Aus dem Institut für
Pflanzenzüchtung, Saatgutforschung und Populationsgenetik
der Universität Hohenheim
Fachgebiet Nutzpflanzenbiodiversität und Züchtungsinformatik
Prof. Dr. Karl J. Schmid

Evaluation of association mapping and genomic prediction in diverse barley and cauliflower breeding material

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften
vorgelegt
der Fakultät Agrarwissenschaften

von
Master of Science
Patrick Thorwarth
aus Göppingen

Stuttgart-Hohenheim
2017

Die vorliegende Arbeit wurde am 11. Oktober 2017 von der Fakultät Agrarwissenschaften als “Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften (Dr. sc. agr.)” angenommen.

Tag der mündlichen Prüfung: 22. Februar 2018

1. i.V.d. Prodekan:	Prof. Dr. J. Wünsche
Berichterstatter, 1. Prüfer:	Prof. Dr. K.J. Schmid
Mitberichterstatter, 2. Prüfer:	Prof. Dr. F. Ordon
3. Prüfer:	Prof. Dr. J. Bennewitz

Contents

[1] Schmid, K.J., Thorwarth, P. 2014. Genomic Selection in Barley Breeding. In: Kumlehn, J., Stein, N. (eds) Biotechnological Approaches to Barley Improvement. Biotechnology in Agriculture and Forestry. Springer, Berlin, Heidelberg 69:367–378

[2] Gawenda, I., Thorwarth, P., Günther, T., Ordon, F., Schmid, K.J. 2015. Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. Plant Breeding 134:28–39

[3] Thorwarth, P., Ahlemayer, J., Bochard, A.-M., Krumnacker, K., Blümel, H., Laubach, E., Knöchel, N., Cselényi, L., Ordon, F., Schmid, K.J. 2017. Genomic prediction ability for yieldrelated traits in German winter barley elite material. Theor Appl Genet 130:1669–1683

[4] Thorwarth, P. Yousef, E.A.A, Schmid, K.J. 2017. Genomic prediction and association mapping of curd-related traits in genebank accessions of cauliflower. Submitted to G3: Genes, Genomes, Genetics

Abbreviations

ANOVA	Analysis of variance
BL	Bayesian least absolute shrinkage and selection operator
BLUP	Best linear unbiased prediction
BRR	Bayesian ridge regression
CV	Cross validation
EN	Elastic net
GP	Genomic prediction
GS	Genomic selection
GWAM	Genome wide association mapping
LASSO	Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
MAS	Marker assisted selection
MSE	Mean squared error
PCA	Principle component analysis
QTL	Quantitative trait locus
RR	Ridge regression
RRBLUP	Ridge regression best linear unbiased prediction
SNP	Single nucleotide polymorphism

1. General Introduction

Plant breeding is a key factor to cope with the demand for increased production of high quality agricultural products under changing environmental conditions and limited resources. Due to technological progress made in sequencing and information technology, marker arrays and whole genome information are now available, providing information on the genetic constitution of individuals and enabling an increase in genetic gain. Genetic gain is defined as $\Delta G = \frac{i \times h^2 \times \sigma}{L}$, where i is the selection intensity, h^2 the accuracy, σ the variability of the population and L the breeding cycle length. The increase in genetic gain due to the usage of molecular markers is mainly driven by increasing i and decreasing L. The uses of molecular markers for plant breeding are versatile and the investigation and application an integral part of this thesis. Roughly, a classification into three groups can be made for the application of molecular markers in plant breeding: (1) The identification and localization of loci that affect genetic variation or of regions affecting a loci linked to a quantitative trait (QTL), (2) the usage of molecular markers for identifying genotypes with a favourable genetic makeup for the purpose of selection and (3) the assessment of genetic differentiation of individuals or populations. In the following sections a general introduction to the methods used and their development is given along with how these methods are integrated into the framework of this thesis.

From QTL mapping to Genome Wide Association Mapping

The idea of QTL mapping is to identify QTLs due to linkage disequilibrium (LD), the non-random association of alleles at different loci in a given population (Hill and Robertson 1968), between a known genetic marker and the unknown QTL. Therefore, an experimental population is created. This population is derived, in the simplest case, of progenies (F_1) of two inbred lines, by e.g. selfing the F_1 to generate the F_2 mapping population. The parents should differ in the mean value of the trait characteristic under investigation (Lander and Botstein 1989). The cross of two inbred lines leads to the creation of complete LD between loci, that differ between the lines (Lynch and Walsh 1998). To use LD for the identification of QTLs a high number of genetic markers, such as single nucleotide polymorphisms (SNPs), is required. If those marker cover the genome or region of interest, the identification of markers linked to QTLs is possible, due to differences in the genetic makeup at the locus influencing the trait under investigation.

Statistical methods are applied to perform QTL mapping. One of the earliest methods proposed is based on analysis of variance (ANOVA). There, the F_2 population is separated into groups based on the marker genotype and an F-statistic is used to compare the average trait performance of the groups (Broman 2001). This method has several disadvantages and methods based on the conditional probability of a QTL based on an observed marker, such as interval mapping or composite interval mapping, are commonly used (Mackey 2001). These methods, developed around the methodology of linear models and maximum likelihood estimates, provide a robust framework for QTL detection and localization. QTL mapping has some limitations such as a low mapping resolution due to the limited amount of recombinations occurring in the creation of the experimental population and the limited amount of allelic diversity between the parents (Korte et al. 2013).

Genome wide association mapping (GWAM) is a statistical method that can overcome the limitations of QTL mapping. Similar to QTL mapping

GWAM uses LD between markers and QTLs to detect associations between phenotype and genotype, but instead of using a controlled, experimental cross between selected parents GWAM relies on diverse, natural populations taking advantage of historical recombination events (Korte et al. 2013). Due to this, the mapping resolution is increased and a larger part of the genetic variation, which is segregating in the population can be revealed (Zhu et al. 2008). The usage of natural populations introduces the problem of confounding effects due to population structure, the occurrence of LD due to admixture and migration. This leads to significant marker-trait associations even though markers are not linked to QTLs which cause the phenotypic variation (Flint-Garcia et al. 2003). A further problem occurring particularly in plant breeding populations is cryptic relatedness of individuals (Devlin and Roeder 1999, Voight and Pritchard 2005). Cryptic relatedness refers of the occurrence of covariance between related individuals in the population under investigation. Several methods to cope with population structure and cryptic relatedness have been implemented (Sillanpäa 2011).

The general approach to correct for confounding effects is based on detecting population structure (Q) and kinship (K) calculated from genetic markers and include Q as fixed effect and K as random effect in the framework of a linear mixed model (Yu et al. 2006). Each marker is included as fixed effect in the model to test for a significant association between the marker and the phenotype and a correction for multiple testing, in order to control the false discovery rate, is applied (Benjamini and Hochberg 1995).

A crucial point in the detection of significant associations is a large population size to achieve a high enough power, especially if the detection of QTLs with a small effect is desired (Zhang et al. 2010). Power is defined as the ability to detect the causal QTL, or a marker in close linkage with the causal QTL. Power depends on the LD in the population, the genetic architecture, the sample size and data quality (phenotypic and genotypic, Abdurakhmonov and Abdukarimov 2008). One way to increase the power of a GWA study is to group linked markers together into a haplotype and to use it instead of a single-marker in the GWA model (Calus et al. 2009).

Nowadays, GWA studies are a well described and well investigated method for the identification and localization of QTLs, but not without limitations and challenges such as small population sizes, missing genotypes, rare alleles occurring at low frequencies, a complex genetic architecture and a limited power to detect QTLs with small effects. Investigating the impact of several parameters influencing the results of GWAM was one objective of this thesis.

From Marker Assisted Selection to Genomic Selection

The next step after the identification and localization of promising QTLs is to integrate their information into the breeding program. Marker assisted selection (MAS) provides the methodological framework for using the information of markers linked to QTLs, that predict the phenotypic value. This information can be used for early stage selection of single plants, especially for traits that are difficult to assess in field or greenhouse experiments (Collard and Mackill 2008). The inherent problem, which is passed down from the QTL detection stage to the QTL utilization stage, is the limited amount of genetic variance explained by the detected QTLs. Many traits used in plant breeding are complex traits, which do not follow Mendel's laws of inheritance as they are not controlled by a single gene with a large effect on the phenotype, but rather are polygenic and thus controlled by many genes each with a small effect (Fisher 1918). Due to this MAS is a useful tool for traits with a simple, monogenic architecture, the pyramiding of resistance genes and some other applications (Collard and Mackill 2008), but is of limited use for the improvement of complex traits, when compared to classical phenotypic selection (Moreau et al 2004; Bernardo 2008).

