Genomic Selection in Bread Wheat

DISSERTATION

zur Erlangung des Doktorgrades an der
Universität für Bodenkultur, Wien

eingereicht von

Christian Ametz

Betreuer:

Univ.Prof. Dipl.Ing. Dr. Hermann Bürstmayr

A.o. Univ.Prof. Dipl.Ing. Dr. Heinrich Grausgruber

Univ.Prof. Dipl.Ing. Dr. Johann Sölkner

Angefertigt am
Institut für Biotechnologie in der Pflanzenproduktion
Universität für Bodenkultur, Wien

Zusammenfassung

Die Entwicklung robuster, ertragreicher Kulturpflanzen mit breiter Anpassung an variable Umweltbedingungen ist unverzichtbar um ausreichend sichere Nahrung für eine wachsende Weltbevölkerung bereitzustellen. Konventionelle Züchtung hat zu erheblichen Verbesserungen in der Produktivität und Qualität geführt. Um mit der projizierten steigenden Nachfrage in den kommenden Jahrzehnten Schritt halten zu können, ist es unausweichlich, den Fortschritt weiter zu erhöhen, um der Nachfrage der kommenden Jahrzehnte, wenn die Weltbevölkerung voraussichtlich 9 Milliarden erreicht, gerecht zu werden.

Weizen ist die wichtigste Kulturart im Hinblick auf die Weltanbaufläche und die drittgrößte in Bezug auf die Weltproduktion. Ackerland ist jedoch eine begrenzte Ressource und die notwendige Erhöhung der Kulturpflanzenproduktion kann nur zu einem geringen Prozentsatz durch die Kultivierung von neuem Land erreicht werden.

Die jüngsten Fortschritte in der molekularen Markeranalyse und das Aufkommen von Hochdurchsatz-Genotypisierungsplattformen ermöglichen die genomweite Analyse einer Weizenzuchtlinie zu den Kosten einer einzelnen Parzelle in einem herkömmlichen Feldversuch. Genomische Selektion, ein neues Paradigma das zuerst in der Tierzucht eingeführt wurde, nutzt diese dichten Markerinformationen und kombiniert sie mit phänotypischen Informationen, um den Zuchtwert von neuen Kandidatenlinien vorherzusagen. Selektion auf Basis dieser Zuchtwerte hat die Hoffnung geweckt, die Rate des Zuchtfortschritts in der Tier- und Pflanzenzucht deutlich zu erhöhen.

Die vorliegende Arbeit stellt eine der ersten Anwendungen von genomischer Selektion in einem realen Weizenzüchtungsprogramm dar. Die neuesten Technologien zur Genotypisierung wurden für den Einsatz in der genomischen Selektion ausgewertet. Mittels genotyping-by-sequencing konnten Elite-Weizenzuchtlinien auf Basis von Einzelbasenunterschieden charakterisiert und genomische Fingerabdrücke erstellt werden. Diese Fingerabdrücke wurden mit herkömmlichen phänotypischen Informationen kombiniert und statistische Vorhersagemodelle erstellt. Basierend auf diesen Modellen war es möglich, genomische Zuchtwerte für ungetestete Weizenlinien ausschließlich anhand ihres genomischen Fingerabdrucks vorherzusagen. Verschiedene statistische Modelle wurden auf ihre Fähigkeit zur Vorhersage von Zuchtwerten und ihre Annahmen über die genetische Varianz verglichen. Bei diesen Vergleichen stellte sich RR-BLUP als stabile und leistungs-

fähige Methode zur Vorhersage heraus, die vor allem durch kurze Berechnungszeiten besticht.

Zudem wurden in dieser Arbeit weitere Faktoren, die die genomische Vorhersage beeinflussen können, untersucht. Eine Trainingspopulationsgröße von 250 Weizenlinien war bereits ausreichend um gute Vorhersageergebnisse zu erzielen, wohingegen der Einsatz von mehr als 3000 Markern empfehlenswert erscheint.

In einem abschließenden Experiment, das im Zuge eines laufenden Weizenzüchtungsprogramms durchgeführt wurde, konnte gezeigt werden, dass Zuchtwerte aufgrund eines einzelnen genetischen Fingerabdrucks ebenbürtig mit den Zuchtwerten aus konventionellen Feldversuchen sind. Die Ergebnisse lassen darauf schließen, dass Genomische Selektion in der Pflanzenzüchtung ebenso erfolgreich sein kann, wie sie in der Tierzucht bereits ist.

Abstract

The development of more robust, productive crops with broad adaptation to variable environmental conditions is indispensable for providing safe and sufficient food to a growing world population. Genetic improvement by conventional breeding has been leading to substantial enhancements in productivity and quality. Yet to keep up with the projected increasing demand in the coming decades further increases seem inevitable to meet the demand of the coming decades when the global population is expected to reach 9 billion.

Wheat is the most important crop plant species in terms of world acreage and the third largest in terms of produced grain. But arable land is a limited resource and without doubt the needed increases in crop plant production can only to a small percentage come from cultivating new land.

Recent advances in molecular marker analysis and the advent of high-throughput genotyping platforms now enable genome-wide marker analysis of one breeding line to be carried out at the cost of one plot in a conventional field trial. Genomic selection, a new paradigm first introduced in animal breeding makes use of this dense marker information and integrates it with phenotypic information to accurately predict the breeding value of new candidate lines. Selection based on those breeding values has raised hopes to substantially increase the rate of genetic gain in animal and plant breeding.

The work at hand represents one of the first applications of genomic selection in an actual plant breeding program. The latest genotyping technologies were evaluated for their use in genomic selection. Using genotyping-by-sequencing it was possible to characterize elite wheat breeding lines based on single base differences to create genomic fingerprints. These fingerprints were integrated with conventional phenotypic information to create statistical prediction models. Based on these models it was possible to predict genomic breeding values for untested wheat lines solely based on their genomic fingerprint. Different statistical models were compared for their predictive ability and their assumptions on genetic variance. Using RR-BLUP, a statistical method for prediction, it was possible to create stable and well performing models with the advantage of short computation times.

Moreover, other factors that may influence the genomic predictions were investigated in this work. A training population size of 250 wheat lines was already sufficient for achieving accurate prediction results whereas the use of at least 3000 markers seems advisable.

Finally, in experiment within an actual plant breeding setting, it could be shown that breeding values based on a single genomic fingerprint were on par with breeding values obtained by conventional field trials. The results of this work indicate that genomic selection can be as successful in plant breeding as it is already in animal breeding.

Contents

Part I

Introduction

1 Project *Genowheat*: Genomic selection in bread wheat

The present work was performed as part of the co-operational project, *Genowheat,* run in close collaboration of several partners, which are part of the plant breeding and seed industry, in the the European Union *EUROSTARS* initiative.

The aim was to evaluate and implement models for predicting the breeding performance of elite winter wheat lines using genomic selection (GS), a new paradigm in plant breeding. Different traits, from traditional grain yield to resistance traits like rust resistance should be assessed for their utilization in genomic selection. Extensive multi-environment field trials in conjunction with high-throughput genotyping technology provide the phenotypic and genotypic data for use in statistical models shaped to predict the future performance of several hundred current elite winter wheat lines.

The project was largely designed and executed in co-operation with the Austrian breeding company Saatzucht Donau (SZD) and the Austrian Millers Association (ÖMV). The other participating enterprises were ProGen Seed, Probstdorfer Romania and the Hungarian breeding company Lajtamag. All partners were interested in taking advantage of robust winter wheat varieties capable of compensating year to year grain yield and price variations due to different stress conditions.

Genowheat represented one of the first applications of GS in a practical wheat breeding program and exposed all project partners for the first time to this new paradigm in plant breeding. Therefore the focus was to gain novel insights into the merits and limitations of GS in wheat breeding and its implementation into current breeding schemes. The duration of the project was set to three years, from late 2011 to the end of 2014.

Phenotypic data were obtained from multi-environmental field testing in six countries with substantial differences regarding environmental factors such as rainfall, drought or heat stress. This diverse set of locations should allow the simultaneous observation of adaptional characteristics, diseases resistance, grain quality and yield. Due to the nature of the ongoing breeding program not all wheat varieties were tested in every location in every year.

Genotypic data were obtained by analyzing more than thousand wheat lines with dense molecular markers by genotyping-by-sequencing. The large number of markers generated by this genotyping method should provide a uniform representation of the wheat lines' underlying genotypes and should allow simulating even complex traits like grain yield.

2 Substituting phenotypes - The state of research on genomic selection

Starting nearly 30 years ago, molecular marker analysis led to a revolution of molecular plant breeding and gave rise to a plethora of new methods (Moose and Mumm, 2008; Sharma et al., 2002; Varshney et al., 2006; Bernardo, 2008). At first molecular marker systems developed for crop plants were used to relate genetic marker information with important crop traits. Edwards et al. (1987) evaluated 82 traits in two F_2 populations of maize with 17-20 marker loci respectively for their genetic behavior linked to quantitative trait loci (QTL). Paterson et al. (1988) were the first to use a linkage map of tomato constructed by restriction fragment length polymorphisms (RFLP) markers at 57 loci to resolve quantitative traits into discrete Mendelian factors. Koziel et al. (1993) evaluated the potential of crosses from transgenic hybrid maize plants with commercial inbred lines, showing the successful application of molecular marker technology in plant breeding programs. Lusser et al. (2012) investigated the deployment of latest methods in biotechnology including zinc finger nuclease technology, oligonucleotide directed mutagenesis, cisgenesis and intragenesis, reverse breeding and agro-infiltration for plant breeding.

The essential part of plant breeding however is to select superior progeny to improve desirable traits. Conventional selection based on phenotypes has been effective throughout history and remains the gold standard in crop improvement. However for certain traits there is little progress in improvement due to challenges in assessing the true breeding value (Moose and Mumm, 2008). Time and cost intensive field trials have to be conducted and analyzed (see chapter 2.3) to obtain highly reliable phenotypic data.

Recent advances in molecular marker analysis and the advent of high-throughput genotyping platforms enable genome-wide marker analysis of one breeding line to be carried out at the cost of one plot in a common field trial. Consequently, when first introduced, technologies that use dense marker information to substitute phenotypic observation, i.e.

marker-assisted-selection (MAS) and genomic selection (GS) led to some excitement in the plant breeding community.

2.1 Marker-assisted-selection

The concept of marker-assisted-selection reaches back to the early last century when Sax (1923) discussed the theoretical advantage of indirect selection by genetic markers. It took however till the advent of high-throughput marker technologies before a sufficiently larger number of markers for most crop plants was available and the theory of MAS could be feasibly applied in plant breeding programs (Sharma et al., 2002).

Marker-assisted-selection is an indirect selection approach comprising of two essential steps, QTL mapping and QTL validation, in conjunction with the development of the mapping population.

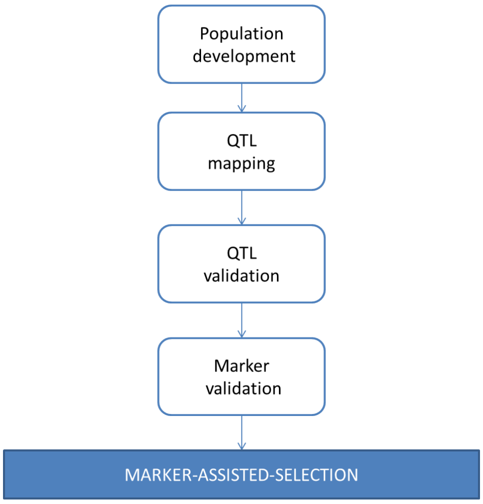


Figure 2.1: Workflow of marker-asssisted-selection (redrawn after (Collard and Mackill, 2008))

2.1.1 Population development

Contrasting parents are used to create a segregating population since only QTL for which the parents possess different allelic combinations can be detected by marker-assisted-selection. Generating many progenies from a certain parental cross assures a random assortment of the parental alleles, yet not all allele-gene combinations are completely random. Combinations that are in linkage disequilibrium, where marker alleles are tightly linked with the genes, fail to segregate independently and thereby are observed more frequently than one would expect by chance. Frequently used types of mapping populations are F_2 populations which can be quickly generated by selfing F_1 plants, back-cross populations where F_1 plants are crossed with one of their parents, recombinant-inbred-lines generated by single-seed-descent over several generations or doubled-haploid lines generated by micro-propagation in tissue cultures. Immortal populations comprising of at least 150-200 doubled-haploid lines or of later generations of back-crossed lines or recombinant-inbred-lines are of advantage compared to F_2 or F_3 derived populations since they can be permanently re-produced and allow repeated experiments (Collard and Mackill, 2008).

2.1.2 QTL mapping

Molecular markers that detect single genetic differences in the genomic sequence of the plants in the generated mapping population are used to create a linkage map of the population. Although most markers will be in the non-coding region of the genome and do not directly affect the trait, some will be linked to a QTL that controls the desired trait. Next to genotyping, phenotypic evaluation of the plants in the mapping population is one of the most critical tasks in the process of marker-assisted-selection. Very precise estimates of the true genetic value of each line are necessary to correctly identify QTL effects. QTL analysis then exploits the phenomenon of linkage disequilibrium between marker-QTL combinations where the marker locus and the QTL don't segregate independently and relates marker with phenotypic information (Kearsey, 1998).

The simplest method in QTL analysis to verify if a QTL is associated with a certain trait is a single factor analysis of variance (ANOVA) to test if the allelic means are significantly different for the trait (Sax, 1923). Since the test is performed separately for each marker, the map information is not taken into account so that the method only confirms the presence of a QTL in a given region. It provides an approximate QTL position, the explained phenotypic variance by each marker and the donor parent of the desirable allele. However, the further a QTL is away from a marker the less likely it will be detected by single marker analysis which can lead to an underestimation of marker effects (Collard et al., 2005; Tanksley, 1993). To overcome this limitation, dense genotyping with a high number of markers is desirable (see chapter 2.4.1).

A more commonly used method is interval mapping, first introduced by Lander et al. (1989), which uses the linkage map information to test if there is a QTL within the interval of two adjacent markers. Using theoretical markers spaced at narrow intervals,

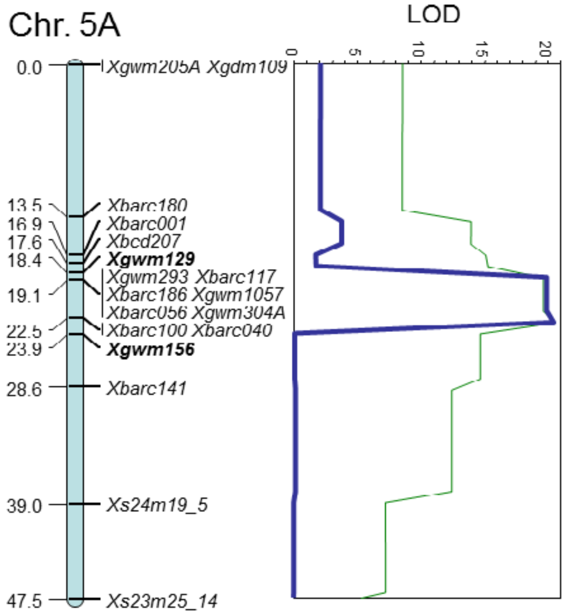


Figure 2.2: Composite interval mapping (blue) results in sharper LOD curves compared to standard interval mapping (green)

e.g. 1 cM between two markers, a more accurate estimation of the QTL position is possible. Interval mapping provides the same information as single marker analysis but in addition to the estimates of additive and dominance effects also an estimate of the epistatic effects across several QTL. Composite interval mapping, a more elaborate method of interval mapping uses markers near the QTL as co-factors in the analysis to give an even more accurate estimation of the QTL position (see Figure 2.2).

2.1.3 QTL and marker validation

Markers that don't reliably predict the phenotype are essentially useless in plant breeding programs. Hence, to use the markers found by QTL analysis effectively, further validation steps have to be performed (Collard et al., 2005). These steps may include the validation of the markers' effectiveness in different genetic backgrounds or independent breeding populations (Cakir et al., 2003; Avenue et al., 2003) or identification of a set of markers in a limited window spanning or flanking a QTL in order to tackle limited polymorphisms of individual markers (Collard and Mackill, 2008).

Once QTL which largely affect the phenotypic variation of a given trait are identified, including diagnostic markers for the beneficial alleles, breeders can use these markers for indirect selection. According to Collard et al. (2005) the advantages of marker-assisted-selection include among others:

- time saving from the substitution of laborious field trials with molecular markers;
- selection of lines at the seedling stage;
- avoid the transfer of undesirable genes, i.e. linkage drag, when introgressing new material;
- selecting for traits with low heritability.

The universal nature of the marker-assisted-selection scheme allows for different applications in plant breeding (Collard and Mackill, 2008). Molecular markers can be used prior to crossing to assess the genetic diversity and selection of parental lines for population generation. The efficiency of back-crossing could be increased by choosing the parent with the highest number of desirable alleles for certain traits or by reducing the linkage drag by controlling the size of the donor segment with flanking markers. Marker-assisted-selection could also be used to pyramid genes from multiple parents together into a single genotype. Most importantly, marker-assisted-selection could already be applied in early stages of the breeding process, eliminating undesirable alleles from the selection population, thus allowing the breeder to focus on a smaller number of plants in subsequent generations.

2.2 Genomic selection

Despite all theoretical advantages of marker-assisted-selection and thousands of reported marker-trait associations for different traits and plants, very few successful implementations have been reported in literature (Bernardo, 2008). While QTL mapping worked well for non-complex traits, especially for biotic stress tolerance traits that are governed by only few loci, utilizing QTL identified by MAS for complex traits was of limited success (Xu and Crouch, 2008). This may be partly due to the small contribution of each QTL to the total genetic variance or due to limitations to transfer QTL to different genetic backgrounds (Crossa et al., 2013).

When tackling the problem of many small effect QTL of most complex traits, Meuwissen et al. (2001) observed that it would be of advantage to use all contributing QTL simultaneously in marker-assisted-selection. They concluded that using a dense marker map that covers all chromosomes, it was possible to accurately predict the breeding value of animals and that selection based on those breeding values could substantially increase the rate of genetic gain in animal and plant breeding. The first successful implementations of this new method called genomic selection (GS) consequently came from animal breeding (VanRaden, 2007, 2008) and already some dairy breeding companies market bulls that have been selected solely based on the predicted genomic breeding value without pedigree testing information (Storlie and Charmet, 2013; Hayes et al., 2009). Few empirical results of GS in plants have been published yet, but several studies show the potential of GS in plant breeding. Lorenzana and colleagues (Lorenzana and Bernardo, 2009) used maize, Arabidopsis and barley bi-parental populations to study the accuracy of genotypic predictions for different traits, finding that for all but two of the tested traits the response per genomic selection cycle was at least half the response of phenotypic selection. A barley case study by Zhong et al. (2009) showed that the accuracy of genomic estimates was similar to phenotypic estimates.

Genomic selection uses genomic estimated breeding values (GEBV) that substitute previously used phenotypic breeding values for selection of individuals. In contrast to the marker effects obtained by marker-assisted-selection, these GEBV represent the sum of all underlying genetic effects captured by markers (Poland et al., 2012). GS thereby cancels the necessity to map individual marker effects (Zhao et al., 2012a). For complex traits where marker-assisted-selection often couldn't live up to the expectations, GS proved to be of advantage for different crops including maize. (de los Campos et al., 2009; Crossa et al., 2010; Jannink et al., 2010; Zhao et al., 2012a)

Figure 2.3 depicts the diagram of the GS process. Based on a training population which consists of plants that have been genotyped and phenotyped for the desired traits, the statistical model parameters are estimated. These parameters are subsequently used to calculate the genomic estimated breeding values for candidate breeding lines having only genotypic data (Heffner et al., 2009). It is possible to perform multiple cycles of genomic selection based on the same model parameters estimated from the training population. Since the GEBV are used for selection of candidate lines, a high accuracy is desirable. Therefore the training population has to be representative of the selection candidates and the model parameters have to be re-estimated if the training population is changed.

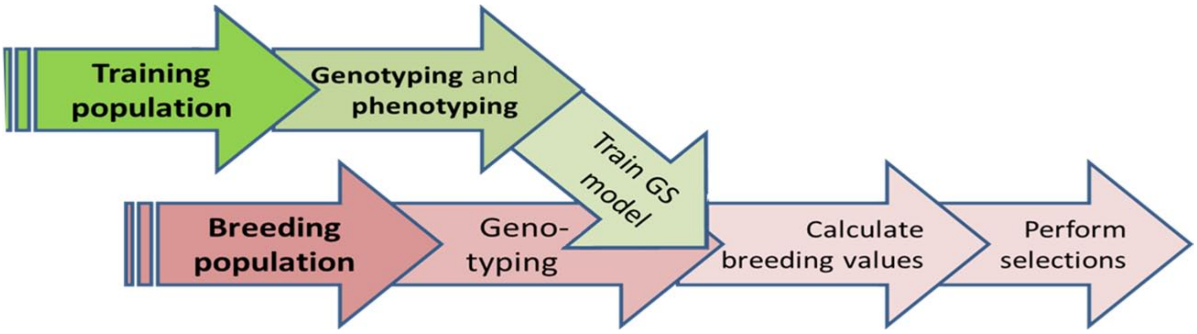


Figure 2.3: Diagram of the GS process (Figure taken from Heffner et al. (2009))

2.2.1 The influence of the training population

There are two types of population used in genomic selection, a training population by which the parameters of the selection model are estimated and a genetically broader breeding population for which the genomic estimated breeding values are calculated and selection is performed. Phenotypic information about the used plant material is only necessary for the training population whereas the calculated GEBV substitute for the phenotypic information in the breeding population. Given the size of the training population is much less than the size of the breeding population, the amount of cost and labor intensive field trials reduces drastically when only the training population has to be phenotyped. Taken together with the stagnant phenotyping costs and decreasing genotyping costs it is clear why genomic selection is believed to be able to revolutionize plant breeding (Jannink et al., 2010).

The setup of the training population is essential since the derived model parameters determine the GEBV which are used as the selection criterion. General findings, e.g. from trials at the International Maize and Wheat Improvement Center (CIMMYT), show that the prediction accuracy between genetically unrelated populations is low (Crossa et al., 2013). Consequently, as considering family structures increases the prediction accuracy, a careful choice of breeding lines used for the training population must be taken. Riedelsheimer et al. (2013) analyzed five interconnected doubled haploid (DH) populations of maize and came to the conclusion that for high-heritable traits within-population prediction is feasible even with moderate training population sizes. Unrelated crosses with opposite linkage phase similarities, i.e. the proportion of marker pairs with the same linkage phase in both populations, with the breeding population can have negative or reduced prediction accuracies and should be excluded from the training population. Similar results were found earlier by Habier et al. (2007) in animal breeding, who concluded, that the accuracy of GEBV can result in a large part from genetic relationship captured by markers. Other factors that influence the prediction accuracy are the statistical selection model, the training population size and composition and the marker density (Lorenzana and Bernardo, 2009; Lorenz et al., 2012). Recently Endelman et al. (2014) investigated the optimal design of preliminary yield trials with genome-wide markers available and found that the prediction accuracy slightly decreased as more phenotyping resources were devoted to existing lines in the training population instead of phenotyping new lines.

2.2.2 Estimating genomic breeding values

When introducing the concept of genomic selection, Meuwissen et al. (2001) made use of the high-density genotyping technologies and concluded, that due to the high density of markers covering the genome, population-wide estimates of the marker effects would be meaningful (Jannink et al., 2010). Because of the large number of markers, usually in the range of thousands, compared to the number of plants in the training population, usually in the range of hundreds, the statistical challenges of high-dimensional data arise (Johnstone and Titterington, 2009). The so-called large p small n paradigm where the number of parameters p, in this case markers, is much larger than the number of observations n, the phenotypes, is well-known in statistics and machine learning. Since variable selection that reduces the number of markers contradicts the original aim of Meuwissen et al. (2001) of avoiding marker selection so that the estimated marker effects are unbiased, multiple linear regression can't be used (Jannink et al., 2010). A number of statistical methods such as best linear unbiased prediction (Kolbehdari et al., 2007), ridge regression (Kennard and Hoerl, 1970), Bayesian models (Meuwissen et al., 2001) and other machine learning methods like support vector machines and Random Forest have been proposed as prediction models for genomic selection (Heslot et al., 2012). All these models use the available phenotypic and genotypic data of the samples in the training population to estimate the model parameters of which the GEBV of the untested population are calculated. A detailed description of the models and their performance in the training and validation set at hand can be found in chapter 9.

2.2.3 The accuracy of genomic breeding values

To assess the goodness of fit of the genomic selection model, i.e. the accuracy of the GEBV, the Pearson correlation coefficient measuring the degree of correlation between the GEBV predicted by the selection model and the observed phenotypic data is often used (Storlie and Charmet, 2013). When the accuracy achieved by GS models, expressed as $r_{GS} = \frac{r_m}{h}$, equals the square root of the heritability h of the trait in question, one cycle of genomic selection is as effective as phenotypic selection (Dekkers, 2007). Several studies investigated the influence of different factors on the accuracy of the genomic predictions. Zhao et al. (2012b) used 788 testcross progenies of an elite maize breeding program to study the influence of phenotypic selection on the accuracy and bias of genomic selection, which showed a substantial loss in accuracy in populations with unidirectional selection. Heslot et al. (2013) investigated the role of marker ascertainment bias, essentially a sampling bias, on GS accuracy, and found it significant for one of the tested traits.

A large number of reported accuracies come from cross-validation experiments. Cross-validation is used in statistical analysis to test how well the results can be generalized to an independent data set. Since phenotypic and genotypic values are commonly just available for the training population, the set of possible plants is randomly split into a a given number, e.g. k , of disjunct subsets. The genomic selection model is then trained on $k - 1$ subsets to estimate the model parameters, while the remaining subset is used for validation since genomic estimated predictions as well as observed phenotypic data are available. This process is repeated k-times so that every subset is once used as the validation set and the results can be averaged. Hofheinz et al. (2012) used 310 sugar beet lines phenotyped for a high ($h~_P = 0.9$) and a low ($h~_P = 0.4$) heritable trait to investigate how well the cross-validation results reflect the results in the next generation. While the accuracies for both traits in the cross-validation set were > 0.8, correlation accuracies in the next breeding cycle were only 0.8 for one trait and dropped to 0.4 for the other trait under investigation. The authors therefore concluded that the prediction of lines of the next cycle in the breeding program is most effective for traits with high heritability.

2.2.4 Prospects of genomic selection in plant breeding

Initial theoretical findings on genomic selection rose hopes that it could revolutionize plant breeding as it has dairy cattle breeding and led to an eminent amount of research activity in both areas of breeding (Heffner et al., 2009; Weng et al., 2012). A prime example of GS in breeding can be seen in the first large-scale implementation of GS in dairy cattle (Poland et al., 2012). Phenotyping of dairy bulls is performed by daughter progeny testing which is time and very cost intensive, typically thousands of dollars per bull. Genotyping of the bull however costs a few hundred dollars, depending on the used technology. Heffner et al. (2010) concluded that GS could drastically accelerate genetic gain through shorter breeding cycles when moderate accuracies are achieved and up to two years of field trials could be eliminated. Designing a GS-based plant breeding program however is a complex process. Endelman et al. (2014) argued that breeders may need the help of decision support software to account for all caveats that could arise. Storlie and Charmet (2013) concluded that though genomic selection may not substitute the need of field trials it could take a more interactive role in plant breeding programs.

2.3 Evaluating field trial designs for high quality phenotypes

Accurate phenotypic information has been the key for successful selection in plant breeding throughout history. Genomic selection uses phenotypic information only for estimating the model parameters in the training population. However the precision of the phenotypic information is of paramount importance for the quality of the genomic estimated predictions. Therefore, even in a lower number, carefully planned and analyzed field trials are all the more necessary.

For certain traits which are mainly governed by single genes or quantitative trait loci even one trial may be enough to correctly assess the performance of a variety, e.g. infection trials for disease resistance. Other traits on the other hand are controlled by many genes or QTL and often there is a genotype by environment (GxE) interaction. If so, the change in phenotype of a variety differs between some of the locations the field experiment is conducted in. More generally ,a genotype by environment interaction refers to a change in phenotype reaction specific for a certain combination of a genotype and an environment (Hinkelmann, 2011; Des Marais et al., 2013). Special cases of GxE interactions are genotype by year, genotype by location and genotype by treatment interactions or combinations of interactions with specific QTL.

As most of today's plant breeding programs aim to create varieties with broad adaption to diverse environmental conditions, variety trials have to be conducted in multiple environmental conditions. In the case of wheat, new varieties are tested at several locations over multiple years and treatments to accommodate a broad range of different conditions. These early tests are usually performed by breeders and the most promising candidates may enter the official variety tests performed by most countries. Consequently, the time from crossing to create new genetic variation to the registration of a new variety is cost intensive and takes several years. Well laid out field trials that increase the chance of a variety to be registered by enabling the breeder to select the actually best performing varieties are therefore of high importance in any breeding program.

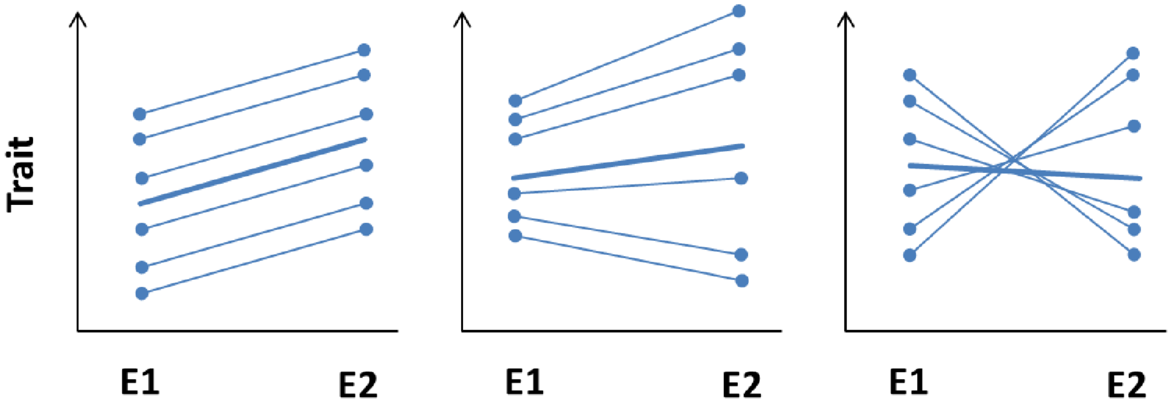


Figure 2.4: Types of GxE interactions (Figure redrawn after Des Marais et al., 2013)
Figure a) Change of scale with no change in variance
Figure b) Change of scale with change in variance
Figure c) Change of ranks

2.3.1 A short history of field trial designs

Unreplicated trials where fields were divided into large subsets - roughly equal sized areas - were common in the beginning of 20th century. Later field trial designs that incorporated replication started to become an active field of research (Lupton and Others, 1987), but the real foundation was laid by R.A. Fisher in his work about the arrangement of field experiments and the design of experiments (Fisher, 1936, 1992) whose methods are up to now widely used. He observed that experimental observations are only experience carefully planned in advance, and designed to form a secure basis of new knowledge . Since then a plethora of articles and books have been published on various experimental designs. The key principles yet stay the same and will be shortly described.

