Data Integration Strategies for Genomic Prediction and Association Studies in Wheat Breeding

Dissertation
zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät III
Agrar- und Ernährungswissenschaften,
Geowissenschaften und Informatik

der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Herrn Lell, Moritz Johannes

Eingereicht am 05.05.2025

Verteidigt am 03.11.2025

Gutachter:
Prof. Dr. Jochen Reif
Prof. Dr. Tobias Würschum

Zusammenfassung

Fortschritte bei der Hochdurchsatz-Genotypisierung und der Phänotypisierung haben die datengestützte Weizenzüchtung vorangebracht. Die daraus gewonnenen phänotypischen und genotypischen Daten lassen auf eine genauere Vorhersage vielversprechender Sortenkandidaten hoffen, wenn eine einheitliche Auswertung über Datensilos hinweg gelingt. In dieser Arbeit wurden integrative Strategien für genomische Vorhersagen bei Weizenhybriden und Inzuchtlinien sowie für die genomweite Assoziationskartierung untersucht. Überlappende Genotypen und gemeinsame Methoden in bestehenden Weizenzuchtprogrammen erwiesen sich als vorteilhaft für die Integration. Es zeigte sich, dass sich bei genomweiten Assoziationsstudien weniger signifikante Assoziationen zwischen Markern und Merkmalen finden ließen als bei individuellen Datensätzen, welche aber eine höherere Vorhersagekraft besaßen. Die Leistungsfähigkeit der genomischen Vorhersage nahm durch integrative Analyse deutlich zu, wobei mit zunehmender Datensatzgröße weitere Fortschritte schwieriger wurden. Dies zeigt, dass die Kombination von Daten über Silos hinweg vorteilhaft ist, aber auch weitere Faktoren verbleiben, die die Vorhersagekraft einschränken. Möglicherweise wurden Wechselwirkungen zwischen Genotyp und Umgebung zu limitierenden Faktoren, die aufgrund der stark unausgewogenen Daten schwer zu erfassen waren. Methodische Innovationen, wie *balanced environmental sampling*, können auf der Grundlage erster Ergebnisse dieser Arbeit weiter erforscht werden.

Abstract

Advancements in genotyping and phenotyping technologies have allowed for progress in data-driven wheat breeding. Resulting phenotypic and genotypic data can improve the prediction of promising variety candidates if a unified evaluation across data silos succeeds. This thesis explores integrative strategies for genomic prediction in wheat hybrids and inbred lines, as well as for genome-wide association mapping. Common checks and methodology in existing wheat breeding programs proved beneficial for integration. Genome-wide association studies were found to yield fewer significant marker-trait associations than for individual data sets, albeit with higher predictive power. The predictive power of genomic prediction increased markedly, showing decreasing additional benefits as dataset sizes grew. This shows that combining data across silos is beneficial, but unresolved factors remain that limit the predictive power. This is potentially due to genotype-times-environment interactions, which are difficult to track due to the strongly imbalanced data. Methodological innovations, like balanced environmental sampling, can be further explored based on initial results from this work.

Table of contents

Abbreviations

ASCII	American Standard Code for Information Interchange
bp	Base pairs
Gb	Gigabases (10^9 bp)
DNA	deoxyribonucleic acid
GBLUP	Genomic best linear unbiased prediction
GWAS	Genome-wide association study
MAF	Minor allele frequency
MAGIC	Multi-parent advanced generation intercross
NAM	Nested association mapping
RKHS	Reproducing Kernel Hilbert Space Regression
RR-BLUP	Ridge regression best linear unbiased prediction
REML	Restricted maximization likelihood
SNP	Single nucleotide polymorphism

1 Introduction

1.1 Breeding for food security in a changing world

Wheat is one of the most important staple crops, providing about 19% of the world's demand in calories either directly as food or as livestock feed, with the relative shares varying strongly between wealthy and poorer countries (Shiferaw et al., 2013). As earth's population is still growing and increasing wealth promises to reduce hunger, the global food demand compared to 2005 is expected to increase by about 30 to 60 percent by 2050 (Tilman et al., 2011; Tian et al., 2021; Van Dijk et al., 2021) due to diet diversification and a trend towards animal-based food. Since the 1960s wheat yields have tripled and current yield gains are about 1.7 percent per year. The main drivers behind this are changed farming practices, in particular much higher nitrogen input and irrigation, and the improvement of wheat genetics by breeding, at about equal shares (Fischer et al., 2022). Raising carbon dioxide concentrations and resulting temperature increase will also further improve yields mostly in high-latitude regions, however, ramifications of climate change will have detrimental impacts on food security overall by various means like increased droughts and floods, damage to pollinator populations, spread of pests, and lower nutritional quality of the harvested food (Intergovernmental Panel On Climate Change, 2022). Moreover, intensive agriculture transgresses several other planetary boundaries besides the impact on the climate system, like those of an intact biosphere and sufficient freshwater, so that the current projection is a „Food for all but not forever" (Tian et al., 2021) world, and a safe livelihood of humanity is endangered. Therefore, agriculture must perform a „U-turn" towards a balance of intensification and resource use (Gerten et al., 2020). Large shares of yield gains since the green revolution have been achieved by increasing the resource usage of agriculture. As those resources are becoming increasingly constrained due to their associated planetary boundaries, breeding to improve the genetic performance of varieties given restricted input gains importance.

In the 20th century, crop breeding has become a dedicated profession, as compared to earlier-time on-farm selection. This has gone together with a steady refinement of breeding methods. For example, breeders started to rely on targeted crosses of wheat varieties, which are normally self-pollinated. This generates some progeny with more extreme phenotypes than their parents (transgressive segregation), which can be selected for to achieve breeding progress (Rasheed and Xia, 2019). Initially, wheat

breeding was mainly driven by public institutions while nowadays the focus is shifting more towards private breeding. Comparison of pedigree and genetic data suggests that public institutions have put more emphasis on sharing genetic material across geographic regions, however, it is not clear whether more recent private programs reuse more of the local elite material or whether pre-breeding using local landraces has partially replaced crossing across geographic boundaries (Fradgley et al., 2019).

1.2 Molecular Breeding

The cycle of crossing, observation and selection has achieved significant progress during the green revolution and afterwards. The most famous innovations among many are a reduction in plant height, allowing for intensive fertilization without inducing lodging, and introducing flowering that is independent of the photoperiod, which allowed for cultivars that could be grown in a wide geographic range. These traits were introduced based on observed phenotypes and later, the genetic basis for this was found in mutated alleles of the *Rht* and *Ppd* gene families, respectively (Trethowan et al., 2007). The elucidation of causal genes for these and many other traits and the advent of cost-effective genotyping technology have led to the development of molecular breeding, which uses genetic data for more accurate selection. One of the first applications was marker-assisted backcrossing, where genetic markers are used to speed up the introgression of desired novel genetic material into a pre-existing elite genetic background, while avoiding introducing undesired properties via linkage drag. Other methods include quality control of seed stocks, evaluation of heterotic groups for hybrid breeding and stacking multiple desired genes by pyramiding (Collard and Mackill, 2008). Especially for disease resistances and quality traits, several markers have been developed that are linked to good performance in these respects or are even functional markers which directly target causal genes (Liu et al., 2012; Hasan et al., 2021; Song et al., 2023).

1.3 Genome-Wide Association Studies

The hunt for genome regions which influence traits of interest initially started with crosses between two or multiple parents that show variation in that trait. In the segregating offspring, linkage blocks of parent genetic material would appear in different combinations. That allowed to determine genetic maps from the linkage disequilibrium of different genetic loci and to associate loci with phenotypes (reviewed by Würschum, 2012). On the downside, the genetic diversity of the mapping population was limited by the choice of the parents. To generate successful mapping populations,

parents must be chosen *a priori* to possess variation in as many interesting genetic loci as possible, even though those loci might not be known. Moreover, large amounts of crosses are necessary to produce sufficient numbers of offspring, in particular when producing more advanced mapping populations like MAGIC or nested association mapping panels (Gage et al., 2020).

In addition to selection bias by parent choice, the lack of linkage between causal variants influencing traits of interest and the available genetic markers challenged mapping efforts, particularly as natural populations of higher genetic diversity were targeted. This was because a higher diversity is associated with smaller linkage blocks, as more meiotic events lie between individuals and their common ancestors than in designed mapping populations (Hamblin et al., 2011). Thus, more diverse populations allow for finer mapping, however the marker density must be sufficient to ensure most linkage blocks are in linkage disequilibrium with some markers. With the availability of dense genetic marker data because of cost-effective high-throughput genotyping technology like DNA microarrays and the dramatic price drop of whole-genome resequencing technologies (Wetterstrand, 2024), it became viable to screen large populations for associations of genomic loci and phenotypes in genome-wide association studies (GWAS). This allowed to screen diversity panels, large numbers of diverse individuals that were available in the domain of interest, instead of self-generated mapping populations, for hitherto unknown marker-trait associations.

The greater potential of diversity panels to detect unknown genetic loci compared to mapping populations comes at the cost of several potential sources of bias in the results. One is that the population structure is not controlled for in a diversity panel. This can lead to spurious associations: Subpopulations can have specific alleles both at causal and non-causal loci, which thereby can become confounded (Sul et al., 2018). Several statistical approaches have been developed to reduce the influence of heterogenous genetic background, of which the mixed linear model approach is the most widespread. This method explicitly models the correlation between trait phenotypes because of their genetic background based on genomic relationship matrices (Yu et al., 2006). Numerous researchers have improved upon this method for computational efficiency and power and have also provided alternative approaches (Tibbs Cortes et al., 2021). In summary, a successful GWAS depends on a population that is diverse enough to cover relevant genetic variation, not too structured to reduce synthetic and correlated associations, covered by genetic markers densely enough to have high linkage between markers and causal alleles, and sampled by enough individuals to determine

effect sizes correctly. A further limiting factor is the heritability of the trait, which can be reduced for example by strong genotype-times-environment interactions and quality or difficulty of the trait measurement.

1.4 Genomic prediction

Instead of detecting causative alleles for certain traits as a basis to predict genotype performance, another approach is to use a training set of genotypes for which both genotypic and phenotypic data is available to derive phenotype predictions for a test set of individuals that are only genotyped. As such methods tend to use many or all of the genomic loci for which data is available, rather than only those with proven influence, they are subsumed under the term genomic prediction. One of the first such models was Ridge Regression Best Linear Unbiased Prediction (RR-BLUP, Meuwissen et al., 2001; Whittaker et al., 2000). It decomposes the observed phenotype into a sum of estimated effects for all genetic loci. Moreover, it assumes that there is a very large number of genetic loci influencing the trait but the influences are mostly small. Assuming further that most of these loci are in strong linkage disequilibrium with at least one of the available genetic markers, a linear mixed model is used with the phenotypes as the independent variable and the marker effects as a random explanatory variable. The variance of the random variable is commonly estimated using the iterative restricted maximum likelihood (REML) procedure (Gilmour et al., 1995). The name of RR-BLUP stems from its equivalence to a ridge regression of the phenotypes on the marker data with the shrinkage parameter λ being determined by the narrow-sense heritability h^2 as $\lambda = 1/h^2 - 1$ (De Vlaming and Groenen, 2015).

From the comparison to ridge regression it can be seen how genomic prediction approaches like RR-BLUP deal with the problem that the number of parameters normally exceeds the sample size, also known as the $p > n$ problem. Without restrictions on the parameter estimates, there would be an infinite number of potential marker effect estimates that explain the observed phenotypes equally well. By introducing the precondition of normally distributed marker effects (from a linear mixed model view) or a penalization of large parameter estimates (from a ridge regression view), the model is driven to prefer smaller parameter estimates, thereby resolving the ambiguity. This also opens the door to numerous alternative model setups. For example, many authors used Bayesian approaches and defined alternative prior probability distributions for the marker effects to mimic more closely the genetic architecture of the traits under study (Gianola, 2013).

Another line of genomic prediction set off from the genetic theory of related individuals, in particular the correlations of the genetic trait components between relatives. Henderson (1988) proposed to use a linear mixed model that decomposes the observed phenotypes into a random additive genetic component (the breeding value), potentially fixed additional group effects, and residual effects. The relatedness of individuals was introduced as the covariance matrix of the additive genetic random variable, which was set to the numerator relationship matrix (the matrix of consanguinity coefficients). The model originated in animal breeding where pedigree records were available to derive the numerator relationship matrix. It became applicable in plant breeding when it became possible to derive estimates of relatedness from marker data (VanRaden, 2008; Hayes et al., 2009a) and became known as Genomic Best Linear Unbiased Prediction (GBLUP). Further theoretical work on the correlation of dominance and epistatic effects allowed extensions in the form of additional random variables representing those effects (Alvarez-Castro and Carlborg, 2007). Additionally, a reformulation inspired by Kernel Ridge Regression that is also known as Reproducing Kernel Hilbert Space Regression (RKHS, Jacquin et al., 2016) allowed to implicitly model multiple levels of epistatic interactions (Gianola and van Kaam, 2008). As genomic prediction lies at the confluence of genetics, frequentist and Bayesian statistics and machine learning, multiple independently invented methods have been shown to be equivalent, for example RR-BLUP and GBLUP given normally distributed marker effects (Hayes et al., 2009b), or multiple Bayesian methods and their classical counterparts (De Vlaming and Groenen, 2015; Jacquin et al., 2016).

There are many factors that influence the accuracy of genomic prediction and some can be tuned to improve predictions. Firstly, the marker density must be sufficient to capture most linkage blocks in the population (Meuwissen, 2009). This means that more diverse populations require more dense sampling of the genome by markers, however an analysis of a large and diverse data set of wheat elite cultivars and land races found that accuracy plateaus at about 5,000 markers (Zhao et al., 2021), which is easily achievable by current technology. Furthermore, additive genetic relationship between the test and the training set is an important foundation of successful Genomic Prediction. This is because this method is not only based on physical linkage between individual causative loci and genetic markers but also extracts information from the total number of causative loci that have co-segregated between individuals in the test set and training set, that is, the additive genetic relationship or pedigree (Gianola et al., 2009; Habier et al., 2007, 2013). As not all of those loci are known, the total fraction

of shared markers between a test and training individual serves at an estimate. Therefore, training sets with individuals that are closely related to the test set individuals confer a large boost in genomic prediction ability as the fraction of shared markers is dominated by co-segregation and thus the estimation of additive genetic relationship is near to reality.

1.5 Balancing sample size and heritability for accurate genomic prediction poses a dilemma given breeding experimental designs

Two further training set properties crucially influence the capabilities of genomic prediction: Training set size and heritability of the trait (Meuwissen, 2009). The higher the diversity of the population in question, the more individuals in the training set are required for an accurate prediction. However, there are multiple approaches on how to compute how large that number of required individuals is, and the resulting estimates vary substantially (Brard and Ricard, 2015). The relationship between the trait heritability and the genomic prediction accuracy is more straightforward, as a lower heritability means that measured phenotypic values of the training set have a smaller genetic component compared to other influencing factors, that are not genetic and thus non-heritable. Beyond additive genetic effects, more complex genetic interactions and genotype-by-environment interactions, could be tackled by more advanced genomic prediction models, advances in phenotyping methodology, or specialized experimental designs, as reviewed by for example Voss-Fels *et al.,* (2019). However, especially for complex traits, repeated observation in different environments lies at the heart of a robust picture of a candidate's performance.

These two requirements of genomic prediction, high sample size and high heritability, turn out to be conflicting when considering the layout of a multi-generational and multi-environmental plant breeding program. These programs evaluate a large number of progenies of initial crosses with large genetic diversity. Over multiple generations, favourable candidates are selected and their offspring (in wheat, by self-fertilization) re-sown. As the number of candidates to test is high in early generations and is reduced with every following generation, the remaining candidates can be tested in an increasing number of environments (Figure 1A). This results in a higher heritability of the observed phenotypes, as genetic performance becomes more clearly visible with environment-specific effects averaging out. At the same time, the number of candidates sharply declines over generations, so that across a program, for most individuals only

low-heritability observations are available (Figure 1B). The merging of data sets is therefore a promising strategy (Zhao et al., 2021).

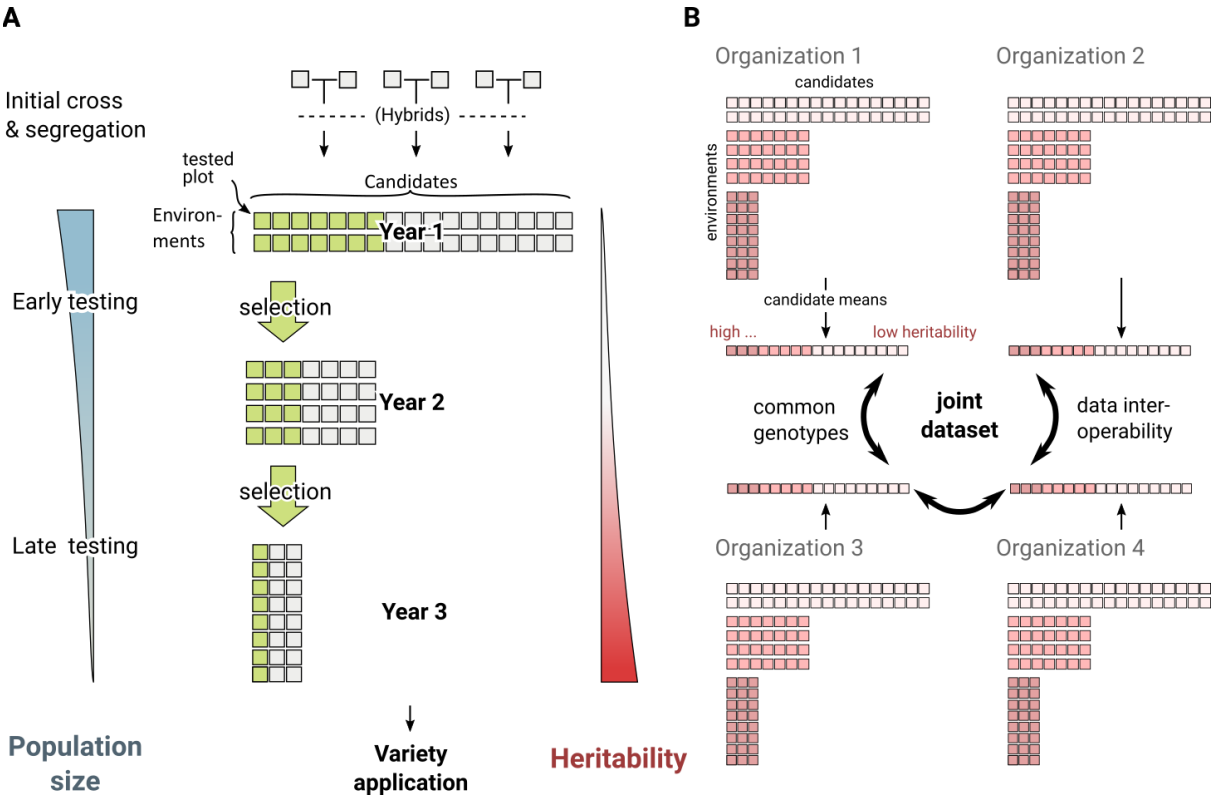

Figure 1: *Potentials to form joint data sets across data silos.* ***A:*** *Schematic structure of a variety breeding program. The progression of generations from crossing, over multi-environmental field trials and selection, to variety application is shown from top to bottom. The decline of population size and increase of heritability towards later generations is sketched in the left and right margins.* ***B:*** *Concept of joining candidate information across multiple data silos, for example public and private breeding programs. Each program provides estimates of variable heritability due to its internal experimental structure that allows to envision different merging strategies.*

1.6 Objectives

The field of potential strategies for merging data sets is large and it includes the case studies of genomic prediction and GWAS on merged data sets, as well as explorations of altered experimental designs that cater for evaluation pipelines based on genomic prediction. In this thesis, my objectives were to

(1) find the best of several ways to sparsely distribute a fixed number of plots across different environments to increase the accuracy of predicting hybrid performance,

(2) compare several commonly used biometrical models for phenotypic data analysis and identify the one that most accurately estimates the performance of the candidates in such unbalanced trials,

(3) examine the impact of including a common set of genotypes across all environments,

(4) determine whether, integrating and re-evaluating breeding data sets improves the detection power of GWAS,

(5) evaluate quality-control measures including a cross-validation using genomic prediction,

(6) connect differences in the GWAS results between individual experimental series and the combined data to the established theory,

(7) investigate whether it is possible to perform an integrated analyses of disparate phenotypic and genotypic data sets and how to perform quality control of such a task,

(8) examine what prediction abilities can be expected when using genomic prediction beyond the confines of individual experimental series and how well multiple series can be combined to form larger training sets for genomic prediction,

(9) test approaches to improve the training set by drawing subsets from the full data, distilling the most reliable data and potentially increasing prediction ability.

2 Peer-reviewed scientific articles

2.1 Optimizing the setup of multienvironmental hybrid wheat yield trials for boosting the selection capability

Published: In 2021, *The Plant Genome* 14(3):e20150.

DOI: 10.1002/tpg2.20150

Authors: Moritz Lell, Jochen Reif, Yusheng Zhao

Abstract:

The accuracy of genomic prediction increases with increasing heritability, and thus the challenge of optimizing the design of multienvironment yield trials under a limited budget arises. With this in mind, we aimed to find the best of several options to sparsely distribute a fixed number of plots across different environments to increase the accuracy of hybrid performance prediction. We used a comprehensive published genomic and phenotypic data set of 1,604 winter wheat (Triticum aestivum L.) hybrids and compared several commonly used biometric models for phenotypic data analysis in a resampling study to identify the one that most accurately estimated the hybrid performance in different imbalanced trials. Our results showed that when using information about genotypic relationships, genotypic values were more strongly associated with the reference values than when this information was ignored. In addition, a balanced environmental sampling resulted in an adequate characterization of each environment and increased the accuracy for estimating the hybrid performance. One promising design involved dividing the genotypes into equally sized subgroups that were tested in a subset of environments, with the constraint that the subgroups overlapped with respect to the environments. This scenario appears to be particularly appropriate, as it provided both high accuracies in the estimates of genotypic values and had low variability resulting from the data sample used. Thus, we were able to clearly demonstrate the utility for optimizing the design of multienvironment hybrid wheat yield trials in times of genomic selection.

Received: 13 May 2021 | Accepted: 22 July 2021

DOI: 10.1002/tpg2.20150

The Plant Genome

SPECIAL ISSUE: ADVANCES IN GENOMIC SELECTION AND APPLICATION OF MACHINE LEARNING IN GENOMIC PREDICTION

Optimizing the setup of multienvironmental hybrid wheat yield trials for boosting the selection capability

Moritz Lell | **Jochen Reif** | **Yusheng Zhao**

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland D-06466, Germany

Correspondence
Jochen Reif, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany, D-06466.
Email: reif@ipk-gatersleben.de

Assigned to Associate Editor Rajeev Varshney.

Abstract

The accuracy of genomic prediction increases with increasing heritability, and thus the challenge of optimizing the design of multienvironment yield trials under a limited budget arises. With this in mind, we aimed to find the best of several options to sparsely distribute a fixed number of plots across different environments to increase the accuracy of hybrid performance prediction. We used a comprehensive published genomic and phenotypic data set of 1,604 winter wheat (*Triticum aestivum* L.) hybrids and compared several commonly used biometric models for phenotypic data analysis in a resampling study to identify the one that most accurately estimated the hybrid performance in different imbalanced trials. Our results showed that when using information about genotypic relationships, genotypic values were more strongly associated with the reference values than when this information was ignored. In addition, a balanced environmental sampling resulted in an adequate characterization of each environment and increased the accuracy for estimating the hybrid performance. One promising design involved dividing the genotypes into equally sized subgroups that were tested in a subset of environments, with the constraint that the subgroups overlapped with respect to the environments. This scenario appears to be particularly appropriate, as it provided both high accuracies in the estimates of genotypic values and had low variability resulting from the data sample used. Thus, we were able to clearly demonstrate the utility for optimizing the design of multienvironment hybrid wheat yield trials in times of genomic selection.

1 | INTRODUCTION

The use of genomic selection has led to several modifications in wheat (*Triticum aestivum* L.) breeding programs. One major change is the increase in the value of historical genomic and phenotypic data from previous breeding cycles, which are used in genomic selection to train the genome-wide prediction models (Storlie & Charmet, 2013). The use of historical genomic data is not difficult as long as the genotyping

Abbreviations: BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; GBLUP, genomic best linear unbiased prediction; GCA, general combining ability; GEGV, genomically estimated genotypic value; PBLUP, pedigree-based best linear unbiased prediction; PEGV, phenotype-based estimate of the genotypic value; S-All, all hybrid observations of data set; SCA, specific combining abilities; SNP, single nucleotide polymorphism.

platform remains largely unchanged or new developments are compatible with those previously used. This was the case, for example, when the 9k single nucleotide polymorphism (SNP) array (Cavanagh et al., 2013) in wheat was updated by the 90k SNP array (Wang et al., 2014). In contrast, using historical phenotypic data to train genome-wide prediction models can be challenging.

Historical phenotypic data are typically not orthogonal across years (e.g., Storlie & Charmet, 2013) and come from multistage selection programs in which extensive populations are tested in a few environments and a selected fraction are then evaluated in a large number of environments. On the one hand, data from early selection cycles are valuable to train the genome-wide prediction models because they have a large phenotypic diversity (He et al., 2017); on the other hand, genotypes in early selection cycles are tested in only a few environments, which reduces the heritability and thus the accuracy of the genome-wide prediction models (He et al., 2017). The optimal design of a wheat breeding program that uses genome-wide predictions can lead to adjustments to increase the number of environments used for yield testing in early selection cycles as a function of the accuracy of genome-wide predictions (Longin et al., 2015). Nevertheless, the number of environments is limited by the moderate multiplication coefficient in wheat, limiting the amount of seed available for the next stage. This also hampers the implementation of phenotype imputation, where indirect traits are used to enable sparse yield testing (e.g., Rutkoski et al., 2016; Ward et al., 2019). An important side effect of a small number of environments used in early yield tests is a potential underrepresentation of the diversity of environments, which can lead to a bias in the decomposition of phenotypic variance into effects of genotypes, environments, and their interaction effects (Utz et al., 2000).

Extensive research has been conducted to optimize the design and biometric analyses of plant breeding trials (Cullis et al., 2006; Patterson & Williams, 1976; Smith & Cullis, 2018; Yates, 1940). The use of relationship information has become an important underpinning for crop candidate performance evaluation, as observations of one genotype can inform decisions about related genotypes. Estimations of genotypic values that take known relationships into account can be modeled using best linear unbiased prediction (BLUP). As the genotypic value is therein modeled using one or more random variables, relationship can enter the model in two ways: Either explicit covariance structures for genotype-related random variables are defined or related genotypes are defined to share common effects. Details about both approaches can be found for example in a review by Piepho et al. (2008). An example of the first approach is the decomposition of the genotypic effect into an additive and dominance component, for which expected covariances result from quantitative genetic theory or are approximated using genomic markers (Alvarez-Castro & Carlborg, 2006; VanRaden, 2008). The second approach is suitable for evaluation of hybrids, where the genotypic effect of a candidate is decomposed into the sum of the general combining abilities (GCAs) of its two parents and the specific combining ability (SCA) of the specific cross. The GCA of potential parents and thus the most beneficial future crosses are often predicted using crosses to some tester lines, but Seye et al. (2020) recently pointed out that a sparse factorial crossing of parent lines might be superior. This example shows how GCA/SCA-based models incorporate relationship information without defining a covariance structure of the individual effects. By use of relationship information, BLUP allows an increased flexibility in breeding trial layout as some genotypes can be phenotyped in fewer environments or even not at all (Longin et al., 2015). Computationally, these models are fitted by restricted maximum likelihood estimation (Gilmour et al., 1995) or as a special case of reproducing kernel Hilbert space regression (de los Campos et al., 2009).

Core Ideas

- In preliminary trials, evaluating candidates sparsely in more environments improves accuracy.
- A prerequisite for that are biometric models that consider candidate relatedness.
- Joint analysis of concurrent trials can improve accuracy without logistic changes.

The potential to increase the diversity of environments in early yield trials in the context of genomic selection was suggested in a pioneering resampling study using genomic and phenotypic data from a biparental family for barley (*Hordeum vulgare* L.) and maize (*Zea mays* L.; Endelman et al., 2014). The authors observed that genome-wide prediction models improved when genotypes were distributed across multiple locations rather than testing all entries in one location and concluded that genome-wide markers in such imbalanced designs establish connectivity by modeling relatedness among entries. Also, Jarquin et al. (2020) demonstrated in two sparse three-environment maize testcrosses that the prediction accuracy of genome-wide prediction models can benefit from connecting the environments with common genotypes. However, an in-depth study on the potential to optimize the design of multienvironmental yield trials in times of genomic selection in diverse populations is lacking. This is very promising, for example, in hybrid breeding with a pronounced degree of relatedness resulting from factorial mating designs.

Our survey is based on a comprehensive published data set of wheat hybrids comprising 1,604 single crosses and their 135 parental lines. We used a resampling strategy to investigate the optimum allocation of resources in multienvironmental field trials assuming restricted budgets. In particular, our

objectives were (a) to find the best of several ways to sparsely distribute a fixed number of plots across different environments to increase the accuracy of predicting hybrid performance, (b) to compare several commonly used biometrical models for phenotypic data analysis and identify the one that most accurately estimates the performance of the candidates in such unbalanced trials, and (c) to examine the impact of including a common set of genotypes across all environments.

2 | MATERIALS AND METHODS

All calculations have been performed using R 4.0.2 (R Core Team, 2020).

2.1 | Phenotypic data

Our study is based on previously published phenotypic and genomic data of an F_1 hybrid population resulting from 1,604 out of 1,800 potential single crosses with sufficient seeds for multienvironmental yield trials between 120 female and 15 male winter wheat elite lines that were bred for wheat growing in central Europe (Jiang et al., 2017; Longin et al., 2013; Zhao et al., 2015). Briefly, the hybrids, their parents, and 10 check cultivars were phenotyped for grain yield in 2 year (2012 and 2013) and six locations (Supplemental Figure S4, 11 location-year combinations in total). The genotypes were divided randomly for each environment into three trials for reasons of field management. The trial designs were partially replicated alpha lattice designs that were connected by 10 replicated check cultivars. The same seeding rate was used for lines and hybrids. Grain yield was adjusted to a moisture concentration of 140 g H_2O kg^{-1}. Further environment details are shown in Supplemental Tables S1 and S2. The parental lines were fingerprinted using a 90K SNP array (Würschum et al., 2013; Zhao, Gowda, et al., 2013). The SNP profiles of the hybrids were derived from the information of the SNP profiles of the parental lines. Our study is based on 16,937 polymorphic SNPs. The phenotypic and genomic data was used in an in silico study to investigate the optimum allocation of resources in multienvironmental field trials assuming restricted budgets. The available hybrid observations of the data set are shown in Supplemental Figure S3.

2.2 | Phenotypic values adjusted for experimental design effects

As a first step, we analyzed the raw data using the following linear mixed model:

$$y_{ijklm} = g_i + e_j + t_{jk} + r_{jkl} + b_{jklm} + \varepsilon_{ijklm} \quad (1)$$

decomposing each observation y_{ijklm} into a genotypic value g_i for each genotype i, an environment effect e_j for each environment (year × location) j; a trial effect t_{jk} for each trial k, nested within the environment; a replication effect r_{jkl} for each replication l, nested within the trial; a block effect b_{jklm} for each block m, nested within the replication; and a residual error term for each observation. The genotypic values $\mathbf{g}$ are modeled as a fixed effect and all other effects as random; that is, $\mathbf{x} \sim N(0; \mathbf{I}\sigma_x)$ for each $\mathbf{x}$ in $\mathbf{e}$, $\mathbf{t}$, $\mathbf{r}$, $\mathbf{b}$, $\boldsymbol{\varepsilon}$.

The model was fitted using AsReml (Gilmour et al., 2015). The estimated design effects for $\mathbf{t}$, $\mathbf{r}$, $\mathbf{b}$ were then subtracted from the raw data. The resulting data set $\mathbf{y}^{(\mathbf{d})}$, which has the same size as $\mathbf{y}$, was then used for further analysis. The clustering of environments was analyzed in detail in a previous study (See Figure Supplementary Note f-2 in Zhao et al., 2015) and the absence of distinct clusters was reported.

2.3 | In silico scenarios of allocation of multienvironmental field trials

As a reference, all hybrid observations of our data set (S-All) were used to compare different schemes of allocating plots in multienvironmental hybrid grain yield trials. We then defined six in silico scenarios (S1–S6) that contained only 38% of the observations of the S-All data set (Supplemental Methods). The number of plots corresponds to grain yield assessment of the whole population in about four environments. The yield evaluations in four environments represents a typical scenario for resource allocation for first multienvironmental yield trials in Central European wheat breeding programs.

The first scenarios were designed for conceptual simplicity, modifying as few aspects of the data set as possible at the same time (Figure 1):

- S1: With this scenario we tested the influence of reducing the number of observations in a balanced fashion. It was generated as a random sample from the data set and included data of all 11 environments. The sampling was constrained in that all environments had a similar number of observations, and all genotypes were tested in a similar number of environments.
- S2: With this scenario we tested the influence of a reduced number of environments. It was generated as a balanced data set that consisted of data from four environments randomly selected from the full data set.

The remaining four scenarios were designed to resemble field trials of potential practical relevance to breeders:

- S3: The set of genotypes was divided in five groups of genotypes. Each group was tested in four environments. The groups overlapped with respect to the environments. In total, the scenario included 10 environments.

- S4: One-fifth of the genotypes were tested in 10 environments. The remaining genotypes were clustered into four groups that were evaluated in separate sets of environments. This pattern may arise from combining trials from early and late stages of two multistage selection programs that are running concurrently, where one program has progressed farther.
- S5: One-fifth of the genotypes were tested in 10 environments. The remaining genotypes were tested in up to three environments. As for S4, this pattern may arise from combining early- and late-stage trials.
- S6: Half of the genotypes were tested in 4 to 10 environments, the other half in 1 to 3 environments. This pattern was repeated for the hybrid offspring of each female parent so that for every female parent, some offspring were tested in many and some were tested in few environments.

2.4 | Predicting and estimating reference genotypic values

To have a benchmark for the different scenarios, we obtained estimates or predictions of the performance of the hybrids using the data set S-All applying two statistical models.

We obtained the best linear unbiased estimations (BLUEs), fitting the following linear model:

$$y^{(d)}_{ijklm} = g_i + e_j + \varepsilon_{ijklm} \quad (2)$$

where $\boldsymbol{y}^{(\mathbf{d})}$ is the observations without design effects obtained following Model 1. The estimated/predicted variables are $\mathbf{g}$, a vector of fixed genotype effects and $\mathbf{e}$ and $\boldsymbol{\varepsilon}$, vectors of random environment and residual effects, respectively. As in Model 1, the covariance structure of the random effects is the identity matrix.

The model was fitted using AsReml 4 (Gilmour et al., 2015). The BLUEs from the S-All data set can be interpreted as the purely phenotype-based estimates of the genotypic value (PEGV).

