Incorporating Interactions and Gene Annotation Data in Genomic Prediction

Dissertation

for the Doctoral Degree

Dr. sc. agr.

at the Faculty of Agricultural Sciences,

Department of Animal Sciences,

Georg-August Universität Göttingen

presented by

Johannes Wolfgang Robert Martini

born in Kronach

Göttingen, September 2017

1^{st} Referee: Prof. Dr. Henner Simianer
Animal Breeding and Genetics
Department of Animal Sciences
Georg-August Universität Göttingen

2^{nd} Referee: Prof. Dr. Thomas Kneib
Statistics and Econometrics
Georg-August Universität Göttingen

Date of disputation: 3^{rd} of November, 2017

Table of contents

Preliminaries

Acknowledgments

First, I thank Prof. Henner Simianer for making this project possible and for the constant support during the last three years. In particular, I also thank him for encouraging his students to present on conferences and to get into contact with external scientists.

I also thank KWS SAAT SE for financing the project, and in particular Valentin Wimmer and Andreas Menze for their support and advice.

I thank Prof. Thomas Kneib and Prof. Tatyana Krivobokova for being member of my thesis committee and for their valuable advices.

I thank Prof. Rodolfo "Fito" Juan Carlos Cantet and his group at the University of Buenos Aires, for welcoming me twice, for visiting the KWS SAAT SE breeding station in Chivilcoy with me, for introducing me to the Argentinean cattle breeders, and for many other things. In particular, I am very thankful to Carolina Andrea García Baccino and Martín Emilio Mosbruker for the organizational support in Argentina.

Moreover, I thank the Animal Breeding and Genetics Group of Göttingen for the interaction and the pleasant working atmosphere. In particular, I would like to thank here the co-authors of my publications Malena Erbe, Diercles F. Cardoso, Torsten Pook, Christian Reimer, and Ning Gao. I also thank Swetlana Berger for her support in familiarizing me with the topic of genomic selection and Frau Döring for all the help with administration issues.

Finally, I thank María Emilia Barreyro for her support in the last two years.

Summary

With the broad availability of genomic data, the concept of predicting *genetic values* of individual animals or plant lines from their genomic data (*genomic prediction*) has become an everyday tool in plant and animal breeding programs in the last decade. The standard model of quantitative genetics is built upon a linear model in which locus effects are usually considered to be additive and the (additive) genetic value is the sum of all these locus effects.

In this work, we consider theoretical and practical aspects of models incorporating interactions, so called "epistasis models". Moreover, we follow approaches towards the incorporation of biological prior knowledge into prediction methods, and discuss the potential usefulness of a higher prediction accuracy for the *total genetic value*. The total genetic value represents the complete genetic basis of a phenotype and not only the part captured by additive effects. The three main chapters are given by the corresponding published articles.

After the general introduction, the section "Epistasis and covariance: How gene interaction translates into genomic relationship" addresses the correspondence between interaction effect models and models based on *genomic relationship matrices* as genetic covariance matrix. In particular, this leads from an interaction effect model in which interactions are modeled by products of markers as predictor variables to Hadamard products of relationship matrices (extended GBLUP or EGBLUP). As an example for the usefulness of variable selection by prior experimental data in epistasis models and for the prediction of the performance of plant lines under different environmental conditions, we use a well-studied wheat data set which provides phenotype records of the same lines grown under four different environmental conditions. The interaction effects are estimated with the data of one environment and their absolute effect size is used as an indicator for their relevance. The complete model with all pairwise interactions is then reduced to the interactions which are more relevant (according to absolute effect sizes). This subsystem of pairwise interaction effects is translated into a relationship matrix which is used for genomic prediction within the data of the plants being grown

under different environmental conditions. The results show that an epistasis model with pairwise interactions can improve predictive ability and that data from previous experiments can be used in a beneficial way as external information for variable selection.

The section "Genomic prediction with epistasis models: On the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE)" is motivated by the fact that ridge regression models with interactions modeled as products of the marker values (EGBLUP) are not invariant to translations of the marker coding. Although the effect on predictive ability may be small when a high number of markers is used and when both the additive effects and the interactions are modeled, the coding dependence provides a motivation to consider other types of epistasis models. We demonstrate the coding-dependent performance of EGBLUP and investigate coding-dependent theoretical properties of the interaction effect model. Moreover, we show that adapting the coding systematically to prior data can also capture the covariance structure and thus can be used to incorporate prior experimental knowledge. However, the observed effect is smaller than by means of variable selection which has been described in the previous section. To define an alternative to EGBLUP, we introduce the categorical marker effect model (CM) which assigns an independent effect to each possible state 0, 1 or 2 of each locus. This model is then extended to the categorical epistasis model (CE), which assigns for each pair of loci an independent effect to each combination of this pair. Considering marker k and l, each of the nine tuples $(k, l) \in \{0, 1, 2\}^2$ is modeled as having its own independent effect. We investigate theoretical properties of these models and show that they can improve predictive ability for some traits of the wheat and a published mouse data set.

The categorical epistasis model also provides a framework for our approaches of "Incorporating gene annotation into genomic prediction of complex traits". Here, we define the linear model on biological units, in this case on protein coding genes. We assign markers to genes and construct haploblocks for each set of markers characterizing one gene. For the definition of haploblocks, we use a previously published approach of restricting the maximal number of haplotypes in the population. However, this purely

data driven definition is applied to each set of markers characterizing a gene, thus fusing the biological prior knowledge with the data structure driven approach. We test different approaches based on categorical or numerical intra-locus additive haplotype models and with or without epistasis on the previously mentioned mouse data, a *Drosophila* data set, and a rice data set. Our results demonstrate a systematic improvement of predictive ability compared to GBLUP for the mouse and the rice data, but not for the *Drosophila* data set.

In the final discussion, the coding dependence of ridge regression methods is reviewed and the usefulness of the prediction of total genetic values in different breeding schemes is discussed. Both sections give answers but also raise other questions which are supposed to be addressed in the future. In the last part of the discussion, we review the results of this work in a broadened context, which is followed by an outlook on future work.

Introduction

Genomic prediction in breeding

Today, the concept of *genomic selection* (Haley and Visscher 1998; Meuwissen et al. 2001) is widely established in different variants in commercial breeding programs. The general underlying approach is to use broad genetic information of individual animals or plant lines to predict individual characteristics such as their *genetic values* for different traits. The broad genetic information can for instance be given by the state of *single nucleotides polymorphisms* (SNPs) at a set of base pair positions distributed across the whole genome. In particular, the *additive genetic value*, which is also called the *breeding value*, is often the quantity to be predicted in statistical models. The breeding value reflects the average improvement of the phenotype of the population under selection, when genes of the considered individual are randomly enriched in the population. For instance in dairy cattle, the breeding value of a sire with respect to traits related to milk production is defined by the corresponding phenotypes of its daughters in terms of the deviation from the mean performance of the reference population (Mrode 2014). In this particular case, this indirect measurement of a sire's value is necessary because the sires on which the selection decisions are made do not have an own phenotype. However, also when a direct measurement of the individual's phenotype is possible, a selection for high breeding values incorporating the performance of relatives in an appropriate way may exhibit a stronger response to selection than a program using the phenotype as selection criterion. Due to the incorporation of the information on the performance of relatives, the breeding value also reflects the effects of the genes of the respective individual in a changed genetic background and thus may be the relevant quantity for the improvement of the population over time (Fisher 1918; Henderson 1975; 1977; Falconer and Mackay 1996).

For young animals without an own performance or performance-tested offspring, the breeding values can also be predicted using a (pedigree based) relationship matrix relating older animals with highly accurate predictions of their breeding values to the young animals under consideration (Henderson 1975; 1977; Henderson and Quaas 1976). Genomic prediction offers here the option to use the realized relationship, which is inferred from the genomic data, instead of the expected relationship provided by the

pedigree (Habier et al. 2007; Hayes et al. 2009b). The resulting more accurate predictions of the breeding values of young animals are especially of interest, since selection decisions can thus be made in early stages of their lives, saving costs for maintaining animals which would not be used for breeding in the future, and more importantly saving time and consequently increasing the selection gain by shortening the generation interval (Schaeffer 2006).

Due to its convincing economic potential (and historical and structural reasons), the concept of genomic selection has been quickly implemented in animal, in particular in dairy cattle breeding (Hayes et al. 2009a; Reinhardt et al. 2009; Harris and Johnson 2010; Hayes and Goddard 2010; Mrode 2014; Gianola and Rosa 2015), but has also become an essential tool in plant breeding (Jannink et al. 2010; Heslot et al. 2012; Nakaya and Isobe 2012; Hayes et al. 2013; Newell and Jannink 2014; Heslot et al. 2015; Hickey et al. 2017). In tree breeding, shortening the generation interval and thus the breeding cycle can also be considered to be the most relevant aspect of genomic selection (Grattapaglia and Resende 2011; Resende et al. 2012; Isik 2014). In crops, the generation interval is naturally already much shorter, but the more accurate predictions of genetic values induced by the finely resolved relationship can generate a higher selection response than for instance compared to marker-assisted recurrent selection (Bernardo and Yu 2007), can shorten the breeding cycle (Heffner et al. 2009) by rearranging its modular structure with the new options provided by higher prediction accuracies, and can also reduce the expanses for field experiments by predicting the performance of a certain subset of lines instead of performing the corresponding trials. This reduction of field experiments will be especially of interest when for instance crosses of lines from different heterotic pools are supposed to be evaluated for their hybrid performance (Technow et al. 2012; Albrecht et al. 2014; Technow et al. 2014; Xu et al. 2014) and where consequently the number of potential crosses increases quadratically with the number of lines in the pools, but the genotyping costs increase only linearly (Zhao et al. 2015). Additionally to the essential questions of at which steps of a breeding program to use genomic prediction in which way, the questions arise which statistical model, and which method to determine the corresponding parameters should be used to predict the quantity of interest appropriately.

The additive marker effect model and epistasis

The standard reference model of quantitative genetics used to describe the effects of alleles on the phenotype is built upon a linear model in which gene effects are modeled additively (Falconer and Mackay 1996). In more detail, given a locus with alleles **a** and **A** in a diploid species, the common single locus model for the effect of the genotype is

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{M}\beta + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is the $n \times 1$ vector of n observations of the phenotype, and $\mathbf{1}_n$ an $n \times 1$ vector with each entry equal to 1. Moreover, μ is the fixed effect modeling the y-intercept, $\mathbf{M}$ the $n \times 1$ vector of marker states of the n individuals coded as 0 (**aa**), 1 (**aA** or **Aa**) or 2 (**AA**). The allele substitution effect of the locus under consideration is denoted by β, and $\boldsymbol{\epsilon}$ represents the $n \times 1$ independent and identically distributed (i.i.d.) random errors with mean 0. Expanding Eq.(1) to a model considering p loci simultaneously gives the same equation, but with $\mathbf{M}$ being an $n \times p$ matrix and $\boldsymbol{\beta}$ a $p \times 1$ vector. Here, μ could also be included in $\boldsymbol{\beta}$ and the vector $\mathbf{1}_n$ as a column in $\mathbf{M}$. However, Eq. (1) is the notation usually used in the mixed model approaches of quantitative genetics (Henderson 1975; Meuwissen et al. 2001; Mrode 2014), where the genetic effects are separated from the intercept and treated in a different way in the estimation / prediction process. The most popular method which is based on Eq. (1) and which uses certain additional assumptions on $\mu, \boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$, which we will explain in the next section, is called *genomic best linear unbiased prediction* (GBLUP). The genetic data can be any type of data, such as DNA sequence data, a subset of pre-selected base positions which are read by a genotyping array ("marker"), or other types of data such as the methylation state of certain nucleotides.