Randomization

Randomization in the context of field experiments is the process of randomly assigning plots into different blocks to avoid undesired effects such as:

- systematic bias, e.g. a certain treatment is only applied to a rocky area of the field whereas the other treatment is applied to a normal area;
- selection bias, e.g. only the most resistant varieties are chosen for a certain treatment;
- accidental bias, e.g. using seeds from one source for one treatment and seeds from another source for another treatment;
- cheating by the experimenter (Bailey, 2008)

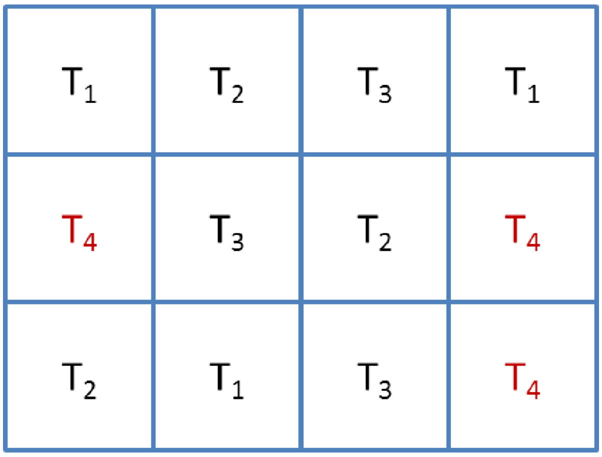
and to give unbiased estimates of the error variance.

Replication

Replication refers to treatments that are repeated more than one time to measure experimental error (Compton, 1994). In field trials replication often refers to repetition of varieties within the field trials to estimate the variability of the trial site and to account for the uncertainty of measurements. If each variety is replicated the same number of times, the design is called to be balanced.

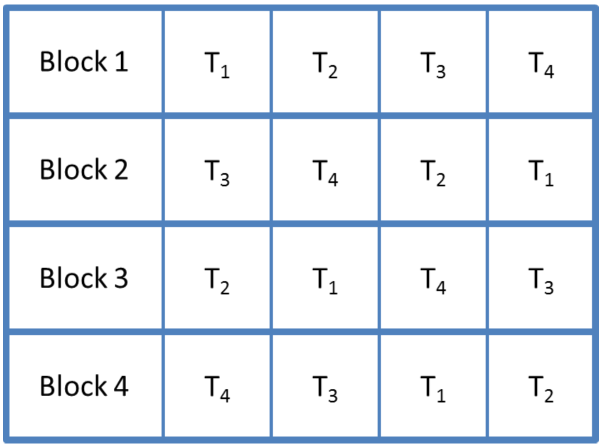
Blocking

Blocking refers to the grouping of experimental units into blocks in a way that the units within a block are alike and thereby reducing variation within such blocks. Grouping into uniform blocks provides a better estimate of treatment effects by removing otherwise unaccounted variation from the experimental error which increases the precision (Trigiano and Gray, 1999). In field trials such blocks comprise small areas where it is assumed that



T_1	T_2	T_3	T_1
T_4	T_3	T_2	T_4
T_2	T_1	T_3	T_4

Figure 2.5: Balanced replication of four varieties



Block 1	T_1	T_2	T_3	T_4
Block 2	T_3	T_4	T_2	T_1
Block 3	T_2	T_1	T_4	T_3
Block 4	T_4	T_3	T_1	T_2

Figure 2.6: Blocking of four varieties

plots closely located to each other are likely to behave more similar than plots that are farther apart. According to Bailey, 2008, if possible, blocking should be performed in a way that:

- all blocks should have the same size;
- blocks should be large enough to include each treatment.

2.3.2 Replicated control design

The replicated control design (RCD) is one of the simplest designs suited for a large number of test varieties for which just little seed is available. The test varieties are

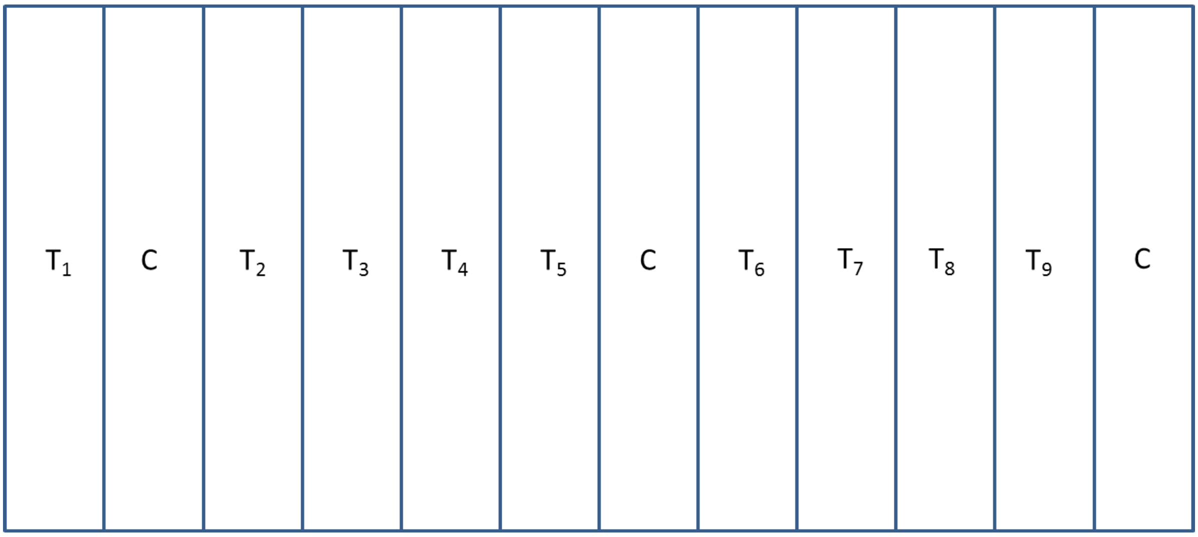


Figure 2.7: Replicated control design
Test varieties are indicated as T_1 to T_9
Check variety is indicated as C

plotted without replications separated by a replicated check variety every three to five plots. Since this design is unreplicated in terms of the test varieties it is not possible to estimate error variances for the test varieties or perform analysis of variance. Instead the check variety is used to estimate a field heterogeneity for which the testing varieties are corrected.

Considering a field trial that is used to test nine different test varieties, a layout similar to Figure 2.7 can be used. If test variety T_1 yielded 55 decitonnes/hectare (dt/ha) and test variety T_9 yielded 70 dt/ha, one could simply assume that T_9 outperformed T_1. If however the control variety next to T_1 yielded 60 dt/ha and the one next to T_9 yielded 75 dt/ha, the difference between the test varieties may reflect a heterogeneity of the field rather than a true difference based on the genotype.

2.3.3 Completely randomized designs

Completely randomized designs (CRD) rely to a great extent on the principles of replication and randomization. The treatments are assigned at random to the experimental units and are not limited in their number. CRD are among the most powerful statistical designs

Table 2.1: Yield of four wheat varieties with four replicates in a CRD design

A: 42	D: 28	C: 38	B: 36
C: 38	B: 32	A: 44	D: 26
A: 38	D: 22	A: 36	D: 24
C: 32	B: 26	C: 28	B: 30

as the degrees of freedom used for estimating the experimental error are maximized which increases the power to detect statistical differences among treatments (Trigiano and Gray, 1999). However if there is a large amount of variation not assigned to treatment effects this variation will be allocated to the experimental error resulting in a lower power when deciding to reject the null hypothesis of no differences between treatments. If parts of this non-accounted variation can be identified, e.g. environmental field heterogeneity, designs that use blocking to account for this variation are of advantage. Statistical analysis is often carried out using standard analysis of variance. The test statistics for e.g. a treatment effect is to test the null hypothesis of no difference in treatments

$$\begin{aligned} H_0 :\ & {}_1 \\ = \ & {}_1 \\ = \ & {}_2 \\ = \ & \dots \\ = \ & {}_t \end{aligned}$$

against the alternative hypothesis of at least one mean is different

$$H_1 : \text{at least one mean is different.}$$

In the simplest case this is done by dividing the treatment variance by the error variance. If the null hypothesis is true the test statistic follows a F-distribution with $(t-1)$ degrees of freedom for the numerator and $(b-1)*(t-1)$ degrees of freedom for the denominator (Crawley, 2005).

Table 2.1 shows the yield performance of four winter wheat varieties in dt/ha. Each variety was replicated four times and plotted in a CRD. The statistical analysis gives an F-value of 10.17 for the null-hypothesis of no differences between the wheat varieties which therefore can be safely rejected. However a large proportion (reflected by the error sum of squares) of the total variance is not accounted for by the model (see Table 2.2).

Table 2.2: ANOVA of the CRD of table 2.1

Source	Sum of squares (SS)	Degrees of freedom (df)	Mean Square (MS)	F
Variety	468	3	156	10.17
Error	184	12	15.33	
TOTAL	652	15		

2.3.4 Randomized complete block designs

Randomized complete block designs (RCBD) have become widely used to overcome the problem of completely randomized designs that all unrecognized variation is allocated to the experimental error. As stated above, grouping into uniform blocks provides a better estimate of treatment effects by removing otherwise unaccounted variation from the experimental error (Trigiano and Gray, 1999). In the context of variety tests each variety is plotted in several replications such that each replication forms a block in which the varieties are randomly plotted to avoid bias. Each block has to contain each variety once, which minimizes the within block variation while maximizing the between block variation. Furthermore the randomization of the varieties within the blocks should be done in a way that:

- each test variety is plotted in each block once;
- within each block the varieties are randomly plotted;
- each block has a different randomization.

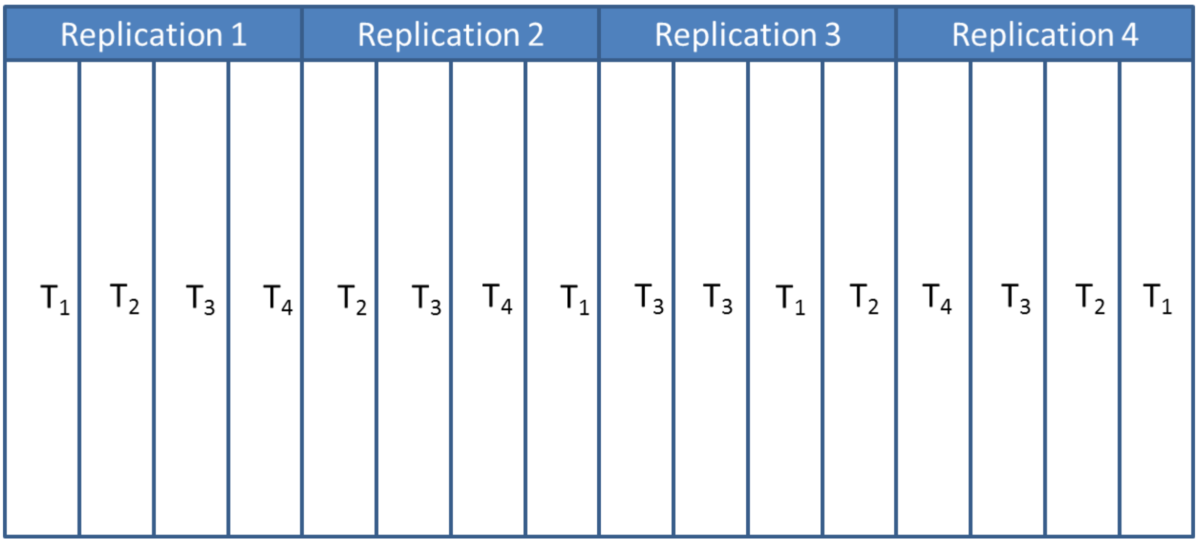


Figure 2.8: Randomized complete block design

If the experiment shown in table 2.1 was performed utilizing a randomized complete block design, a possible layout could look like:

Block	A	B	C	D
1	42	36	38	28
2	44	32	38	26
3	38	30	32	24
4	36	26	28	22

The statistical analysis is similar to the CRD with the block number added as factor to the ANOVA model. Both factors, the variety and the block number are significant but the unaccounted variance, i.e. the error variance is greatly reduced which increases the power to detect differences between the treatments though with a loss in degrees of freedom. Due to the loss of freedom the test statistic could lose sensitivity if the blocks are inconsiderately chosen (Bärlocher, 2008).

Table 2.3: ANOVA of the RCBD

Source	Sum of squares (SS)	Degrees of freedom (df)	Mean Square (MS)	F
Variety	468	3	156	70.2
Block	164	3	54.66	24.6
Error	20	9	2.22	
TOTAL	652	15		

2.3.5 Incomplete block designs

As a consequence of the increasing block size of randomized complete block designs, as more varieties are tested, such designs become less suitable for large scale plant breeding trials. Larger block sizes also lead to a contradiction of the ideal property of a block of being homogeneous (Lentner and Bishop, 1986). Hence, when the number of varieties becomes too large for a single homogeneous block the use of RCBD becomes inefficient and incomplete block designs (IBD) using blocks that do not contain all treatments may be better suited. Two types of incomplete block designs are commonly used (Chahal and Gosal, 2002):

- balanced incomplete block designs and
- partially balanced incomplete block designs.

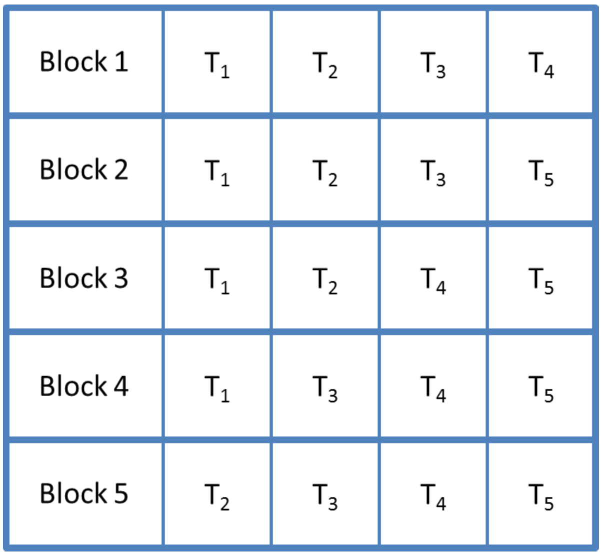
Balanced incomplete block designs are defined in terms of number of genotypes g plotted in b blocks of size k in r replications with λ possible direct comparisons between each treatment (see Figure 2.9 for an example) (Chahal and Gosal, 2002)(Sharma, 2008):

(i) Each treatment pair occurs together in an equal number of times in the same block

(ii) Each treatment pair is compared with the same precision

(iii) The number of replications needed for balance is determined by g and k

Other types of incomplete block designs, e.g. rectangular lattice designs or alpha lattice designs allow for more flexibility in choice of b blocks of size k.

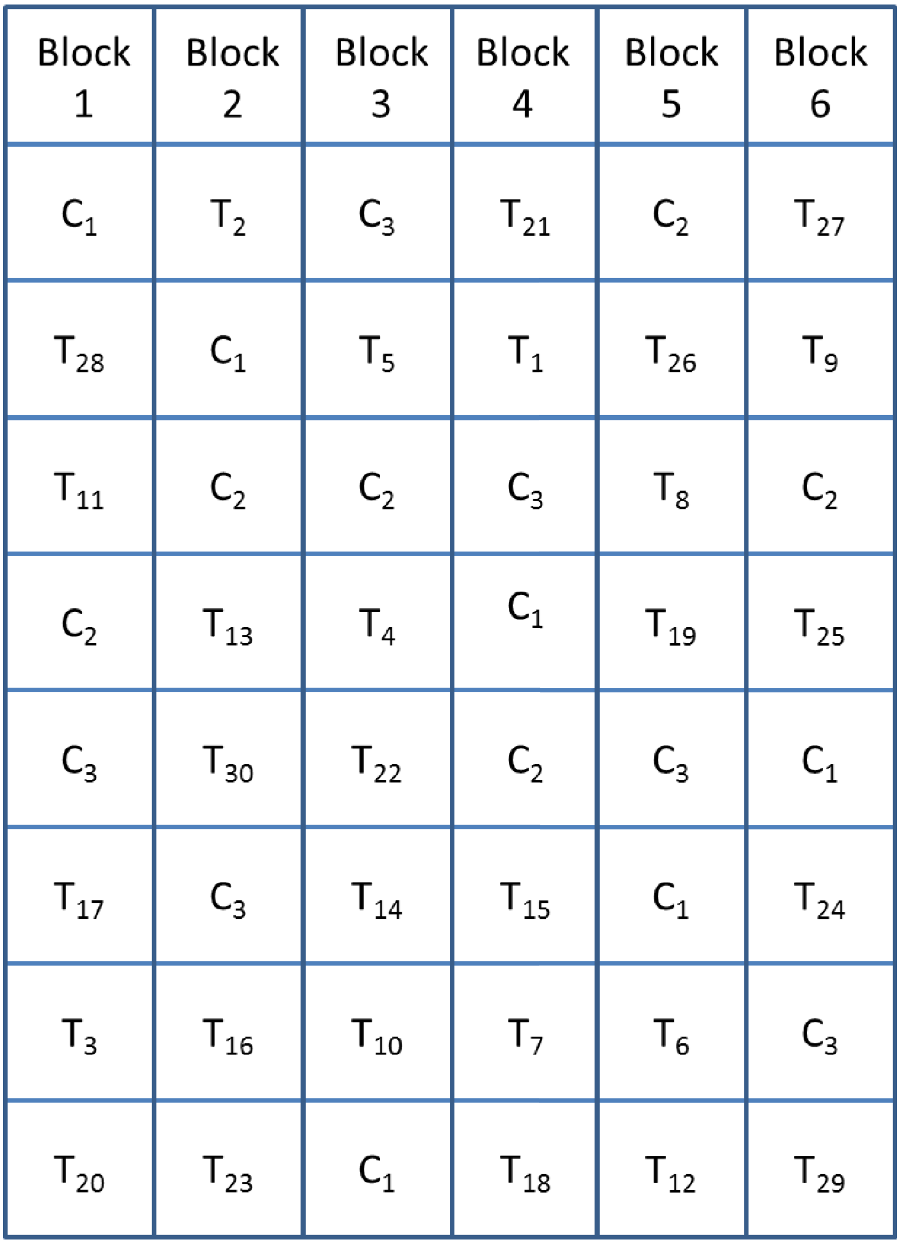


Block 1	T_1	T_2	T_3	T_4
Block 2	T_1	T_2	T_3	T_5
Block 3	T_1	T_2	T_4	T_5
Block 4	T_1	T_3	T_4	T_5
Block 5	T_2	T_3	T_4	T_5

Figure 2.9: Example of a balanced incomplete block design with $g = b = 5$, $r = k = 4$ and $\lambda = 3$

2.3.6 Augmented designs

In the early stages of a breeding program where a large number of new varieties are tested, a completely replicated design may not be feasible in terms of cost and seed availability. Without replication a meaningful estimate of the error associated with the test varieties cannot be calculated (Chahal and Gosal, 2002). Federer (Federer, 1956; Federer et al., 1975; Faraway, 2002) proposed another form of complete block designs where check varieties are replicated to form blocks. The design for the check treatments can be chosen freely, e.g. a RCBD, IBD or a row-column design (Federer et al., 1999). These blocks are then augmented by plots of unreplicated - or just replicated in a small number - test varieties, hence the name. The resulting blocks all consist of the same replicated check varieties but differ in the set of unreplicated test varieties (Chahal and Gosal, 2002). The replicated check varieties provide an estimate of the standard error used to compare among the test varieties or between the check and test varieties and the blocking allows for adjustment of spatial field heterogeneity. Also the block size does not have to be the same for all blocks. However, due to the fewer degrees of freedom for the experimental error the power to detect differences between varieties is reduced. Statistical analysis of augmented designs in this work is carried out using linear mixed models (Henderson, 1975) together with a row-column design. This allows for the new varieties to be treated either as fixed or random effects and to model two sources of field heterogeneity.



Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
C_1	T_2	C_3	T_{21}	C_2	T_{27}
T_{28}	C_1	T_5	T_1	T_{26}	T_9
T_{11}	C_2	C_2	C_3	T_8	C_2
C_2	T_{13}	T_4	C_1	T_{19}	T_{25}
C_3	T_{30}	T_{22}	C_2	C_3	C_1
T_{17}	C_3	T_{14}	T_{15}	C_1	T_{24}
T_3	T_{16}	T_{10}	T_7	T_6	C_3
T_{20}	T_{23}	C_1	T_{18}	T_{12}	T_{29}

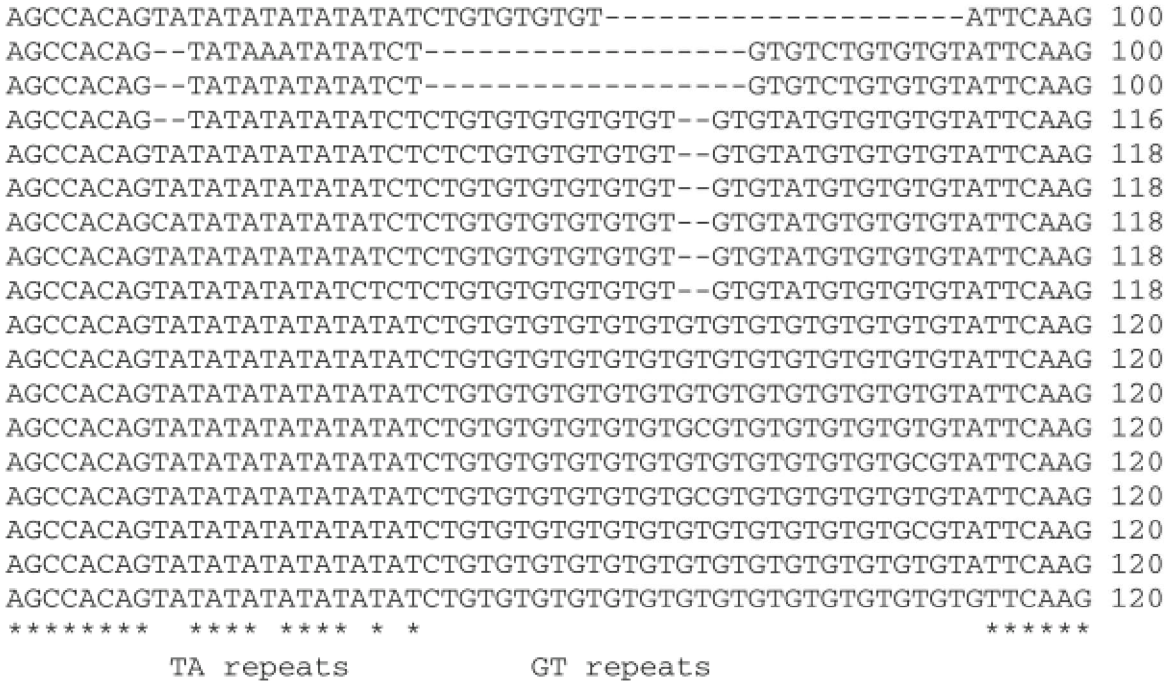
Figure 2.10: Example layout of an augmented design
3 check varieties are replicated over 6 blocks and augmented by 30 test varieties

2.4 Genomic fingerprinting

The advancements in molecular analysis made it feasible to assess the genetic variability of whole breeding populations at an unprecedented level based on the genomic sequence of individual plants rather than infer it from the observed phenotype. These advances led the genomics revolution in molecular plant breeding and gave rise to numerous new methods which all rely on markers that detect polymorphisms found in the genomic sequence of the DNA. Genomic selection uses statistical models to relate the phenotypic information of plants in the training population to their respective molecular marker information obtained by genotyping to estimate the effects of the individual markers on the trait under investigation. These molecular markers detect variants at homologous sites in the DNA sequence of individuals among a group of individuals. Often these polymorphisms are located in the non-coding region of the DNA and are therefore not phenotypically observable. Added to the fact that the number of detectable polymorphisms is ultimately limited only by the genome size, molecular markers became of high interest in research in the field of genomics in animal and plant breeding (Albrechtsen et al., 2010; Ignal et al., 2002; Heslot et al., 2013). Numerous technologies of DNA markers exist who all detect polymorphisms that are either due to point mutations of the DNA or insertions or deletions of parts of the DNA including variations in the number of tandem repeats (Ignal et al., 2002).

Insertions and deletions

Insertions or deletions (indels) or a combination thereof can affect stretches of DNA that range from single base pairs to several hundred base pairs. Markers based on these polymorphisms have been the method of choice for many studies dealing with relationship between individuals or populations. A fairly small number is usually sufficient for a broad range of studies in population genetics due to the high information content (Väli et al., 2008). However, single nucleotide polymorphisms are recently becoming the most commonly used type of marker used for studies of animal and plant populations.



```
AGCCACAGTATATATATATATATCTGTGTGTGT--------------------ATTCAAG 100
AGCCACAG--TATAAATATATCT------------------GTGTCTGTGTGTATTCAAG 100
AGCCACAG--TATATATATATCT------------------GTGTCTGTGTGTATTCAAG 100
AGCCACAG--TATATATATATCTCTGTGTGTGTGTGT--GTGTATGTGTGTGTATTCAAG 116
AGCCACAGTATATATATATATCTCTCTGTGTGTGTGT--GTGTATGTGTGTGTATTCAAG 118
AGCCACAGTATATATATATATATCTCTGTGTGTGTGT--GTGTATGTGTGTGTATTCAAG 118
AGCCACAGCATATATATATATATCTCTGTGTGTGTGT--GTGTATGTGTGTGTATTCAAG 118
AGCCACAGTATATATATATATATCTCTGTGTGTGTGT--GTGTATGTGTGTGTATTCAAG 118
AGCCACAGTATATATATATATCTCTCTGTGTGTGTGT--GTGTATGTGTGTGTATTCAAG 118
AGCCACAGTATATATATATATATATCTGTGTGTGTGTGTGTGTGTGTGTGTGTATTCAAG 120
AGCCACAGTATATATATATATATATCTGTGTGTGTGTGTGTGTGTGTGTGTGTATTCAAG 120
AGCCACAGTATATATATATATATATCTGTGTGTGTGTGTGTGTGTGTGTGTGTATTCAAG 120
AGCCACAGTATATATATATATATATCTGTGTGTGTGTGTGCGTGTGTGTGTGTATTCAAG 120
AGCCACAGTATATATATATATATATCTGTGTGTGTGTGTGTGTGTGTGTGCGTATTCAAG 120
AGCCACAGTATATATATATATATATCTGTGTGTGTGTGTGCGTGTGTGTGTGTATTCAAG 120
AGCCACAGTATATATATATATATATCTGTGTGTGTGTGTGTGTGTGTGTGCGTATTCAAG 120
AGCCACAGTATATATATATATATATCTGTGTGTGTGTGTGTGTGTGTGTGTGTATTCAAG 120
AGCCACAGTATATATATATATATATCTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTCAAG 120
********  **** **** * *                               ******
         TA repeats          GT repeats
```

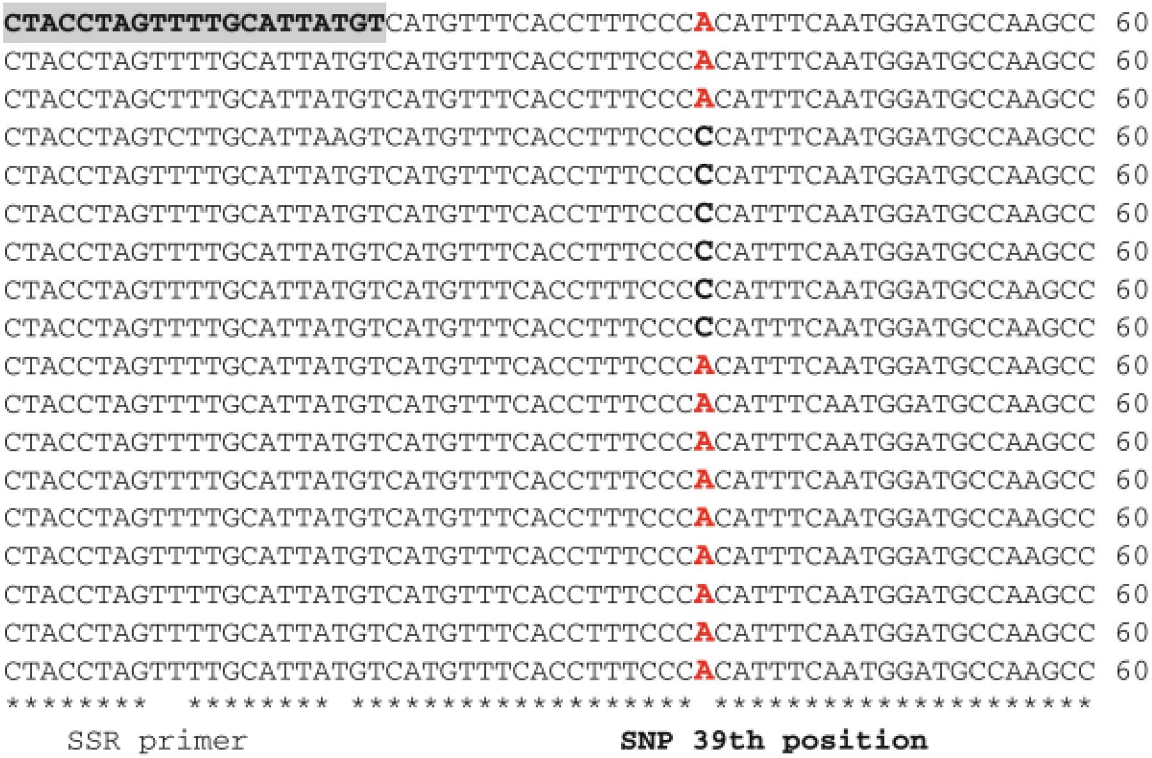
Figure 2.11: Example of an indel (Figure taken from (Fang et al., 2010))

Point mutations

Single nucleotide polymorphisms (SNP) are single base differences in the genomic sequence among a group of individuals. These polymorphisms are much more prevalent in the genomic sequence, by far outnumbering markers based on indels, since they occur likewise in the coding and non-coding regions.

