sites relative to the expressed genes. Together, these findings suggest that downstream predictors capture physiological epistasis that is transmitted from parents to their hybrid offspring. We conclude that the use of downstream 'omics' data in prediction can exploit important information beyond structural genomics for leveraging the efficiency of hybrid breeding.

3. Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize

Tobias A. Schrag[1,*] Matthias Westhues[1,*], Wolfgang Schipprack[1], Felix Seifert[2], Alexander Thiemann[2], Stefan Scholten[2], Albrecht E. Melchinger[1]

[*]These authors contributed equally

[1]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, D-70599 Stuttgart, Germany
[2]Biocenter Klein Flottbeck, Developmental Biology and Biotechnology, University of Hamburg, D-22609 Hamburg, Germany

Genetics, Vol. 208, 1373-1385 (April 2018)

The original publication is available at
https://doi.org/10.1534/genetics.117.300374

Abstract

The ability to predict the agronomic performance of single-crosses with high precision is essential for selecting superior candidates for hybrid breeding. With recent technological advances, thousands of new parent lines, and consequently, millions of new hybrid combinations are possible in each breeding cycle, yet only a few hundred can be produced and phenotyped in multi-environment yield trials. Well established prediction approaches, such as best linear unbiased prediction (BLUP) using pedigree data and whole-genome prediction using genomic data are limited in capturing epistasis and interactions occurring within and among downstream biological strata such as transcriptome and metabolome. Because mRNA and small RNA (sRNA) sequences are involved in transcriptional, translational and post-translational processes, we expect them to provide information influencing several biological strata. However, using sRNA data of parent lines to predict hybrid performance has not yet been addressed. Here, we gathered genomic, transcriptomic (mRNA and sRNA) and metabolomic data of parent lines to evaluate the ability of the data to predict the performance of untested hybrids for important agronomic traits in grain maize. We found a considerable interaction for predictive ability between predictor and trait, with mRNA data being a superior predictor for grain yield and genomic data for grain dry matter content, while sRNA performed relatively poorly for both traits. Combining mRNA and genomic data as predictors resulted in high predictive abilities across both traits and combining other predictors improved prediction over that of the individual predictors alone. We conclude that downstream ‘omics’ can complement genomics for hybrid prediction, and thereby, contribute to more efficient selection of hybrid candidates.

4. Efficient Genetic Value Prediction Using Incomplete Omics Data

Matthias Westhues[1], Claas Heuer[2,3], Georg Thaller[2], Rohan Fernando[4], Albrecht E. Melchinger[1]

[1]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, D-70599 Stuttgart, Germany

[2]Institute of Animal Breeding and Husbandry, Christian-Albrechts-University Kiel, D-24098 Kiel, Germany

[3]Inguran, LLC dba STGenetics, 22575 SH6 South, Navasota, TX 77868

[4]Department of Animal Science, Iowa State University, 50011 Ames, Iowa, U.S.A.

Theor. Appl. Genet. (2019), 132(4):1211-1222

The original publication is available at
https://doi.org/10.1007/s00122-018-03273-1

Abstract

Predicting genetic values with high accuracy is pivotal for effective candidate selection in animal and plant breeding. Novel ‘omics’-based predictors have been shown to improve upon established genome-based predictions of important complex traits but require laborious and expensive assays. As a consequence, there are various datasets with full genetic marker coverage of all studied individuals but incomplete coverage with other ‘omics’ data. In animal breeding, single-step prediction was introduced to efficiently combine pedigree information, collected on a large number of animals, with genomic information, collected on a smaller subset of animals, for breeding value estimation without bias. Using two maize datasets of inbred lines and hybrids, we show that the single-step framework facilitates imputing transcriptomic data, boosting forecasts when their predictive ability exceeds that of pedigree or genomic data. Our results suggest that covering only a subset of inbred lines with ‘omics’ predictors and imputing all others using pedigree or genomic data could enable breeders to improve trait predictions while keeping costs under control. Employing ‘omics’ predictors could particularly improve candidate selection in hybrid breeding because the success of forecasts is a strongly convex function of predictive ability.

5. General Discussion

Endophenotypes are assumed to capture a plethora of biological information, which may not be available from routinely used predictors such as pedigree and genomic information (Mackay et al., 2009; Sackton and Hartl, 2016). To improve selection gain in hybrid breeding programs, our objective was to assess the utility of these endophenotypes for predicting genetic values of single crosses and to explore avenues for maximizing cost-efficiency of these alternative predictors.

High predictive abilities in the estimation of genetic values for hybrids are primarily driven by large genetic distances between parents, linkage disequilibrium, relatedness between training and test set and properties of the predictors themselves.

Heterotic patterns and Linkage Disequilibrium

Heterotic patterns refer to two genetically diverse populations, one being used for generating maternal and the other generating paternal lines for the production of hybrid offspring. The separate improvement of each heterotic pool through recurrent selection has two beneficial outcomes: (1) genetic drift and selection for hybrid performance procure complementary gametes in the hybrid offspring and (2) it ensures that the performance of superior hybrids can predominantly be predicted on the basis of general combining ability (GCA) effects (Reif et al., 2007).

All maize hybrids used throughout our studies were based on the Dent x Flint heterotic pattern, which was established by exploiting two independent introductions of material from North America and the Caribbean (Brandenburg et al., 2017). The genetic divergence between the Dent and Flint materials used in our studies could clearly be gleaned from principal component analyses using pedigree, genomic markers, gene expression transcripts and single-stranded RNA, which is consistent with two studies using U.S. maize lines (Gerke et al., 2015; Hall et al., 2016). Moreover, ratios of specific combining ability (SCA) variance to the total genetic variance in our material were less than ten percent for both, silage maize and grain maize traits. This result coincides with observations from previous studies on the genetic architecture of hybrid breeding programs in maize, which reported SCA variance that typically amounted to five to twenty percent of the total genetic variance for various traits and heterotic patterns from European as well as U.S. hybrid maize (Argillier et al., 2000; Bernardo, 1996; Fischer et al., 2008; Geiger et al., 1986; Kadam et al., 2016; Parisseaux and Bernardo, 2004; Schrag et al., 2006; Technow et al., 2014). It should be noted, though, that this low ratio is not exclusively due to the exploitation of heterotic patterns but also affected by the makeup of a typical hybrid breeding scheme. Commonly, the final selection of single crosses in factorial mating designs is preceded by a selection stage where parents from each heterotic group are selected based on their general combining ability with one or more testers from the opposite heterotic group, thereby reducing the magnitude of SCA variance (Giraud et al., 2017).