In a second approach, the following random linear model was fitted that decomposes the effect of the genotype into additive and dominance components. This model is referred to as GBLUP (genomic best linear unbiased prediction) in the following:

$$y^{(d)}_{ijklm} = a_i + d_i + e_j + \varepsilon_{ijklm} \quad (3)$$

The estimated effects include the additive ($\mathbf{a}$) and dominance ($\mathbf{d}$) genetic effects. The notation of the other effects is analogous to Model 2.

The two genetic effects had covariance structures derived from the SNP profile of the hybrids ($\mathbf{a} \sim N[0, \mathbf{K_a}\sigma_a]$ and ($\mathbf{d} \sim N[0, \mathbf{K_d}\sigma_d]$). The SNP status of the hybrids was calculated as the mean of the respective parents' SNP status. Then, the covariance structure of the additive effect $\mathbf{K_a}$ was calculated according to VanRaden (2008, p. 4116, "first method"). The dominance effect had a covariance structure $\mathbf{K_d}$ according to Alvarez-Castro and Carlborg (2006).

We fitted this model using the R package BGLR (Pérez & de los Campos, 2014) that solves mixed models as a special case of reproducing kernel Hilbert space regression in a Bayesian framework, using a Gibbs sampler. We used 5,000 iterations of which 900 iterations were treated as burn-in phase. The performance of the Gibbs sampling was checked by calculating the effective sample size of the resulting Markov chains, which corrects the number of iterations with the autocorrelation of the chain, as well as the Geweke (1992) diagnostic, which checks whether the empirical distributions of parameter values at the beginning and the end of the chain are different. The effective sample sizes were high enough to calculate the predicted values by taking the mean of the iteration steps (Supplemental Figure S1A). For the majority of model runs, the Geweke scores were within what was expected from a standard normally distributed random variable. This result conforms to the Markov chains converging to a stationary distribution and thus the burn-in phase being long enough (Supplemental Figure S1B). We also manually checked the Markov chain trace of the GBLUP of S-All wherein the convergence to a stationary distribution well within the burn-in phase can be seen (Supplemental Figure S2).

The obtained predictions, when done on the S-All data set, can be interpreted as the genomically estimated genotypic value (GEGV).

2.5 | Statistical models to predict genotypic values from in silico scenarios

We obtained genotypic values from the data sets generated according to the six in silico scenarios described above and S-All. The statistical models used for this were BLUE (Model 2), GBLUP (Model 3) and two additional models that are described in this section.

In the first model, we obtained the BLUP using the following model with two random terms assuming no specific covariance structure between the genotypes:

$$y^{(d)}_{ijklm} = g_i + e_j + \varepsilon_{ijklm} \quad (4)$$

The model is similar to Model 2 and the same terms are used with the only difference that $\boldsymbol{g}$ is a random variable with $\mathbf{g} \sim N(0, \mathbf{I}\sigma_a)$.

In the second model, we decomposed the hybrid performance into GCA and SCA effects by fitting the following random effect model:

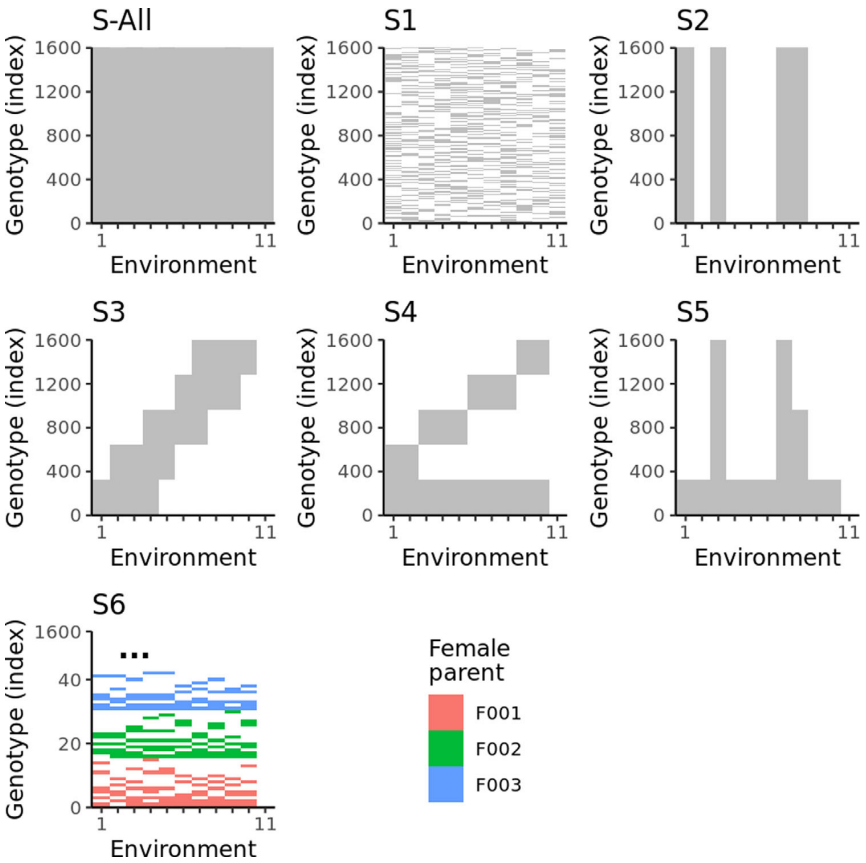

FIGURE 1 Graphical illustration of how genotypes were assigned to different environments in the six in silico scenarios (S1–S6). A filled square indicates that a genotype was evaluated in an environment. In S6, the female parent of the genotype is indicated by the color. The plot for S6 shows the offspring of the first 3 of the 120 female parents for a clear display of the scenario structure

$$y^{(d)}_{ijklm} = f_{p(i)} + m_{q(i)} + s_i + e_j + \varepsilon_{ijklm} \quad (5)$$

As before, the term $\mathbf{y^{(d)}}$ is the observations without design effects (Model 1). All factors are random without covariance structure analogous to Model 1. The parental GCAs are **f** and **m** that have one value for each parent genotype. The functions $p(i)$ and $q(i)$ yield the female and male parent index for the hybrid genotype i, respectively. The SCA is **s**, with one entry for each hybrid genotype i. The remaining effects (environment and residual) are defined similarly to the previous models.

This model takes the pedigree of the hybrids into account implicitly by estimating the influences of their common parents. Thus, we call it pedigree-based BLUP (PBLUP) in this text. Both Models 4 and 5 do not use SNP data. Nevertheless, Model 5 considers relatedness, as multiple hybrids share the same parent and thus common effects. If Model 5 would be constrained in that σ_f and σ_m must be the same value, the effects **f** and **m** could be coalesced to one random variable whose covariance structure would be the matrix of coefficients of coancestry of the hybrids, assuming a population of unrelated parental genotypes. The Models 4 and 5 were fitted using AsReml 4 (Gilmour et al., 2015)

2.6 | Evaluating estimates of genotypic values obtained from in silico scenarios

We evaluated the precision of the estimated/predicted genotypic values of the hybrids obtained by Models 2–5 in the Scenarios S1–S6 by correlating these predictions to the reference values, the PEGVs and GEGVs. As each scenario was realized 100 times (Supplemental Methods), for each scenario and model 100 correlation values depending on the data sample result, from which the median and the quartiles were taken.

Moreover, we estimated the reliability with which the best genotypes, according to their PEGV and GEGV, respectively, are found in the best genotypes of the in silico scenarios. We considered for this the 10% best genotypes in terms of PEGV and GEGV, respectively. We then measured the percentage of

realizations for each scenario and model in which these genotypes were also among the top 10%. As an example, say 100 hybrid genotypes are studied. From selecting the 10% best by their GEGVs and PEGVs, respectively, two sets of 10 genotypes result. They represent the breeder's choice if the full data set would be available to them. Let us call these two sets G* and P*. To judge a certain scenario and model against these standards, for each of the 100 scenario realizations the best 10 genotypes are taken, resulting in a set of sets S with 100 elements and each element being a set of 10 genotypes. For every genotype in G*, it is now calculated in how many elements of S this genotype is present.

2.7 | Joint analysis of different series of trials conducted in the same environments

In the Scenarios S4 and S5, not all genotype groups were tested in the same number of environments. Rather, the first four genotype groups were tested in two to three environments while the fifth genotype group was tested in 10 environments.

This situation is similar to an environment in which breeders test two trials where one trial is in a later stage than the other. The late-stage trial is tested in more environments but comprises fewer genotypes than the earlier trial. We explored whether the accuracy of the early-stage trial can profit from the inclusion of the concurrent late-stage trial into a joint analysis.

To answer this question, we split the genotypes of the Scenarios S4 and S5 in two groups, respectively. The genotypes that were measured in two to three environments were designated as early-stage genotypes. When referring to the early-stage genotype group of a specific scenario, we use the suffix "e" (e.g., S4e and S5e). Similarly, the genotypes that were measured in 10 environments were designated as late-stage genotypes, with the suffix "l" (e.g., S4l and S5l).

We focused on the correlations of the yield predictions of the early-stage genotypes S4e and S5e to the PEGVs and the GEGVs where the predictions were generated with the same collection of models as in the previous section. The data available to the model were either the early-stage genotypes only, or the early-stage plus the late-stage genotypes. Regardless of the data available to the model, only the correlation of the early-stage genotypes to the PEGVs/GEGVs was considered.

3 | RESULTS

3.1 | Definition of genotypic values used as references

For the S-All data set, all biometric models yielded very similar estimations/predictions of the hybrid performances

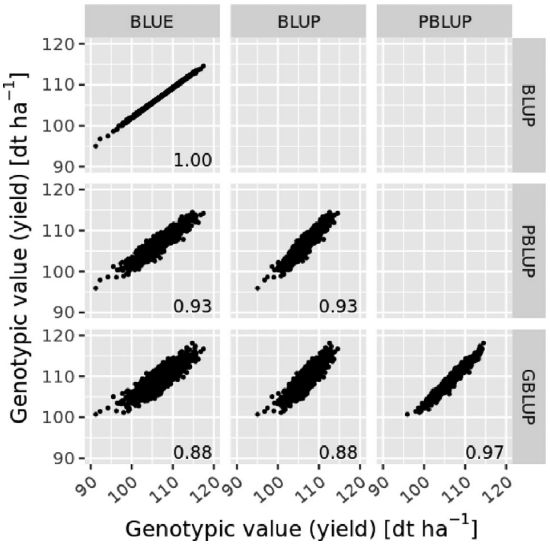

FIGURE 2 Correlations of hybrid performance predictions/estimates of different models using the S-All data set. The models are shown on the plot borders. See the model equations in the Materials and Methods section for their definitions: best linear unbiased estimation (BLUE, Model 2), best linear unbiased prediction (BLUP, Model 4), pedigree-based best linear unbiased estimation (PBLUP, Model 5), genomic best linear unbiased prediction (GBLUP, Model 3). The predictions/estimations of BLUE and GBLUP in this figure are the phenotypically estimated genotypic values (PEGVs) and genomic estimated genotypic values (GEGVs), respectively. The Pearson correlation coefficients are shown in each plot

(Figure 2). The PEGVs and GEGVs, estimations/predictions resulting from the BLUE and GBLUP models, respectively, correlated to each other with 0.88. The BLUPs correlated very strongly to the PEGVs (>0.99). Moreover, PBLUP showed correlations of 0.93 to 0.97 to the predictions from the other models. Summarizing, the values from BLUE and GBLUP form the two extremes of the studied predictions as their correlations to each other was the lowest of all models. They therefore are complementary approaches to define the genotypic values used as references for the in silico scenario predictions. This is elaborated in more detail in the discussion.

Genotypic values obtained with the PBLUP and GBLUP models were more strongly associated with the reference values than those of the BLUE and BLUP models

Unless otherwise stated, we focused on the median of correlations between estimates/predictions and genotypic reference values for comparisons below. Also, we refer to the genotypic value predictions/estimations of a model from an in silico scenario by writing the model name with an "s" appended (e.g., "BLUEs").

Across all scenarios, the BLUEs never outperformed the BLUPs. Depending on the choice of reference values and scenario, the correlations for Scenarios S4 to S6 were 0.03 to 0.06

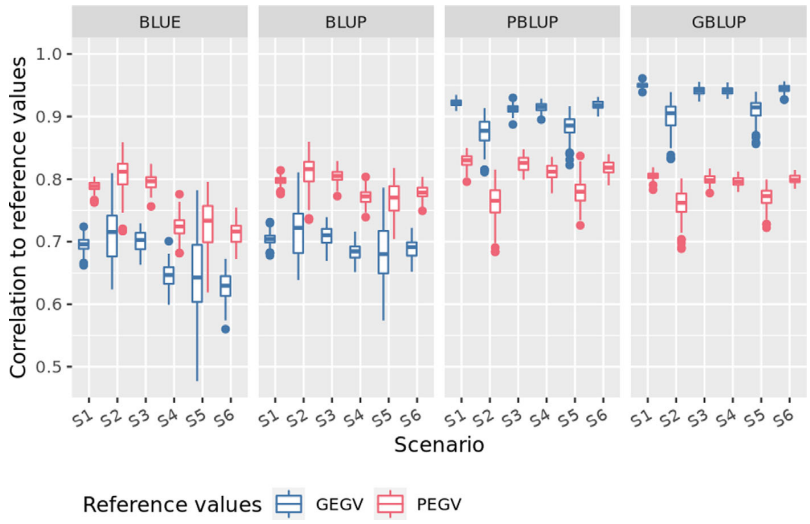

FIGURE 3 Correlations of the hybrid performance predictions/estimations of different models and scenarios to the phenotypically estimated genotypic values (PEGVs) and genomic estimated genotypic values (GEGVs) (colors). The models are shown above the plots. For the structure of the scenarios, see Figure 1. Each plot shows the correlations resulting from one model in the different scenarios. Each box plot summarizes 100 realizations. BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; PBLUP, pedigree-based best linear unbiased estimation; GBLUP, genomic best linear unbiased prediction

TABLE 1 Median correlations of the hybrid performance predictions/estimations of the scenarios and models to the reference values ("Ref.")

Ref.	Scenario	Model order	BLUE	BLUP	PBLUP	GBLUP
GEGV	S1	BLUE = BLUP < PBLUP < GBLUP	0.70	0.70	0.92	0.95
GEGV	S2	BLUE = BLUP < PBLUP < GBLUP	0.72	0.72	0.88	0.91
GEGV	S3	BLUE = BLUP < PBLUP < GBLUP	0.70	0.71	0.91	0.94
GEGV	S4	BLUE < BLUP < PBLUP < GBLUP	0.65	0.68	0.92	0.94
GEGV	S5	BLUE < BLUP < PBLUP < GBLUP	0.64	0.68	0.89	0.91
GEGV	S6	BLUE < BLUP < PBLUP < GBLUP	0.63	0.69	0.92	0.95
PEGV	S1	BLUE = BLUP = GBLUP < PBLUP	0.79	0.80	0.83	0.81
PEGV	S2	GBLUP = PBLUP < BLUE = BLUP	0.81	0.82	0.77	0.76
PEGV	S3	BLUE = GBLUP = BLUP < PBLUP	0.80	0.81	0.83	0.80
PEGV	S4	BLUE < BLUP < GBLUP < PBLUP	0.72	0.77	0.81	0.80
PEGV	S5	BLUE < BLUP = GBLUP = PBLUP	0.73	0.77	0.78	0.77
PEGV	S6	BLUE < BLUP < GBLUP < PBLUP	0.72	0.78	0.82	0.80

Note. The column „Model order" summarizes which models lead to higher correlations for the same scenario. Any difference above 0.01 is denoted by "<", smaller differences by "=". BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; PBLUP, pedigree-based best linear unbiased estimation; GBLUP, genomic best linear unbiased prediction; GEGV, genomic estimated genotypic value; PEGV, phenotypically estimated genotypic value.

lower for the BLUEs than for the BLUPs and approximately the same for Scenarios S1 to S3.

Within the scenarios, the predictions of the different biometric models correlated more strongly with the reference values of the model to which they are conceptually most similar: The BLUEs and BLUPs from the scenarios are stronger correlated to the PEGVs (0.72–0.82) than to the GEGVs (0.63–0.70). Conversely, the PBLUPs and GBLUPs are more similar to the GEGVs (0.88–0.95) than to the PEGVs (0.76–0.83; Figure 3 and Table 1).

Regardless, averaged across all scenarios, PBLUPs or GBLUPs showed a higher correlation to the reference values PEGV and GEGV than BLUEs or BLUPs: the correlation to the GEGVs is about 0.2 higher for the PBLUPs and GBLUPs than for the BLUPs and BLUEs. The correlations to the PEGVs, on the other hand, differed less as a function of the model, but again, in most cases PBLUPs and GBLUPs correlated equally well or stronger to the PEGV in most of the cases than predictions by BLUE or BLUP (difference 0.00–0.08) The only case where this model ranking was less clear was in Scenario S2, in which BLUEs and BLUPs predictions

correlated with the PEGVs with 0.81 on average, whereas the PBLUPs and GBLUPs predictions correlated with 0.76 on average. Considering the GEGV correlation, however, the GBLUPs or PBLUPs of Scenario S2 performed about 0.2 higher off than the other two models. In conclusion, in our study, the genomic models PBLUP and GBLUP were at an advantage against BLUE and BLUP.

3.2 | Balanced environmental sampling boosts the accuracy of PBLUP and GBLUP

The choice of the genotypic reference value did not affect which in silico scenario yielded the best correlations to the reference values for a certain model (Figure 3). In contrast, the choice of biometric model did impact the optimal choice of scenarios. Here, the results of BLUE and BLUP were similar, and the results of PBLUP and GBLUP were also similar.

The in silico Scenario S2 was, on average, the scenario with the highest correlations of the BLUEs and BLUPs to the genotypic reference values. However, there was a high variation among the individual realizations. The predictions for the Scenarios S1 and S3 had only slightly lower correlations (differences of the medians of 0.02 or less) but the variance among realizations was much lower for these scenarios. The BLUPs and BLUEs from the remaining Scenarios S4 to S6 were 0.01 to 0.10 lower than those from the other scenarios (Figure 3).

For the PBLUPs and GBLUPs, the Scenario S1 showed the highest correlation for both reference values than the predictions of the other scenarios (Figure 3). However, Scenarios S3, S4, and S6 also nearly reached this level with differences between the correlations of less than 0.02. The PBLUPs and GBLUPs of the Scenarios S2 and S5 were less correlated with the reference than the predictions from S1, S3, S4, and S6 with differences ranging from 0.02 to 0.06.

In summary, the performance of the BLUE and BLUP benefited when the genotypes were tested in a comparable number of environments (Scenarios S1, S2, and S3). In contrast, the PBLUP and GBLUP benefited when each environment had a comparable number of observations, adequately characterizing each environment (Scenarios S1, S3, S4, and S6).

3.3 | Selection by PBLUPs or GBLUPs had the highest average probability to recover the top genotypes

Up to this point, we have evaluated the merits of the models and scenarios by looking at the entire population. However, breeders are more interested in finding the top or a proportion of the best genotypes. For example, if we consider the best 10% of genotypes, the following picture emerges. (a) Taking PEGV as reference genotypic values, no large differences were found between the biometric models for finding the best 10% genotypes. The values varied between 49 and 60% (Figure 4A and Table 2). Except for Scenario S2, PBLUP had slight advantages of 3–5 percentage points compared with BLUP. When the selection intensity was increased by, for example, to the top 2%, 82–90% were identified on average across the scenarios by PBLUP or GBLUP, but only 71–84% by BLUP (see the moving window averages in Figure 4A). (b) When considering the GEGV as reference values, the picture was different, and the choice of the biometric model strongly influenced the results. The recovery rate of the best 10% of genotypes was 45–48% when analyzed with BLUP, whereas the recovery rate was markedly higher with 65–78% for the PBLUP and GBLUP models (Figure 4B and Table 2). Summarizing, the PBLUP and GBLUP models were superior to BLUP and BLUE in that they either yielded a higher rate of recovered top 10% genotypes from the scenarios or, in cases where this figure was similar between the models, the advantage of the PBLUP and GBLUP models over BLUP increased when the group of top genotypes was defined increasingly smaller.

TABLE 2 Mean percentage of realizations in which the top 10% of genotypes by PEGV or GEGV, respectively, were ranked among the top 10% of scenario (S) predictions

Reference	Model	S1	S2	S3	S4	S5	S6
PEGV	BLUE	54.8	56.5	55.5	49.0	48.8	49.2
PEGV	BLUP	55.5	56.9	56.1	53.4	53.1	55.7
PEGV	PBLUP	60.4	52.9	59.4	59.4	55.6	59.8
PEGV	GBLUP	56.4	51.9	54.7	55.9	53.2	56.2
GEGV	BLUE	46.3	47.1	46.9	42.0	42.4	42.5
GEGV	BLUP	46.8	48.0	47.7	45.2	45.0	46.4
GEGV	PBLUP	73.0	65.4	71.2	72.2	66.8	72.5
GEGV	GBLUP	78.0	68.5	75.6	76.6	70.5	76.6
Mean		58.9	55.9	58.4	56.7	54.4	57.4

Note. BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; PBLUP, pedigree-based best linear unbiased estimation; GBLUP, genomic best linear unbiased prediction; GEGV, genomic estimated genotypic value; PEGV, phenotypically estimated genotypic value.

Comparing different in silico scenarios, which reflect different allocations of plots in a multienvironmental field trial, the predictions from Scenarios S1 and S3 were the most consistent, as they showed high proportions of recovered top genotypes compared with the predictions of other scenarios (Table 2). In Scenario S2, the aforementioned advantage of GBLUP and PBLUP over BLUP and BLUE was the smallest. In recovering the best PEGV genotype, BLUP outperformed PBLUP by 4 percentage points in S2.

However, in summary, the choice of the scenario was less important than the choice of the biometric model. The PBLUP and GBLUP models were superior to the BLUE and BLUP

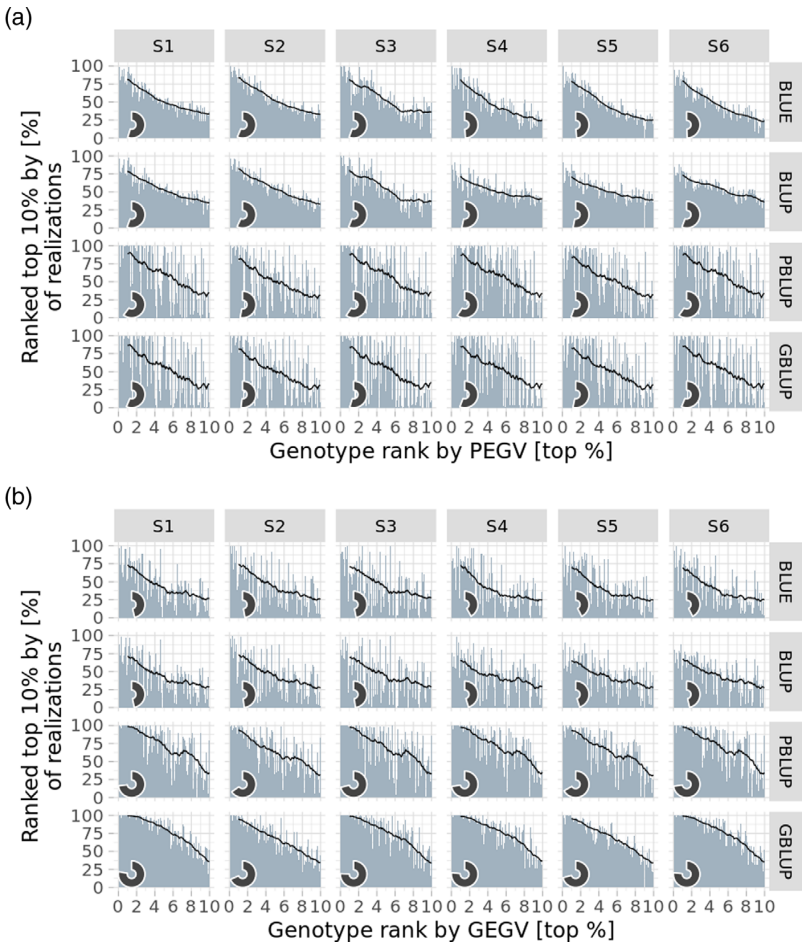

FIGURE 4 Fractions of realizations in which the top 10% genotypes by yield phenotypically estimated genotypic values (PEGV) (a) or genomic estimated genotypic value (GEGV) (b) are found among the top 10% of the predictions from different scenarios and models (fraction of realizations) as well. Each bar corresponds to one genotype, the circular bar in the plot corner shows the average amount of top 10% genotypes that were recovered to be in the top 10% of scenario predictions. See Table 2 for numeric values. The black line shows the average within a moving window of 2% (32 genotypes). BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; PBLUP, pedigree-based best linear unbiased estimation; GBLUP, genomic best linear unbiased prediction

in that they either yielded a higher rate of recovered top 10% genotypes from the scenarios or, in cases where this figure was similar between the models, the advantage of the PBLUP and GBLUP models over BLUP increased when the top genotype group was defined increasingly smaller.

3.4 | An early-stage trial can profit from joint analysis with related, later-stage trials

The correlation of the predictions of the early-stage genotypes S4e and S5e to the two reference values was higher when the early- and late-stage genotypes were jointly analyzed than when only the early-stage genotypes were used. This joint analysis of early- and late-stage genotypes lead to higher correlations to the reference values for S5e candidates (0.04–0.09 higher) than for the S4e candidates (0.00–0.04 higher, Figure 5C). Dissecting this trend further, the correlations to the GEGVs increased for all cases except the S4e BLUPs, while the correlation to the PEGVs increased only for S5e whereas PEGV correlations of S4e candidates remained unchanged. However, jointly using early- and late-stage genotypes instead of early-stage genotypes alone never led to inferior results, so averaging across the two references, using early- and late-stage genotypes jointly, was the safer choice to achieve high correlations of early-stage genotypes to the reference values.

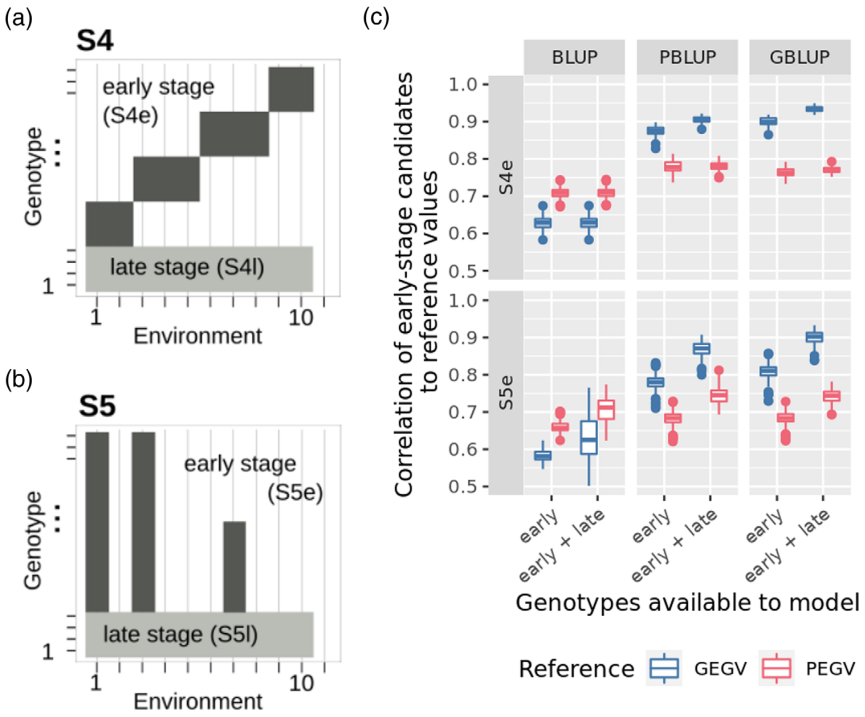

FIGURE 5 (a,b): Schematics of the early- and late-stage genotypes in Scenarios S4 and S5. (c): Correlations of the hybrid predictions from the early-stage genotypes of Scenarios S4 and S5 (thus suffixed with "e") to the PEGVs or GEGVs. The two plot rows show the correlations for the two genotype groups S4e and S5e (see Methods section) and the three plot columns show the results for each of the models used to estimate candidate performance. The horizontal plot axis lists the different sets of genotypes available to the model, detailed in (a) and (b)

4 | DISCUSSION

4.1 | The choice of proper reference values for the genotypic values

We used two different reference values, PEGVs and GEGVs, to compare different arrangements of field trials analyzed with four popular biometric models. This raises the question of which of the two reference values approximates the genotypic value more precisely and is therefore more informative for a breeder. The model underlying PEGVs is also used by the Federal Variety Office in Germany for candidate evaluations, and therefore a formal argument is that the PEGVs are of direct interest to breeders. The PEGVs reflect estimates of the genotypic values based on only a few assumptions, such as, for example, randomly distributed residual errors, but disregard information on genetic relationships among candidates and implies that all genotypes have a homogeneous genotypic variance and are independent. The latter is particularly important for hybrid wheat because factorial mating designs (Bernardo, 2010, p. 123), as used, for example, in our study to generate the hybrids, result in extensive half-sib families.

In contrast to the PEGVs, the predictions of GEGVs consider the relatedness between genotypes. Conceptually, the hybrid performance is decomposed in the applied model into a component of the breeding value and the dominance deviation, assuming that epistatic and genotype × environment effects do not play a role (Falconer, 1989, p. 125). The GBLUP model is equivalent to ridge regression BLUP for large numbers of SNPs (Habier et al., 2007), which means that GEGVs assume equal contributions of all loci to the genetic variance (Meuwissen et al., 2001) and effect size estimates are shrinked stronger for markers with low minor allele frequency (Gianola, 2013). A further assumption made in the decomposition of the hybrid performance is that the two genetic effects are orthogonally defined, which unfortunately is not always the case due to linkage disequilibrium (Alvarez-Castro & Carlborg, 2006). Furthermore, the prediction of the hybrid performance uses variance components of the genetic effects (Bernardo, 2010, p. 283). Estimations of variance components, especially in small data sets, can be quite erroneous (Huang & Mackay, 2016) and, thus, influence the prediction of the hybrid performance. In summary, GEGVs use more information compared with PEGVs, but they are also based on numerous assumptions. Therefore, the superiority of GEGVs

over PEGVs depends very much on whether the information used is accurate and whether the assumptions can be made. Thus, the question of superiority cannot be definitely answered and, as a result, the comparison between different arrangements of field trials analyzed with four different biometric models was performed on both GEGV and PEGV.

4.2 | PBLUP and GBLUP outperform BLUE and BLUP under most circumstances

Considering the ambiguity described above regarding the optimal reference values, the PBLUP and GBLUP models stand out as analysis models that provide stable and superior results in terms of correlations in all tested scenarios except S2 (Figure 3, Figure 4). For all other scenarios, the advantage of PBLUP and GBLUP over BLUP and BLUE in predicting the true genotypic values depends on where the true values fall in the spectrum between PEGVs and GEGVs. Assuming that the true values would be best captured by the PEGVs, the only settings that outperform the "classical" approach, S2 analyzed with BLUP, are S1 and S3, analyzed by PBLUP. The better the GEGVs capture the true genotypic values, the larger the number of scenarios grows in which PBLUP and GBLUP estimate the true values better than BLUP and BLUE.

Further refining the winner between PBLUP and GBLUP is not possible in this study. The PBLUP prediction accuracies are influenced less by the reference value choice so that, within a scenario, PBLUP outperforms GBLUP in terms of PEGV correlation but is inferior in terms of GEGV correlation. This pattern is essentially reflected when assessing the fraction of top genotypes that are also recovered as such by the scenario predictions (Figure 4). When comparing predictions from these two models with respect to their GEGV correlation one must consider that GBLUP is also the underlying model of the GEGVs. However, the extent to which this influences the result cannot be answered in this study.

4.3 | Scenario S3 represents an interesting alternative to balanced testing of genotypes in a subset of environments

Scenario S2 represents a current standard of a preliminary yield trial for many breeding programs (Zhao, Zeng, et al., 2013). It is interesting to note that for both reference values, one of the alternative scenarios (S1, S3, or S6) in which the genotypes are distributed over more environments leads to an improved prediction accuracy and, moreover, a significantly reduced variability in the estimations or predictions of the genotypic values (Figure 3). In this context, the advantage of the above alternative scenarios over a classical balanced experiment in a few environments (S2) becomes much higher if one uses the GEGVs instead of the PEGVs as reference values. Therefore, our results indicate that wheat breeders could improve their preliminary yield trials by changing the design of preliminary yield trials. As an alternative design, Scenario S3 seems to us to be particularly suitable, because it provided both high accuracies in the estimations/predictions and yet showed low variability resulting on the used data sample. Also, as the genotypes are not completely randomly distributed like in the case of Scenario S1, it might be easier to implement from a logistic standpoint. Finally, if implementation of S3 is not possible, and the preliminary trials have to be tested in the same few environments, some of the benefits of pedigree-based or genomic modeling could still be harvested by including candidates that are related but tested in different trials and environments in a joint phenotypic analysis (Figure 5).

4.4 | Conclusions and Outlook

The final decision on the design of multienvironment hybrid wheat yield trials also depends on cost scenarios where increasing the number of locations could cause additional costs besides the cost per plot. However, a breeder might have infrastructure already present in many environments when advanced trials are evaluated there. Therefore, implementing sparse preliminary testing might not mean using additional environments but redistributing plots of preliminary trials to environments that are already in use, albeit only for advanced and not preliminary trials. Our results clearly demonstrated the utility for optimizing the design of multienvironment hybrid wheat yield trials in times of genomic selection. However, our results are not limited to hybrid breeding programs but are also of interest for line breeding. The study of Endelman et al. (2014) already pioneered extended environmental sampling in a design similar to our Scenario S1. They tested candidates from extensive biparental populations of maize and barley in a sparse fashion in up to three environments. In this case, genome-wide markers established the connectivity between full sibs and allowed better control for interaction effects between genotypes and environment. In principle, this is also expected for comprehensive diversity panels of wheat inbred line breeding programs, as high prediction accuracies have also been reported in such situations (He et al., 2016).

ACKNOWLEDGMENTS

The authors acknowledge funding within the Wheat BigData project (German Federal Ministry of Food and Agriculture, FKZ2818408B18).

AUTHOR CONTRIBUTIONS

Moritz Lell: Conceptualization; Formal analysis; Investigation; Methodology; Visualization; Writing-original

draft. Jochen Reif: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing-review & editing. Yusheng Zhao: Conceptualization; Methodology; Supervision; Writing-review & editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Moritz Lell https://orcid.org/0000-0002-2428-5157
Jochen Reif https://orcid.org/0000-0002-6742-265X
Yusheng Zhao https://orcid.org/0000-0001-6783-5182

REFERENCES

Alvarez-Castro, J. M., & Carlborg, O. (2006). A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics*, *176*(2), 1151—1167. https://doi.org/10.1534/genetics.106.067348

Bernardo, R. (2010). *Breeding for quantitative traits* (2nd ed.). Stemma Press.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*(1), 375–385. https://doi.org/10.1534/genetics.109.101501

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G. L., Akhunova, A., See, D., Bai, G., Pumphrey, M., Tomar, L., Wong, D., Kong, S., Reynolds, M., Lopez da Silva, M., Bockelman, H., Talbert, L., … Akhunov, E. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(20), 8057–8062. https://doi.org/10.1073/pnas.1217133110

Cullis, B. R., Smith, A. B., & Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*(4), 381–393. https://doi.org/10.1198/108571106x154443 https://doi.org/10.1198/108571106X154443

Endelman, J. B., Atlin, G. N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M. E., & Jannink, J.-L. (2014). Optimal design of preliminary yield trials with genome-wide markers. *Crop Science*, *54*(1), 48–59. https://doi.org/10.2135/cropsci2013.03.0154

Falconer, D. S. (1989). *Introduction to quantitative genetics*. Longman.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (Vol. 4, pp. 625–631). Clarendon Press.

Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics*, *194*(3), 573–596. https://doi.org/10.1534/genetics.113.151753

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J., & Thompson, R. (2015). *ASReml user guide release 4.1 functional specification*. VSN International Ltd. https://www.vsni.co.uk

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, *51*(4). 1440–1450 https://doi.org/10.2307/2533274

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*(4), 2389–2397. https://doi.org/10.1534/genetics.107.081190

He, S., Reif, J. C., Korzun, V., Bothe, R., Ebmeyer, E., & Jiang, Y. (2017). Genome-wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to Central Europe. *Theoretical and Applied Genetics*, *130*(4), 635–647. https://doi.org/10.1007/s00122-016-2840-x

He, S., Schulthess, A. W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., Ebmeyer, E., Reif, J. C., & Jiang, Y. (2016). Genomic selection in a commercial winter wheat population. *Theoretical and Applied Genetics*, *129*(3), 641–651. https://doi.org/10.1007/s00122-015-2655-1

Huang, W., & Mackay, T. F. C. (2016). The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLOS Genetics*, *12*(11), e1006421. https://doi.org/10.1371/journal.pgen.1006421

Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., Pazaran, G. C., Burgueño, J., Pacheco, A., Grondona, M., Wimmer, V., & Prasanna, B. M. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3 Genes|Genomes|Genetics*, *10*(8), 2725–2739. https://doi.org/10.1534/g3.120.401349

Jiang, Y., Schmidt, R. H., Zhao, Y., & Reif, J. C. (2017). A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. *Nature Genetics*, *49*(12), 1741–1746. https://doi.org/10.1038/ng.3974

Longin, C. F. H., Gowda, M., Mühleisen, J., Ebmeyer, E., Kazman, E., Schachschneider, R., Schacht, J., Kirchhoff, M., Zhao, Y., & Reif, J. C. (2013). Hybrid wheat: Quantitative genetic parameters and consequences for the design of breeding programs. *Theoretical and Applied Genetics*, *126*(11), 2791–2801. https://doi.org/10.1007/s00122-013-2172-z

Longin, C. F. H., Mi, X., & Würschum, T. (2015). Genomic selection in wheat: Optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. *Theoretical and Applied Genetics*, *128*(7), 1297–1306. https://doi.org/10.1007/s00122-015-2505-1

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819. https://doi.org/10.1093/genetics/157.4.1819

Patterson, H. D., & Williams, E. R. (1976). A new class of resolvable incomplete block designs. *Biometrika*, *63*(1), 83–92. https://doi.org/10.1093/biomet/63.1.83

Pérez, P., & de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/genetics.114.164442

Piepho, H. P., Möhring, J., Melchinger, A. E., & Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, *161*(1-2), 209–228. https://doi.org/10.1007/s10681-007-9449-8

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3 Genes|Genomes|Genetics*, *6*(9), 2799–2808. https://doi.org/10.1534/g3.116.032888

Seye, A. I., Bauland, C., Charcosset, A., & Moreau, L. (2020). Revisiting hybrid breeding designs using genomic predictions: Simulations highlight the superiority of incomplete factorials between segregating families over topcross designs. *Theoretical and Applied Genetics*, *133*(6), 1995–2010. https://doi.org/10.1007/s00122-020-03573-5

Smith, A. B., & Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, *214*(8), 143. https://doi.org/10.1007/s10681-018-2220-5

Storlie, E., & Charmet, G. (2013). Genomic selection accuracy using historical data generated in a wheat breeding program. *The Plant Genome*, *6*(1). https://doi.org/10.3835/plantgenome2013.01.0001

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org

Utz, H. F., Melchinger, A. E., & Schön, C. C. (2000). Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics*, *154*(4), 1839–1849. https://doi.org/10.1093/genetics/154.4.1839

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., Cattivelli, L., Mastrangelo, A. M., Whan, A., Stephen, S., Barker, G., Wieseke, R., Plieske, J., International Wheat Genome Sequencing Consortium, Lillemo, M., Mather, D., Appels, R., … Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90∼000 single nucleotide polymorphism array. *Plant Biotechnology Journal*, *12*(6), 787–796. https://doi.org/10.1111/pbi.12183

Ward, B. P., Brown-Guedira, G., Tyagi, P., Kolb, F. L., Sanford, D. A., Sneller, C. H., & Griffey, C. A. (2019). Multienvironment and multitrait genomic selection models in unbalanced early-generation wheat yield trials. *Crop Science*, *59*(2), 491–507. https://doi.org/10.2135/cropsci2018.03.0189

Würschum, T., Langer, S. M., Longin, C. F. H., Korzun, V., Akhunov, E., Ebmeyer, E., Schachschneider, R., Schacht, J., Kazman, E., & Reif, J. C. (2013). Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. *Theoretical and Applied Genetics*, *126*(6), 1477–1486. https://doi.org/10.1007/s00122-013-2065-1

Yates, F. (1940). The recovery of inter-block information in balanced incomplete block designs. *Annals of Eugenics*, *10*(1), 317–325. https://doi.org/10.1111/j.1469-1809.1940.tb02257.x

Zhao, Y., Gowda, M., Würschum, T., Friedrich, C., Longin, H., Korzun, V., Kollers, S., Schachschneider, R., Zeng, J., Fernando, R., Dubcovsky, J., & Reif, J. C. (2013). Dissecting the genetic architecture of frost tolerance in central European winter wheat. *Journal of Experimental Botany*, *64*(14), 4453–4460. https://doi.org/10.1093/jxb/ert259

Zhao, Y., Li, Z., Liu, G., Jiang, Y., Maurer, H. P., Würschum, T., Mock, H.-P., Matros, A., Ebmeyer, E., Schachschneider, R., Kazman, E., Schacht, J., Gowda, M., Longin, C. F. H., & Reif, J. C. (2015). Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(51), 15624–15629. https://doi.org/10.1073/pnas.1514547112

Zhao, Y., Zeng, J., Fernando, R., & Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Science*, *53*(3), 802–810. https://doi.org/10.2135/cropsci2012.08.0463

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Lell, Moritz, Reif, Jochen, & Zhao, Yusheng Optimizing the setup of multienvironmental hybrid wheat yield trials for boosting the selection capability. *Plant Genome*. 2021;*14*:e20150. https://doi.org/10.1002/tpg2.20150

2.2 Leveraging the potential of big genomic and phenotypic data for genome-wide association mapping in wheat

Published: In 2024, *The Crop Journal* 12(3):803.

DOI: 10.1016/j.cj.2024.03.005

Authors: Moritz Lell, Yusheng Zhao, Jochen Reif

Abstract:

Genome-wide association mapping studies (GWAS) based on Big Data are a potential approach to improve marker-assisted selection in plant breeding. The number of available phenotypic and genomic data sets in which medium-sized populations of several hundred individuals have been studied is rapidly increasing. Combining these data and using them in GWAS could increase both the power of QTL discovery and the accuracy of estimation of underlying genetic effects, but is hindered by data heterogeneity and lack of interoperability. In this study, we used genomic and phenotypic data sets, focusing on Central European winter wheat populations evaluated for heading date. We explored strategies for integrating these data and subsequently the resulting potential for GWAS. Establishing interoperability between data sets was greatly aided by some overlapping genotypes and a linear relationship between the different phenotyping protocols, resulting in high quality integrated phenotypic data. In this context, genomic prediction proved to be a suitable tool to study relevance of interactions between genotypes and experimental series, which was low in our case. Contrary to expectations, fewer associations between markers and traits were found in the larger combined data than in the individual experimental series. However, the predictive power based on the marker-trait associations of the integrated data set was higher across data sets. Therefore, the results show that the integration of medium-sized to Big Data is an approach to increase the power to detect QTL in GWAS. The results encourage further efforts to standardize and share data in the plant breeding community.

The Crop Journal 12 (2024) 803–813

Contents lists available at ScienceDirect

The Crop Journal

journal homepage: www.keaipublishing.com/en/journals/the-crop-journal/

Leveraging the potential of big genomic and phenotypic data for genome-wide association mapping in wheat

Moritz Lell, Yusheng Zhao, Jochen C. Reif*

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstraße 3, D-06466 Seeland, Germany

ARTICLE INFO

Article history:
Received 17 November 2023
Revised 22 March 2024
Accepted 29 March 2024
Available online 19 April 2024

Keywords:
Big Data
Genome-wide association study
Data integration
Genomic prediction
Wheat

ABSTRACT

Genome-wide association mapping studies (GWAS) based on Big Data are a potential approach to improve marker-assisted selection in plant breeding. The number of available phenotypic and genomic data sets in which medium-sized populations of several hundred individuals have been studied is rapidly increasing. Combining these data and using them in GWAS could increase both the power of QTL discovery and the accuracy of estimation of underlying genetic effects, but is hindered by data heterogeneity and lack of interoperability. In this study, we used genomic and phenotypic data sets, focusing on Central European winter wheat populations evaluated for heading date. We explored strategies for integrating these data and subsequently the resulting potential for GWAS. Establishing interoperability between data sets was greatly aided by some overlapping genotypes and a linear relationship between the different phenotyping protocols, resulting in high quality integrated phenotypic data. In this context, genomic prediction proved to be a suitable tool to study relevance of interactions between genotypes and experimental series, which was low in our case. Contrary to expectations, fewer associations between markers and traits were found in the larger combined data than in the individual experimental series. However, the predictive power based on the marker-trait associations of the integrated data set was higher across data sets. Therefore, the results show that the integration of medium-sized to Big Data is an approach to increase the power to detect QTL in GWAS. The results encourage further efforts to standardize and share data in the plant breeding community.

1. Introduction

Plant breeding is being transformed by the success and low-cost availability of high-throughput genotyping techniques and the fast pace of machine learning research [1]. As a result, data are becoming an asset whose value outlives their original use-case, giving rise to more systematic data retention and reuse [2]. A successful example is the reuse of data by merging several medium-sized, existing data sets for genomic best linear unbiased prediction (GBLUP) in wheat, which doubled the prediction accuracy of hybrid grain yield compared to the single data sets [2]. Genomic prediction is suited for grain yield because in that case a large number of small-effect genetic loci influence the trait of interest [3]. This is also the assumption underlying GBLUP, which estimates genetic values as random effects that are normally distributed as they are influenced by many genetic loci whose effects are themselves normally distributed [4].

Genomic prediction does not select important loci that shape phenotypic variation, nor does it seek accurate estimates of the effects of single genetic loci; instead, predictions of genetic values often rely on relatedness between individuals [5]. Nevertheless, identifying specific genetic loci that influence a trait is required to answer a multitude of questions: Loci that correlate with a trait are starting points for experimental approaches to elucidate the biological mechanisms underlying the trait [6]. Furthermore, a trait of interest may not follow the infinitesimal model of many small-effect loci, but may be influenced by a few loci with large effects, in which case the data do not fit the GBLUP assumptions well. When individual loci of decent effect size are uncovered, they can be used to identify candidates for pre-breeding based on genomic data, even if the desired phenotype is masked by other traits or genetic interactions [7]. Finally, they can be used in situations where insufficient data are available for genomic prediction, which depends on a large training set of individuals with available genomic and phenotypic data [8]. To detect marker-trait associations (MTAs), genome-wide association studies (GWAS) are performed to test each marker in the data set to assess whether it correlates with

* Corresponding author.
E-mail address: reif@ipk-gatersleben.de (J.C. Reif).

https://doi.org/10.1016/j.cj.2024.03.005

the phenotype of interest. A popular model used for this purpose is the Q + K model [9] that reduces false-positives caused by subpopulations because if distinct sub-populations with different trait means exist, differences in the subpopulation's respective genetic backgrounds would be reported as MTAs. Therefore, effects of the population structure are considered in the model as fixed effects (Q), suitable for distinct categorization of genotypes, and a variance structure (K), suitable for including relatedness matrices, e.g., numerator relationship matrices or estimates thereof. The success of GWAS can be compromised by inadequate population size and/or population structure, which can lead to spurious and off-target reported hits due to linkage disequilibrium blocks prevailing in the population. These problems are exacerbated if sought-for variants are rare in the population [10].

In the past years, several medium-size data sets mostly of a few hundred individuals have been published addressing heading date in wheat [2,11]. Their integration enables studies on the potential of Big Data approaches for GWAS. Combining diverse experimental series, however, requires overcoming several hurdles in data curation: If genomic and phenotypic data for genotypes come from different sources, reconciling them can be challenging due to ad-hoc data management practices in many research projects. Phenotyping and genotyping protocols can differ, requiring careful quality control and possibly conversion of values to ensure interoperability. Finally, population structure resulting from the combination of multiple populations may pose challenges to the interpretation of the results. In this study, we aim to determine whether, despite the practical and theoretical challenges outlined above, integrating and reevaluating those experimental series improves the detection power of GWAS. In this process, we aim to evaluate quality-control measures including a cross-validation using genomic prediction. Finally, differences in the GWAS results between individual experimental series and the combined data need to be connected to the established theory.

2. Material and methods

2.1. Plant material, SNP genotyping, and field trials

Three data sets of elite winter wheat populations (further referred to as experimental series) were used in our study (Table S1): Two of them (H1 and H2) include multi-environmental field trials with wheat hybrids, their parents, and check varieties. The experimental series (L) includes multi-environmental field trials with inbred lines. The inbred lines as well as parents of the hybrids were genotyped using a 90K SNP array [12] or subsets thereof. Lines from H1 were genotyped with the full 90K array. Series H2 were genotyped using a reduced set of 15K markers. Of series L, 1699 lines were genotyped using the 90K SNP array and the remaining 2934 lines based on the 15K array. The fraction of missing values per experimental series ranged from 0.4% to 17.6%. After merging the genomic data, the final fraction of missing values filled by imputation was 36.8%. The gaps were filled by imputation (IMPUTE2, [13]). The resulting data had 18,566 high-quality SNPs. The SNP profiles of the hybrids were derived based on the SNP profiles of their homozygous parental inbred lines. Details on the genotyping and genomic data curation were described in [2]. We used a slightly more stringent threshold to define duplicate genotypes with pairwise Rogers' distances below 0.025.

Experimental series H1 was based on 135 elite winter bread wheat lines and their 1604 single-cross hybrid progenies. The composition of the hybrid population has been outlined in detail previously [2,11]. Briefly, the parental lines were selected to reflect a diverse spectrum of genetic diversity present in Central Europe. The 135 lines were divided into a female pool of 120 lines and a male pool of 15 lines, and seeds of 1604 single-cross hybrids were produced (Table S1). The parental lines, hybrids, and 10 additional checks were evaluated in 10 environments (6 locations in 2012 and 4 locations in 2013) in Germany for heading time. Heading time was recorded as the number of days from 1st of January to the day when half of the heads had emerged from flag leaves [14]. The genotypes were evaluated in a partially replicated alpha lattice design (for details, see [11]).

Experimental series H2 was based on 224 elite winter bread wheat lines and their 1812 single-cross hybrids [2]. Parental lines were selected, as in experimental series set H1, to reflect a broad spectrum of diversity present in Central Europe. Seeds for 1812 single-cross hybrids were generated based on 184 male lines and 40 female lines (Table S1). The parental lines, hybrids, and 11 checks were tested for heading time in 12 environments (Table S2) with 6 locations in 2016 and 6 locations in 2017 in Germany. Heading time was recorded as the number of days from 1st of January to the day when half of the heads had emerged from flag leaves [14]. The experimental design was a non-replicated alpha lattice design.

Experimental series L was based on 4633 Central European elite winter wheat lines of the breeding program of KWS LOCHOW GmbH (Einbeck, Germany). The composition of the population has been outlined in detail previously ([2], experimental series VI). Briefly, the lines were evaluated in the routine breeding trials in the years 2012 to 2015 for heading time in Germany. The experimental design for each trial followed an alpha design with one to three replications per site, with the number of entries per trial ranging from 33 to 607. On average, a single line was tested in 3.5 environments (Table S2). Heading time was recorded as the developmental stage (BBCH, see [15]) at that time when ears of approximately half of the genotypes were fully visible. To merge experimental series L with H1 and H2, we transformed the data of experimental series L to flowering time in days since January 1st. The scales exhibit a linear relationship. Therefore, we estimated parameters necessary for the transformation using linear regression based on data from overlapping genotypes of L and H1 or H2.

2.2. Genomic data analyses

All calculations of this study were performed on a Linux machine with 4 Intel Xeon CPU E7-4890 v2 processors (120 logical cores), using R 4.0.4 [16] and AsReml 4.1.0 [17], and required about 200 GB of RAM.

The population structure was investigated by computing the pairwise Rogers' distances [18] between individuals. Principal coordinate analyses were performed based on the matrix of pairwise Rogers' distances [19]. In addition, we performed complete-linkage hierarchical clustering. For each of the experimental series the linkage disequilibrium was calculated as the squared correlation between markers [20]. Effective population size was estimated as $N_e = k/(3(r^2 - n^{-1}))$ [21], where n is the sample size, k = 1 for inbred lines and 2 for hybrids, and r^2 is the linkage disequilibrium between unlinked loci.

Linkage disequilibrium decay was calculated by determining pairwise marker distances for each chromosome and fitting the respective linkage disequilibrium values on the distances using Hill and Weir's formula [22] using nonlinear regression. The regression was performed separately for each experimental series. When grouping marker-trait-associations into quantitative trait loci (QTL), the distance where linkage disequilibrium $r^2 \leq 0.2$ were used as thresholds.

2.3. Phenotypic data analyses

Linear mixed models were applied for data of every environment separately assuming genotypes as fixed effects and design effects such as trials, replicates, or incomplete blocks as random effects. All data were checked for outliers using method 4 "Bonferroni-Holm with rescaled standardized residuals of mean absolute deviation" as previously suggested [23]. Outliers were removed for further analyses. Best linear unbiased estimates (BLUEs) of genotypes in each environment and each experimental series were obtained and served as the input for the subsequent analyses. Moreover, for individual replicated environments repeatability was assessed as quality measure.

The across-environment BLUEs within H1, H2, and L but also across experimental series (A) were obtained from the respective set of per-environment BLUEs y_{ijk} as:

$$y_{ijk} = d_i + g_j + e_k + \varepsilon_{ijk} \quad (1)$$

where d_i and g_j are the effect of the i^{th} experimental series and the j^{th} genotype BLUEs, both fixed effects, and e_k and ε_{ijk} are the effect of the k^{th} environment and residual effects, respectively, both random effects. The effect of the experimental series d was omitted for the analyses within H1, H2, and L. The broad-sense heritability was calculated assuming genotypes as random effects in the above outlined model as:

$$H^2 = \sigma_g^2 / \left(\sigma_g^2 + \frac{\sigma_r^2}{n_e}\right) \quad (2)$$

where σ_g^2 is the genotypic variance and σ_r^2 is the variance of the residuals and n_e is the average number of environments a genotype is measured in.

2.4. Genomic prediction for quality control

The phenotype-genotype match was assessed by genome-wide predictions combined with cross validations. As genome-wide prediction model, we used Reproducing Kernel Hilbert Spaces Regression (RKHS) that was implemented in the BGLR R package [24] with two kernels, one for the additive effects, where the covariance between individuals is defined using the VanRaden matrix [25], and one for the dominance effects, where the covariance is defined using the matrix described by Alvarez-Castro et al. [26].

To estimate the prediction ability within the experimental series, we applied a classical fivefold cross-validation strategy. Moreover, because in factorial mating designs the relatedness between training and test set strongly influences the prediction ability, we followed previous suggestions [11] and selected training sets consisting of 80% of female and male lines. The hybrids were used to form test sets with three successively decreasing degrees of relatedness to the training set. Test set T2, which was most closely related to the training set, contained only hybrids descended from the parents in the training set, whereas the less related test set T1 contained hybrids sharing a parent with the hybrids in the training set, and the least related test set T0 contained only hybrids with no parents in common with the training set. We also assessed the prediction ability across experimental series H1, H2, and L using one of these series as training population to predict heading date of genotypes evaluated in another series. The prediction ability was calculated as the correlation of the predicted phenotypes to the measured phenotypes of the test set.

2.5. Genome-wide association mapping

Association mapping was performed for each of the three individual experimental series, the joint data set A, and for subsets of A containing inbred lines only ("Al") or hybrids only ("Ah"). We applied a filtering of minor allele count of at least 30 genotypes and used the following linear mixed model [27]:

$$y = Xh + S_a a + S_d d + g + \varepsilon \quad (3)$$

Using the phenotypic data $y \in \mathbb{R}$, the model contains as fixed effects the additive ($a \in R^p$) and dominance ($d \in \mathbb{R}^p$) effects of the SNPs, the overall intercept of hybrids and non-hybrids $h \in \mathbb{R}^2$ and as random effects the genetic effects g and the residuals $\mathcal{N}(0, I_n)$. $X \in \mathbb{R}^{n\times 2}$ is a design matrix, assigning the status of an inbred or a hybrid to each individual. $S_a \in [-1; 0; 1]^{n\times p}$ holds the SNP alleles (coding: homozygous allele 1; heterozygous; homozygous allele 2) for each individual. $S_d = 1 - |S_a|$, where $|\cdot|$ denotes the element-wise absolute values of the matrix. SNPs of hybrid genotypes were not measured directly by an SNP array, instead the mean of their parents' SNP codes were used. To correct for population structure, the genetic effects were included in two fashions, depending on whether a genotype is a hybrid or an inbred line. The genetic effects of inbred lines were modelled using their Rogers' distance; those of hybrids were modelled using the Rogers' distance of their parents. Thus, formally, g can be defined as follows: For a total population of n_ℓ inbreds and $n - n_\ell$ hybrids that are derived from n_f female and n_m male parents,

$$g = Z_\ell \ell + Z_f f + Z_m m \quad (4)$$

The random vectors $\ell \in \mathbb{R}^{n_\ell}$, $f \in \mathbb{R}^{n_f}$ and $m \in \mathbb{R}^{n_m}$ are mutually independent and have a multivariate normal distribution $x\ N(0, 2(1 - K_x)\sigma_x^2), (x = l, f, m)$, while K_ℓ, K_f, and K_m are the Rogers' distance matrices of inbreds, female and male parents, respectively. $Z_\ell \in \{0; 1\}^{n\times n_\ell}$ is the design matrix for inbreds, $Z_f \in \{0; 1\}^{n\times n_f}$ and $Z_m \in \{0; 1\}^{n\times n_m}$ are the design matrices for female and male parents of hybrids, respectively. To save computation time, GWAS was performed using the P3D approach, which estimates the genomic variance once and does not re-estimate it for every marker [28]. Subsequently, this simplification was removed for all resulting P values smaller than 0.01, using AsReml-R to fit the mixed linear models (3) and (4). The p value of a marker-trait association (MTA) was obtained by performing a Wald test [29]. The significance level was determined using the Benjamini-Hochberg procedure [30] to control the false discovery rate (FDR) at 0.05 individually for each data set and effect type (additive or dominance).

To group physically near markers into quantitative trait loci (QTL), we employed single-linkage clustering based on the physical distance of the markers. Within a series, the linkage disequilibrium decay distance of that dataset (H1: 1.4 cM, H2: 1.41 cM, L: 1.07 cM, Fig. S1) was used as cutoff for clustering. Between series, the average linkage disequilibrium decay distance of the two series was used. For grouping with markers found in the joint data set A, we used the average linkage disequilibrium decay distance of all three experimental series.

We assessed the predictive value of the significant MTAs by calculating their phenotypic variance explained and their prediction ability in the three experimental series and the joint data set A. The phenotypic variance explained was derived as the "Wherry-1" [31] adjusted coefficient of determination $\widehat{R^2} = 1 - \left(1 - R^2\right)(n-1)/(n-p-1)$, given coefficient of determination R^2, sample size n and number of predictors p from an analysis of variance (ANOVA) within each experimental series. We chose as explanatory variables at first all markers that were deemed significant in the GWAS using the same experimental series. For comparison, we added the markers that were deemed as significant using the joint data set A to the set of explanatory variables. For H1, H2, and A the hybrid/inbred status of the genotype

was used as an additional predictor in the ANOVA but it was excluded for the calculation of the adjusted coefficient of determination so that different heading date means of the hybrid and inbred line groups do not inflate the measure. As adding the significant markers of data set A increases the number of explanatory variables, we separately created negative controls where we added a sample of random markers to the markers found significant in the individual data sets. The sample size was equal to the number of significant markers in A so that the number of explanatory variables in the negative control was equal to the number of explanatory variables of the experimental series plus A case. The sampling was repeated 50 times to generate a variance estimate. The prediction ability was calculated by first estimating within each experimental series (H1, H2, L) and the joint dataset A the effect sizes of the markers found significant in that series using a linear model with the markers as predictors. Given these marker effect estimates, we predicted heading dates of other experimental series using respective genomic data. Prediction ability was defined as the correlation of true versus predicted values.

We also tested for presence of digenic epistasis in the joint data set A. The screen included epistatic interaction effects between all markers that were found to be significant in any data set (a_1 and d_1, additive and dominance effects, respectively) and all other markers, referred to as background markers (a_2 and d_2). We fitted a mixed linear model in which the main effects of markers and their interaction effects were modelled as fixed effects. The design matrices $S_{a_1}, S_{d_1}, S_{a_2}, S_{d_2}$ for the main effects a_1, d_1, a_2, d_2 are denoted as in (3), while for the interaction effects $aa_{12}, ad_{12}, ad_{21}, dd_{12}$, the design matrices are the Hadamard product "$\circ$" between the design matrices relevant to the main effects. To correct for population structure, a random genotypic effect g was included, defined in the same way as in the GWAS model (4). The effects h and ε were also defined like in (3).

$$\begin{aligned} y = Xh + S_{a_1}a_1 + S_{d_1}d_1 + S_{a_2}a_2 + S_{d_2}d_2 + \\ + \left(S^{\circ}_{a_1}S_{a_2}\right)aa_{12} + \left(S^{\circ}_{a_1}S_{d_2}\right)ad_{12} + \\ + \left(S^{\circ}_{a_2}S_{d_1}\right)ad_{21} + \left(S^{\circ}_{d_1}S_{d_2}\right)dd_{12} + g + \epsilon. \end{aligned} \quad (5)$$

Because of the large number of effects to be estimated, as for single-SNP effects, the P3D GWAS approach was used for the estimation. The significance level that controls the FDR at 0.05 was chosen.

2.6. Simulation study

Computer simulations were performed to assess the QTL detection rate using data from experimental series L. Three simulated traits, QTL_minor, QTL_medium, and QTL_large, were defined to be affected each by 5 markers so that each marker explains 2% (QTL_minor), 5% (QTL_medium), or 10% (QTL_large) of the genetic variance. The simulated phenotypic values were computed using following model:

$$y = \mu + X\beta + V\gamma + e \quad (6)$$

The simulated phenotypic values are composed of a mean value μ of 95 (corresponding to the mean observed for experimental series L), a vector of effects of the 5 QTL β, a vector of effects of 18,523 background markers γ and a residual effect vector e. X and V refers to the respective design matrices. The simulation was run 20 times, and for each run, five random markers with minor allele frequencies greater than 0.05 were designated as QTL in the simulated data. The effects of the background marker were independently drawn as $\gamma \sim \mathcal{N}(0,\ 0.0016)$. The effects of the QTL β were determined using iterative optimization to achieve the specified explained genetic variance for each QTL. Residual effects were drawn independently as $e \sim \mathcal{N}\left(0,\ \frac{4}{9}\sigma_g^2\right)$, where σ_g^2 is the simulated genetic variance, with the goal of mimicking a trait with the average heritability of 0.69 that has been observed for experimental series L. We examined two scenarios: First, we sampled from the simulated populations 100, 200, 300, 400, 500, 1000, 2000, and all 3404 individuals and performed P3D GWAS. The average rate of simulated QTL that could be recovered at a false discovery rate of 0.05 was counted for each population size and simulated trait. Second, the population size was set at 300 individuals, but contrasting subpopulations (low N_e and high N_e) were sampled aiming to minimize or maximize the effective population size (N_e). We performed GWAS in these 100 subpopulations for both traits, QTL_minor, QTL_medium and QTL_large, and recorded the average rate of QTL detected for each subgroup.

3. Results

3.1. Genetic diversity of Series L surpasses H1 and H2, which share some ancestry

We investigated the population structure the experimental series H1 (1604 F_1 and their 135 parents) and H2 (1812 F_1 and their 244 parents), and L (4633 lines). We performed a principal coordinate analysis of the inbred lines of the experimental series. (Fig. 1A). With respect to the first three principal coordinates, it can be seen that the parents of the hybrids have a lower genetic diversity than the lines of the experimental series L (Fig. 1A). This limited diversity was particularly pronounced in the parent lines that were used as males in the hybrids. The limited diversity resulted in a mean linkage disequilibrium between adjacent markers within a 10 cM windows that was 10%–16% higher in the experimental series H1 and H2 than in the experimental series L (Table S3; Figs. S2, S3). In addition, linkage disequilibrium between markers on different chromosomes was 2–3 times higher in experimental series H1 and H2 than in the L. In fact, hardly any marker in experimental series L showed a linkage disequilibrium above 0.1, whereas this was the case for about 1% of the marker pairs in the hybrid series (Figs. S4–S9). Consequently, the effective population size N_e, which is a function of the linkage disequilibrium between unlinked loci, was lower for H1 and H2 than for L (Table S4).

The distribution of the eigenvalues of the principal coordinate analyses within experimental series using inbred lines and/or hybrids, which reflect the proportion of molecular variance explained by the principal coordinates, revealed clear differences in the degree of complexity of the population structure (Fig. 1B): For experimental series H1 and H2, 99% of the genetic variance can be explained by half as many eigenvalues as for L. Thus, the population structure of H1 and H2 is much less complex than that of L.

Inspecting in detail the relationship between the male parents of the two hybrid experimental series in a cluster analysis (Fig. 1C), it can be seen that there was a total of 4 overlapping lines that were used as male parental lines in both H1 and H2. Furthermore, 2 clusters with a number of up to 9 genetically similar male parental lines from H1 or H2 can be recognized. This relatedness has to be considered when interpreting the results of genome-wide predictions across experimental series.

3.2. Quality assessment demonstrates high quality of underlying data

For each of the three experimental series, various quantitative genetic parameters were estimated to assess the quality of the phenotypic information for heading date as well as the match with the genomic data. At the lowest level of aggregation, repeatabilities

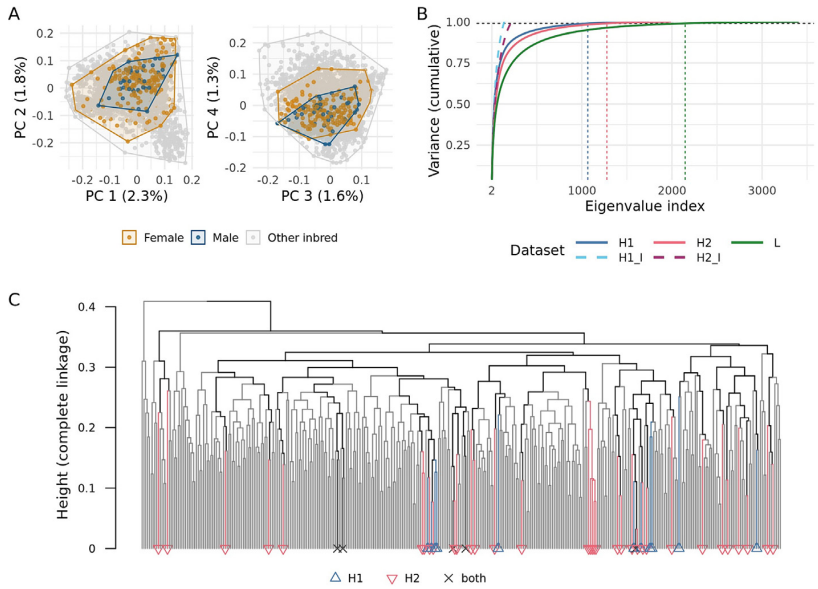

Fig. 1. Population structure and diversity. (A) Principal coordinate analysis (PCoA) across the three experimental series based on pairwise Rogers' distances. The parent lines of experimental series H1 and H2 are highlighted separately in color for male and female parents. Inbred lines from experimental series L are shown in gray. Principal coordinates (PC) are identified on the axis labels along with their explained molecular variances. (B) Cumulative genetic variance explained by the eigenvalues of the Rogers' distance matrices. The experimental series are denoted by the colors, the suffix "_I" denotes the inbred lines of the respective series. The vertical lines indicate the number of eigenvalues that make up for 99% of the total variance. (C) Cluster analysis of the parent and check genotypes of experimental series H1 and H2. Male parents used in H1, H2, or both are colored. Female parents and checks are colored in gray. The dendextend R package [52] was used to generate this figure.

provide an indication of phenotypic data quality. Estimated repeatability within individual environments averaged 87% (Fig. 2A). Only 5 of the total 35 replicated environments displayed repeatability estimates between 58% and 80%. In the other 30 environments, the values were above 0.8, which impressively demonstrates the high quality of the phenotypic data.

In a second step, the importance of genotypic variance for heading date was evaluated in relation to the total phenotypic variance, i.e., the heritability. Within each experimental series, heritability always exceeded 90% (Fig. 2A), suggesting precise estimates of the genotypic values for heading date. Nevertheless, due to the few overlapping genotypes between the three experimental series (Fig. S10), special attention was paid to the quality assessment of the estimated values for heading date. Interestingly, the overlapping genotypes showed a wide distribution with respect to heading date within all three experimental series (Fig. 2B). The correlations of the overlapping genotypes between pairs of experimental series were very high, with Pearson moment correlation coefficients of $r = 0.86$ ($P < 0.05$) between H1 and H2, $r = 0.95$ ($P < 0.05$) between H1 and L, and $r = 0.98$ ($P < 0.05$) between H2 and L. Consequently, it is not surprising that the heritability in the joint data set A was also very high at 89% (Fig. 2A). This clearly

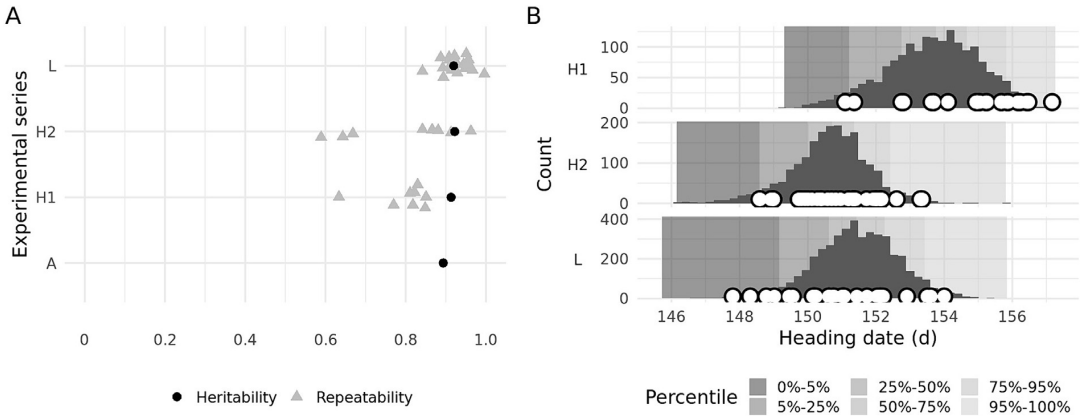

Fig. 2. Phenotypic data quality. (A) Repeatability (gray) and heritability estimates (black), corresponding to per-environment and across-environment analyses, respectively, for experimental series H1, H2, and L as well as for the joint data set. (B) Heading dates of genotypes overlapping across experimental series compared to those of the joint data set. The histograms show the distribution of heading date of the experimental series H1, H2, and L. The values of the overlapping genotypes are shown as dots below the histograms.

suggests that the non-orthogonal design of the experimental series did not result in a systematic bias in the estimation of genotypic effects.