Instead of focusing on the detection of single QTLs with large effects, Meuwissen et al. (2001) suggested to use all available markers, linked to the unknown QTLs, for selection. The idea is to estimate the effect of all QTLs and sum them up to the breeding value of an individual. This breeding value can be used for genomic selection (GS) of superior genotypes without the necessity of a QTL identification and localization step of limited power. First, to utilize the framework of GS, a training population has to be phenotyped for a trait of interest and genotyped with genome-wide markers. The required marker density depends on the LD decay in the population. Then, marker effects are estimated based on a statistical model. These are then used to predict the genotypic value of individuals, which form the validation population. The individuals in the validation population are ranked according to their genotypic value and the best individuals can be selected without assessing their phenotypic value in field trials (Heffner 2009). The methodology of GS comprises several theoretical advantages which can potentially benefit a breeding program such as: (1) increase in genetic gain through reduction of the breeding cycle length and increased accuracy, (2) reduction of costs of a breeding program by decreasing the amount of genotypes that have to be tested in field trials, (3) prediction of all potential offspring genotypes and their performance (e.g. in hybrid breeding all factorial combinations can be predicted), (4) screening of genetic resources for genotypes with a promising genotypic value for a given trait, without the necessity to observe them in field trials (Nakaya and Isobe 2012, Daetwyler et al. 2013, Yu et al. 2016).

The basic model for the estimation of breeding values is based on the separation of the phenotype of an individual into components influencing its expression such as the genetic effects (additive, dominance and epistasis) and the environmental effects. Each parent inherits a random sample of half of its genes (additive genetic value) to its progeny. The sum of the additive genetic values of both parents is the breeding value of the offspring and thus the criteria for selection. Henderson (1949) developed the theoretical framework for the calculation of breeding values, called Best Linear Unbiased Prediction (BLUP). The linear mixed model has the following notation as described by Mrode (2014):

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is a vector of phenotypic observations, $\boldsymbol{b}$ a vector of fixed effects, $\boldsymbol{a}$ a vector of random effects, $\boldsymbol{e}$ a vector of residual effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices relating phenotypic observations to fixed and random effects, respectively. The simplified solution of the mixed model equation was presented by Henderson (1950) and has the following form as described by Mrode (2014):

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix} \tag{2}$$

Nowadays BLUP is the most important method in animal breeding for the genetic evaluation of animals (Mrode 2014) and is finding its standing in plant breeding, providing the statistical framework for the genomic prediction of breeding values. Several statistical models have been suggested for GS, since the paper of Meuwissen (2001), which mainly deals with the small n large p problem ($n << p$). Due to advances in the development of next-generation sequencing technology, a large number of genome wide SNP markers are available (p), which by far exceed the number of genotypes (n). This, leads to underdetermined systems of linear equations, which cannot be directly solved (de los Campos et al. 2013). Several solutions for this problem are available such as regularization, variable selection and Bayesian statistic, which can be categorized into parametric, semi-parametric and non parametric models. The main difference concerning the parametrization of a model is the assumptions made about the probability distributions of the variables in the model (Howard et al. 2014). The investigation of the influence of statistical models on the estimation of marker effects or genomic estimated breeding values is called genomic prediction (GP), where the final ranking and genomic selection of genotypes in the candidate population is not of direct concern. SNP marker can be directly included as predictor variables into the model, or used to calculate a realized relationship ($\boldsymbol{G}$) matrix, which describes the covariance among individuals. Using SNP markers directly, allows the estimation of marker effects, whereas including $\boldsymbol{G}$ provides a direct prediction of genomic estimated breeding values. A commonly used method for judging the performance of a model is the Pearson's correlation between genomic estimated breeding values and the true genotypic values ($\rho_{pac} = r(\hat{g}, g)$) in the training population. This measure is called prediction accuracy. As the true genotypic value is unknown, the prediction ability, which is the correlation between the estimated breeding value and the phenotypic value ($\rho_{pa} = r(\hat{g}, y)$) in the training set, is often assessed. A common approach to approximate the prediction accuracy is, to standardize the prediction ability with the square root of the heritability ($\hat{\rho_{pac}} = \frac{\rho_{pa}}{h}$). Several factors have an influence on the prediction ability such as LD and marker density (Zhong et al. 2009; Wientjes et al. 2013), relatedness of genotypes within and between training and validation sets (Wientjes et al. 2013; Habier et al. 2007), population structure (Guo et al. 2014; Isidro et al. 2015), genetic architecture and QTL number affecting a trait (Daetwyler et al. 2010), performance of statistical models (de los Campos et al. 2013), and the adjustment of phenotypic data (Bernal-Vasquez et al. 2014).

Investigating factors influencing the prediction ability and comparing statistical models for the estimation of marker effects and genomic estimated breeding values in barley (*Hordeum vulagare* L.) and cauliflower (*Brassica oleracea var. botrytis*) was a central goal of this thesis.

Increasing genetic variation by utilization of Genetic resources

Genetic variation, the heritable variation within and between populations, provides the basis for crop improvement (Rao and Hodgkin 2001). Genetic resources provide a natural richness of allelic variation and play an important part in the history of plant breeding as several important developments such as the introduction of dwarfing genes in wheat are based on allelic variation detected in exotic germplasm (Hedden 2003). The efforts for the conservation of plant genetic resources have steadily increased over the years and currently about 7.4 million accessions are conserved in more than 1,750 genebanks (Yu et al. 2016). Large parts of the conserved genetic variation remains underutilized as it is difficult and costly to evaluate the hidden potential of plant genetic resources (Wang et al. 2017). Recently, strategies based on genomic prediction methodology have been proposed to cope with these challenges (Longin and Reif 2014, Yu et al. 2016, Muleta et al. 2017, Wang et al. 2017) as genomic prediction provides a relatively cheap alternative for phenotyping large genebank collections for specific traits.

One part of this thesis was the evaluation of genome-wide association mapping and genomic prediction in elite and germplasm material.

Objectives

The goals of my research thesis were to investigate the feasibility of genome-wide association mapping and genomic prediction in self-fertilizing barley and outcrossing cauliflower populations, persisting of either elite material or a mixture of elite material and genetic resources. In particular, the objectives were to:

1. Compare single-marker and haplotype based methods for genome wide association studies

2. Investigate the effects of marker density, sample size and GWAS methods for detecting QTLs with additive and epistatic effects using simulated data.

3. Compare parametric, semi-parametric and non-parametric models for Genomic prediction

4. Assess the accuracy of phenotypic selection in comparison to genomic selection

5. Analyse the linkage disequilibrium and persistence of the linkage phase to derive the optimal marker density in a given population

6. Investigate the effect of relatedness and population structure on the accuracy of Genomic prediction

7. Assess the usefulness of Genotyping-by-sequencing for the characterization of genetic resources and elite breeding material

8. Evaluate the effect of genotype imputation on GWAS and Genomic prediction

References

Abdurakhmonov, I.Y., Abdukarimov, A. 2008. Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources. International Journal of Plant Genomics. 2008:1–18

Benjamini, Y. Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 57:289–300

Bernal-Vasquez, A.M., Möhring, J., Schmidt, M., Schönleben, M., Schön, C.C., Piepho, H.P. 2014. The importance of phenotypic data analysis for genomic prediction - a case study comparing different spatial models in rye. BMC Genomics 15:646

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. CropSci 48:1649–1664.