With the progressing adoption of new technologies for rapidly generating fully homozygous inbred lines, such as single-seed descent (SSD) and the doubled-haploid (DH) technology (Wedzony et al., 2009), vast numbers of hybrids could theoretically be produced. Hybrids can be partitioned into three categories: T2, T1 and T0 hybrids, where the digit refers to the number of parents that have previously been tested in other hybrids. Given that 81% of hybrids will be T0 hybrids, sharing no parental gametes with the training set (TRN), efforts should be directed towards improving predictive ability for this subset of hybrids. The importance of similarities between the TRN and the test set (TST) for high predictive abilities was pointed out in several studies (Albrecht et al., 2014; Crossa et al., 2014; Technow et al., 2014). Technow et al. (2014) observed that predictive ability increases with the number of hybrids in the TRN, but that the benefit of a large number of hybrids in the TRN is reduced if the ratio of SCA variance to the total genetic variance is small, as typically observed in established heterotic patterns. Predictive abilities of T2 hybrids were particularly high in our studies with comparatively small dependence on the predictor type. This can be explained by the fact that T2 hybrids were covered by multiple copies of their parental gametes in the TRN so that predictive ability was primarily driven by the low ratio of SCA variance to total genetic variance as argued by Technow et al. (2012). Remarkably high predictive abilities observed in experiments on the utility of metabolites for predicting T2 hybrids in rice (Dan et al., 2016) corroborate this hypothesis. Conversely, predictive ability for T0 and T1 improved when increasing the number of parental lines in the TRN while keeping the number of hybrids constant; probably due to more precise estimates for GCA effects (Technow et al., 2014). This ranking of T2 over T1 and T0 hybrids, with respect to predictive abilities, was confirmed in independent studies (Kadam et al., 2016; Zenke-Philippi et al., 2017). To match the genetic diversity in the TRN with that of the target population, Bustos-Korts et al. (2016) suggested an algorithm used by gene banks (Odong et al., 2013). This uniform sampling procedure successfully improved predictive ability over that with random sampling of the TRN without introducing population stratification. With a constant number of hybrids that can be evaluated, a higher number of properly sampled parent lines would particularly benefit the prediction of T0 and T1 hybrids compared to T2 hybrids (Technow et al., 2014).

Ideally, studies should be designed such that markers and causative QTL are in high LD, otherwise predictive ability will primarily be based on relationships between individuals occurring in both, TRN and TST (Habier et al., 2013; Schopp et al., 2017; Yang et al., 2017). Unfortunately, relationships between TRN and TST disintegrate faster than LD between markers and QTL, thereby necessitating more resources for recurring estimation of marker effects or breeding values (Heslot et al., 2012). Moreover, kinship-driven selection of candidates increases the risk of inbreeding, which is equivalent to a loss of genetic diversity, thereby threatening sustainable genetic gain (Heslot et al., 2012). Technologies for the interrogation of the entire genome are now emerging, raising expectations of perfectly tagging QTL with markers (Kahvejian et al., 2008). Simulations have shown that gains from whole-genome sequencing (WGS) data over SNP genotyping-arrays are merely modest (Pérez-Enciso et al., 2015), which is due to technical limitations such as sequencing errors but also to the 'curse of dimensionality' (Libbrecht and Noble, 2015), which is particularly problematic in plant breeding programs where sample sizes are rather limited. While new sequencing technologies improve tagging of causal QTL, there may be no algorithms that can successfully cope with this plethora of information given limited sample sizes. While algorithms, that can set the size of effects to zero (*e.g.* BayesB, LASSO, elastic net) can outperform GBLUP for traits with a small number of well-tagged QTL with large effects (Daetwyler et al., 2010), they suffer from identifiability issues for a large number of features (Heslot et al., 2012; Technow et al., 2014). Our own studies have found that predictive ability with genomic information quickly reached a plateau at marker densities around 10,000 SNPs, which is in agreement with previous studies (Riedelsheimer et al., 2012; Technow et al., 2012, 2014) and BayesB could not improve upon GBLUP, which is why we did not present those results in our publications.

Pedigree-based and genomic prediction

Bernardo (1994) pioneered the use of pedigree information and genomic markers for best linear unbiased prediction of genetic values in plant hybrids. Since then, numerous studies on genetic value prediction in hybrid material have been conducted (see Zhao et al. (2015) for a review), most of them using either (genomic) BLUP, BayesB or reproducing kernel Hilbert spaces (RKHS) as prediction algorithms. In our studies, pedigree information was typically outperformed by genomic information in the prediction of genetic values for hybrids, which is consistent with findings from wheat (Crossa et al., 2010). This was particularly the case for T0 hybrids, where only distantly related individuals were shared between TRN and TST, and coincides with observations on Merino sheep as well as simulated data (Clark et al., 2012) and a study using genetially diverse maize and wheat material from CIMMYT (Crossa et al., 2014). Intriguingly, in the latter study, nonlinear algorithms yielded improvements in predictive ability over those obtained with GBLUP when marker density increased (Crossa et al., 2014). This supports a finding by Technow et al. (2014), who used a subset of the material presented in Schrag et al. (2018), where BayesB yielded a slightly higher predictive abiliy in T0 hybrids compared to GBLUP.

Value added by endophenotypes

Neither pedigree nor genomic data capture intricate physiological epistasis arising from complex interactions among and between endophenotypes (Brem et al., 2005; Brown et al., 2014; Guo et al., 2016; Hill et al., 2008). Our studies suggested that the transcriptome, which is the first endophenotype following the genome, integrates interactions between SNP loci, both on the same as well as across chromosomes. Genetic value prediction for hybrids using transcriptomic data yielded improved predictive abilities compared to those achieved with genomic information for yield in silage and grain maize as well as for protein in silage maize (Schrag et al., 2018; Westhues et al., 2017). In our study using maize inbred lines (Guo et al., 2016), ear diameter and kernel width, two yield-related traits, were predicted with higher precision by transcriptomic data compared to genomic data (Westhues et al., 2017). Small RNAs, were included in one of our studies (Schrag et al., 2018) because they are intricately interwoven with gene expression by influencing transcriptional, post-transcriptional and post-translational processes (Franks et al., 2017; Lappalainen et al., 2013; Li et al., 2016; Mortimer et al., 2014), thereby potentially augmenting gene expression data. While they did not outperform any of the other evaluated predictors they yielded a predictive ability for grain yield that nearly matched that of genomic data and were on par with gene expression-based predictive ability for grain dry matter content (Schrag et al., 2018). Metabolites are the final endophenotype prior to the manifestation of the phenotype, and therefore, could potentially integrate all previous main and interaction effects within and between upstream endophenotypes (Patti et al., 2012). For the prediction of grain dry matter yield, root metabolites were superior to genomic information but trumped by gene expression information (Schrag et al., 2018). Conversely, root metabolites were the worst predictor for grain dry matter content in the same material. Competitive predictive abilities could be realized with root metabolites for the traits 'fat' and 'dry matter yield' in silage maize while leaf metabolites performed poorly for all assessed traits (Westhues et al., 2017).

Balance between phenotype-proximity and endophenotype perturbation

As previously noted, metabolites, as the final endophenotype in the genotype- phenotype cascade, should potentially be the superior predictor by representing all upstream effects acting on the phenotype. However, this was clearly not the case throughout our studies; a result that is in agreement with other studies (Guo et al., 2016; Xu et al., 2017) comparing metabolites to different predictor types. One explanation could be the very limited subset of metabolites that were measured in these studies compared to the vast array estimated to be present in plants (Fernie, 2007). Additionally, metabolites belonging to the elementary Calvin-Benson cycle have extremely fast turnover rates (Arrivault et al., 2009), making it nearly impossible to quantify them precisely. Finally, metabolites are very susceptible to biotic (Rudd et al., 2015; Tzin et al., 2015) and abiotic (Asiago et al., 2012; Caldana et al., 2011; Witt et al., 2012) perturbations. Notwithstanding, other endophenotypes are also affected by external factors and such deviations of endophenotype levels from what is expected based on levels observed in upstream biological layers is generally referred to as 'phenotypic buffering' (Civelek and Lusis, 2014). The expression of genes is known to be influenced by small RNAs, which couple with an Argonaute (AGO) protein and bind to a matching messenger-RNA sequence, thereby partially or completely suppressing the matched gene (Mortimer et al., 2014). Involvement of sRNAs has been reported for various processes in plants such as (i) leaf polarity, (ii) leaf serration, (iii) development phase change, (iv) flowering time, (v) root meristem development and (vi) responses to biotic and abiotic stresses (see Li et al. (2016) for a review). Two studies on rape seed (Shen Yifei et al., 2017) and maize Seifert et al. (2018a) found that sRNA expression levels displayed moderate to strong negative correlations with the expression of their pertaining mRNAs. This is in opposition to positive correlations of genomic and gene expression information in maize (Seifert et al., 2018a) and suggests that sRNA expression profiles capture different information (Seifert et al., 2018b). Gene expression patterns in maize hybrids revealed a preponderance of additive expression in the F1 generation (Springer and Stupar, 2007b; Stupar et al., 2008), which was recently confirmed for gene expression in hybrids of rape seed (Shen Yifei et al., 2017) and Arabidopsis, where 95% of expressed genes were at the mid-parent level (Alonso-Peral et al., 2016). Similar observations were made at the metabolite level in maize hybrids (Lisec et al., 2011) and taken together these findings could be explained within the context of metabolic flux theory, which posits that detrimental expression levels of endophenotypes in the parents are counterbalanced in their hybrid progeny (Kacser and Burns, 1981; Springer and Stupar, 2007a). Further support is provided by co-localization of additively expressed genes with heterotic QTL in hybrid maize (Thiemann et al., 2014). Interestingly, genes associated with defense-mechanisms were significantly suppressed in rape seed hybrids compared to their parent lines, possibly adding to plant growth heterosis given that plant immunity and plant growth are negatively associated (Shen Yifei et al., 2017).