2.4.1 High-throughput genotyping platforms

The first molecular marker systems used for genetic mapping were cost and labor intensive while only a small number of DNA variants could be detected in one genotyping task. For a feasible use in genomic selection however, a large number of such variants had to be identified. For quite some time genotyping technologies advanced slowly with only incremental advances (Edwards and Henry, 2011) and only the advent of high-throughput genotyping platforms enabled genome-wide marker analysis at reasonable costs. Initially, allele specific microarrays were used to detect thousands of single base differences (SNP)



```
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGCTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTCTTGCATTAAGTCATGTTTCACCTTTCCCCCATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCCCATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCCCATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCCCATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCCCATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCCCATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
CTACCTAGTTTTGCATTATGTCATGTTTCACCTTTCCCACATTTCAATGGATGCCAAGCC 60
********  ******** ******************* *********************
   SSR primer                      SNP 39th position
```

Figure 2.12: Example of an SNP (Figure taken from (Fang et al., 2010))

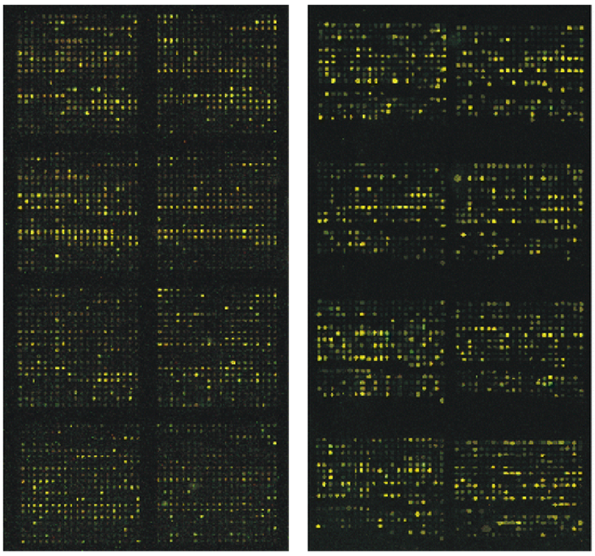


Figure 2.13: Sample of a cDNA microarray (picture taken from (Nazar et al., 2010))

in one hybridization step (Miller et al., 2007) followed by technologies that relied on the advances in next generation sequencing (NGS) technology (Baird et al., 2008).

2.4.1.1 Chip-based technologies

Chip-based genotyping approaches rely on high-density molecular arrays to detect polymorphisms with oligonucleotide probes that have been plotted to the arrays (Albrechtsen et al., 2010). Oligonucleotide microarrays commonly consist of oligonucleotide probes in the range of 10 to 25 bases while cDNA microarrays usually comprise 500 to 2000 bases of cDNA amplified by polymerase chain reaction (PCR) (Lemieux et al., 1998). Commonly used genotyping platforms use probes based on SNP that were previously discovered by sequencing, e.g. of expressed-sequence-tag libraries (Heller, 2002). As a result from this pre-selection of available SNP these chip-based technologies may suffer from ascertainment bias (Foll et al., 2008; Guillot and Foll, 2009; Albrechtsen et al., 2010). Heslot et al. (2013) investigated the impact of the ascertainment bias, essentially a sampling bias, on the accuracy of genomic selection and estimates of genetic diversity. They concluded that the used chip-based markers overestimated the genetic diversity which is a clear indication of ascertainment bias. However, with the same number of used markers, genomic predictions were not significantly different when sequenced-based markers and chip-based markers were used.

2.4.1.2 Sequencing based technologies

The advent of next generation sequencing (NGS) made it possible to sequence complete genomes of almost any species. While the technology initially was used for gene expression analysis, recently it is often used to discover SNP directly on genomic DNA at a genome-wide level in animals, humans and plants (Davey et al., 2011). Genotyping-by-sequencing in general uses restriction enzymes to generate reduced representation libraries of the target genomes (Baird et al., 2008). Elshire et al. (2011) presented a method that uses methylation-sensitive restriction enzymes to avoid repetitive regions and target lower copy regions with higher efficiency. The restriction enzymes cut the DNA frequently to generate randomly sheered fragments of DNA with overhanging ends, i.e. restriction-site associated (RAD) DNA. Barcoded adapters, short DNA sequences, are then annealed to the fragments upstream of the cut site of the restriction enzymes. The barcoded adapters allow sequencing multiple samples, typically 96 or 384, in parallel on a single run (Poland et al., 2012). Based on this principle, different protocols varying in adapter choice and number have been proposed (Baird et al., 2008; Elshire et al., 2011; Davey et al., 2011).

Sequencing based genotyping has become popular of late and GBS has been successfully applied to a variety of species, e.g. and Arabidopsis, barley, Eucalyptus, maize and wheat (Poland et al., 2012; Romay et al., 2013; Crossa et al., 2013; Lorenzana and Bernardo, 2009; Sansaloni et al., 2011).

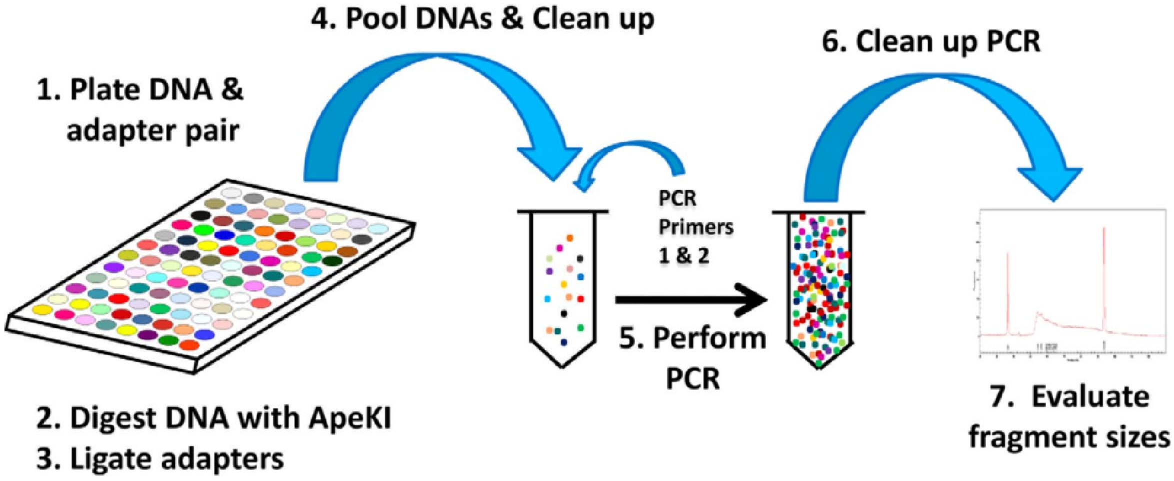


Figure 2.14: Genotyping-by-sequencing library construction (Figure taken from (Elshire et al., 2011))

Part II

Materials and Methods

3 Field trials and phenotyping

Prior to the implementation of genomic selection a typical pedigree-trial breeding scheme was used. The first generation of crosses (F1) between parents showing favorable characteristics was grown in Chile in autumn with the subsequent F2 generation planted after vernalisation in the following spring. Single ears were cut out and planted as rows in a block (F3) for scoring of heading, plant height, disease resistance and early quality analyses based on grain observations. The F4 generation was grown in blocks of double row units of single ear descendants of the F3 generation. Phenotyping at this stage already included more quality analyses, e.g. thousand grain weight (TGW) and near-infrared spectroscopy (NIRS). First yield trials were conducted in the F5 generation. This so called observation (OBS) set was tested in Austria and Turkey. The F6 generation comprised of unreplicated field trials in the so called YOUNG set at ten locations for yield and quality traits. The further advanced pre-commercial (PRECW) set was tested in replicated trials in 18 different locations. Promising new cultivars were then selected for the first of three years of official testing for variety registration. A figure showing the breeding cycle in more detail can be found in the Appendix (18).

3.1 Locations and experimental designs

The wheat lines were tested in six different countries, i.e. Austria, Hungary, Romania, Serbia, Slovakia and Turkey. A comparison of all locations performed in the preceding project showed no consistent correlation among locations over years. Due to the more extreme climate conditions reflected by large variations in grain yield levels, Konya in Turkey correlated least with the other locations. To evaluate the broad adaption of the wheat lines in this project, an even more drought-prone testing location in the Central Anatolian Plateau, Karapinar, which is east of Konya, is included. The YOUNG set comprising of 180 entries was planned as an unreplicated incomplete block design, e.g. a $4x45$ generalized lattice design, evaluated in eight to ten locations. The pre-commercial PRECW set comprising of 25 to 30 advanced breeding lines was planned as a replicated $2x30$ or $4x25$ generalized lattice design evaluated in all 18 locations.

Table 3.1: Climate data of the Genowheat testing locations
Rainfall and elevation data are approximate values of the regions (www.climate-data.org; Last accessed: 16.02.2015)

Country	Location	Coordinates	Elevation	Rainfall	Temp.
Austria	Probstdorf	48°12'N 16°33'E	156 m (512 ft)	622 mm	10.0°C
	Leopoldsdorf	48°07'N 16°23'E	161 m (528 ft)	630 mm	9.8°C
	Weikendorf	48°21'N 16°46'E	152 m (499 ft)	634 mm	9.8°C
	Aumühle	48°28'N 15°88'E	180 m (590 ft)	625 mm	9.7°C
	Reichersberg	48°20'N 13°21'E	347 m (1138 ft)	1038 mm	8.7°C
Hungary	Székkutas	46°30'N 20°32'E	84 m (275 ft)	533 mm	10.8°C
	Jászboldogháza	47°37'N 20°00'E	90 m (295 ft)	519 mm	10.7°C
	Mosonmagyaróvár	47°52'N 17°16'E	120 m (394 ft)	594 mm	10.0°C
Romania	Modelu	44°12'N 27°23'E	19 m (61 ft)	493 mm	11.5°C
	Constanta	44°10'N 28°38'E	25 m (82 ft)	423 mm	11.6°C
	Dr g nesti	44°10'N 24°31'E	93 m (304 ft)	496 mm	10.6°C
	Livada	47°52'N 23°07'E	130 m (426 ft)	585 mm	10.9°C
	Fundulea	44°28'N 26°30'E	70 m (230 ft)	567 mm	10.9°C
Serbia	Sombor	45°47'N 19°07'E	90 m (295 ft)	581 mm	9.8° C
Slovakia	Haniska	48°57'N 21°14'E	241 m (790 ft)	622 mm	8.8°C
	Rado ina	48°32'N 17°56'E	192 m (630 ft)	668 mm	8.9°C
Turkey	Babaeski	41°25'N 27°05'E	55 m (180 ft)	614 mm	14.0°C
	Hayrabolu	41°12'N 27°06'E	81 m (266 ft)	632 mm	13.8°C
	Karapinar	37°43'N 33°33'E	987 m (3238 ft)	432 mm	9.0°C
	Konya	37°52'N 32°29'E	1200 m (3900 ft)	337 mm	11.3°C

3.2 Phenotyping

Phenotyping of the breeding material was performed for a large number of traits of different groups. Robustness mediating traits included morphological characteristics like tillering capacity or plant height, abiotic stress resistance and physiological characteristics like frost resistance or heading date and biotic stress resistance characteristics like powdery mildew resistance or Fusarium head blight. The group of quality traits included among others protein content in dry matter and thousand grain weight. Depending on the trait, phenotyping was performed only in suitable locations, for detailed information see the phenotyping plan in the appendix 17.

Genomic selection models were created for four of the traits. Grain yield and protein yield were both scored in all trials, while yellow rust (*Puccinia striiformis*) resistance and head blight (*Fusarium spp*) resistance were scored in artificial inoculation trials in Austria.

3.3 Spatial analysis of field trials

The planning and analysis of the field trials was performed with PLABSTAT ((Utz, 2001)) using analysis of variance, where adjusted mean values of lines in the field trials were calculated for each location separately. Not all wheat lines were grown in each location and year, resulting in an unbalanced phenotype matrix. To account for this problem efficiently, the analysis of field trials was changed to linear mixed model analysis, since the analysis of such unbalanced data sets is a typical application of these models (Cnaan et al., 1997; Searle, 2012). Historically incomplete block designs were used to account for spatial variation of field trials, but as more sophisticated methods for analyzing spatial trends have evolved it has been found that block designs often failed to account for the multiple sources of the spatial variation (Gilmour et al., 1997).

A total of 14.500 single plots have been phenotyped and analyzed by estimating variance components of linear mixed models using ASREML following the strategy employed by the International Maize and Wheat Improvement Center (CIMMYT) (Burgueño et al., 2000). In the first step each field trial was analyzed separately to identify spatial trends while in the second step a multi-site-model of all field trials incorporating all spatial trends identified before was used to calculate a best linear unbiased prediction (BLUP) for each wheat line.

3.3.1 Identification of field variation

The phenomenon of spatial variation arises when field trial sites are not completely homogeneous or contain a large number of treatments so that the growing conditions

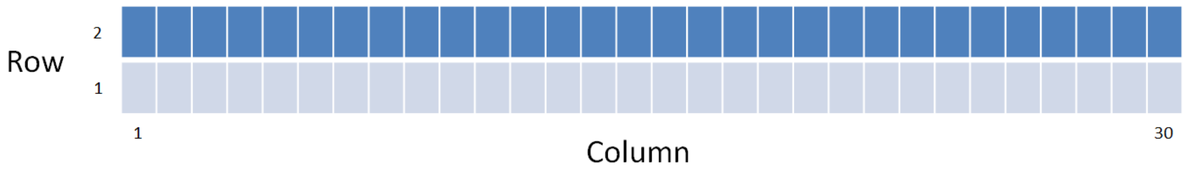


Figure 3.1: Trial layout of a PRECW trial

throughout the site may vary (Stringer et al., 2011). Usually plant breeding field trials are arranged in a rectangular shape of plots, e.g. the replicated pre-commercial set with 2 rows and 30 columns (see Figure 3.1) or the YOUNG set with 4 rows and 45 columns. Cullis and Gleeson (1991) extended the one-dimensional approach by Gleeson and Cullis (1987), who used a general neighbor model expressed as an autoregressive moving average model, to two dimensions and showed the potential gain of spatial designs over traditional row-column designs using a tobacco field experiment.

Later Gilmour et al. (1997) identified three major components of spatial variation

(i) Non stationary large scale (global) variation across the field;

(ii) Stationary variation within the field (natural variation or local trend);

(iii) Extraneous variation often induced by experimental procedures;

and provided a strategy to account for them by extending the two-dimensional approach of Cullis and Gleeson (1991). To account for global effects across the field, linear trends, splines, row and column contrasts and covariates were used. To account for local variation, e.g. uneven soil depth, the authors propose to use an autoregressive correlation structure, while the extraneous effects were modelled with random row (ROW) and column (COL) effects. To identify extraneous variation Gilmour et al. (1997) use a variogram, essentially the complement of the spatial autocorrelation matrix (Burgueño et al., 2000; Gilmour et al., 2009). Clearly visible patterns in the variogram indicate the presence of extraneous variation while otherwise the variogram is flat. Other tools used by the authors included plots of random effects including a trellis plot of residuals.

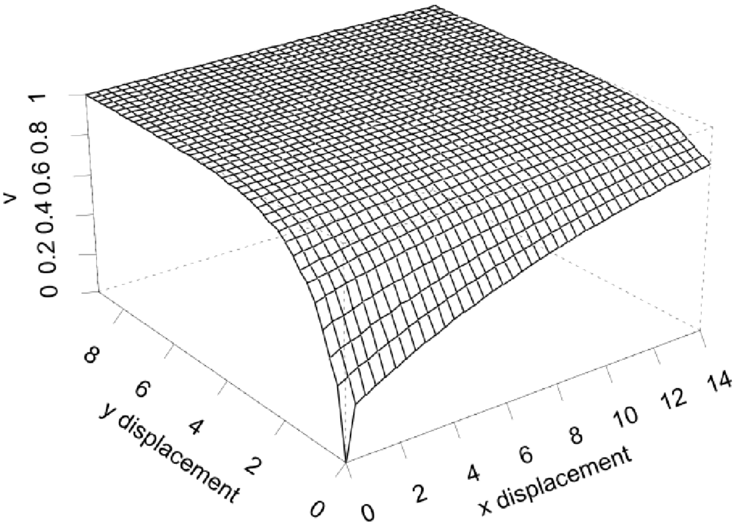


Figure 3.2: Variogram of a standardized AR1xAR1 process (Figure taken from (Gilmour et al., 1997))

The analysis of spatial trends of the individual field trials was done after (Burgueño et al., 2000) using the basic spatial model

$$y = mu + variety\ \ effects + global\ \ trend + design\ \ effects + error.$$

For each trial site, a two-dimensional spatial model was fitted as a basic model. This model included a first-order autoregressive correlation structure that is capable of representing a number of different spatial patterns. This structure models natural variation by the direct product of an autoregressive correlation structure for columns as well as rows, denoted by AR1xAR1. As suggested by Gilmour et al. (1997) the variogram was used to identify problems with the fit of the model. If it had the standard AR1xAR1 shape (see Figure 3.2) and no other problems were observable, the model was accepted. Otherwise, the model was adapted to account for the observed characteristics in the variogram. Extraneous variation, e.g. sowing directions or irrigation flow, were modelled as spatial covariates or function of the spatial coordinates using cubic splines. For each site this process was iterated until no obvious structure other than the standard AR1xAR1 structure was observed or the model could not be improved any further without over-fitting to the data. An example of a fitted spatial model for a given field trial could then be

$$y = mu + lin(COL) + GEN + ROW + spl(COL),$$

were additional to the basic AR1xAR1 model a linear fixed effect as well as random cubic smoothing spline was fitted for one dimension and a random effect for the other dimension. In each model the genotype (GEN) was included as a random effect to get correlations between the performances of the genotype at different sites in the multi-site analysis.

3.3.2 Multi-site analysis

In the second step of the field trial analysis the individual models of each single field trial were merged together into one single model (see Annex 19 for details). The genotype effects were treated as random to obtain best linear unbiased predictions (BLUP) compared to best linear unbiased estimates (BLUE) obtained by treating genotypes as fixed effects. Burgueño et al. (2000) noted that BLUE are the best estimate of the performance of a given genotype in a given trial, while BLUP estimates predict the future performance. Piepho et al. (2007) reviewed applications of BLUP estimates in plant breeding and variety testing and showed that these estimates had a good predictive ability.

3.3.3 Heritability filtering and field trial numbers

Individual field trials having a heritability for a given trait of less than 0.2 were discarded from the analysis. Of the 127 single field trials, 107 remained for grain yield and were used in the multi-site model. As protein content was not scored in every field trial the multi-site model comprised 39 field trials for this trait. Yellow rust (*Puccinia striiformis*) resistance and head blight (*Fusarium spp*) resistance were scored in artificial inoculation trials in Austria with five field trials each making up the multi-site analysis model.

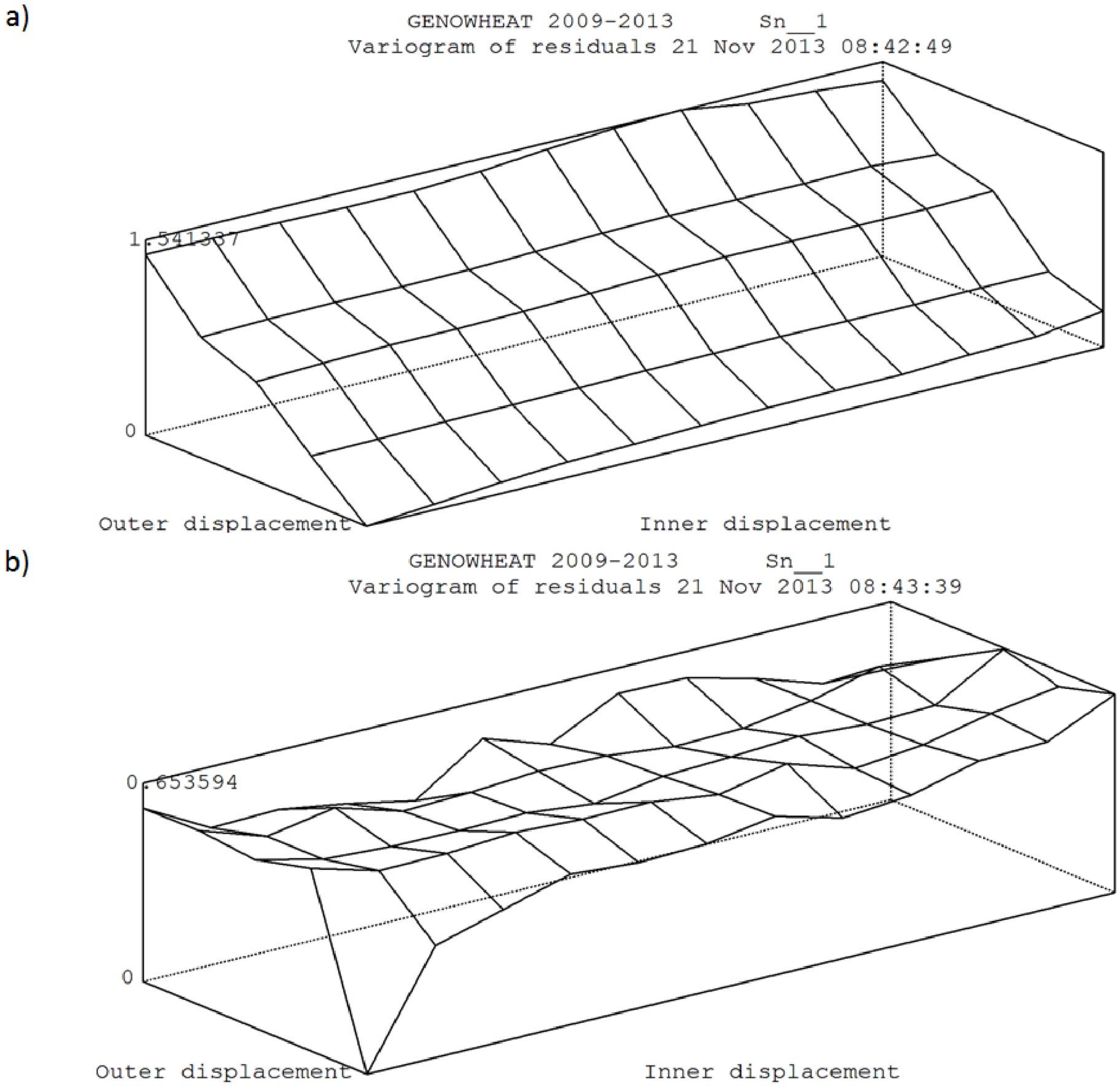


Figure 3.3: Sample variogram calculated before (a) and after (b) fitting the linear mixed model

3.3.4 Distribution of phenotypes

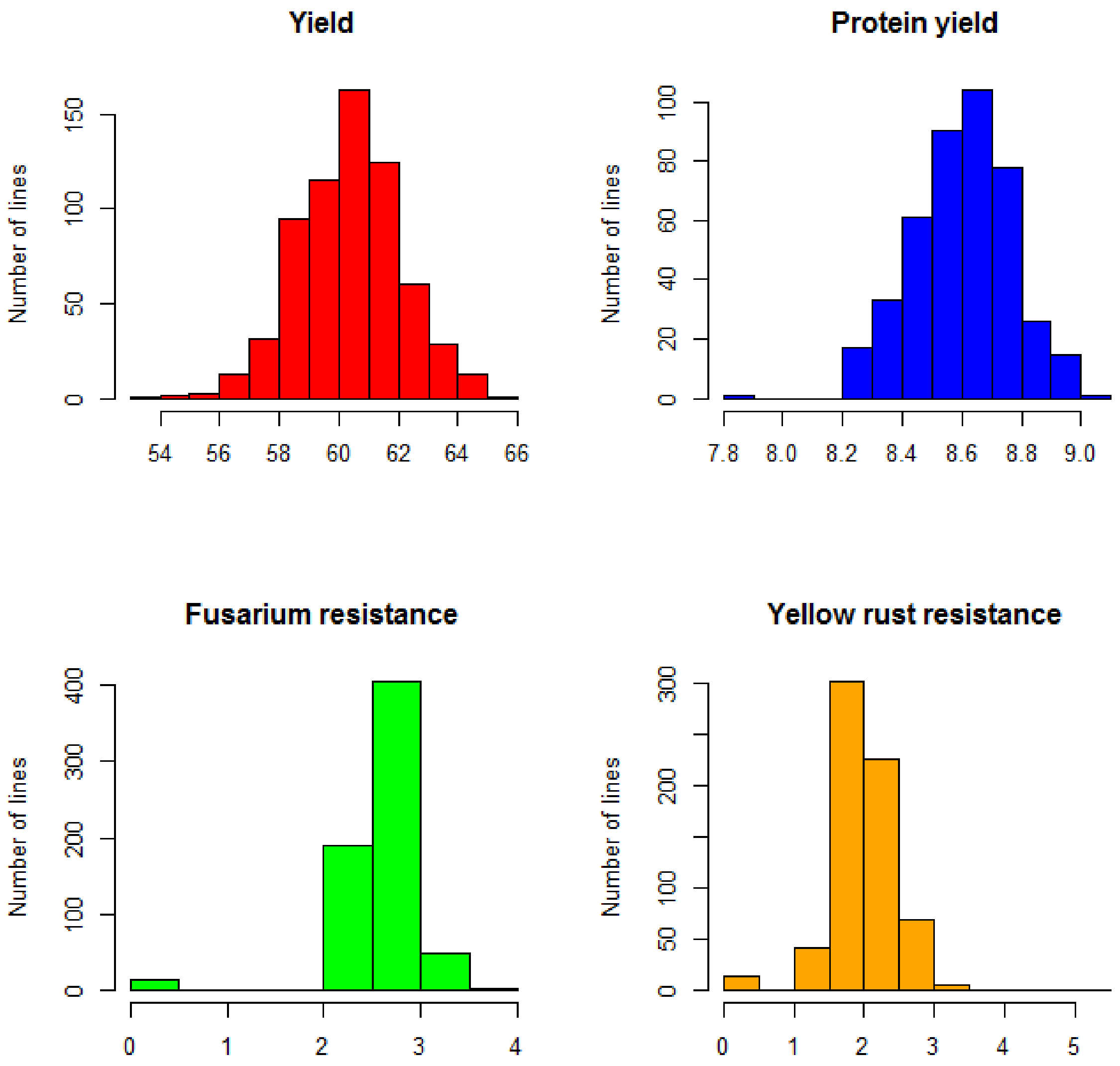


Figure 3.4: Phenotypic distribution of the traits used in the GS models

4 Genotyping

For genotyping of the breeding material a sequencing based approach was chosen for this project after reviewing several genotyping platforms including chip-based technologies. In total 1488 wheat lines were genotyped resulting in 15778 co-dominant markers which can identify heterozygous loci and about 45000 dominant markers scoring either absence or presence of the variant allele. Only the 15778 co-dominant markers were used since the addition of the dominant markers led to no significant increase in accuracy. A similar result was observed by Poland et al. (2012) who compared a set of 34739 genotyping by sequencing (GBS) markers to a set of 1827 GBS markers and found that the accuracy given by the smaller set was lower although not significantly. However in an outbred population with less population structure than the one at hand a higher marker number might lead to higher accuracy in prediction.

4.1 DNA extraction of leaf samples

Extraction of DNA from the wheat lines was done after Saghai-Maroof et al. (1984) by colleagues at our institute. Short samples were cut from the leaves of seedlings regrown from kept seed and dried at a temperature of 35°C for one or two days. 2 ml tubes were then filled with cut leaf material and five to seven glass beads and put into a Retsch Mill (Retsch, Haan, Germany) for grinding for ten minutes. 1000-1200 µl of CTAB buffer was then added to each tube following incubation for 60 to 90 minutes in a warm water bath at 65°C with gentle rocking. After cooling down to room temperature 500 to 600 µl of chloroform:isoamylalcohol (24:1) were added to each tube with subsequent gentle shaking by inversion for about five minutes. The tubes were then centrifuged for ten minutes at 10000 rcf to form two distinct phases. 700 to 800 µl, depending on the amount of CTAB, of the uppermost aqueous phase were pipetted into new 1.2 ml tubes. After adding of 300 µl of chloroform:isoamylalcohol (24:1) the samples were centrifuged for ten minutes at 3500 rcf. Again 400µl of the top aqueous phase were pipetted into new tubes containing 10 µl RNase A (1.2mg/ml), incubated at room temperature for eight minutes following mixing by gentle inversion and another incubation at room temperature for 30 minutes. 400 µl of isopropyl alcohol were then added, mixed by gentle inversion and centrifuged for eight minutes at about 600 rcf. After discarding the supernatant, 100 µl of a 76% ethanol/10 mM NaOAc solution were added, gently mixed for 5 minutes and centrifuged for eight minutes at about 600 rcf. After discarding the supernatant the DNA pellet was let to dry overnight and dissolved in 100 µl ddH_2O or 0,1 x TE buffer. Lastly, for complete dissolution of the DNA pellet the samples were mixed for a few hours following determination of DNA concentration.

4.2 DArT-seq

DArT-seq uses a genotyping by sequencing approach, representing a combination of previously used complexity reduction methods and next generation sequencing (Elshire et al., 2011; Kilian et al., 2012; Von Mark et al., 2013; Raman et al., 2014). It relies on sequencing of complexity reduced representations introduced by Altshuler et al. (2000) and more recent implementations using next generation sequencing (Baird et al., 2008; Elshire et al., 2011). Following steps were carried out by the genotyping provider Diversity Arrays Technology, Australia. For wheat, PstI-Hpall were chosen as restriction enzymes and DNA samples were processed for digestion and ligation reactions similarly to (Kilian et al., 2012). Following Raman et al. (2014), a single PstI-compatible adapter was replaced by two different adapters corresponding to two different restriction enzyme overhangs. This PstI-compatible adapter comprises of a varying length barcode region as well as a flowcell attachment sequence and a sequencing primer sequence. Likewise, the reverse adapter contains a flowcell attachment sequence and HpaII-compatible overhang sequence. Only mixed fragments (PstI-HpaII) were effectively amplified in 30 rounds of PCR using the following reaction conditions:

1. 94°C for one minute
2. 30 cycles of:
    a) 94°C for 20 seconds

   b) 58°C for 30 seconds

   c) 72°C for 45 seconds

3. 72°C for seven minute

Following PCR equal amounts of amplification products from each sample were bulked and applied to c-Bot bridge PCR followed by a sequencing run of 77 cycles.

The downstream analysis of the sequenced lanes was carried out by filtering out poor quality sequences from the fastq files where a more stringent filter criterion was emphasized on the barcode region compared to the rest of the read sequence. Per barcode/sample about two million sequences were used for marker calling.

4.2.1 Marker data pre-processing

Prior to the use in genomic selection models, quality filtering was applied to all 15778 co-dominant SNP markers. The genotyping report by the genotyping provider included several quality metrics such as allele frequency, average read count, call rate or reproducibility. Some markers were called with strong statistical support but had a low call rate as they represent just unique alleles. After discussion with the provider these markers were omitted for genomic selection since they would not be useful for prediction but for other analysis, e.g. haplotype analysis. Therefore only markers that had been called with a very high certainty (call rate > 0.9) were kept for further analysis.