Broman, K.W. 2001. Review of Statistical Methods for QTL Mapping in Experimental Crosses. Lab Anim 30:44-52

Calus, M.P.L., Meuwissen, T.H.E., Windig, J.J., Knol, E.F., Schrooten, C., Vereijken, A.L., Veerkamp, R.F. 2009. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. Genetics Selection Evolution 41:11

Collard, B.C.Y., Mackill, D.J. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Phil. Trans. R. Soc. 363:557–572

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

Daetwyler, H.D., Calus, M.P.L., Pong-Wong, R., de los Campos, G., Hickey, J.M. 2013. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193:347-365

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., Calus, M. P. L. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193:327–345

Devlin, B., Roeder, K. 1999. Genomic Control for Association Studies. Biometrics 55:997–1004

Flint-Garcia, S.A., Thornsberry J.M., Buckler, E.S. 2003. Structure of Linkage Disequilibrium in Plants. Annu. Rev. Plant Biol. 54:357–374

Fisher, R.A., 1918. The correlation between relatives on the supposition of mendelian inheritance. Trans R Soc Edin 52:399–433

Guo, Z., Tucker, D.M., Basten, C.J., Gandhi, H., Ersoz, E., Guo, B., Xu, Z., Wang, D., Gay, G. 2014. The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet 127:749–762

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

Hedden, P., 2003. The genes of the Green Revolution. Trends in Genetics 19:5–9

Heffner, E. L., Sorrells, M. E., Jannink, J. L. 2009. Genomic selection for crop improvement. Crop Scienc 49:1–12.7

Henderson, C.R. 1949. Estimation of changes in herd environment. Journal of Diary Science 32:709

Henderson, C.R. 1950.Estimation of genetic parameters. Annals of Mathematical Statistics 21:309

Hill, W.G., Robertson, A. 1968. Linkage disequilibrium in finite populations. Theoret. Appl. Genetics 38:226–231.

Howard, R., Carriquiry, A. L., Beavis, W. D. 2014. Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures. G3: Genes, Genomes, Genetics, 4:1027–1046

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

Korte, A., Farlow, A. 2013. The advantages and limitations of trait analysis with GWAS: a review.Plant Methods 9:29

Lander, E.S., Botstein, D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Longin, C.F.H., Reif, J.C. 2014. Redesigning the exploitation of wheat genetic resources. Trends in Plant Science 19:631–636

Lynch, M., Walsh,B. 1998. Genetics and Analysis of Quantitative traits. Sinauer Associates, Sunderland, Massachusetts.

Mackay, T.F.C. 2001. The genetic architecture of quantitative traits. Annu. Rev. Genet. 35:303–309

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819-1829

Moreau, L., Charcosset, A., Gallais, A. 2004. Experimental evaluation of several cycles of marker-assisted selection in maize. Euphytica. 137:111–8

Mrode, R.A., 2014. Linear models for the prediction of animal breeding values. 3rd ed. CABI

Muleta, K.T., Bulli, P., Zhang, Z., Chen, X., Pumphrey, M. 2017. Unlocking Diversity in Germplasm Collections via Genomic Selection: A Case Study Based on Quantitative Adult Plant Resistance to Stripe Rust in Spring Wheat. Plant Genome 10:1–15

Nakaya, A., Isobe, S.N. 2012. Will genomic selection be a practical method for plant breeding?. Annals of Botany 110:1303–1316

Rao, V.R., Hodgkin, T. 2001. Genetic diversity and conservation and utilization of plant genetic resources. Plant Cell, Tissue and Organ Culture 68:1–19

Sillanpäa,M.J., 2011. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. Heredity 106:511–519

Voight B.F., Pritchard, J.K. 2005. Confounding from Cryptic Relatedness in Case-Control Association Studies. PLoS Genet 1(3):e32

Wang, C., Hu, S., Gardner, C., Lübberstedt, T. 2017. Emerging Avenues for Utilization of Exotic Germplasm. Trends in Plant Science 22:624–637

Wientjes, Y.C.J., Veerkamp, R.F., Calus, M.P.L. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621–631

Yu, J., Pressoir, G., Briggs W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203–208

Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S.E., Roozeboom, K. L., Wang, D., Wang, M. L., Pederson, G. A., Tesso, T. T., Schnable, P. S., Bernardo, R., Yu, J. 2016. Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nature Plants 2:1–7

Zhang, Z., Ersoz, E., Lai, C.-Q.,Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., Buckler, E.S. 2010.Mixed linear model approach adapted for genome-wide association studies. Nature Genetics 42:355-360

Zhong, S., Dekkers, J.C.M., Fernando, R.L., 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

Zhu, C.,Gore, M., Buckler, E.S., Yu, J. 2008. Status and Prospects of Association Mapping in Plants. The Plant Genome 1:5–20

2. Genomic selection in barley breeding

Karl J. Schmid[1], Patrick Thorwarth[1]

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

Biotechnological Approaches to Barley Improvement. Biotechnology in Agriculture and Forestry, vol 69:367–378. Springer, Berlin, Heidelberg, 2014

The original publication is available at

`https://link.springer.com/chapter/10.1007/978-3-662-44406-1_19`

Abstract

Genomic Selection is used to improve breeding populations by using genome-wide markers for selection. Therefore, a training set is phenotyped for traits of interest and genotyped with a high density set of markers. Then, a genomic prediction model is trained using both the phenotypic and genotypic data. In the following selection cycles, individuals from a breeding population are only genotyped with the same marker set, and their genomic estimated breeding values (GEBV) are calculated. Individuals with a high GEBV are selected for the next cycle. Genomic Selection leads to significant

cost savings and to an increased selection gain per time unit as costly and time-consuming phenotypic selection does not have to be performed in every selection cycle. Both simulations and empirical studies showed a high accuracy of genomic prediction in barley breeding populations. The high level of linkage disequilibrium and the close genetic relationship present in barley breeding material allow the use of relatively small marker sets to test populations for Genomic Selection in barley breeding and suggest that this method will be highly useful for barley breeding.

3. Genomewide association studies in elite varieties of German winter barley using singlemarker and haplotypebased methods

Inka Gawenda[1*], Patrick Thorwarth[1*], Torsten Günther[1], Frank Ordon[2], Karl J. Schmid[1]

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

[2] Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance, Julius-Kühn Institute (JKI), 06484 Quedlinburg, Germany

[*] These authors contributed equally to this work.

Plant Breeding, 134:28–39, 2015

The original publication is available at https://onlinelibrary.wiley.com/doi/abs/10.1111/pbr.12237

Abstract

Genome wide association studies (GWAS) are a commonly used method to map qualitative and quantitative traits in plants. We compared existing single marker and haplotype based methods for association mapping with a focus on barley. Based on German winter barley genotypes, four different GWAS methods were tested for their power to detect significant associations in a large genome with a limited number of markers. We identified significant associations for yield and quality related traits using the iSelect array with 3886 mapped single nucleotide polymorphism (SNP) markers in a structured population consisting of 109 genotypes. Genome simulations with different numbers of genotypes, marker densities and marker effects were used to compare different GWAS methods. Results of simulations revealed a higher power in detecting significant associations for haplotype than for single marker approaches, but showed a higher false discovery rate for SNP detection, due to lack of correction for population structure. Our simulations revealed that a population size of about 500 individuals is required to detect QTLs explaining a small trait variance ($< 10\%$).

4. Genomic prediction ability for yield-related traits in German winter barley elite material

Patrick Thorwarth[1], Jutta Ahlemeyer[2], Anne-Marie Bochard[3], Kerstin Krumnacker[3], Hubert Blmel[4], Eberhard Laubach[5], Nadine Knöchel[6], Lszl Cselnyi[7], Frank Ordon[6], Karl J. Schmid[1]

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

[2] Deutsche Saatveredelung AG, Weissenburger Str. 5, 59557 Lippstadt, Germany

[3] Limagrain GmbH, Salder Strasse 4, 31226 PeineRosenthal, Germany

[4] Secoba Saatzucht GmbH, Feldkirchen 3, 85368 Moosburg, Germany

[5] Nordsaat-Saatzucht GmbH, Hofweg 8, 23899 GudowSegrahn, Germany

[6] Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance, Julius-Kühn Institute (JKI), 06484 Quedlinburg, Germany

[7] W. von Borries-Eckendorf GmbH & Co. KG, Hovedisser Str. 92, 33818 Leopoldshöhe, Germany

Theor Appl Genet (2017) 130: 1669.