To further elucidate the role of interaction effects between genotypes and experimental series, we examined genome-wide prediction abilities within and between different experimental series. Within the experimental series H1 and H2, the relatedness between training and test population played an important role: With increasing relatedness, the prediction ability increased from an average of 0.4 for the T0 scenario to 0.8 for the T2 scenario (Fig. 3A). The high prediction ability of on average 0.74 observed for the T1 scenarios, in which hybrids of the training and test populations contained overlapping females (T1_F), can be explained by the low diversity in the male pool and the resulting high relatedness between training and test populations for the T1_F scenario. In comparison, the prediction ability within experimental series L was 0.6 (Fig. 3B). The prediction ability between experimental series H1 and H2 averaged 0.5, which is within the range of values observed in the T0 scenario in the chess-board like cross-validations within experimental series H1 and H2. Using experimental series L as the training population and H1 or H2 as the test population, we also observed prediction abilities of on average 0.5. Thus, the body of results clearly suggests that the lack of orthogonality resulting from the pooling of experimental series H1, H2, and L, does not cause a strong systematic bias in the estimation of genotypic effects. The low prediction ability, with means of 0.25, observed when H1 or H2 were used as the training population and L as the test population can be explained by the small number of parental lines used to generate the hybrids.

3.3. Genome-wide association mapping revealed large discrepancies between marker-trait associations between experimental series

The location and number of MTAs varied greatly between experimental series H1 (1604 F_1 and their 135 parents), H2 (1812 F_1 and their 244 parents), and L (4633 lines) and overlapped only slightly (Figs. 4, S12). A similar trend could be seen considering correlations between P values of different experimental series. The $-\log_{10}P$ values of additive effect MTAs of the three experimental series showed correlations below 0.1 (Fig. S11A). The overlaps when markers that are closer to each other than the linkage decay threshold were grouped into QTL are shown in Table 1, Fig. 4, and Fig. S12: In experimental series H1 and H2, 64 and 79 QTL were found, respectively, but only 8 of them occurred in both experimental series. In experimental series L, which included only inbred lines, a total of 9 QTL were found, 3 of which were unique to that series. The overlap of QTL between the two experimental series H1 or H2 and experimental series L was small, with only 1 QTL occurring in both H1 and L, and 2 in both H2 and L. Of the QTL found in H1 or H2, only a small fraction was also found in the combined analysis across the three experimental series (Table 1): 4 of the 64 QTL in H1 and 10 of the 79 QTL in H2 were found in the joint experimental series A. Of the experimental series L, 6 of 9 QTL were detected in the joint data set. Conversely, of the 26 QTL found in the combined experimental series, 16 were also detected in individual experimental series, and 10 were found exclusively in the combined analysis.

To investigate the influence of the combination of hybrids and inbred lines on the GWAS result, we performed two additional GWAS runs with inbred lines only (abbreviated "Al") and hybrids only ("Ah"), of the joint data set A and correlated the resulting *P* values to those obtained from the data set A (Fig. S11B). We observed a strong correlation (0.72) between the results of data set Al and A for additive effects and similarly a strong correlation (0.85) between the results of data set Ah and A for dominance effects. The GWAS using data set Al resulted in one hit outside of QTL that were also found by data set A whereas multiple QTL were found using A only but not Al (Fig. S13). Using the hybrid-only dataset Ah no additive MTAs were found. A few dominance MTAs were detected using Ah that were not found using A, mainly on chromosome 7B but also on 1B, 2D, and 3A. Conversely, there were no dominance MTAs that were found only by A and not Ah.

Because inconsistencies between MTAs among different experimental series may be caused by epistasis, we estimated interaction effects between markers that were found to be significant in one experimental series and all other loci as genetic background (Fig. S14). The quantile–quantile plot shows no deviation from the diagonal line, indicating that the resulting P values are what is expected by chance. No P values are found to be significant at an FDR of 5%.

In a further step, we examined the predictive value of MTAs within experimental series by calculating the phenotypic variance explained by the significant markers detected within the series (Fig. 5A). We contrasted this scenario with one in which we added

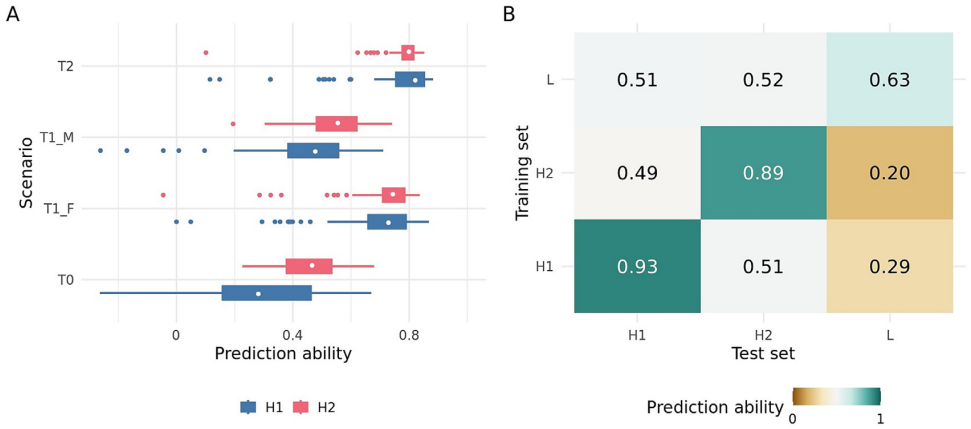

Fig. 3. Genomic prediction abilities. (A) Boxplot of genomic prediction abilities of the hybrid performance within experimental series H1 and H2. The relationship between the training and test population is shown on the vertical axis: T2: Both male and female parents of the test set are included in the training set; T1_F: Female parents only are related to the hybrids in the training set, male lines are unrelated; T1_M: Only male parents are related; T0: Lines in the training set are unrelated to the hybrids in the test set. 100 rounds of cross-validation with 80% of the hybrids were performed. (B) Genome-wide prediction abilities between (diagonal values) and among experimental series H1, H2, and L (off-diagonal values).

Table 1
QTL per experimental series and QTL shared between experimental series ("Exp. Srs.").

Exp. Srs.	H1	H2	L	A	(only)	Exp. Srs.	*n*
H1	64	8	1	4	53	H1 + H2 + L	0
H2		79	2	10	62	H2 + L + A	2
L			9	6	3	H1 + L + A	1
A				26	10	H1 + H2 + A	1
						H1 + H2 + L + A	0

The left part of the table shows how many QTL are shared between two experimental series. On the diagonal the number of QTL found in a single series can be seen. The column "(only)" shows how many QTL are found exclusively in the respective experimental series. Note that the off-diagonal values for a series do not add up to the diagonal value because QTL that are found in more than two series are shown multiple times. The right part of the table shows how many QTL that are shared by three or all four groups ("*n*").

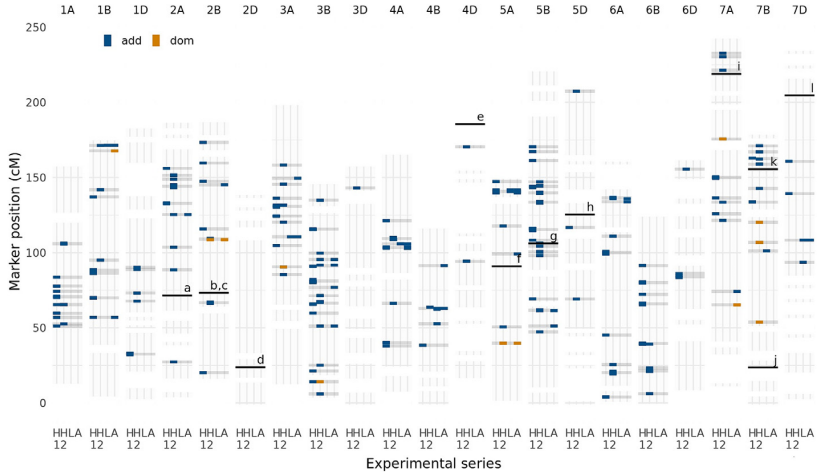

Fig. 4. Chromosomal regions with significant marker-trait associations for heading date in experimental series H1, H2, and L as well as the total (A) data set. The wheat chromosomes are arranged next to each other and denoted above the plot. For each chromosome, significant marker-trait associations found in the four data sets are shown next to each other as colored marks, where the color distinguishes additive (add) from dominance (dom) effects. The gray horizontal bars indicate quantitative trait locus (QTL) regions. Estimated locations of some well-known heading date-related genes (Table S5) are indicated with black letters: a = PPD-A1; b,c = PPD-B1; d = PPD-D1; e = VRN2; f = VRN-A1; g = VRN-B1; h = VRN-D1; i = PIE1-like; j = FT/VRN-B3; k = PIE1-like; l = PIE1-like.

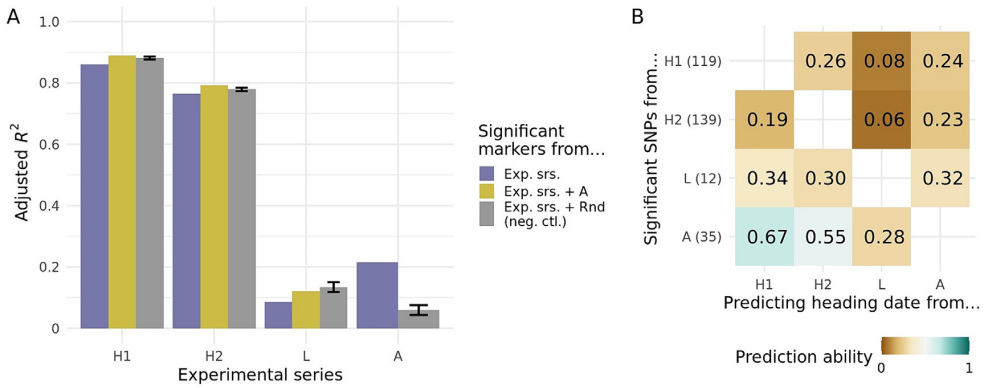

Fig. 5. Explained variance and prediction ability of the significant markers. (A) Phenotypic variance explained by QTL (R^2) with and without markers found significant exclusively in the joint data set A. Blue indicates the explained variance using markers with significant marker-trait-associations as found using only data from the respective experimental series (Exp. srs.). Yellow indicates the explained variance when considering in addition the markers found significant using the joint data set A. Gray is the negative control, error bars show the standard deviation of 50 sampling runs. (B) Prediction ability of heading time using effect estimates of significant markers. The training set (vertical axis) provides the marker estimates and the test data set (horizontal axis) genotypes are predicted. All significant markers of a data set are used for prediction. The vertical axis shows the numbers of non-collinear markers.

to the MTAs within experimental series those discovered in the joint data set A. While the adjusted explained variance increased by this, it did not exceed the negative control (dataset-specific MTAs plus random markers instead of MTAs of data set A) by a meaningful margin for series H1, H2 and L (Fig. 5A). The significant MTAs of the joint data set A could exceed a negative control (equal number of random markers). Next, we estimated the prediction ability of heading date between experimental series in the context

of marker-assisted selection (Fig. 5B). On average, the largest experimental series L showed almost twice the prediction ability compared to H1 and H2. The prediction ability of the joint data set A increased this further by a factor of 1.5.

3.4. Computer simulations revealed dependency between effective population size and QTL detection rate

Based on the genomic data of experimental series L (4633 lines), a simulation study was conducted to investigate the interplay between population size, QTL effect size, and QTL detection rate (Fig. 6A). The detection rate plateaued at a population size of approximately 2000 lines for QTL explaining 5% (QTL_medium) or 10% (QTL_large) of the phenotypic variation. Interestingly, the plateau was observed for QTL_medium at around 75% detection rate which was substantially lower than for the QTL_large, which approached 100%. In contrast, the detection rate of QTL explaining 2% of the phenotypic variation (QTL_minor) did not reach a plateau even when approaching a population size of 3404, the size of the largest individual data set of the study. This clearly highlights the benefits of combining medium-sized data into Big Data to detect small-effect QTL.

In a further step, we fixed the sample size to 300 lines and examined contrasting subpopulations with low (low N_e) and large effective population size (high N_e). The low N_e group included 50 subpopulations with effective population sizes ranging from 28.5 to 31.2 and a mean of 29.7. The high N_e group comprised 50 subpopulations with effective population sizes ranging from 58.8 to 61.2 and a mean of 59.3 (Fig. 6B). Interestingly, we observed a significant difference in the distribution of QTL detection rates between the two subpopulations (Fig. 6C): The detection rate was significantly larger in the subpopulations with high compared to low effective population size. The difference was more pronounced for QTL_large compared to QTL_minor. The QTL detection rates for QTL_minor were close to 0 for both subpopulations as the population size was not sufficient to detect QTL of that small scale (Fig. 6A). Sampling two subpopulations that were larger but still having substantially different N_e proved infeasible with the available data.

4. Discussion

The potential of Big Data to boost the accuracy of genome-wide predictions was recently documented in wheat [2]: Integrating medium-sized data into Big Data doubled the accuracy to predict grain yield of hybrid wheat. The study also showed that prediction accuracy particularly benefits from increasing sample sizes when diverse germplasm with a large effective population size N_e is involved. Published theoretical (e.g., [32,33]) and simulation studies (e.g., [34]) showed that a higher rate of QTL detection can also be expected in genome-wide association mapping when the sample size is increased. In addition to this, several other factors affect the QTL detection rate, such as population and family structure and their relationship with the phenotypic diversity, minor allele frequency, the extent of linkage disequilibrium, the precision of phenotyping, and the heritability of individual QTL [32–34]. Our small-scale simulation study builds on genomic data observed in the experimental series L and makes the simplifying assumption that the causal SNPs are included. Interestingly, we found that the increase in QTL detection rate scaled with the ratio N/N_e of the number of samples N to the effective population size N_e (Fig. 6A), which is the opposite trend than what is observed in genome-wide predictions [2]. An important conclusion is that under a fixed number of samples, maximizing diversity in the mapping population increases QTL detection rates. Maximum diversity is for example strived for when assembling core collections [35,36]. Moreover, the simulation results showed that the QTL detection rate for loci

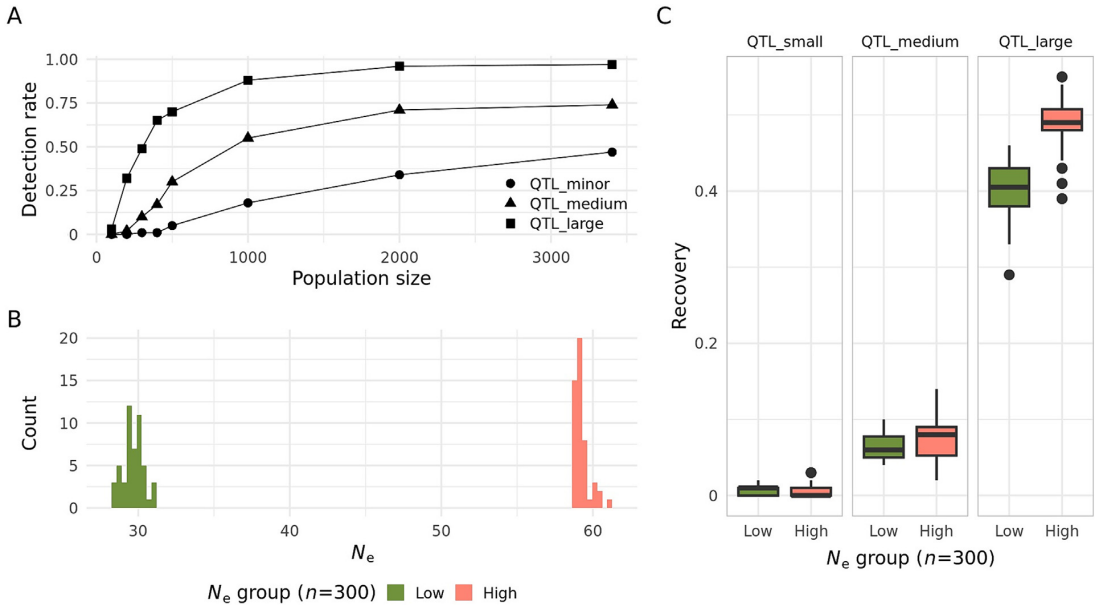

Fig. 6. Power study of genome-wide association mapping using three simulated traits and subpopulations of experimental series L with differing characteristics: (A) Detection rate of the QTL of three simulated traits by genome-wide association mapping (FDR = 5%). Each of the three traits ("QTL_minor", "QTL_medium", "QTL_large") is influenced by ten QTL where each QTL makes up for, depending on the trait, 2%, 5%, or 10% of the total phenotypic variance. The detection power is shown for different population sizes. (B) Effective population sizes (N_e) of 100 populations with 300 genotypes each that were sampled from the genotypes of experimental series L with the aim of achieving 50 population samples with high N_e and 50 populations with low N_e. (C) Detection rates in genome-wide association mapping (FDR = 5%) using populations with low ("Low") or high ("High") effective population size (N_e, see Fig. 6B).

explaining 2% (QTL_minor) of the phenotypic variation is less than 50% (Fig. 6A). Thus, in order to map small effect QTL, the integration of medium-sized data into Big Data is a promising strategy.

4.1. Potential pitfalls when integrating medium-sized data into Big Data

A major challenge in integrating data from different sources is establishing interoperability. The experimental series used in our study have been genotyped with different subsets of markers so that imputation was required to fill in marker information. A previous study showed that this kind of imputation is possible with high accuracy with a reference population of 300 genotypes [37]. As we have more than 1800 genotypes available as reference set for imputation, we expect a higher imputation accuracy. Another potential challenge for interoperability arises if experimental series use different genotyping technologies. For instance, when integrating data generated on different SNP array platforms, integration can become more challenging, but is in principle possible, as shown in a recent study [38]. Interoperability can also be established when combining different sequence-based genotyping approaches such as genotyping-by-sequencing and DArTseq [39]. An increasing complexity is expected when data are mapped against different versions of the reference genome or even when varying reference sequences from pangenome approaches are used. Establishing interoperability is even possible between SNP arrays and sequence data [39]. Nevertheless, this requires detailed *meta*-information about the underlying design of the SNP array. In summary, establishing interoperability of genomic data from disparate sources is possible and sometimes challenging, but should not hinder Big Data approaches to deciphering genetic architecture of important traits in crops.

Regarding phenotypic data, two requirements must be met to establish interoperability between different sources: First, phenotyping protocols must be comparable, and second, overlapping genotypes are needed to link the different experimental series. The phenotyping protocols for heading date varied in our study: Experimental series H1 and H2 used an alternative phenotyping protocol than experimental series L. With up to 22 overlapping genotypes, the relationship between the different phenotyping protocols could be tested (Fig. S15). Fortunately, the relationship was linear. The regression explained a large part of the variance between the two pairs of comparisons H1 versus L and H2 versus L, i.e., 91% and 95%, respectively. Thus, in our study, it was possible to establish interoperability of the data from the different experimental series without major limitations. Of course, this can be much more challenging depending on the trait being studied. Here, it is not only important to have good documentation of the metadata, e.g., according to the MIAPPE standards [40], but also to use overlapping genotypes. This requires intensive discussion within the crop communities to agree on representative standard genotypes to be used in different studies. We consider the definition of standard genotypes, which could be changing over years, to be very fruitful for meta genome-wide association mapping studies. To reduce complexity, agreeing on standard genotypes in similar mega-environments would already be a great success.

Another challenge in integrating data generated in different experimental series is to assess the role of interaction effects between genotypes and experimental series. For example, in a previous study [2], although no systematic bias was observed for grain yield, the moderate correlation between grain yield data of overlapping genotypes of each experimental series was a clear indication of interaction effects between genotypes and experimental series. In contrast, in our study, the correlation between heading dates of overlapping genotypes between experimental series was very high with Pearson moment correlation coefficients above 0.86 (Fig. S15). Therefore, analyses based on the combined heading date data should not be affected by interaction effects between genotypes and experimental series.

When only a very few genotypes link the experimental series or when the overlapping genotypes cover only a narrow range of phenotypic values, potential bias and interaction effects between genotypes and experimental series can only be investigated to a limited extent using the approach presented above. Therefore, we tested the potential of genome-wide predictions as a tool to assess interaction effects between genotypes and experimental series. The hypothesis to be tested is whether cross-validated prediction abilities within experimental series differ greatly from those between experimental series and, hence, point to substantial interaction effects between genotypes and experimental series. As outlined in detail in the results section, information on relatedness between genotypes, population size, and diversity must be considered to interpret the predictions properly. In our study, the outcomes of genome-wide predictions revealed that cross-validated prediction abilities within experimental series could approach those attained by across-experimental series genomic prediction, given that the training set had a big enough sample size and diversity (Fig. 3B, training set L). Thus, using genome-wide predictions is a further tool to inspect the potential bias and interaction effects between genotypes and experimental series but requires a very detailed analysis. In summary, our study profited from documented phenotyping protocols that facilitated interoperability, high (genomic) repeatabilities and heritabilities that were achieved by intensive data curation, and common genotypes between series that allowed to correct for series main effects.

4.2. Association mapping within the experimental series

The overlap of MTAs that we found in the original H1, H2 and L experimental series was low (Fig. 4), which is surprising at first glance because all lines originate from a similar genetic pool (Fig. 1A). One explanation is that heading date is a complex quantitative trait when dealing with germplasm adapted to the Central European target environment, and small effect earliness *per se* genes are responsible for the phenotypic variation [41] but major effect photoperiod [42–44] and vernalization response genes [44–50] are rather fixed. Concordant to this model, we see almost no association to known central photoperiod genes (Table S5 and Supplementary Methods). Our simulation study highlights the moderate QTL detection rate when they individually explain less than 5% of the phenotypic variation. Thus, a complex genetic architecture of heading date could be the reason for the low overlap of MTAs across experimental series.

One of the most striking features of the genome-wide association mapping performed within the experimental series is that the number of MTAs was higher for the experimental series H1 and H2 than for L and the joint data set A (Fig. 4). These differences can be explained by a substantially lower N_l/N_e ratio for experimental series H1 and H2 compared with L and A (Table S4) and the relationship between this ratio and the average QTL detection rate (Fig. 6C). Moreover, the use of factorial mating designs leads to a higher linkage disequilibrium between markers (Table S3) and causative alleles in H1 and H2. Also, these series have shown a much simpler population structure, which reduces the complexity of resolving collinearity by population structure between other markers and causative alleles (Fig. 1B) and can thus increase the QTL detection rate.

Interestingly, the predictive ability of significant markers of L was higher than that of the significant markers of H1 and H2, despite significantly more MTAs were found for the latter two series (Fig. 5B). A likely explanation is that the larger number of lines in L compared to H1 and H2 allows for more precise estimates of

marker effects and, hence, higher prediction accuracies favoring Big Data approaches.

4.3. Association mapping in the integrated data set

Opposed to our expectation, using the larger joint data set A lead to less MTAs being discovered than in the individual series H1 and H2. One possible hypothesis is that this behavior is due to genotypes being included that have potentially different genetic backgrounds as some of the experimental series consist predominantly of hybrids while one has inbred lines only. However, we did not see indications that would support this worry. Omitting the hybrids from the joint dataset (Al) lead to only one additional significant MTA being found while several others were lost. Using hybrids alone for GWAS did yield several MTAs that were unseen using A. However, using hybrids alone for GWAS might diminish the accuracy of the dominance effect estimation, if only few genotypes with the two homozygous states of the locus are in the population. In that case, dominance effects might be overfitted to account for what could be explained by additive effects. We therefore do not have high confidence in those additional MTAs without further study.

Another explanation for the decrease in the number of significant MTAs when integrating multiple experimental series could be based on higher extent of intrachromosomal linkage disequilibrium and a high persistence of haplotype phases within H1 or H2, as discussed previously. Although it can be beneficial to identify more QTL from H1 and H2, care must be taken when transitioning predictions from within to across populations. QTL that are based on extensive haplotype blocks that combine multiple causative alleles are detrimental to predictions across populations where that haplotype block is not preserved. This can lead to incorrect decisions when prioritizing genes for cloning as their individual effects will be different than that of the whole haplotype block. This could be the reason why the significant MTAs derived from the common data set A do not explain additional phenotypic variance beyond the series-specific MTAs (Fig. 5A), but perform much better than the dataset-specific MTAs in cross-dataset prediction despite their smaller number (Fig. 5B).

Comparing the joint data set A to the series L, a third factor influencing the number of detected MTAs can be illustrated: More MTAs were detected in A than in L that also had a higher prediction ability. Compared to series L, the N_e of A is higher, which goes along with a lower degree of linkage disequilibrium, but the data set A also has double the number of individuals than L. Considering that in the simulation study both a higher N_e, given a constant sample size, and a higher sample size on its own were beneficial for GWAS power, the joint data set A might strike a good balance between sufficient linkage disequilibrium to link causative loci to markers and diverse, independent population sampling to achieve higher predictive ability beyond data set borders. Additionally, as shown in the simulation study, the use of Big Data is beneficial to identify small-size QTL and the results of genome-wide association mapping studies based on Big Data therefore provide a promising entry point for marker-assisted selection and dissecting the genetic architecture of heading date in wheat.

4.4. Outlook

In our study, we integrated data from research projects with data from breeding programs at the raw data level and were thus able to increase the power of GWAS. For this purpose, we had all the necessary information for the genomic and phenotypic data and could successfully use this high-quality data for association mapping. The bulk of data is collected in Central Europe as part of private wheat breeding programs. In addition, there is also very deep data from variety trials as part of the official approval process. The data of private breeding programs or official variety testing have a sensitivity that can severely limit sharing of all the necessary information (for example, details on the design of SNP arrays), thus, hampering integrated analyses. As an alternative, methods in the field of human genetics have been developed for this purpose. Besides simple approaches to integrate data sets on the level of summary statistics, there are advanced methods like sPLINK, where privacy-aware GWAS are possible without having to share critical information [51]. These approaches are promising to enable Big Data approaches in plant breeding also for analyses across different competing breeding programs. Developing the necessary data ecosystem with clear incentive schemes for data sharing is a necessary prerequisite and deserves further efforts.

CRediT authorship contribution statement

Moritz Lell: Software, Formal analysis, Data curation, Writing – original draft, Visualization. **Yusheng Zhao:** Conceptualization, Methodology, Software, Writing – review & editing, Supervision. **Jochen C. Reif:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge funding within the Wheat BigData Project (German Federal Ministry of Food and Agriculture, FKZ2818408B18).

Appendix A. Supplementary data

Supplementary data for this article can be found online at https://doi.org/10.1016/j.cj.2024.03.005.

References

[1] Y. Xu, X. Liu, J. Fu, H. Wang, J. Wang, C. Huang, B.M. Prasanna, M.S. Olsen, G. Wang, A. Zhang, Enhancing genetic gain through genomic selection: from livestock to plants, Plant Commun. 1 (2020) 100005.
[2] Y. Zhao, P. Thorwarth, Y. Jiang, N. Philipp, A.W. Schulthess, M. Gils, P.H.G. Boeven, C.F.H. Longin, J. Schacht, E. Ebmeyer, V. Korzun, V. Mirdita, J. Dörnte, U. Avenhaus, R. Horbach, H. Cöster, J. Holzapfel, L. Ramgraber, S. Kühnle, P. Varenne, A. Starke, F. Schürmann, S. Beier, U. Scholz, F. Liu, R.H. Schmidt, J.C. Reif, Unlocking big data doubled the accuracy in predicting the grain yield in hybrid wheat, Sci. Adv. 7 (2021) eabf9106.
[3] S. Cao, D. Xu, M. Hanif, X. Xia, Z. He, Genetic architecture underpinning yield component traits in wheat, Theor. Appl. Genet. 133 (2020) 1811–1823.
[4] B.J. Hayes, P.M. Visscher, M.E. Goddard, Increased accuracy of artificial selection by using the realized relationship matrix, Genet. Res. 91 (2009) 47–60.
[5] P. Schopp, D. Müller, F. Technow, A.E. Melchinger, Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium, Genetics 205 (2016) 441–454.
[6] M.E. Cannon, K.L. Mohlke, Deciphering the emerging complexities of molecular mechanisms at GWAS loci, Am. J. Hum. Genet. 103 (2018) 637–653.
[7] S. McCouch, Z.K. Navabi, M. Abberton, N.L. Anglin, R.L. Barbieri, M. Baum, K. Bett, H. Booker, G.L. Brown, G.J. Bryan, L. Cattivelli, D. Charest, K. Eversole, M. Freitas, K. Ghamkhar, D. Grattapaglia, R. Henry, M.C. Valadares Inglis, T. Islam, Z. Kehel, P.J. Kersey, G.J. King, S. Kresovich, E. Marden, S. Mayes, M.N. Ndjiondjop, H.T. Nguyen, S.R. Paiva, R. Papa, P.W.B. Phillips, A. Rasheed, C. Richards, M. Rouard, M.J. Amstalden Sampaio, U. Scholz, P.D. Shaw, B. Sherman, S.E. Staton, N. Stein, J. Svensson, M. Tester, J.F. Montenegro Valls, R. Varshney, S. Visscher, E. Von Wettberg, R. Waugh, P. Wenzl, L.H. Rieseberg, Mobilizing crop biodiversity, Mol. Plant 13 (2020) 1341–1344.