The marker data was re-formatted using Visual Basic for Applications to a commonly used format where -1 and 1 denote homozygous marker alleles whereas 0 denotes heterozygous allele calls.

In the next step of quality filtering, markers with a minor allele frequency, i.e. the frequency of the least common allele present in the population, of less than 5% were discarded.

After quality filtering of the markers, the fraction of missing data points per line was calculated. Missing data points arise if one or more markers could not be called for a given line, e.g. due to sequencing errors. Only lines with less than 5% missing data points were kept fur further analysis. This cut-off is stringent but well within the range of 1 to 10% found in literature and only a few lines had more than 5% missing data.

Missing data were imputed for all remaining lines and filtered 7055 high quality markers using a multivariate normal expectation maximization (EM) algorithm developed by Poland et al. (2012). This kinship-based imputation algorithm assumes that the marker genotypes follow a multivariate normal distribution which is a working approximation with a realized relationship matrix in the context of breeding value prediction. All steps were performed using the free statistical software R and the rrBLUP package developed by Endelman et al. (R Development Core Team, 2008; Endelman, 2011).

4.2.2 Chromosomal mapping

Additional to the quality metrics the sequenced part of the DNA fragments following the barcoded adapter was included in the report. This 69bp sequence information was extracted using a custom Perl script for all available co-dominant markers. The resulting multi-fasta file consisted of marker id and sequence information for each marker. Colleagues of the Plant genome and systems biology department of the Helmholtz Zentrum München , Germany, blasted the multi-fasta file against the latest chromosome-arm sorted reference gene set by the International Wheat Genome Sequencing Consortium (IWGSC) (Mayer et al., 2014). This allowed for anchoring more than 50% of the markers to chromosome arms including their physical position on the chromosome. For further analysis only the marker ids and the chromosomal positions were kept.

4.3 Genetic relationships

Relationships among the wheat lines were estimated using the realized relationship matrix proposed by VanRaden (2008). The incidence matrix defined as $M \in [-1, 0, 1]^{nxm}$ specifies which allele each individual inherited with dimension for the number of individuals (n) and the number of loci (m). The cross-product of the incidence matrix with it's transposed, MM', has diagonal elements equal to the degree if the individual's relationship to itself, i.e. inbreeding, while off-diagonal elements equal the number of shared alleles by relatives (VanRaden, 2007). The centered genotype matrix Z is obtained by subtracting the marker mean from each data point: $Z_{ij} = M_{ij} - p_j$, where p_j is the frequency of the 1 allele (Poland et al., 2012). This subtraction gives more weight to rare alleles when calculating genomic relationships by setting the mean values of allele effects to zero. The genomic relationship matrix G can now be obtained by dividing through $2\sum_k p_k(1-p_k)$,

$$G = \frac{ZZ'}{2\sum_k p_k(1-p_k)}.$$

The scaling of ZZ' makes the genomic relationship matrix G analogous to the widely used numerator relationship matrix A (VanRaden, 2008). Genomic relationship coefficients were obtained by dividing elements off-diagonal elements G_{ij} by the square root of diagonal elements G_{ii} and G_{jj}. Hierarchical clustering of the genomic relationship matrix was done in R using the package gplots (Warnes et al., 2014).

4.4 Superior progeny values

Zhong and Jannink (2007) proposed the superior progeny value (SPV), a linear combination of the mean of the progeny of a cross and their standard deviation, to determine the value of a distinct cross. Endelman (2011) implemented this concept in the scope of genomic selection so that the marker effects determined by the genomic selection model can be used to estimate the mean GEBV of the cross as well as the variance of the population.

Let $P_1 = [-1, 1, 0, -1, -1, 0, 0, -1]$ and $P_2 = [0, 1, -1, 1, 1, -1, 1, 1]$ denote the genotypes of both parents. The genotype of the cross of those parents could be calculated as the arithmetic mean denoted as $C = [-0.5, 1, -0.5, 0, 0, -0.5, 0.5, 0]$. Given marker effects of $ME = [0.16, 0.86, 0.13, 0.40, 0.12, 0.83, 0.44, 0.82]$ and an additive selection model, the GEBV are

$$\begin{vmatrix} -1 & 1 & 0 & -1 & -1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\ -0.5 & 1 & -0.5 & 0 & 0 & -0.5 & 0.5 & 0 \end{vmatrix} \cdot \begin{vmatrix} 0.16 \\ 0.86 \\ 0.13 \\ 0.40 \\ 0.12 \\ 0.83 \\ 0.44 \\ 0.82 \end{vmatrix} = \begin{vmatrix} -0.64 \\ 1.68 \\ 0.52 \end{vmatrix},$$

where the estimated breeding value of the cross is the mean of the breeding values of the parents. The standard deviation can be calculated as the square root of the cross-product of $1 - C^2$ and the squared marker effects ME^2

$$\sqrt{\begin{vmatrix} 0.75 & 0 & 0.75 & 1 & 1 & 0.75 & 0.75 & 1 \end{vmatrix} \cdot \begin{vmatrix} 0.0256 \\ 0.7396 \\ 0.0169 \\ 0.1600 \\ 0.0144 \\ 0.6889 \\ 0.1936 \\ 0.6724 \end{vmatrix}} = 1.241189.$$

The superior progeny value for a selected population is depending on the selection intensity and is given by $\mu_{SPV} = \mu + i \quad$ (Endelman, 2011). For a selection intensity of 20% i equals 1.4 and the SPV of cross C is calculated as $0.52 + 1.4 * 1.241189 = 2.257665$.

5 Statistical models

Genomic estimated breeding values are often estimated in a two-step approach (Sun et al., 2012b). In the first step the effects of allele-substitutions at the marker loci are estimated using statistical models that use both the phenotypic and genotypic information of the training population. GEBV for individuals of the test population are then calculated by summing up the effects according to the individual's genotypic makeup. The accuracy of the predictions is assessed in the second step using the correlation between genomic predictions and true breeding values. Since these true breeding values usually remain unknown phenotypic information are used instead.

When Meuwissen et al. (2001) introduced the concept of genomic selection, the authors proposed four methods to estimate the effects of certain alleles at the marker loci. Since then a variety of methods have been proposed to overcome the large p small n paradigm when relating a larger number of genomic markers to a smaller number of phenotypes (Heffner et al., 2009; Moser et al., 2009; González-Camacho et al., 2012; Heslot et al., 2012; Neves et al., 2012; Sun et al., 2012a). Sun et al. (2012b) divided the set of methods into two classes. The first class consists of random SNP effects that are independent and identically distributed (i.i.d.) with the same effect variance and GEBV are estimated as BLUP from a linear mixed model, e.g. ridge regression BLUP or GBLUP. The second class consists of methods with prediction rules that are not necessarily linear in the observed phenotypes such as Bayesian models.

For comparison of the performance with the data at hand the following models were implemented in the free statistical software R (R Development Core Team, 2008). Accuracies were calculated as Pearson correlation between the estimated breeding values and phenotypes using repeated 10-fold cross-validation.

5.1 Ridge regression BLUP

One of the first methods proposed by Meuwissen et al. (2001) was BLUP estimation with linear mixed models. To tackle the problem of over-parametrization due to the large p small n paradigm, shrinkage factors can be used. Classical ordinary least squares (OLS) and maximum likelihood (ML) estimates are obtained by maximizing the fit of the model to the data, e.g. in the case of OLS solving the optimization problem $_{OLS} =_{argmin} \sum_i (y_i - \sum_j x_{ij} \; _j)$, where y_i is the phenotype of individual i; x_{ij} is the genotype of individual i at marker j and $_j$ is the effect of marker j (de los Campos et al., 2013). However, when the number of predictors p is large compared to the number of samples n, methods like OLS and ML can largely sample variance and consequently high mean-squared-errors. Ridge regression best linear unbiased predictor (RR-BLUP) is one approach using a penalization parameter λ that controls the trade-off between the fit of the model and the complexity. The estimator of marker effects becomes $argmin_b \{\sum_i (y_i - \sum_j x_{ij} \; _j) \; + \lambda \sum \; _y\}$. When $\lambda = 0$ the solution is the OLS solution of , while when $\lambda \to \infty$ the solution becomes 0. RR-BLUP is equivalent to BLUP estimation when the penalization parameter is defined accordingly (Hofheinz and Frisch, 2014). The basic ridge regression model after Endelman (2011) was implemented as

$$y = 1 \quad + Zu + e,$$

were y is the one-dimensional vector of length n corresponding to the phenotypes of the training population; is a fixed intercept, u is the one-dimensional vector of length m of marker effects; $Z = WG$ where G is the nxm-dimensional genotype matrix coded as {-1,0,1} for genotypes $\{A_1A_1, A_1A_2, A_2A_2\}$ and W is the design matrix relating genotypes to phenotypes and e is the one-dimensional vector of length n of residuals. Marker effects u and residuals e are expected to be independent $(cov(u_i, u_{j)}) = 0 \mid i \neq j$; $cov(e_k, e_{l)}) = 0 \mid k \neq l)$ and follow normal distributions with mean 0 and known variances ^2_m and ^2_e. Shrinkage of marker effects was controlled by the ridge parameter $\lambda = \frac{\sigma^2_e}{\sigma^2_m}$ and marker effects u where estimated as BLUP solutions

$$\begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} 1'1 & 1'Z \\ Z'1 & Z'Z + \lambda I \end{bmatrix}^{-1} \begin{bmatrix} 1'y \\ Z'y \end{bmatrix}.$$

RR-BLUP assumes a common penalization parameter and shrinks all markers equally towards zero (Heslot et al., 2012). Although all marker effects are constrained to the same variance the markers don't necessarily have the same effect (Bernardo and Yu, 2007). RR-BLUP keeps all marker which is favorable for traits that follow the infinitesimal model of quantitative genetics but incorrectly treats all effects equally (Hofheinz and Frisch, 2014; Heffner et al., 2009). However studies showed that the BLUP prediction performed well even for traits not supposed to follow the infinitesimal model, e.g. *Fusarium* resistance in barley by Lorenz et al. (2012) or by (Heffner et al., 2011) in wheat.

5.2 Bayesian Approaches

Ridge regression BLUP keeps all markers and assumes fixed and equal variance for all which can lead to over-shrinking of large effects. Bayesian approaches as proposed by Meuwissen et al. (2001) relax the assumption of equal variance and employ variable selection and shrinkage of marker effect estimates using different prior and posterior distributions. A prior distribution $p(\;)$ represents information about the parameters before the data are analyzed. After analysis of the data the prior distribution is updated with the gathered information and called posterior distribution which can be expressed as $\mathit{posterior} \propto \mathit{likelihood} \times \mathit{prior}$ or as $p(\;|y) \propto f(y|\;)p(\;)$ (Hamada et al., 2008). Prior and posterior distributions are together with the likelihood function the main components of Bayesian models.

In standard genomic regression phenotypes were regressed on marker covariates using a linear model such as

$$y_i = \quad + \sum_{j=1}^{p} x_{ij} \;_j + \;_i,$$

where y_i is the phenotype of individual i; µ is the intercept common to all individuals; x_{ij} is the genotype of individual i at marker j; $\;_j$ is the effect of marker j and is the vector of residuals. If the residuals are independent and identically distributed random variables following a normal distribution the likelihood function is

$$p(y|\;,\;,\;^2) = \prod_{i=1}^{n} N(y_i|\; + \sum_{j=1}^{p} x_{ij} \;_j, \;^2),$$

where $N(y_i|\; + \sum_{j=1}^{n} x_{ij} \;_j, \;^2)$ is a normal distribution with mean $\; + \sum_{j=1}^{n} x_{ij} \;_j$ and variance $\;^2$. In the Bayesian context, shrinking is controlled by the choice of the prior density. Without assumption of a prior density for shrinking, the general Bayesian linear model ((de los Campos et al., 2013)) is

$$p(\ , \ , \ ^2|y, \ ) \propto p(y|(\ , \ , \ ^2)p((\ , \ , \ ^2| \ ) \propto$$
$$\prod_{i=1}^{n} N(y_i| \ + \sum_{j=1}^{p} x_{ij} \ _j, \ ^2) \prod_{j=1}^{p} p(\ _j| \ )p(\ ^2).$$

Commonly the residual variance $p(\ ^2)$ is assigned a scaled inverse chi-square distribution, $^{-2}(v, S)$ with known parameters v, the number of degrees of freedom, and S, a scale parameter (Heffner et al., 2009). Assuming that the marker variances are independent random variables that following this distribution (de los Campos et al., 2009) the joint prior distribution could then be written as

$$p(\ , \ , \ ^2|v, S, \ ) \propto [\prod_{j=1}^{p} p(\ _j| \ _{\beta_j}, \ ^2)p(\ _{\beta_j}| \ )] \ ^{-2}(\ ^2|df, S).$$

$p(\ _j| \ _{\beta_j}, \ ^2)$ is the prior density of marker effect j and $p(\ _{\beta_j}| \ )$ is the prior density of the parameters given the hyper-parameters that controls the amount of shrinkage and if variable selection is performed. The posterior density given the data can be written as the product of the likelihood and the joint prior distribution as

$$p(\ , \ , \ ^2|y, v, S, \ ) \propto$$
$$\prod_{i=1}^{n} N(y_i| \ + \sum_{j=1}^{p} x_{ij} \ _j, \ ^2) \times [\prod_{j=1}^{p} p(\ _j| \ _{\beta_j}, \ ^2)p(\ _{\beta_j}| \ )] \ ^{-2}(\ ^2|df, S).$$

5.2.1 BayesA

The first method proposed by Meuwissen et al. (2001), termed BayesA, uses a mixture distribution sample marker effects from a normal distribution with mean 0 and marker specific variance $^2_{\beta_j}$. This allows for shrinking each marker effect back to 0 to a different extent (Lorenz et al., 2011). $^2_{\beta_j}$ is sampled from a scaled inverse chi-square distribution, $^{-2}(v, S)$ with degrees of freedom v, and the scale S (Heslot et al., 2012). The thick-tail prior distribution of BayesA can be written as ((de los Campos et al., 2009))

$$\begin{aligned} p(\ , \ ^2_\epsilon, \ ^2_\beta) &= p(\ | \ ^2_\beta)p(\ ^2_\epsilon)p(\ ^2_\beta) \\ &= [\prod_{j=1}^{p} N(\ _j 0, | \ ^2_{\beta_j})] \ ^{-2}(\ ^2_\epsilon|v, S) \times [\prod_{j=1}^{p} \ ^{-2}(\ ^2_{\beta_j}|v, S_b). \end{aligned}$$

5.2.2 BayesB

The second method proposed by Meuwissen et al. (2001), BayesB, allows for zeroing out marker effects by assuming that the effects are equal to 0 with probability and sampled as in BayesA from a normal distribution with 0 and variance ^2_j with probability$(1 - \ )$ to better model the underlying genetic architecture when many loci have no genetic variance. The authors derived from the ratio of the expected heterozygosity and the expected heterozygosity when a loci is segregating and set it to $= 0.947$. When $= 0$ the BayesB model becomes essentially BayesA (Heslot et al., 2012).

5.2.3 BayesCPi

Another Bayesian approach by Habier et al. (2011) treats the probability of a marker effect being 0 as unknown and estimates from the data to better model real organisms and traits. BayesCPi assumes common variance for all markers that remain in the model so that $^2_{\beta_j}= \ ^2_\beta$ and jointly estimates $^2_\beta$ over all non-zero markers (Lorenz et al., 2011).

A problem of the original Bayesian implementation is that there is no iterative solution for the parameter estimates. Instead Markov chain Monte Carlo (MCMC) algorithms such as Gibbs sampling ((Casella and George, 1992)) are used to sample parameter estimates from their posterior distributions which is computationally expensive (Lorenz et al., 2012).

5.3 Random forest

Random forest (RF) originated in the field of machine-learning and is quite different from linear approaches of RR-BLUP and the Bayesian methods (Heslot et al., 2012). A random forest is an ensemble of unpruned regression trees which are constructed from bootstrap samples, i.e. samples generated by random sampling with replacement, of the training data (Breiman, 2001; Ogutu et al., 2011). The trees are grown after following procedure

1. Randomly sample p variables with replacement from of the training data
2. At each node randomly select k out of the p remaining variables

3. Determine the best split given k
4. Repeat steps 2 and 3 to fully grow the tree without pruning

The prediction of the random forest is then calculated as the ensemble average of predictions over the trees (Poland et al., 2012). Due to the non-linearity of the model RF may be of advantage if non-additive effects account for a large amount of genetic variation (Jannink et al., 2010).

5.4 Support Vector Machines

Support vector machines (SVM) try to minimize model complexity and error simultaneously. The models map the input space, e.g. the marker data, to a feature space of different dimension to implement non-linear regression with linear models (Heslot et al., 2012). This is done using a non-linear kernel function followed by linear regression in the feature space (Jannink et al., 2010). The performance varies depending on the trait but is generally found to perform not as well as parametric methods (Jannink et al., 2010; Ogutu et al., 2011; Heslot et al., 2012).

5.5 Cross-validation procedure

To assess the accuracy of the model's ability to predict future outcomes cross-validation (CV) simulations were performed. Repeated 10-fold cross-validation was performed on a training set of 651 lines with genotypic and phenotypic information available using following steps.

1. Randomly split the training data into 10 disjunct partitions
2. Train the statistical model on 9 partitions, i.e. 586 wheat lines
3. Predict the GEBV of the remaining partition, i.e. 65 wheat lines
4. Determine the accuracy as the Pearson correlation between GEBV and phenotypes
5. Repeat steps 1 to 4 ten times so that each partition was once used for prediction of GEBV
6. Average the obtained accuracies

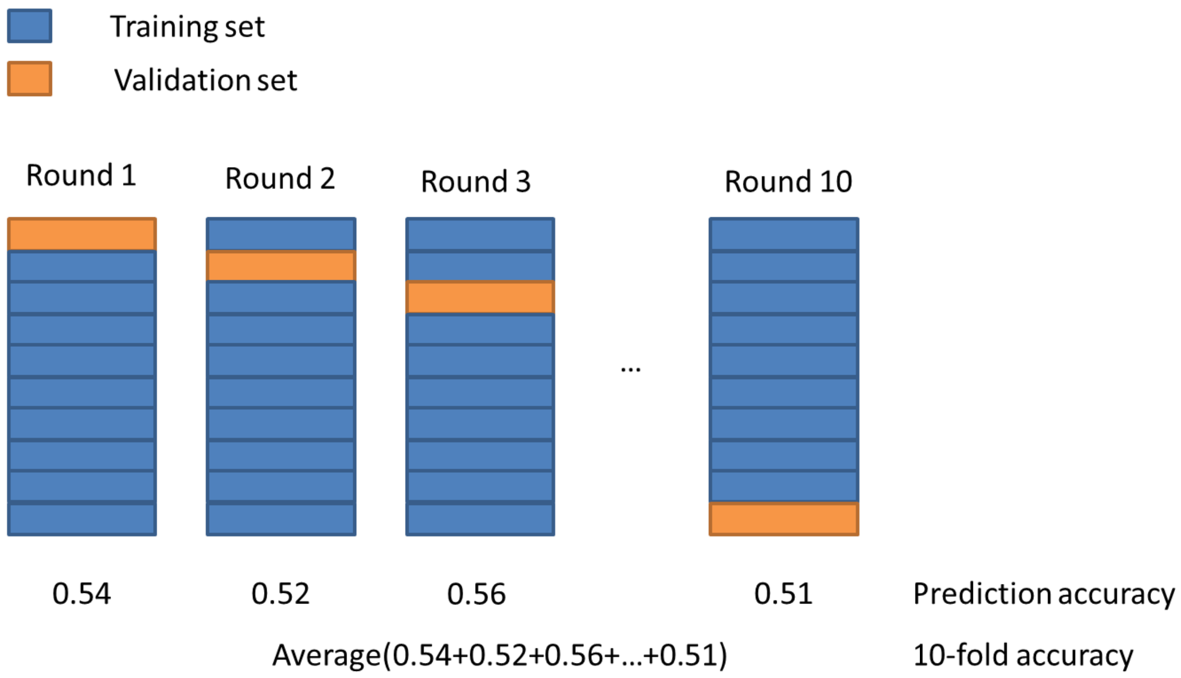


Figure 5.1: 10-fold cross validation procedure

This 10-fold CV was again repeated ten times and the obtained accuracies were averaged to reduce variance of the single 10-fold CV simulations.

6 Implementing GS in R

6.1 RR-BLUP

The basic ridge regression model was implemented using the package *rrBLUP* Endelman (2011).

RR-BLUP

```
### GS using RR-BLUP
### Accuracy determined using cross-validaton

library(rrBLUP)

### Set the working directory
setwd ("D:\\Studium\\BOKU\\GBS\\analysis\\August2013")

### Load the pre-processed training data
mydata <- as.data.frame(read.table("traindata28082013_YLD.csv",sep=";",
            fill=T,header=T,row.names=1,stringsAsFactors=F))
mydata <- data.matrix(mydata)

### Remove test data
hasPheno <- which(mydata[,1]>0)

### Extract phenotypes and genotypes
Y <- mydata[hasPheno,1]
X <- mydata[hasPheno,2:ncol(mydata)]

### RR-BLUP core function
### Returns the correlation between training and test data
ridgerBLUP <- function(traindata,trainmarker,testdata,testmarker)
{
    ans <- mixed.solve(y=traindata,Z=trainmarker)
    x <- testmarker
    a <- testdata
    aHat <- x %*% ans$u
    corr <- cor(a,aHat)
    return (corr)
}

### Cross-validation core function
### Returns the average accuracy after a given rounds of cross-validation
doCrossValidation <- function(crossvalnumber)
{
    cat (paste("Starting ",crossvalnumber,"-fold CV for rrBLUP with ",
            ceiling(nrow(X)/crossvalnumber*(crossvalnumber-1)),"  out of ",
            nrow(X)," as training data at ",date(),"\n",sep=""))
    results = vector(mode="numeric",length=crossvalnumber)
    testgroup <- (sample(1:nrow(X),nrow(X),replace=FALSE) %%
    crossvalnumber) + 1
    for (i in 1:crossvalnumber)
    {
        results[i] <- ridgerBLUP(Y[testgroup!=i],X[testgroup!=i,],Y[
        testgroup==i],X[testgroup==i,])
        cat(".")
    }
    cat (paste("\n","Finished ",crossvalnumber,"-fold CV for rrBLUP at ",
            date()," with rgs=",mean(results),"\n\n",sep=""))
    return (mean(results))
}

### Perform repeated cross-validation
stratnumber <- 10
stratresults = vector(mode="numeric",length=stratnumber)
cat (paste("Starting stratified CV for rrBLUP at ",date(),"\n",sep=""))
for (i in seq(1:stratnumber))
{
    stratresults[i] <- doCrossValidation(10)
}
cat (paste("\n","Finished ",stratnumber,"-fold stratificated CV for
rrBLUP at ",
        date()," with rgs=",mean(stratresults),"\n\n",sep=""))
```

6.2 Bayesian methods

All Bayesian approaches were run as single chains of 10000 iterations, of which the first 2000 discarded as burn-in, using custom scripts and the package *BGLR (de los Campos and Paulino Perez Rodriguez, 2014)*

BayesA

```
### GS using Bayes A
### Accuracy determined using cross-validaton

library(BGLR)

### Set the working directory
setwd ("D:\\Studium\\BOKU\\GBS\\analysis\\August2013")

### Load the pre-processed training data
mydata <- as.data.frame(read.table("traindata28082013_YLD.csv",sep=";",
            fill=T,header=T,row.names=1,stringsAsFactors=F))
mydata <- data.matrix(mydata)

### Remove test data
hasPheno <- which(mydata[,1]>0)

### Extract phenotypes and genotypes
Y <- mydata[hasPheno,1]
X <- mydata[hasPheno,2:ncol(mydata)]

### BayesA core function
### Returns the correlation between training and test data
bayesA <- function(traindata,trainmarker,testdata,testmarker,nIter=10000,
burnIn=2000)
{
    ETA <- list(list(X=trainmarker,model="BayesA"))
    ans <- BGLR(y=traindata ,ETA=ETA, nIter=nIter, burnIn=burnIn ,saveAt=
    "", verbose=FALSE)
    x <- testmarker
    a <- testdata
    aHat <- x %*% ans$ETA[[1]]$b
    corr <- cor(a,aHat)
    return (corr)
}

doCrossValidation <- function(crossvalnumber)
{
    cat (paste("Starting ",crossvalnumber,"-fold CV for BayesA with ",
            ceiling(nrow(X)/crossvalnumber*(crossvalnumber-1))," out of ",
            nrow(X)," as training data at ",date(),"\n",sep=""))
    results = vector(mode="numeric",length=crossvalnumber)
    testgroup <- (sample(1:nrow(X),nrow(X),replace=FALSE) %%
    crossvalnumber) + 1
    for (i in 1:crossvalnumber)
    {
        results[i] <- bayesA(Y[testgroup!=i],X[testgroup!=i,],Y[testgroup
        ==i],X[testgroup==i,])
        cat(".")
    }
    cat (paste("\n","Finished ",crossvalnumber,"-fold CV for BayesA at ",
            date()," with rgs=",mean(results),"\n\n",sep=""))
    return (mean(results))
}

### Perform repeated cross-validation
stratnumber <- 10
stratresults = vector(mode="numeric",length=stratnumber)
cat (paste("Starting stratified CV for BayesA at ",date(),"\n",sep=""))
for (i in seq(1:stratnumber))
{
    stratresults[i] <- doCrossValidation(10)
}
cat (paste("\n","Finished ",stratnumber,"-fold stratificated CV for
BayesA at ",
        date()," with rgs=",mean(stratresults),"\n\n",sep=""))
```

BayesB

```
### GS using Bayes B
### Accuracy determined using cross-validaton

library(BGLR)

### Set the working directory
setwd ("D:\\Studium\\BOKU\\GBS\\analysis\\August2013")

### Load the pre-processed training data
mydata <- as.data.frame(read.table("traindata28082013_YLD.csv",sep=";",
            fill=T,header=T,row.names=1,stringsAsFactors=F))
mydata <- data.matrix(mydata)

### Remove test data
hasPheno <- which(mydata[,1]>0)

### Extract phenotypes and genotypes
Y <- mydata[hasPheno,1]
X <- mydata[hasPheno,2:ncol(mydata)]

### BayesA core function
### Returns the correlation between training and test data
bayesB <- function(traindata,trainmarker,testdata,testmarker,nIter=10000,
burnIn=2000)
{
    ETA<-list(list(X=trainmarker,model="BayesB",probIn=0.05))
    ans <- BGLR(y=traindata,ETA=ETA, nIter=nIter, burnIn=burnIn,saveAt="",
    verbose=FALSE)
    x <- testmarker
    a <- testdata
    aHat <- x %*% ans$ETA[[1]]$b
    corr <- cor(a,aHat)
    return (corr)
}

doCrossValidation <- function(crossvalnumber)
{
    cat (paste("Starting ",crossvalnumber,"-fold CV for BayesB with ",
            ceiling(nrow(X)/crossvalnumber*(crossvalnumber-1)),"  out of ",
            nrow(X)," as training data at ",date(),"\n",sep=""))
    results = vector(mode="numeric",length=crossvalnumber)
    testgroup <- (sample(1:nrow(X),nrow(X),replace=FALSE) %%
    crossvalnumber) + 1
    for (i in 1:crossvalnumber)
    {
        results[i] <- bayesB(Y[testgroup!=i],X[testgroup!=i,],Y[testgroup
        ==i],X[testgroup==i,])
        cat(".")
    }
    cat (paste("\n","Finished ",crossvalnumber,"-fold CV for BayesB at ",
            date()," with rgs=",mean(results),"\n\n",sep=""))
    return (mean(results))
}

### Perform repeated cross-validation
stratnumber <- 10
stratresults = vector(mode="numeric",length=stratnumber)
cat (paste("Starting stratified CV for BayesB at ",date(),"\n",sep=""))
for (i in seq(1:stratnumber))
{
    stratresults[i] <- doCrossValidation(10)
}
cat (paste("\n","Finished ",stratnumber,"-fold stratificated CV for
BayesB at ",
        date()," with rgs=",mean(stratresults),"\n\n",sep=""))
```

BayesCPi

```
### GS using Bayes CPi
### Accuracy determined using cross-validaton

library(BGLR)

### Set the working directory
setwd ("D:\\Studium\\BOKU\\GBS\\analysis\\August2013")

### Load the pre-processed training data
mydata <- as.data.frame(read.table("traindata28082013_YLD.csv",sep=";",
            fill=T,header=T,row.names=1,stringsAsFactors=F))
mydata <- data.matrix(mydata)

### Remove test data
hasPheno <- which(mydata[,1]>0)

### Extract phenotypes and genotypes
Y <- mydata[hasPheno,1]
X <- mydata[hasPheno,2:ncol(mydata)]

### BayesA core function
### Returns the correlation between training and test data
bayesCPi <- function(traindata,trainmarker,testdata,testmarker,nIter=10000
,burnIn=2000)
{
    ETA<-list(list(X=trainmarker,model="BayesC"))
    ans <- BGLR(y=traindata,ETA=ETA, nIter=nIter, burnIn=burnIn,saveAt="",
    verbose=FALSE)
    x <- testmarker
    a <- testdata
    aHat <- x %*% ans$ETA[[1]]$b
    corr <- cor(a,aHat)
    return (corr)
}

doCrossValidation <- function(crossvalnumber)
{
    cat (paste("Starting ",crossvalnumber,"-fold CV for Bayes CPi with ",
            ceiling(nrow(X)/crossvalnumber*(crossvalnumber-1)),"  out of ",
            nrow(X)," as training data at ",date(),"\n",sep=""))
    results = vector(mode="numeric",length=crossvalnumber)
    testgroup <- (sample(1:nrow(X),nrow(X),replace=FALSE) %%
    crossvalnumber) + 1
    for (i in 1:crossvalnumber)
    {
        results[i] <- bayesCPi(Y[testgroup!=i],X[testgroup!=i,],Y[
        testgroup==i],X[testgroup==i,])
        cat(".")
    }
    cat (paste("\n","Finished ",crossvalnumber,"-fold CV for Bayes CPi at
    ",
            date()," with rgs=",mean(results),"\n\n",sep=""))
    return (mean(results))
}

### Perform repeated cross-validation
stratnumber <- 10
stratresults = vector(mode="numeric",length=stratnumber)
cat (paste("Starting stratified CV for Bayes CPi at ",date(),"\n",sep=""))
for (i in seq(1:stratnumber))
{
    stratresults[i] <- doCrossValidation(10)
}
cat (paste("\n","Finished ",stratnumber,"-fold stratificated CV for Bayes
CPi at ",
        date()," with rgs=",mean(stratresults),"\n\n",sep=""))
```

6.3 RandomForest

The RF method was implemented using the package *randomForest (Liaw and Wiener, 2002).* An ensemble of 1000 regression trees was grown with one third of the marker number sampled at each split and minimal node size of three.