The original publication is available at

`https://link.springer.com/article/10.1007%2Fs00122-017-2917-1`

Abstract

To warrant breeding progress under different environmental conditions the implementation and evaluation of new breeding methods is very important. Modern breeding approaches like genomic selection may significantly accelerate breeding progress. We assessed the potential of genomic prediction in a training population of 750 genotypes, consisting of multiple six-rowed winter barley (*Hordeum vulgare* L.) elite material families and old cultivars, which reflect the breeding history of barley in Germany. Crosses of parents selected from the training set were used to create a set of double-haploid families consisting of 750 genotypes. Those were used to confirm prediction ability estimates based on a cross validation with the training set material using 11 different genomic prediction models. Population structure was inferred with dimensionality reduction methods like discriminant analysis of principle components and the influence of population structure on prediction ability was investigated. In addition to the size of the training set, marker density is of crucial importance for genomic prediction. We used genome wide linkage disequilibrium and persistence of linkage phase as indicators to estimate that 11,203 evenly spaced markers are required to capture all QTL effects. Although a 9k SNP array does not contain a sufficient number of polymorphic markers for long term genomic selection, we obtained fairly high prediction accuracies ranging from 0.31 to 0.71 for the traits earing, hectoliter weight, spikes per square meter, thousand kernel weight and yield and show that they result from the close genetic relatedness of the material. Our work contributes to designing long-term genetic prediction programs for barley breeding.

5. Genomic prediction and association mapping of curd-related traits in genebank accessions of cauliflower

Patrick Thorwarth[1] , Eltohamy A.A. Yousef[2] and Karl J. Schmid[1]

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

[2] Department of Horticulture, Faculty of Agriculture, University of Suez Canal, Ismailia (41522), Egypt

G3: Genes, Genomes, Genetics, 2018 vol. 8 no. 2 707-718

The original publication is available at

http://www.g3journal.org/content/8/2/707.long

Abstract

Genetic resources are a valuable source of genetic variation for plant breeding. Genome-wide association studies (GWAS) and genomic prediction facilitate the analysis and utilization of useful genetic diversity for improving complex phenotypic traits in crop plants. We explored the potential of these methods for improving curd-related traits in cauliflower (*Brassica oleracea* var. *botrytis*) by combining 174 randomly selected cauliflower gene bank accessions from two different gene banks. The collection was genotyped with genotyping-by-sequencing (GBS) and phenotyped for six curd-related traits at two locations and three growing seasons. A GWAS analysis based on 120,693 single nucleotide polymorphisms identified a total of 24 significant associations for curd related traits. The potential for genomic prediction was assessed with a genomic best linear unbiased prediction model and BayesB. Prediction abilities ranged from 0.10 to 0.66 for different traits and did not differ between prediction methods. Imputation of missing genotypes only slightly improved prediction ability. Our results demonstrate that GWAS and genomic prediction in combination with GBS and phenotyping of highly heritable traits can be used to identify useful quantitative trait loci and genotypes among genetically diverse gene bank material for subsequent utilization as genetic resources in cauliflower breeding.

6. General Discussion

Genome-wide association mapping and genomic prediction are closing the gap in the toolbox of plant breeders and scientists by extending classical methods such as QTL mapping and marker assisted selection to natural populations and quantitative traits with a complex genetic architecture, respectively. In this study the usefulness of GWAM and GP were investigated in two different species composed of elite breeding material and genetic resources. In the following sections the general findings of the publications included in this thesis and the used methods are discussed.

Data generation

A high data quality is of crucial importance for all analysis. The phenotypic data quality influences all subsequent analysis, thus a carefully planned field design and precise phenotyping are of utmost importance. Heritability is an often used quality measure of field trial data (Piehpho and Möhring 2007) as it describes the proportion of phenotypic variance to genotypic variance. Broad sense heritability (H^2) is the proportion of the phenotypic variance on the sum of the additive, dominance and epistatic effects, whereas narrow sense heritability (h^2) relates the phenotypic variance to the additive effect (Falconer and Mackay 1996). Phenotypic data quality covered a wide range over all experiments included in this thesis and ranged from 0.05 to 0.99. Two traits exhibited low heritabilities in this thesis. They were all observed in the cauliflower experiment (Thorwarth et al. 2017a) and reflect the difficulties of taking phenotypic measurement of traits such as e.g. the nearest branch length and the apical length. Traits with such low heritability should be excluded from the data analysis as the phenotypic data is not robust enough for subsequent analysis. In the cauliflower paper we investigated the genomic prediction ability of traits that only have a small heritability, thus we did not exclude such traits from our analysis. Otherwise phenotypic data qualities are moderate to high, reflecting a good data quality for the investigation of subsequent analysis.

Influence of population size on Genome-wide association mapping and Genomic prediction

The sample size under study has a strong influence on the power of GWAM (Gawenda and Thorwarth et al. 2015) and the prediction ability of genomic prediction (Thorwarth et al. 2017). Many studies show that relatively large sample sizes for GWA studies are required to be able to detect causal variants with small effects. In our study (Gawenda and Thorwarth at al. 2015) we could show based on simulations, that sample sizes of at least 500 genotypes are necessary for detecting small effects that explain less than 10% of the phenotypic variation in a diverse barley population. For the detection of QTLs explaining a large proportion of the phenotypic variance about 100 individuals are enough (Atwell et al. 2010). Even though it is of theoretical interest to fine-map even small effects, the introgression of them would remain challenging. It is already the case that, the introgression of QTLs with large effects using marker assisted backcross selection is a difficult, time consuming and costly task (Lande and Thompson 1990, Ribaut and Ragot 2007). It would need at least six generations to recover 99.2% of the recurrent parent. The detection of QTLs and the development of methods is an important research field, but from a practical point of view the usability of results of

GWAS are often limited (Bernardo 2008). Also for genomic prediction the size of the training population plays an important role. We assessed the effect of increasing the population size on the prediction ability (Schmid and Thorwarth 2014, Thorwarth et al. 2017) and observed an increase of the prediction ability in for all traits with an increased training set size. In dependence of the trait, prediction accuracy increased by 165% to 235% by increasing the training set size from 25 individuals to 750 individuals (Thorwarth et al. 2017). This is in accordance with results of Zong et al. (2009) who showed that the prediction accuracy depends on the population size and on the presence of QTLs. The authors observed a stronger increase in prediction accuracy when the prediction accuracy was mainly based on the relatedness of the individuals. Thus, an increase of the population size is especially beneficial if not enough markers are present to capture all QTL effects segregating in the population but are enough for inference of the relatedness of the individuals. This was also demonstrated by a study of Hayes et al. (2009) who showed, that the prediction accuracy strongly depends on the effective population sizes (N_e) and that large N_e requires larger training populations to achieve high prediction accuracies and also a higher marker density to capture all segregating QTL effects.

Influence of population structure and relatedness on Genomic prediction

The confounding effect of population structure in GWA studies is well studied (Gawenda and Thorwarth et al. 2015) and has also a strong influence on genomic prediction (Thorwarth et al. 2017, Thorwarth et al. 2017a). In our studies we could show, that the presence of population structure can strongly influence the prediction ability in barley and cauliflower (Thorwarth et al. 2017, Thorwarth et al. 2017a) and that the relationship of genotypes is the main driver of the prediction ability at the current marker density in barley. Thus, genomic prediction works best within populations of similar relatedness while being less effective between populations that share few related individuals or are diverged for only a few generations. Whereas, the prediction of unrelated individuals, for which the LD level differs or the linkage phase is not the same, is at the current marker density not recommendable in barley. Similar results were obtained by Ross et al. (2009), Windhausen et al. (2012) and Guo et al. (2014). They could show that prediction accuracy is highest if training and validation population are built up of related individuals and decreases if the training and validation set are unrelated and marker density is not high enough. These results are in line with our observations but in different species such as maize, rice and cattle.

Several methods for correcting for population structure have been tested, such as inferring clusters based on pedigree or molecular markers, including the highest principle components as fixed effects in a linear mixed model, or methods that perform a reparametrization of the genomic prediction model (Windhausen et al. 2012, Janss et al. (2012), Guo et al. 2014). Recent methods for correcting for population structure in the framework of genomic prediction rely on the identification of clusters using principle component analysis (PCA) or eigenvalue decomposition of the relationship matrix to obtain eigenvectors. The eigenvectors explaining most of the variation present in the sample are then included as fixed effects in a linear mixed model (Yang et al. 2010), which was shown to be disadvantageous if also a relationship matrix is included as random effect, as double-counting occurs (Janss et. al 2012).

Alternatively, all eigenvectors are used as random effects in a linear mixed model to obtain regression coefficients of the principle components. The first regression coefficients, which represent the effect of the top eigenvectors, are set to zero. All regression coefficients are then used for weighting the eigenvalue decomposition of the marker matrix in the marker effect estimation step of the standard genomic prediction model. This leads to a correction for population structure, which overcomes the limitations of including the top eigenvectors as fixed effects (Janss et al. 2012, Guo et al. 2014).