Complementing endophenotypes

Combining endophenotypes was shown to improve upon predictive abilities of individual predictors in a study on the prediction of breast cancer risk (Vazquez et al., 2016) and a study predicting genetic values in maize inbred lines (Guo et al., 2016). In concordance with our observations, Guo et al. (2016) showed that predictive abilities, based on either genomic or metabolite information, were dependent upon multiple factors: (i) the trait to be predicted, (ii) the type of tissue from which endophenotypes were sampled, (iii) the age of the tissue at sampling and (iv) the number of features. The combination of genomic with gene expression data improved the stability of predictions across traits (Guo et al., 2016; Schrag et al., 2018; Westhues et al., 2017) possibly due to the compensation of missing or unreliable information in any individual predictor (Ritchie et al., 2015). While the simultaneous inclusion of all endophenotypes in predictions maximized predictive abilities for maize inbred lines (Guo et al., 2016), this was not strictly true for maize hybrids where combinations of just gene expression data with genomic information yielded the best average predictive abilities across traits. To investigate the optimal contribution of individual predictors for maximizing predictive abilities, Schrag et al. (2018) placed weights ranging from 0 to 1 on each of three predictors (pedigree, genomic information, gene expression information) so that the sum of weights was equal to one. A previous simulation study suggested that adjusting the variance of individual markers with known large effects could improve predictive abilities (Bernardo, 2014). Adding markers from high-density genotyping chips, which were found to be significantly associated with complex traits in GWAS, to low-density genotyping chips yielded gains in reliability of production trait prediction ranging from 1.5 to 4 percentage points in Nordic Holsteins (Brøndum et al., 2015). In our study (Schrag et al., 2018), a high weight on gene expression data and almost no weight on pedigree information, maximixed the ability to predict genetic values for grain yield whereas weights for the three predictors were roughly balanced in the prediction of grain dry matter content (GDMC). Ashraf et al. (2016) also observed that the optimal weight placed on pedigree and genomic relationship matrices varied depending on the trait and hypothesized that pedigree information likely received higher weights for traits where a high fraction of variance was generated by QTL with low minor allele frequency, which might apply to GDMC. Finally, modelling SCA effects did not improve predictive abilities for any predictor (Schrag et al., 2018; Westhues et al., 2017) and interactions between predictors were not modeled because the studies by Guo et al. (2016) and Vazquez et al. (2016) had indicated that our sample sizes would have been too small to benefit from including SCA effects.

Properties of valuable predictors

Our studies aimed to explore what constitutes a valuable predictor for the prediction of genetic values in hybrids and identify properties that should be considered when designing experiments. The four major properties were: (i) sampling costs and throughput, (ii) repeatability, (iii) tissue and (iv) cell age. Pedigree information can be collected at practically no cost except for the effort of record-keeping itself. Assuming that pedigree records are correct, it follows that repeatability of pedigree information is perfect. Genotyping has become increasingly affordable for larger plant breeding companies that can exploit economies of scale. Repeatability of genomic information is nearly perfect since the genome is static and technologies are sufficiently mature for producing high quality data at high throughput. Tissue type and cell age are irrelevant factors for genomic information, making it an almost ideal predictors apart from the previously mentioned caveats: i) their inability to capture physiological epistasis, ii) sequencing errors and iii) their large number compared to typical sample sizes in plant breeding. In contrast to pedigree and genomic information, all of the listed factors must be weighted when choosing endophenotypes for the prediction of genetic values. Because endophenotypes are not static layers of biological information, it is necessary to sample them from multiple biological replicates to deal with external perturbations (Asiago et al., 2012; Caldana et al., 2011; Rudd et al., 2015; Tzin et al., 2015; Witt et al., 2012) for improving their repeatability. With the advent of RNA-seq (Martin and Wang, 2011; Wang et al., 2009), throughput and coverage for gene expression have reached promising levels although the technology still requires fine-tuning (Franks et al., 2017). For metabolites, the situation is currently less promising given that the sampled range of metabolites using recent technologies is nowhere close to the number of metabolites expected to exist in the plant kingdom (Fernie and Stitt, 2012). Our first study showed that metabolites sampled from leaf tissue were worse predictors than metabolites sampled from young roots (Westhues et al., 2017). Mediocre predictive ability of genetic values using metabolites sampled from leaf tissue were previously reported in maize inbred lines (Riedelsheimer et al., 2012) and confirmed in two recent studies using a maize association panel (Wen Weiwei et al., 2018) and a genetically diverse set of rice inbred lines (Wei et al., 2018) whereas metabolites sampled from root tissue could match

(Westhues et al., 2017) or even exceed (Schrag et al., 2018) predictive abilities achieved with genomic information for the trait 'yield'. In human genetics, the Genotype-Tissue Expression (GTEx) consortium was initiated to study multi-tissue gene regulation using close to 50 tissues from almost 900 deceased donors (Lonsdale et al., 2013). Regarding tissue type, some of the major findings of GTEx-based studies are: (i) over half of the detected gene expression QTL (eQTL) are tissue specific (Gibson, 2015), (ii) not all previously detected eQTL were biologically plausible based on the causative tissue (Gibson, 2015) and (iii) shared eQTL effects were confirmed across tissues having biologically meaningful similarities (Ongen et al., 2017). In addition to tissue type, cell age at sampling is another important determinant of predictor properties. Our own studies could not discriminate effects arising from cell age from those originating from tissue type since the only predictor that was sampled from two different tissues was also sampled from different development stages on each tissue. However, studies in maize (Meng Dexuan et al., 2018), rice (Narsai et al., 2017), vervet monkeys (Jasinska et al., 2017) and humans (Gopalan et al., 2017) indicate that such effects act on the expression of genes.