Independent of the method used to determine the parameters of Eq. (1), the model defines a framework in which each marker effect is independent of the states of the other markers. Having determined the additive effect $\hat{\beta}_i$ of locus i, the model predicts that a change at this position from marker value 0 to 1 or from 1 to 2 will have an effect

$\hat{\beta}_i$ on the genetic value, independently of the genetic background that is of the states of the other markers. This characteristic is fundamentally contrary to the principle of phenotype-forming biological systems in which interaction is omnipresent on the molecular biological level, and which may also translate to the statistical effect level (Sohail et al. 2017). This big discrepancy between the prevalent interaction in molecular biological mechanisms and the absence of interaction in statistical standard models used for prediction of complex traits, provides a motivation to consider "epistasis" models for the prediction of total genetic values. The terms "epistatic" and "epistasis" have been introduced more than one hundred years ago in the context of interaction effects of alleles on a phenotype (Bateson 1909; Fisher 1918) and different precise definitions of epistasis are not totally coinciding (Cordell 2002). In this work, any statistical model in which the effect of a change at a locus depends in any way on the state of any other locus, or which is based on a "non-additive" covariance model, will be called an epistasis model.

Several publications have shown that the incorporation of interaction effects (Ober et al. 2015; Forsberg et al. 2017) or the modeling of "non-additive" relationships (de los Campos et al. 2009; 2010; Ober et al. 2011; Crossa et al. 2010; Gianola et al. 2014; Morota and Gianola 2014) can improve predictive ability. The latter type of "non-additive" relationship models is based on *Reproducing kernel Hilbert space* (RKHS) regression models (Gianola and Van Kaam 2008) but can also be interpreted as a spatial statistics approach with general (isotropic) covariance functions. Among them, the Gaussian kernel has been the most popular in genetics and thus will be used at some points as a reference for the performance of other models.

The first epistasis model considered more deeply in this work, will be the extended GBLUP model (EGBLUP) (Su et al. 2012; Ober et al. 2015; Jiang and Reif 2015) which models interactions by a polynomial of degree two in the marker data (and which is based on some additional assumptions on $\mu, \boldsymbol{\beta}, \mathbf{h}$ and $\boldsymbol{\epsilon}$)

$$y_i = \mu + \mathbf{M}_{i,\bullet}\boldsymbol{\beta} + \sum_{k=1,\ldots,p;l>k} M_{i,k}M_{i,l}h_{k,l} + \epsilon_i. \tag{2}$$

Here, all variables are defined as before, $\mathbf{M}_{i,\bullet}$ denotes the i-th row of $\mathbf{M}$, that is the genomic data of individual i, and $h_{k,l}$ is the interaction effect of loci k and l. This model includes interactions of markers, since the effect of a change at locus k on the expected phenotype also depends on the genetic background, that is on the states of the other markers. If $\hat{h}_{k,l}$ is determined, the effect of a change at locus k will depend on the value of $M_{i,l}$. This model has the obvious interesting aspect of modeling the characteristic of biologically omnipresent interaction, but has the relevant statistical disadvantage of increasing the number of parameters drastically (which we will discuss more deeply in the next section). In the following, we will address the general problem of statistical genetics that is that the number of markers is usually much higher than the number of observations.

The p > n problem and mixed models

To predict $\hat{\mathbf{y}}$ of genotypes without measurements of their phenotypes, the parameters of Eq. (1), that is the intercept μ and the additive effects $\boldsymbol{\beta}$, have to be estimated from a data set. Having determined these coefficients as $\hat{\mu}$ and $\hat{\boldsymbol{\beta}}$, the genetic value of any individual from which genomic information is available can be predicted by Eq. (1). Here, the accuracy of the prediction will depend on how "similar" the individual is to the set from which the coefficients have been derived, and this similarity again is determined by the marker data, but also by the genetic architecture of the trait (Solberg et al. 2008; Shengqiang et al. 2009; Daetwyler et al. 2010).

In the case of having more individuals than parameters that is if $n > p + 1$ (and the corresponding relevant matrix having a rank of $p + 1$), we can use the well-known ordinary least squares (OLS) method. Its solution is defined by the minimization of the quadratic Euclidean distance of the vector of observations and the vector of the corresponding fitted values:

$$\begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}_{OLS} := \underset{(\mu,\boldsymbol{\beta})\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{M}_{i,\bullet}\boldsymbol{\beta} - \mu)^2 \tag{3}$$

$\mathbf{M}_{i,\bullet}$ denotes here the i-th row of $\mathbf{M}$, that is the genomic data of individual i. Pro-

vided that the respective matrices are invertible where necessary, the solution to the minimization problem of Eq. (3) is given by the well-known OLS estimate

$$\begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}_{OLS} = \left(\begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix}^t \begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix}^t \mathbf{y} \tag{4}$$

In problems of statistical genetics, we often deal with a high number of loci and a relatively low number of observations. In this situation of $p+1 > n$, the solution to Eq. (3) is not unique but a vector subspace of which each point minimizes Eq. (3) to zero. Due to this overfitting, that is fitted noise, the quality of predictions $\hat{\mathbf{y}}$ for genotypes which have not been used to estimate the parameter $(\hat{\mu}, \hat{\boldsymbol{\beta}})$ are usually poor. Moreover, if $p+1 > n$, the maximal rank of $\left(\begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix}^t \begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix} \right)$ of Eq. (4) is n, which implies that the matrix is not invertible and thus Eq. (4) is not defined.

A solution to the $p+1 > n$ problem is the use of a penalized regression method which minimizes

$$\begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}_{RR_\lambda} := \underset{(\mu,\boldsymbol{\beta})\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{M}_{i,\bullet}\boldsymbol{\beta} - \mu)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{5}$$

for a chosen penalty factor $\lambda > 0$. Note here that Eq. (5) is not introduced here as a pure ridge regression (RR), but already as a mixed model (MM) (Henderson 1975; 1977) which treats μ and $\boldsymbol{\beta}$ differently by not penalizing the size of μ. The term mixed model actually refers to a model in which some effects are treated as being fixed (but unknown) and others as coming from a random distribution. However, this prior assumption of being generated by a certain distribution introduces the penalty term of Eq. (5). Thus, Eq. (5) represents a simple version of a mixed model with an intercept μ as fixed effect and one class of random effects $\boldsymbol{\beta}$.

The corresponding solver is given by

$$\begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}_{RR_\lambda} = \left(\begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix}^t \begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix} + \lambda \begin{pmatrix} 0 & \mathbf{0}_p^t \\ \mathbf{0}_p & \mathbf{I}_p \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{1}_n & \mathbf{M} \end{pmatrix}^t \mathbf{y}. \tag{6}$$

Here, $\mathbf{0}_p$ denotes the $p \times 1$ vector of zeros and $\mathbf{I}_p$ the p-dimensional identity matrix. The effect of the introduction of the penalty term $\lambda \sum_{i=1}^{p} \beta_i^2$ is that for the minimization of Eq. (5), we have a trade-off between fitting the data optimally and shrinking the squared effects to 0. The method will only "decide" to increase the estimate $\hat{\beta}_j$, if the gain from improving the fit is greater than the penalized loss generated by the increase of $\hat{\beta}_j$.

In the context of quantitative genetics, the mixed model of Eq. (5) is usually built with the additional assumption of $\beta_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\beta^2)$ and $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ and called *ridge regression best linear unbiased prediction* (RRBLUP) or *genomic best linear unbiased prediction* (GBLUP) when it is rewritten with $\mathbf{g} := \mathbf{M}\boldsymbol{\beta}$ (Habier et al. 2007). The latter version is then usually formulated on the level of a *genomic relationship matrix* defining the covariance of $\mathbf{g}$ (VanRaden 2008). Considering the joint distribution of $(\mathbf{y}, \boldsymbol{\beta})$ in this setup and applying an approach of maximizing the joint density defines the penalty factor as $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\beta^2}$ (Henderson 1975). There are many alternative methods which have been applied to the additive effect model of Eq. (1), including Bayesian approaches with modified assumptions on the distribution of $\boldsymbol{\beta}$ (Gianola et al. 2009; Habier et al. 2011; Gianola 2013). Among all these approaches, RRBLUP is the one most commonly used and usually the reference when different methods are compared. In this work, all central methods will be based on variations of this ridge regression approach (for instance with interactions or with alternative predictor variables).

Statistical and computational problems arising with the consideration of epistasis

As previously mentioned, EGBLUP of Eq. (2) models all pairwise interactions which means that instead of only dealing with p additive effects which have to be determined in the additive model, we are facing a problem of having p additive effects and $\frac{p(p-1)}{2}$ interactions. For genomic prediction this is not necessarily a computational problem since, as we will see in this work, we can easily calculate the corresponding relationship matrices as Hadamard products of the additive models. However, with more variables and relatively less data points, the prior assumptions of the model may play a more

important role. For instance, when a model with pairwise interactions is applied to data generated by a large additive effect, the model may tend to distribute the main effect over the pairwise interactions to minimize the penalty of the ridge regression. Thus, when interactions are detected, this may not necessarily be the result of their extraordinary biological importance, but can also be a consequence of the combination of data structure and method. This circumstance may be a critical point for variable selection methods based on identifying interactions from given data. Whereas for *genome wide association studies* (GWAS) testing for main effects, different methods are quite well understood and standardized, the pool of methods for epistatic interactions is still developing. Thus, for instance for the detection of main effects, the single marker model is today often applied as a mixed-linear-model association (MLMA) method correcting for the genetic background by adding a random term with the relationship matrix defined by the remaining markers, that is all markers except for the one considered in the test (Yang et al. 2014). Whether this concept can be transferred analogously to tests for interactions by modeling a random term with the Hadamard product of the relationship matrix as covariance is not clear. The dynamics in the field of detecting interactions is also illustrated by the large number of publications of the last years addressing epistasis from a computational (González-Domínguez et al. 2016; Kässens et al. 2016; Jünger et al. 2017) or from a methodological or a conceptual perspective (Frost et al. 2016; Hung et al. 2016; Li et al. 2016; Sung et al. 2016; Uppu and Krishna 2016; Uppu et al. 2016; Wang et al. 2016; Wong et al. 2016; Xu et al. 2016). The main problem of having only a few observations of each state of a pair of markers, but performing many tests with collinear variables remains (Cordell 2009; Aschard 2016).

The relevance of epistasis in the formation of phenotypes and in breeding

In spite of the difficulties in detecting interactions statistically significantly and repeatably across different data sets, many studies of the last years have addressed the topic of detecting interactions in agriculturally relevant species (Carlborg et al. 2003; Chen et al. 2016; Ehrenreich 2017), but also from an epidemiological or a medical perspective in humans (Su et al. 2016; Galarza-Muñoz et al. 2017; Jing et al. 2017). Moreover, it has

been shown in many instances that epistasis models have the potential to improve the prediction of phenotypes (Ober et al. 2015; Zhang et al. 2015; Forsberg et al. 2017). In spite of this partial success, the relevance of epistasis for breeding is not clear. Criticism of the relevance of epistasis includes that the major genetic variance is additive (Hill et al. 2008), that parts of the epistatic variance will be converted into additive variance over time (Carlborg et al. 2006; Hill 2017), that beneficial combinations of markers may not be stable across generations due to recombination and segregation, and that consequently the quantity of main interest is the additive genetic value (regressed by Eq. (1)), and not the total genetic value. The fact that the major part of genetic variance can be expressed as additive variance may not necessarily be a reason to neglect epistasis in every context, since the overall variance does not directly reflect the genetic architecture (Huang and Mackay 2016), the small portion of epistatic variance can also be a result of the fitted main effects "obscuring" epistatic effects (Sackton and Hartl 2016), and because there is evidence that epistasis is not only present on the molecular biological mechanistic level, but that indeed statistical effects on the phenotype are not additive (Sohail et al. 2017; Tyler et al. 2016). Moreover, epistasis may also be responsible for maintaining the additive genetic variance of a population during long-term selection (Carlborg et al. 2006; Hallander and Waldmann 2007; Hill 2017) and may influence the long-term response to selection, which however will depend on details of the genetic architecture of the trait (Paixão and Barton 2016) and which also may be in part a result of not reducing the effective population size that drastically when alternatives to the additive breeding value are used as selection criterion (Esfandyari et al. 2017). Overall, the biological importance of interactions cannot be doubted, and there is also evidence for the statistical marker effects not being purely additive (Sohail et al. 2017; Tyler et al. 2016). Yet, it is not clear whether a more accurate prediction of the total genetic value can be used beneficially in breeding programs. There are good arguments illustrating that at least for standard line breeding, the total genetic value will not play an important role. However, models with a higher predictive ability for the phenotype may be in general interesting for crossbred and hybrid programs and also for predicting individual risks in personalized medicine.