One problem that can occur with the inclusion of all genotypes in a PCA are "artifactual principle components" (Conomos et al. 2015), which are created due recent genetic relatedness and confound the principle components. This is a situation that frequently occurs in breeding populations as can be seen in Thorwarth et al. (2017). Thus, the clear separation of ancestral genetic relatedness, which creates population structure, from recent genetic relatedness is required (Thorwarth et al. 2017, Thorwarth et al. 2017a).

In our studies of the influence of population structure, relatedness and LD on genomic prediction ability (Thorwarth et al. 2017, Thorwarth et al. 2017a) we applied a method suggested by Conoms et al. (2016). It relies on the identification of population structure based on a principle component analysis. A sample of the whole population, consisting of unrelated individuals that represent the ancestral genetic relatedness, is used to calculate a weighted relationship matrix based on "individual-specific allele frequencies" (Conoms et al. 2016). This leads to a direct correction of the genetic relationship matrix for population structure, which can be integrated into a GBLUP model (Thorwarth et al. 2017, Thorwarth et al. 2017a) to obtain prediction abilities that are not influenced by population structure.

Comparison of methods and models for Genome-wide association mapping and Genomic prediction

Single-marker versus haplotybe based methods in Genome-wide association mapping

An important research field is the development of improved models for GWAM and GP. Many different GWAM methods exist. For example multivariate analysis where several correlated traits, that would normally be evaluated in an univariate analysis, are analysed together or haplotype based methods where, in comparison to single-marker analysis, multi-allelic markers are used to perform the GWAM. Further, methods that account in different ways for population structure are used. They differ in their power and efficiency. In this thesis we focused on the comparison of single marker based methods with haplotype based methods and the comparison of the effect of a correction for population structure in the single marker case. Therefore, we conducted a simulation study to validate our results based on empirical data (Gawenda and Thorwarth at al. 2015). In our study we could demonstrate an increase in power due to the use of haplotype based methods for GWAM, the advantage of correcting for population structure in the single marker analysis and provide an estimate of the required population sizes for detecting QTLs explaining less than 10% of the phenotypic variation. The power of haplotype based methods was also demonstrated in other publications, especially if a correction for population structure is applied (Lorenz et al. 2010, N'Diaye et al.2017). In our study we could demonstrate that haplotype based methods have a higher power of detecting QTLs with very small effects in comparison to single-marker analysis, but that larger populations are required increasing the computational burden and decreasing the efficiency of haplotype based methods (Gawenda and Thorwarth at al. 2015).

The crux of cross-validation in Genomic prediction

In genomic prediction four main categories of models exist: parametric, semi-parametric and non-parametric models and their Bayesian interpretations (Pérez-Rodríguez et al. 2012). In Thorwarth et al. (2017) we compared 11 models belonging to either of these categories and could not find large differences in the model performance using a 10 fold cross-validation (CV) with 5 replications. In a CV scheme with predefined folds we observed differences in the predictive performance of the different models and the Elastic Net (EN) outperformed the other methods for all traits, indicating a model - training set - trait interaction. While the standard CV provides a good measure of the average performance of models, the obtained value of the prediction accuracy is not necessarily representative for the true prediction of unobserved genotypes in a breeding program. In plant breeding programs crosses between individuals are performed that differ in their degree of relatedness, resulting into complex family structures (Thorwarth et al. 2017, Yu et al. 2006), thus the underlying assumption of independence between training set and validation set does not hold, which is a likely explanation of the observed similarity in predictive performance of the tested models (Pérez-Cabal et al. 2012).

It is difficult to create independent data sets using plant breeding populations. In an attempt to do so, we used the material of a specific breeding company, which is a subset of the whole data set, and used it for the prediction of the remaining sets (Thorwarth et al. 2017). Independence between data sets cannot be achieved in this way as often genotypes among breeding companies share a common ancestry. A better approach would be to use the inferred clusters from the DAPC, which groups closely related individuals together with minimal variation within the group but maximized variation between groups. Prediction ability decreased on average strongly by using the material of a specific breeding company as training set, especially after correcting for population structure. For specific training set - validation set combinations an increase in the prediction ability was observed and the predictive performance of the single models often differed, too. Similar observations were made by e.g. Heslot et al. (2012) and Windhausen et al. (2012). Differences in the genetic architecture are one explanation for the differences in the prediction abilities for the respective traits and are also an explanation for the observed differences in the prediction ability of specific training set - validation set combinations (Heslot et al. 2012). We observed moderate to high levels of genetic differentiation (F_{ST}) between the whole training population and the offspring population, which contains 33 families derived from specific crosses of parental lines from the training population. Correlation between F_{ST} values and prediction accuracy were mostly negative and not significant, which hinders a strong conclusion, but indicates that the relationship between the respective training set - validation set combination has an influence. We often observed a higher prediction accuracy of a family that has a great genetic distance to the training set compared to another family. One explanation is that similar selection pressure applied in the different breeding programs leads to similar changes in key genomic regions resulting in a stronger increase in the prediction accuracy than expected based on the genetic differentiation. The reason for this observation is that even for polygenic traits some QTLs that explain a larger proportion of the phenotypic variance are expected. Those can account for the higher prediction accuracy, whereas F_{ST} is always based on the whole genome as was demonstrated by Scutari et al. (2016).

Difference in Genomic prediction models and further factors influencing prediction ability

Several methods for the purpose of genomic prediction have been developed (Heslot et al. 2012, de los Campos et al. 2013), which all try to solve the small n large p problem, a situation in which less observations (n) than predictors (p) are available ($n << p$), making a model underdetermined. This is a situation commonly encountered nowadays in the fields of plant and animal breeding, where hundreds of individuals are genotyped with thousands of SNPs. This, raises the problem of multicollinearity due to LD between markers (Ogutu and Piepho 2014). In this thesis the exact reason for the observed differences of the prediction ability of the models could not be analysed. Due to the complexity of the problem simulation studies would be necessary, as they are the only possibility to get control of the influential parameters. But even if simulation studies are used a clear separation of the influential parameters is not always possible (Heslot et al. 2012). In the following section differences in the models and their possible influence on the observed training set - validation set - trait interactions, are discussed and related to the results obtained in this thesis.

Regularization and the bias-variance tradeoff

Regularization is a commonly applied solution to the problem of multicollinearity. By extending the ordinary least square solution of the regression coefficients $\hat{\beta}$ by a penalty term that induces regularization, a solution to an ill-posed model can be obtained. Ridge Regression (RR) takes advantage of regularization by shrinking the regression coefficients towards zero. Following Hastie et al. (2009), the notation of the RR model is:

$$\hat{\boldsymbol{\beta}}^{\lambda} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_{\ell 2}^2 + \lambda||\boldsymbol{\beta}||_{\ell 2}^2, \tag{3}$$

where $\lambda||\boldsymbol{\beta}||_{\ell 2}^2$ represents the penalty term and λ is the shrinkage parameter, where larger values of λ lead to a greater shrinkage (Hastie et al. 2009). By adding a small constant to the diagonal of the X^TX matrix of the matrix notation of the regression solution $\hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty$, the problem of singularity in matrix inversion is circumvented. Ridge Regression was not directly tested in this thesis, but the Elastic Net incorporates the Ridge Regression in its model formulation and can be obtained by setting $\alpha = 0$ in equation 7 of Thorwarth et al. 2017. By performing cross validation for a sequence of α-values it is possible to obtain a specific amount of shrinkage and variable selection, which minimizes the mean squared error. For most of the tested models (Thorwarth et al. 2017; Supplementary Tables) the smallest mean squared error was obtained for $\alpha = 0$ (data not shown), thus resulting into the Ridge Regression interpretation of the Elastic Net.