Random forest

```
### GS using Random Forest
### Accuracy determined using cross-validaton

library(randomForest)

### Set the working directory
setwd ("D:\\Studium\\BOKU\\GBS\\analysis\\August2013")

### Load the pre-processed training data
mydata <- as.data.frame(read.table("traindata28082013_YLD.csv",sep=";",
            fill=T,header=T,row.names=1,stringsAsFactors=F))
mydata <- data.matrix(mydata)

### Remove test data
hasPheno <- which(mydata[,1]>0)

### Extract phenotypes and genotypes
Y <- mydata[hasPheno,1]
X <- mydata[hasPheno,2:ncol(mydata)]

### Random forest multiprocessor function
### Returns a random forest object
rFMulticore <- function (y, x, numtrees=500, cores=1, mtry = NULL,
nodesize = NULL) {
    require(randomForest, quietly=T)
    char2fac <- function (data)
    {
        coli <- colnames(data)
        for (cols in 1:length(coli)) {
            if(mode(data[,cols])=="character") data[,cols]=as.factor(data[
            ,cols])
        }
        return(data)
    }
    if(!class(y)=="numeric") y <- as.factor(y)
    x<-char2fac(x)

    ### Multiprocessor function
    if (!cores==1) {
      require(foreach, quietly=T)
      require (doSNOW, quietly =T)
      require(parallel, quietly = T)

      available_cores <-  detectCores()
      limit_cores <- ifelse(cores==0,Inf,cores)
```

```
   cores <- min(available_cores, limit_cores)
   cluster <- makeCluster(cores, type="SOCK")
   registerDoSNOW(cluster)
   core_trees <- ceiling(numtrees/cores)
   m1 <- foreach(ntree=rep(core_trees, cores), .combine=combine, .
   packages="randomForest") %dopar% randomForest(y=y, x=x, importance=T
   , keep.forest=T, do.trace=50, ntree=ntree, na.action=na.omit, mtry=
   mtry, nodesize=nodesize)
   stopCluster(cluster)
  }
  ### Singleprocessor function
  else {
   m1 <- randomForest(y=y, x=x, importance=T, keep.forest=T, do.trace=
   50, ntree=numtrees, na.action=na.omit, mtry=mtry, nodesize=nodesize)
  }

 return(m1)
}

### Random forest core function
### Returns the correlation between training and test data
calcRandomForest <- function(marker,traindata,testdata,startMarker,
nmarkers,ntree,mtry,nodesize,cores=1)
{
  #rf <-
  randomForest(marker,traindata,ntree=ntree,mtry=mtry,nodesize=nodesize)
  rf <- rFMulticore(x=marker,y=traindata,numtree=ntree,cores=cores,mtry=
  mtry,nodesize=nodesize)
  yhat <- predict(rf,testdata[,startMarker:nmarkers])
  y <- testdata[,nmarkers+1]
  corr <- cor(yhat,y)
  return (corr)
}

### Cross-validation core function
### Returns the average accuracy after a given rounds of cross-validation
doCrossValidation <- function(crossvalnumber)
{
  cat (paste("Starting ",crossvalnumber,"-fold CV for RF with ",ceiling(
  nrow(X)/crossvalnumber*(crossvalnumber-1))," out of ",nrow(X)," as
  training data at ",date(),"\n",sep=""))
  results = vector(mode="numeric",length=crossvalnumber)
  testgroup <- (sample(1:nrow(X),nrow(X),replace=FALSE) %%
  crossvalnumber) + 1
  for (i in 1:crossvalnumber)
```

```
    {
        results[i] <- calcRandomForest(X[testgroup!=i,],Y[testgroup!=i],
        cbind(X[testgroup==i,],Y[testgroup==i]),1,ncol(X),1000,floor(ncol(
        X)/3),3,cores=8)
        cat(".")
    }
    cat (paste("\n","Finished ",crossvalnumber,"-fold CV for RF at ",date
    ()," with rgs=",mean(results),"\n\n",sep=""))
    return (mean(results))
}

### Perform repeated cross-validation
stratnumber <- 10
stratresults = vector(mode="numeric",length=stratnumber)
for (i in seq(1:stratnumber))
{
    stratresults[i] <- doCrossValidation(10)
}
cat (paste("\n","Finished",stratnumber,"-fold stratificated CV for RF at "
,date()," with rgs=",mean(stratresults),"\n\n",sep=""))
```

6.4 Support Vector Machines

SVM were implemented using the package *e1071* using a linear kernel with an epsilon-intensive loss function of 0.1(Meyer et al., 2014).

Support vector machines

```
### GS using Support vector machines
### Accuracy determined using cross-validaton

library(e1071)

### Set the working directory
setwd ("D:\\Studium\\BOKU\\GBS\\analysis\\August2013")

### Load the pre-processed training data
mydata <- as.data.frame(read.table("traindata28082013_YLD.csv",sep=";",
            fill=T,header=T,row.names=1,stringsAsFactors=F))
mydata <- data.matrix(mydata)

### Remove test data
hasPheno <- which(mydata[,1]>0)

### Extract phenotypes and genotypes
Y <- mydata[hasPheno,1]
X <- mydata[hasPheno,2:ncol(mydata)]


### SVM core function
### Returns the correlation between training and test data
calcSVM <- function(marker,traindata,testdata,startMarker,nmarkers)
{
    suppvm <- svm(x=marker,y=traindata,kernel="linear",cross=10,cost=0.001
    , epsilon=0.1) #Ogutu 2011, A comparison
    yhat <- predict(suppvm,testdata[,startMarker:nmarkers])
    y <- testdata[,nmarkers+1]
    corr <- cor(yhat,y)
    return (corr)
}

### Cross-validation core function
### Returns the average accuracy after a given rounds of cross-validation
doCrossValidation <- function(crossvalnumber)
{
    cat (paste("Starting ",crossvalnumber,"-fold CV for SVM with ",ceiling
    (nrow(X)/crossvalnumber*(crossvalnumber-1))," out of ",nrow(X)," as
    training data at ",date(),"\n",sep=""))
    results = vector(mode="numeric",length=crossvalnumber)
    testgroup <- (sample(1:nrow(X),nrow(X),replace=FALSE) %%
    crossvalnumber) + 1
    for (i in 1:crossvalnumber)
    {
        results[i] <- calcSVM(X[testgroup!=i,],Y[testgroup!=i],cbind(X[
        testgroup==i,],Y[testgroup==i]),1,ncol(X))
        cat(".")
    }
    cat (paste("\n","Finished ",crossvalnumber,"-fold CV for SVM at ",date
    ()," with rgs=",mean(results),"\n\n",sep=""))
    return (mean(results))
}

### Perform repeated cross-validation
stratnumber <- 10
stratresults = vector(mode="numeric",length=stratnumber)
cat (paste("Starting stratified CV for SVM at ",date(),"\n",sep=""))
for (i in seq(1:stratnumber))
{
    stratresults[i] <- doCrossValidation(10)
}
cat (paste("\n","Finished ",stratnumber,"-fold stratificated CV for SVM
at ",date()," with rgs=",mean(stratresults),"\n\n",sep=""))
```

Part III

Results

7 Genotyping-by-sequencing data

Genotyping of wheat lines was performed in six batches totaling in 1488 unique samples. To validate the technical reproducibility, previously extracted DNA from already genotyped samples was sent again for genotyping in subsequent batches. The estimated technical error was acceptable with around 1% diverging markers of which only less than ten markers per line showed a change in homozygous allele calls. After data pre-processing a total of 1488 wheat lines together genotyped with 7055 co-dominant SNP markers was available for data analysis. Additionally, 45.000 dominant markers scoring either absence or presence of the variant allele were available but kept from analysis since the addition of the dominant markers led to no significant increase in prediction accuracy but increased the computational time drastically.

7.1 Statistics on marker data

The final marker matrix was denoted by $X \in [-1, 0, 1]^{nxm}$ with $n = 1448$ wheat lines and $m = 7055$ markers and alleles scored as -1 or 1 when being homozygous while 0 denotes heterozygous loci. The average heterozygosity rate per line was 15.8% which is lower than the expected 25% heterozygosity of F3 plants while being higher than the expected 12.5% heterozygosity rate of F4 plants. A possible explanation could be that heterozygous alleles were called with less certainty and were therefore filtered out during marker pre-processing. Figure 7.1 shows the distribution of all 10.215.640 allele calls of which only 2.5% were missing data points and had to be imputed. Without quality filtering the fraction of missing data points would be 20%.

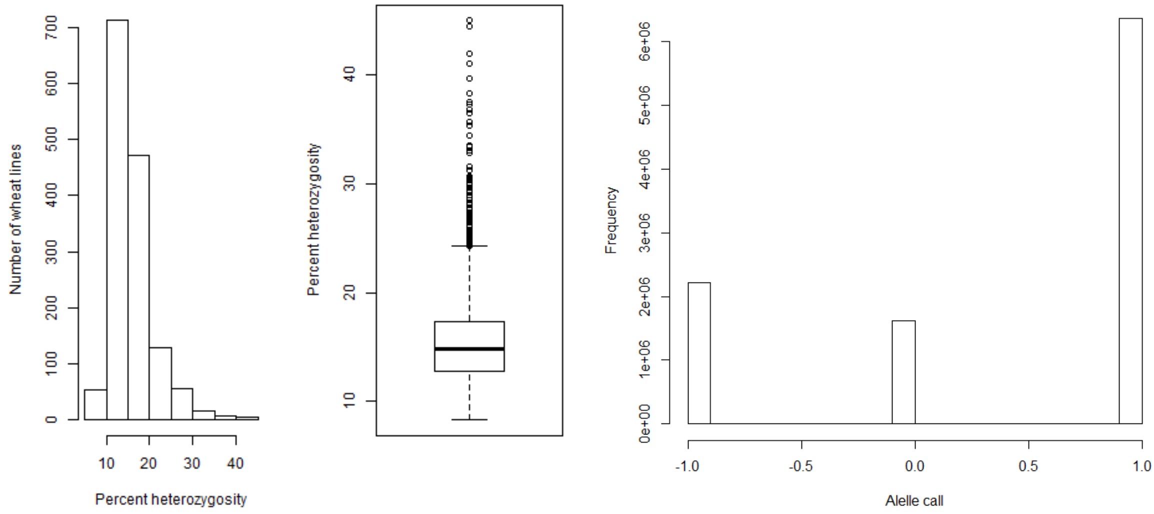


Figure 7.1: Distribution of heterozygous marker and allele calls

7.2 Evaluating the coverage of the SNP markers

To evaluate the coverage of the SNP markers over the bread wheat genome the idea was to use the provided short sequence information of the markers to infer their chromosomal position. However during the first year of the project there was no complete bread wheat reference genome available. So in the first step the draft genome of the wild diploid grass *Aegilops tauschii* (Jia et al., 2013), the donor of D-genome, was used as a mapping reference. The sequence information of the co-dominant SNP markers was mapped against the draft genome to obtain chromosomal position for about 2800 markers. Chromosome 4D was known to have less variability which was reflected in a lower number of mapped markers (see Figure 7.2). In general the markers were evenly distributed over all chromosomes.

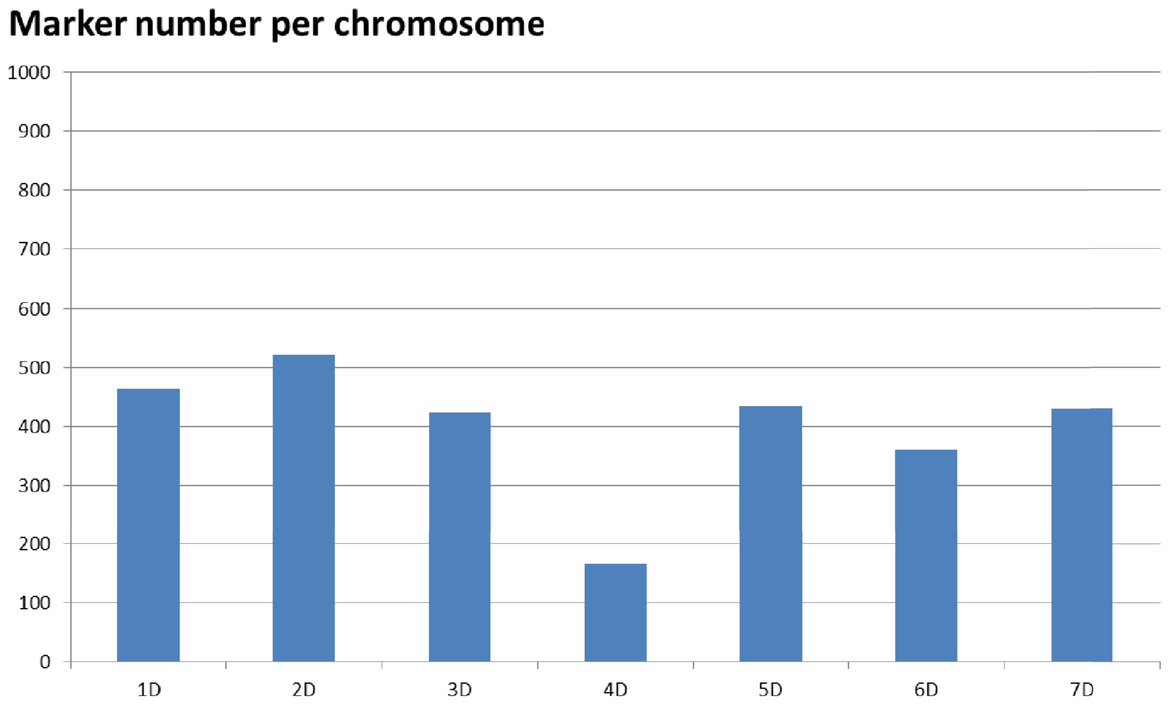


Figure 7.2: Mapping of the co-dominant SNP markers on the draft genome of *Aegilops tauschii*

The release of the IWGSC bread wheat reference genome made it later possible to evaluate the position of the SNP markers on a chromosome arm level on all A,B and D sub-genomes. Using the same approach more than 8500 markers were anchored onto chromosome arms (see Figure 7.3). When neglecting the position on the chromosome arms the distribution over the IWGSC genome looked similar to the distribution over the *Aegilops tauschii* genome. The least number of markers was anchored on chromosome 4 while chromosome arm 3b was by far the chromosome arm with the highest number of anchored markers. This distinct peak might be due to the intensive work put into sequencing of this chromosome arm.

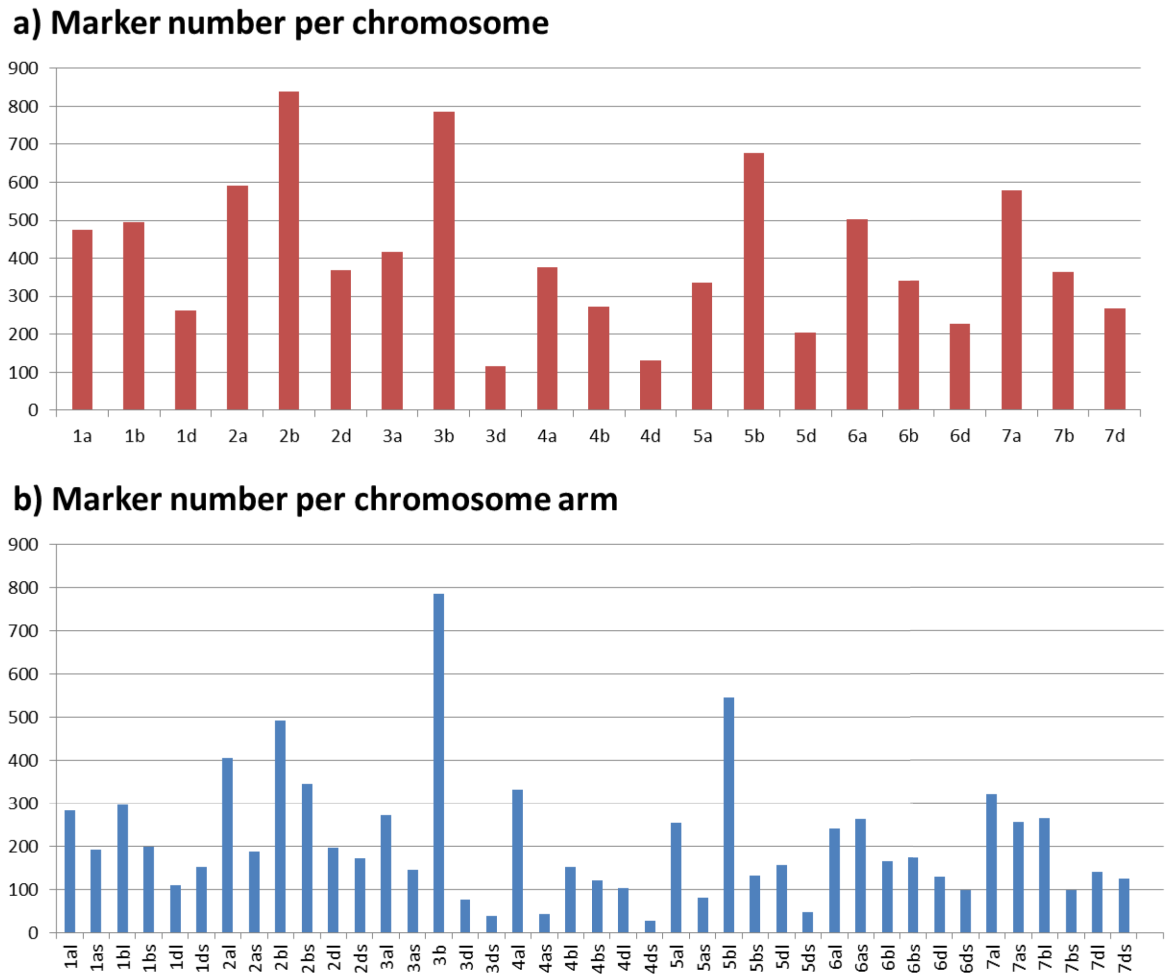


Figure 7.3: Mapping of the SNP markers to the IWGSC wheat reference genome

8 Population structure

For genomic selection, a set of dense molecular markers covering all chromosomes is required to provide accurate estimates of breeding values. Consequently, the first task of exploring the breeding population was to investigate the distribution of the SNP markers over the bread wheat genome. Another question was whether the observed phenotypic variation in the desired trait was caused by genetic or by other factors, e.g. environmental factors including stress or disease pressure. Statistical identification and quantification of the genetic variation based on the observed phenotype was routinely performed as part of the ongoing breeding program. However quantification of the genetic variation alone might not be enough for a constant improvement of elite breeding populations and difficulties arise when introducing new genetic material with undesirable alleles. Therefore the second task was to estimate the genetic relationship of the breeding lines to gain insights about the structure of the breeding population and to assist in resolving issues regarding the discriminability and homogeneity of breeding lines as well as in the choice of parental lines for crossing to generate new genetic variability.

8.1 Genetic relationships

Kinship between breeding lines plays an important role in the prediction of breeding values. An ideal GS model should be based on marker - gene or marker - QTL linkage disequilibrium rather than on kinship. If the breeding population consists of different sub-populations or families which are more related to each other than to the rest of the breeding population, genomic predictions of breeding values may be confounded as within-family predictions provide more accurate estimates than predictions over different groups of families. It is also vital to avoid allele fixation and loss of genetic variability by choosing distantly related breeding lines as parents for crossing to ensure long-term efficiency of GS.

To investigate grouping of families, the widely used realized relationship matrix, that is the cross-product of the marker matrix with its transposed, was calculated. This matrix was centered and scaled to give rare alleles more weight. The scaling also let the mean of the diagonal elements of the matrix equal $1 + f$, where f is the population's inbreeding coefficient. The population structure was first inspected by calculating the eigenvectors of the realized relationship matrix. Figure 8.1 shows 2D and 3D plots of the first two and first three eigenvectors. Families or sub-populations would show up as clusters of adjacent data points. No obvious separation of data points into distinct clusters is visible using either the first two eigenvector dimensions or the first three dimensions. This indicates that there is no population structure and the breeding population is relatively well balanced..

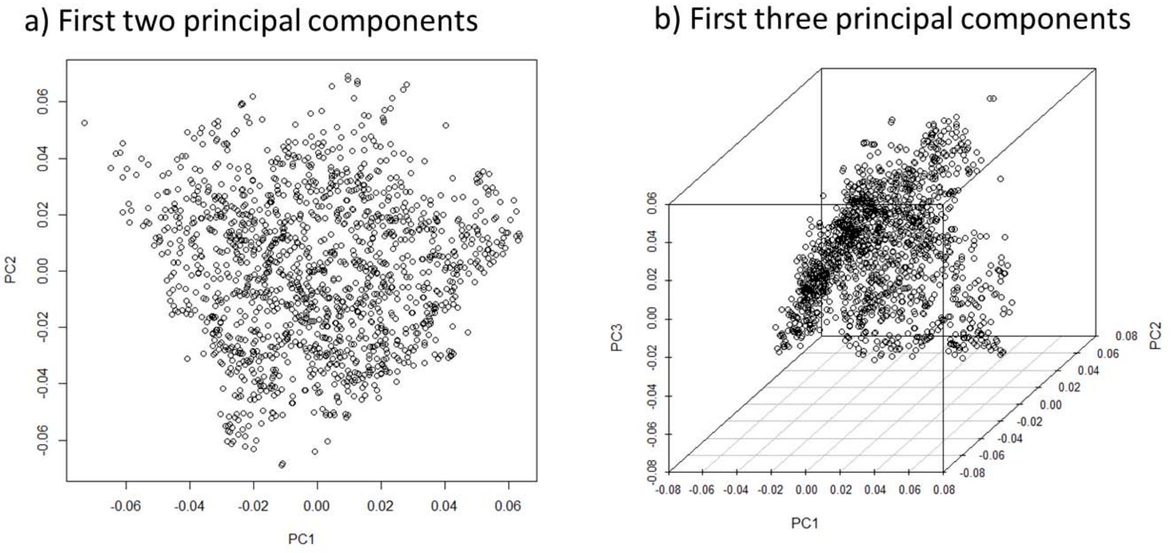


Figure 8.1: Eigenvector analysis of the realized relationship matrix

Pedigree information reaching back several generations was available for a large number of wheat lines. Having a closer look at the kinship of the lines in the breeding population led to the believe that some larger families or sub-populations should have been present. So to further analyze the relationship between breeding lines, the matrix of the first two eigenvectors was investigated using k-means clustering. Figure 8.2a depicts the partitioning around medoids for 1 to 15 possible clusters. The bending of the within cluster sum of squares curve indicated the presence of around 3 to 4 sub-populations. This was supported by the the number of clusters estimated by optimum average silhouette in Figure 8.2b and Figure 8.2c.

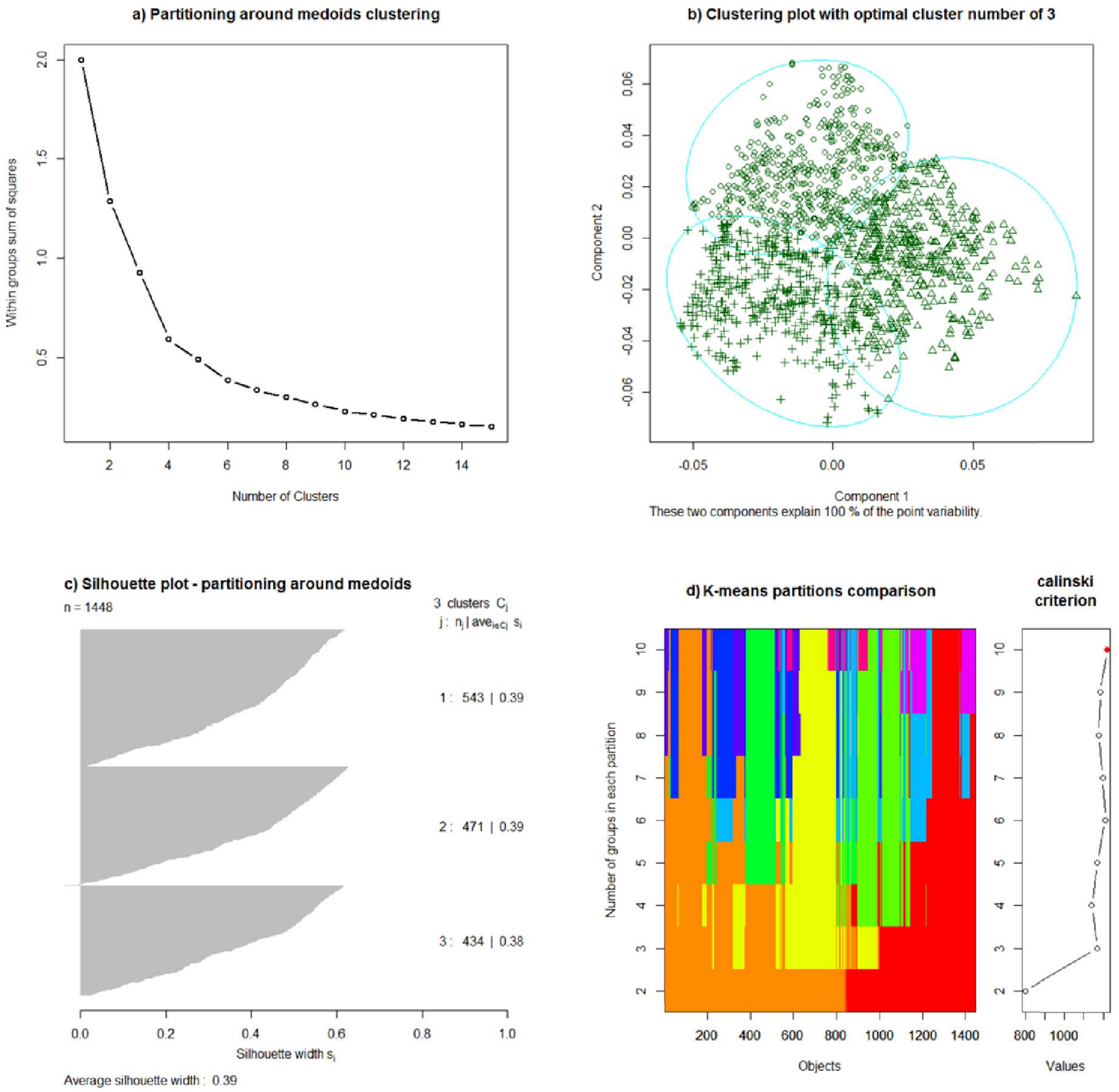


Figure 8.2: Cluster plots of the breeding population

Using the Calinski-Harabasz criterion (Calinski and Harabasz, 1974), essentially an F statistic on the ratio of sum of squares within a cluster by the sum of squares among the clusters, to create several partitions using a range of k values for the k-means clustering (see Figure 8.2d) resulted in 10 sub-populations. The difference in the criteria at 3 and

10 clusters was not too large to reject the number of clusters being 3, so the number of possible sub-populations present in the breeding populations should fall into the range from 3 to 10.

A heatmap of the realized relationship matrix gave similar results. The dendrogram at the top of Figure 8.3 indicates two larger clusters, the first located at the bottom-left of the heatmap, which is more separated from the others. The second cluster is split again into two larger clusters, thereby supporting the number of sub-populations being 3 indeed, but there are a also a number of sub-clusters present that could favor the number of sub-populations being around 10. Prediction accuracies for breeding lines within one of the two clusters were not significantly different from prediction accuracies for other breeding lines again indicating that the breeding population is generally well balances.

Table 8.1: Lines sorted by their relationship coefficients

	Line 1			**Line 2**
Line 1	1		Line 2	1
Parent 1	0.483		Sister line 2	0.593
Line 23	0.482		**Parent 1**	0.583
Parent 2	0.472		**Parent 2**	0.472
Line 48	0.464		Line 82	0.458
Line 37	0.377		Line 64	0.454
Line 46	0.375		Line 55	0.436
Line 50	0.365		Line 49	0.429

Another interesting property of the realized relationship matrix is that it captures the mendelian sampling effect. By calculating the relationship coefficients of Wright it was possible to easily rank and interpret the relationship among the breeding lines. Using this spreadsheet the breeder could easily distinguish between parental and sister lines as well detect possible selfings (see Table 8.1).