[8] M. Erbe, B. Gredler, F.R. Seefried, B. Bapst, H. Simianer, A function accounting for training set size and marker density to model the average accuracy of genomic prediction, PLoS ONE 8 (2013) e81046.
[9] J. Yu, G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, E.S. Buckler, A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, Nat. Genet. 38 (2006) 203–208.
[10] L. Tibbs Cortes, Z. Zhang, J. Yu, Status and prospects of genome-wide association studies in plants, Plant Genome 14 (2021) e20077.
[11] Y. Zhao, Z. Li, G. Liu, Y. Jiang, H.P. Maurer, T. Würschum, H.P. Mock, A. Matros, E. Ebmeyer, R. Schachschneider, E. Kazman, J. Schacht, M. Gowda, C.F.H. Longin, J.C. Reif, Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 15624–15629.
[12] S. Wang, D. Wong, K. Forrest, A. Allen, S. Chao, B.E. Huang, M. Maccaferri, S. Salvi, S.G. Milner, L. Cattivelli, A.M. Mastrangelo, A. Whan, S. Stephen, G. Barker, R. Wieseke, J. Plieske, M. Lillemo, D. Mather, R. Appels, R. Dolferus, G. Brown-Guedira, A. Korol, A.R. Akhunova, C. Feuillet, J. Salse, M. Morgante, C. Pozniak, M.C. Luo, J. Dvorak, M. Morell, J. Dubcovsky, M. Ganal, R. Tuberosa, C. Lawley, I. Mikoulitch, C. Cavanagh, K.J. Edwards, M. Hayden, E. Akhunov, Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array, Plant Biotechnol. J. 12 (2014) 787–796.
[13] B.N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, PLoS Genet. 5 (2009) e1000529.
[14] Y. Zhao, M.F. Mette, M. Gowda, C.F.H. Longin, J.C. Reif, Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat, Heredity 112 (2014) 638–645.
[15] H. Hack, H. Bleiholder, L. Buhr, U. Meier, U. Schnock-Fricke, E. Weber, A. Witzenberger, Einheitliche Codierung der phänologischen Entwicklungsstadien mono- und dikotyler Pflanzen-Erweiterte BBCH-Skala, Allgemein. Nachrichtenblatt Dtsch. Pflanzenschutzdienstes 44 (1992) 265–270.
[16] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021.
[17] A.R. Gilmour, B.J. Gogel, B.R. Cullis, S.J. Welham, R. Thompson, ASReml User Guide Release 4.1 Structural Specification, VSN International Ltd., Hemel Hempstead, UK, 2015, www.vsni.co.uk.
[18] J.S. Rogers, Measures of genetic similarity and genetic distance, in: M.R. Wheeler (Ed.), Stud. Genet. VII, University of Texas, Austin, TX, USA, 1972, pp. 145–153.
[19] J.C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (1966) 325–338.
[20] W.G. Hill, A. Robertson, Linkage disequilibrium in finite populations, Theor. Appl. Genet. 38 (1968) 226–231.
[21] R.S. Waples, A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci, Conserv. Genet. 7 (2006) 167–184.
[22] W.G. Hill, B.S. Weir, Variances and covariances of squared linkage disequilibria in finite populations, Theor. Popul. Biol. 33 (1988) 54–78.
[23] A.M. Bernal-Vasquez, H.F. Utz, H.P. Piepho, Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML, Theor. Appl. Genet. 129 (2016) 787–804.
[24] P. Pérez, G. de los Campos, Genome-wide regression and prediction with the BGLR statistical package, Genetics 198 (2014) 483–495.
[25] P.M. VanRaden, Efficient methods to compute genomic predictions, J. Dairy Sci. 91 (2008) 4414–4423.
[26] J.M. Alvarez-Castro, O. Carlborg, A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis, Genetics 176 (2006) 1151–1167.
[27] Y. Zhao, M. Gowda, T. Würschum, C.F.H. Longin, V. Korzun, S. Kollers, R. Schachschneider, J. Zeng, R. Fernando, J. Dubcovsky, J.C. Reif, Dissecting the genetic architecture of frost tolerance in Central European winter wheat, J. Exp. Bot. 64 (2013) 4453–4460.
[28] Z. Zhang, E. Ersoz, C.Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, J. Yu, D.K. Arnett, J.M. Ordovas, E.S. Buckler, Mixed linear model approach adapted for genome-wide association studies, Nat. Genet. 42 (2010) 355.
[29] Y.L.B. Rubio, J.L.G. Duarte, R.O. Bates, C.W. Ernst, D. Nonneman, G.A. Rohrer, A. King, S.D. Shackelford, T.L. Wheeler, R.J.C. Cantet, J.P. Steibel, Meta-analysis of genome-wide association from genomic prediction models, Animal Genet. 47 (2015) 36–48.
[30] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Methodol. 57 (1995) 289–300.
[31] P. Yin, X. Fan, Estimating R^2 shrinkage in multiple regression: a comparison of different analytical methods, J. Exp. Educ. 203–224 (2001).
[32] P.M. Visscher, N.R. Wray, Q. Zhang, P. Sklar, M.I. McCarthy, M.A. Brown, J. Yang, 10 Years of GWAS discovery: biology, function, and translation, Am. J. Hum. Genet. 101 (2017) 5–22.
[33] M. Wang, S. Xu, Statistical power in genome-wide association studies and quantitative trait locus mapping, Heredity 123 (2019) 287–306.
[34] J. Yang, N.A. Zaitlen, M.E. Goddard, P.M. Visscher, A.L. Price, Advantages and pitfalls in the application of mixed-model association methods, Nat. Genet. 46 (2014) 100–106.
[35] J. Franco, J. Crossa, M.L. Warburton, S. Taba, Sampling strategies for conserving maize diversity when forming core subsets using genetic markers, Crop Sci. 46 (2006) 854–864.
[36] S.G. Milner, M. Jost, S. Taketa, E.R. Mazón, A. Himmelbach, M. Oppermann, S. Weise, H. Knüpffer, M. Basterrechea, P. König, D. Schüler, R. Sharma, R.K. Pasam, T. Rutten, G. Guo, D. Xu, J. Zhang, G. Herren, T. Müller, S.G. Krattinger, B. Keller, Y. Jiang, M.Y. González, Y. Zhao, A. Habekuß, S. Färber, F. Ordon, M. Lange, A. Börner, A. Graner, J.C. Reif, U. Scholz, M. Mascher, N. Stein, Genebank genomics highlights the diversity of a global barley collection, Nat. Genet. 51 (2019) 319–326.
[37] S. He, Y. Zhao, M.F. Mette, R. Bothe, E. Ebmeyer, T.F. Sharbel, J.C. Reif, Y. Jiang, Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat, (Triticum aestivum L), BMC Genomics 16 (2015) 168.
[38] A. Gogna, A.W. Schulthess, M.S. Röder, M.W. Ganal, J.C. Reif, Gabi wheat a panel of European elite lines as central stock for wheat genetic research, Sci. Data 9 (2022) 538.
[39] A.W. Schulthess, S.M. Kale, F. Liu, Y. Zhao, N. Philipp, M. Rembe, Y. Jiang, U. Beukert, A. Serfling, A. Himmelbach, J. Fuchs, M. Oppermann, S. Weise, P.H.G. Boeven, J. Schacht, C.F.H. Longin, S. Kollers, N. Pfeiffer, V. Korzun, M. Lange, U. Scholz, N. Stein, M. Mascher, J.C. Reif, Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement, Nat. Genet. 54 (2022) 1544–1552.
[40] E.A. Papoutsoglou, D. Faria, D. Arend, E. Arnaud, I.N. Athanasiadis, I. Chaves, F. Coppens, G. Cornut, B.V. Costa, H. Ćwiek-Kupczyńska, B. Droesbeke, R. Finkers, K. Gruden, A. Junker, G.J. King, P. Krajewski, M. Lange, M.A. Laporte, C. Michotey, M. Oppermann, R. Ostler, H. Poorter, R. Ramı'rez-Gonzalez, Ž. Ramšak, J.C. Reif, P. Rocca-Serra, S.A. Sansone, U. Scholz, F. Tardieu, C. Uauy, B. Usadel, R.G.F. Visser, S. Weise, P.J. Kersey, C.M. Miguel, A.F. Adam-Blondon, C. Pommier, Enabling reusability of plant phenomic datasets with MIAPPE 1.1, New Phytol. 227 (2020) 260–273.
[41] M. Zikhali, S. Griffiths, The Effect of earliness per se (Eps) genes on flowering time in bread wheat, in: Y. Ogihara, S. Takumi, H. Handa (Eds.), Advances in Wheat Genetics: from Genome to Field, Springer, Tokyo, Japan, 2015, pp. 339–345.
[42] J. Beales, A. Turner, S. Griffiths, J.W. Snape, D.A. Laurie, A pseudo-response regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (*Triticum aestivum* L.), Theor. Appl. Genet. 115 (2007) 721–733.
[43] E.P. Wilhelm, A.S. Turner, D.A. Laurie, Photoperiod insensitive *Ppd-A1a* mutations in tetraploid wheat (*Triticum durum* Desf.), Theor. Appl. Genet. 118 (2009) 285–294.
[44] A. Díaz, M. Zikhali, A.S. Turner, P. Isaac, D.A. Laurie, Copy number variation affecting the *photoperiod-b1* and *vernalization-a1* genes is associated with altered flowering time in wheat (*Triticum aestivum*), PLoS ONE 7 (2012) e33234.
[45] L. Yan, A. Loukoianov, G. Tranquilli, M. Helguera, T. Fahima, J. Dubcovsky, Positional cloning of the wheat vernalization gene *VRN1*, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 6263–6268.
[46] B. Trevaskis, M.N. Hemming, E.S. Dennis, W.J. Peacock, The molecular basis of vernalization-induced flowering in cereals, Trends Plant Sci. 12 (2007) 352–357.
[47] A. Distelfeld, C. Li, J. Dubcovsky, Regulation of flowering in temperate cereals, Curr. Opin. Plant Biol. 12 (2009) 178–184.
[48] A. Distelfeld, G. Tranquilli, C. Li, L. Yan, J. Dubcovsky, Genetic and molecular characterization of the *VRN2* loci in tetraploid wheat, Plant Physiol. 149 (2009) 245–257.
[49] S. Shimada, T. Ogawa, S. Kitagawa, T. Suzuki, C. Ikari, N. Shitsukawa, T. Abe, H. Kawahigashi, R. Kikuchi, H. Handa, K. Murai, A genetic network of flowering-time genes in wheat leaves, in which an *APETALA1/ FRUITFULL*-like gene, *VRN1*, is upstream of *FLOWERING LOCUS T*, Plant J. 58 (2009) 668–681.
[50] A. Distelfeld, J. Dubcovsky, Characterization of the maintained vegetative phase deletions from diploid wheat and their effect on *VRN2* and *FT* transcript levels, Mol. Genet. Genomics 283 (2010) 223–232.
[51] R. Nasirigerdeh, R. Torkzadehmahani, J. Matschinske, T. Frisch, M. List, J. Späth, S. Weiss, U. Völker, E. Pitkänen, D. Heider, N.K. Wenke, G. Kaissis, D. Rueckert, T. Kacprowski, J. Baumbach, sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies, Genome Biol. 23 (2022) 32.
[52] T. Galili, dendextend: An R package for visualizing, adjusting, and comparing trees of hierarchical clustering, Bioinformatics 31 (2015) 3718–3720.

2.3 Breaking down data silos across companies to train genome-wide predictions -- a feasibility study in wheat

Published: In 2025, *Plant Biotechnology Journal*, in press.

DOI: 10.1111/pbi.70095

Authors: Moritz Lell, Abhishek Gogna, Vincent Kloesgen, Ulrike Avenhaus, Jost Dörnte, Wera Maria Eckhoff, Tobias Eschholz, Mario Gils, Martin Kirchhoff, Michael Koch, Sonja Kollers, Nina Pfeiffer, Matthias Rapp, Valentin Wimmer, Markus Wolf, Jochen Reif, Yusheng Zhao

Abstract:

Big Data, combined with artificial intelligence (AI) techniques, holds the potential to significantly enhance the accuracy of genome-wide predictions. Motivated by the success reported for wheat hybrids, we extended the scope to inbred lines by integrating phenotypic and genotypic data from four commercial wheat breeding programs. Acting as an academic data trustee, we merged these data with historical experimental series from previous public-private partnerships. The integrated data spanned twelve years, 168 environments, and provided a genomic prediction training set of up to ~9,500 genotypes for grain yield, plant height, and heading date. Despite the heterogenous phenotypic and genotypic data, we were able to obtain high-quality data by implementing rigorous data curation, including SNP imputation. We utilized the data to compare genomic best linear unbiased predictions with convolutional neural network-based genomic prediction. Our analysis revealed that we could flexibly combine experimental series for genomic prediction, with prediction ability steadily improving as the training set sizes increased, peaking at around 4,000 genotypes. As training set sizes were further increased, the gains in prediction ability decreased, approaching a plateau well below the theoretical limit defined by the square root of the heritability. Potential avenues, such as designed training sets or novel non-linear prediction approaches, could overcome this plateau and help to more fully exploit the high-value big data generated by breaking down data silos across companies.

Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2025), pp. 1–16 doi: 10.1111/pbi.70095

Breaking down data silos across companies to train genome-wide predictions: A feasibility study in wheat

Moritz Lell[1], Abhishek Gogna[1], Vincent Kloesgen[1], Ulrike Avenhaus[2], Jost Dörnte[3], Wera Maria Eckhoff[4], Tobias Eschholz[5], Mario Gils[5], Martin Kirchhoff[5], Michael Koch[3], Sonja Kollers[4], Nina Pfeiffer[6], Matthias Rapp[2], Valentin Wimmer[4], Markus Wolf[7], Jochen Reif[1] and Yusheng Zhao[1,*]

[1]*Leibniz Institute for Plant Genetics and Crop Plant Research, Seeland, Germany*
[2]*W. von Borries-Eckendorf GmbH & Co. KG, Leopoldshöhe, Germany*
[3]*Deutsche Saatveredelung AG, Lippstadt, Germany*
[4]*KWS SAAT SE & Co. KGaA, Einbeck, Germany*
[5]*Nordsaat Saatzucht GmbH, Langenstein, Germany*
[6]*KWS LOCHOW GmbH, Northeim, Germany*
[7]*SU BIOTEC GmbH, Gatersleben, Germany*

Received 6 August 2024;
revised 7 March 2025;
accepted 7 April 2025.
**Correspondence* (Tel +49 39482 5-404; fax +49 39482 5-137; email zhao@ipk-gatersleben.de)

Keywords: wheat, genomic prediction, data integration, big data, imputation.

Summary

Big data, combined with artificial intelligence (AI) techniques, holds the potential to significantly enhance the accuracy of genome-wide predictions. Motivated by the success reported for wheat hybrids, we extended the scope to inbred lines by integrating phenotypic and genotypic data from four commercial wheat breeding programs. Acting as an academic data trustee, we merged these data with historical experimental series from previous public–private partnerships. The integrated data spanned 12 years, 168 environments, and provided a genomic prediction training set of up to ~9500 genotypes for grain yield, plant height and heading date. Despite the heterogeneous phenotypic and genotypic data, we were able to obtain high-quality data by implementing rigorous data curation, including SNP imputation. We utilized the data to compare genomic best linear unbiased predictions with convolutional neural network-based genomic prediction. Our analysis revealed that we could flexibly combine experimental series for genomic prediction, with prediction ability steadily improving as the training set sizes increased, peaking at around 4000 genotypes. As training set sizes were further increased, the gains in prediction ability decreased, approaching a plateau well below the theoretical limit defined by the square root of the heritability. Potential avenues, such as designed training sets or novel non-linear prediction approaches, could overcome this plateau and help to more fully exploit the high-value big data generated by breaking down data silos across companies.

Introduction

In the last decade, genome-wide prediction has revolutionized plant breeding by providing an estimate of the genotypic value of a new candidate variety from its genomic profile and phenotype observations of related genotypes (Meuwissen *et al*., 2001). This allows some of the expensive field trials to be omitted, as early breeding trial stages with low observed heritability can be replaced using genomic prediction (Riedelsheimer and Melchinger, 2013). Thereby, the time required to select superior genotypes can be shortened (Beyene *et al*., 2021), even for complex traits controlled by many genes.

Given the large number of genes and their interactions that influence most agronomic traits, feasible population sizes in breeding programs do not allow us to infer the influence of each individual locus and locus interactions. Modern statistical techniques have been developed to address the shortcomings of traditional linear models in capturing complex gene interactions and genotype relationships, effectively reducing the influence of large amounts of noisy data (Chafai *et al*., 2023). The most practical and widely used methods are parametric or semi-parametric models (Montesinos-López *et al*., 2022). These models generally introduce *a priori* assumptions about the genetic effects, either by regularization of parameter estimates (de los Campos *et al*., 2013) or by selecting informative prior distributions in a Bayesian framework (Gianola, 2013). Originally, the parameters to be estimated were effects of genetic loci, as in Ridge-Regression Best Linear Unbiased Prediction (rrBLUP, Meuwissen *et al*., 2001), but genetic effects of individuals can instead be modelled directly by including the expected correlation of their breeding values in the model. When pedigrees are unknown, these relationships can be inferred from genomic data, giving rise to Genomic BLUP (GBLUP, VanRaden, 2008). The resulting genomic kinship matrix considers both the additive-genetic relationship, that is, the pedigree of individuals, and shared linkage groups, which link the tested Single Nucleotide

Please cite this article as: Lell, M., Gogna, A., Kloesgen, V., Avenhaus, U., Dörnte, J., Eckhoff, W.M., Eschholz, T., Gils, M., Kirchhoff, M., Koch, M., Kollers, S., Pfeiffer, N., Rapp, M., Wimmer, V., Wolf, M., Reif, J. and Zhao, Y. (2025) Breaking down data silos across companies to train genome-wide predictions: A feasibility study in wheat. *Plant Biotechnol. J.*, https://doi.org/10.1111/pbi.70095.

Polymorphism (SNP) loci to causative loci (Habier *et al.*, 2013). As a consequence, a more diverse population may be a more difficult target for genomic prediction, as more and smaller linkage groups have to be accounted for (Daetwyler *et al.*, 2013).

In recent years, deep learning approaches for genomic predictions have gained attention (Ma *et al.*, 2018; Montesinos-Lopez *et al.*, 2021). In contrast to conventional methods mentioned above, these do not use quantitative genetic models but are based on flexible arrangements of many non-linear transformations of the input data (neurons) to detect (1) patterns in the input data, and (2) their relationship to the phenotype. The parameters of those transformations are optimized by supervised learning on a test set. This is expected to provide advantages where crop traits are strongly influenced by complex interaction effects that are not covered by the theory behind one of the more classical models (Pérez-Enciso and Zingaretti, 2019). Besides this, neural networks training has linear time complexity with respect to sample size. This avoids the computing time explosion that researchers face whenever kinship matrices have to be inverted, like in the case of GBLUP (Pook *et al.*, 2020). While in the beginning, multilayer perceptrons have been used for genomic prediction, recently convolutional networks have demonstrated their ability to capture linkage patterns (Pook *et al.*, 2020). However, the full potential of deep learning has yet to be fully realized by providing much larger training sets than are available to single institutions. As such, a comprehensive evaluation of the advantages and limitations of deep learning techniques when applied to large data sets is urgently needed, particularly in the domain of plant breeding.

The prediction ability of genomic prediction, defined as the correlation between true and predicted phenotype, is significantly influenced by several characteristics of the training and test set, such as (1) the size of the population, (2) the diversity and relatedness between the genotypes, (3) the degree of linkage disequilibrium (LD) and (4) the quality of the phenotypic and genotypic data (Schopp *et al.*, 2016). Increasing the training set is a straightforward and promising strategy to achieve high levels of prediction ability for diverse populations in hybrid wheat breeding practice (Zhao *et al.*, 2021). Several steps in this direction have already been taken. For example, in a study including more than 8000 wheat landraces, prediction abilities of 0.68 for thousand kernel weight within the population could be achieved (Crossa *et al.*, 2016). In another study, a massive data set of more than 10 000 wheat lines was phenotyped in an unreplicated single-year design, and prediction abilities close to one for grain yield could be attained in a cross-validation (Norman *et al.*, 2018). These results are very encouraging but require either a large investment of resources beyond the reach of most institutes and companies or reduced phenotyping intensities that are below the standards of commercial wheat breeding for variety development in terms of numbers of environments and replications.

Multiple institutional and/or across-company collaborations for mutual benefit are an attractive concept to increase the populations for training genome-wide prediction models but are hampered by heterogeneous and non-orthogonal (unbalanced) data. Commercial breeding trials are unbalanced in that they screen a large number of genotypes and evaluate their phenotypes in only a small number of environments. Selected genotypes from the first breeding stages are then evaluated in more environments in the next season. Therefore, as the reliability of the estimate of a candidate's performance rises, the number of available candidates drops. Combining several of such trials would produce a data set that includes a large number of early-stage genotypes, but also late-stage data for a larger number of candidates than what is feasible for each individual actor. In an earlier study, on which we build here, combining multiple historic wheat trials doubled the prediction ability for grain yield for hybrids (Zhao *et al.*, 2021). Combining different data sets is therefore promising. In this study, we investigated the impact on genomic prediction of combining such historical data with routine breeding data from four companies. Our objectives were (1) to investigate whether it is possible to perform an integrated analysis of disparate phenotypic and genotypic data sets and how to perform quality control of such a task, (2) to examine what prediction abilities can be expected when using genomic prediction beyond the confines of individual experimental series and how well multiple series can be combined to form larger training sets for genomic prediction as well as to explore the potential of deep learning models for enhancing this process and (3) to test approaches to improve the training set by drawing subsets from the full data, distilling the most reliable data and potentially increasing prediction ability.

Results

Absence of genetically divergent subpopulations revealed by accurately imputed genotypic data

Given the block-wise gaps in the SNP data resulting from the integration of the heterogeneous SNP array platforms (Figures 1 and 2a), we conducted a validation of the imputation accuracy. This was estimated by masking and imputing some SNP data in a blocked and a random approach and then calculating the ratio of correctly imputed SNP calls to the masked SNP calls. In the blocked approach, almost all markers were imputed with accuracies above 0.75 (95th percentile) and most (75th percentile) even above 0.93 (Figure 2b). Imputation using the random masking approach was possible with even higher accuracy. The 95% and 75%-percentiles of the imputation accuracy were 0.89 and 0.99, respectively.

The imputed SNP data covered most of the wheat genome (Figure 3a). Markers satisfying the liberal missing value criterion (at most 80% missing and imputed values) were found at densities of about 1–100 markers per 10 Mbp (Figure 3a). The smaller marker set resulting from the strict missing value criterion, that is, with at most 30% missing and imputed values, covered the genome at 1–30 markers per Mbp (Figure 3b). Coverage near the chromosome centres was markedly weaker than near the ends, and especially for the strict missing value criterion, large gaps were present near the chromosome centres.

The principal coordinate analysis based on the Rogers' distances revealed added diversity by combining the series of the study (Figure 4). Parts of some series were found in regions of the diversity space that were only sparsely covered by other series. For example, parts of series 6 and 7 were outside of the diversity space of series 1–3 but were quite similar to each other. Besides this tendency towards complementarity, none of the series formed a clearly distinct cluster separated from all other series.

Phenotypic data are of high quality and consistent with genotypic data

We have integrated and curated phenotypic data generated in 105 000 grain yield plots as part of large public–private partnerships or wheat breeding programs in Central Europe. As a measure of the quality of the phenotypic data, broad-sense

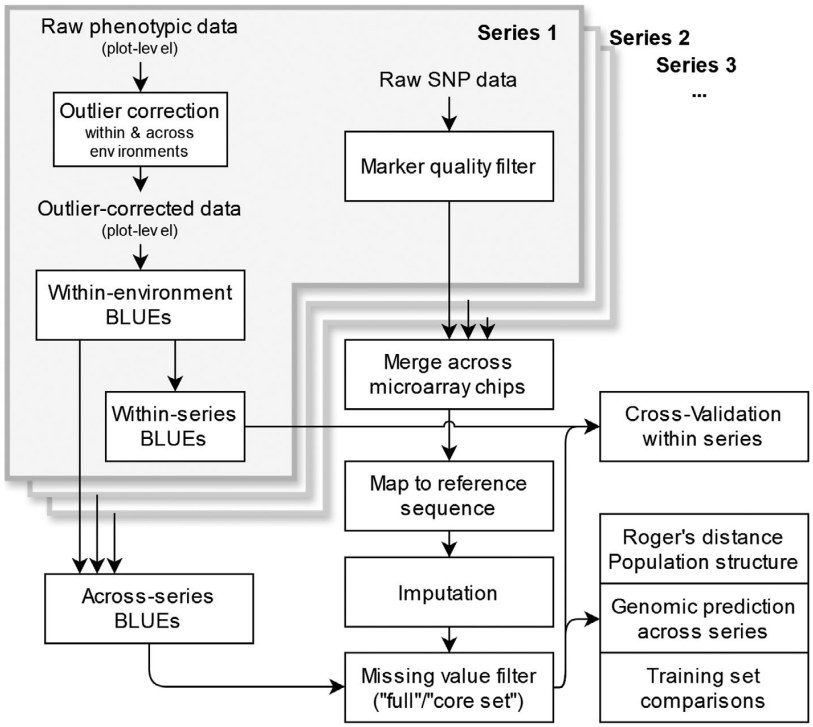

Figure 1 Schematic overview of the data-processing steps.

heritability was estimated for three traits for each experimental series (Figure 5a). Heritabilities ranging from 0.86 to almost 0.98 were obtained for heading date, from 0.81 to 0.99 for plant height, and from 0.74 to 0.93 for grain yield. Thus, the quality of the phenotypic data was excellent for all traits.

Within-series cross-validation showed moderate to high prediction abilities. Medians across the 20 replications ranged from 0.50 to 0.98, except for series 5 for heading date (0.26) and series 3 for grain yield (0.36, Figure 5b). The latter can be explained by the fact that series 3 consists of only 142 inbred lines (Table 2). For series 5–8, which include data from commercial wheat breeding programs, the prediction abilities for heading date and plant height were lower than for the historical series. This can be explained by the unbalanced nature of these series. Due to economic efficiency constraints, these traits were evaluated in fewer environments than grain yield; in some cases, the average number of environments per genotype was even <2 (Table 1). Under these circumstances, an impact on the prediction ability is to be expected. Overall, the SNP data provided good predictions of the phenotypic data, indicating a successful data integration.

Convolutional neural networks become competitive with larger training set sizes

Training set sizes greatly impacted performance achieved with convolutional neural networks (CNN). For small training set sizes of 10% (for yield: around 950 genotypes, Table 2), the CNN showed lower prediction abilities by a margin of about 0.15 compared to GBLUP (difference of the medians, Figure 5c). Interestingly, as the training set sizes increased to 80% (about 7.600 genotypes for yield), this margin narrowed and CNN performed similarly to GBLUP, even exceeding it at a few individual iterations (Figure 5c). The rate at which this gap closed varied with the trait and for yield, where the largest training sets were available, the gap closed earlier than others.

We also benchmarked the performance of the CNN against GBLUP, with the data siloed into the individual experimental series. The training set sizes for these predictions were 90% of the genotype counts of the individual experimental series, which were 128 to 3332 genotypes in the example of yield (Table 2). High GBLUP prediction abilities for series were associated with a strong CNN performance, leaving only a small difference between CNN and GBLUP prediction abilities (Figure 5b): For example, for yield, the series 2 and 4 lead to the most accurate GBLUP predictions (0.85 and 0.69) and also the smallest gap (0.06 and 0.13) between GBLUP and CNN predictions. The picture was reversed for series 3, with both the worst median GBLUP prediction ability (0.36) and the largest gap to the median CNN prediction ability (0.46). The only outlier to this pattern is the experimental series 2 for heading date and plant height, where a high GBLUP prediction ability coincides with a large gap to the CNN performance.

We did not investigate the performance of the CNN in the remainder of the study but used GBLUP because of the high computational burden and the comparable or superior performance of GBLUP.

Experimental series can be flexibly combined in genomic prediction training sets, approaching a plateau in their prediction ability

To test the ability of predicting unknown genotypes given the collection of series in this study, we chose different combinations

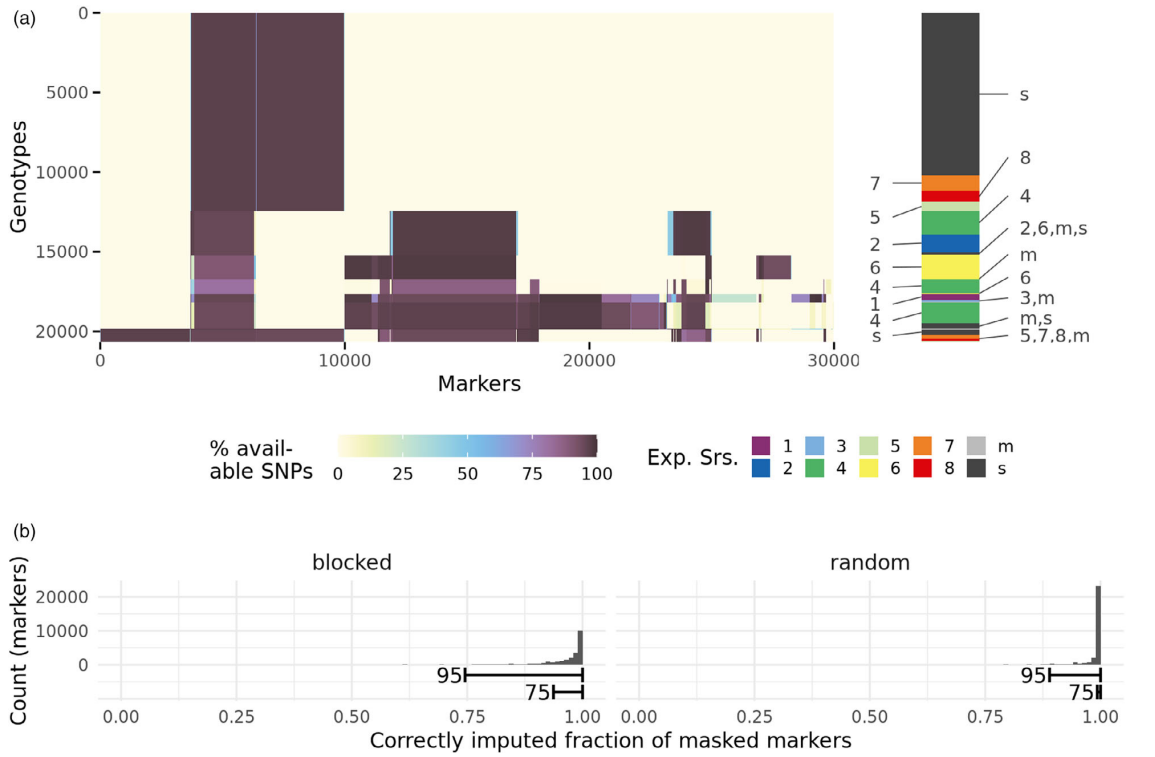

Figure 2 Imputation data basis and accuracy. (a) Available single-nucleotide polymorphism (SNP) calls for different experimental series after quality filtering and merging of individual genotyping batches and before imputation. Dark areas indicate available genomic data for the respective markers and genotypes. Both are ordered to cluster markers/genotypes with similar availability together. The experimental series that genotypes belong to is indicated on the right side by colours and labels. Labels of multiple adjacent small groups are merged to reduce visual cluttering. The digits indicate the experimental series, 's' indicates genotypes that have no phenotypic data (only genomic data), and 'm' indicates genotypes that appear in multiple experimental series, for example, common reference genotypes. (b) Imputation performance estimated by partially masking and imputing known genomic data. Histograms show the fractions of correctly imputed calls per marker, so the baseline for random guessing is 0.5. The markings '95' and '75' encompass the respective percentiles: 95% and 75% quantiles, respectively. The two subplots 'blocked' and 'random' discern the two different masking strategies (mask many SNP calls for single genotypes or randomly mask SNP calls, see 'Methods').

of the series as training sets and assigned all other series as test sets. We then derived the genomic prediction ability by comparing the predictions with the across-series BLUEs of the test sets, calculating the prediction abilities separately for each series in the test set. Cumulating different combinations and numbers of series resulted in training sets of different sizes, by which the prediction abilities were ordered (Figure 6a). As a general trend, we observed an increase in prediction ability with an increasing training set size, but this increase approached a plateau beyond training set sizes of about 4000 individuals. For many test sets, the prediction abilities approached those obtained by using the series' own data in the cross-validated genomic prediction (Figure 6a, dashed lines) but did not come close to the upper limit defined by the square root of the heritabilities (solid lines).

The above scenarios are based on imputed marker data and the liberal SNP filtering criterion (<80% of pre-imputation missing values). Consistent with the high imputation accuracy of the missing SNP data, we also observed a decrease in prediction ability in genome-wide prediction on average across the three traits when the same liberal filtering criterion was used but no missing values were imputed, or when the strict SNP filtering criterion (<30% of pre-imputation missing values) was used (Figure 6b). Thus, liberal filtering combined with imputation also appears to be the most successful strategy for the block-wise missing marker data underlying the used data set.

We then investigated whether the number of series included in the training set had a significant effect on prediction ability, given an unchanged training set size. We generated training sets of 800 genotypes coming from either single or multiple experimental series and compared the resulting prediction abilities: On average, using a single series resulted in a lower prediction ability of 0.02 for heading date and below 0.01 for plant height and grain yield (Figure S1). The standard deviations between replications were much larger, ranging from 0.07 to 0.08. The differences in standard deviation were below ±0.01 for all traits. Thus, for a medium-sized training set, the number of experimental series included in the training set had no meaningful effect on the prediction ability.

Most experimental series are compatible with each other for prediction

Observing the increase in genomic prediction ability with increasing training set size, we investigated whether the inclusion

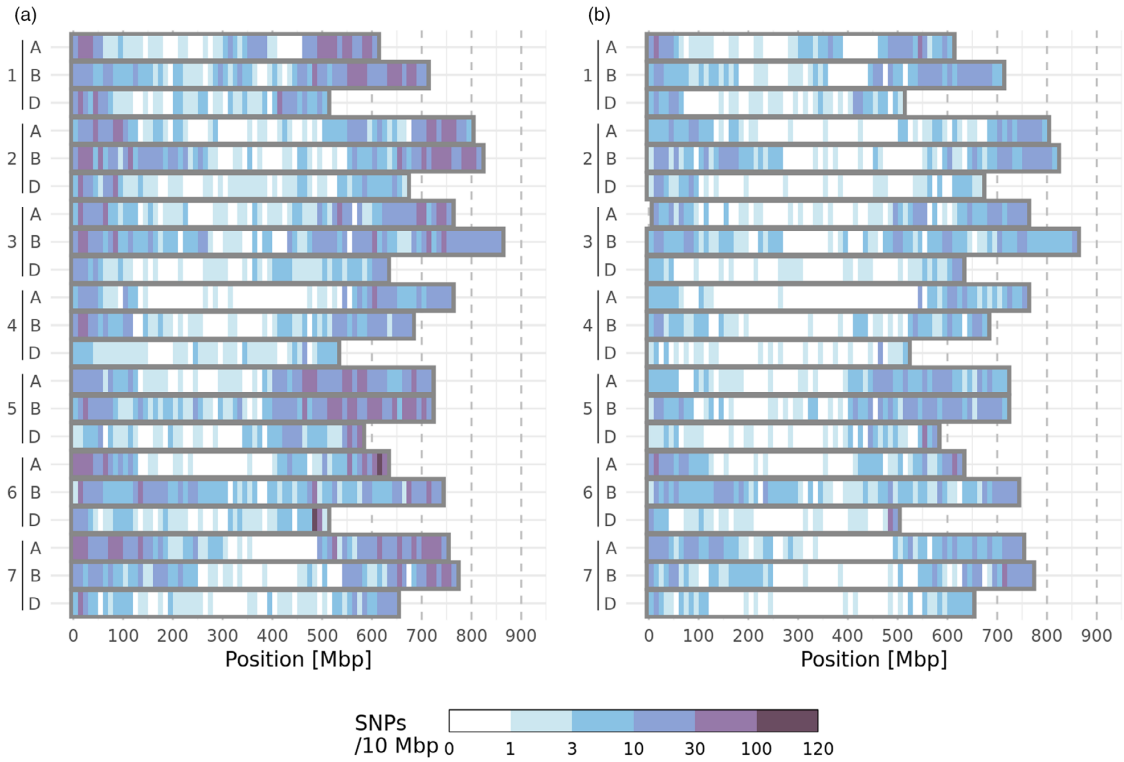

Figure 3 Density of SNPs used for genomic prediction when ordered according to their genomic position. Chromosome names are denoted on the vertical axes, nucleotide position on the horizontal axis. Density is shown by colour for 10 Mbp chunks. (a) Markers passing the liberal missing value threshold (<80%, 13 692 markers). (b) Markers passing the strict missing value threshold (<30%, 5913 markers).

of individual series in the training set benefits the prediction ability for all or some test sets. As a starting point for this analysis, we derived the average increase of prediction ability with training set size using an empirical model (Equation 5, Figure 6a). We considered the residuals, which can be interpreted as the performance of the genomic prediction runs, corrected for training set size. We noticed that these deviations were small compared to the effect of the training set size, with a standard deviation of 0.09 for grain yield and heading date, and 0.05 for plant height. This can also be seen visually, with the genomic prediction abilities gathering closely around a common tendency for most test sets (Figure 6a).

We then decomposed the deviations of the genomic prediction runs into contributions of individual series in their training sets (6) to find out whether certain experimental series caused prediction runs to systematically over- or underperform. The influence of the choice of series corrected for training set size was small compared to the importance of increasing the training set size (Figure S2b). The total variance attributable to individual series or their combinations was higher for heading date (0.009) than for plant height and grain yield (0.003). Judging from the relative sizes of the variance components, the interaction effects of specific pairs of experimental series (one being part of the training set, the other being the test set) were the dominating factor. Main effects of individual series being in the training set, representing experimental series that improved or decreased prediction ability for most test sets, played only a minor role (Figure S2b). Only a few individual training set—test set combinations were unusually beneficial or detrimental for prediction ability. The series 1 and 2 showed particularly high compatibility, improving prediction ability by 0.1–0.3 above the values expected for the training set size for all three studied traits (Figure S2a). Apart from this, the combinations with the strongest deviations from the expectation showed no particular pattern (Figure S2a), for example, series 5 → 7 (training set → test set), with a prediction ability of −0.09 compared to the expectation, or series 7 → 8 and vice versa for heading date (+0.07/+0.09).

Assuring environmental diversity improves training set performance

To test strategies that could further improve the performance of training sets for genomic predictions, we chose training sets encompassing defined numbers of environments or years, while keeping the training set sizes constant. The moderate training set sizes (300 and 600 for the first and second approach, respectively) resulted from restrictions from the unbalanced nature of the data. Following the first strategy, training sets backed by higher numbers of environments tended to yield better prediction accuracies (Figure 7a). However, compared to a training set of equal size sampled from the full data set without the environment number restrictions, these selections were not or only marginally better. For heading date and plant height, training sets backed by at least six environments per genotype were sufficient to reach the prediction ability of the random set. For grain yield, the prediction ability of the random set was reached with at least four environments per genotype.