The quality of model can be assessed by observing the means square error (MSE, Hastie et al. 2009). The MSE is defined as:

$$\begin{aligned} Err(X_0) &= E[(Y - \hat{f}(x_0))^2|X = x_0] \\ &= \sigma_E^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\ &= \sigma_E^2 + \text{ Bias}^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \\ &= Irreducible\ Error + Bias^2 + Variance \end{aligned}$$

The irreducible error term σ_E^2 is the variance of the given observations around their true mean $f(x_0)$ and represents a fixed parameter, which can not be reduced for a given data set. The squared bias term $[E\hat{f}(x_0) - f(x_0)]^2$ describes the deviation of the expected values estimator from the true mean and the variance term $E[\hat{f}(x_0) - E\hat{f}(x_0)]^2$ describes the expected squared deviation of $\hat{f}(x_0)$ around the true mean. Overall we look for a model with a good fit, that has a balanced set up between bias and variance while minimizing the generalization error for a given model. To find the right trade-off between bias and variance is a common problem in statistics (Hastie et al. 2009) as it is closely related to the concept of underfitting due to a to high bias and overfitting due to high variance. Both situations lead to a drop in prediction ability of unseen data, thus the right balance between both values and a minimization of the mean square error is required. In Ridge Regression, the multiplication of λ with the l_2-norm increases the bias of the model but decreases the variance and thus can lead to a reduction of the mean squared error and a more stable solution in the case of $n << p$ (de los Campos et al. 2013). The right choice of the regularization parameter can have a strong influence on the predictive performance (Wimmer et al. 2013) and often cross validation is used to find the λ value that minimizes the mean squared error (Hastie et al. 2009). In the genomic prediction literature, the BLUP interpretation of the RR approach (RRBLUP) is frequently used (Piepho 2009). Here $\lambda = \frac{\sigma_\epsilon}{\sigma_G}$ and the variance components are either based on previous knowledge or based on e.g. restricted maximum likelihood estimates. If the λ-value for the RR is the same as the one derived based on the variance components then both models are equivalent (Vlaming and Groenen 2014). Further, the estimation of the variance components based on one REML fit can be inaccurate, especially if the population size is small (Mrode et al. 2014, page 183) or in the presence of population structure, where the variance component estimates within a subpopulation can differ from the estimates of the whole population (Guo et al. 2014). This is important to note as this kind of parametrization based on the signal-to-noise ratio is also used for other methods such as Bayesian Ridge Regression (BRR) and the Bayesian Least Absolute Shrinkage and Selection Operator (BL, Thorwarth et al. 2017), but with the differences that variance components are obtained from a posterior distribution based on draws from the prior distribution with defined hyperparameters that control the shape of the distribution, given a specific data sample. These descriptions of the differences in the choice of λ are one explanation for the observed differences in the model performance of e.g. the RRBLUP, the BRR and EN model with $\alpha = 0$ in Thorwarth et al. (2017).

Advantages and pitfalls of variable selection

A further, method to cope with the problem of multicollinearity or $n << p$ is variable selection. In this thesis, we compared two methods that can perform stringent variable selection by setting non-influential coefficients exactly to zero, namely the Least Absolute Shrinkage and Selection Operator (LASSO) and the Elastic Net (Thorwarth et al. 2017). There are several reasons why variable selection methods can be advantageous. A reduced set of variables is following Ockham's principle of parsimony always preferable and can support the biological interpretation of the predictors retained in the model. Collinearity among predictor variables is removed and the reduction of noise can increase the future prediction accuracy, by decreasing prediction variance at the cost of an increased bias (Ding and Peng 2003, Hastie et al. 2009, Wimmer et al. 2013). The computational cost are reduced with a sparse model, especially if several models have to be tested. Additionally, a decrease in costs for genotyping can be obtained with a reduced set of markers and the selected variables could be used as a basis to construct e.g. a SNP array. The problem of variable selection is a well known statistical problem (O'Hara and Silanpää 2009), for which a variety of possible solution exists (Fan and Jinchi 2010).

The LASSO (Tibshirani 1996, Thorwarth et al. 2017) is a method which is often used to handle high-dimensional data. It produces a sparse solution by setting some of the regression coefficients to zero, while shrinking the remaining coefficients towards zero. Drawbacks of the LASSO are its incapability of handling correlated predictors as it selects one out of them and sets the remaining ones to zero. Further, it lacks the oracle property and can not select more variables than the sample size before it saturates when $n << p$ (Zou et al. 2005, Friedman et al. 2010). To overcome the limitations of the LASSO, (Zou et al. 2005) developed the EN. Due to the combined use of the ℓ_1 and ℓ_2 norms, the EN performs variable selection due to the ℓ_1 penalty and overcomes the problem of correlated predictors encountered in the LASSO by using the ℓ_2 penalty to shrink correlated predictors towards each other. As a result a group of correlated predictors is jointly in- or excluded (Friedman et al. 2010). Additionally the EN can produce a sparse solution with more than n predictors in $n << p$ scenarios.

In Thorwarth et al. (2017) we show, that the prediction accuracy between the two models differ strongly. In the ranking of the best model (Thorwarth et al. 2017, Table S11) the Elastic Net is the best performing method, over a wide range of training set - validation set - trait combinations, whereas the LASSO performed worst. The main reason for this strong difference is the way how stringent the two methods select variables. The LASSO selects at most n variables before it saturates (Zou et al. 2005). Thus, in some scenarios in which the training set size was very small the retained number of markers was very low, leading to a strong drop of prediction accuracy in comparison to methods that do not perform variable selection, or select variables in a less stringent way. This observation is supported by our general finding in Thorwarth et al. (2017), that the current marker density used in barley breeding is not high enough, as about 11,203 markers would be necessary to predict even genetically distant populations.

Bayesian models

Bayesian models used for genomic prediction (Meuwissen et al. 2001) are built up around *a priori* assumptions about the distribution of marker effects, their variances and the probability of a marker having zero variance, or not (Gianola et al. 2009). The necessary information (hyperparameters) to set up the respective distributions are based on theoretical assumption about the biological distributions of marker effects and QTLs, for which no clear evidence exists (Brem and Kruglyak 2005). Also the definition of genetic complexity is not consistent in literature (Thompson and Galitski 2012) and the assumptions about the complexity of continuous trait seems to be to simplistic (Goddard et al. 2016), which makes a proper specification of the prior information difficult. As shown by Gianola et al.(2009) Bayesian methods that assume marker-specific variances do not allow Bayesian learning, or only to a very small extent. Thus, the prior information often dominates the model and cannot be overcome by the data, due to e.g. limited sample size used for model training or model specifications, which hinder a move away from the prior (Gianola et al. 2009). This is one reason for the often observed similarity between models in this thesis and in the genomic prediction literature in general. In a comparative study Heslot et al. (2012) showed that Bayesian methods are either prone to overfitting or very similar in their accuracy to simpler models such as RRBLUP, but with a much higher computational time and model complexity. In our study, on average, similar prediction accuracies for all models were observed. One explanation for this can be the usage of default hyperparameters, which were set according to default specification of Pérez et al. (2014) leading to uniformative priors. The combination of the small subpopulation sizes (Thorwarth et al. 2017) and the limited amount of Bayesian learning due to wrong prior specification (Gianola et al. 2009), are one explanations of the differences observed in the Bayesian prediction models in this thesis, similar to Heslot et al. (2012).

Conclusion

In this thesis we could demonstrate the feasibility and limitations of genome-wide association mapping and genomic prediction in self-fertilizing barley and outcrossing cauliflower populations, persisting of either elite material or a mixture of elite material and genetic resources.

In particular we could show, the strength of haplotype based methods for GWAM and the importance of correcting for population structure and cryptic relatedness in the single-marker and haplotype-based case. The size of the population under study should be at least 500 individuals to achieve a high enough power to reliable detect significant associations, whereas the number of markers is less important if the marker density is high enough to cover the whole genome, due to a high extend of linkage disequilibrium.

Genomic prediction provides in theory a valuable tool for estimating the genomic breeding value of genotypes. In our studies we obtained moderate to high cross-validation prediction abilities, but those dropped rapidly if small families or subpopulations were used to predict another family or subpopulation. This indicates that the main driver of the prediction ability is the relatedness of individuals. Further, we could demonstrate a strong influential effect of the population structure on the prediction ability in our studies, which has to be taken into account, especially in breeding populations with a complex population structure. Linkage phase was not consistent among barley families, indicating the necessity of much higher marker densities than expected from linkage disequilibrium analysis.

The comparison of GP methods demonstrated, based on cross-validation, only minor differences in the model performance. If small subpopulations or families are used as training set to predict other families or subpopulations, differences in the performance of the models could be observed. Based on our results we can not recommend the use of a single model, but rather suggested the usage of several methods on a routine basis. One method that seems to be particularly interesting is the Elastic Net, which can adapt relatively flexible to a given data set, further a method that directly calculates the genomic estimated breeding values such as GBLUP, a non parametric method such as Random Forest and a Bayesian method such as e.g. BayesC should be included in all analysis. Further, it is important to not only compare the prediction ability but also e.g. the mean squared error and the choice of regularization parameter. Methods that perform strong variable selection such as the LASSO can not be recommended for genomic prediction at the current marker density particularly in barley. Such methods produce a sparse model, which can lead to a drop in prediction ability, especially if small populations are used as training set.