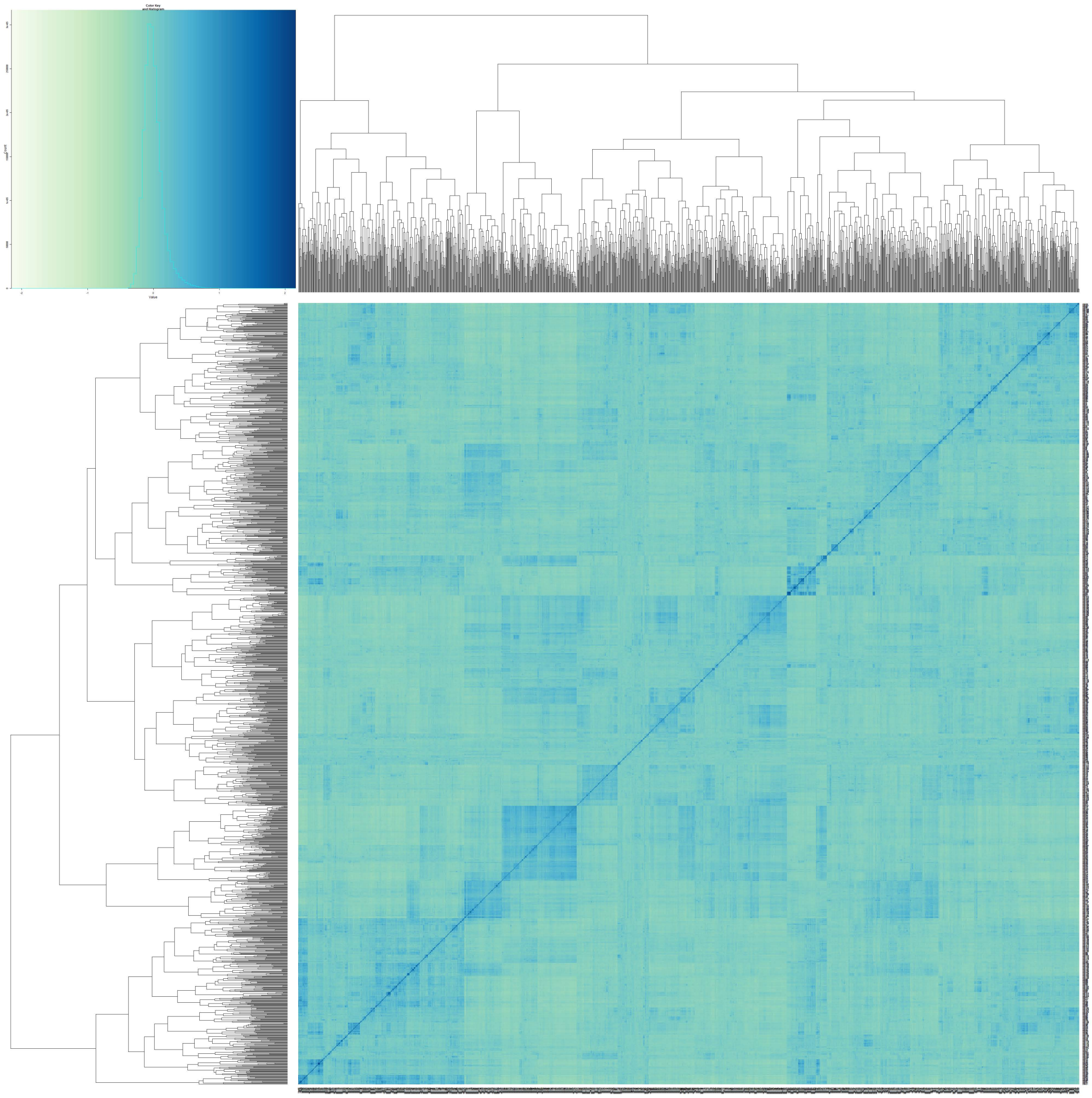


Figure 8.3: Heatmap of the realized relationship matrix

9 Evaluation of prediction methods

There are two approaches of assessing the accuracy of GS found in published literature. Common to both methods is the use of the correlation between observed phenotypes and genomic estimated breeding values to determine the accuracy of the methods. In some publications this correlation is then divided by the square root of the heritability, e.g. Dekkers (2007); Lorenzana and Bernardo (2009):

$$\begin{aligned} r_{G,T} &= \frac{r_{G,O}}{r_{O,T}} \\ &= \frac{r_{G,O}}{\sqrt{h}}, \end{aligned}$$

where $r_{G,T}$ is the correlation of the genomic estimated breeding values and the unknown true breeding values, $r_{G,O}$ is the correlation of the genomic estimated breeding values and the observed phenotypic values and $r_{O,T}$ is the correlation of the observed phenotypic values with the unknown true breeding values. Dividing by the square root of the heritability accounts for the unknown true breeding values and relates the efficiency of GS with the efficiency of traditional phenotypic selection. Since the heritability per trait is identical for all GS methods this scaling would not have any influence on the relative performance of the methods. Therefore the accuracy is estimated as the Pearson correlation between observed phenotypic values and predicted genomic breeding values without adjusting for the heritability.

9.1 Cross validation accuracy

The performance of the GS methods was evaluated in the second year of the project where a subset of 651 wheat lines had already been phenotyped and genotyped. Of the four traits only grain yield had been recorded in all locations and all years so it was chosen as the phenotypic trait that was correlated with the genomic estimated breeding values. To reduce the sampling bias the 10-fold cross-validation procedure was repeated again ten times and the results of the single cycles were averaged. One cycle of cross-validation was implemented using following steps:

1. The 651 wheat lines were randomly partitioned into 10 equally large disjunct subsets
2. 9 subsets were used as the training set to estimate the model parameters
3. The trained model was used to predict the breeding values for the remaining subset
4. The accuracy of the prediction was calculated as the Pearson correlation between predictions and available phenotypes
5. Steps 1-4 were repeated in a way that each subset was used 9 times for training of the GS model and 1 time for prediction of the breeding values
6. The obtained accuracies were averaged and reported

9.1.1 Results

Figure 9.1 depicts the performance of the different statistical methods on the training data set using the repeated cross-validation method. Support vector machines were significantly inferior to the other methods. The achieved cross-validation accuracies ranged from 0.415 to 0.467 with a median of 0.450. All tested Bayesian methods performed similarly with achieved accuracies ranging from 0.462 to 0.498 with no significant difference between their medians. Although BayesCPi had the lowest median of all three Bayesian methods, the single cross-validation cycles showed less variability than the other two methods. As expected from published literature, RR-BLUP performed similarly well than as Bayesian methods. Of all tested methods, RandomForest performed best. Although the minimum with 0.464 was in the same range as the Bayesian methods and RR-BLUP, about 8 of 10 cross-validation cycles achieved a higher accuracy than the average accuracy achieved by the other methods.

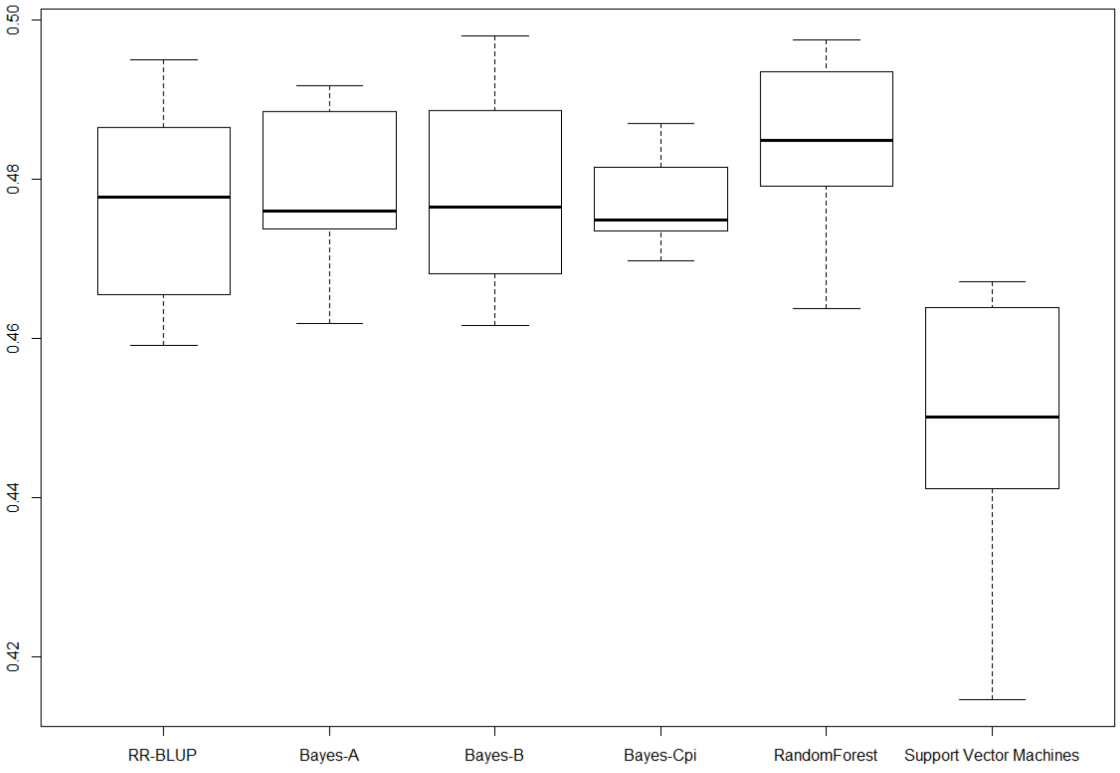


Figure 9.1: Performance of the statistical methods

9.1.2 Comparison of the different statistical methods

Comparing the predicted values for the wheat lines, it could be clearly seen, that the predictions of different methods were highly correlated (see Table 9.1).

Table 9.1: Correlation of predicted breeding values between the statistical methods

	RR-BLUP	BayesA	BayesB	BayesCPi	RF	SVM
RR-BLUP	1	0.9993299	0.9988789	0.9717142	0.8760442	0.9464235
BayesA	0.9993299	1	0.9983623	0.9722375	0.8753431	0.9457349
BayesB	0.9988789	0.9983623	1	0.9776306	0.8685065	0.9518308
BayesCPi	0.9717142	0.9722375	0.9776306	1	0.7987837	0.9572482
RF	0.8760442	0.8753431	0.8685065	0.7987837	1	0.7842741
SVM	0.9464235	0.9457349	0.9518308	0.9572482	0.7842741	1

RR-BLUP as well as the Bayesian methods had a Pearson correlation of at least 0.97. The linear kernel of the SVM could explain the relatively high correlation with the Bayesian methods and RR-BLUP. The GEBV predicted by RandomForest however were more distinct to the other predictions can nicely be seen in in the hierarchical clustering in Figure 9.2. Very similar results were obtained by Heslot et al. (2012).

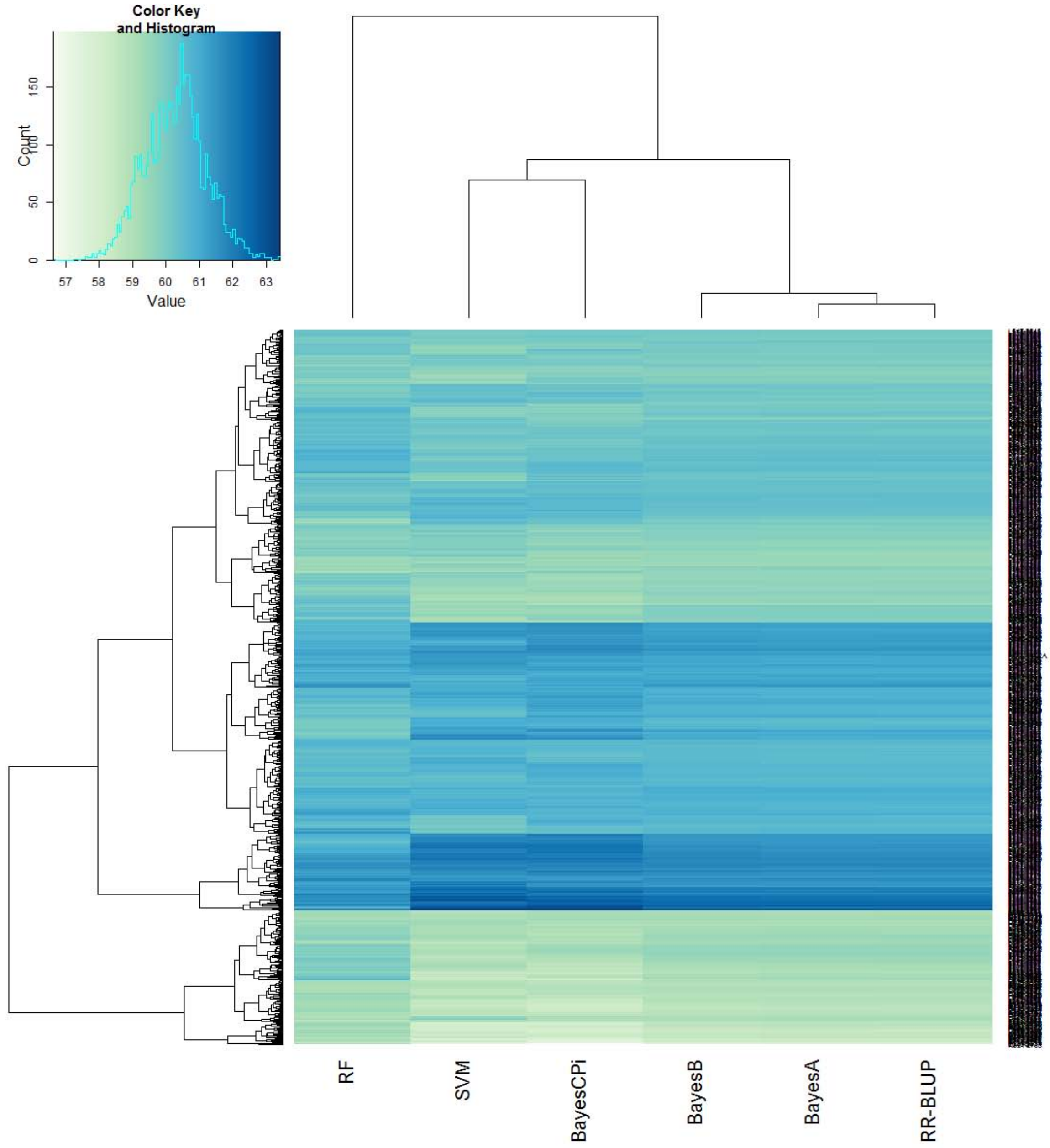


Figure 9.2: Heatmap of the predicted GEBV of the statistical methods

9.2 Computation time

The optimal statistical GS method should provide accurate predictions calculated in reasonable time. The computation complexity increases drastically with increasing number of markers and wheat lines in the data set. To have an estimate of the computation time required by the different statistical methods, the time to perform 10 cycles of 10-fold cross-validation with a training set of 651 wheat lines and 7055 SNP markers was measured. The used workstation was equipped with an Intel Core i7 3770S CPU with a peak frequency of 3.90GHz. The CPU had 4 physical cores with HyperThreading technology, hence 8 logical cores. The workstation was equipped with 16GB DDR3-1333 memory and a standard hard-drive.

Table 9.2: Computational requirements for different statistical models

Computation time for 10 cycles of 10-fold cross-validation						
	RR-BLUP	BayesA	BayesB	BayesCPi	RF	SVM
Mean accuracy	0.477	0.478	0.479	0.476	0.485	0.449
Computation time	00:07:26	03:00:46	03:37:23	03:05:49	03:01:38	01:53:13

Table 9.2 shows a striking performance advantage of RR-BLUP coupled with competitive achieved prediction accuracy. While support vector machines only need about half the time of the Bayesian methods and RandomForest, the prediction accuracy was significantly inferior to those methods. There was no significant difference in the mean accuracy between RR-BLUP, the Bayesian methods and RandomForest. Taken together with the replicable predictions accuracies shown in the single cross-validation cycles in Figure 9.1 it became clear why many workgroups chose RR-BLUP or GBLUP as the method of choice.

9.3 Prediction ranks

To compare the predicted values of the different statistical methods, the 100 wheat lines with the highest predicted breeding value per method were selected, since a breeder would normally be interested in the highest ranked wheat lines. When looking at the Venn diagram in Figure 9.3 it can be seen that 59 out of the 100 top ranked lines were shared by all methods. Each line selected by the RR-BLUP predictions was also selected by at least on of the other methods, where on the other hand 28 lines selected by RandomForest would not have been chosen by any other model. Similarly to the high correlations of all predicted breeding values there was a nice agreement on the highest predicted lines between RR-BLUP and BayesA. Support vector machines were more similar to RR-BLUP and BayesA than to RandomForest.

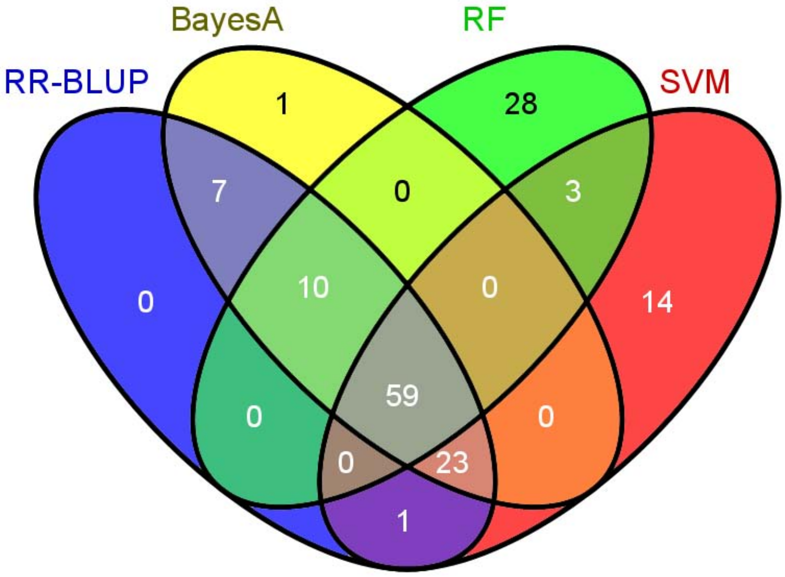


Figure 9.3: Venn-diagram of the top 100 predicted lines per statistical method

9.4 Influence of the training population size

To test the influence of the size of the training population on the prediction accuracy, random subsets of 50, 100 and 250 were chosen together with the full training population. To reduce the bias introduced by the random sampling, 10 rounds of 10-fold cross-validation were performed, each round using a different random sample. As can be seen in Figure 9.4 the predictions become stable at training population size of about 250. However the increase from 250 lines to the full number of lines of 651 increased the median accuracy significantly and reduced the overall bandwidth of the predictions. Hence a minimum training size of about 250 seems advisable.

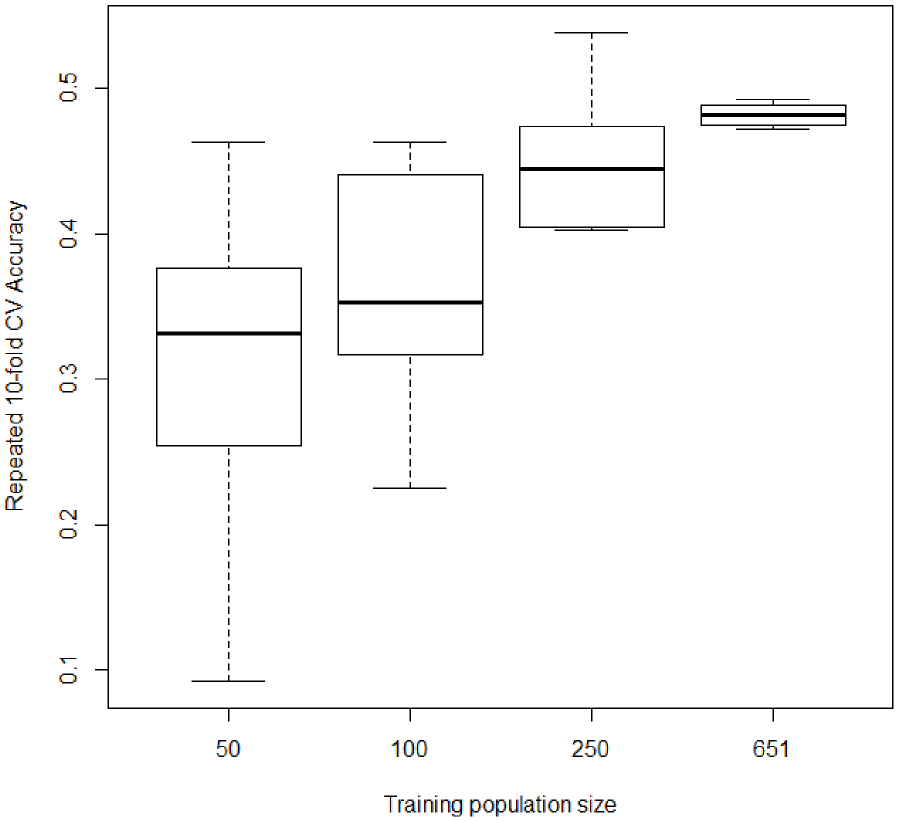


Figure 9.4: Influence of the training population size on the prediction accuracy using RR-BLUP

9.5 Influence of the number of markers

A similar test as in chapter 9.4 was performed to test the influence of the number of markers on the prediction accuracy. Random subsets of 100, 500, 1000, 3500 together with the full set of markers were used to test for the effect on the prediction accuracy.

Again 10 rounds of 10-fold cross-validation were performed and the results plotted in Figure 9.5. There is a steep increase in prediction accuracy going from 100 to 500 markers and again when going from 1000 to 3500 markers. The decrease in median prediction accuracy using the full set raised the question of performing variable selection, e.g. using the BayesCPi method. A minimum number of about 500 markers seemed advisable when using RR-BLUP with a recommended number of markers of at least 3500.

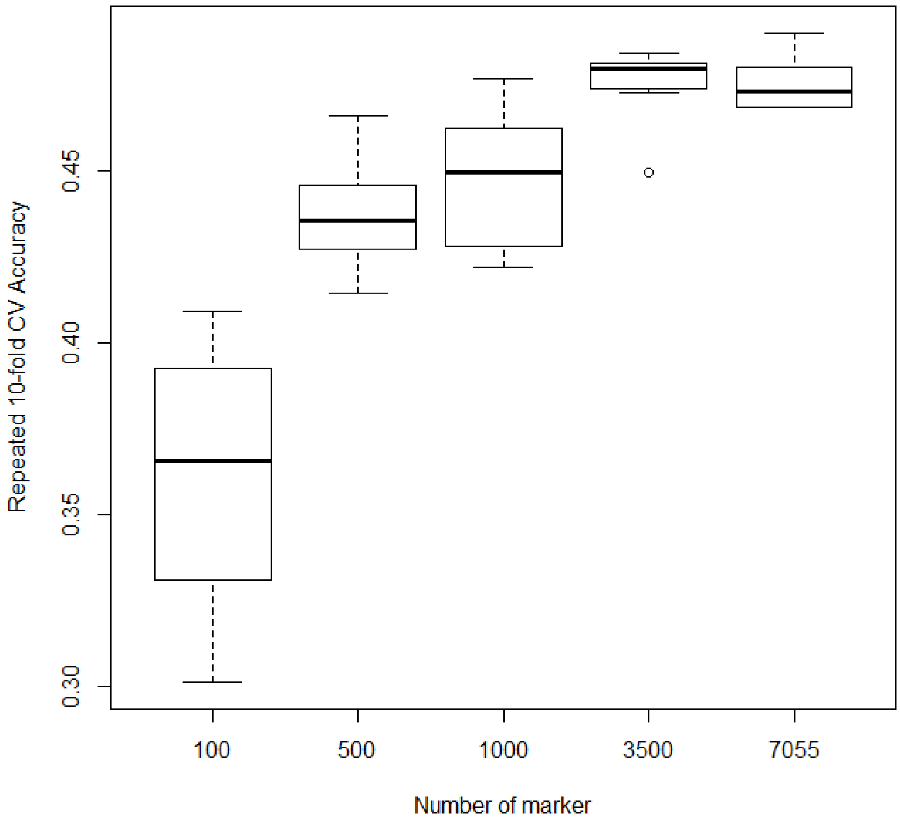


Figure 9.5: Influence of the number of markers on the prediction accuracy using RR-BLUP

9.6 Choice of the prediction method

To choose one of the statistical methods as the GS model in the selection experiment, the performance of the methods as well as their computation time was compared. Table 9.2

shows a drastic advantage in computation time for RR-BLUP. The prediction accuracies found in Figure 9.1 testify to the robustness of the algorithm. Taken together with the results of the prediction rankings it was a fairly easy choice to pick RR-BLUP as the standard model for the final selection experiment. Similarly many other work-groups chose RR-BLUP or GBLUP as the method of choice for its robustness and computational efficiency.

10 Breeding value prediction

After the decision for RR-BLUP as the statistical method genomic estimated breeding values for 800 wheat lines had to be computed based on model parameters estimated from the training set of 651 wheat lines.

10.1 Optimizing the prediction accuracy

Prior to the calculation of breeding values the selection model was optimized to better fit the data at hand using an algorithm similar to leave-one-out cross-validation implementing following steps:

1. Train the model on the full training set
2. Predict the GEBV with the trained model and compute the accuracy
3. Remove 1 wheat line from the training set and train the model again
4. Store the computed accuracy
5. Repeat steps 3 and 4 until each wheat line has once been removed from the training set
6. Sort the stored accuracies
7. If the highest stored accuracy is higher than accuracy computed in step 2 then
    a) remove the corresponding wheat line from the training set
    b) start over again from step 1 with the reduced training set as the new full training set
8. If the accuracy is lower than the computed accuracy in step 2 then
    a) revert the wheat line back to the training population

b) exit the algorithm

Since the statistical model had to be trained at least n times, where n is the number of wheat lines in the training population, and each repeated cycle needed again $n - 1$ training steps the drawback of the algorithm above was the demand in computation time. Any other method than RR-BLUP would have taken far too long to finish the algorithm. Another caveat of the proposed algorithm is that it is very prone to over-fitting, i.e. the estimated breeding values have less accuracy for all but the wheat lines in the training population. To counter this drawbacks a limit of at most 10% removed lines in total was set and after each run of the algorithm the predictive ability of the reduced model was tested using 10 repeated cycles of 10-fold cross-validation.

Though the proposed algorithm is fairly naive, it was possible to find a number of lines having an incorrect phenotype and their removal from the training model truly increased the predictive ability of the model (see Table 10.1).

Table 10.1: Cross-validation accuracies before and after the optimization algorithm

	Grain yield	Protein yield	Rust resistance	Fusarium resistance
Before	0.51	0.39	0.47	0.49
After	0.66	0.46	0.63	0.61
No. removed lines	8%	4%	5%	4%

11 One-year results

The final part of the project was to test genomic selection against conventional selection and, to investigate the impact of directed selection and random selection of wheat lines. Out of the 2000 wheat lines which had been planted as the so called observation set (OBS), about 800 lines were chosen as selection candidates based on visual inspection by the breeder until the end of June. After extraction, DNA of these selection candidates was sent for genotyping middle of July and genotyping results were obtained by end of August. During the first two weeks of September all necessary computations were conducted and selection was performed using following steps:

- Random selection (No. of selected lines = 31)
  - 31 wheat lines were randomly sampled from selection candidates using no phenotypic or genotypic information
- Conventional selection (No. of selected lines = 73)
  - 70 wheat lines were selected by the breeder using all available phenotypic information including early yield scores from the short plots in the observation set
  - Of these 70 lines 3 lines were also selected by random selection and 3 additional lines where therefore selected by the breeder
- Genomic selection (No. of selected lines = 73)
  - 70 wheat lines were selected based on the ranks of the genomic estimated breeding values for the 4 different traits. These 70 wheat lines comprised of

    the 10 top ranked lines for grain yield,

    the 10 top ranked lines for protein yield, the 50 top ranked lines according to a naive index using 40% grain yield, 40% protein yield, 10% rust resistance and 10% fusarium resistance as weight factors.

  - Of these 70 lines 3 lines were also selected by random selection and 3 additional lines where therefore selected by the breeder

- 5 worst predicted lines according to the GS index

11.1 Results

The selected lines were tested in replicated augmented design field trials and phenotypes were estimated as BLUP after spatial analysis as pointed out in the chapter 3.3. Figure 11.1 depict the BLUP for grain yield for the wheat lines selected by the different selection models. Except for one outlier wheat line, which was selected because of being one of the 10 lines with the highest protein content, the lines selected by genomic selection had higher grain yield than the other selection methods. However neither the lines selected by genomic selection nor by conventional selection had significantly higher median grain yield than the randomly selected lines. The top yielding wheat line was also selected by genomic selection whereas random selection failed to select one of the top yielding lines. Genomic selection was not only able to select high yielding wheat lines but was also able to identify low yielding lines as can be seen in the boxplot furthest to the right. The median yield performance of those lines was significantly lower than the medians of genomic, conventional and random selection.

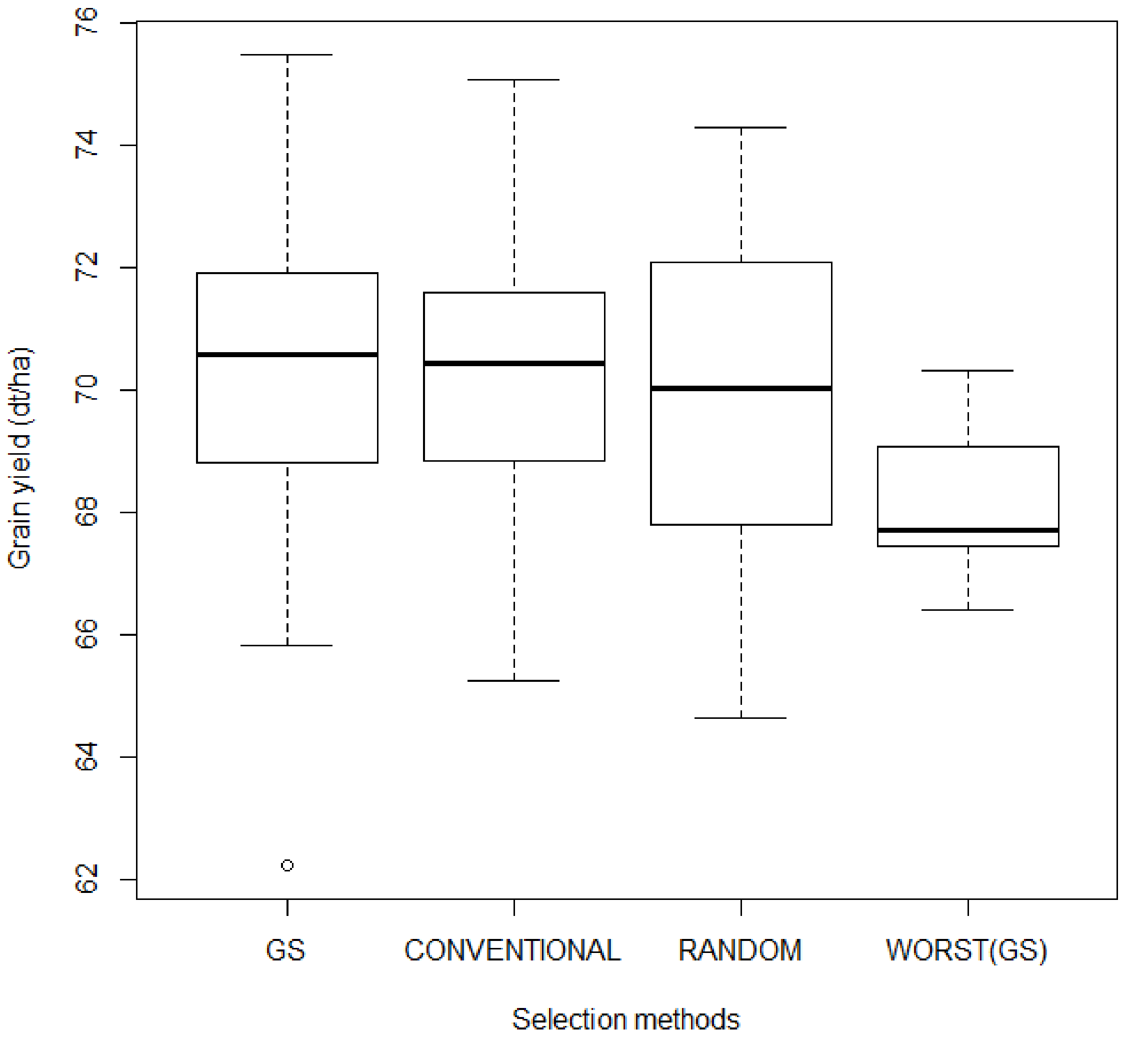


Figure 11.1: One year performance - selection results for the different methods

In Figure 11.2 the centered genomic estimated breeding values were plotted against the observed phenotypes. The horizontal line indicates the median phenotypic value of 70.45dt/ha while the vertical line indicates the median GEBV of 70.57dt/ha. Of the worst predicted lines by genomic selection only one line achieved a performance above the median phenotypic value, while both genomic selection and conventional selection shared 37 selected lines that performed above the median phenotypic value with genomic selection having 23 selected lines in the top 5% and conventional selection having 18 lines selected.