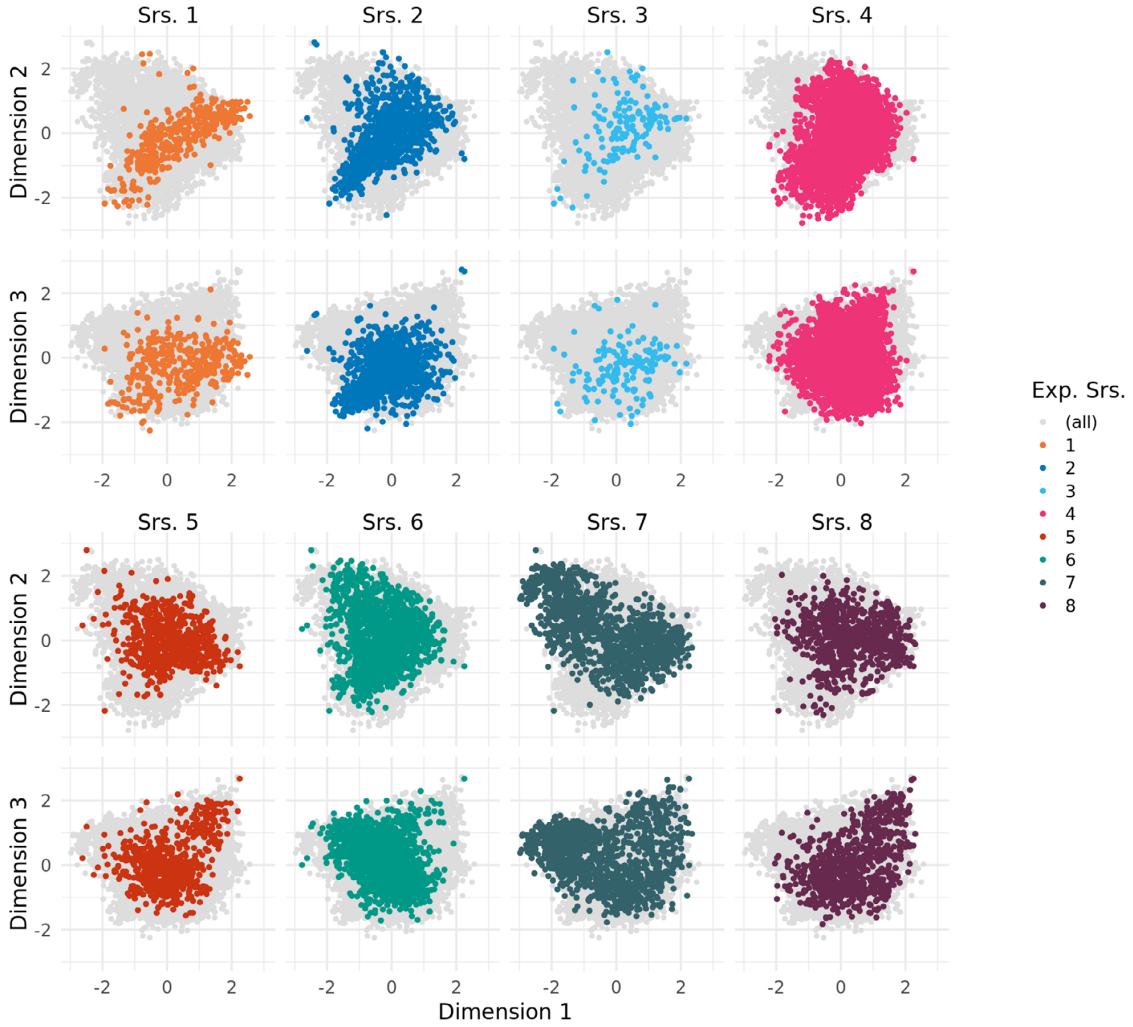

Figure 4 First three principal coordinates of the experimental series (Srs.), based on the Rogers' distances between lines.

A higher number of years in the training set was associated with a better prediction ability. For all years in the available data, including five preceding years in the training set proved superior to including 2 years in the training set at a constant training set size of 600 genotypes (Figure 7b).

Discussion

Sufficient interoperability to ensure successful integration of genotypic and phenotypic data

Previously, it has been shown that genotypic and phenotypic data from different public-private partnership projects can be successfully integrated, doubling the prediction ability of hybrid performance in wheat (Zhao *et al.*, 2021). In this study, we extended these efforts to data from inbred lines generated in four commercial breeding programs (series 5–8) and another published data set (series 1, Gogna *et al.*, 2022). The first question to address is whether the new genotypic and phenotypic data are interoperable, which could be hampered either by differences of employed methods and protocols or by biological reasons.

The genotyping platforms employed are a significant factor influencing the interoperability of genomic data (Gogna *et al.*, 2022; Schulthess *et al.*, 2022a). In this study, all experimental series utilized SNP arrays, specifically the Illumina 90 k iSelect array (Wang *et al.*, 2014). However, other genotyping platforms used in wheat research may have limited or partial overlap (Sun *et al.*, 2020) and these platforms may be more suitable for characterizing populations with smaller linkage disequilibrium or distinct genetic substructures. As a result, ensuring interoperability of genotyping data across platforms may require additional efforts. Furthermore, the availability of marker sequences is critical for imputation of genomic data in our approach, although alternative methods exist that do not rely on marker sequences, albeit with a trade-off in accuracy (He *et al.*, 2015). Genotyping-by-Sequencing (GBS) technologies, which were not employed in this study, generate high-density genotypic data and exhibit random patterns of missing values. These gaps can be imputed with higher accuracy compared to systematic gaps arising from the use of different marker platforms, as demonstrated in this and previous studies (He

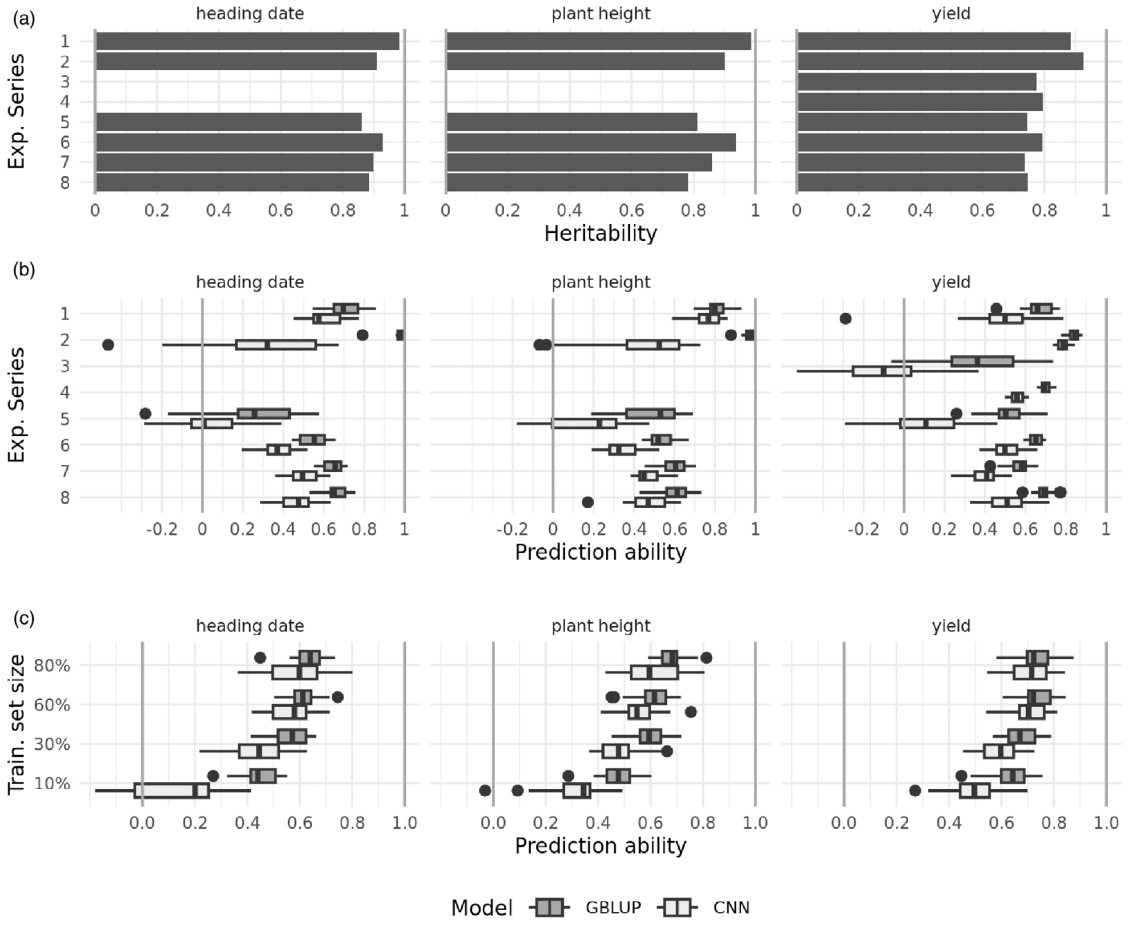

Figure 5 (a) Heritabilities of the individual experimental series. (b) Prediction abilities within each experimental series, 20 cross-validation replications. (c) Prediction abilities across experimental series, 20 cross-validation replications. Training set sizes are percentages of the full number of available genotypes for the respective trait (see Table 2). Cross-validations in subfigures b and c were performed with both GBLUP and a CNN (boxplot colours).

Table 1 Numbers of genotypes (gt), environments, that is, location times year combinations (env), per-genotype average (avg) and total (tot), and number of tested years (y) by the experimental series. The second column shows the calendar year ranges. For series that contain both hybrids (h.) and inbred lines (i.) those two groups are shown on distinct rows

		Heading date				Plant height				Grain yield			
Series	Years (20)	gt	env avg	env tot	y	gt	env avg	env tot	y	gt	env avg	env tot	y
1	09–10	380	8.0	8	2	380	8.0	8	2	380	8.0	8	2
2 (h.)	16–19	3639	17.6	38	4	3639	18.3	39	4	5051	10.1	61	4
2 (i.)	16–19	469	18.5	38	4	469	19.6	39	4	1099	9.6	61	4
3 (h.)	12–13									1604	11.0	11	2
3 (i.)	12–13									144	11.0	11	2
4	12–15									4958	3.9	30	4
5	20–21	781	1.3	7	1	1001	2.1	8	1	1911	4.4	26	2
6	20–21	1631	5.3	13	2	1631	5.6	15	2	1631	5.7	17	2
7	20–21	1707	4.1	9	2	1742	4.0	10	2	1742	4.4	12	2
8	20–21	3516	1.4	7	2	3512	1.6	7	2	3505	2.1	15	2
(all)	09–21	12 096	8.1	81	8	12 347	8.3	83	8	21 891	6.2	168	12

Table 2 Effective population sizes (N_e) and numbers of genotypes (inbred lines only) available for prediction with both phenotypic and SNP data available for each experimental series

		Number of genotypes			
Series	N_e	Heading date	Plant height	Grain yield	(all)
1	34.5	371	371	371	371
2	51.6	467	467	1081	1081
3	49.1	0	0	142	142
4	58.0	0	0	3703	3703
5	61.9	214	418	641	641
6	55.0	1614	1614	1614	1614
7	29.7	1178	1213	1213	1213
8	40.7	848	848	848	848
(all)	79.4	4665	4904	9480	9480

et al., 2015; Torkamaneh and Belzile, 2015). Therefore, we are optimistic about the integration of such data sources. When working with diverse landrace populations, where linkage disequilibrium is often reduced and/or genetically distinct subpopulations are present (Schulthess *et al*., 2022b), dedicated analytical approaches may be necessary to ensure the interoperability of genotyping data.

Interoperability of phenotypic data also represents challenges for data integration. There needs to be agreement on the definition of the recorded traits. In our case, we benefited from the fact that all data were focused on Germany, and thus the methods used in the German official variety tests (Test of Value for Cultivation and Use) served as a *de facto* standard of trait evaluation. Similarly, for the same reason, farming regimes in our study were similar, using intensive management practices to assess grain yield potential. To go beyond this scope, common method documentation standards need to be agreed, used and improved, balancing the fine line between covering many use cases and being understood and correctly employed by practitioners (Darnala *et al*., 2023; Papoutsoglou *et al*., 2020; Selby *et al*., 2019). Even then, objective phenotyping can be a challenging task in a commercial breeding program considering the trial network size, speed of plant development and restricted staff size.

In order to enable interoperability between series, common genotypes are essential to estimate environmental effects. The experimental series of this study were mostly connected by more than five genotypes, which is a standard also used in commercial plant breeding. Some combinations of series fall below this standard (Table S1), particularly for the historical series with the traits heading date and plant height. Fortunately, the connectivity of data sets can also be transitory, so that one experimental series which is well connected to two other series can serve as a common reference, relating them to each other even when they are mutually only weakly connected. Declaring genotypes with a Rogers' distance below a certain threshold as equal (Zhao *et al*., 2021) can further increase these numbers slightly (by on average three individuals in our case) for pairwise comparisons.

Potentials and limitations of genome-wide prediction across series

For predicting hybrid wheat performance, it was noticed that across-population prediction profits from combining multiple different experimental series into one training set (Zhao *et al*., 2021): Prediction abilities of up to 0.4 were achieved. Here, we focused on inbred lines and drastically increased the number of available genotypes by cooperating with commercial breeders. We confirmed also for inbred lines that cross-population prediction profits from combining multiple different experimental series into one training set and were positively surprised by the flexibility with which the different experimental series could be combined. The increases in prediction ability were largely determined by the size of the training set (Figure 6a), but which series were included in the training set was less important (Figure S2b). An important factor that contributes to this is the weak population structure in our joint population (Figure 4). Other studies, in particular those focusing on gene bank material like landraces, have found populations that form more distinct clusters (Crossa *et al*., 2016; Ramstein and Casler, 2019; Schulthess *et al*., 2022a). As the relatedness between the genotypes in the test and the training set drops, the prediction ability of GBLUP drops as well (Alemu *et al*., 2023; Habier *et al*., 2010; Lorenz and Smith, 2015). Therefore, when populations are found to diverge in their genetic diversity, the power of across-population predictions will decrease. However, Central European elite wheat breeding pools seem to share large parts of genetic diversity. We conclude that in such populations, combining data from multiple breeding programs is a very promising strategy to improve genomic prediction ability and shows a way for small to medium-size breeding programs to achieve shared benefits through cooperation. Experimental series 2 partly contains genetic resources from the IPK gene bank, but the PCoA (Figure 4) does not show a markedly larger genetic space covered by this series than by the other. Reasons could be that this diversity, weighted by the number of genotypes, is less prominent compared to the distances between the individual data sets and thus is not shown by the PCoA. Another reason could be the marker filtering before the Roger's distance was computed, which might have removed rare alleles that contributed to that population's diversity. However, within- and across-prediction abilities are not lower for this series (Figures 5b and 6a), indicating that there is no strong effect of the removed markers on the measured traits. Traits like disease resistances, which can show higher dependence on individual loci, would require a more careful approach in this respect.

Opportunities to further boost the prediction abilities

Expanding the training set beyond 4000 lines yielded diminishing returns in prediction ability, with performance approaching a plateau. However, it is worth noting that this plateau was reached well below the heritability of the test set, which represents the theoretical maximum (Figure 6a). The reasons for this are not clear yet and warrant further studies. First, when moving from within-series to across-series prediction, one loses the additional prediction ability which GBLUP confers by capturing co-segregation when training and test genotypes are only a few generations apart (Habier *et al*., 2013). Another possible cause could be that the increased diversity of the joint data weakens the linkage between causative loci and the loci in the SNP data (Meuwissen, 2009). In a previous study, in order to quantify this effect in an applied setting, subsets of different nominal and effective population sizes were drawn from experimental series 4 and cross-validated prediction accuracies, that is, prediction abilities divided by the square root of heritability, were obtained (Zhao *et al*., 2021). Extrapolating this

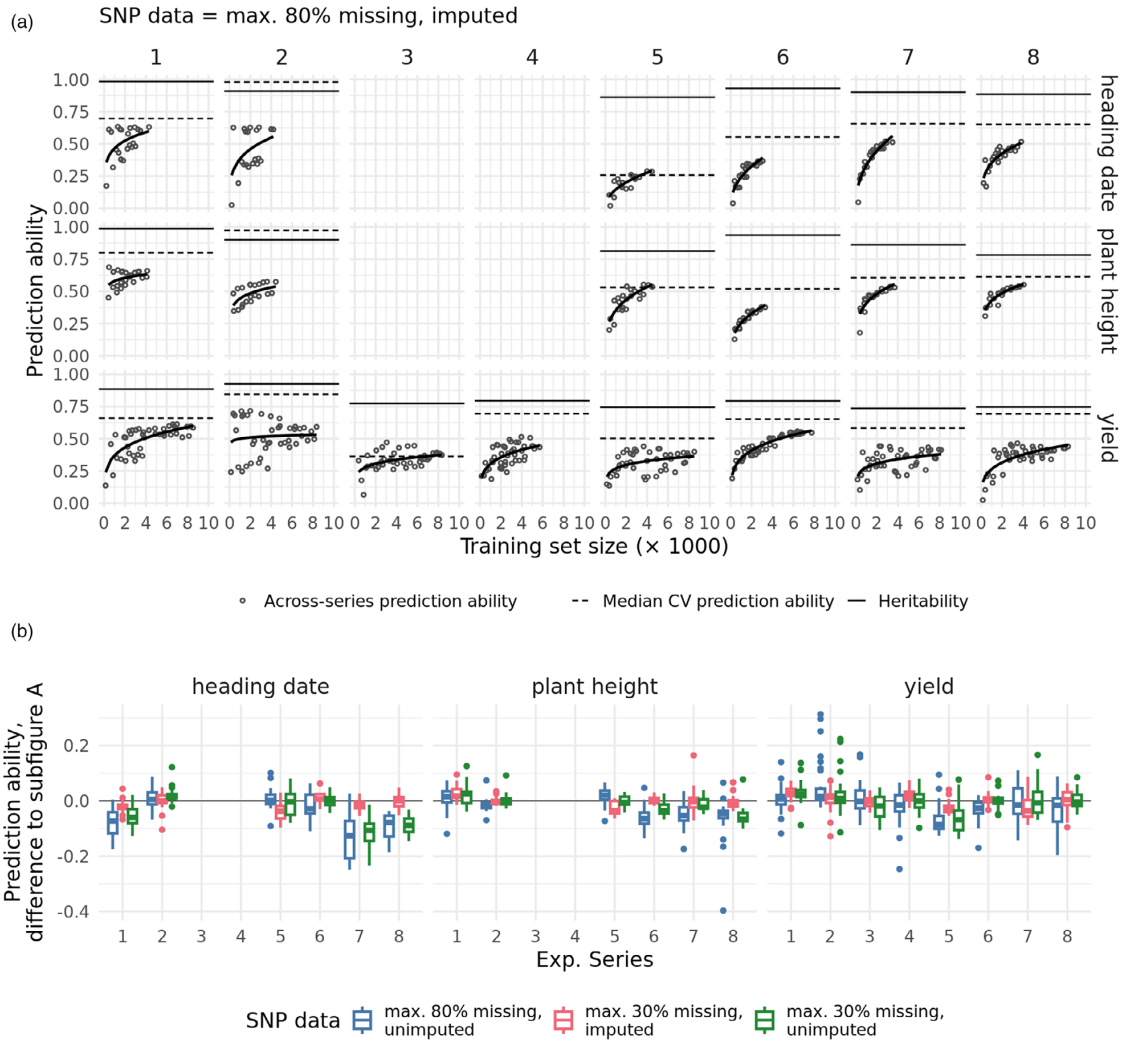

Figure 6 Across-experimental-series prediction abilities. (a) Prediction abilities using the large and imputed marker set (see 'Methods') and different combinations of experimental series as training sets. The training sets are ordered by their size (number of training genotypes) on the horizontal axis and by their prediction ability on the vertical axis. Each plot shows one experimental series as the test set and one trait. Heritability and median cross-validation (CV) prediction ability (see Figure 5a,b) are annotated as horizontal lines for reference. The empirical model fit (Equation 5) is denoted by a solid line. (b) Difference in prediction ability when using any other genomic data than the large and imputed data. Each boxplot summarizes genomic prediction runs of the same test set and different series combinations as the training set. The colours show the results for different missing value thresholds and imputation of the SNP data (colours). The difference to the prediction abilities of subfigure a is shown.

relationship to the nominal and effective population sizes that were seen in our study for the joint population (Table 2), an expected prediction accuracy of about 0.75 would result. This would account for a large part of the missing prediction accuracies of the training sets in this study (Figure S3). According to this theory, increasing the ratio of the nominal to the effective population size should improve prediction ability. This could be done by selecting subsets for the full training data that are highly related to the test population. However, prior attempts to do this have rarely succeeded in surpassing the prediction ability of GBLUP using the full training data and mostly focused on achieving comparable power with a smaller training set size (Fernández-González *et al.*, 2024; Isidro y. Sánchez and Akdemir, 2021; Lopez-Cruz and de los Campos, 2021).

A further avenue might be to test alternative genomic prediction methods; for example, BayesB has been reported to perform slightly better than GBLUP in estimating marker effects based on LD to causative loci (Habier *et al.*, 2010). In recent times, attempts to use non-linear machine learning methods, such as Random Forest, Support Vector Machine or Neural Networks to predict genotype performance, have reached parity with GBLUP and sometimes even achieve better results (Abdollahi-Arpanahi *et al.*, 2020; Montesinos-López *et al.*, 2024; Sandhu *et al.*, 2021). To connect to these findings, we have employed a Convolutional

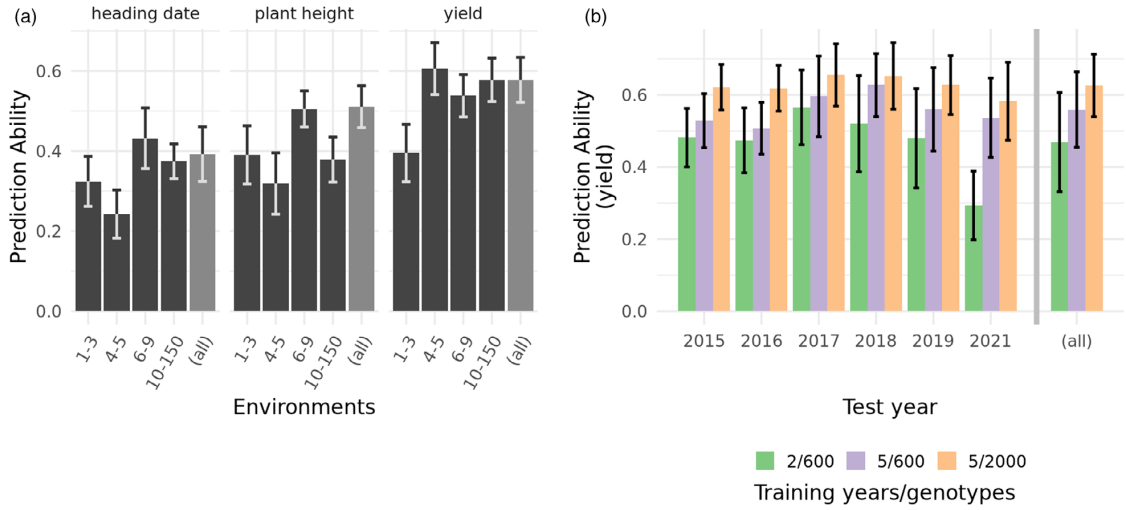

Figure 7 Prediction abilities for yield of across-series training sets sharing breeding-related characteristics. (a) Training sets of 300 genotypes each, formed of genotypes that were phenotyped in similar numbers of environments. For comparison, an equal number of training genotypes is drawn from all experimental series ('(all)'). The test set was 100 genotypes, sampled to equal fractions from all experimental series. Error bars show the standard deviation of all replications (25). (b) Effect of using recent versus historical data for prediction. The test set was 100 genotypes drawn from a single year. Training sets were drawn from 2 to 5 years before the test set year. Additionally, results of a test set from the preceding 5 years and a higher number of genotypes are shown. Error bars show the standard deviation across the replications (50 per test set year). The label '(all)' shows the mean and standard deviation of the pooled runs.

Neural Network (CNN) for genomic prediction and compared the results to the GBLUP. We found that for our task, the CNN was able to predict phenotypes equally well as the GBLUP, where GBLUP provides high prediction abilities. Subpar GBLUP prediction accuracies were associated with CNN performances that were not only weaker in absolute terms but also showed a wider performance gap to GBLUP. It seems that some factors that are detrimental for GBLUP affect CNN disproportionately stronger. The most likely factor is the training set size. For the across-series predictions with small training sets (Figure 5c) and within-series predictions of small experimental series (Figure 5b), poor CNN performance is to be expected. In order to fit a CNN, a high number of parameters and hyper-parameters have to be estimated. On the contrary, for GBLUP, only a mean and a variance parameter have to be estimated, and predictions can then be derived by means of quantitative genetic theory. Therefore, small training sets are expected to be insufficient for an adequate training of a CNN, and consequently, GBLUP remains the preferred model for training set sizes smaller than about 4.000 genotypes. An interesting exception to this trend is experimental series 1, where both the CNN and the GBLUP attain high prediction accuracies despite the small training set size.

The observed comparable prediction performances of GBLUP and CNNs is consistent with examples from the literature (Montesinos-López *et al*., 2024; Sandhu *et al*., 2021). However, those studies used only a fraction of the genotypes that were available for this study. Therefore, it stands to reason that there might be more factors than sample size that inhibit an even better prediction ability and eventual superiority of neural network-based predictions over linear methods like GBLUP. Interestingly, there could be a counterpart for neural networks to the above-mentioned hypothesis that the diminishing role of additive-genetic relationships in multi-origin data sets might hamper GBLUP predictions. The neural networks could strongly base its predictions on additive-genetic relationship, neglecting the effects and interactions of individual alleles. This phenomenon of 'shortcut learning' was observed before for genomic prediction with neural networks (Ubbens *et al*., 2021). Neural networks have many properties that hold great promise for genomic prediction, like a linear time complexity with increasing sample size, their flexibility to incorporate many different data types like environmental covariates (Washburn *et al*., 2021), and their flexibility in cases where the additive genetic model falls short of describing the genetic architecture (Pérez-Enciso and Zingaretti, 2019). The more interoperable training data can be made available, possibly by cooperation across institutional borders, the more of this potential can be realized. In this study, the IPK has served as an 'academic data trustee'. By providing a neutral and confidential data deposit, commercial stakeholders could contribute data. The results suggest that such a model could facilitate innovation in breeding.

Besides the method, one could consider a more detailed genomic picture by increasing the marker density. In theory, the number of required markers for high genomic prediction abilities rises with the effective population size of the data set because the size of linkage blocks decreases (Meuwissen, 2009). However, the benefits of increasing marker density have already been found to diminish beyond about 5000 markers (Zhao *et al*., 2021), using experimental series 2–4 of this study. As the effective population size of this study remains at the same level, it seems unlikely that using more than the 13 000 markers used in this study would hold much potential for improvement. A limiting factor when comparing the effective population sizes of this study and that of Zhao *et al*. (2021) is that as hybrids are not used for genomic prediction in this study, the number of used markers differs and imputed genomic data plays a larger role in our study. Another

approach to better model the correlation of genetic values for GBLUP would be to create trait-specific correlation matrices. Employing the genomic relationship matrix to this end is a simplification which assumes that effective loci are very large in number and pervade the whole genome evenly. Instead, genetic correlation matrices could be computed, for example, from results of Genome-Wide Association Studies (GWAS). To define the genetic correlation, only markers would be considered that are associated with the trait of interest, or by defining synthetic markers that are associated with specific haploblocks (Jiang *et al.*, 2018; Weber *et al.*, 2023).

Besides considering whether the training set data have sufficient power to capture the genetic architecture of the trait, one must also consider whether the genotypic values themselves are measured accurately enough. In particular, genotype-times-environment interactions could limit the prediction ability in this study. Where many environments are sampled, the proportion of genotype-times-environment interactions in the phenotypic variance of grain yield has been found to be large, sometimes even larger than the additive genetic effect (De Los Campos *et al.*, 2020; Jarquín *et al.*, 2014; Lado *et al.*, 2016). When fewer environments are sampled, the impact of genotype-times-environment interactions is estimated to be less important (Montesinos-López *et al.*, 2024). Continuing this trend, where only one environment is sampled, and thus the experimental design confounds genotype-times-environment interactions with the genotypic main effect, GBLUP can achieve grain yield prediction accuracies of up to one (Norman *et al.*, 2018). Interestingly, a larger number of sampled environments does not necessarily require a larger number of plots to be tested. The overlap between environments could also be improved by performing preliminary grain yield tests using more environments of a testing network, sparsely distributing candidates and thus keeping the overall number of plots constant. This has been shown to be advantageous in previous studies for hybrid series 2 and 3 (Lell *et al.*, 2021; Zhao *et al.*, 2021), as well as for biparental populations of barley and maize (Endelman *et al.*, 2014). In the unbalanced breeding trials of this study, the largest share of the data is made up of early-stage candidate varieties which are tested in less than five environments, and almost no series test genotypes in more than 10 environments per year (Table 1). Moreover, the number of years per tested genotype is small. Therefore, one can expect that the genotype-times-environment interactions are partially confounded with the genetic main effects in the across-environment BLUEs. This assumption is supported by our observation that given an equal number of training genotypes, increasing the number of years in the training set improves the prediction results (Figure 7b). The across-series predictions in this study use BLUEs of a single series in the test set. Therefore, it is conceivable that the across-series genomic prediction will yield breeding value estimates that are in fact less affected by confounding genotype-times-environment interactions than the single-series BLUEs to which they are compared. The confounding genotype-times-environment interactions in the test set BLUEs would therefore decrease the prediction ability. If this effect proves to be relevant, across-series predictions would also be of value to breeders whose data sets exceed the aforementioned threshold of a training set size of about 4000 genotypes. Candidate varieties are tested in an extensive set of environments during official registration procedures, which exceeds the number of environments of the final breeding stage. Therefore, benchmarking genomic prediction methods using the final-stage genotypes of one's own breeding program may give overly optimistic results due to residuals confounding genotype–times–environment interactions. Obtaining phenotypic and genomic data from official variety tests would therefore be a gold standard for assessing the impact of cross-series predictions on prediction ability. This endeavour is easily hampered by legal issues, but seems worthwhile.

Methods

Plant material and field trials

Eight different experimental series were assembled for this study. The first four of them consist of historical data that have been studied previously. In addition, four more were contributed by breeding companies for this study. Within the experimental series, the number of genotypes ranges from 380 to 5051 and the number of environments (location – year combinations) from 8 to 61 (Table 1). The series are characterized in the following:

- Series 1: Orthogonal grain yield trials of 380 genotypes representing a broad diversity of the European elite breeding pool (Gogna *et al.*, 2022).
- Series 2: Non-orthogonal grain yield trials of 1099 diverse inbred lines including elite lines and plant genetic resources along with 5051 hybrids (Zhao *et al.*, 2021, Experimental series II to V).
- Series 3: Orthogonal grain yield trials of 135 elite parent lines, their 1604 hybrids and 10 released varieties (Zhao *et al.*, 2021, Experimental Series I, and Zhao *et al.*, 2015).
- Series 4: Non-orthogonal grain yield trials, generated in the course of a commercial inbred line breeding program, comprising 4958 genotypes (Zhao *et al.*, 2021, Experimental Series VI).
- Series 5–8: Four experimental series, provided by four breeding companies. The series consisted of 781–3516 genotypes tested in an average of 1.3–6.7 locations in 2 years (Table 1). All four series were excerpts from the companies' usual breeding activities and were therefore non-orthogonal grain yield trials in which lines were phenotyped and selected in up to 3 years. There are three sub-trials per series, reflecting the breeding trial stages. The candidates were evaluated for grain yield under intensive treatment and partially for heading date and plant height. Connectivity between data from each series was ensured by several common genotypes (Table S1).

Phenotypic data analysis

We applied a two-step approach to analyse the phenotypic data as described previously (Lell *et al.*, 2024) including outlier correction following the method M5 from Bernal-Vasquez *et al.* (2016). See Figure 1 for an overview. Best linear unbiased estimator of the genotypes (BLUEs) within and across the environments was obtained after outlier correction. The within-environment BLUEs were generated using the following mixed linear model:

$$y_{ijkl} = \mu + g_i + t_j + r_{jk} + b_{jkl} + \varepsilon_{ijkl}, \quad (1)$$

where y is the plot-level grain yield data, μ is the overall mean, g is the genotypic effects, t, r and b are design effects for trials, replications and blocks, ε are residual effects, and i, j, k and l are indices for model effects. Depending on the experimental design, only some or none of the effects t, r or b were estimated. When

phenotyping was done in an unreplicated trial in an environment, design effects were estimated and subtracted from the measurements. The design effects were estimated as random effects, g was estimated as a random effect to calculate BLUPs and estimate the genetic variance and as a fixed effect to obtain BLUEs. All random effects were identically and independently distributed. The repeatability was estimated as $r^2 = \sigma_g^2 / \left(\sigma_g^2 + \sigma_\varepsilon^2 / n\right)$, where σ_g^2 is the estimated genotypic variance, σ_ε^2 is the estimated residual variance, and n is the average number of plots per genotype. Environments that have a repeatability of <0.3, or that show <0.1 correlation to all other environments on average were discarded.

The across-environment BLUEs were obtained using the following mixed model:

$$y_{ijk} = \mu + s_i + g_j + e_k + \varepsilon_{ijk}, \quad (2)$$

where y are the within-environment genotype means, μ is the overall mean, s are the experimental series effects, g are the genotypic effects, e are the environment (location × year) effects, ε are the residual effects, and i, j, and k are indices for model effects. The effects s are estimated as fixed effects, e are random effects, and g is fixed for BLUEs and random for BLUPs. The broad-sense heritability was estimated as $H^2 = \sigma_g^2 / \left(\sigma_g^2 + \sigma_\varepsilon^2 / n\right)$, where n is the average number of environments a genotype is measured in, and σ_g^2 and σ_ε^2 are the estimated genetic and residual variances, respectively. All computations were performed using R 4.0.4 (R Core Team, 2021) linked to OpenBLAS 0.2.20 (https://www.openblas.net) and AsReml-R 4.1.0.110 (Gilmour *et al.*, 2015) on a Linux machine with 4 Intel Xeon CPU E7-4890 v2 processors (120 logical cores) and required about 500 GB of RAM.

Genotypic data analyses

All experimental series had SNP data available for varying fractions of genotypes (Table 2). The SNP data were generated according to each data provider's own processes and thus a heterogeneous selection of SNP markers was available for the different experimental series (Figure 2a). The SNP calls were filtered individually for each genotyping batch according to the following thresholds: Genotypes were discarded if they had more than 30% missing values or more than 10% heterozygous calls. Markers were discarded if they had more than 90% missing values or more than 20% heterozygous calls. As the SNP data were generated by different providers, we further checked and corrected for consistency of strand designation using overlapping genotypes before integrating data across experimental series. The computational efficiency of this process was improved by modelling SNP array overlaps as graphs and determining maximum spanning trees (Markowski *et al.*, 2021). Marker positions were derived using BLAST (Camacho *et al.*, 2009) and the Chinese Spring Reference Sequence v.2.1 (Zhu *et al.*, 2021), removing markers whose sequences showed mismatches to the reference sequence or mapped to more than one position. Of the markers that showed variation for the population, a unique physical position was found for 29 970 markers and those marker's data were subsequently imputed. Imputation was performed using BEAGLE 5.2 (Browning *et al.*, 2018) without the use of a reference panel for phasing information, with a window size of 1000, an overlap size of 100, and 10 burn-in iterations.