Incorporating (external) prior knowledge into genomic prediction approaches

The key step in the development of genomic selection for complex traits was to discard the concept of building a statistical model only on a few genetic markers which have been identified to have a significant association with the trait, but to use instead all available genomic information provided by markers covering the whole genome (Meuwissen et al. 2001). Thus, a major part of the variables included in the model will in real data examples not show a statistically significant effect on the trait, since their effects are too small to be detected statistically significantly at the available sample sizes. This step can partly be interpreted as rejecting the concept of having to clarify the biological mechanism underlying the phenotypic response, or at least of having to demonstrate its statistical association with the predictor variables before a model for the prediction of a trait is built. However, there is plenty of biological knowledge available in public data bases, and an approach currently followed is to incorporate the knowledge not by restricting to a few variables, but by expanding the whole genome prediction approach with additional prior knowledge. Most of the approaches found in literature aiming at incorporating prior knowledge are based on single markers as units in an additive effect setup, but there is also some work on using prior knowledge to incorporate interactions. For instance, there are several publications addressing the topic of how to use results from genome wide association studies (GWAS) from external data bases or from the data under consideration by weighting markers (Su et al. 2012; 2014; Zhang et al. 2014; 2015; Veroneze et al. 2016) or by modeling selected markers as having a fixed effect (Spindel et al. 2016; Bian and Holland 2017; Lopes et al. 2017). Moreover, there are several publications which use external information of gene ontology categories to subdivide the markers into classes which are treated separately, most often in the sense that each class has its own variance component (Morota et al. 2014; Do et al. 2015; Abdollahi-Arpanahi et al. 2016; MacLeod et al. 2016; Sarup et al. 2016; Fang et al. 2017a;b). Other approaches extended the additive effect model by selected interactions (Ober et al. 2015; Forsberg et al. 2017). These approaches are certainly promising, since traits may for instance be related to different genes belonging to the same biological pathways (Edwards et al. 2015). This may also mean that using prior knowledge on

which genes belong to which pathway and giving markers associated to genes of the corresponding biochemical pathways a higher weight, may lead to an increase in predictive ability. Indeed, the cited references report an improvement of predictive ability in several instances, but a clear protocol of which type of information to use how to improve genomic prediction systematically does not exist so far. The topic is currently strongly being investigated and whether or not an approach will increase predictive ability also depends on the respective data set (Do et al. 2015).

The focus of the work on hand

In this work, we consider theoretical and practical aspects of different epistasis models and follow approaches to modify the statistical models towards biological mechanisms. The performances of different models are compared on simulated and real, publicly available data sets.

We first discuss properties of the extended GBLUP of Eq. (2). The chapter "Epistasis and covariance: How gene interaction translates into genomic relationship" deals with the question of how we can translate this marker and interaction effect based model into a genomic relationship matrix based approach. This reformulation of the statistical model facilitates some computational aspects and permits us to use the pairwise interaction model with a given GBLUP implementation, but with a relationship matrix based on the Hadamard product of the additive relationship matrix $\mathbf{G}$. We use a publicly available data set which offers phenotype records of the trait grain yield of 599 wheat lines, each grown under four different environmental conditions (Crossa et al. 2010) to test whether the predictive ability can be improved when all pairwise interactions are modeled. Moreover, we follow an approach of selecting pairwise interactions based on prior experimental data.

In the second chapter "On the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE)", we discuss the coding dependence of the predictive ability of EGBLUP which is caused by the use of a penalized regression approach to estimate / predict the effects of β_j and $h_{j,k}$. The

influence of the marker coding on predictive ability provides a motivation to consider other types of epistasis models. We start by defining the categorical marker effect model (CM) which assumes that each category of a marker state –that is 0, 1 or 2– has its own effect which is independent of the effect of the other categories. This model provides an alternative to GBLUP incorporating the freedom to model deviations from the allele dosage model implemented by multiplying the marker value with the corresponding marker effect in Eq. (1). However, CM does not model interactions of different loci. We extend CM then to the categorical epistasis model (CE) which models each state of a pair of markers as an independent variable coming from the same distribution. This model increases the number of variables further, but a corresponding relationship matrix approach is computationally not more demanding than the additive GBLUP. We compare the predictive ability of the different models on simulated data, on the wheat data which has already been used in the first chapter, and on a mouse data set (Valdar et al. 2006a;b).

In chapter "Incorporating gene annotation into genomic prediction of complex phenotypes", the allele dosage models and the categorical models are then transferred to haploblocks defined by gene positions as predictor variables. Here, we fuse an approach of defining haploblocks according to statistical considerations and characteristics of the data (Meuwissen et al. 2014) with the information on gene location, provided by public data bases. We test the different approaches on publicly available data sets.

In the final discussion, we review the coding-dependent performance of ridge regression approaches, and we discuss the results of a simulation of different breeding programs, thus considering the question whether an improved prediction of the total genetic value may be used somehow to generate a higher selection gain in breeding programs. In particular, we compare the selection gains of different programs of truncated selection with random mating, and of different targeted mating programs. We close with a general discussion of the results of this work.

Epistasis and covariance: How gene interaction translates into genomic relationship

Theor Appl Genet (2016) 129:963–976
DOI 10.1007/s00122-016-2675-5

ORIGINAL ARTICLE

Epistasis and covariance: how gene interaction translates into genomic relationship

Johannes W. R. Martini[1] · Valentin Wimmer[2] · Malena Erbe[1,3] · Henner Simianer[1]

Received: 2 June 2015 / Accepted: 16 January 2016 / Published online: 16 February 2016

Abstract
Key message **Models based on additive marker effects and on epistatic interactions can be translated into genomic relationship models. This equivalence allows to perform predictions based on complex gene interaction models and reduces computational effort significantly.**
Abstract In the theory of genome-assisted prediction, the equivalence of a linear model based on independent and identically normally distributed marker effects and a model based on multivariate Gaussian distributed breeding values with genomic relationship as covariance matrix is well known. In this work, we demonstrate equivalences of marker effect models incorporating epistatic interactions and corresponding mixed models based on relationship matrices and show how to exploit these equivalences computationally for genome-assisted prediction. In particular, we show how models with epistatic interactions of higher order (e.g., three-factor interactions) translate into linear models with certain covariance matrices and demonstrate how to construct epistatic relationship matrices for the linear mixed model, if we restrict the model to interactions defined a priori. We illustrate the practical relevance of our results with a publicly available data set on grain yield of wheat lines growing in four different environments. For this purpose, we select important interactions in one environment and use this knowledge on the network of interactions to increase predictive ability of grain yield under other environmental conditions. Our results provide a guide for building relationship matrices based on knowledge on the structure of trait-related gene networks.

Communicated by J. Crossa.

✉ Johannes W. R. Martini
jmartin2@gwdg.de

[1] Department of Animal Sciences, Animal Breeding and Genetics Group, Georg-August University, Göttingen, Germany

[2] KWS SAAT SE, Einbeck, Germany

[3] Institute of Animal Breeding, Bavarian State Research Centre for Agriculture, Grub, Germany

Introduction

In the last decades, newly developed methods and tools from other quickly expanding scientific fields such as molecular biology, genetics and statistics have been introduced into breeding procedures. Among these new breeding concepts, *genomic selection* (Meuwissen et al. 2001), which combines the availability of genetic data from individual animals or plant lines, in the following generally referred to as "*genotype*", with appropriate statistical evaluation methods for large data sets, has the potential to accelerate breeding progress and to reduce its costs at the same time. Instead of raising all animals and measuring their performance or growing all variants of a crop, the mean phenotype of some genotypes can be predicted based on genetic information. Assuming that the prediction method is reliable, this procedure can partly substitute expensive experiments and / or save time, thus increasing selection gain. Prediction methods based on pedigree data have already been used since the 1970s (Henderson 1975, 1984; Henderson and Quaas 1976) and their importance has further increased after the availability of genomic data in form of high throughput single-nucleotide polymorphisms (SNPs) panels has allowed for using estimated realized, instead of pedigree-based expected relationships (Hayes et al. 2009; Habier et al. 2007; Piepho et al. 2008; Gianola and Rosa

2015). Moreover, with the increased availability of substantial genomic datasets, the approaches for genome-assisted prediction have diversified (see e.g., Gianola and Rosa 2015; Morota and Gianola 2014).

Among these different approaches, the correspondence between an additive marker effect model and a model with a certain kind of genomic relationship matrix is of central importance in the theory of genome-assisted prediction (Habier et al. 2007). In more detail, a widely used linear regression model is given by

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is the $n \times 1$ vector of phenotypes of the n genotypes, $\mathbf{1}$ is the $n \times 1$ vector with each entry 1, μ is a fixed effect, and $\mathbf{M}$ is the $n \times p$ matrix giving the p marker values of the n genotypes. Moreover, we assume here that the entries of the $p \times 1$ vector of unknown effects $\boldsymbol{\beta}$ and the $n \times 1$ error vector $\boldsymbol{\epsilon}$ are independent and Gaussian distributed with certain variance components $\beta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\beta^2)$ and $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. Considering the numbers of an allele of a diploid organism, each $M_{i,j}$ (the value of the j-th marker of genotype i) can be coded for instance as $\{0, 1, 2\}$ or $\{-1, 0, 1\}$, or rescaled by subtracting the population mean of each of the p marker values from the respective column of the matrix $\mathbf{M}$ (VanRaden 2008). Having predicted the marker effects as $\hat{\boldsymbol{\beta}}$ and having estimated the fixed effect as $\hat{\mu}$, the predicted average phenotype of a genotype is then given by $\hat{y}_i = \hat{\mu} + \mathbf{M}_{i,\bullet}\hat{\boldsymbol{\beta}}$, where $\mathbf{M}_{i,\bullet}$ denotes the i-th row of $\mathbf{M}$. The assumptions made on $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ give the equivalence to the model

$$y = \mathbf{1}\mu + \mathbf{g} + \boldsymbol{\epsilon} \tag{2}$$

with $\mathbf{g} \sim \mathcal{N}_n(0, \sigma_\beta^2 \mathbf{M}\mathbf{M}')$, $\mathbf{M}\mathbf{M}'$ being the genomic relationship, and all other variables as defined before (Habier et al. 2007). In the following, we will call $\mathbf{g}$ the genetic values of the genotypes. The prediction method based on Eq. (2) or its corresponding analog with marker effects as variables (Eq. (1)) is known as *genomic best linear unbiased prediction* (GBLUP) which is currently the most widely used method and a reference for any other approach (Gianola and Rosa 2015). The advantage of Eq. (2) compared to Eq. (1) is of computational nature if $p \gg n$: instead of solving a system with p variables (the marker effects), we deal with a reduced system in which only the n genetic values have to be determined. However, regarding Eq. (1) in which the genetic value is the sum of the product of the marker codes and of their respective effects, makes obvious that the model is based on statistical considerations (a direct linear regression approach with the marker values as predictor variables) and does not capture any kind of interaction which may be present in biochemical pathways that produce the phenotype. Consequently, the question arises as to how to build statistical models that can incorporate interactions between loci. Among many other approaches, a straightforward extension of Eq. (1) is the introduction of products of two genotype codes:

$$y_i = \mu + \sum_{j=1}^{p} M_{i,j}\beta_j + \sum_{k=1}^{p}\sum_{j=k+1}^{p} M_{i,j}M_{i,k}h_{j,k} + \epsilon_i, \tag{3}$$

where $h_{j,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_h^2)$ is the pairwise interaction effect of markers j and k ($j > k$) and all other terms as defined before. We will call Eq. (3) the *pair epistasis model*. This model incorporates additive marker effects and pairwise interactions defined by the regression coefficients $h_{j,k}$ of the products $M_{i,j}M_{i,k}$ of different markers. The impact of the incorporation of the products $M_{i,j}M_{i,k}$ on the phenotype y_i might be clearer when marker codes are either 0 or 1: the effect $h_{j,k}$ is only present if both markers are of value 1. Note here that Eq. (3) is a polynomial function of degree two in the marker variables but it is still linear in the regression coefficients. As a variant of the *pair epistasis model*, we consider the *pair epistasis and dominance model*

$$y_i = \mu + \sum_{j=1}^{p} M_{i,j}\beta_j + \sum_{k=1}^{p}\sum_{j=1}^{p} M_{i,j}M_{i,k}h_{j,k} + \epsilon_i. \tag{4}$$

The main difference between Eqs. (3) and (4) is the fact that we also model marker effects depending quadratically on the value of the marker codes in the latter equation ($j = k$ in Eq. (4)). The model thus can incorporate dominance effects, since this kind of genetic interaction produces a nonlinear impact of the marker code on the expected phenotype. Fig. 1 illustrates the differences between the two epistasis models. Moreover, for a comparison of this parameterization of dominance to the classical one used by Falconer (Falconer and Mackay 1996) see Fig. 4 and the "Supporting information". A second difference between Eqs. (3) and (4) is that in the *pair epistasis and dominance model*, the effect of an interaction between marker k and j is incorporated by the sum $h_{j,k} + h_{k,j}$. Assuming $h_{j,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_h^2)$ consequently means that the "effective" variance is twice the variance of $h_{r,r}$. This variant of the epistasis model is of importance since it serves as a reference. As we will show later, Eq. (4) is a marker effect model where the corresponding relationship matrix is the Hadamard product of the usual genomic relationship matrix with itself. This illustrates a useful way to calculate relationship matrices in other variants of the epistasis model easily by expressing its deviation from this reference model. Moreover, this relationship matrix has also been used as an approximation to the model of Eq. (3) in asymptotic considerations (Jiang and Reif 2015).