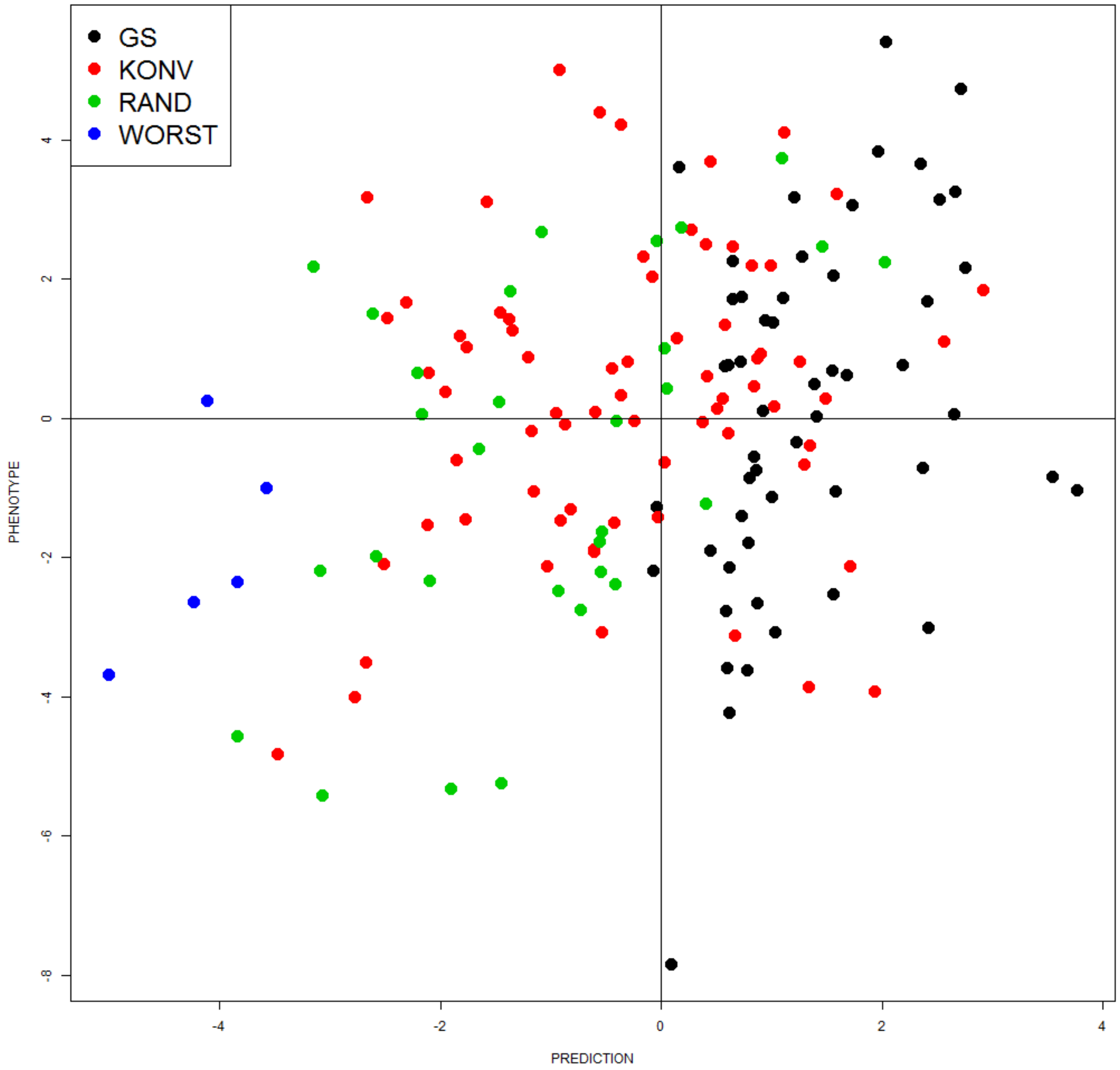


Figure 11.2: Scatterplot between genomic estimated breeding values and observed phenotypes

Table 11.1: Correlations of genomic breeding values and observed phenotypes of different selection methods

	GS	CONVENTIONAL	RANDOM	WORST(GS)
Prediction accuracy	0.26	0.19	0.49	0.62
Critical value ($p < 0.05$)	0.23	0.23	0.36	0.88

11.2 Prediction accuracy

The correlation between genomic estimated breeding values and the observed phenotypes for the set of worst predicted lines was not significant ($p = 0.265$; for critical values see Table 11.1). The accuracy of predictions for the set of randomly selected lines was lower but due to the larger set of lines (n=31) significant ($p = 0.005$). The accuracy of the set of lines by genomic selection was also significant ($p = 0.02$) while the correlation between predictions and phenotypes for the set of conventional selected lines was not significant ($p = 0.107$).

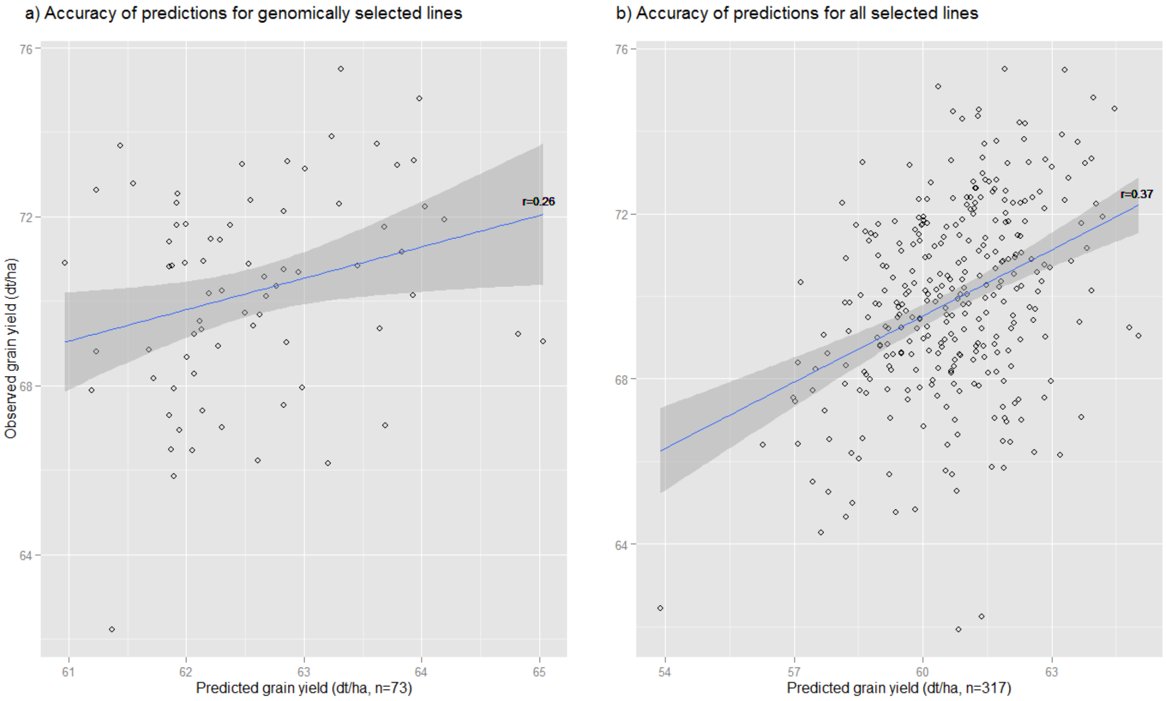


Figure 11.3: Regression of the observed phenotypes on the genomic estimated breeding values

Part IV

Discussion

12 High-density genotyping

While this work was ongoing, several papers on genomic selection were published and it is now more or less degreed on that single nucleotide polymorphism markers are most suitable for large genome-wide high-density genotyping. On the other hand it is not yet decided if genotyping-by-sequencing, using complexity reduction protocols, or pre-defined chip array will be the method of choice for the future's application in breeding programs. While GBS markers lack ascertainment bias and can be produced in nearly unlimited numbers their anonymous origin creates challenges in data handling.

The same set of markers has to be scored in each genotyping batch leading to the requirement of re-scoring previously sent batches of wheat lines. Although re-scoring was performed by the genotyping provider, additional tasks had to be carried out after each genotyping cycle to include the new information. Most prominently, the training population had to be updated each time after obtaining a new set of markers to estimate the new model parameters. While re-training of the training population alone is a simple process which provides little challenge, all downstream analyzes that depend on the model parameters have to be updated as well. Superior progeny value predictions from one year for example would not match up with previous years making it hard to compare between years. Also the optimization step (see chapter 10.1) to increase the prediction accuracy has to be carried out once more and the new additional markers have to be anchored on chromosomes again.

The major problem of re-doing all these analyses is the short time frame between obtaining the genotyping results and the need for the new predictions. In this work this time frame was about two weeks and the optimization step alone took about one week. Improvements in the processing power of newer workstations might help to alleviate this particular problem but the number of possible tasks to be carried out is ultimately limited by this rather short window of about two weeks. Furthermore, and most substantially, the renewal of the model parameters prevents or at least increases the difficulty to use the off-seasonal time of the breeding program to improve the GS program.

To overcome this problem the same set of markers should be used. Chapter 9.5 dealt with how the prediction accuracy plateaus with sufficient high markers number and follow up results using 10.000 instead of 7000 markers also showed no further gain in accuracy. However, not including new markers may prevent new, maybe important, alleles to be accounted for. An alternative and more advisable way would be to use the same set of markers during the limited time frame for selection and re-train the prediction model and all other analyses afterwards.

Another more drastic way would be to switch the genotyping platform to chip-based technologies that use a fixed set of markers (see chapter 2.4.1.1). Improvements made over the last couple of months have yielded arrays that include a set of well-defined markers of which many have been annotated with mapping positions. Furthermore known markers for certain genes (e.g. for reduced height) are now included on such arrays making their usage promising. Another fact that could provide a major advantage in the future is that the more work-groups start using the same markers and publish their results, the more information will be available to breeders. However this would be only the case if there are a fairly limited number of arrays using different markers with ideally one array used by most work-groups.

In any case the aims of optimized marker technologies in the future are clearly defined:

- They should yield high-quality markers that are or can be easily annotated with mapping information.
- The markers should ideally work in a large number of gene-pools at a reasonable allele frequency.
- The markers should ideally include known markers of importance.
- Results should be able to be interchanged between different work-groups.
- They should be available at a reasonable price

The genotyping platform used in this work fulfils the first two points nearly complete although some bioinformatics efforts are needed to make full use of all markers. While there is the chance that one or more markers are in linkage with known markers or genes, with the current platform it is not so easy to identify such markers and with every update of the marker set these have to be obtained again. The most substantial drawback is the initial anonymity of the obtained markers and the difficulty to transfer results from one group that uses the same platform to the other.

In summary, it can be concluded that the choice of the genotyping platform was fully justified, as the data quality was nearly excellent, and even though there are new developments on genotyping technologies, it still remains very competitive and there is no immediate need for a change.

13 Prediction methods

Looking at the results of chapter 9, there are two striking properties of the different methods. Firstly, the median predictions of the different methods are fairly similar, and secondly, there are substantial differences in computation time. Additionally, predictions from the different methods were highly correlated, making the choice for the best method not trivial.

Based on these results, RR-BLUP seems to have the best trade-off between these two criteria. Similar results were first obtained by Meuwissen et al. (2001) in animal breeding, and subsequently verified in plant breeding throughout published literature (Moser et al., 2009; Ogutu et al., 2011; Heslot et al., 2012; de los Campos et al., 2013). Together with the short computation time, RR-BLUP (or GBLUP) has soon become one of the most widely used methods for GS. Still, this method is somewhat a double-edged sword. On the one hand there are the fairly good prediction results and the short computation time. On the other hand, the concept of each marker accounting the same amount of genetic variance contradicts the basics of genetics. While this assumption may holds true for quantitative traits controlled by a large number of genes, it is not an accurate assumption for monogenic traits. And yet the prediction accuracies for such traits are still not far off from models with heteroscedastic marker variance, e.g. BayesA or BayesB. According to Hofheinz and Frisch (2014) a possible explanation could be that for a finite population of related wheat lines it is expected that long stretches of the chromosomes are in linkage disequilibrium and it is sufficient that the sum of marker effects of these stretches are predicted with a high accuracy rather than accurate predictions of single markers. Therefore, disadvantages arise if the marker effects are used for other applications than breeding value prediction. In certain cases, e.g. for planning of crosses, methods with heteroscedastic marker variance would be of advantage.

From a plant breeding perspective, of the Bayesian models BayesCPi seems least favorable due to its variable selection process, which is prone to over-fitting (Heslot et al., 2012).

BayesA and BayesB are special cases of BayesCPi that avoid variable selection and are generally considered as well suited in genomic selection, as the marker effects are very accurately estimated. Other Bayesian approaches, such as the Bayesian Lasso, performed well in animal breeding and are also evaluated for plant breeding applications. The major drawback of all Bayesian methods is their high computation time due to the Markov chain Monte Carlo algorithms used for sampling the model parameter.

Other machine learning algorithms, such as the presented support vector machines or neuronal nets have been investigated (Ogutu et al., 2011), but in general they show no significant advantage. Similar conclusions were drawn for ensemble algorithms like bagging or boosting.

RandomForest is the exemption of these machine learning algorithms. It uses ensembles of unpruned classification and regression trees for prediction and is therefore not prone to overfitting. The computation time is similar to the Bayesian methods. In this work the achieved prediction accuracy was the highest of all methods. One reason could be the ability of the method to capture non-additive effects. Similar findings were made by Heslot et al. (2012) who concluded that more development for using RandomForest in GS should be done.

What this all amounts to is that RR-BLUP is and probably will remain a benchmark for future prediction methods. However, it is vital to acknowledge the possibilities and limitations of the different methods in different situations and to use them accordingly.

14 Breeding value prediction

A significant increase in breeding value prediction was obtained using the proposed naive optimization algorithm. The caveat of such optimization processes based on a given training population is that one can easily overfit the statistical model to the trained data, i.e. the predictions are more accurate for the training set than for other test set. The same applies to methods that perform variable selection on the markers. Consequently, such methods have to be used with care. Yet the increase in predictive ability of 30% can't be neglected. Given that this increase was observed as an average of 10 cycles of 10-fold cross-validation and less than 10% of lines were removed, there should be little overfitting to the training set present. A possible explanation could lie in the quality of phenotypes and in the composition of the training set. Figure 9.4 points out the need of having a sufficiently large training population, and when starting this project, the main objective was to achieve a sufficiently large number of genotyped and phenotyped wheat lines. However, as enough phenotypic and genotypic data are available now, the optimal design of the training set should be the focus of further investigation with more sophisticated methods than the proposed algorithm.

Nevertheless, the algorithm found some illustrative examples of how substantially incorrect phenotypic data influence the prediction results. Five highly related sister lines were in the training set of which four had nearly the same phenotypic value. One line, however, had by mistake a significantly lower phenotypic value and discarding this line of the training set improved the predictive ability.

Consequently, it is advisable to put more work into the optimal design of the breeding population, as for example already mentioned by Endelman et al. (2014). Clusters of genetically related lines that share a similar phenotypic value for the trait of interest could for example be replaced by one representative line. This would help to balance the training population, could potentially increase the predictive ability of the model and decrease computation time.

15 One year results

The idea behind the selection experiment was to test the efficiency of genomic selection compared to phenotypic selection from a practical point of view, rather than looking at cross-validated prediction accuracies. Only grain yield has been evaluated in the presented results, since this was the trait with the most complete set of phenotypic records and it is one of the agronomically most important traits.

It was no surprise that only three selected wheat lines were shared between genomic and random selection and between conventional and random selection. Somewhat surprising was though that genomic and conventional selection only shared 17 lines which were selected by both methods. One possible explanation, which should to a certain degree apply to all the results, is the way how the lines were chosen based on the genomic estimated breeding values. The focus had been laid on grain yield and protein yield. The 10 top ranked lines for both traits were selected in addition to the lines selected by a naive index of 40% grain yield, 40% protein yield, 10% yellow rust resistance and 10% fusarium resistance. Since no other parameters influenced selection other than these four traits, the results for grain yield might be a little biased upwards. Breeder's conventional selection, on the other hand, considered next to grain yield a number of other factors such as morphological and quality scores. Moreover, the only phenotypic data on grain yield that was available to the breeder at the time of selection came from short plots in a field trial at only one location.

Nevertheless, the results in Figure 11.1 suggest that genomic selection should be able to perform on par with conventional selection in this setting.

Interestingly, the median grain yield of the selected lines neither for genomic selection nor for conventional selection was significantly higher than for random selection. Typically the breeder is interested in the top selected lines, and looking at the top 25% of the selected lines, indicated by the whiskers, both directed selection methods surpass random selection. Also the phenotypic records are based on field trials of only one year, and in this particular year there was a heavy yellow rust stress which confounds the results to some extent.

The five worst performing lines according to the genomic selection index were also selected to test their performance. Performance estimates of such lines are not commonly found in published literature since it contradicts the actual breeding target of improving the gene-pool by selecting superior progeny. Yet selection of those lines was helpful to investigate the discriminative ability of the genomic predictions. Four out of the five lines performed worse than the median performance of the conventionally selected lines. Furthermore, the genomic predictions had shown a high correlation with the observed breeding values so that it is possible to conclude that the genomic selection model is not only suited for finding superior progeny but even more so for discarding inferior progeny. This ability is of high interest in the early stages of the breeding cycles where large numbers of selection candidates have to be downscaled to a manageable number of candidate lines.

The correlations of genomical and conventional selection were substantially lower than for the randomly selected lines and only the correlations of GS were significant ($p < 0.05$). The low correlation with the set of conventional selected lines could be an explanation for the small set of lines that was selected by genomic and conventional selection ($n = 17$). However ,the performances of both methods were fairly similar. An other reason for the small overlaps in selected lines could be the different importance of grain yield in the selection decisions. Moreover severe yellow rust stress was present in this year and the rust races had changed so previously established resistance worked only partially.

16 Outlook

Looking back at three years of work what this all amounts to is that the foundations for an ongoing application of genomic selection in the breeding program are laid. Regarding the three main parts of successful genomic selection, genotyping, phenotyping and statistical methods, in each part there is still room for improvement.

The quality of the DNA-samples and the quality of the data obtained by the chosen genotyping platform were nearly excellent. No severe problems were encountered during data processing and a number of new insights into the genetic makeup of the breeding population were gained. However, most of the used markers are still anonymous and lack good annotation. To fully exploit the possibilities offered by the high density genome-wide marker analysis, further work has to be put in annotation of the markers.

Most improvements should be possible in the planning of the field trials and processing of the phenotypic records. In this work phenotypic records of a variety of sometimes fairly different environments were joined together into single predictions. Optimizing the field trials, e.g. by using smaller sets of lines, that are orthogonally tested in similar environments, could improve the data quality substantially. Furthermore, it seems now appropriate to discard lines from the training set, so that the remaining training set is balanced.

Regarding the choice of the statistical methods, there is less chance of significantly improving the prediction results. RR-BLUP turned out to be a robust, well performing prediction method. Obtaining more precise marker effects and having better annotations on those markers however might help substantially in the selection of new parents for crossing. At any rate, the research on machine learning algorithms is constantly increasing and further gains may be obtained by better prediction methods.

Part V

Acknowledgement

Die vorliegende Arbeit wäre ohne den Rat und die Unterstützung einer Vielzahl von Personen nicht möglich gewesen, denen ich hiermit meinen Dank aussprechen möchte.

Vielen Dank an alle Kollegen vom Institut für Biotechnologie in der Pflanzenproduktion am IFA Tulln, die mir während der gesamten Zeit beiseite standen. Besonderer Dank gilt dabei Maria Bürstmayr, die unzählige Stunden mit der Extraktion von DNA-Proben meiner Weizenlinien verbracht hat und der dabei keine einzige Probe missglückt ist. Herzlichen Dank ebenso an Gerlinde Kindler, die dieses Projekt wunderbar organisatorisch betreut hat.

Vor allem möchte ich mich bei Barbara Steiner für all die Hilfe und guten Ratschläge in schwierigen Zeiten und bei Wolfgang Schweiger für die Unterstützung und unermüdliche Bereicherung meines Allgemeinwissens bedanken. Ich hätte mir nie ein besseres Arbeitsklima mit euch beiden wünschen können.

Bei allen Mitarbeitern der Saatzucht Donau möchte ich mich ebenfalls herzlich bedanken. Allen voran Johann Birschitzky und Herbert Hetzendorfer, die mich immer unterstützt haben, aber besonders bei Franziska Löschenberger, die mich über das gesamte Projekt hinweg angespornt hat und mir bei jeder Frage mit Rat und Tat zur Seite gestanden ist. Vor dem Enthusiasmus mit dem du meinen Erkenntnissen begegnet bist und im besonderen Maße vor deinem wirklich großartigen, unermüdlichen Einsatz zur Verbesserung meiner gesamten Arbeit kann ich nur den Hut ziehen.

Großer Dank gilt meinen Betreuern dieser Arbeit. Heinrich Grausgruber für all die Unterstützung bei der Verrechnung und Johann Sölkner für die Erkenntnisse der Tierzucht sowie Clay Sneller für die Zeit, die er für mich während deines Aufenthaltes geopfert hat.

Zu guter Letzt und von ganzem Herzen vielen Dank an Hermann Bürstmayr. Ohne dich wäre es nie zu dieser Arbeit gekommen. Danke, dass du mir dieses Möglichkeit gegeben hast und besonders für deine Unterstützung und Hilfe in all den Stunden, die du für mich aufgebracht hast. Ebenso für all die schönen Stunden, die ich in deinem Institut verbringen durfte.

Dieses Projekt wurde durch das Eurostars Programm der Europäischen Union im Rahmen des Projekts E! 6399 Genomic selection of wheat varieties for robustness, yield and quality sowie der FFG Österreich gefördert, herzlichen Dank.

Part VI

Bibliography

Bibliography

Albrechtsen, A., Nielsen, F. C., and Nielsen, R. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular biology and evolution*, 27(11):2534 47, 2010.

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., and Lander, E. S. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803):513 516, 2000.

Avenue, G., Zhou, W.-C., Kolb, F. L., Bai, G.-H., Domier, L. L., Boze, L. K., and Smith, N. J. Validation of a major QTL for scab resistance with SSR markers and use of marker-assisted selection in wheat. *Plant breeding*, 122(1):40 46, 2003.

Bailey, R. *Design of Comparative Experiments*, volume 25. Cambridge University Press Cambridge, Cambridge, UK, 2008.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. a., Selker, E. U., Cresko, W. a., and Johnson, E. a. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*, 3(10):e3376, 2008.

Bärlocher, F. *Biostatistik*. Georg Thieme Verlag KG, Stuttgart, 2008.

Bernardo, R. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science*, 48(5):1649 1664, 2008.

Bernardo, R. and Yu, J. Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47(3):1082 1090, 2007.

Breiman, L. Random forests. *Machine learning*, 45(1):1 33, 2001.

Burgueño, J., Cadena, A., Crossa, J., Banziger, M., Gilmour, A. R., and Cullis, B. *User s guide for spatial analysis of field variety trials using ASREML*. Cimmyt, Mexico, 2000.

Cakir, M. A., Gupta, S. B., Platz, G. J. C., Ablett, G. A. D., Loughman, R. B., Emebiri, L. C. E., Poulsen, D. C., Li, C. D. B., Lance, R. C. M. B., Galwey, N. W. A., Jones,

M. G. K. F., Appels, R. B., and Others. Mapping and validation of the genes for resistance to Pyrenophora teres f. teres in barley (Hordeum vulgare L.). *Crop and Pasture Science*, 54(12):1369 1377, 2003.

Calinski, T. and Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1 27, 1974.

Casella, G. and George, E. I. E. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167, 1992.

Chahal, G. S. and Gosal, S. S. *Principles and procedures of plant breeding: Biotechnological and conventional approaches.* Alpha Science Int'l Ltd., Oxford, UK, 2002.

Cnaan, A., Laird, N. A. N. M., and Slasor, P. Tutorial in biostatistics: Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat Med*, 16:2349 2380, 1997.

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., and Pang, E. C. K. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, 142(1-2):169 196, 2005.

Collard, B. C. Y. and Mackill, D. J. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1491):557 72, 2008.

Compton, M. E. Statistical methods suitable for the analysis of plant tissue culture data. *Plant cell, tissue and organ culture*, 37(3):217 242, 1994.

Crawley, M. *Statistics: an introduction using R.* John Wiley & Sons, Hoboken, NJ, USA, 2005.

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., Bonnett, D., and Mathews, K. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, 112:1 13, 2013.

Crossa, J., Campos, G. D. L., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., and Braun, H.-J. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186(2):713 24, 2010.

Cullis, B. R. and Gleeson, A. C. Spatial analysis of field experiments-an extension to two dimensions. *Biometrics*, 47:1449 1460, 1991.

Davey, J. W., Hohenlohe, P. a., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics*, 12(7):499 510, 2011.

de los Campos, G. and Paulino Perez Rodriguez. BGLR: Bayesian Generalized Linear Regression, 2014.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375 85, 2009.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327 345, 2013.

Dekkers, J. C. M. Prediction of response to marker-assisted and genomic selection using selection index theory. *Journal of Animal Breeding and Genetics*, 124(6):331 341, 2007.

Des Marais, D. L., Hernandez, K. M., and Juenger, T. E. Genotype-by-environment interaction and plasticity: Exploring genomic responses of plants to the abiotic environment. *Annual Review of Ecology, Evolution, and Systematics*, 44:5 29, 2013.

Edwards, M. D., Stuber, C. W., and Wendel, J. F. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics*, 116(1):113 125, 1987.

Edwards, M. and Henry, R. DNA sequencing methods contributing to new directions in cereal research. *Journal of Cereal Science*, 54(3):395 400, 2011.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. a., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PloS one*, 6(5):e19379, 2011.

Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal*, 4(3):250, 2011.

Endelman, J. B., Atlin, G. N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M. E., and Jannink, J.-L. Optimal Design of Preliminary Yield Trials with Genome-Wide Markers. *Crop Science*, 54(1):48, 2014.

Fang, D. D., Xiao, J., Canci, P. C., and Cantrell, R. G. A new SNP haplotype associated with blue disease resistance gene in cotton (Gossypium hirsutum L.). *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 120(5):943 53, 2010.

Faraway, J. J. Practical Regression and Anova using R. University of Bath, 2002.

Federer, W. T. Augmented (or hoonuiaku) designs. *Biometrics Unit Technical Reports*, 55:191 208, 1956.

Federer, W. T., Raghavarao, D., and Federer. On augmented designs. *Biometrics*, 31(1): 29 35, 1975.

Federer, W. W. T., Reynolds, M., and Crossa, J. Combining Results from Augmented Designs over Sites. *Agronomy Journal*, 93:55 59, 1999.

Fisher, R. A. The Design of Experiments. *The American Mathematical Monthly*, 43(3): 180, 1936.

Fisher, R. A. The arrangement of field experiments. In *Breakthroughs in Statistics*, pages 82 91. Springer, 1992.

Foll, M., Beaumont, M. a., and Gaggiotti, O. An approximate Bayesian computation approach to overcome biases that arise when using amplified fragment length polymorphism markers to study population structure. *Genetics*, 179(2):927 39, 2008.

Gilmour, A., Cullis, B., and Verbyla, A. Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, 2:269 293, 1997.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., and Thompson, R. ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*, 2009.

Gleeson, A. C. and Cullis, B. R. Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. *Biometrics*, 43:277 287, 1987.

González-Camacho, J. M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., Babu, R., and Crossa, J. Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and applied genetics.*, 125(4): 759 71, 2012.

Guillot, G. and Foll, M. Correcting for ascertainment bias in the inference of population structure. *Bioinformatics*, 25(4):552 554, 2009.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389 97, 2007.

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics*, 12(1):186, 2011.

Hamada, M. S., Wilson, A., Reese, C. S., and Martz, H. *Bayesian reliability.* Springer, Berlin, Germany, 2008.

Hayes, B. J., Bowman, P. J., Chamberlain, a. J., and Goddard, M. E. Invited review: Genomic selection in dairy cattle: progress and challenges. *Journal of dairy science*, 92(2):433 43, 2009.

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. Genomic Selection for Crop Improvement. *Crop Science*, 49(1):1, 2009.

Heffner, E. L., Lorenz, A. J., Jannink, J.-L., and Sorrells, M. E. Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. *Crop Science*, 50(5):1681, 2010.

Heffner, E. L., Jannink, J.-l., and Sorrells, M. E. Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. pages 65 75, 2011.

Heller, M. J. DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4:129 53, 2002.

Henderson, C. R. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423 447, 1975.

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Science*, 52(1):146, 2012.

Heslot, N., Rutkoski, J., Poland, J., Jannink, J.-L., and Sorrells, M. E. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PloS one*, 8(9):e74612, 2013.

Hinkelmann, K. *Design and Analysis of Experiments, Special Designs and Applications*, volume 3. John Wiley & Sons, Hoboken, NJ, USA, 2011.

Hofheinz, N. and Frisch, M. Heteroscedastic Ridge Regression Approaches for Genome-Wide Prediction with a Focus on Computational Efficiency and Accurate Effect Estimation. *G3: Genes Genomes Genetics*, 4(3):539 546, 2014.

Hofheinz, N., Borchardt, D., Weissleder, K., and Frisch, M. Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theoretical and applied genetics*, 125(8):1639 45, 2012.

Ignal, A. V., Ilan, D. M., Vignal, A., Milan, D., SanCristobal, M., Eggen, A., et al. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, 34(3):275 306, 2002.

Jannink, J.-L., Lorenz, A. J., and Iwata, H. Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics*, 9(2):166 77, 2010.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X., Jing, R., Zhang, C., Ma, Y., Gao, L., Gao, C., Spannagl, M., Mayer, K. F. X., Li, D., Pan, S., Zheng, F., Hu, Q., Xia, X., Li, J., Liang, Q., Chen, J., Wicker, T., Gou, C., Kuang, H., He, G., Luo, Y., Keller, B., Xia, Q., Lu, P., Wang, J. J., Zou, H., Zhang, R., Xu, J., Gao, J., Middleton, C., Quan, Z., Liu, G., Yang, H., Liu, X., He, Z., and Mao, L. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, 496(7443):91 5, 2013.

Johnstone, I. M. and Titterington, D. M. Statistical challenges of high-dimensional data. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367(1906):4237 53, 2009.

Kearsey, M. J. The principles of QTL analysis (a minimal mathematics approach). *Journal of Experimental Botany*, 49(327):1619 1623, 1998.