As the SNP data are heterogeneous in terms of marker density, there are non-random gaps in the SNP data, so that large numbers of markers are available for only a fraction of individuals. To assess the accuracy of imputing those systematic gaps, we chose a blocked masking approach: We divided the available markers into a low-density set and a high-density set by the fraction of genotypes that had data for the respective marker. Markers for which more than 70% of genotypes had data were grouped into a low-density set; all other markers formed the high-density set (low-density set: 2582 markers, high-density set: 13604 markers). For a random 10% of genotypes that had data for more than half of the high-density markers, the high-density marker data were masked. The resulting data set was imputed using the same BEAGLE parameters as described previously. For each marker, the imputation accuracy was calculated as the sum of masked SNP calls whose imputed data matched the original calls divided by the total number of masked SNP calls. This process was repeated 20 times. In an additional experiment, we employed a random masking approach to mimic missing calls as occurring, for example, in genotyping-by-sequencing approaches. We masked 1% of all marker calls randomly throughout the whole data set, imputed them, and calculated the imputation accuracy per marker as described above. This process was also repeated 20 times.

To visualize the population structure, we calculated the Rogers' distance (Rogers, 1972) using the SNP data set before imputation (Figure 2a). We filtered the data for a maximum missing value rate of 80% per marker and a minimum minor allele frequency of 0.05. This retained 13 720 markers of the initial data. Rogers' distances were computed for each pair of genotypes based on markers that were available for both. The Rogers' distances were then subjected to principal coordinate analysis using the R function 'cmdscale'.

Genomic predictions within and across experimental series

Within each experimental series, the correspondence of phenotypic and genotypic data was assessed. For this, a random sample of 90% of genotype means within each series were taken as training set to predict the remaining 10%. This process was repeated 20 times for each experimental series. To obtain the genotype means within each series, model (2) was fitted on the within-environment BLUEs without the series effect s. The resulting genotypic effects (within-series BLUEs) were then used as measurements in a genomic BLUP (GBLUP) model:

$$y_i = \mu + g_i + \varepsilon_i, \quad (3)$$

where y were the within-series BLUEs, μ is the mean, ε are the residual effects, and g are the genotypic effects that were modelled to be correlated using the VanRaden (2008) genomic relationship matrix K, so $Cov(g) = K\sigma_g^2$, where the variance σ_g^2 was estimated by the model. The SNP data used to compute K were filtered to include only markers that have SNP calls for at least 80% of that series' genotypes and in addition a minor allele frequency of at least 0.05. The prediction was performed using the BGLR R package (Pérez and de los Campos, 2014).

We created many different training sets for genomic prediction by choosing different combinations of experimental series as training sets. From the large number of possible combinations, we chose a subset which was numerically optimized for D-optimal experimental design using Fedorov's algorithm (Wheeler, 2004),

with 135–349 different series combinations as training sets, depending on maximum training set size for the trait. We then determined the prediction ability of experimental series that were not included in the respective training sets.

To test the influence of quality-control steps during data integration, we performed the predictions with four different versions of the SNP data, which influenced the genomic relationship matrix K: Markers were filtered using either a liberal threshold of at most 80% missing values per marker or a strict threshold of at most 30% missing values per marker. The data resulting from the liberal threshold had 13 692 markers. The data resulting from the strict threshold had 5913 markers. Furthermore, we used either the SNP data with or without imputation by BEAGLE. In the unimputed case, gaps in the marker data were filled by the respective marker means as described above. For the VanRaden distance, which involves centring the SNP data, this is equivalent to basing the pairwise distances only on markers for which there is information in both involved individuals. The resulting kinship matrices are used as covariance matrices for a genomic BLUP using the following model:

$$y_i = \mu + g_i + \varepsilon_i, \tag{4}$$

where y, μ, g and ε were defined analogous to (3). Like for the within-series cross-validations, the computation was performed using the BGLR R package.

Subsequently, we focused on a more detailed analysis of the influence of individual series on the prediction ability. We based this on the prediction runs with the imputed SNP data with the liberal missing value threshold. For each test set i ($i \in \{1, \ldots, d\}$, where $d = 8$, the number of series in this study), a model

$$y_{ip} = \alpha_i x_{ip}^{(1/\beta_i)} + r_{ip} \tag{5}$$

was fitted to approximate an average increase in prediction ability with training set size, where y_{ip} was the prediction ability of the genomic prediction run of test set i that used the set of experimental series p as training set. x_{ip} was the number of genotypes in its training set, α_i and β_i were empirical parameters, estimated separately for each test set, and r_{ip} was the deviation of the prediction run (i, p) from the empirical average. Most of the genomic prediction runs had more than one series as training set, which is why i is an index and p is a set of indices.

We studied how individual experimental series being in the training set influence the prediction ability of a particular test set by fitting a linear mixed model on the vector r of deviations from the empirical fit (5) for all training and test sets. In the following, we denoted the deviation of a single genomic prediction run as r_{ip}, indexed by the experimental series as test set i and the set p of experimental series indices in the training set. Based on this, the linear model was

$$r_{ip} = \mu + \eta_i + \sum_j^d \delta_{j \in p}\left(\theta_j + \kappa_{ij}\right) + \varepsilon_{ip} \quad \text{with } \delta_x = (1 \text{ if } x \text{ is true, else } 0). \tag{6}$$

The model decomposes the deviations of the genomic prediction runs into a mean μ and three groups of random effects: The main effects of the test sets (η) and training sets (θ) with d effects each, and the effect of each combination effects κ of two experimental series as test set i and training set j. For each of η, θ and κ, one variance parameter is estimated. As one run has more than one training set, the term $\delta_{j \in p}$ selects for each measurement r_{ip} the relevant parameters. The residuals are denoted as ε. The model was fitted using the BGLR R package as Bayesian Ridge Regression (model setting 'BRR').

Convolutional neural network for genomic prediction

To compare with GBLUP, we used a Convolutional Neural Network (CNN) to assess the prediction ability of both within-series BLUEs (previous section) and across-series BLUEs. The Python framework 'Keras' was used for model development. The CNN operated on a one-dimensional sequence of marker states, ordered by their mapping position in the genome. The two homozygous states were coded as 0 and 2, and the heterozygous state as 1. Both phenotypic and marker state vectors were rescaled to a range of [0, 1].

For the training set, we selected random samples of 10%, 30%, 60% and 80% of the total data available for each trait. The test set consisted of 100 genotypes that were not in the training set, randomly chosen for each iteration. The process was repeated for 20 runs for each trait and training set size. The CNN was designed to allow a flexible network architecture for each run, with neurons as edges in an acyclic directed graph, organized into several layers to capture linkage and haplotype structure, and genetic interactions, respectively.

The first set of layers focused on feature extraction, followed by layers for pattern recognition. The specific network structure was influenced by a variety of hyperparameters, whose values were optimized using the Hyperband Tuner (Li *et al*., 2017). The hyperparameter space for the feature extraction section included (1) number of alternating convolution and average pooling layers (three to five), (2) number of filters (ranging from 64 to 512 with a step size of 64) and (3) kernel size (between three and 36 with a step size of three) in the convolution layer. The pool size for the pooling layer was also varied between two and 32 with a step size of 4. The feature extraction output was flattened and relayed into the pattern recognition section.

The hyperparameter space for the pattern recognition section comprised (1) the number of alternating dense and dropout layers (one to four), (2) the number of dense layer units (between 32 and 256 with a step size of 32) and (3) a dropout rate (between 0.1 and 0.5 with a step size of 0.01). Dropout was applied to reduce overfitting by preventing the model from becoming too reliant on specific neurons. The Rectified Linear Unit (ReLU) function was used as the activation function for all layers except the final prediction neuron. A single neuron with a hyperbolic tangent (tanh) activation function performed trait prediction, receiving input from all neurons in the final pattern recognition layer.

For each set of hyperparameters chosen by the tuner, the model was fitted with a batch size of 32 genotypes. The training set was split into 90% to be used for this purpose and 10% that were used exclusively to compare the prediction performances of the trained models resulting from the hyperparameter choices (validation set). The goal of the tuner was to minimize the mean squared error between predicted and observed genotype means in the validation set. Hyperparameter tuning stopped when the error did not decrease by more than 0.01 over 5 or more iterations. The resulting model was used for genomic prediction.

For comparison with GBLUP, the same across-series data as used for the Neural Network was subjected to GBLUP following Equation (4).

Influence of number of experimental series, environments and years in training set

We measured the effect of including a single versus multiple experimental series into the training set, at a constant number of 800 genotypes in the training set and 100 genotypes in the test set. Both sets were chosen randomly (25 replications). In the first scenario, we choose all training genotypes from one random experimental series only. Only the experimental series 6, 7 and 8 for heading date and plant height and series 2, 4, 6, 7 and 8 for grain yield were large enough for this. In the second scenario, we sampled the same number of genotypes randomly from all but one series. The test set was sampled from the series that were not in the training set.

We further investigated potential strategies to subset genotypes based on data quality or number of phenotyping environments to improve the predictions obtained from a large integrated data set by restricting the training set to fulfil different criteria. To test the influence on prediction ability of the number of environments a training set is based on, we assigned all genotypes to one of four environment groups: 1–3, 4–5, 6–9 and 10 or more environments. These were chosen based on the available data to have groups as even in size as possible while still including genotypes from multiple experimental series into one environment group. The training set consisted of 300 genotypes randomly chosen from the genotypes of an environment group. The test set were 100 genotypes that were randomly drawn while ensuring that genotypes from each experimental series were included at equal shares in the test set. For each environment group, 25 replications of the prediction were performed. For the prediction, the GBLUP model (3) was used, and the kinship matrix was based on the imputed SNP data using the liberal missing value criterion (13 692 markers).

To test whether historic data were still useful for predictions, we selected training sets that had the same number of genotypes but differed in the number of years covered by the training data. We chose 2-year ranges from which we sampled 600 training genotypes for each range: 1–2 years before the test set and 1–5 years before the year of the test set. Our data covered 12 different years, so we could generate test sets for 7 years (year 6 to year 12). Analyses was only done for grain yield because the available data for heading date and plant height was missing for several environments. We also excluded one test set (2020) because the preceding years did not allow for a training set of adequate size. The test set encompassed 100 random genotypes that were measured in the respective year. As mentioned before, we used the same set of pre-computed across-environment BLUEs as data for the predictions, so technically information from more than 1 year was included in some of these BLUEs. However, the majority of genotypes in our data were measured in 1 year only. We made sure that genotypes that were included in the training set were not included in the test set. The prediction was done using the GBLUP model (3) and the kinship matrix was based on the imputed large SNP data.

Acknowledgements

The authors acknowledge funding within the Wheat BigData Project (German Federal Ministry of Food and Agriculture, FKZ2818408B18). The authors thank Alexandre Pinheiro and Uwe Scholz for their contributions to a smooth data curation and transmission. Open Access funding enabled and organized by Projekt DEAL.

Author Contributions

M.L., Y.Z. and J.C.R. conceived and designed the study. U.A., J.D., W.M.E., T.E., M.G., M.Ki., M.Ko., S.K., N.P., M.R., V.W. and M.W. acquired, curated and contributed data. M.L. curated and processed the data, performed the analyses and analysed the results. Y.Z. supported data analyses. A.G. conceptualized the neural network. M.L., V.K. and A.G. performed neural network-based data analysis. M.L., Y.Z. and J.C.R. interpreted the results and wrote the manuscript, with the help of all coauthors. Since the study was conducted, Valentin Wimmer has moved to Aardevo B.V., Johannes Postweg 8, 8308 PB Nagele, Netherlands, and Martin Kirchhoff has moved to Nordzucker AG, Küchenstraße 9, 38 100 Braunschweig, Germany.

Competing interests

The authors declare that they have no competing interests.

Data availability statement

Data of series 1–4 were presented earlier. Data and data access are documented for experimental series 1 in Gogna *et al*. (2022) and for experimental series 2–4 in Zhao *et al*. (2021). Data of experimental series 5–8 are not available publicly as they are trade secrets of the respective breeding companies. They are archived at the IPK.

References

Abdollahi-Arpanahi, R., Gianola, D. and Peñagaricano, F. (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel. Evol.* **52**, 12.

Alemu, A., Batista, L., Singh, P.K., Ceplitis, A. and Chawade, A. (2023) Haplotype-tagged SNPs improve genomic prediction accuracy for Fusarium head blight resistance and yield-related traits in wheat. *Theor. Appl. Genet.* **136**, 92.

Bernal-Vasquez, A.-M., Utz, H.-F. and Piepho, H.-P. (2016) Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. *Theor. Appl. Genet.* **129**, 787–804.

Beyene, Y., Gowda, M., Pérez-Rodríguez, P., Olsen, M., Robbins, K.R., Burgueño, J., Prasanna, B.M. *et al*. (2021) Application of genomic selection at the early stage of breeding pipeline in tropical maize. *Front. Plant Sci.* **12**, 685488.

Browning, B.L., Zhou, Y. and Browning, S.R. (2018) A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chafai, N., Hayah, I., Houaga, I. and Badaoui, B. (2023) A review of machine learning models applied to genomic prediction in animal breeding. *Front. Genet.* **14**, 1150596.

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., Vikram, P. *et al*. (2016) Genomic prediction of gene bank wheat landraces. *G3 (Bethesda)*, **6**, 1819–1834.

Daetwyler, H.D., Calus, M.P.L., Pong-Wong, R., De Los Campos, G. and Hickey, J.M. (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics*, **193**, 347–365.

Darnala, B., Amardeilh, F., Roussey, C., Todorov, K. and Jonquet, C. (2023) C3PO: a crop planning and production process ontology and knowledge graph. *Front. Artif. Intell.* **6**, 1187090.

De Los Campos, G., Pérez-Rodríguez, P., Bogard, M., Gouache, D. and Crossa, J. (2020) A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. *Nat. Commun.* **11**, 4876.

Endelman, J.B., Atlin, G.N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M.E. and Jannink, J.-L. (2014) Optimal design of preliminary yield trials with genome-wide markers. *Crop Sci.* **54**, 48–59.

Fernández-González, J., Haquin, B., Combes, E., Bernard, K., Allard, A. and Isidro Y Sánchez, J. (2024) Maximizing efficiency in sunflower breeding through historical data optimization. *Plant Methods*, **20**, 42.

Gianola, D. (2013) Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics*, **194**, 573–596.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. and Thompson, R. (2015) *ASReml user guide release 4.1 structural specification*.

Gogna, A., Schulthess, A.W., Röder, M.S., Ganal, M.W. and Reif, J.C. (2022) Gabi wheat a panel of European elite lines as central stock for wheat genetic research. *Sci. Data*, **9**, 1–17.

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P. and Thaller, G. (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* **42**, 5.

Habier, D., Fernando, R.L. and Garrick, D.J. (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics*, **194**, 597–607.

He, S., Zhao, Y., Mette, M.F., Bothe, R., Ebmeyer, E., Sharbel, T.F., Reif, J.C. *et al*. (2015) Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (*Triticum aestivum* L.). *BMC Genomics*, **16**, 1–12.

Isidro y. Sánchez, J. and Akdemir, D. (2021) Training set optimization for sparse phenotyping in genomic selection: a conceptual overview. *Front. Plant Sci.* **12**, 715910.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F. *et al*. (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* **127**, 595–607.

Jiang, Y., Schmidt, R.H. and Reif, J.C. (2018) Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3 (Bethesda)*, **8**, 1687–1699.

Lado, B., Barrios, P.G., Quincke, M., Silva, P. and Gutiérrez, L. (2016) Modeling genotype × environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Sci.* **56**, 2165–2179.

Lell, M., Reif, J. and Zhao, Y. (2021) Optimizing the setup of multienvironmental hybrid wheat yield trials for boosting the selection capability. *Plant Genome*, **14**, e20150.

Lell, M., Zhao, Y. and Reif, J.C. (2024) Leveraging the potential of big genomic and phenotypic data for genome-wide association mapping in wheat. *Crop J.* **12**, 803–813.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. and Talwalkar, A. (2017) Hyperband: a novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **18**, 6765–6816.

Lopez-Cruz, M. and de los Campos, G. (2021) Optimal breeding-value prediction using a sparse selection index. *Genetics*, **218**, iyab030.

Lorenz, A.J. and Smith, K.P. (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Sci.* **55**, 2657–2667.

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

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J. and Ma, C. (2018) A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, **248**, 1307–1318.

Markowski, J., Kempfer, R., Kukalev, A., Irastorza-Azcarate, I., Loof, G., Kehr, B., Pombo, A. *et al*. (2021) GAMIBHEAR: whole-genome haplotype reconstruction from Genome Architecture Mapping data. *Bioinformatics*, **37**, 3128–3135.

Meuwissen, T.H. (2009) Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* **41**, 35.

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

Montesinos-Lopez, O.A., Montesinos-Lopez, J.C., Salazar, E., Barron, J.A., Montesinos-Lopez, A., Buenrostro-Mariscal, R. and Crossa, J. (2021) Application of a Poisson deep neural network model for the prediction of count data in genome-based prediction. *Plant Genome*, **14**, e20118.

Montesinos-López, O.A., Montesinos-López, A. and Crossa, J. (2022) General elements of genomic selection and statistical learning. In *Multivariate statistical machine learning methods for genomic prediction* (van Eeuwijk, F., ed), pp. 1–34. Cham: Springer International Publishing.

Montesinos-López, A., Crespo-Herrera, L., Dreisigacker, S., Gerard, G., Vitale, P., Saint Pierre, C., Govindan, V. *et al*. (2024) Deep learning methods improve genomic prediction of wheat breeding. *Front. Plant Sci.* **15**, 1324090.

Norman, A., Taylor, J., Edwards, J. and Kuchel, H. (2018) Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3 (Bethesda)*, **8**, 2889–2899.

Papoutsoglou, E.A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I.N., Chaves, I., Coppens, F. *et al*. (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytol.* **227**, 260–273.

Pérez, P. and de los Campos, G. (2014) Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, **198**, 483–495.

Pérez-Enciso, M. and Zingaretti, L.M. (2019) A guide on deep learning for complex trait genomic prediction. *Genes*, **10**, 553.

Pook, T., Freudenthal, J., Korte, A. and Simianer, H. (2020) Using local convolutional neural networks for genomic prediction. *Front. Genet.* **11**, 561497.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramstein, G.P. and Casler, M.D. (2019) Extensions of BLUP models for genomic prediction in heterogeneous populations: application in a diverse switchgrass sample. *G3 (Bethesda)*, **9**, 789–805.

Riedelsheimer, C. and Melchinger, A.E. (2013) Optimizing the allocation of resources for genomic selection in one breeding cycle. *Theor. Appl. Genet.* **126**, 2835–2848.

Rogers, J.S. (1972) Measures of genetic similarity and genetic distance. In *Studies in Genetics VII, University of Texas Bulletin* (Wheeler, M.R., ed), pp. 145–153. Austin: University of Texas.

Sandhu, K., Aoun, M., Morris, C. and Carter, A. (2021) Genomic selection for end-use quality and processing traits in soft white winter wheat breeding program with machine and deep learning models. *Biology*, **10**, 689.

Schopp, P., Müller, D., Technow, F. and Melchinger, A.E. (2016) Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium. *Genetics*, **205**, 441–454.

Schulthess, A.W., Kale, S.M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., Jiang, Y. *et al*. (2022a) Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. *Nat. Genet.* **54**, 1544–1552.

Schulthess, A.W., Kale, S.M., Zhao, Y., Gogna, A., Rembe, M., Philipp, N., Liu, F. *et al*. (2022b) Large-scale genotyping and phenotyping of a worldwide winter wheat genebank for its use in pre-breeding. *Sci. Data*, **9**, 1–21.

Selby, P., Abbeloos, R., Backlund, J.E., Basterrechea Salido, M., Bauchet, G., Benites-Alfaro, O.E., Birkett, C. *et al*. (2019) BrAPI—an application programming interface for plant breeding applications. *Bioinformatics*, **35**, 4147–4155.

Sun, C., Dong, Z., Zhao, L., Ren, Y., Zhang, N. and Chen, F. (2020) The Wheat 660K SNP array demonstrates great potential for marker-assisted selection in polyploid wheat. *Plant Biotechnol. J.* **18**, 1354–1360.

Torkamaneh, D. and Belzile, F. (2015) Scanning and filling: ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data. *PLoS One*, **10**, e0131533.

Ubbens, J., Parkin, I., Eynck, C., Stavness, I. and Sharpe, A.G. (2021) Deep neural networks for genomic prediction do not estimate marker effects. *Plant Genome*, **14**, e20147.

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423.

Wang, Q., Tian, F., Pan, Y., Buckler, E.S. and Zhang, Z. (2014) A SUPER powerful method for genome wide association study. *PLoS One*, **9**, e107684.

Washburn, J.D., Cimen, E., Ramstein, G., Reeves, T., O'Briant, P., McLean, G., Cooper, M. *et al*. (2021) Predicting phenotypes from genetic, environment, management, and historical data using CNNs. *Theor. Appl. Genet.* **134**, 3997–4011.

Weber, S.E., Frisch, M., Snowdon, R.J. and Voss-Fels, K.P. (2023) Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. *Front. Plant Sci.* **14**, 1217589.

Wheeler, R.E. (2004) *AlgDesign. The R project for statistical computing*.

Zhao, Y., Li, Z., Liu, G., Jiang, Y., Maurer, H.P., Würschum, T., Mock, H.-P. *et al*. (2015) Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proc. Natl Acad. Sci. USA*, **112**, 15624–15629.

Zhao, Y., Thorwarth, P., Jiang, Y., Philipp, N., Schulthess, A.W., Gils, M., Boeven, P.H.G. *et al*. (2021) Unlocking big data doubled the accuracy in predicting the grain yield in hybrid wheat. *Sci. Adv.* **7**, eabf9106.

Zhu, T., Wang, L., Rimbert, H., Rodriguez, J.C., Deal, K.R., De Oliveira, R., Choulet, F. *et al*. (2021) Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. *Plant J.* **107**, 303–314.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1 Number of overlapping genotypes between experimental series.

Figure S1 Comparing prediction abilities using only one ('single') versus multiple ('multi') experimental series as training set.

Figure S2 Differences in prediction ability attributed to presence of individual experimental series in the training set.

Figure S3 Prediction accuracy (Prediction ability, divided by the square root of the heritability).

3 General Discussion

3.1 Preparing for a Wheat Data Warehouse

This thesis lays the groundwork to incrementally collect and integrate pre-existing wheat breeding data sets into an interconnected data source that allows to reuse those data sets and obtain new knowledge from the combination of multiple data sets. The disparate aims and institutions who have authored the individual data sets require unified yet flexible data structures that can provide a unified frame for a joint analysis. Inspiration on best practices and future extensions are already available. As with statistical methods, the field of animal breeding has pioneered Big Data application in breeding. Practices which are nowadays subsumed under this term, like automatized data collection from routine work for decision support, have been common in dairy farming since at least the 1950s (Newton et al., 2020), despite the used media being punch cards instead of cloud databases at the time. Nowadays, many dairy farms are connected by national or international Herd Testing Organizations that provide data-backed decision aids to farmers and valuable large data sets about cow genomics and phenotypes to researchers. Newton et al. (2020) have surveyed Big Data practices in the Australian dairy industry and note that there is a strong integration of genomic and phenotypic data collection and sharing in daily agricultural practice. Compared to that track record, data integration in many plant breeding programs is a relatively young field. One reason for this could be that in plant breeding, the variety registration process stands like a barrier between breeding and farming activities. Information about candidate varieties is not gathered from production agriculture but from evaluation trials of breeding companies. Consequently, enterprises striving to facilitate data evaluation and data sharing in wheat breeding face a much smaller market, as they have to target wheat breeders and not wheat farmers.

3.1.1 Integrating inconsistent data models and structures by use of a data catalogue and automated integrity checking

While in the Australian dairy sector 90% of participating farms use the same software to analyse their phenotypic and genomic data (Newton et al., 2020), in wheat breeding a diversity of in-house solutions of varying technical sophistication is adopted for trial data. To integrate data coming from multiple plant breeding companies and academic

research projects therefore means to struggle with a large diversity of ways how the individual data sets are organized. This thesis is based on newly generated data (Lell et al., 2025) which required for its curation several measures, both technical and organizational, to ensure both data transfer and project coordination. Therefore, a guide was required that could communicate the correct format data to data providers. Therefore, the data transfer template that was developed in cooperation with the data providers included a data catalogue, that is, descriptions for the fields that were to be filled by data. Setting this explicit standard was beneficial for smooth and unified project communication.

The data catalogue also proved valuable for the next quality control step, an automated data constraint checking mechanism that provided feedback to data providers in short time after the upload was completed. The fact that the data catalogue already contained acceptable value ranges and example values proved useful in developing checks and user messages of the checking mechanism. As the project progressed, the data catalogue turned out to be most useful where it intertwined documentation and software logic, serving as a single source of truth for both humans and automated pipelines. This highlights the benefits of striving for self-documenting data that is designed to be readable by humans and machines alike.

However, checks that ensured internal data consistency required developing a more comprehensive data model of the relevant entities and their relationships (Figure 2). An entity of the data model can be represented by a data table (so that an individual is a row), or a single field of a table (so that an individual is a value). In addition to the data catalogue, the data model defines the relationships between the entities, that is, how entities have properties that reference other entities. For the relationships, a data model defines the cardinalities, that is how many instances are part of a relationship on either side. As an example, the data model used in this thesis specified that a trial is conducted in exactly one environment, however, in the same environment an arbitrary number of trials can be conducted (Figure 2). The data model was laid down in a declarative programming style and encapsulated into an R package. This architecture communicates the intent of the code and allows for future extensions, for example automated generation of an entity-relationship diagram like shown in Figure 2. Using this model, adherence of the data to constraints was ensured. Such constraints are well known from relational databases. As an example, a primary key constraint on a table column enforces a unique and non-missing identifier for an instance (row) of the entity (table). If the data model mandates a relationship to another entity, that relationship

can be modelled by a foreign key constraint in the database. The fact that these constraints were checked automatically on data upload turned out to be instrumental for error correction due to short feedback time for data uploaders. Data could be corrected instantly as the notion of "correct data" could be tested automatically.

After the formal correctness of the data had been ensured, outlier corrections on multiple levels were employed, from the plot to the environment. This was implemented to guard against systematic errors, be they small-scale like typos or large-scale like wrong genotype labelling within an environment. In addition, where plot coordinates were provided, the spatial distribution of measured values on the field was also a valuable check, underlining the importance of transmitting the grid coordinates of plots. Outlier correction always carries the risk of erroneously discarding interesting

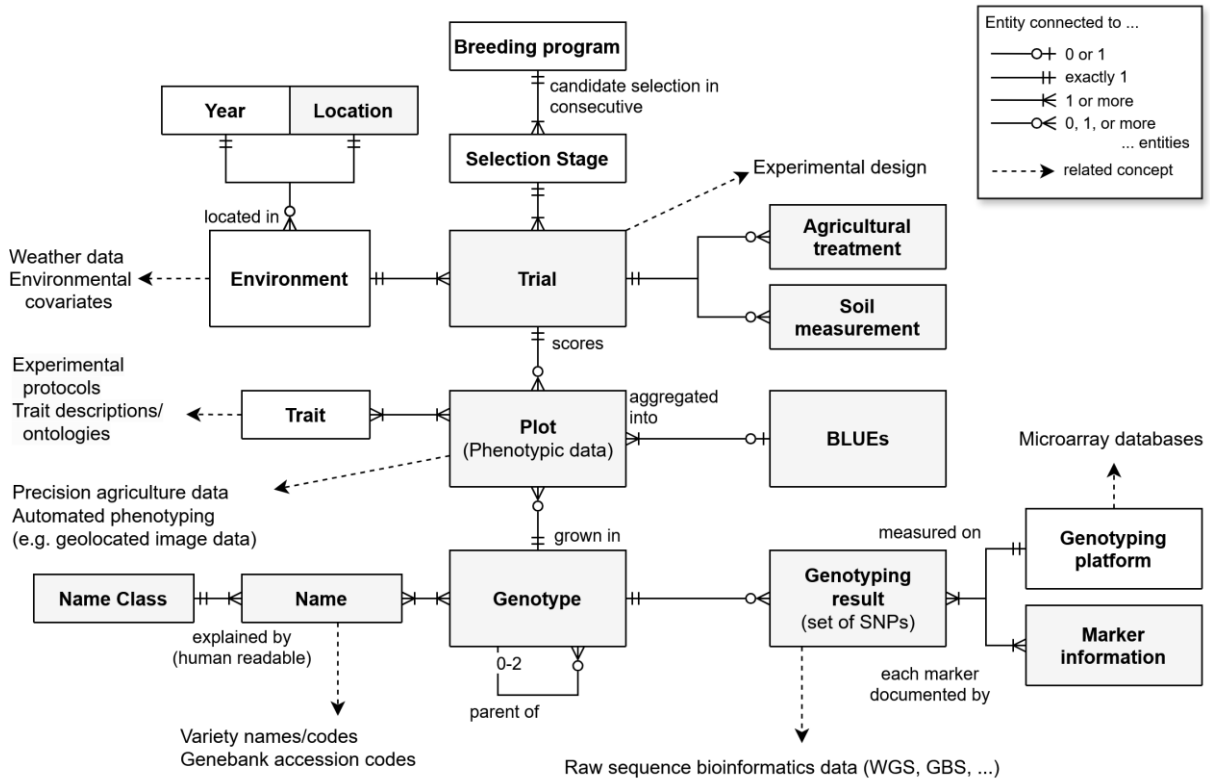

Figure 2: *Entity-relationship diagram of the data model underlying the data of the Wheat BigData project (Lell et al., 2025). Each box represents a concept which was represented in the data either as a table or a field. Relationships between the concepts are shown using the crow-foot notation, indicating how many instances of a concept are referenced by another concept (cardinality). Related concepts to be considered for future extensions of the data model are marked with dashed lines and arrows. Data about agricultural treatments and soil measurements were not available in sufficient detail to be studied in this thesis.*

individuals, be it genotypes with rare genetics or environments with underrepresented environmental conditions. This has to be weighed against the risk of including erroneous data.

After data upload and correction, second-degree statistics and heritability estimates were reported back to the data uploaders. As with error correction a fast turn-around time by an automated process was desirable, because data providers could compare the results to their internal calculations and corrections could be discussed while memory of scientists and technicians of the experiments was still fresh. However, this endeavour entailed the challenge of automatically choosing an appropriate statistical model for a given dataset to remove design effects and calculate entry means. This is even more relevant when envisioning a future, larger and fully automated data curation system. In this thesis, several steps in this direction were taken (Lell et al., 2025). To specify appropriate ways of removing design effects and forming entry means, two different approaches are conceivable: One possibility is to ask the data providers for a suitable statistical model for the data. As an experimental design is usually chosen with a statistical evaluation in mind, the advantage of this approach is that it ensures that a well-suited statistical evaluation is performed and a large variety of evaluations are possible. On the flip side, the approach communicates the experimental design metadata of the trial only indirectly, via the desired statistical model. Other context, for example relating to the physical location of plots might be lost if it was not deemed relevant for the analysis at upload time. If many data sets in a data warehouse followed this standard, deviating from the originally desired statistical model for a reanalysis would require a lot of dataset-specific manual work to infer or guess such missing context. Therefore, another approach was taken in this thesis which stored the spatial positions and groupings in the field by design effect fields whose meaning was pre-defined across all trials: The Series delineates trial parts that test disjoint sets of candidates. The Replication delineates trial parts that test the same set of candidates. The Block contains overlapping subsets of the candidates. All design effects are expected to relate to spatial grouping of the contained plots, nested within each other in the above order of mention. Using this data model, heterogeneity of experimental design can be depicted by omitting certain design effects in a data set.

Enforcing such pre-defined semantics onto diverse phenotyping data can be too restrictive at times. This approach assumes that all relevant experimental designs estimate the same kinds of design effects, an assumption which does not hold in general, as can be shown by comparing for example an alpha-lattice and a partially replicated

design. While the alpha-lattice design uses spatially separated replications, the partially replicated design does not. Therefore, the term “Replication” can have different meanings, depending on the experimental design. This can entail completely different statistical treatment. To allow for such diverse experimental designs while still communicating the intention behind the design, the data model of a future wheat data warehouse should include an object describing the experimental design. This will determine which design effects must be specified for plots in a data upload, and provide documentation about their meaning. Equally, it will determine how the observed phenotypes can be summarized. For example, for an experimental design involving hybrids, general and special combining abilities can be estimated instead of a per-environment entry mean. For the historical data sets analysed in Lell et al. (2025), which have more diverse experimental designs, the shortcomings of the data model were mitigated by custom analysis procedures and by integrating the data only at the within-environment entry-mean level. In the future, a more standardized approach for such data sets is desirable.

3.1.2 Separating different genotype identifiers by systematics and usage

To merge breeding data from different origins, finding common genotypes is crucial. Genomic data is one way to achieve this. For example, a threshold fraction of identical SNPs can be defined above which two genotypes are treated as equal (Schulthess et al., 2022b). This approach is indirect, which is dependent on sufficiently compatible genomic data to compare genotypes across data set borders. While imputation seems to be suited to match low to high density SNP profiles (He et al., 2015; Lell et al., 2025) and intersecting SNP with GBS data appears to be a viable option as well (Schulthess et al., 2022a), the selection of compared genetic loci remains a factor influencing the result. Moreover, identifying common genotypes via genomic data mandates that phenotypic and genomic data come from the same source and have identical genotype identifiers. When approaching an automated data sharing regime involving early- and late-stage trials and many sharing parties, more elaborate handling of genotype identifiers beyond a simple name or code is important.

In commercial variety breeding, a candidate will be given multiple names, which can lead to same or related genotypes not being recognized as such in a later data analysis. Worse, different genotypes could be mistaken for the same if they happen to get the same name or code in different projects. This means that some genotype names must

be valid across data sets in order to link the data sets, but many genotype names are also dataset-specific and must not be matched across data set borders. Therefore, data curation based on knowledge of different genotype identifiers must be performed. The earliest names given to variety candidates are breeder's references, codes that usually follow a system, sometimes including pedigree information. In early breeding stages, these are assigned anew each year, as the heterozygosity of early candidates will cause offspring of self-fertilization to differ from the parents. Doubled haploids, F1-hybrids and late-stage, highly homozygous candidates can retain their breeder's reference code due to their genetic stability. Upon application, some breeders confer an additional code to their genotypes, an application code, in order to prevent others from concluding from the systematics of the breeder's code to the logistics of the underlying breeding program. Finally, upon application, the German Plant Variety Office assigns a numeric application code and finally, if the variety accepted, it is marketed using a variety name defined by the breeder and registered by the Plant Variety Office. German and European variety protection laws require that variety names are distinct from each other, however for example in the European Union, this does allow names to be reused 10 years after expiry of a variety (European Commission regulation No 637/2009 of 22 July 2009) or immediately if a variety application has been withdrawn. In addition to these different genotype identifiers, projects might choose to assign project-specific names to genotypes, for example to enumerate parents of hybrid trials.