Improving cost efficiency of alternative predictors

As shown, endophenotypes, even when used in isolation, have the potential to improve upon the prediction of genetic values compared to either pedigree or genomic information. Few studies have used endophenotypes for more than the identification of molecular trait QTL, yet. Multivariate analyses, treating gene expression variables as phenotypes, are theoretically well-suited for enriching predictions with information from partially missing features but quickly become computationally prohibitive with features sets as large as those provided by RNA-seq, particularly when used in conjunction with cross-validation techniques. PrediXcan offered, for the first time, a practical framework going beyond endophenotype-based QTL mapping by inferring gene expression data for individuals having only genetic information from individuals having both, genetic information as well as gene expression data to correlate these features with disease traits (GTEx Consortium et al., 2015). We consider this study exciting because it shows a way for reconciling higher costs of endophenotypes, compared to pedigree and genomic data, with their suggested potential for improving genetic value prediction. Unfortunately, their method suffers from an inability to account for the incurred imputation error. The single-step framework, which is widely used in animal breeding, has been intensively studied and improved to address this particular problem in the imputation of genotypic information from animals having both, pedigree and genomic information (Christensen and Lund, 2010; Fernando et al., 2014; Legarra et al., 2009). Our third study is the first investigation of the applicability of the single-step framework to the imputation of a quantitative predictor. While this framework has been developed assuming heritabilities equal to one for the features to be imputed, gene expression data used throughout our study were sampled under highly standardized conditions, thereby reducing the influence of environmental perturbations. Applying this framework in the prediction of maize inbred lines, using complete genomic and incomplete gene expression data, yielded merely minor improvements over those observed when using genomic information alone. Similar results were obtained for our hybrid data when comparing predictive abilities based on genomic data to those achieved from imputed gene expression data, which we attributed to the considerably higher predictive ability for genomic data in the full set of 1,521 hybrids compared to that achieved with the core set of 685 hybrids. When imputing transcripts in the hybrid data set via pedigree information, however, a substantial improvement in predictive ability over that achieved with pedigree data was observed. In conclusion, using endophenoypes in the single-step framework is highly promising for genetic value prediction of both, inbred lines and hybrids, when the discrepancy in predictive ability between the two predictors is moderate to large.

Conclusions

Our research demonstrated that endophenotypes, and gene expression data in particular, can be of great utility for improving the prediction of genetic values for complex traits recorded on hybrid plants. Our specific conclusions are:

(a) endophenotypes seem to be particularly suited for the prediction of yield-related traits,

(b) combining different predictors yields stable predictive abilities across diverse agronomic traits,

(c) to be highly predictive, endophenotypes should be sampled under highly standardized conditions from young tissue,

(d) covering merely a subset of all genotypes with both, endophenotypes and pedigree/genomic information can leverage the positive properties of endophenotypes for genetic value prediction at reduced costs.

6. Summary

One of the great challenges for plant breeders is dealing with the vast number of putative candidates, which cannot be tested exhaustively in multi-environment field trials. Using pedigree records helped breeders narrowing down the number of candidates substantially. With pedigree information, only a subset of candidates need to be subjected to exhaustive tests of their phenotype whereas the phenotype of the majority of untested relatives is inferred from their common pedigree. A caveat of pedigree information is its inability to capture Mendelian sampling and to accurately reflect relationships among individuals. This shortcoming was mitigated with the advent of marker assays covering regions harboring causal quantitative trait loci. Today, the prediction of untested candidates using information from genomic markers, called 'genomic prediction', is a routine procedure in larger plant breeding companies. Genomic prediction has revolutionized the prediction of traits with complex genetic architecture but, just as pedigree, cannot properly capture physiological epistasis, referring to complex interactions among genes and endophenotypes, such as RNA, proteins and metabolites. Given their intermediate position in the genotype-phenotype cascade, endophenotypes are expected to represent some of the information missing from the genome, thereby potentially improving predictive abilities.

In a first study we explored the ability of several predictor types to forecast genetic values for complex agronomic traits recorded on maize hybrids. Pedigree and genomic information were included as the benchmark for evaluating the merit of metabolites and gene expression data

in genetic value prediction. Metabolites, sampled from maize plants grown in field trials, were poor predictors for all traits. Conversely, root-metabolites, grown under controlled conditions, were moderate to competitive predictors for the traits 'fat' as well as 'dry matter yield'. Gene expression data outperformed other individual predictors for the prediction of genetic values for 'protein' and the economically most relevant trait 'dry matter yield'. A genome-wide association study suggested that gene expression data integrated SNP interactions. This might explain the superior performance of this predictor type in the prediction of 'protein' and 'dry matter yield'.

Small RNAs were probed for their potential as predictors, given their involvement in transcriptional, post-transcriptional and post-translational regulation. Regardless of the trait, small RNAs could not outperform other predictors. Combinations of predictors did not considerably improve the predictive ability of the best single predictor for any trait but improved the stability of their performance across traits. By assigning different weights to each predictor, we evaluated each predictor's optimal contribution for attaining maximum predictive ability. This approach revealed that pedigree, genomic information and gene expression data contribute equally when maximizing predictive ability for 'grain dry matter content'. When attempting to maximize predictive ability for 'grain yield', pedigree information was superfluous.

For genotypes having only genomic information, gene expression data were imputed by using genotypes having both, genomic as well as gene expression data. Previously, this single-step prediction framework was only used for qualitative predictors. Our study revealed that this framework can be employed for improving the cost-effectiveness of quantitative endophenotypes in hybrid prediction. We hope that these studies will further promote exploring endophenotypes as additional predictor types in breeding.

7. Zusammenfassung

Eine der größten Herausforderungen der Pflanzenzüchtung ist der Umgang mit der enormen Anzahl von Kandidaten, die nicht vollständig in mehrortigen Versuchen geprüft werden können. Die Nutzung von Verwandtschaftsbeziehungen hilft Züchtern die Anzahl dieser Kandidaten erheblich zu reduzieren. In diesem Fall muss nur ein Teil der Kandidaten phänotypisch geprüft werden. Für die übrigen, ungetesten Verwandten wird der Phänotyp hingegen mit Hilfe des Stammbaums vorhergesagt. Ein Nachteil von Stammbauminformationen ist, dass sie Zufallsprozesse Mendelscher Vererbung nicht erfassen und somit nicht präzise die genetische Ähnlichkeit zwischen Individuen wiedergeben. Die Nutzung von Marker-Chips, welche Genomregionen mit kausaler Beziehung zur Ausprägung phänotypischer Merkmale abdecken, konnte an dieser Stelle eine Verbesserung erzielen. Inzwischen ist die Nutzung von Markerinformationen zur Vorhersage ungetester Kandidaten - gemeinhin als 'Genomische Selektion' bezeichnet - in größeren Pflanzenzüchtungsunternehmen bereits Routine. Genomische Selektion hat die Vorhersage von Merkmalen mit komplexer genetischer Architektur revolutioniert. Wie Stammbauminformationen, so können auch genomische Informationen physiologische Epistasie, welche komplexe Interaktionen zwischen Genen und Endophänotypen wie RNA, Proteinen und Metaboliten beschreibt, nicht adäquat abbilden. Aufgrund ihrer Einbettung innerhalb der Genotyp-Phänotyp-Kaskade wird erwartet, dass sie Informationen, die nicht durch das Genom repräsentiert werden, abbilden. Auf diesem Weg könnten Endophänotypen möglicherweise die Vorhersagegenauigkeit gegenüber genomischen Informationen verbessern.