Even though it has been shown that an additive model explained a major part of genetic variance in different

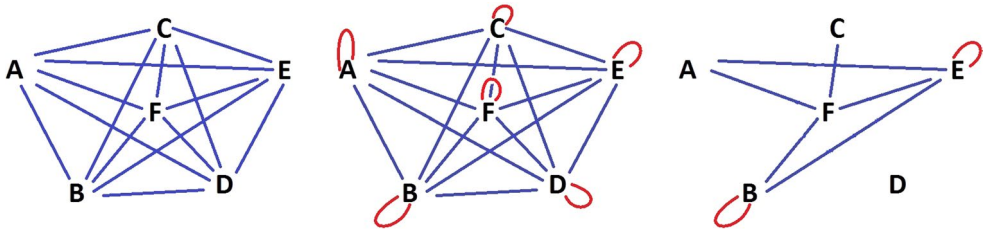

Fig. 1 Illustration of the differences between the *pair epistasis model* (*left* scheme) and the *pair epistasis and dominance model* (*middle* scheme). In the latter, the interactions are modeled as the sum of two variables and dominance is included by an interaction of each locus with itself. The *right* scheme illustrates a subnetwork which might fit the underlying biology of the trait best

data sets (Hill et al. 2008), approaches incorporating interactions have been viewed as potentially beneficial for genome-assisted prediction (Hu et al. 2011; Mackay 2013; Wang et al. 2012; Wittenburg et al. 2011). In this work, we consider different variants of the epistasis models and present analogs using relationship matrices, which offers computational advantages as the squared number of markers is usually much larger than the number of genotypes. In particular, we investigate theoretical aspects which lay the ground for building relationship matrices based on knowledge on trait-related gene networks and demonstrate gains in predictive ability when the model is reduced to certain subnetworks. Finally, we discuss connections of our results to other work on the relation between epistasis, dominance and relationship matrices and to classical results of quantitative genetics.

Material and methods

Data used to compare different models

We used the wheat data set described by Crossa et al. (2010) and found in the R-package BGLR (Core Team 2014; de los Campos and Perez-Rodriguez 2014). We chose this data set because these authors compared Reproducing Kernel Hilbert Space (RKHS) methods to GBLUP and reported improved predictive ability, suggesting the presence of non-additive effects that might also be captured by epistasis models (3,4). The trait in this data set is grain yield recorded for 599 different CIMMYT inbred lines grown under four different environmental conditions. The genotypes are typed with 1279 DArT markers coded as 0 or 1, indicating the absence (0) or presence (1) of the respective marker. The phenotypic correlations between the records in the different environments are

$$\begin{pmatrix} & \text{Env 2} & \text{Env 3} & \text{Env 4} \\ \text{Env 1} & -0.020 & -0.193 & -0.123 \\ \text{Env 2} & & 0.661 & 0.411 \\ \text{Env 3} & & & 0.388 \end{pmatrix}.$$

Prediction and evaluation of predictive ability of different models

To evaluate the different prediction methods, we used the following approach: Out of all 599 wheat lines, 60 lines were chosen randomly to be a test set (using function `sample()` of R (version 3.1.1)). The variance components σ^2_ϵ and σ^2_β (or σ^2_h) were estimated from the training set (539 lines) using version 1.3.14 of the `R` package `regress`. The corresponding function finds an extreme of the likelihood with the Newton-Raphson algorithm (Clifford and McCullagh 2006, 2014). The relationship matrix was calculated only for the genotypes of the training set and all entries were divided by the maximum entry to standardize the matrix, to guarantee numerical stability. Having estimated the variance components, we used them together with the full relationship matrix relating all lines, rescaled by the same factor as the one for the training set, in prediction Eq. (5). This procedure was repeated 200 times at random. The correlation r between observed and predicted phenotypes in the test set and its Fisher's z-transformation were used as an indicator of predictive ability.

The prediction equation

$$\begin{pmatrix} \hat{\mathbf{g}}_{\text{train}} \\ \hat{\mathbf{g}}_{\text{test}} \end{pmatrix} = \left[\mathbf{T}_{\text{train}} - s^{-1} \begin{pmatrix} \mathbf{J}_{s\times s} & 0 \\ 0 & 0 \end{pmatrix} + \sigma^2_\epsilon \left(\frac{1}{\sigma^2_\beta} \mathbf{G}^{-1} \right) \right]^{-1} \times \left(\begin{pmatrix} \mathbf{y}_{\text{train}} \\ 0 \end{pmatrix} - \begin{pmatrix} \mathbf{1}_s \bar{y}_{\text{train}} \\ 0 \end{pmatrix} \right). \quad (5)$$

This prediction equation is based on the assumption of known variance components σ^2_β and σ^2_ϵ and unknown μ. The fixed effect μ is implicitly estimated (for a derivation of Eq. (5) see the "Supporting information"). We use the notation $\mathbf{y}_{\text{train}}$ for the phenotypes of the lines in the training set, which are used to predict the genetic values of all genotypes. The matrix $\mathbf{G}$ is the genomic relationship matrix, $\hat{\mathbf{g}}_i$ ($i \in \text{test}, \text{train}$) are the predicted genetic values of the respective set, s is the number of genotypes in the training set (here $s = 539$), $\mathbf{1}_s$ the vector of length s with each entry 1, $\mathbf{J}_{s\times s}$ the analogous $s \times s$ matrix with each entry equal to 1, and $\bar{y}_{\text{train}}$ the unweighted mean of

training set phenotypes, which is not in general equal to $\hat{\mu}$. Moreover, $\mathbf{T}_{\text{train}}$ denotes a diagonal matrix of dimension $n = 599$ with zeros on the diagonal at the positions of the test set genotypes and ones for genotypes of the training set. The prediction of the phenotypes of the test set is $\hat{\mathbf{y}}_{\text{test}} = \mathbf{1}_{n-s}\hat{\mu} + \hat{\mathbf{g}}_{\text{test}}$. Note here that $\mathbf{G}$ can be replaced by any other relationship matrix.

Variable selection and prediction based on a subsystem of interactions

We used our theoretical results for building relationship matrices based on interaction subnetworks. To select important variables, we estimated interaction effects based on the 599 genotypes in one environment, and used this information to predict phenotypes in the other environments. We calculated the effects for instance in environment 1, and reduced the full model by removing the 5 % of the pariwise interactions that had the smallest absolute value. We re-estimated the interactions with the reduced model, to quantify the importance of the remaining variables. We continued this procedure in 5 % steps until at most 90 % of the variables were removed. We then used these submodels for prediction of phenotypes in the other environments.

Results

We start by presenting the statistical equivalence of marker effect models (Eqs. (3) and (4)) and their corresponding models based on relationship (covariance) matrices. The derivations of some results are based on the assumption of non-singularity of the respective covariance matrices which is required for the existence of the density of the multivariate Gaussian distribution.

Equivalences of the epistasis effect models and linear mixed models with two random terms

An equivalent model to the pair epistasis model. The model described by Eq. (3) is statistically equivalent to the model

$$y = \mathbf{1}\mu + \mathbf{g} + \mathbf{g}_2 + \boldsymbol{\epsilon} \tag{6}$$

with the n-dimensional Gaussian terms $\mathbf{g} \sim \mathcal{N}_n(0, \sigma^2_\beta \mathbf{G})$ and $\mathbf{g}_2 \sim \mathcal{N}_n(0, \sigma^2_h \mathbf{H})$; $\mathbf{1}$, μ and $\boldsymbol{\epsilon}$ as previously defined, and $\{\mathbf{g}, \mathbf{g}_2, \boldsymbol{\epsilon}\}$ stochastically independent. Moreover, the relationship matrix based on additive effects is $\mathbf{G} = \mathbf{M}\mathbf{M}'$ and the relationship matrix $\mathbf{H}$ for the epistatic effects is given by

$$H_{i,l} = \sum_{k=1,j>k}^{p} M_{i,k}M_{i,j}M_{l,k}M_{l,j} = \sum_{k=1}^{p}\left(M_{i,k}M_{l,k}\sum_{j>k}^{p} M_{i,j}M_{l,j}\right) \tag{7}$$

Recall here that $\mathbf{M}$ is the $n \times p$ genotype matrix. The derivation of this statement can be found in the "Supporting information".

An equivalent model to the pair epistasis and dominance model. Eq. (4) is statistically equivalent to the model of Eq. (6) with $\mathbf{g} \sim \mathcal{N}_n(0, \sigma^2_\beta \mathbf{G})$, $\mathbf{g}_2 \sim \mathcal{N}_n(0, \sigma^2_h \tilde{\mathbf{H}})$ and $\mathbf{1}$, μ and $\boldsymbol{\epsilon}$ as previously defined and $\{\mathbf{g}, \mathbf{g}_2, \boldsymbol{\epsilon}\}$ stochastically independent. Moreover, the relationship matrix $\tilde{\mathbf{H}}$ is given by

$$\tilde{H}_{i,l} = \sum_{j,k=1}^{p} M_{i,j}M_{i,k}M_{l,j}M_{l,k} = \left(\sum_{k=1}^{p} M_{i,k}M_{l,k}\right)^2 = \left(G_{i,l}\right)^2 = (\mathbf{G}\circ\mathbf{G})_{i,l} \tag{8}$$

with $\circ$ denoting the Hadamard product (the matrices of equal size are multiplied component-wise which means here that $\mathbf{G}\circ\mathbf{G}$ translates to matrix $\mathbf{G}$ with each entry squared). The derivation of Eq. (8) can also be found in the "Supporting information". Again, Eqs. (7, 8) allow us to use genome-assisted prediction methods incorporating epistatic effects on the level of relationship matrices, which is of computational advantage if $p^2 \gg n$.

Relationship between the different covariance matrices

Equation (8) shows that $\tilde{\mathbf{H}} = \mathbf{G}\circ\mathbf{G}$. The following equation describes the relationship between the covariance matrices $\mathbf{H}$ and $\tilde{\mathbf{H}}$:

$$\mathbf{H} = 0.5\tilde{\mathbf{H}} - 0.5(\mathbf{M}\circ\mathbf{M})(\mathbf{M}\circ\mathbf{M})' = \tag{9}$$

$$0.5\underbrace{(\mathbf{M}\mathbf{M}'\circ\mathbf{M}\mathbf{M}')}_{=\mathbf{G}\circ\mathbf{G}} - 0.5(\mathbf{M}\circ\mathbf{M})(\mathbf{M}\circ\mathbf{M})'.$$

See the "Supporting information" for the proof of Eq. (9). In the special case of just two possible marker values for each locus, as in haploid organisms or fully homozygous diploid individuals, the genotypes can be coded as 0 and 1. Then $M^2_{i,k} = M_{i,k}$ and Eq. (9) is the linear combination

$$\mathbf{H} = 0.5\tilde{\mathbf{H}} - 0.5\mathbf{G}. \tag{10}$$

Equation (9) can also be found in the work of Jiang and Reif (2015) who argue asymptotically that $\tilde{\mathbf{H}}$ can also be used as an approximation to the covariance matrix of the *pair epistasis model* without dominance described by Eq. (3).