Kennard, R. W. and Hoerl, A. E. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55 67, 1970.

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., Caig, V., Heller-Uszynska, K., Jaccoud, D., Hopper, C., and Others. Diversity arrays technology: a generic genome profiling technology on open platforms. In *Data production and analysis in population genomics*, pages 67 89. Springer, 2012.

Kolbehdari, D., Schaeffer, L. R., and Robinson, J. A. B. Estimation of genome-wide haplotype effects in half-sib designs. *Journal of Animal Breeding and Genetics*, 124(6): 356 361, 2007.

Koziel, M. G., Beland, G. L., Bowman, C., Carozzi, N. B., Crenshaw, R., Crossland, L., Dawson, J., Desai, N., Hill, M., Kadwell, S., and Others. Field performance of elite transgenic maize plants expressing an insecticidal protein derived from Bacillus thuringiensis. *Nature biotechnology*, 11(2):194 200, 1993.

Lander, E. S., B, D. B., and Botstein, D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185 99, 1989.

Lemieux, B., Aharoni, A., and Schena, M. Overview of DNA chip technology. *Molecular Breeding*, 4(4):277 289, 1998.

Lentner, M. and Bishop, T. *Experimental Design and Analysis.* Valley Book Company, Blacksburg, VA, USA, 1986.

Liaw, A. and Wiener, M. Classification and Regression by randomForest. *R News*, 2(3): 18 22, 2002.

Lorenz, a. J., Smith, K., and Jannink, J.-L. Potential and Optimization of Genomic Selection for Fusarium Head Blight Resistance in Six-Row Barley. *Crop Science*, 52 (4):1609, 2012.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E., and Jannink, J.-L. Genomic Selection in Plant Breeding: Knowledge and Prospects. *Advances in agronomy*, 110:77, 2011.

Lorenzana, R. E. and Bernardo, R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *TAG. Theoretical and applied genetics.*, 120(1):151 61, 2009.

Lupton, F. G. H. and Others. *Wheat breeding-its scientific basis.* Chapman & Hall Ltd., London, UK, 1987.

Lusser, M., Parisi, C., Plan, D., and Rodríguez-cerezo, E. Deployment of new biotechnologies in plant breeding. *Nature biotechnology*, 30(3):231 9, 2012.

Mayer, K. F. X., Rogers, J., Dole el, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A. J., Sourdille, P., and Others. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, 345(6194):1251788, 2014.

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819 29, 2001.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014.

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. a., and Johnson, E. a. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome research*, 17(2):240 8, 2007.

Moose, S. P. and Mumm, R. H. Molecular plant breeding as the foundation for 21st century crop improvement. *Plant physiology*, 147(3):969 77, 2008.

Moser, G., Tier, B., Crump, R. E., Khatkar, M. S., and Raadsma, H. W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet Sel Evol*, 41(1):56, 2009.

Nazar, R. N., Chen, P., Dean, D., and Robb, J. DNA chip analysis in diverse organisms with unsequenced genomes. *Molecular biotechnology*, 44(1):8 13, 2010.

Neves, H. H., Carvalheiro, R., and Queiroz, S. a. A comparison of statistical methods for genomic selection in a mice population. *BMC genetics*, 13(1):100, 2012.

Ogutu, J. O., Piepho, H.-P., and Schulz-Streeck, T. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC proceedings*, 5 Suppl 3(Suppl 3):S11, 2011.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., and Tanksley, S. D. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, 335(6192):721 726, 1988.

Piepho, H. P., Möhring, J., Melchinger, a. E., and Büchse, a. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1-2):209 228, 2007.

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., and Jannink, J.-L. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome Journal*, 5(3): 103, 2012.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

Raman, H., Raman, R., Kilian, A., Detering, F., Carling, J., Coombes, N., Diffey, S., Kadkol, G., Edwards, D., McCully, M., Ruperao, P., Parkin, I. a. P., Batley, J., Luckett, D. J., and Wratten, N. Genome-wide delineation of natural variation for pod shatter resistance in Brassica napus. *PloS one*, 9(7):e101673, 2014.

Riedelsheimer, C., Endelman, J. B., Stange, M., Sorrells, M. E., Jannink, J.-L., and Melchinger, A. E. Genomic Predictability of Interconnected Bi-parental Maize Populations. *Genetics*, 194(2):493 503, 2013.

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. a., Swarts, K. L., Casstevens, T. M., Elshire, R. J., Acharya, C. B., Mitchell, S. E., Flint-Garcia, S. a., McMullen, M. D., Holland, J. B., Buckler, E. S., and Gardner, C. a. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, 14(6):R55, 2013.

Saghai-Maroof, M., Biology, P., and Allard, R. W. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proceedings of the . . .*, 81(December):8014 8018, 1984.

Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., and Kilian, A. Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. *BMC Proceedings*, 5(Suppl 7):P54, 2011.

Sax, K. The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. *Genetics*, 8(6):552, 1923.

Searle, S. R. *Linear models*. John Wiley & Sons, Hoboken, NJ, USA, 2012.

Sharma, H. C., Crouch, J. H., Sharma, K. K., Seetharama, N., and Hash, C. T. Applications of biotechnology for crop improvement: prospects and constraints. *Plant Science*, 163(3):381 395, 2002.

Sharma, V. K. Incomplete Block Desgins. In *Advances in Data Analytical Techniques*, pages 163 176. Indian Agricultural Statistics Research Institute, 2008.

Storlie, E. and Charmet, G. Genomic Selection Accuracy using Historical Data Generated in a Wheat Breeding Program. *The Plant Genome*, 6(1):1 9, 2013.

Stringer, J. K., Smith, A. B., and Cullis, B. R. Spatial Analysis of Agricultural Field Experiments. *Design and Analysis of Experiments, Special Designs and Applications*, 3:109, 2011.

Sun, C., Wu, X.-l., Weigel, K. A., Rosa, G. J. M., Bauck, S., Woodward, W., Schnabel, R. D., Taylor, J. F., and Gianola, D. Ensemble-Based Imputation for Genomic Selection: an Application to Angus Cattle. *Interbull Bulletin*, (46), 2012a.

Sun, X., Qu, L., Garrick, D. J., Dekkers, J. C. M., and Fernando, R. L. A fast EM algorithm for BayesA-like prediction of genomic breeding values. *PloS one*, 7(11): e49157, 2012b.

Tanksley, S. D. Mapping polygenes. *Annual review of genetics*, 27(1):205 233, 1993.

Trigiano, R. N. and Gray, D. J. *Plant tissue culture concepts and laboratory exercises*. CRC press, Boca Raton, FL, USA, 1999.

Utz, H. F. PLABSTAT, ein Computerprogramm zur statistischen Analyse pflanzenz{ü}chterischer Experimente. *Universit{ä}t Hohenheim*, 2001.

Väli, U., Brandström, M., Johansson, M., Ellegren, H., and Väli, U. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC genetics*, 9 (1):8, 2008.

VanRaden, P. M. Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414 23, 2008.

VanRaden, P. M. Genomic measures of relationship and inbreeding. *INTERBULL bulletin*, 25(37):33, 2007.

Varshney, R. K., Hoisington, D. a., and Tyagi, A. K. Advances in cereal genomics and applications in crop breeding. *Trends in biotechnology*, 24(11):490 9, 2006.

Von Mark, V. C., Kilian, A., and Dierig, D. A. Development of DArT Marker Platforms and Genetic Diversity Assessment of the US Collection of the New Oilseed Crop Lesquerella and Related Species. *PloS one*, 8(5):e64062, 2013.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. gplots: Various R programming tools for plotting data, 2014.

Weng, Z., Zhang, Z., Ding, X., Fu, W., Ma, P., Wang, C., and Zhang, Q. Application of imputation methods to genomic selection in Chinese Holstein cattle. *Journal of animal science and biotechnology*, 3(1):6, 2012.

Xu, Y. and Crouch, J. H. Marker-Assisted Selection in Plant Breeding: From Publications to Practice. *Crop Science*, 48(2):391, 2008.

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H. P., Longin, F. H., Ranc, N., and Reif, J. C. Accuracy of genomic selection in European maize elite breeding populations. *TAG. Theoretical and applied genetics.*, 124(4):769 76, 2012a.

Zhao, Y., Gowda, M., Longin, F. H., Würschum, T., Ranc, N., and Reif, J. C. Impact of selective genotyping in the training population on accuracy and bias of genomic selection. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 125(4):707 13, 2012b.

Zhong, S. and Jannink, J.-L. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics*, 177(1):567 76, 2007.

Zhong, S., Dekkers, J. C. M., Fernando, R. L., and Jannink, J.-L. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics*, 182(1):355 364, 2009.

Part VII

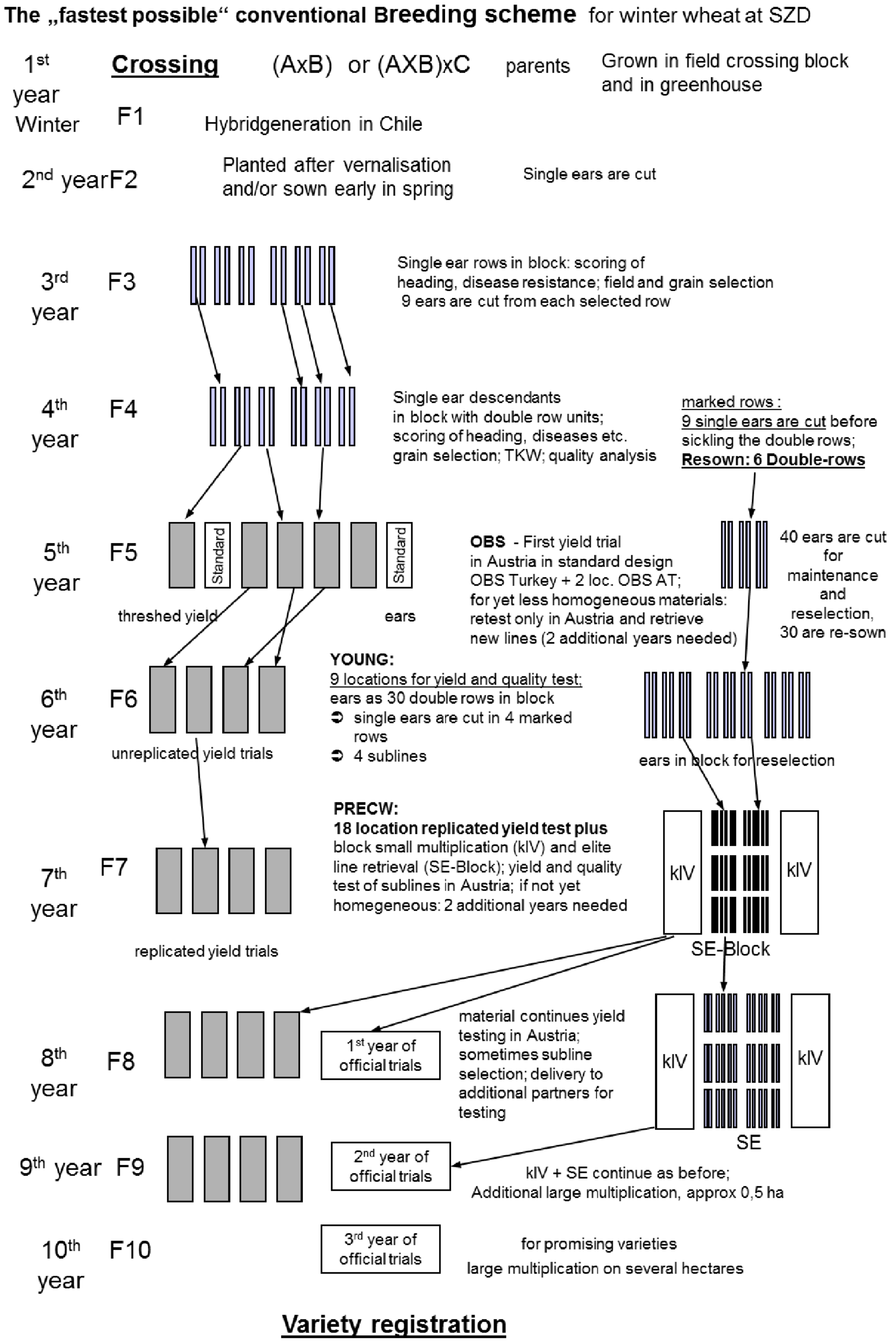
Appendix

17 Phenotyping plan

GENOWHEAT 2011/2012 SCORING TABLE 13.12.2011

TRAITS for phenotyping	**CODE**		
	for		
Robustness mediating traits	**reporting**		
- morphological characteristics		**Score**	
tillering capacity	**TILL**	1 to 9	1= only main tiller, "estimated" numbers of tillers
early vigorous growth	**VIGOR**	1 to 9	1=very vigorous, 9= retarded slow growth
waxiness of ears and leaves	**WAX**	1 to 9	1=without wax
awnedness	**AWNS**	0, 1	0= awnless, 1= awned
plant height	**PH**	in cm	top of canopy
inclination of leaves before anthesis	**INCL**	1 to 9	1= upright, 9 = curved
salinity resistance score	**SAL**	1 to 9	1= resistant
germination under salinity conditions	**GERSAL**	1 to 9	1= fully germinated, 9= no germination
- abiotic stress resistance and physiology			
winter survival - frost resistance	**WINT1**	1 to 9	1= very good, very dense, 9= dead, no plant left
winter survival - frost resistance	**WINT2**	1 to 9	1= very good, very dense, 9= dead, no plant left
canopy temperature depression	**CT**		?
leaf chlorophyll content	**CHL**		?
heading date	**HD**		date of heading - e.g. days after May 1st
Zadocks growth stage instead of heading date	**ZDKS**		*Alternative to heading date in Turkey*
resistance to lodging (Zadoks 1 - 49)	**LOD1**	1 to 9	1= upright, 9= totally lodged
resistance to lodging (Zadoks 50 - 70)	**LOD2**	1 to 9	1= upright, 9= totally lodged
resistance to lodging (Zadoks 70 - 95)	**LOD3**	1 to 9	1= upright, 9= totally lodged
appearance during drought	**DRO1**	1 to 9	1 = resistant, 9= totally diseased, dead
appearance during drought	**DRO2**	1 to 9	1 = resistant, 9= totally diseased, dead
- biotic stress resistance			
powdery mildew resistance (Erysiphe graminis)	**PM1**	1 to 9	1 = resistant, 9= totally diseased, dead
powdery mildew resistance (Erysiphe graminis)	**PM2**	1 to 9	1 = resistant, 9= totally diseased, dead
yellow rust resistance (Puccinia striiformis)	**YR1**	1 to 9	1 = resistant, 9= totally diseased, dead
yellow rust resistance (Puccinia striiformis)	**YR2**	1 to 9	1 = resistant, 9= totally diseased, dead
leaf rust resistance (Puccinia recondita)	**LR1**	1 to 9	1 = resistant, 9= totally diseased, dead
leaf rust resistance (Puccinia recondita)	**LR2**	1 to 9	1 = resistant, 9= totally diseased, dead
tan spot (Helminthosporium tritici repentis)	**HTR1**	1 to 9	1 = resistant, 9= totally diseased, dead
tan spot (Helminthosporium tritici repentis)	**HTR2**	1 to 9	1 = resistant, 9= totally diseased, dead
septoria leaf spot (Septoria tritici)	**STRIT1**	1 to 9	1 = resistant, 9= totally diseased, dead
septoria leaf spot (Septoria tritici)	**STRIT2**	1 to 9	1 = resistant, 9= totally diseased, dead
septoria nodorum leaf	**SNOD1**	1 to 9	1 = resistant, 9= totally diseased, dead
septoria nodorum leaf	**SNOD2**	1 to 9	1 = resistant, 9= totally diseased, dead
head blight (Fusarium spp)	**FUS1**	1 to 9	1 = resistant, 9= totally diseased, dead
head blight (Fusarium spp)	**FUS2**	1 to 9	1 = resistant, 9= totally diseased, dead
eye spot (Pseudocercosporella/Oculimacula sp)	**CERC**	1 to 9	1 = resistant, 9= totally diseased, dead
- yield potential			
grain yield dt/ha	**YLD**	dt/ha	
protein yield dt/ha	**PROTY**	dt/ha	
Quality traits			
Protein content in dry matter %	**PROT**		
Hectolitre weight kg/hl	**HLW**		
Thousand grain weight g	**TGW**		
SDS - Sedimentation volume	**SDS**		
Zeleny - Sedimentation volume	**ZEL**		
Gluten Index	**GLUTI**		
Hagberg Falling number	**HFN**		
Farinogram stability	**FSTAB**		
Extensogram - Water uptake	**EXWA**		
Extensogram - Energy	**EXE**		
Extensogram - Extensibility	**EXEXT**		
Extensogram - ratio resistance/extensibility	**EXRA**		
Alveogram W-value	**ALW**		
Alveogram P/L - value	**ALPL**		
Yellow pigment content by Near Infrared	**YEL**		
Grain hardness	**HARD**		
specific requirements per region yet to be defined			
list to be extended by the partners			

18 Breeding scheme



The „fastest possible“ conventional Breeding scheme for winter wheat at SZD

1st year Winter

Crossing (AxB) or (AXB)xC parents

Grown in field crossing block and in greenhouse

F1 Hybridgeneration in Chile

2nd year F2

Planted after vernalisation and/or sown early in spring

Single ears are cut

3rd year F3

Single ear rows in block: scoring of heading, disease resistance; field and grain selection
9 ears are cut from each selected row

4th year F4

Single ear descendants in block with double row units; scoring of heading, diseases etc. grain selection; TKW; quality analysis

marked rows :
9 single ears are cut before sickling the double rows;
Resown: 6 Double-rows

5th year F5

threshed yield

ears

OBS - First yield trial in Austria in standard design
OBS Turkey + 2 loc. OBS AT;
for yet less homogeneous materials: retest only in Austria and retrieve new lines (2 additional years needed)

40 ears are cut for maintenance and reselection, 30 are re-sown

6th year F6

unreplicated yield trials

YOUNG:
9 locations for yield and quality test;
ears as 30 double rows in block
- single ears are cut in 4 marked rows
- 4 sublines

ears in block for reselection

7th year F7

replicated yield trials

PRECW:
18 location replicated yield test plus
block small multiplication (klV) and elite line retrieval (SE-Block); yield and quality test of sublines in Austria; if not yet homegeneous: 2 additional years needed

8th year F8

1st year of official trials

material continues yield testing in Austria; sometimes subline selection; delivery to additional partners for testing

9th year F9

2nd year of official trials

klV + SE continue as before;
Additional large multiplication, approx 0,5 ha

10th year F10

3rd year of official trials

for promising varieties
large multiplication on several hectares

Variety registration

19 Multi-site analysis linear mixed model

```
YLD ~ mu,
at(ENV,1,5,12,14,18,24,26,28,43,53,62,64,69,71,73,74,75,80,82,85,86,88,89,99,106).lin(COL),
at(ENV,4,14,62,75,88).lin(ROW),
!r GEN ENV.GEN,
at(ENV,2,3,5,7,13,16,17,20,21,22,23,28,29,30,31,32,33,34,36,37,38,41,45,47,48,49,50,54,56,59,60,61
,63,64,66,67,68,70,71,72,79,82,83,85,87,89,91,92,93,94,96,97,98,99,102,103,104,105,107).ROW
8.79762 1.49566 15.26 25.736 12.6157 1.21082 37.8532 0.974436 22.6909 18.5879 11.2928 8.30439
1.49603 38.7708 32.5966 12.5841 5.68974 17.9589 35.9291 53.7492 2.91524 4.13032 2.27683
3.75157 2.59267 5.49876 11.2643 24.1825 7.46341 5.75834 4.33337 8.99291 0.901243 3.93235
2.18493 27.0366 12.6652 23.5546 467.368 2.60217 2.04516 6.20886 7.44993 11.1311 24.6405
4.99482 3.29107 16.9499 2.81121 14.4871 6.07401 20.6225 0.216127 7.1652 10.1314 15.5106
7.67513 3.76606 4.73148,
at(ENV,2,4,21,22,25,29,32,33,34,35,37,38,39,41,42,44,45,46,50,54,55,59,66,78,79,91,92,93,94,100,1
04).COL
1.17055 9.86605 1.50179 2.21248 9.64147 7.52045 8.4745 1.73995 1.62074 30.498 1.51557 1.29369
5.41841 1.28569 3.03962 3.29585 15.9508 16.7196 3.02123 1.629 0.767554 2.32163 0.799926
11.4842 1.11676 2.63452 7.72884 3.7719 6.62553 4.80423 1.6465,
at(ENV,1,5,18,24,26,28,43,53,62,64,69,74,75,80,82,88,89,99,105,106).spl(COL)
 0.342035 0.0000015823 2.9955 0.0000162127 406.478 0.00553533 12.4901 20.57 7.51553 3.5438
0.0352912 0.0630464 4.12213 0.104207 0.84673 14.1067 1.59433 1.74742 4.80718 43.2373,
!f mv
ENV 2
6 ROW AR1 0.32804 !S2=23.8336 // 60 COL AR1 0.324149
4 ROW ID !S2=11.186 // 15 COL ID
2 ROW AR1 -0.932485 !S2=2.72222 // 90 COL ID
8 ROW AR1 0.243227 !S2=11.6437 // 25 COL AR1 0.19072
4 ROW AR1 0.449002 !S2=15.6365 // 45 COL AR1 0.43967
1 ROW ID !S2=15.3467 // 40 COL ID
19 ROW ID !S2=19.8233 // 15 COL ID
2 ROW ID !S2=68.889 // 30 COL AR1 0.259831
2 ROW AR1 0.105569 !S2=15.8392 // 30 COL AR1 0.170023
4 ROW AR1 0.501795 !S2=55.0459 // 50 COL AR1 0.480745
3 ROW AR1 0.261676 !S2=41.3731 // 20 COL AR1 0.372021
4 ROW AR1 0.257492 !S2=44.4573 // 50 COL AR1 0.968688
3 ROW ID !S2=8.0521 // 20 COL ID
3 ROW AR1 0.461391 !S2=7.87683 // 75 COL ID
1 ROW ID !S2=53.2915 // 40 COL ID
6 ROW AR1 0.339828 !S2=6.10933 // 60 COL AR1 0.197897
3 ROW AR1 0.390852 !S2=11.8144 // 30 COL AR1 0.458912
3 ROW ID !S2=19.3315 // 30 COL AR1 0.340156
8 ROW AR1 0.396863 !S2=22.2876 // 25 COL AR1 0.249917
4 ROW ID !S2=8.65078 // 45 COL ID
3 ROW AR1 0.319117 !S2=7.45025 // 22 COL AR1 0.411152
6 ROW ID !S2=6.0618 // 33 COL AR1 0.214931
22 ROW ID !S2=9.5971 // 5 COL ID
2 ROW AR1 0.56635 !S2=30.8818 // 55 COL AR1 0.806794
```

```
1 ROW ID !S2=29.2555 // 180 COL AR1 0.972636
6 ROW ID !S2=16.19 // 33 COL AR1 0.173738
3 ROW AR1 0.626379 !S2=26.9956 // 55 COL AR1 0.610154
5 ROW AR1 0.196198 !S2=5.02443 // 75 COL AR1 0.555339
3 ROW ID !S2=14.572 // 30 COL AR1 0.508503
3 ROW AR1 0.476302 !S2=62.3331 // 70 COL AR1 0.673291
5 ROW AR1 0.63905 !S2=36.2432 // 50 COL AR1 0.491735
6 ROW AR1 0.411598 !S2=25.7858 // 60 COL AR1 0.24513
8 ROW ID !S2=2.12295 // 25 COL AR1 0.482215
11 ROW ID !S2=22.2198 // 51 COL AR1 0.295641
1 ROW ID !S2=27.4446 // 40 COL AR1 0.799092
4 ROW ID !S2=35.8055 // 45 COL AR1 0.729696
4 ROW ID !S2=9.0708 // 10 COL ID
8 ROW ID !S2=12.0517 // 30 COL AR1 0.295919
2 ROW AR1 0.35367 !S2=6.19513 // 20 COL AR1 0.913167
1 ROW ID !S2=5.26 // 90 COL ID
2 ROW AR1 0.999 !S2=6.08629 // 45 COL AR1 0.999
1 ROW ID !S2=7.7443 // 60 COL AR1 0.706604
4 ROW AR1 0.45997 !S2=34.766 // 50 COL AR1 0.601447
4 ROW ID !S2=9.45324 // 25 COL AR1 0.355627
4 ROW ID !S2=17.8183 // 15 COL ID
2 ROW ID !S2=12.1917 // 30 COL ID
8 ROW AR1 0.358477 !S2=4.46392 // 25 COL AR1 0.579082
4 ROW AR1 0.398257 !S2=11.2327 // 15 COL AR1 0.308569
4 ROW AR1 0.176302 !S2=22.4791 // 75 COL AR1 0.18079
4 ROW AR1 0.230192 !S2=15.3892 // 15 COL ID
4 ROW AR1 0.225996 !S2=15.4469 // 25 COL AR1 0.613934
4 ROW AR1 0.40822 !S2=6.24515 // 50 COL AR1 0.867521
4 ROW AR1 0.641721 !S2=19.9827 // 50 COL AR1 0.692025
4 ROW AR1 0.402926 !S2=30.9814 // 25 COL AR1 0.731436
2 ROW AR1 0.241512 !S2=8.32587 // 30 COL AR1 0.904702
4 ROW AR1 0.359351 !S2=9.62247 // 45 COL AR1 0.831895
3 ROW AR1 0.0236525 !S2=20.3335 // 20 COL AR1 0.683504
3 ROW AR1 0.494805 !S2=3.82117 // 20 COL AR1 0.319915
2 ROW ID !S2=32.443 // 30 COL AR1 0.582177
4 ROW AR1 0.298553 !S2=13.084 // 45 COL AR1 0.678558
4 ROW AR1 0.254035 !S2=11.7852 // 15 COL ID
10 ROW AR1 0.664449 !S2=38.4534 // 26 COL AR1 0.744553
4 ROW ID !S2=2.06312 // 15 COL AR1 0.476212
6 ROW AR1 0.605964 !S2=3.26535 // 30 COL AR1 0.598866
4 ROW AR1 -0.00675665 !S2=18.7136 // 45 COL AR1 -0.166443
4 ROW AR1 0.497995 !S2=4.4771 // 25 COL AR1 0.528292
6 ROW ID !S2=18.7932 // 10 COL ID
4 ROW ID !S2=7.80066 // 15 COL ID
1 ROW ID !S2=3.59571 // 60 COL AR1 0.162954
2 ROW AR1 0.186714 !S2=50.788 // 30 COL AR1 0.571719
```

```
2 ROW ID !S2=34.7792 // 30 COL AR1 0.364309
2 ROW AR1 0.164883 !S2=15.5198 // 30 COL ID
2 ROW AR1 0.117064 !S2=7.67437 // 30 COL AR1 0.248654
3 ROW AR1 0.367788 !S2=2.90754 // 30 COL AR1 0.539236
3 ROW ID !S2=11.7926 // 30 COL AR1 0.375979
2 ROW ID !S2=4.72006 // 35 COL AR1 0.723293
4 ROW ID !S2=5.30369 // 15 COL ID
4 ROW ID !S2=3.77101 // 15 COL AR1 0.40596
4 ROW ID !S2=33.1212 // 15 COL ID
3 ROW ID !S2=5.92291 // 20 COL AR1 0.141009
4 ROW ID !S2=11.3979 // 15 COL AR1 0.375848
4 ROW ID !S2=13.5274 // 15 COL ID
4 ROW AR1 0.187096 !S2=9.82813 // 15 COL AR1 0.38768
1 ROW ID !S2=49.5624 // 60 COL AR1 0.627281
3 ROW AR1 0.337009 !S2=30.0224 // 30 COL AR1 0.570524
3 ROW AR1 0.336828 !S2=31.8177 // 30 COL AR1 0.379176
4 ROW ID !S2=22.1768 // 15 COL AR1 0.166897
3 ROW AR1 0.419041 !S2=87.1677 // 30 COL AR1 0.616951
2 ROW AR1 0.414766 !S2=15.9321 // 30 COL AR1 0.62151
4 ROW AR1 -0.490417 !S2=12.2852 // 15 COL AR1 0.567602
2 ROW ID !S2=29.3894 // 30 COL AR1 0.79172
4 ROW AR1 -0.604327 !S2=13.8828 // 15 COL AR1 -0.435511
2 ROW ID !S2=14.4638 // 30 COL AR1 0.601888
4 ROW ID !S2=11.142 // 15 COL ID
4 ROW ID !S2=7.50599 // 15 COL AR1 0.151789
4 ROW ID !S2=42.9919 // 15 COL ID
3 ROW ID !S2=29.0686 // 30 COL AR1 0.582243
2 ROW ID !S2=2.92514 // 30 COL ID
4 ROW ID !S2=17.5381 // 45 COL AR1 0.680887
4 ROW ID !S2=14.0847 // 45 COL AR1 0.7264
8 ROW AR1 0.701142 !S2=18.9358 // 25 COL AR1 0.753636
7 ROW ID !S2=36.3199 // 30 COL ID
4 ROW AR1 0.175435 !S2=75.7671 // 50 COL AR1 0.124113
4 ROW ID !S2=19.0702 // 50 COL AR1 0.396103
4 ROW AR1 0.657356 !S2=34.7487 // 50 COL AR1 0.419925
4 ROW AR1 0.557668 !S2=45.4682 // 50 COL AR1 0.267989
4 ROW AR1 0.25576 !S2=14.0717 // 45 COL ID
PREDICT GEN
```

20 Abbreviations

ANOVA	Analysis of variance
BOKU	University of Natural Resources and Life Sciences Vienna
CIMMYT	International Maize and Wheat Improvement Center
CRD	Completely randomized designs
CV	Cross validation
GBS	Genotyping-by-sequencing
GEBV	Genomic estimated breeding values
GS	Genomic Selection
GxE	Genotype by environment
IBD	Incomplete block design
INDEL	Insertions and deletions
IWGSC	International Wheat Genome Sequencing Consortium
MAS	Marker assisted selection
NGS	Next generation sequencing
NIRS	Near-infrared spectroscopy
OBS	Observation set
ÖMV	Österreichische Mühlenvereinigung
PCR	Polymerase chain reaction
PRECW	Pre-commercial winter set
QTL	Quantitative trait loci
RAD	Restriction site associated
RCBD	Randomized complete block design
RCD	Replicate control design
RFLP	Restriction fragment length polymorphisms
RFLP	Random forest
RR-BLUP	Ridge regression best linear unbiased prediction
SNP	Single nucleotide polymorphisms
SPV	Superior progeny value
SVM	SZD Saatzucht Donau
TGW	Thousand grain weight

List of Figures

List of Tables