The method of genomic prediction was hyped in the last years, but a strong proof of concept in plant breeding is in my opinion still missing. One reason for it is the complexity of plant breeding and the difficulty to transfer this complexity into appropriate statistical models. In this work we could demonstrate that genome-wide association mapping and genomic prediction can be powerful tools to support breeding decisions but that a model training based on some hundred individuals assessed in a limit amount of locations and years is currently not enough to completely replace the breeder's eye.

"Gedanken ohne Inhalt sind leer, Anschauungen ohne Begriffe sind blind" (Immanuel Kant)

References

Atwell, S., Huang, Y.S., Vilhjlmsson, B.J., Willems, G., Horton, M., Li, D.M., Platt, A., Tarone, A.M., Hu, T.T., Jiang, R., Wayan, N., et al. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–631

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

Brem, R.B., Kruglyak, L. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. PNAS 102:1572–1577

Conomos, M.P., Miller, M.B., Thornton, T.A. 2015. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. Genetic Epidemiology 39:276–293

Conomos, M.P., Reiner, A.P., Weir, B.S., Thornton, T.A. 2016. Model-free Estimation of Recent Genetic Relatedness. The American Journal of Human Genetics 98:127–148

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

de Roos, A.P.W., Hayes B.J., Goddard, M.E. 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183:1545–1553

Ding, C., Peng, H. 2003. Minimum redundancy feature selection from microarray gene expression data. Computational Systems Bioinformatics 3:185–205

Falconer, D.S., Mackay, T.F.C. 1996. An Introduction to Quantitative Genetics. Ed. 4. Prentice Hall, London.

Fan, J., Jinchi, L. 2010. A Selective Overview of Variable Selection in High Dimensional Feature Space. Statistics Sinicia 20: 101–148

Friedman, J., Hastie, T., Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software 33:1–22

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., Fernadno, R. 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183:347–363

Gawenda, I., Thorwarth, P., Günther, T., Ordon, F., Schmid, K.J. 2015. Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. Plant Breeding 134:28–39

Goddard, M.E., Kemper, K.E., MacLeod, I.M., Chamberlain, A.J., Hayes, B.J. 2016. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. Proc. R. Soc. B 283:1–12

Guo, Z., Tucker, D.M., Basten, C.J., Gandhi, H., Ersoz, E., Guo, B.,Xu, Z., Wang, D., Gay, G. 2014. The impact of population structure on genomic prediction in stratified populations. Theoretical and Applied Genetics 127:749–762

Hastie, T., Tibshirani, R., Friedman, J. 2009. The Elements of Statistical Learning. Springer Series in Statistics

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

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

Janss, L., de los Campos, G., Sheehan, N., Sorensen, D. 2012. Inferences from Genomic Models in Stratified Populations. Genetics 192:693–704

Lande, R., Thompson, R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics:124743–756

Lorenz, A.J., Hamblin, M.T., Jannink, J.L. 2010. Performance of Single Nucleotide Polymorphisms versus Haplotypes for Genome- Wide Association Analysis in Barley. PLoS ONE 5:1–11

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

Mrode, R.A., 2014. Linear models for the prediction of animal breeding values. 3rd ed. CABI

N'Diaye, A., Haile, J.K, Cory, T.A., Clarke, F.R., Clarke, J.M., Knox, R.E., Pozniak, C.J. 2017. Single Marker and Haplotype- Based Association Analysis of Semolina and Pasta Colour in Elite Durum Wheat Breeding Lines Using a High-Density Consensus Map. PLoS ONE 12:1–24

Ogutu, J.O., Piepho, H.P., Regularized group regression methods for genomic prediction: Bridge, MCP, SCAD, group bridge, group lasso, sparse group lasso, group MCP and group SCAD. BMC Proc 8:1–9

O'Hara, R.B., Sillanpää, M.J. 2009. A review of bayesian variable selection methods: What, how and which. Bayesian Analysis 4:85–118

Pérez-Cabal, M.A., Vazquez, A.I., Gianola, D., Rosa, G.J.M., Weigel, K.A. 2012. Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. Front. Genet. 27:1–7

Pérez-RodrguezP., Gianola, D., Gonzlez-Camacho, J.M., Crossa, J., Mans, Y., Dreisigacker, S. 2012. Comparison Between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat. G3: Genes, Genomes, Genetics 2:1595–1605

Pérez, P., de los Campos, G. 2014. Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198:483–495

Piepho, H.P., Möhring, J. 2007. Computing Heritability and Selection Response From Unbalanced Plant Breeding Trials. Genetics 177:1881–1888.

Piepho, H.P. 2009. Ridge Regression and Extensions for Genomewide Selection in Maize. Crop Science 49:1165–1176

Ribaut, J.-M., Ragot, M. 2007. Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. Journal of Experimental Botany 58:251-360

Scutari, M., Mackay, I., Balding, D. 2016. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. PLoS Genetics 12:1–19

Schmid, K.J., Thorwarth, P. 2014. Genomic Selection in Barley Breeding. In: Kumlehn, J., Stein, N. (eds) Biotechnological Approaches to Barley

Improvement. Biotechnology in Agriculture and Forestry. Springer, Berlin, Heidelberg 69:367–378

Thompson E.G., Galitski, T. 2012. Quantifying and analyzing the network basis of genetic complexity. PLoS Comp. B. 8:1–10

Thorwarth, P., Ahlemayer, J., Bochard, A.-M., Krumnacker, K., Blümel, H., Laubach, E., Knöchel, N., Cselényi, L., Ordon, F., Schmid, K.J. 2017. Genomic prediction ability for yieldrelated traits in German winter barley elite material. Theor Appl Genet 130:1669–1683

Thorwarth, P. Yousef, E.A.A, Schmid, K.J. 2017a. Genomic prediction and association mapping of curd-related traits in genebank accessions of cauliflower. Submitted to G3: Genes, Genomes, Genetics

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society 58:267–288

de Vlaming, R., Groenen, P.J.F., The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. BioMed Research International 2015:1–18

Wimmer,V., Lehermeier, C., Albrecht, T., Auinger, H.-J., Wang, Y., Schn, C.-C. 2013. Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection. Genetics 195:573–578

Windhausen, V.S., Atlin, G.N., Crossa, J., Hickey, J.M., Grudloyma, P., Terekegne, A. et al. 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3:Genes,Genomes,Genet 2:1427–1436

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203–208

Zhong, S., Dekkers, J.C.M., Fernando, R.L., 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

Zou, H., Hastie, T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society 67:301–320

7. Summary

Due to the advent of new sequencing technologies and high-throughput phenotyping an almost unlimited amount of data is available. In combination with statistical methods such as Genome-wide association mapping (GWAM) and Genomic prediction (GP), these information can provide valuable insight into the genetic potential of individuals and support selection and crossing decisions in a breeding program. In this thesis we focused on the evaluation of the aforementioned methods in diverse barley (*Hordeum vulgare L.*) and cauliflower (*Brassica oleracea var. botrytis*) populations consisting of elite material and genetic resources. We concentrated on the dissection of the influence of specific parameters such as marker type, statistical models, influence of population structure and kinship, on the performance of GWAM and GP. For parts of this thesis, we additionally used simulated data to support findings based on empirical data.

First, we compared four different GWAM methods that either use single-marker or haplotypes for the detection of quantitative trait loci in a barley population. To find out the required population size and marker density to detect QTLs of varying effect size, we performed a simulation study based on parameter estimates of the empirical population. We could demonstrate that already in small populations of about 100 individuals, QTLs with a large effect can be detected and that at least 500 individuals are necessary to detect QTLs with an effect $< 10\%$. Furthermore, we demonstrated an increased power of haplotpye based methods in the detection of very small QTLs.

In a second study we used a barley population consisting of 750 individuals as training set to compare different GP models, that are currently used by scientists and plant breeders. From the training set 33 offspring families were derived with a total of 750 individuals. This enabled us to assess the prediction ability not only based on cross-validation but also in a large offspring population with varying degree of relatedness to the training population. We investigated the effects of linkage disequilibrium and linkage phase, population structure and relatedness of individuals, on the prediction ability. We could demonstrate a strong inflating effect of the population structure on the prediction ability and show that about 11,203 evenly spaced SNP markers are necessary to predict even genetically distant populations. This implies that at the current marker density prediction ability is based on the relatedness of the individuals.