Inferring the additive and epistatic marker effects

To see how the additive effects can be inferred when prediction is based on Eq. (6), we consider the model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{M}\boldsymbol{\beta} + \mathbf{g}_2 + \boldsymbol{\epsilon} \tag{11}$$

which represents an "intermediate" between the effect models Eqs. (3, 4) and (6). Assuming that μ, $\mathbf{y}$ and the epistatic

genetic values $\mathbf{g}_2$ are known, we can maximize the conditional density $f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{g}_2, \mu)$, which gives as solution the ridge-regression equation for prediction of additive marker effects

$$\hat{\boldsymbol{\beta}} = \left(\lambda \mathbf{I}_p + \mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'(\mathbf{y} - \mathbf{1}\mu - \mathbf{g}_2) \tag{12}$$

with $\lambda = \frac{\sigma^2_\epsilon}{\sigma^2_{\hat{\beta}}}$. This is equivalent to

$$\hat{\boldsymbol{\beta}} = \mathbf{M}'\left(\lambda \mathbf{I}_n + \mathbf{M}\mathbf{M}'\right)^{-1}(\mathbf{y} - \mathbf{1}\mu - \mathbf{g}_2) \tag{13}$$

The computational advantage of Eq. (13) over Eq. (12) is that the matrix which has to be inverted is of size $n \times n$, and not of size $p \times p$ (the number of genotypes n is often much smaller than the number of markers p; for the derivation of the dual equations for ridge regression see Shawe-Taylor and Cristianini 2004).

With $\mathbf{h}$ denoting the vector of interactions $h_{j,k}$, an analogous procedure for maximizing the joint density gives

$$\hat{\mathbf{h}} = \left(\lambda_2 \mathbf{I}_p + \mathbf{N}'\mathbf{N}\right)^{-1}\mathbf{N}'(\mathbf{y} - \mathbf{1}\mu - \mathbf{g}), \tag{14}$$

with $\lambda_2 = \frac{\sigma^2_\epsilon}{\sigma^2_h}$ which is equivalent to

$$\hat{\mathbf{h}} = \mathbf{N}'\left(\lambda_2 \mathbf{I}_n + \mathbf{N}\mathbf{N}'\right)^{-1}(\mathbf{y} - \mathbf{1}\mu - \mathbf{g}). \tag{15}$$

$\mathbf{N}$ denotes here the initial matrix of the *pair epistasis model* with n rows and $0.5p(p-1)$ columns. The i-th row is

$$\mathbf{N}_{\mathbf{i},\bullet} = \left(M_{i,1}M_{i,2}, M_{i,1}M_{i,3}, ..., M_{i,1}M_{i,p}, M_{i,2}M_{i,3}, ..., M_{i,p-1}M_{i,p}\right).$$

For the *pair epistasis and dominance model*, $\mathbf{N}$ has to be substituted by the $n \times p^2$ matrix $\mathbf{Q}$ with

$$\begin{aligned}\mathbf{Q}_{i,\bullet} &= \left(M_{i,1}M_{i,1}, M_{i,1}M_{i,2}, ..., M_{i,1}M_{i,p}, M_{i,2}M_{i,1}, ..., M_{i,p}M_{i,p}\right)\\ &= \mathbf{M}_{i,\bullet} \otimes \mathbf{M}_{i,\bullet}\end{aligned}$$

where $\otimes$ denotes the Kronecker product. Also here, for the epistatic effects, we have the advantage of the dual equation that $\mathbf{N}\mathbf{N}'$ is an $n \times n$ matrix, whereas $\mathbf{N}'\mathbf{N}$ is a $0.5p(p-1) \times 0.5p(p-1)$ matrix (note that $\mathbf{M}\mathbf{M}' = \mathbf{G}$, $\mathbf{N}\mathbf{N}' = \mathbf{H}$ and $\mathbf{Q}\mathbf{Q}' = \tilde{\mathbf{H}}$). At the level of individual interactions, Eq. (15) means that in the *pair epistasis and dominance model*, $h_{j,k}$ is predicted by

$$\hat{h}_{j,k} = (\mathbf{M}_{\bullet,j} \circ \mathbf{M}_{\bullet,k})'\left(\lambda_2 \mathbf{I}_n + \mathbf{Q}\mathbf{Q}'\right)^{-1}(\mathbf{y} - \mathbf{1}\mu - \mathbf{g}).$$

In particular this implies that the predictions $\hat{h}_{j,k}$ and $\hat{h}_{k,j}$ are identical.

Equivalent linear models for higher order polynomial functions of the markers

Equation (4) is a polynomial of degree two in the genotypes, its analog of Eq. (6) has two random effects and the corresponding covariance matrix of the epistatic terms is $\sigma^2_h \mathbf{G} \circ \mathbf{G}$. Here, we show how to generalize this statement to any degree D of epistatic interaction:

For $j \leq D$, let $\boldsymbol{\kappa} = (k_1, ..., k_j) \in \{1, ..., p\}^j$ be a j-dimensional vector with entries in $\{1, ..., p\}$ (the entries are not necessarily different). Moreover, let $M_{i,\boldsymbol{\kappa}} = M_{i,k_1}M_{i,k_2}...M_{i,k_j}$ denote the product of the marker values at loci $k_1, ...k_j$ of genotype i. Let

$$f(\mathbf{M}_{i,\bullet}) = \sum_{j=1}^{D}\left(\sum_{\boldsymbol{\kappa}\in\{1,...,p\}^j} \beta_{\boldsymbol{\kappa},j} M_{i,\boldsymbol{\kappa}}\right)$$

be a polynomial function of the marker data and for each degree $j \in \{1, ..., D\}$, let the coefficients $\beta_{\boldsymbol{\kappa},j}$ be random and distributed as $\beta_{\boldsymbol{\kappa},j} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2_j)$ (and all random terms independent).

Then the model $y_i = \mu + f(\mathbf{M}_{i,\bullet}) + \epsilon_i$ is statistically equivalent to

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{g} + \mathbf{g}_2 + \mathbf{g}_3 ... + \mathbf{g}_D + \boldsymbol{\epsilon} \tag{16}$$

with $\mathbf{g}_j \sim \mathcal{N}_n(0, \sigma^2_j \mathbf{G}^j)$ where $\mathbf{G}^j$ denotes the jth power with respect to the Hadamard product and $\mathbf{G} = \mathbf{M}\mathbf{M}'$ (see the "Supporting information" for the proof of this statement).

In particular, extending the model to all three-factor interactions $M_{i,k_1}M_{i,k_2}M_{i,k_3}$ translates into adding another summand $\mathbf{g}_3 \sim \mathcal{N}_n(0, \sigma^2_3 \mathbf{G}^3)$ with its own variance component σ^2_3 to Eq. (6). An important point in the structure of polynomial f is that a summand is added for each $\boldsymbol{\kappa} \in \{0, ..., p\}^j$. This means that, for instance, for three-factor interactions, the product $M_{i,1}M_{i,2}M_{i,3}$ of the polynomial function f has coefficient $\beta_{(1,2,3),3} + \beta_{(1,3,2),3} + \beta_{(2,1,3),3} + \beta_{(2,3,1),3} + \beta_{(3,1,2),3} + \beta_{(3,2,1),3}$. This is analogous to the pair epistasis and dominance model where interactions were incorporated as the sum of $h_{l,k}$ and $h_{k,l}$. Moreover, in the case of $j = 3$, this polynomial includes the factor $M^3_{i,j}$ for each locus j. There is no obvious meaningful interpretation of these higher order terms in a diploid organism, but such terms are implicitly included if relationship matrix $\mathbf{G}^3$ is used. For this reason, note here that if we want to approximate a model that only incorporates products $M_{i,k_1} \cdots M_{i,k_j}$ of different loci by the Hadamard power $\mathbf{G}^j$, the quality of the approximation will deteriorate with increasing j, since the fraction of variables that are incorporated in $\mathbf{G}^j$ but which are not present in the model we want to approximate is increasing with j. For modeling interactions of higher order, products that are not required should be subtracted from the Hadamard power of $\mathbf{G}$. In particular, this circumstance will reduce the quality of the approximation, when the exponential power series is used with Hadamard products of $\mathbf{G}$ to approximate a model with interactions only between different loci, which has been done to connect the epistasis effect model to an RKHS approach with the Gaussian kernel (Jiang and

Reif 2015). For multi-kernel approaches similar to Eq. (16) see also Morota et al. (2013) and Abdollahi-Arpanahi et al. (2014).

Restriction to a specified subset of interactions

The epistasis model described by Eq. (4) incorporates all possible pairwise interactions and dominance terms for each marker. This guarantees that interactions are included, but also inflates model complexity by potentially accounting for a huge number of unimportant variables. Thus, the question arises of how a subset of relevant pairwise interactions can be incorporated into the model (Fig. 1 illustrates a full network and a reduced subnetwork). Note, that by reducing the marker matrix to the entries which will be involved in the model, and then calculating the covariance matrix according to Eqs. (7, 8), means restricting the model to a Cartesian product. For instance, if we assume that marker 1 has a relevant dominance effect and that markers 2 and 3 interact, reducing $\mathbf{M}$ to the columns 1, 2, 3 and then applying Eq. (8) means that we build a model with pairwise interactions $h_{1,1}, h_{1,2}, h_{1,3}, h_{2,1}, h_{2,2}, h_{2,3}, h_{3,1}, h_{3,2}, h_{3,3}$ instead of only incorporating the effects $h_{1,1}, h_{2,3}$. Thus, the following result may be helpful for improving prediction by building the covariance matrix as the sum of covariances from pairwise interactions only.

If markers k and j interact, this is captured by the term $M_{i,k}M_{i,j}h_{k,j}$ in Eq. (3). Then, the corresponding covariance matrix is

$$\left(\mathbf{M}_{\bullet,k}\mathbf{M}'_{\bullet,k}\right) \circ \left(\mathbf{M}_{\bullet,j}\mathbf{M}'_{\bullet,j}\right)\sigma_h^2 \tag{17}$$

with $M_{\bullet,k}$ denoting the k-th column of $\mathbf{M}$ and which contains the marker codes of all genotypes at position k. The full covariance matrix corresponding to the model with all relevant interactions is then the sum of all pairwise interaction covariance matrices as in of Eq. (17) (since $h_{k,j}$ are assumed to be *i.i.d.* random variables). See the "Supporting information" for the derivation of this statement. This procedure has already been used by Ober et al. (2015).

An example with wheat data

Our examples focus on the calculation of the interactions with Eq. (15), and on a case of variable selection and building network based relationship matrices using Eq. (17). Recall that the wheat data set we use is based on presence / absence markers coded as 0 and 1. Thus, the value of a marker remains unchanged if it is squared, and the effect of the variants 0, 1 can be fully described by a linear function or by a polynomial of degree two without a linear term. This means using the model of Eq. (4) with additive and epistatic effects incorporates the additive effects of each marker twice (with a different variance component), since the additive effects are also incorporated as the dominance terms in Eq. (4). Therefore, for this data, we used the variant

$$y_i = \mu + \sum_{k=1}^{p}\sum_{j=k}^{p} M_{i,j}M_{i,k}h_{j,k} + \epsilon_i, \tag{18}$$

of the epistasis marker effect model, which incorporates the additive effects and all pairwise interactions modeled as a single term. The corresponding relationship matrix can be derived in the same way as Eqs. (9, 10) and it is given by $\bar{\mathbf{H}} = 0.5(\mathbf{G} \circ \mathbf{G}) + 0.5\mathbf{G}$.