In einer ersten Studie untersuchten wir die Eignung unterschiedlicher Klassen von Prädiktoren zur Vorhersage genetischer Werte für komplexe agronomische Merkmale bei Hybridmais. Stammbaum- sowie genomische Informationen wurden als Referenz zur Bewertung der Eignung von Metabolit- und Genexpressionsdaten für die Vorhersage genetischer Werte herangezogen. Metabolite, die von Maispflanzen aus dem Feld entnommen wurden, erwiesen sich als wenig geeignet für die Vorhersage der untersuchten Merkmale. Im Gegensatz dazu erwiesen sich Wurzelmetabolite, entnommen von Maispflanzen, welche unter kontrollierten Bedingungen im Gewächshaus angezogen wurden, als akzeptable Prädiktoren für die Vorhersage der Merkmale "Fett" und "Trockensubstanzgehalt". Genexpressionsdaten waren der überlegene Prädiktor zur Vorhersage genetischer Werte für die Merkmale "Protein" sowie das ökonomisch wichtigste Merkmal "Trockenmasseertrag". Eine genomweite Assoziationskartierung deutete darauf hin, dass Genexpressionsdaten Interaktionen zwischen Genorten integrieren. Dies könnte die überlegene Eignung dieser Prädiktorenklasse zur Vorhersage der Merkmale "Protein" und "Trockenmasseertrag" erklären.

Small RNAs wurden in einer zweiten Studie auf ihre Eignung als Prädiktoren untersucht, da sie an der Regulierung transkriptionaler, post-transkriptionaler und post-translationaler Prozesse beteiligt sind. Unabhängig vom Merkmal konnten small RNAs andere Prädiktoren nicht übertreffen. Obwohl keine Kombination von Prädiktoren deutlich die Vorhersagegenauigkeit der besten einzelnen Prädiktorenklasse übertreffen konnte, gewährleistete die Nutzung mehrer Prädiktoren die höchste Stabilität der Vorhersagen über Merkmale hinweg. Indem wir jedem Prädiktor ein unterschiedliches Gewicht zuwiesen, konnten wir deren optimale Beiträge zur Maximierung der Vorhersagegenauigkeit bestimmen. Dieser Ansatz zeigte, dass Stammbauminformationen, genomische Informationen sowie Genexpressionsdaten zu gleichen Anteilen zur Maximierung der Vorhersagegenauigkeit beim Merkmal "Korntrockensubstanzgehalt" beitrugen. Zur Maximierung der Vorhersagegenauigkeit des Merkmals "Kornertrag" waren Stammbauminformationen hingegen unerheblich.

Für Genotypen, die lediglich mit genomischer Information abgedeckt waren, imputierten wir Genexpressionsdaten mit Hilfe solcher Genotypen für die sowohl genomische Informationen als auch Genexpressionsdaten vorlagen. Bis dato wurde dieser ”single-step” Vorhersageansatz lediglich für qualitative Prädiktoren verwendet. Unsere Studie zeigte, dass dieser Ansatz zur Verbesserung der Kosteneffizienz quantitativer Prädiktoren in der Hybridleistungsvorhersage genutzt werden kann. Wir hoffen mit diesen Studien einen Anstoß für weiterführende Forschungsarbeiten über den Einsatz von Endophänotypen als zusätzliche Prädiktoren in der Züchtung gegeben zu haben.

Bibliography

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci 93(2):743–52, DOI 10.3168/jds.2009-2730, NIHMS150003

Albrecht T, Auinger HJ, Wimmer V, Ogutu JO, Knaak C, Ouzunova M, Piepho HP, Schön CC (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor Appl Genet 127(6):1375–1386, DOI 10.1007/s00122-014-2305-z

Alonso-Peral MM, Trigueros M, Sherman B, Ying H, Taylor JM, Peacock WJ, Dennis ES (2016) Controls of gene expression in developing embryos of Arabidopsis hybrids. Plant J 89:927–939, DOI 10.1111/tpj.13432

Argillier O, Méchin V, Barrière Y (2000) Inbred line evaluation and breeding for digestibility-related traits in forage maize. Crop Sci 40(6):1596–1600, DOI 10.2135/cropsci2000.4061596x

Arrivault S, Guenther M, Ivakov A, Feil R, Vosloh D, Van Dongen JT, Sulpice R, Stitt M (2009) Use of reverse-phase liquid chromatography, linked to tandem mass spectrometry, to profile the Calvin cycle and other metabolic intermediates in Arabidopsis rosettes at different carbon dioxide concentrations. Plant J 59(5):824–839, DOI 10.1111/j.1365-313X.2009.03902.x

Ashraf B, Edriss V, Akdemir D, Autrique E, Bonnett D, Crossa J, Janss L, Singh R, Jannink JL (2016) Genomic prediction using phenotypes from pedigreed lines with no marker data. Crop Sci 56(3):957–964, DOI 10.2135/cropsci2015.02.0111

Asiago VM, Hazebroek J, Harp T, Zhong C (2012) Effects of Genetics and Environment on the Metabolome of Commercial Maize Hybrids: A Multisite Study. J Agric Food Chem 60(46):11,498–11,508, DOI 10.1021/jf303873a

Bernardo R (1994) Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Sci 34(1):20, DOI 10.2135/cropsci1994.0011183X003400010003x

Bernardo R (1996) Best Linear Unbiased Prediction of Maize Single-Cross Performance. Crop Sci 36:50–56

Bernardo R (2008) Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. Crop Sci 48(5):1649, DOI 10.2135/cropsci2008.03.0131

Bernardo R (2010) Breeding for Quantitative Traits in Plants, 2nd edn. Stemma Press, Woodbury, MN

Bernardo R (2014) Genomewide Selection when Major Genes Are Known. Crop Sci 54(1):68, DOI 10.2135/cropsci2013.05.0315

Brandenburg JT, Mary-Huard T, Rigaill G, Hearne SJ, Corti H, Joets J, Vitte C, Charcosset A, Nicolas SD, Tenaillon MI (2017) Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. PLoS Genet 13(3):e1006,666

Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436(7051):701–703, DOI 10.1038/nature03865

Brøndum R, Su G, Janss L, Sahana G, Guldbrandtsen B, Boichard D, Lund M (2015) Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci 98(6):4107–4116, DOI 10.3168/jds.2014-9005

Brown AA, Buil A, Vinuela A, Lappalainen T, Zheng HF, Richards JB, Small KS, Spector TD, Dermitzakis ET, Durbin R (2014) Genetic interactions affecting human gene expression identified by variance association mapping. eLife 2014(3):1–16, DOI 10.7554/eLife.01381

Bustos-Korts D, Malosetti M, Chapman S, Biddulph B, van Eeuwijk F (2016) Improvement of Predictive Ability by Uniform Coverage of the Target Genetic Space. G3 6(November):g3.116.035,410, DOI 10.1534/g3.116.035410

Caldana C, Degenkolbe T, Cuadros-Inostroza A, Klie S, Sulpice R, Leisse A, Steinhauser D, Fernie AR, Willmitzer L, Hannah Ma (2011) High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions. Plant J 67(5):869–84, DOI 10.1111/j.1365-313X.2011.04640.x

de los Campos G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11(12):880–6, DOI 10.1038/nrg2898

Christensen OF, Lund MS (2010) Genomic prediction when some animals are not genotyped. Genet Sel Evol 42:2, DOI 10.1186/1297-9686-42-2

Civelek M, Lusis AJ (2014) Systems genetics approaches to understand complex traits. Nat Rev Genet 15(1):34–48, DOI 10.1038/nrg3575

Clark Sa, Hickey JM, Daetwyler HD, van der Werf JH (2012) The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol 44(1):4, DOI 10.1186/1297-9686-44-4, URL http://www.gsejournal.org/content/44/1/4

Crossa J, De Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2):713–724, DOI 10.1534/genetics.110.118521

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112(1):48–60, DOI 10.1038/hdy.2013.16

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams Ja (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185(3):1021–1031, DOI 10.1534/genetics.110.116855

Dalchau N, Baek SJ, Briggs HM, Robertson FC, Dodd AN, Gardner MJ, Stancombe MA, Haydon MJ, Stan GB, Gonçalves JM, Webb AAR (2011) The circadian oscillator gene GIGANTEA mediates a long-term response of the Arabidopsis thaliana circadian clock to sucrose. Proc Natl Acad Sci USA 108(12):5104–9, DOI 10.1073/pnas.1015452108, arXiv:1408.1149