If these many different kinds of identifiers are not separated and documented appropriately, data curation becomes very laborious and error-prone. Different naming systems require completely different approaches to match genotypes across different data origins. For example, matching variety names to each other could be done using fuzzy string matching, which allows for inexact matches in order to identify misspellings or text encoding errors when non-ASCII letters are used in the name. This approach is not viable for systematic names that involve monotonically rising numeric IDs. When a naming scheme for F1 hybrids that involve the parent names is used, F1 hybrids from one project could be identified as equal to an F1 hybrid from another project of the same parents, given that the female and male role is assigned equally. However, this cannot be done when matching 3-way hybrids from different experiments, which are genetically different from each other despite originating from the same parents.

Given the multitude of different naming schemes, it is unlikely that a completely automated process of genotype matching is feasible. However, data structures and metadata should communicate the used naming schemes and systematics to enable

future data reuse. In the data underlying this thesis (Lell et al., 2025), the concept of a "name class" was introduced (Figure 2), that states for each type of identifier from which naming system it originates. For each name class, human-readable information could be provided, which explains the creator of the name, its purpose, the domain within it is meant to be unique, and its structure if it is composed of multiple parts or systematically. By using internally-generated synthetic keys for genotypes during the analysis, multiple names could be linked to the same genotype and for different data curation steps only names of suitable classes could be used. However, this system, while being superior to a single unqualified genotype name property, is still dependent on the cooperation of the data providers, and must be supervised. The properties of the name classes can provide a schema for orientation when describing genotype naming systematics. However, objective criteria that those free-form descriptions have to fulfil and that can be checked like data constraints are not yet attainable.

3.1.3 Comparison of the data model developed in the frame of this thesis with other approaches highlights importance of documented data semantics

There have been several prior attempts of to integrate breeding data. For example, Germeier and Unger (2019) describe the data model underlying the structure of the crop databases for oat and beet of the European Cooperative Programme for Plant Genetic Resources (ECPGR). They are proponents of data models whose entities are as close as possible to concepts known by domain experts. The data model is similar to that used in this thesis (Lell et al., 2025), which is shown in Figure 2. Both approaches have in their centre a description of a phenotypic observation, obtained from the field. In both approaches, the spatial location of the plot in the experimental grid is determined by a pair of dedicated properties of the plot. Germeier and Unger do not introduce specific properties for other design effects like Series and Replication, in contrast to this thesis, thereby circumventing the aforementioned complications arising from different experimental designs. To model those design effects in the systematics of Germeier and Unger, separate experiments at the same location could be defined, each representing one Replication within one Series. Analysis routines would need to be informed separately about the fact that some experiments are replications of each other, the data model does not depict this fact. With regards to agricultural interventions like fertilization, pathogen inoculation or pesticide application, the model of Germeier and Unger binds those data to the plot level, allowing in theory for a different treatment regime for each

plot. This is not the case in the model of this thesis, where interventions are expected to apply to whole trials. This highlights that data models always carry a subjective component and that documentation on the data structure is important to communicate these assumptions. Finally, Germeier and Unger share with this thesis the notion that computed summaries on observation data, like entry means, are a different entity than the raw observations, and should be modelled by a dedicated data table with different properties from the raw observations.

A very notable further instance of breeding data modelling is the Breeding Application Programming Interface (BrAPI) project, which was initiated almost 10 years ago (Selby et al., 2019). The project does not aim to form a single integrated database to harbour all breeding data, but rather strives to establish a common interface for as many partners as possible to enable their respective services to exchange data using queries that follow a common format. Nevertheless, the project has established a data model to enable users to understand the structure in which queries have to be formed and the meaning of the responses from the data sources. Owing to the diversity of the project partners, the data model is significantly more complex than the one employed in this thesis (Lell et al., 2025). It shows some interesting deviations, using slightly more abstract terms. The project has started multiple efforts to convey the meaning of the entities and their relations despite their complexity. These comprise of an entity-relationship diagram, a glossary where terms related to the project are explained (Guignon and Selby, 2023), and a declarative definition of the API calls in the OpenAPI format, which is annotated by human-readable descriptions (Selby, 2022). The BrAPI project employs collaborative tools for development by its members, similarly to a wiki. This lowers the barrier to correct errors and propose improvements to the API and the documentation and highlights that data integration across different parties is not only a technical question to be solved, but also a social question about means of collaboration and participation.

The BrAPI data model unifies multiple concepts that are modelled differently by Germeier and Unger (2019) and Lell et al. (2025). For example, the concepts of a field, a block, or a plot are collated into the concept of an Observation Unit. The reasoning behind this is that different projects use different observation units and nest them into each other differently. In addition, it allows to link measurements not only to a plot, but to any observation unit. For example, heritability estimates could refer to a field or an environment rather than individual plots. Besides observational units, the BrAPI data model considers observations and statistical computations as fundamentally

similar and models both of them with the single entity “Observation”. Distinguishing between raw and computed data is done within the same framework as distinguishing observations of different traits. Via an entity called “ObservationVariable” trait and method entities are linked, that describe what is observed and how it is done. The entities regarding methods and traits can also be linked to ontologies like the Crop Ontology (Matteis et al., 2013), which strives to unify meanings of plant trialling terms and their relationships to each other across the whole scientific field. A similar approach is seen with Unger and Germeier (2019), however in Lell et al. (2025) such a detailed model of traits and methods is still missing.

3.2 Population structure and its influence on genomic prediction accuracy

Genomic prediction has revolutionized plant breeding by its broader applicability compared to pedigree-based predictions. Its key innovation is replacing the numerator relationship matrix, which is derived from the genotype pedigree, with a matrix of marker states (Meuwissen et al., 2001) or the realized or genomic relationship matrix, which is derived therefrom (VanRaden, 2008). This allows predictions to be based on a richer source of data that also encodes the relatedness of the founding generation and the influence of Mendelian sampling. This depicts relatedness in a more fine-grained way than the model of additive genetic relationship in the numerator relationship matrix. In Genomic Prediction, estimated effects of individual linkage blocks shared by related genotypes were considered the basis for prediction, rather than the expected global fraction of shared DNA that is the additive genetic relationship (Meuwissen et al., 2001). Interestingly though, later it was recognized that the genotype pedigree and additive genetic relationship re-emerge as a predictive factor in genomic prediction “through the back door” as the number of common markers approximates the number of unlinked causative alleles shared by chance between a pair of individuals. Therefore, the marker states carry predictive value even if no linkage blocks would exist in a population (Habier et al., 2007). The power of genomic prediction therefore rests on both the ability to indirectly estimate the effect of individual linkage blocks on the phenotype as well as to find related genotypes with known phenotype in the population to use as predictors.

Consequently, the expected prediction accuracy is influenced by the degree of relationship in the population as well as by the ratio of the number of observed genotypes to the effective number of independent chromosome segments (linkage blocks) in the population (Hayes et al., 2009b). This means that the number of genotypes that have

to be observed grows with the population diversity. Moreover, the study of Hayes et al. (2009b) lists heritability as another decisive factor for an accurate genomic prediction. For complex traits, achieving high heritabilities in crop plants therefore requires sampling many environments. This multiplies the work amount of observing a sufficient number of genotypes and potentially goes beyond the reach of individual breeding programs. The central task of this thesis therefore was to establish the potential and limits of merging data that was obtained through the efforts of multiple individual breeding programs. The most extensive data set was available in Lell et al. (2025), where eight different sources of data were joined into a common training set. As a main result, this merger proved feasible. Increasing the training set size improved the prediction ability on average, while no single experimental series proved particularly good or bad as training set, beyond its size. Deviations from this relationship were caused by individual combinations of series that proved particularly useful or detrimental if one was in the training set and the other was the test set, for example experimental series 1 and 2. At the same time, the prediction ability plateaus off well below the theoretical limit of the heritability. Therefore, it might be worthwhile to search for factors that pairs of experimental series share in order to identify potentials for further improving prediction abilities.

One aspect to explore further is the amount of relationship between the individual data sets. Given that several populations in the study share the same time range, it seems likely that part of the crossing partners should be shared among them and some crosses should even have been done in multiple breeding programs. Assuming no close relationship between the individual breeding programs, the observed prediction accuracies are close to the expectation as modelled by Zhao, et al. (2021, see also Figure 3). For families of genotypes where close relatives exist in other breeding programs it should be able to achieve better prediction accuracies, leveraging the pedigree information encoded in the SNP data in addition to linkage-based prediction (De Roos et al., 2009; Habier et al., 2007). Attempting to reconstruct the genealogy from the linkage blocks observed in the data could be an interesting further study avenue (Fan et al., 2022) as it could allow to directly identify offspring of crosses that were performed in multiple experimental series.

3.3 Benchmarking Big Data-based genomic prediction approaches could be skewed by confounded genotype-times-environment interactions

Grain yield is a complex trait that is a combination of multiple components. These can be significantly influenced by the environment, particularly by extreme conditions during short but critical time periods during grain development (Sabir et al., 2023). When gathering multiple data sets for secondary use like it has been done in this work, inevitably unbalanced data sets result, as different studies evaluate different genotypes in different environments. This problem is exacerbated by the internally unbalanced character of many of the studied breeding data sets, which evaluate many candidates in few environments in early stages and fewer candidates in more environments in later stages. Therefore, especially in the early breeding stages, it has to be expected that genotype-times-environment effects are confounded with the estimated main genotypic effects. Including early-stage genotypes can therefore have conflicting influence on prediction results: While the large number of diverse genotypes can improve training set size and thus the prediction accuracy, the confounded genotype-times-environment interactions can limit the accuracy of the estimate genetic values and thus lower the trait heritability and prediction accuracy. The sensitivity of the merged data set to this effect could be estimated in a future study by generating artificial phenotypic data, with a known variance due to genotype-times-environment interactions, and testing the

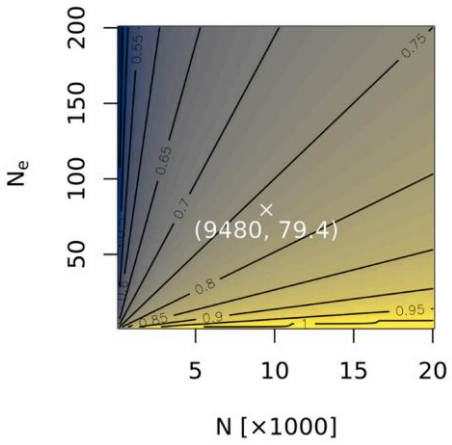

Figure 3: *Relationship between prediction accuracy (background/contours), population size (N) and effective population size (Ne). The data comes from Zhao, et al. (2021, Fig. 4C). The white marker shows the nominal and effective population sizes of the combined data from Lell, et al. (2025). As in Zhao, et al. (2021) the prediction ability is shown, it is divided by the square root of the heritability (0.70, Zhao et al., 2021, Table S1) to obtain the prediction accuracy of this figure.*

accuracy of the recovered genotypic values. The real genomic data can be used to provide a realistic population structure to the simulation.

The impact of genotype-times-environment interactions on genomic prediction could be decreased by sparsely distributing the evaluation of early candidates onto more environments without increasing the total number of genotypes. This approach was tested within this thesis for wheat hybrid breeding (Lell et al., 2021) and showed that sparse environmental sampling can be performed in a way so that the resulting estimates were at least on par with those derived from the classical balanced early-stage trial with very few environments. When adding the assumption that the true breeding values of the candidates were approximated better by GBLUP than by classical BLUE, the environmental sampling approach was even superior. It is interesting to note that the *a priori* decision whether breeding values estimated should be estimated with (GBLUP) or without (BLUE) considering relatedness influenced which environmental sampling strategy was found superior. This inconsistency might hint to the fact that even as the candidates were evaluated in two years and eleven environments in a balanced fashion, undetected genotype-times-environment interactions could have caused deviations of the BLUEs from the genomic breeding values estimated by GBLUP. It is therefore possible that the genotypic values estimated from the late-stage commercial breeding trials in Lell, et al. (2025), that are tested in less environments than the data of Lell, et al. (2021), might still contain an undetected genotype-times-environment component and might not be ideal benchmarks to judge prediction efforts against. A potential way to estimate the validity of either of the two estimation methods could lie in predicting genotypes for which a larger and more diverse set of environments is available. This would reduce the confounding of genotype-times-environment effects with genotypic main effects.

3.4 Big Data can help to obtain a fuller picture of marker-trait associations in wheat

In order to detect even QTL with a small trait contribution via GWAS, sample sizes of 1000 genotypes could already provide enough statistical power (Sul et al., 2018). However, this calculation assumes purely additive effects. Detecting dominance effects requires only slightly higher sample sizes (Wang and Xu, 2019). However, detecting epistatic interactions requires a massively higher number of tests as combinations of markers have to be detected. Therefore, higher sample sizes are

needed to elucidate such interactions. This can also be seen from the results of this thesis, where additive and dominance effects with good predictive ability could be detected but no epistatic interactions were found (Lell et al., 2024).

The higher genetic diversity that results from joining of many data sets from diverse origins is an opportunity and a challenge at the same time. This thesis has shown how a higher genetic diversity, manifesting in a larger effective population size, can increase the detection power (Lell et al., 2024). At the same time, population structure can introduce numerous confounding effects. For example, strong population structure can introduce genetic bottlenecks into the data which can cause false-positive synthetic marker-trait associations (Korte and Farlow, 2013). If merging many data sources into one allows to fill the genetic voids between distinct populations, the confounding effects of population structure could be weakened as it becomes easier to discern the effects of individual loci. A recent study supports this approach by extracting diverse core sets from a large data base of wheat genetic resources (Berkner et al., 2024). The study shows higher GWAS power resulting from a higher genetic diversity. A prerequisite is though that the marker density in the individual data sets is sufficient such that most genome loci are in substantial linkage disequilibrium to markers. This is especially important to detect rare variants, where correlation to nearby markers is sensitive to different minor allele frequencies of causal variant and marker in the population, which limits the power to detect them (Korte and Farlow, 2013). Efforts to achieve denser genotype sampling by providing haplotype maps for imputation are underway and could be used to improve pre-existing marker data of historic data sets (Jordan et al., 2022; Nyine et al., 2019).

However, a larger and more diverse population could also cause a false-negative deception of a kind that could be thought as converse to the false-positive synthetic association: Two different causative loci that occur in distinct sub-populations could mutually weaken the detection power in a population-wide GWAS (Korte and Farlow, 2013). Using GWAS methods that estimate all marker effects simultaneously would ameliorate this problem (Segura et al., 2012; Wen et al., 2018).

3.5 Outlook

It is an encouraging sign that in wheat, commercial breeding populations in Central Europe seem to form no separated individual clusters but rather a large connected space of genetic diversity (Lell et al., 2025). Therefore, a deeper integration of individual breeding programs has the potential to further improve the power of

genomic prediction. A pressing topic to elucidate on the way is the influence of genotype-times-environment interactions on the prediction results and to ensure that there is as little confounding as possible with main effect estimates. The architecture of the data as it is used for this thesis has this confounding as an Achilles heel because of its unbalanced nature. Many genotypes are observed in a low number of environments and many environments are linked by a small number of genotypes. In particular, environments from different breeding programmes often share only a handful of genotypes. These are usually state-of-the-art varieties that breeders use as check varieties in their experimental designs. The connectivity between environments could be significantly improved by agreeing on a core set of genotypes to be evaluated in a large number of environments across different breeding programs. Distributing genotypes sparsely across multiple environments is another approach that can improve the environment connectivity (Lell et al., 2021).

In order to provide data to serve as basis for these investigations, the data management which has been begun in this work must be further formalized and expanded to include environment metadata to enable envirotyping. In this endeavour, compatibility to the BrAPI data model should be retained, however, the BrAPI itself does not specify a way to represent environmental properties, like weather or soil information. Moreover, more detailed trait descriptions must be included as the diversity of data sources increases. Ontologies like the Crop Ontology are emerging to provide a unified set of terms to be used by the community, however, term annotations in many cases are still quite generic and not sufficient to reconstruct methodological differences among data sets. Formalizing knowledge through ontologies can serve as a basis to draw insights from heterogenous data, but only if the focus on human-readable term annotations is not neglected.

4 References

Alvarez-Castro, J.M., Carlborg, O., 2007. A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. Genetics 176, 1151–1167. https://doi.org/10.1534/genetics.106.067348

Berkner, M.O., Jiang, Y., Reif, J.C., Schulthess, A.W., 2024. Trait-customized sampling of core collections from a winter wheat genebank collection supports association studies. Front. Plant Sci. 15, 1451749. https://doi.org/10.3389/fpls.2024.1451749

Brard, S., Ricard, A., 2015. Is the use of formulae a reliable way to predict the accuracy of genomic selection? J. Anim. Breed. Genet. 132, 207–217. https://doi.org/10.1111/jbg.12123

Collard, B.C.Y., Mackill, D.J., 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc. B Biol. Sci. 363, 557–572. https://doi.org/10.1098/rstb.2007.2170

De Roos, A.P.W., Hayes, B.J., Goddard, M.E., 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183, 1545–1553. https://doi.org/10.1534/genetics.109.104935

De Vlaming, R., Groenen, P.J.F., 2015. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. BioMed Res. Int. 2015, 1–18. https://doi.org/10.1155/2015/143712

Fan, C., Mancuso, N., Chiang, C.W.K., 2022. A genealogical estimate of genetic relationships. Am. J. Hum. Genet. 109, 812–824. https://doi.org/10.1016/j.ajhg.2022.03.016

Fischer, T., Ammar, K., Monasterio, I.O., Monjardino, M., Singh, R., Verhulst, N., 2022. Sixty years of irrigated wheat yield increase in the Yaqui Valley of Mexico: Past drivers, prospects and sustainability. Field Crops Res. 283, 108528. https://doi.org/10.1016/j.fcr.2022.108528

Fradgley, N., Gardner, K.A., Cockram, J., Elderfield, J., Hickey, J.M., Howell, P., Jackson, R., Mackay, I.J., 2019. A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. PLOS Biol. 17, e3000071. https://doi.org/10.1371/journal.pbio.3000071

Gage, J.L., Monier, B., Giri, A., Buckler, E.S., 2020. Ten Years of the Maize Nested Association Mapping Population: Impact, Limitations, and Future Directions. Plant Cell 32, 2083–2093. https://doi.org/10.1105/tpc.19.00951

Germeier, C.U., Unger, S., 2019. Modeling Crop Genetic Resources Phenotyping Information Systems. Front. Plant Sci. 10, 728. https://doi.org/10.3389/fpls.2019.00728

Gerten, D., Heck, V., Jägermeyr, J., Bodirsky, B.L., Fetzer, I., Jalava, M., Kummu, M., Lucht, W., Rockström, J., Schaphoff, S., Schellnhuber, H.J., 2020. Feeding ten billion people is possible within four terrestrial planetary boundaries. Nat. Sustain. 3, 200–208. https://doi.org/10.1038/s41893-019-0465-1

Gianola, D., 2013. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics 194, 573–596. https://doi.org/10.1534/genetics.113.151753

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., Fernando, R., 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183, 347–363. https://doi.org/10.1534/genetics.109.103952

Gianola, D., van Kaam, J.B.C.H.M., 2008. Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. Genetics 178, 2289–2303. https://doi.org/10.1534/genetics.107.084285

Gilmour, A.R., Thompson, R., Cullis, B.R., 1995. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. Biometrics 51, 1440–1450.

Guignon, V., Selby, P., 2023. Brapi V2.1 Data Type Browser and Entity Relationship Diagram [WWW Document]. URL https://github.com/plantbreeding/brapi-ontology, git commit c4f2b90693b3 (author date: 2023-02-22), available via GitHub Pages at https://plantbreeding.github.io/brapi-ontology/index.html and https://plantbreeding.github.io/brapi-ontology/uml.html as of 2025-04-30

Habier, D., Fernando, R.L., Dekkers, J.C.M., 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177, 2389–2397. https://doi.org/10.1534/genetics.107.081190

Habier, D., Fernando, R.L., Garrick, D.J., 2013. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics 194, 597–607. https://doi.org/10.1534/genetics.113.152207

Hamblin, M.T., Buckler, E.S., Jannink, J.-L., 2011. Population genetics of genomics-based crop improvement methods. Trends Genet. 27, 98–106. https://doi.org/10.1016/j.tig.2010.12.003

Hasan, N., Choudhary, S., Naaz, N., Sharma, N., Laskar, R.A., 2021. Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. J. Genet. Eng. Biotechnol. 19, 128. https://doi.org/10.1186/s43141-021-00231-1

Hayes, B.J., Bowman, P.J., Chamberlain, A.C., Verbyla, K., Goddard, M.E., 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41, 51. https://doi.org/10.1186/1297-9686-41-51

Hayes, B.J., Visscher, P.M., Goddard, M.E., 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91, 47–60. https://doi.org/10.1017/S0016672308009981

He, S., Zhao, Y., Mette, M.F., Bothe, R., Ebmeyer, E., Sharbel, T.F., Reif, J.C., Jiang, Y., 2015. Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.). BMC Genomics 16, 1–12. https://doi.org/10.1186/s12864-015-1366-y

Henderson, C.R., 1988. Use of an Average Numerator Relationship Matrix for Multiple-Sire Joining. J. Anim. Sci. 66, 1614. https://doi.org/10.2527/jas1988.6671614x

Intergovernmental Panel On Climate Change, 2022. Climate Change and Land: IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems, 1st ed. Cambridge University Press. https://doi.org/10.1017/9781009157988

Jacquin, L., Cao, T.-V., Ahmadi, N., 2016. A Unified and Comprehensible View of Parametric and Kernel Methods for Genomic Prediction with Application to Rice. Front. Genet. 7. https://doi.org/10.3389/fgene.2016.00145

Jordan, K.W., Bradbury, P.J., Miller, Z.R., Nyine, M., He, F., Fraser, M., Anderson, J., Mason, E., Katz, A., Pearce, S., Carter, A.H., Prather, S., Pumphrey, M., Chen, J., Cook, J., Liu, S., Rudd, J.C., Wang, Z., Chu, C., Ibrahim, A.M.H., Turkus, J., Olson, E., Nagarajan, R., Carver, B., Yan, L., Taagen, E., Sorrells, M., Ward, B., Ren, J., Akhunova, A., Bai, G., Bowden, R., Fiedler, J., Faris, J., Dubcovsky, J., Guttieri, M., Brown-Guedira, G., Buckler, E., Jannink, J.-L., Akhunov, E.D., 2022. Development of the Wheat Practical Haplotype Graph database as a resource for genotyping data storage and genotype imputation. G3 GenesGenomesGenetics 12, jkab390. https://doi.org/10.1093/g3journal/jkab390

Korte, A., Farlow, A., 2013. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9, 29. https://doi.org/10.1186/1746-4811-9-29

Lell, M., Gogna, A., Kloesgen, V., Avenhaus, U., Dörnte, J., Eckhoff, W.M., Eschholz, T., Gils, M., Kirchhoff, M., Kollers, S., Pfeiffer, N., Rapp, M., Wimmer, V., Wolf, M., Reif, J., Zhao, Y., 2025. Breaking down data silos across companies to train genome-wide predictions – a feasibility study in wheat. Plant Biotechnol. J. pbi.70095. https://doi.org/10.1111/pbi.70095

Lell, M., Reif, J., Zhao, Y., 2021. Optimizing the setup of multienvironmental hybrid wheat yield trials for boosting the selection capability. Plant Genome 14, e20150. https://doi.org/10.1002/tpg2.20150

Lell, M., Zhao, Y., Reif, J.C., 2024. Leveraging the potential of big genomic and phenotypic data for genome-wide association mapping in wheat. Crop J. 12, 803--813. https://doi.org/10.1016/j.cj.2024.03.005

Liu, Y., He, Z., Appels, R., Xia, X., 2012. Functional markers in wheat: current status and future prospects. Theor. Appl. Genet. 125, 1–10. https://doi.org/10.1007/s00122-012-1829-3

Matteis, L., Chibon, P.Y., Espinosa, H., Skofic, M., Finkers, H.J., Bruskiewich, R., Hyman, J.M., Arnoud, E., 2013. Crop Ontology: Vocabulary For Crop-related Concepts.

Meuwissen, T.H., 2009. Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41, 35. https://doi.org/10.1186/1297-9686-41-35

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

Newton, J.E., Nettle, R., Pryce, J.E., 2020. Farming smarter with big data: Insights from the case of Australia's national dairy herd milk recording scheme. Agric. Syst. 181, 102811. https://doi.org/10.1016/j.agsy.2020.102811

Nyine, M., Wang, S., Kiani, K., Jordan, K., Liu, S., Byrne, P., Haley, S., Baenziger, S., Chao, S., Bowden, R., Akhunov, E., 2019. Genotype Imputation in Winter Wheat Using First-Generation Haplotype Map SNPs Improves Genome-Wide Association Mapping and Genomic Prediction of Traits. G3 GenesGenomesGenetics 9, 125–133. https://doi.org/10.1534/g3.118.200664

Rasheed, A., Xia, X., 2019. From markers to genome-based breeding in wheat. Theor. Appl. Genet. 132, 767–784. https://doi.org/10.1007/s00122-019-03286-4

Sabir, K., Rose, T., Wittkop, B., Stahl, A., Snowdon, R.J., Ballvora, A., Friedt, W., Kage, H., Léon, J., Ordon, F., Stützel, H., Zetzsche, H., Chen, T.-W., 2023. Stage-specific genotype-by-environment interactions determine yield components in wheat. Nat. Plants 9, 1688–1696. https://doi.org/10.1038/s41477-023-01516-8

Schulthess, A.W., Kale, S.M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., Jiang, Y., Beukert, U., Serfling, A., Himmelbach, A., Fuchs, J., Oppermann, M., Weise, S., Boeven, P.H.G., Schacht, J., Longin, C.F.H., Kollers, S., Pfeiffer, N., Korzun, V., Lange, M., Scholz, U., Stein, N., Mascher, M., Reif, J.C., 2022a. Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. Nat. Genet. 54, 1544–1552. https://doi.org/10.1038/s41588-022-01189-7

Schulthess, A.W., Kale, S.M., Zhao, Y., Gogna, A., Rembe, M., Philipp, N., Liu, F., Beukert, U., Serfling, A., Himmelbach, A., Oppermann, M., Weise, S., Boeven, P.H.G., Schacht, J., Longin, C.F.H., Kollers, S., Pfeiffer, N., Korzun, V., Fiebig, A., Schüler, D., Lange, M., Scholz, U., Stein, N., Mascher, M., Reif, J.C., 2022b. Large-scale genotyping and phenotyping of a worldwide winter wheat genebank for its use in pre-breeding. Sci. Data 9, 784. https://doi.org/10.1038/s41597-022-01891-5

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., Nordborg, M., 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44, 825–830. https://doi.org/10.1038/ng.2314

Selby, P., 2022. BrAPI Specification V2.1 [WWW Document]. URL https://github.com/plantbreeding/BrAPI, git commit: a45723dba875 (author date: 2022-07-01)

Selby, P., Abbeloos, R., Backlund, J.E., Basterrechea Salido, M., Bauchet, G., Benites-Alfaro, O.E., Birkett, C., Calaminos, V.C., Carceller, P., Cornut, G., Vasques Costa, B., Edwards, J.D., Finkers, R., Yanxin Gao, S., Ghaffar, M., Glaser, P., Guignon, V., Hok, P., Kilian, A., König, P., Lagare, J.E.B., Lange, M., Laporte, M.-A., Larmande, P., LeBauer, D.S., Lyon, D.A., Marshall, D.S., Matthews, D., Milne, I., Mistry, N., Morales, N., Mueller, L.A., Neveu, P., Papoutsoglou, E., Pearce, B., Perez-Masias, I., Pommier, C., Ramírez-González, R.H., Rathore, A., Raquel, A.M., Raubach, S., Rife, T., Robbins, K., Rouard, M., Sarma, C., Scholz, U., Sempéré, G., Shaw, P.D., Simon, R., Soldevilla, N., Stephen, G., Sun, Q., Tovar, C., Uszynski, G., Verouden, M., The BrAPI consortium, 2019. BrAPI—an application programming interface for plant breeding applications. Bioinformatics 35, 4147–4155. https://doi.org/10.1093/bioinformatics/btz190

Shiferaw, B., Smale, M., Braun, H.-J., Duveiller, E., Reynolds, M., Muricho, G., 2013. Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. Food Secur. 5, 291–317. https://doi.org/10.1007/s12571-013-0263-y

Song, L., Wang, R., Yang, X., Zhang, A., Liu, D., 2023. Molecular Markers and Their Applications in Marker-Assisted Selection (MAS) in Bread Wheat (Triticum aestivum L.). Agriculture 13, 642. https://doi.org/10.3390/agriculture13030642

Sul, J.H., Martin, L.S., Eskin, E., 2018. Population structure in genetic studies: Confounding factors and mixed models. PLOS Genet. 14, e1007309. https://doi.org/10.1371/journal.pgen.1007309

Tian, X., Engel, B.A., Qian, H., Hua, E., Sun, S., Wang, Y., 2021. Will reaching the maximum achievable yield potential meet future global food demand? J. Clean. Prod. 294, 126285. https://doi.org/10.1016/j.jclepro.2021.126285

Tibbs Cortes, L., Zhang, Z., Yu, J., 2021. Status and prospects of genome-wide association studies in plants. Plant Genome 14, e20077. https://doi.org/10.1002/tpg2.20077

Tilman, D., Balzer, C., Hill, J., Befort, B.L., 2011. Global food demand and the sustainable intensification of agriculture. Proc. Natl. Acad. Sci. 108, 20260–20264. https://doi.org/10.1073/pnas.1116437108

Trethowan, R.M., Reynolds, M.P., Ortiz-Monasterio, J.I., Ortiz, R., 2007. The Genetic Basis of the Green Revolution in Wheat Production, in: Janick, J. (Ed.), Plant Breeding Reviews. Wiley, pp. 39–58. https://doi.org/10.1002/9780470168028.ch2

Van Dijk, M., Morley, T., Rau, M.L., Saghai, Y., 2021. A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. Nat. Food 2, 494–501. https://doi.org/10.1038/s43016-021-00322-9

VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Voss-Fels, K.P., Cooper, M., Hayes, B.J., 2019. Accelerating crop genetic gains with genomic selection. Theor. Appl. Genet. 132, 669–686. https://doi.org/10.1007/s00122-018-3270-8

Wang, M., Xu, S., 2019. Statistical power in genome-wide association studies and quantitative trait locus mapping. Heredity 123, 287–306. https://doi.org/10.1038/s41437-019-0205-3

Wen, Y.-J., Zhang, H., Ni, Y.-L., Huang, B., Zhang, J., Feng, J.-Y., Wang, S.-B., Dunwell, J.M., Zhang, Y.-M., Wu, R., 2018. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Brief. Bioinform. 19, 700–712. https://doi.org/10.1093/bib/bbw145

Wetterstrand, K.A., 2024. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) [WWW Document]. URL https://www.genome.gov/sequencingcostsdata (accessed 2024-10-22)

Whittaker, J.C., Thompson, R., Denham, M.C., 2000. Marker-assisted selection using ridge regression. Genet. Res. 75, 249–252. https://doi.org/10.1017/S0016672399004462

Würschum, T., 2012. Mapping QTL for agronomic traits in breeding populations. Theor. Appl. Genet. 125, 201–210. https://doi.org/10.1007/s00122-012-1887-6

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S., 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203–208. https://doi.org/10.1038/ng1702

Zhao, Y., Thorwarth, P., Jiang, Y., Philipp, N., Schulthess, A.W., Gils, M., Boeven, P.H.G., Longin, C.F.H., Schacht, J., Ebmeyer, E., Korzun, V., Mirdita, V., Dörnte, J., Avenhaus, U., Horbach, R., Cöster, H., Holzapfel, J., Ramgraber, L., Kühnle, S., Varenne, P., Starke, A., Schürmann, F., Beier, S., Scholz, U., Liu, F., Schmidt, R.H., Reif, J.C., 2021. Unlocking big data doubled the accuracy in predicting the grain yield in hybrid wheat. Sci. Adv. 7, eabf9106. https://doi.org/10.1126/sciadv.abf9106

5 Acknowledgements

I want to thank my supervisors Prof. Dr. Jochen Reif and Dr. Yusheng Zhao for the countless discussions and ideas and their relentless support over many years that have made this work possible!

I am grateful for the time that I could spend in the working group Quantitative Genetics at the IPK and want to thank all of my colleagues for building an atmosphere of mutual trust, support and inspiration. Thank you, Dr. Albert Schulthess and Dr. Yong Jiang, for your knowledgeable and patient explanations of many mathematical and quantitative genetic topics, and Dr. Renate Schmidt for your valuable suggestions and critique. I am grateful to Abhishek Gogna, Valentin Hinterberger, Dr. Max Rembe, and Dr. Norman Philipp for many critical discussions of this work, breeding, and many more topics, and for your company.

Throughout this work I have profited a lot from the dedication of the colleagues of the IT group of the IPK in maintaining the computational resources of the institute. Thank you for your continuous support.

I want to thank the colleagues of the Wheat BigData project partners in the breeding companies Deutsche Saatveredelung AG, KWS SAAT SE/KWS LOCHOW GmbH, Nordsaat Saatzucht GmbH, SU BIOTEC GmbH, and W. von Borries-Eckendorf GmbH & Co. KG for sharing valuable data and insights of commercial plant breeding, and discussions.

Eidesstattliche Erklärung / Declaration under Oath

Ich erkläre an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content.

_____ _____

Datum / *Date* Unterschrift des Antragstellers / *Signature of the applicant*

Erklärung über bestehende Vorstrafen und anhängige Ermittlungsverfahren / Declaration concerning Criminal Record and Pending Investigations

Hiermit erkläre ich, dass ich weder vorbestraft bin, noch dass gegen mich Ermittlungsverfahren anhängig sind.

I hereby declare that I have no criminal record and that no preliminary investigations are pending against me.

_____ _____

Datum / *Date* Unterschrift des Antragstellers / *Signature of the applicant*

Curriculum Vitae of Moritz Lell

Oct 2017 — Feb 2025

PhD student / Scientific Employee
Leibniz Institute for Plant Genetics and Crop Plant Research (IPK)
Group of Quantitative Genetics (Prof. Dr. Jochen Reif)

Oct 2014– Mar 2017

Studies Bioinformatics, M. Sc.
University of Potsdam, Germany
Thesis: „Bayesian estimation of fitness costs in the context of resistance accumulation during HIV therapy“ (Prof. Dr. Wilhelm Huisinga, Chair of Mathematical Modelling and Systems Biology)

Oct 2010— Feb 2014

Studies Biochemistry, B. Sc.
University of Bayreuth, Germany
Thesis: “Struktur-Funktionsanalysen zur Aktivität von Phytochelatinsynthasen aus Caenorhabditis elegans und Nostoc spec.: Mutagenese und heterologe Expression” (Prof. Dr. Stefan Clemens, Chair of Plant Physiology)

Sep 2009— Jul 2010

Civilian Service
Bund Naturschutz in Bayern e.V., KG Regensburg

2000 — 2009

Goethe-Gymnasium Regensburg

_____ _____

Datum / *Date* Moritz Lell