Calculating the epistatic interactions

As an example, we calculate the epistatic interactions in the four environments. We can use Eq. (15), with $\mathbf{g} = \mathbf{0}$, $\bar{\mathbf{H}}$ and μ, σ_h^2, and σ_ϵ^2 estimated using `regress()`:

$$\hat{h}_{j,k} = (\mathbf{M}_{\bullet,j} \circ \mathbf{M}_{\bullet,k})'\left(\hat{\lambda}_2\mathbf{I}_n + \bar{\mathbf{H}}\right)^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu})$$

We calculated the interactions of all marker pairs in the four different environments. Figure 2 shows the 0.01 % interactions with the largest absolute values in each environment. While the patterns observed in the different environments are diverse, there are some general hints: in many cases, single regions appear to interact with a number of other regions, which would coincide with an assumed function of a regulatory "hub" of such regions. Some of these patterns appear in more than one environment. While it is beyond the scope of our paper to provide a biological interpretation of the observed structures, such results could be used as a staring point to detect complex genomic interactions underlying the phenotype studied. Note here that, among the interactions with the largest absolute values, there are hardly intrachromosomal interactions (the annotation of the markers to chromosomes can be found in Table S1 of the Supporting Information of Crossa et al. (2010)). This may be attributed to linkage disequilibrium, and of a tendency of the model to pick existing interaction between linked loci as additive effects at one or both loci.

Variable selection and prediction based on the reduced model

We tested whether characteristics of the models are preserved across environments. For this purposes, we determined subnetworks in each environment by neglecting interactions with smallest absolute values in 5 % steps from the model (until 90 % of the variables were removed). After each reduction step we recalculated the interactions in

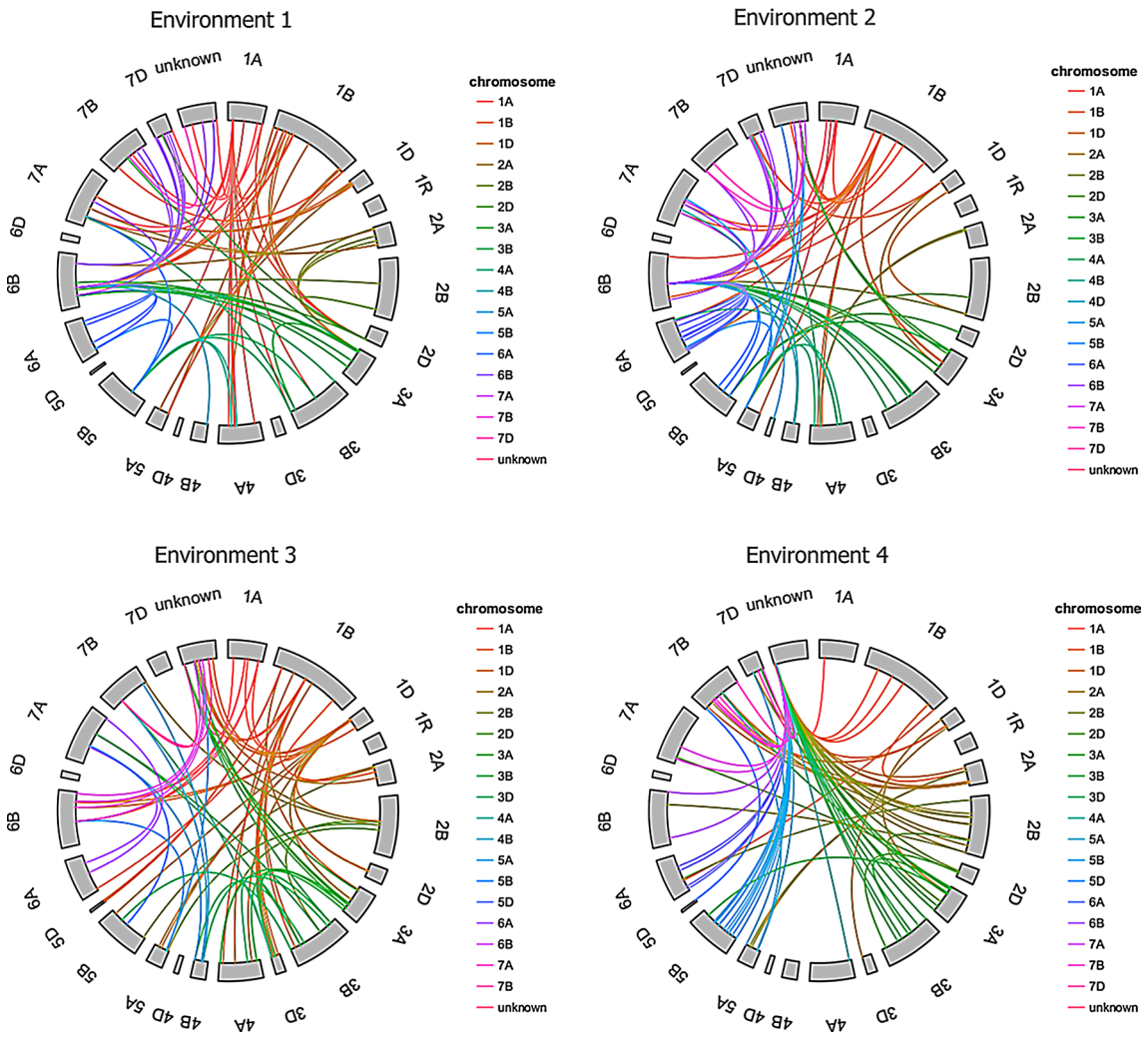

Fig. 2 The 0.01 % most important (highest absolute values) pairwise interactions of the wheat data set, calculated with Eq. (15) for the four different environments (positions of the markers according to Crossa et al. 2010)

the remaining network to allow the model to adapt before choosing the next set of interactions to abandon. We then used these models for prediction in the other environments with 200 randomly drawn test sets. The results are summarized in Table 1 and illustrated in Fig. 3 as Fisher's z-transformed correlation coefficient, which spreads the correlation interval $[-1, 1]$ onto the real line. Considering Fig. 3, we see that neither the information inferred from the data of environment 1 can improve prediction of phenotypes in the other environments, nor does the prediction within environment 1 benefit appreciably from the usage of information from grain yield in the other environments. This result may be expected, since phenotypes of other environments are negatively correlated with phenotypes in environment 1. However, it is remarkable that while the strongest negative correlation is found between the data of environments 1 and 3, the subnetworks of interactions in environment 3 support prediction in environment 1 better than the networks inferred from data in other environments (see Fig. 3; Table 1). Thus, how good inferred networks describe other data might not fully be determined by the correlation of the phenotypes in the respective environments. The pairwise correlations between phenotypes in the other environments are positive, and we observe a gain in predictive ability for all combinations. For environments 2 and 3, the correlation between phenotypes at 0.661 was relatively high. Thus, the observed improvement in predictive ability which increased from 0.502 ± 0.007 up to 0.650 ± 0.005 in environment 2 when the networks are inferred in environment 3 was relatively big, but smaller than the correlation of the phenotypes of both environments. The most interesting cases are the combinations of environment 4 with environments 2 and 3, respectively. In spite of the correlation between the phenotypes of environments 2 and 4 being smaller (0.411)

Table 1 Average correlation between predicted and measured phenotypes obtained by 200 independent draws of a test set consisting of 60 lines in the respective environment

	GBLUP	Full epistasis model	Max 1	Max 2	Max 3	Max 4	Gauss-RKHS
Environment 1	0.511 ± 0.007	0.554 ± 0.007	–	0.554 ± 0.007	0.561 ± 0.007	0.554 ± 0.007	**0.584 ± 0.006**
Environment 2	0.499 ± 0.007	0.502 ± 0.007	0.504 ± 0.007	–	**0.650 ± 0.005**	0.543 ± 0.006	0.500 ± 0.007
Environment 3	0.371 ± 0.008	0.390 ± 0.008	0.396 ± 0.008	**0.561 ± 0.007**	–	0.421 ± 0.008	0.422 ± 0.008
Environment 4	0.463 ± 0.007	0.498 ± 0.007	0.501 ± 0.007	**0.542 ± 0.006**	0.535 ± 0.006	–	0.531 ± 0.006

Average correlation reached by GBLUP, the full epistasis model and the maximum correlation reached by models with a reduced number of interactions selected with phenotypic data of all genotypes in Env1 (Max 1), Env2 (Max 2), Env3 (Max 3) or Env4 (Max 4). Gauss-RKHS describes the predictive ability obtained using the relationship matrix *K* provided by Crossa et al. (2010) which is based on a Gaussian kernel in an RKHS approach. The bold values highlight the highest predictive ability found for the respective data. All values represent the empirical mean with its standard error obtained from 200 independent draws of a test set

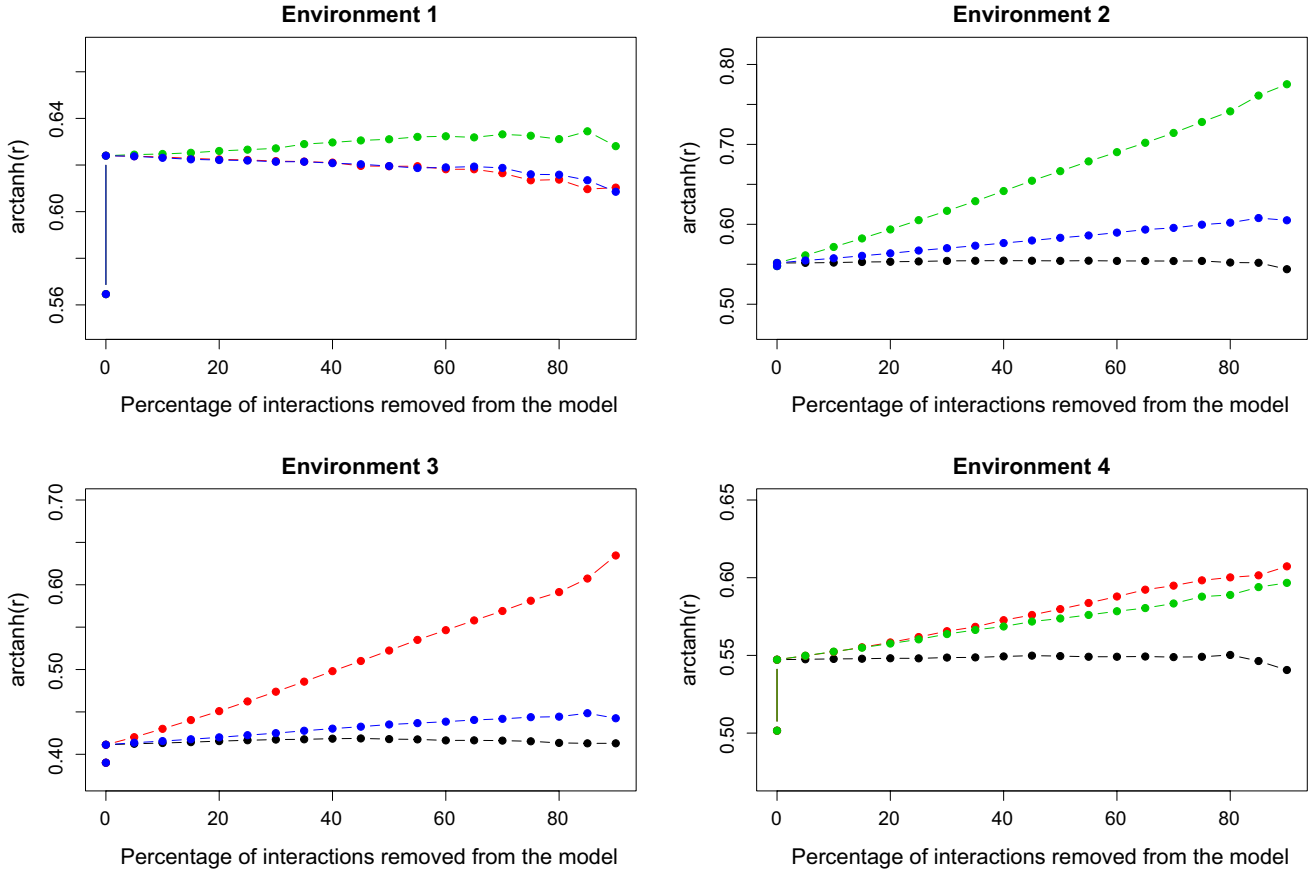

Fig. 3 Prediction accuracy (here Fisher's z-transformed correlation: arctanh(r)) in the different environments of the wheat data set with relationship matrices determined by variable selection in the other environments (0–90 % of the variables were removed in 5 % steps). *Black*, *red*, *green* and *blue dots* reflect the respective environment 1, 2, 3, or 4, which was used to determine the relationship matrices. The initial "jump" at zero represents the difference between GBLUP and the full epistatic model

than the predictive ability in each environment when the full model with all pairwise interaction is used (0.502 ± 0.007 and 0.498 ± 0.007), the predictive ability increased up to 0.543 ± 0.006 and 0.542 ± 0.006. The situation is similar for the combinations of environments 3 and 4 which exhibit a correlation of 0.388, which is smaller than or equal to the predictive ability of the full model in the respective environments (0.390 ± 0.008, 0.498 ± 0.007). The predictive ability increased up to 0.421 ± 0.008 and 0.535 ± 0.006. These examples show that the epistasis model offers a framework for combining results of different experiments to improve genome-assisted prediction by means of variable selection, provided that the phenotypes are positively correlated. We additionally (re)calculated the predictive ability with a RKHS method with Gaussian Kernel (matrix ***K*** provided by Crossa et al. (2010)). The predictive ability