Dan Z, Hu J, Zhou W, Yao G, Zhu R, Zhu Y, Huang W (2016) Metabolic prediction of important agronomic traits in hybrid rice (Oryza sativa L.). Nature Sci Rep 6(October 2015):1–9, DOI 10.1038/srep21732

Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet 3(1):22–32, DOI 10.1038/nrg701

Duenk P, Calus MPL, Wientjes YCJ, Bijma P (2017) Benefits of Dominance over Additive Models for the Estimation of Average Effects in the Presence of Dominance. G3 7(10):3405–3414, DOI 10.1534/g3.117.300113

Duvick DN (2005) Genetic progress in yield of United States maize (Zea mays L .). Maydica 50:193–202 e Souza MB, Cuevas J, Couto EGdO, Pérez-Rodríguez P, Jarquín D, Fritsche-Neto R, Burgueño J, Crossa J (2017) Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype Environment Interaction. G3 p g3.117.042341, DOI 10.1534/g3.117.042341

Falconer D, Mackay TFC (1996) Introduction to Quantitative Genetics, 4th edn. Pearson, Essex

Fernando RL, Grossman M (1989) Marker assisted selection using best linear unbiased prediction. Genet Sel Evol 21:467–477

Fernando RL, Dekkers JC, Garrick DJ (2014) A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel Evol 46(1):50, DOI 10.1186/1297-9686-46-50

Fernie AR (2007) The future of metabolic phytochemistry: Larger numbers of metabolites, higher resolution, greater understanding. Phytochemistry 68(22-24):2861–2880, DOI 10.1016/j.phytochem.2007.07.010

Fernie AR, Stitt M (2012) On the discordance of metabolomics with proteomics and transcriptomics: coping with increasing complexity in logic, chemistry, and network interactions scientific correspondence. Plant Physiol 158(3):1139–45, DOI 10.1104/pp.112.193235

Fischer S, Möhring J, Schön CC, Piepho HP, Klein D, Schipprack W, Utz HF, Melchinger aE, Reif JC (2008) Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim. Plant Breed 127(5):446–451, DOI 10.1111/j.1439-0523.2007.01475.x

Fisher RR (1918) The correlation between relatives on the supposition of mendelian inheritance. Trans Roy Soc Edin 52:399–433

Franks A, Airoldi E, Slavov N (2017) Extensive post-transcriptional regulation across human tissues. PLoS Comp Biol 13(5):e1005,535, DOI 10.1101/020206

Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 41(1):55, DOI 10.1186/1297-9686-41-55

Geiger HH, Melchinger AE, Schmidt G (1986) Analysis of factorial crosses between flint and dent maize inbred lines for forage performance and quality traits. In: Dolstra O, Miedema P (eds) Breeding of Silage Maize, Pudoc, Wageningen, pp 147–154

Gerke JP, Edwards JW, Guill KE, Ross-Ibarra J, McMullen MD (2015) The genomic impacts of drift and selection for hybrid performance in maize. Genetics 201(3):1201–1211, DOI 10.1534/genetics.115.182410, 1307.7313

Gibson BG (2015) GTEx detects genetic effects. Science 348(6235):640–641, DOI 10.1126/science.aab3002

Giraud H, Bauland C, Falque M, Madur D, Combes V, Jamin P, Monteil C, Laborde J, Palaffre C, Gaillard A, Blanchard P, Charcosset A, Moreau L (2017) Reciprocal Genetics: Identifying QTL for General and Specific Combining Abilities in Hybrids Between Multiparental Populations from Two Maize (Zea mays L.) Heterotic Groups. Genetics 207(3):1167–1180, DOI 10.1534/genetics.117.300305

Gopalan S, Carja O, Fagny M, Patin E, Myrick JW, McEwen LM, Mah SM, Kobor MS, Froment A, Feldman MW, Quintana-Murci L, Henn BM (2017) Trends in DNA Methylation with Age Replicate Across Diverse Human Populations. Genetics 206(3):1659–1674, DOI 10.1534/genetics.116.195594

GTEx Consortium, Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ, Im HK (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47(9):1091–1098, DOI 10.1038/ng.3367

Guo Z, Magwire MM, Basten CJ, Xu Z, Wang D (2016) Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theor Appl Genet 129(12):2413–2427, DOI 10.1007/s00122-016-2780-5

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics 194(3):597–607, DOI 10.1534/genetics.113.152207

Hall BD, Fox R, Zhang Q, Baumgarten A, Nelson B, Cummings J, Drake B, Phillips D, Hayes K, Beatty M, Zastrow-Hayes G, Zeka B, Hazebroek J, Smith S (2016) Comparison of genotypic and expression data to determine distinctness among inbred lines of maize for granting of plant variety protection. Crop Sci 56(4):1443–1459

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Genomic selection in dairy cattle: progress and challenges. Dairy Sci 92(2):433–43, DOI 10.3168/jds.2008-1646

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic Selection for Crop Improvement. Crop Sci 49(1):1, DOI 10.2135/cropsci2008.08.0512

Henderson CR (1949) Estimation of changes in herd environment. J Dairy Sci 32:5

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci 52(1):146, DOI 10.2135/cropsci2011.06.0297

Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4(2):1–10, DOI 10.1371/journal.pgen.1000008

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, Burgueño J, de los Campos G (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127(3):595–607, DOI 10.1007/s00122-013-2243-1

Jasinska AJ, Zelaya I, Service SK, Peterson CB, Cantor RM, Choi OW, DeYoung J, Eskin E, Fairbanks LA, Fears S, Furterer AE, Huang YS, Ramensky V, Schmitt CA, Svardal H, Jorgensen MJ, Kaplan JR, Villar D, Aken BL, Flicek P, Nag R, Wong ES, Blangero J, Dyer TD, Bogomolov M, Benjamini Y, Weinstock GM, Dewar K, Sabatti C, Wilson RK, Jentsch JD, Warren W, Coppola G, Woods RP, Freimer NB (2017) Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate. Nat Genet 49(12):1714–1721, DOI 10.1038/ng.3959

Jiang Y, Reif JC (2015) Modelling epistasis in genomic selection. Genetics 201(2):759–768, DOI 10.1534/genetics.115.177907

Kacser H, Burns JA (1981) The molecular basis of dominance. Genetics 97:639–666

Kadam D, Potts S, Bohn MO, Lipka AE, Lorenz A (2016) Genomic Prediction of Hybrid Combinations in the Early Stages of a Maize Hybrid Breeding Pipeline. G3 6:3443–3453, DOI 10.1101/054015

Kahvejian A, Quackenbush J, Thompson JF (2008) What would you do if you could sequence everything? Nat Biotechnol 26(10):1125–1133, DOI nbt1494 [pii]1̊0.1038/nbt1494 [doi]

Kim H, Grueneberg A, Vazquez AI, Hsu S, de los Campos G (2017) Will Big Data Close the Missing Heritability Gap? Genetics 207(3):1135–1145, DOI 10.1534/genetics.117.300271

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

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PaC, Monlong J, Rivas Ma, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501(7468):506–11, DOI 10.1038/nature12531, NIHMS150003

Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. J Dairy Sci 92(9):4656–4663, DOI 10.3168/jds.2009-2061

Li S, Castillo-Gonzalez C, Yu B, Zhang X (2016) The functions of plant small RNAs in development and in stress responses. Plant J pp 654–670, DOI 10.1111/tpj.13444