In a third study we focused on the evaluation of GWAM and GP in cauliflower. We focused on the evaluation of genotyping-by-sequencing and compared the influence of imputation methods on the prediction ability and the number of significant associations. We obtained a total 120,693 SNPs in a random collection of 174 cauliflower genebank accessions. We demonstrated that imputation did not increase prediction ability and that the number of detected QTLs only slightly differed between the imputed and the unimputed data set. GP performed well even in such a diverse gene bank sample, but population structure again inflated the prediction ability.

We could demonstrate the usefulness and limitations of Genome-wide association mapping and genomic prediction in two species. Even though a lot of research in the field of statistical genetics has provided valuable insight, the usage of Genomic prediction should still be applied with care and only as a supporting tool for classical breeding methods

8. Zusammenfassung

Durch die Einführung neuer Sequenzierungstechnologien, welche tausende genetische Marker verfügbar macht und die Hochdurchsatz-Phänotypisierung, steht eine beinahe unbegrenzte Anzahl an Daten zur Verfügung. In Verbindung mit statistischen Methoden wie der Genomweiten Assoziationskartierung (GWA) und der Genomischen Vorhersage (GP), können nützliche Erkenntnisse über das genetische Potential von Individuen erhalten werden. In dieser Doktorarbeit haben wir uns auf die Bewertung dieser Methoden in diversen Gerste (*Hordeum vulgare L.*) und Blumenkohl (*Brassica oleracea var. botrytis*) Populationen, bestehend aus Elitematerial und genetischen Ressourcen, fokussiert. Wir analysierten den Einfluss verschiedener Parameter auf die Ergebnisse der GWA und der GP. Für Teile der Doktorarbeit verwendeten wir simulierte Daten, um unsere Forschungsergebnisse zu unterstützen.

Zuerst verglichen wir vier Methoden zur GWA. Diese verwenden einzelne Marker oder Haplotypen für die Detektion von möglichen Regionen eines quantitativen Merkmals (QTL). Um die Populationsgröße und Markerdichte herauszufinden, welche notwendig ist um QTL mit unterschiedlicher Effektstärke zu entdecken, wurde eine Simulationsstudie verwendet, die auf Parameterschätzungen der empirischen Daten einer Gerstenpopulation beruht. Wir wiesen nach, dass Populationen von 100 Individuen ausreichen um QTLs mit einem großen Effekt zu entdecken und dass mindestens 500 Individuen notwendig sind, um QTLs mit einem Effekt von $< 10\%$

aufzuzeigen. Des Weiteren, zeigten wir, dass eine Erhöhung der Teststärke durch die Verwendung von haplotyp-basierten Methoden zur Detektion von QTLs erreicht werden kann.

In einer zweiten Studie verwendeten wir eine Gerstenpopulation bestehend aus 750 Individuen als Trainingset um verschiedene Methoden zur GP zu vergleichen. Auf Basis der Trainingspopulation wurden 33 Familien entwickelt, die insgesamt aus 750 Individuen bestehen. Dies ermöglichte es uns, die Vorhersagegenauigkeit nicht nur auf Basis von Kreuzvalidierung zu bestimmen, sondern ebenfalls in einer großen Nachkommenspopulation mit unterschiedlichem Verwandtschaftsgrad zum Trainingset. Wir erforschten unter anderem den Einfluss des Kopplungsungleichgewichtes und der Populationsstruktur auf die Vorhersagegenauigkeit. Wir konnten zeigen, dass die Populationsstruktur einen stark überhöhenden Effekt auf die Vorhersagegenauigkeit hat und dass 11,203 SNP Marker notwendig sind, um genetisch entfernte Populationen vorherzusagen.

In einer dritten Studie fokussierten wir uns auf die Evaluierung der GWAM und der GP in Blumenkohl. Hier untersuchten wir den Einflusses von Genotypisierung durch Sequenzierung (GBS) und Methoden zur Imputierung fehlender Werte sowie deren Einfluss auf die Vorhersagegenauigkeit und die Anzahl an signifikanten Assoziationen. Die Verwendung von Imputierungsmethoden führte nicht zu einer Erhöhung der Vorhersagegenauigkeit und die Anzahl der gefundenen QTLs wich nur geringfügig zwischen den imputieren und nicht-imputierten Datensätzen ab. Die GP funktionierte gut in diesem diversen Genbank Material, aber die Populationsstruktur hatte einen stark verzerrenden Einfluss auf die Vorhersagegenauigkeit.

Wir konnten in dieser Doktorarbeit Nutzen und Limitierung der GWA und der GP anhand von Gerste und Blumenkohl aufzeigen. Obwohl die vielen Forschungsbemühungen im Bereich der statistischen Genetik wichtige Erkenntnisse geliefert haben, sollten die hier verwendeten Methoden mit Vorsicht angewendet werden und zur Zeit nur als unterstützende Maßnahme zu klassischen Züchtungsverfahren gesehen werden.

Acknowledgments

My sincere thanks go to my academic supervisor Prof. Dr. Karl J. Schmid for offering me the possibility to join his group to work on the BARSELECT project and on my PhD. Thank you for all the support and freedom you offered me and the many possibilities to present my results on various conferences.

I thank Prof. Dr. Frank Ordon and Prof. Dr. Bennewitz for being part of the graduate committee.

Special thanks I want to address to Fabian Freund. Thank you for your constant support and advice and the many great discussions we had.

Further, I want to thank all my co-authors and project partners of BARSELECT for the good and cooperative work. I thank especially Torsten Günther for his support at the beginning of my PhD thesis .

Thanks to all my former office colleagues Christian Lampei, Thomas Müller, Oliver Simon, Wosene Gebreselassie and Max Haupt for creating a pleasant working atmosphere. Also a lot of thanks to all other colleagues, particularly Ivan Barilar and Markus Stetter. Ivan thanks for all the good discussions, especially the ones not related to work. Markus, thanks a lot for the constant supply of delicious farm products and motivation.

I also would thank Philipp Boeven, Ben Collier and Philipp Thorwarth for proofreading this thesis.

A lot of thanks also to Tobias Würschum and Friedrich Longin for offering me the possibility to use the resources of the State Plant Breeding Institute before and after work to finish my thesis and their general encouragement to finish this work.

Ganz besonderen Dank möchte ich meiner Familie aussprechen. Zuerst möchte ich mich bei meiner Frau Sabina bedanken. Deine bedingungslose und andauernde Unterstützung hat mir die Möglichkeit, den Freiraum und vor allem die Zeit gegeben diese Arbeit fertigzustellen.

Besonderen Dank auch an meine Eltern Gabriele und Rudolf und meinen Bruder Philipp. Danke für all den Rückhalt den ihr mir gebt.

11. Curriculum Vitae

Personal Information:

Name:	Patrick Thorwarth
Family status:	married
Date of birth:	21/12/1984
Place of birth:	Göppingen
Nationality:	German

Academic career:

12/2016 - today	**Research assistant at the State Plant Breeding Institute, University of Hohenheim**
10/2012 - 11/2016	**Phd student in the group Crop Biodiversity and Breeding Informatics, University of Hohenheim**
10/2010 - 08/2012	**M.Sc. in Crop Sciences with specialization in Plant Breeding and Seed Sciences**
10/2006 – 06/2010	**B.Sc. In Agricultural Science with specialization in Economics and Social Sciences**

Professional activities:

04/2011 - 06/2012	**Student assistant in the group Crop Biodiversity and Breeding Informatics, University of Hohenheim**
04/2010 - 07/2010 12/2008 - 09/2009	**Student assistant in the group Agricultural Communication and Consultation Sciences at the University of Hohenheim**
04/2009 - 05/2009	**Student assistant in the group Agricultural Markets and Marketing at the University of Hohenheim**

Other activities:

08/2005 - 08/2006	**Voluntary social service in an Agricultural School in Brasil**

Stuttgart - Hohenheim, 19.09.2017

Statutory declaration
(Eidesstattliche Versicherung)

Bei der eingebrachten Dissertation zum Thema "Evaluation of association mapping and genomic prediction in diverse barley and cauliflower breeding material" handelt es sich um meine eigenständig erbrachte Leistung.

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

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

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

Stuttgart - Hohenheim, 19.09.2017

Patrick Thorwarth