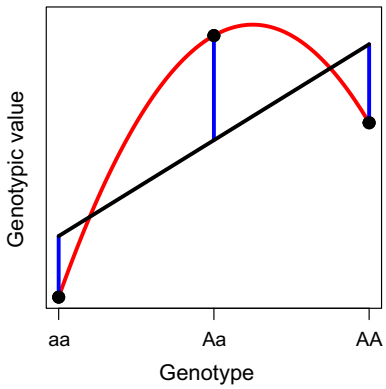

Fig. 4 Comparison of the classical way to incorporate dominance effects (Falconer and Mackay 1996) and the fitting by a polynomial of degree two. *Black points* genotypic values of the genotypes *aa*, *Aa*, *AA* of a diploid organism with the two alleles *a* and *A* present in the population. *Black line* linear regression which defines the additive effect in the classical model. *Blue lines* The dominance terms (one dominance term for each genotype) which are given by the difference of the genotypic values and the regression line. *Red curve* Fit by a polynomial of degree two

of the Kernel method was reached or improved when the phenotypes of the environments used for inferring the subnetwork structure and for prediction were positively correlated (see Table 1). For information on how the RKHS method incorporates nonlinear effects when a Gaussian kernel is used see Gianola et al. (2014).

Discussion

Prediction methods incorporating interactions between genes have been discussed extensively in the last years and have been assessed as potentially useful for prediction of complex traits (Hu et al. 2011; Mackay 2013; Wang et al. 2012). In this work, we demonstrated equivalences between epistatic effect models and linear models with certain covariance structures. These correspondences can be beneficial for computing genome-assisted predictions based on epistatic effect models. Moreover, we showed how to infer additive effects and pairwise interactions from given (additive or epistatic) genetic values. Additionally, we illustrated that polynomials of any degree D with the marker values as variables and random coefficients following a certain distribution can be rewritten as a linear mixed model with D random terms and covariance matrices whose structure is proportional to Hadamard powers of $\mathbf{MM}'$ ("Equivalent linear models for higher order polynomial functions of the markers"). An important point here is that the Hadamard powers of $\mathbf{G}$ implicitly model more variables than desired, and that the fraction of "undesired variables" will increase with the degree of interaction. Moreover, an important and practically relevant result is how to calculate the covariance matrix if only a certain subset of interactions is considered, which allows for building trait-specific, gene-network based relationships. Note here, that our theoretical results are not tied to any specific marker effects model. For instance, if markers are clustered into haplotypes of DNA segments and the linear effect model is built on these variables and their interactions, these equivalences can be used as well. Important questions for future research in this context are: (1) which methods of variable selection are appropriate (e.g., a priori based on available information vs. data driven) and (2) at which levels of a biological hierarchy interactions can be incorporated in the best way (e.g., marker interactions or at the level of interacting genes or pathways). It should be highlighted that if the interactions are incorporated as all pairwise marker interactions, the magnitude of individual interactions should be interpreted with caution since these do not necessarily reflect biological interaction due to the blatant over-parameterization of the model. The question of to which degree the epistasis effect model describes biology or statistical artifacts of the data remains open. As an example of variable selection we used the data of grain yield of wheat in one environment to infer subnetworks, and used these structures for genome-assisted prediction in the other environments. The results are in overall accordance with what one might expect: predictive ability in one environment will be increased by variable selection in another, if the phenotypes under the different conditions are positively correlated. The combinations of environments 2 and 3 with environment 4, which improved predictive ability to values higher than the correlation of the phenotypes in these environments, illustrated well that the epistasis model provides a framework to incorporate results of different experiments to increase predictive ability.

Problems and features of the epistasis model

As already mentioned, a major problem of the full epistasis model with all pairwise interactions is the over-parameterization which makes inference of biological interaction difficult. However, an approach of building the model on a level of clustered variables, such as genes instead of individual markers might already reduce the number of interactions significantly.

Moreover, it is also important here to mention that even if the coding of the markers has been shown to have no effect on predictive ability of the additive model (Strandén and Christensen 2011), encoding was reported to matter for epistasis models, in that different non-equivalent results are obtained with different marker codings (He et al. 2015). Thus, it should be noted that the use of the centered version $\tilde{\mathbf{G}} = (\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'$ with $\mathbf{P}$ being column-wise the matrix of marker genotype frequencies might be problematic

in an epistasis model: Since the epistasis model is affected by the choice of marker encoding, subtracting **P** means that we choose the underlying marker effect model for each pair of markers according to their frequencies. The consequences on predictions of phenotypes, additive effects and interactions are not clear and need to be investigated further.

In this regard, also note that due to the way the interactions are parameterized, the effects are not orthogonal, which means that the estimates of the additive effects will be influenced by whether we estimate dominance and epistasis simultaneously or not (Falconer and Mackay 1996; Zeng et al. 2005; Hallgrímsdóttir and Yuster 2008). Moreover, the non-orthogonality of the variance–covariance matrices can also be a problem when variance components are estimated simultaneously.

Compared to the RHKS approaches that also allow to model a non-linear relation between genotype and genetic value (Gianola and Rosa 2015; Morota and Gianola 2014; Gianola et al. 2014), the full epistasis model with all pairwise interactions was outperformed by a Gaussian kernel approach (see Table 1). However, a clear advantage of the epistasis model is that it provides a framework for incorporating network structures, which led in our example to a better performance in three of the four environments. How to incorporate information of this type into a RKHS approach is not clear. A possibility might be to define a genetic distance between genotypes that depends on the network structure.

The epistasis model in breeding programs

We demonstrated that variants of the epistasis model can improve genome-assisted prediction. However, the question of how a model with interactions can be used in breeding programs arises, since non-additive effects can be "lost" when other alleles change. Assuming that the interactions detected by the epistasis model have a biological significance, we think that models incorporating interactions can be beneficial for breeding programs for homozygous lines, if strong interactions between homozygous loci are detected. Moreover, models incorporating interactions are interesting options for hybrid breeding programs, since the heterosis effect is not a result of additive effects (Technow et al. 2014).

Classical subdivision of the total variance

Classical theory on incorporating dominance and epistasis into relationship matrices (Cockerham 1954; Kempthorne 1954; Henderson 1985) subdivides the genetic variance into additive, dominance, additive by additive, additive by dominance, dominance by dominance variance components and sometimes also terms of higher order (Su et al. 2012; Varona et al. 2014; Muñoz et al. 2014). Our parameterization extends the standard GBLUP model to account for interactions in a prediction model, but not to split the total variance into (orthogonal) components.

Outlook

By allowing to account for specific subsets of interactions, our methodology gives a basis for incorporating knowledge on gene networks and for building gene network specific relationship matrices. The methodology is not restricted to SNP or DArT markers but can also be applied to other types of regressions, for instance if markers are aggregated into haplotypes, which are then used as regression variables. Following this approach of incorporating pairwise interactions, future work should address efficient variable selection and how to code marker values optimally.

Author contribution statement JWRM: Derived the results and their mathematical proofs, analyzed the data, wrote the manuscript. VW: Guided the structure of the research, checked the results for validity, designed Figure 2. ME: Posed the original research question, guided the structure of the research, checked the results for validity. HS: Posed the original research question, guided the structure of the research. All authors read and approved the final version of the manuscript.

Acknowledgments The authors thank Daniel Gianola and another unknown reviewer for helpful suggestions. The comments helped to improve the manuscript immensely. JWRM thanks KWS SAAT SE for financial support and Camila Fabre Sehnem for helpful discussions.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards This manuscript constitutes a first submission to a scientific journal and neither the entire manuscript nor any part of its content has been published or has been accepted by another journal.

Supporting information

Dominance in this marker effect model and in classical quantitative genetics

Different ways of incorporating dominance into effect models exist in literature (see, e.g., Zeng et al. 2005). The most famous approach uses a linear regression of the three genotypic values of a diploid organism and adds a dominance term d_{aa}, d_{Aa}, d_{AA} which describes the difference between the linear fit and the

respective genotypic value (Falconer and Mackay 1996). Figure 4 compares this standard approach and the way dominance is modeled by the *pair epistasis and dominance model* (a fit by a polynomial of degree two). The scheme illustrates that the linear coefficient of the polynomial is not necessarily identical to the coefficient of the linear regression. Moreover, the term $h_{k,k}$ which is the second coefficient of the polynomial regression cannot be identified with $d_{\bullet\bullet}$. The parameterization of the two approaches is different. Analogously, the parameterization of the interactions in a two-locus model differs from the classical subdivision into additive by additive, additive by dominance and dominance by dominance effects. Regarding the effects of the pairs in a two-locus model, the framework used in this work does not give the freedom to choose any arbitrary effect for each allele combination of the two loci: The nine possible combinations of the states of two loci of a diploid organism (each with two alleles) are parameterized by less than nine parameters. Moreover, since the additive parameter of marker j is also present in all other two-loci effect models of j with any other locus k, the individual two-locus models are not independent but connected. For more information on two-locus models and their different parametrization, see the work of Hallgrímsdóttir and Yuster (Hallgrímsdóttir and Yuster 2008).

Derivation of the statistical equivalence of Eq. (3) and Eqs. (6, 7)

We consider the distribution of $\mathbf{y}$. The vector $\mathbf{1}\mu$ is nonrandom (fixed effect), the second summand of Eq. (3) translates into $\mathbf{g}$ of Eq. (6) which has a multivariate Gaussian distribution (with mean zero and covariance matrix $\sigma^2_\beta \mathbf{MM}'$). The vector of errors is multivariate Gaussian, too. What has to be considered in more detail, is the third summand of Eq. (3)

$$\sum_{k=1}^{p}\sum_{j=k+1}^{p} M_{i,j}M_{i,k}h_{j,k}.$$

We rewrite this equation for all n genotypes simultaneously in matrix notation

$$\underbrace{\begin{pmatrix} M_{1,1}M_{1,2} & M_{1,1}M_{1,3} & \cdots & M_{1,1}M_{1,p} & M_{1,2}M_{1,3} & \cdots & M_{1,p-1}M_{1,p} \\ M_{2,1}M_{2,2} & M_{2,1}M_{2,3} & \cdots & M_{2,1}M_{2,p} & M_{2,2}M_{2,3} & \cdots & M_{2,p-1}M_{2,p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ M_{n,1}M_{n,2} & M_{n,1}M_{n,3} & \cdots & M_{n,1}M_{n,p} & M_{n,2}M_{n,3} & \cdots & M_{n,p-1}M_{n,p} \end{pmatrix}}_{=\mathbf{N}} \begin{pmatrix} h_{1,2} \\ h_{1,3} \\ \vdots \\ h_{1,p} \\ h_{2,3} \\ \vdots \\ h_{p-1,p} \end{pmatrix}, \tag{19}$$

and name the left hand matrix $\mathbf{N}$. This presentation shows that the third summand of the covariance matrix analog of Eq. (3) is also multivariate Gaussian distributed with mean zero and covariance matrix $\mathbf{NN}'\sigma^2_h$ (according to a definition of the multivariate Gaussian distribution, since the interactions are assumed to be i.i.d. Gaussian distributed random variables). To see which structure the covariance has, we compare $\mathbf{NN}'$ to $\mathbf{MM}'$ by regarding the entries $(\mathbf{NN}')_{i,l}$ and $(\mathbf{MM}')_{i,l}$.

$$(\mathbf{NN}')_{i,l} = \sum_{k=1,j>k}^{p} M_{i,k}M_{i,j}M_{l,k}M_{l,j} = \sum_{k=1}^{p}\left(M_{i,k}M_{l,k}\sum_{j>k}^{p} M_{i,j}M_{l,j}\right) \tag{20}$$

and

$$(\mathbf{MM}')_{i,l} = \sum_{k=1}^{p} M_{i,k}M_{l,k}. \tag{21}$$

Eqs. (20) and (21) show that the structure of the additional covariance matrix $\mathbf{NN}'$ of the epistasis model resembles $\mathbf{G}$, but the summands of each entry are weighted by products of entries of matrix $\mathbf{M}$.