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 16(6):321–332, DOI 10.1038/nrg3920

Lisec J, Römisch-Margl L, Nikoloski Z, Piepho HP, Giavalisco P, Selbig J, Gierl A, Willmitzer L (2011) Corn hybrids display lower metabolite variability and complex metabolite inheritance patterns. Plant J 68(2):326–336, DOI 10.1111/j.1365-313X.2011.04689.x

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson

E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi L, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struewing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J, Moore HF (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45(6):580–585, DOI 10.1038/ng.2653

Mackay TFC (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet 15(1):22–33, DOI 10.1038/nrg3627

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10(8):565–77, DOI 10.1038/nrg2612

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12(10):671–682, DOI 10.1038/nrg3068, 209

Martini JWR, Wimmer V, Erbe M, Simianer H (2016) Epistasis and covariance: how gene interaction translates into genomic relationship. Theor Appl Genet 129(5):963–976, DOI 10.1007/s00122-016-2675-5

Mathew B, Léon J, Sillanpää MJ (2018) A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. Heredity 120(4):356–368, DOI 10.1038/s41437-017-0023-4

Melchinger AE, Gumber RK (1998) Overview of heterosis and heterotic groups in agronomic crops. In: Lamkey K, Staub J (eds) Concepts and breeding of heterosis in crop plants, CSSA, Madison, p 16

Melchinger AE, Utz HF, Schön CC (1998) Quantitative Trait Locus (QTL) Mapping Using Different Testers and Independent Population Samples in Maize Reveals Low Power of QTL Detection and Large Bias in Estimates of QTL Effects. Genetics 149:383–403

Meng Dexuan, Zhao Jianyu, Zhao Cheng, Luo Haishan, Xie Mujiao, Liu Renyi, Lai Jinsheng, Zhang Xiaolan, Jin Weiwei (2018) Sequential gene activation and gene imprinting during early embryo development in maize. Plant J 93(3):445–459, DOI 10.1111/tpj.13786

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–29

Mortimer Sa, Kidwell MA, Doudna Ja (2014) Insights into RNA structure and function from genome-wide studies. Nat Rev Genet 15(7):469–79, DOI 10.1038/nrg3681

Mrode RA (2014) Linear Models for the Prediction of Animal Breeding Values, 3rd edn. CABI, Oxfordshire, DOI 10.1017/CBO9781107415324.004, arXiv:1011.1669v3

Narsai R, Secco D, Schultz MD, Ecker JR, Lister R, Whelan J (2017) Dynamic and rapid changes in the transcriptome and epigenome during germination and in developing rice (. Plant J 89:805–824, DOI 10.1111/tpj.13418

Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJL (2013) Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor Appl Genet 126(2):289–305, DOI 10.1007/s00122-012-1971-y

Ongen H, Brown AA, Delaneau O, Panousis NI, Nica AC, GTEx Consortium, Dermitzakis ET (2017) Estimating the causal tissues for complex traits and diseases. Nat Genet 49(12):1676–1683, DOI 10.1038/ng.3981

Parisseaux B, Bernardo R (2004) In silico mapping of quantitative trait loci in maize. Theor Appl Genet 109(3):508–14, DOI 10.1007/s00122-004-1666-0

Patti GJ, Yanes O, Siuzdak G (2012) Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 13(4):263–9, DOI 10.1038/nrm3314

Pérez-Enciso M, Rincón JC, Legarra A (2015) Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol 47(1):43, DOI 10.1186/s12711-015-0117-5, URL http://www.gsejournal.org/content/47/1/43

Reif JC, Gumpert F, Fischer S, Melchinger AE (2007) Impact of interpopulation divergence on additive and dominance variance in hybrid populations. Genetics 176(3):1931–1934, DOI 10.1534/genetics.107.074146

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44(2):217–20, DOI 10.1038/ng.1033

Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 16:85–97, DOI 10.1038/nrg3868

Rudd JJ, Kanyuka K, Hassani-Pak K, Derbyshire M, Andongabo A, Devonshire J, Lysenko A, Saqi M, Desai NM, Powers SJ, Hooper J, Ambroso L, Bharti A, Farmer A, Hammond-Kosack KE, Dietrich RA, Courbot M (2015) Transcriptome and metabolite profiling of the infection cycle of Zymoseptoria tritici on wheat reveals a biphasic interaction with plant immunity involving differential pathogen chromosomal contributions and a variation on the hemibiotrophic lifestyle def. Plant Physiol 167(3):1158–85, DOI 10.1104/pp.114.255927

Sackton TB, Hartl DL (2016) Perspective genotypic context and epistasis in individuals and populations. Cell 166:279–287, DOI 10.1016/j.cell.2016.06.047

Schopp P, Müller D, Technow F, Melchinger AE (2017) Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium. Genetics 205(1):441–454, DOI 10.1534/genetics.116.193243

Schrag TA, Melchinger AE, Sørensen A, Frisch M (2006) Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. Theor Appl Genet 113(6):1037–47, DOI 10.1007/s00122-006-0363-6

Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE (2018) Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. Genetics 208(4):1373–1385, DOI 10.1534/genetics.117.300374

Seifert F, Thiemann A, Grant-Downton R, Edelmann S, Rybka D, Schrag TA, Frisch M, Dickinson HG, Melchinger AE, Scholten S (2018a) Parental Expression Variation of Small RNAs Is Negatively Correlated with Grain Yield Heterosis in a Maize Breeding Population. Front Plant Sci 9, DOI 10.3389/fpls.2018.00013

Seifert F, Thiemann A, Schrag TA, Rybka D, Melchinger AE, Frisch M, Scholten S (2018b) Small RNA-based prediction of hybrid performance in maize. BMC Genomics 19(1):371, DOI 10.1186/s12864-018-4708-8

Shen Yifei, Sun Shuo, Hua Shuijin, Shen Enhui, Ye Chu-Yu, Cai Daguang, Timko Michael P, Zhu Qian-Hao, Fan Longjiang (2017)

Analysis of transcriptional and epigenetic changes in hybrid vigor of allopolyploid Brassica napus uncovers key roles for small RNAs. Plant J 91(5):874–893, DOI 10.1111/tpj.13605

Shull GH (1908) The Composition of a Field of Maize. J Heredity 4(1):296–301

Smith OS (1986) Covariance between line per se and testcross performance. Theor Appl Genet 2:540–543

Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? Nat Rev Genet 16(1):33–44, DOI 10.1038/nrg3821

Sprague GF, Tatum LA (1942) General vs. Specific Combining Ability in Single Crosses of Corn. J Am Soc Agron pp 923–932

Springer NM, Stupar RM (2007a) Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. Plant Cell 19(8):2391–2402, DOI 10.1105/tpc.107.052258

Springer NM, Stupar RM (2007b) Allelic variation and heterosis in maize: How do two halves make more than a whole? Genome Res 17(3):264–275, DOI 10.1101/gr.5347007

Stupar RM, Gardiner JM, Oldre AG, Haun WJ, Chandler VL, Springer NM (2008) Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. BMC Plant Biol 8(33):1–19, DOI 10.1186/1471-2229-8-33

Sweetlove, Nielsen Jens, Fernie Alisdair R (2016) Engineering central metabolism – a grand challenge for plant biologists. Plant J 90(4):749–763, DOI 10.1111/tpj.13464

Technow F, Riedelsheimer C, Schrag Ta, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125(6):1181–94, DOI 10.1007/s00122-012-1905-8

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197:1343–1355, DOI 10.1534/genetics.114.165860

Thiemann A, Fu J, Seifert F, Grant-Downton RT, Schrag Ta, Pospisil H, Frisch M, Melchinger AE, Scholten S (2014) Genome-wide meta-analysis of maize heterosis reveals the potential role of additive gene expression at pericentromeric loci. BMC Plant Biol 14(88):1–14, DOI 10.1186/1471-2229-14-88

Tzin V, Fernandez-Pozo N, Richter A, Schmelz EA, Schoettner M, Schäfer M, Ahern KR, Meihls LN, Kaur H, Huffaker A, Mori N, Degenhardt J, Mueller LA, Jander G (2015) Dynamic Maize Responses to Aphid Feeding Are Revealed by a Time Series of Transcriptomic and metabolomic assays. Plant Physiol 169(November):1727–1743, DOI 10.1104/pp.15.01039

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92(1):16–24, DOI 10.3168/jds.2008-1514

Vazquez AI, Veturi YC, Behring M, Shrestha S, Kirst M, Resende Jr MF, de los Campos G (2016) Increased proportion of variance explained and prediction accuracy of survival of breast cancer patients with use of whole-genome multi-omic profiles. Genetics 203(3):1425–1438, DOI 10.1534/genetics.115.185181

Vitezica ZG, Legarra A, Toro MA, Varona L (2017) Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. Genetics 206(3):1297–1307, DOI 10.1534/genetics.116.199406, http://www.genetics.org/content/206/3/1297.full.pdf

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics 10(1):57–63, DOI 10.1038/nrg2484, NIHMS150003

Wedzony M, Forster B, Zur I, Golemiec E, Scechynska-Hebda M, Dubas E, Gotebiowska G (2009) Progress in Doubled Haploid Technology in Higher Plants. In: Touarev A, Forster BP, Mohan JS (eds) Advances in Haploid Production in Higher Plants, Springer, chap 1, p 33

Wei J, Wang A, Li R, Qu H, Jia Z (2018) Metabolome-wide association studies for agronomic traits of rice. Heredity 120(4):342–355, DOI 10.1038/s41437-017-0032-3

Wen Weiwei, Jin Min, Li Kun, Liu Haijun, Xiao Yingjie, Zhao Mingchao, Alseekh Saleh, Li Wenqiang, de Abreu e Lima Francisco, Brotman Yariv, Willmitzer Lothar, Fernie Alisdair R, Yan Jianbing (2018) An integrated multi-layered analysis of the metabolic networks of different tissues uncovers key genetic components of primary metabolism in maize. Plant J 93(6):1116–1128, DOI 10.1111/tpj.13835

Westhues M, Schrag TA, Heuer C, Thaller G, Utz HF, Schipprack W, Thiemann A, Seifert F, Ehret A, Schlereth A, Stitt M, Nikoloski Z, Willmitzer L, Schön CC, Scholten S, Melchinger AE (2017) Omics-based hybrid prediction in maize. Theor Appl Genet 130(9):1927–1939, DOI 10.1007/s00122-017-2934-0

Witt S, Galicia L, Lisec J, Cairns J, Tiessen A, Araus JL, Palacios-Rojas N, Fernie AR (2012) Metabolic and phenotypic responses of greenhouse-grown maize hybrids to experimentally controlled drought stress. Mol Plant 5(2):401–17, DOI 10.1093/mp/ssr102

Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14(7):507–15, DOI 10.1038/nrg3457

Xu S, Xu Y, Gong L, Zhang Q (2016) Metabolomic prediction of yield in hybrid rice. Plant J 88(2):219–227, DOI 10.1111/tpj.13242

Xu Y, Xu C, Xu S (2017) Prediction and association mapping of agronomic traits in maize using multiple omic data. Heredity 119(3):174–184, DOI 10.1038/hdy.2017.27

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden Pa, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565–9, DOI 10.1038/ng.608

Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM (2017) Concepts, estimation and interpretation of SNP-based heritability. Nat Genet 49(9):1304–1310, DOI 10.1038/ng.3941

Zenke-Philippi C, Frisch M, Thiemann A, Seifert F, Schrag TA, Melchinger AE, Scholten S, Herzog E (2017) Transcriptome-based prediction of hybrid performance with unbalanced data from a maize breeding programme. Plant Breed 136:331–337, DOI 10.1111/pbr.12482

Zhao Y, Mette MF, Reif JC (2015) Genomic selection in hybrid breeding. Plant Breed 134(1):1–10, DOI 10.1111/pbr.12231

Zhong S, Toubia-rahme H, Steffenson BJ, Smith KP (2006) Molecular Mapping and Marker-Assisted Selection of Genes for Septoria Speckled Leaf Blotch Resistance in Barley. Phytopathology 96(9):993–999

Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, Tu Z, Brem RB, Bumgarner RE, Schadt EE (2012) Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. PLoS Biol 10(4), DOI 10.1371/journal.pbio.1001301

8. Acknowledgments

I am grateful to my academic supervisor Prof. Dr. Albrecht E. Melchinger for expressing his interest in my work early and for the support throughout my thesis.

Many thanks to Prof. Dr. Fred van Eeuwijk for hosting me at Wageningen University and Research Center. I am grateful to Martin Boer, Marcos Malossetti and Marco Bink for sharing their ideas with me and helping me become more efficient. Thanks to Federico Torretta and all the Ph.D. students there for making it a pleasant stay.

Thanks to Dr. Tobias A. Schrag for insightful discussions, particularly regarding the interpretation of the good performance of endophenotypes.

I would like to thank Claas Heuer for his support with implementing the prediction routines and for being a great mentor.

Many thanks to Beate Devezi-Savula, Franz Mauch, Heralt Pöschel, Jochen Jesse and Birgit Schilling for introducing me to the work at the institute as an undergraduate student. The great impression I gained, mattered a lot. I am indebted to Helga Kösling, Beate Devezi-Savula, Margit Lieb, and Susanne Meyer for resolving many organizational issues during my work.

Thanks to Dominik Müller always being open-minded about exploring the latest software for doing efficient, reproducible research.

I am grateful to Juliane Renner, Dominik Müller, Philipp Boeven, Pascal Schopp, Sen Han, Pedro Brauner-Correa, Felix Sattler, Jose Marulanda, Valheria Castiblanco, Anna Pucher and Alisa Sieber for extending my

student life throughout my Ph.D.

Sincere thanks to my parents and my sister for their constant support and encouragement.

Finally, thanks to Cathy for being sincerely interested in the topic and for being very supportive during the completion of this thesis.

9. Curriculum Vitae

Name:	Matthias Westhues
Date and place of birth:	04.06.1988, Ahlen, Germany

School Education

1999 - 2008	High school degree (Abitur), Städtisches Gymnasium Ahlen, Germany

University Education

2008 - 2011	B.Sc. Biobased Products and Bioenergy, University of Hohenheim, Germany
2011 - 2013	M.Sc. Crop Sciences, University of Hohenheim, Germany
2013 - 2019	Doctorate candidate in Applied Genetics and Plant Breeding (Prof. A. E. Melchinger), University of Hohenheim, Germany

Employment Record

2017 - 2019	Scientist - Phenotypic Data Analysis, KWS Saat SE, Einbeck, Germany

Matthias Westhues

Einbeck, 05.04.2019

10. Erklärung

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

(a) Bei der eingereichten Dissertation zum Thema *Comparison of 'omics' technologies for hybrid prediction* handelt es sich um meine eigentständig erbrachte Leistung.

(b) Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

(c) Ich habe nicht die Hilfe einer kommerziellen Promotionsvermittlung oder -beratung in Anspruch genommen.

(d) Die Bedeutung der eidesstattlichen Versicherung und der strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Erklärung sind mir bekannt.

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

Matthias Westhues

05.04.2019