Derivation of the statistical equivalence of Eq. (4) and Eqs. (6, 8)

Analogously, to the procedure we applied to the first model, we regard the third term

$$\sum_{j=1}^{p}\sum_{k=1}^{p} M_{i,j}M_{i,k}h_{j,k},$$

and write

$$\underbrace{\begin{pmatrix} M_{1,1}^2 & M_{1,1}M_{1,2} & \cdots & M_{1,1}M_{1,p} & M_{1,2}M_{1,1} & M_{1,2}^2 & \cdots & M_{1,p}^2 \\ M_{2,1}^2 & M_{2,1}M_{2,2} & \cdots & M_{2,1}M_{2,p} & M_{2,2}M_{2,1} & M_{2,2}^2 & \cdots & M_{2,p}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ M_{n,1}^2 & M_{n,1}M_{n,2} & \cdots & M_{n,1}M_{n,p} & M_{n,2}M_{n,1} & M_{n,2}^2 & \cdots & M_{n,p}^2 \end{pmatrix}}_{=\mathbf{Q}} \begin{pmatrix} h_{1,1} \\ h_{1,2} \\ \vdots \\ h_{1,p} \\ h_{2,1} \\ h_{2,2} \\ \vdots \\ h_{p,p} \end{pmatrix} \tag{22}$$

where the letter $\mathbf{Q}$ is defined to be the left hand matrix. Thus, we are interested in $(\mathbf{QQ}')_{i,l}$ which is given by

$$(\mathbf{QQ}')_{i,l} = \begin{pmatrix} M_{i,1}^2 & M_{i,1}M_{i,2} & \cdots & M_{i,1}M_{i,p} & M_{i,2}M_{i,1} & M_{i,2}^2 & \cdots & M_{i,p}^2 \end{pmatrix} \times \begin{pmatrix} M_{l,1}^2 \\ M_{l,1}M_{l,2} \\ \vdots \\ M_{l,1}M_{l,p} \\ M_{l,2}M_{l,1} \\ M_{l,2}^2 \\ \vdots \\ M_{l,p}^2 \end{pmatrix} = \sum_{k,j=1}^{p} M_{i,k}M_{i,j}M_{l,k}M_{l,j} = \left(\sum_{k=1}^{p} M_{i,k}M_{l,k}\right)^2 \tag{23}$$

which means that $\mathbf{QQ}'$ represents the Hadamard product $\mathbf{MM}' \circ \mathbf{MM}'$.

Proof of Eq. (9)

From Eq. (20) we know that

$$\begin{aligned} 2(\mathbf{NN}')_{i,l} &= 2\sum_{k=1}^{p}\left(M_{i,k}M_{l,k}\sum_{j>k}^{p} M_{i,j}M_{l,j}\right) \overset{*}{=} \sum_{k=1}^{p}\left(M_{i,k}M_{l,k}\sum_{j\neq k}^{p} M_{i,j}M_{l,j}\right) \\ &= \left(\mathbf{MM}' \circ \mathbf{MM}'\right)_{i,l} - \sum_{k=1}^{p} M_{i,k}M_{l,k}M_{i,k}M_{l,k} \\ &= \left(\mathbf{MM}' \circ \mathbf{MM}'\right)_{i,l} - \left((\mathbf{M}\circ\mathbf{M})(\mathbf{M}\circ\mathbf{M})'\right)_{i,l} \end{aligned}$$

To see equality $*$, consider the left side as sum over products defined by all tuples $\{(k,j)|j > k\}$ with fixed i, l. Since j and k can be exchanged with each other, this is equal to the sum over all products defined by $\{(k,j)|j < k\}$. Thus, multiplying the sum defined by the tuples $\{(k,j)|j > k\}$ by 2 equals adding the sum defined by $\{(k,j)|j < k\}$.

Derivation of Eq. (5)

We maximize

$$F(\mu, \mathbf{g}, \mathbf{y}_{\text{train}}) := -\left(\left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{y}_{\text{test}}\end{pmatrix} - \mathbf{1}\mu - \mathbf{g}\right)' \times \mathbf{I}_n \frac{1}{\sigma_\epsilon^2}\left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{y}_{\text{test}}\end{pmatrix} - \mathbf{1}\mu - \mathbf{g}\right) + \mathbf{g}'\frac{1}{\sigma_\beta^2}\mathbf{G}^{-1}\mathbf{g}\right)$$

with respect to μ, $\mathbf{g}$ and $\mathbf{y}_{\text{test}}$ by calculating the partial derivatives and the corresponding zeros. Let m denote the number of genotypes in the test set. Thus, we have to solve

$$(i) \quad \frac{\partial F}{\partial \mu} \text{ gives: } \begin{pmatrix}\mathbf{1}_{n-m}\\ \mathbf{1}_m\end{pmatrix}' \left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \hat{\mathbf{y}}_{\text{test}}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_{n-m}\\ \mathbf{1}_m\end{pmatrix}\hat{\mu} - \begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix}\right) = 0$$

$$(ii) \quad \frac{\partial F}{\partial \mathbf{g}} \text{ gives: } \left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \hat{\mathbf{y}}_{\text{test}}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_{n-m}\\ \mathbf{1}_m\end{pmatrix}\hat{\mu} - \begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix}\right) - \sigma_\epsilon^2\left(\frac{1}{\sigma_\beta^2}\mathbf{G}^{-1}\right)\begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix} = 0$$

$$(iii) \quad \frac{\partial F}{\partial \mathbf{y}_{\text{test}}} \text{ gives: } \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_m\end{pmatrix}\left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \hat{\mathbf{y}}_{\text{test}}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_{n-m}\\ \mathbf{1}_m\end{pmatrix}\hat{\mu} - \begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix}\right) = 0$$

$\mathbf{1}_i$ here denotes the vector of length i with every entry equal to 1 and $\mathbf{I}_m$ the m-dimensional identity matrix. Eq. (iii) can be rewritten to $\hat{\mathbf{y}}_{\text{test}} = \hat{\mathbf{g}}_{\text{test}} + \mathbf{1}_m\hat{\mu}$. Plugging this into Eq. (i) gives $\hat{\mu} = \mathbf{1}'_{n-m}(\mathbf{y}_{\text{train}} - \hat{\mathbf{g}}_{\text{train}})(n-m)^{-1}$. Using the rewritten version of (iii) in (ii) gives

$$\begin{aligned} (iii) \text{ in } (ii): \quad & \left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}} + \mathbf{1}_m\hat{\mu}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_{n-m}\\ \mathbf{1}_m\end{pmatrix}\hat{\mu} - \begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix}\right) - \sigma_\epsilon^2\left(\frac{1}{\sigma_\beta^2}\mathbf{G}^{-1}\right)\begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix} = \mathbf{0} \\ \Longleftrightarrow \\ & \left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_{n-m}\\ \mathbf{0}\end{pmatrix}\hat{\mu} - \begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \mathbf{0}\end{pmatrix}\right) - \sigma_\epsilon^2\left(\frac{1}{\sigma_\beta^2}\mathbf{G}^{-1}\right)\begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix} = \mathbf{0} \\ \Longleftrightarrow \\ & \left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_{n-m}\\ \mathbf{0}\end{pmatrix}\hat{\mu}\right) - \left(\underbrace{\mathbf{I}_n - \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_m\end{pmatrix}}_{=\mathbf{T}_{\text{train}}} + \sigma_\epsilon^2\left(\frac{1}{\sigma_\beta^2}\mathbf{G}^{-1}\right)\right)\begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix} = \mathbf{0} \end{aligned}$$

The latter equivalence uses the equality $\begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \mathbf{0}\end{pmatrix} = \mathbf{T}_{\text{train}}\begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix}$. The left summand can be rewritten to

$$\begin{aligned}\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_s\\ \mathbf{0}\end{pmatrix}\hat{\mu} &= \begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_s\\ \mathbf{0}\end{pmatrix}\mathbf{1}_s'(\mathbf{y}_{\text{train}} - \hat{\mathbf{g}}_{\text{train}})s^{-1}\\ &= \begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{J}_{s\times s}\\ \mathbf{0}\end{pmatrix}(\mathbf{y}_{\text{train}} - \hat{\mathbf{g}}_{\text{train}})s^{-1}\\ &= \begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_s\bar{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix} + \begin{pmatrix}\mathbf{J}_{s\times s}\\ \mathbf{0}\end{pmatrix}\hat{\mathbf{g}}_{\text{train}}s^{-1}.\end{aligned}$$

by plugging in the rewritten version of (i) into (ii), using $s = n - m$ for the number of genotypes of the training set, and defining the empirical mean $\bar{y}_{\text{train}} = s^{-1}\sum \mathbf{y}_{\text{train}}$. Moreover, writing $\hat{\mathbf{g}}_{\text{train}} = (\mathbf{I}_s, \mathbf{0})\begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix}$, gives

$$\begin{aligned}\left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_s\bar{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix}\right) - \left(\mathbf{T}_{\text{train}} - s^{-1}\begin{pmatrix}\mathbf{J}_{s\times s} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right.\\ \left.+\sigma_\epsilon^2\left(\frac{1}{\sigma_\beta^2}\mathbf{G}^{-1}\right)\right)\begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix} = \mathbf{0}\end{aligned}$$

and thus

$$\begin{aligned}\begin{pmatrix}\hat{\mathbf{g}}_{\text{train}}\\ \hat{\mathbf{g}}_{\text{test}}\end{pmatrix} = \left(\mathbf{T}_{\text{train}} - s^{-1}\begin{pmatrix}\mathbf{J}_{s\times s} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} + \sigma_\epsilon^2\left(\frac{1}{\sigma_\beta^2}\mathbf{G}^{-1}\right)\right)^{-1}\\ \times\left(\begin{pmatrix}\mathbf{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{1}_s\bar{y}_{\text{train}}\\ \mathbf{0}\end{pmatrix}\right)\end{aligned}$$

which represents Eq. (5).

Proof of the statement of "Equivalent linear models for higher order polynomial functions of the markers"

We know that the statement is true for $D = 2$. What has to be shown is that if it is true for $D = j - 1$ then it is also true for $D = j$ (mathematical induction). Let j be the degree. Analogously to Eq. (22), we consider the matrix $\mathbf{Q}^{(j)}$ which has all products of j factors in the respective row of matrix $\mathbf{M}$ as entries. The i-th row $\mathbf{Q}_{i,\bullet}^{(j)}$ of matrix $\mathbf{Q}^{(j)}$ can be written as

$$\mathbf{Q}_{i,\bullet}^{(j)} = \mathbf{M}_{i,\bullet}\otimes\mathbf{Q}_{i,\bullet}^{(j-1)}$$

if the $\beta_{\boldsymbol{\kappa},j}$ is ordered appropriately ($\otimes$ denotes the Kronecker product). Then the (i, l)-th entry of $\mathbf{Q}^{(j)}\mathbf{Q}^{(j)\prime}$ is the matrix product

$$\left(\mathbf{Q}^{(j)}\mathbf{Q}^{(j)\prime}\right)_{i,l} = \left(\mathbf{M}_{i,\bullet}\otimes\mathbf{Q}_{i,\bullet}^{(j-1)}\right)\left(\mathbf{M}_{l,\bullet}\otimes\mathbf{Q}_{l,\bullet}^{(j-1)}\right)'$$

which is equal to

$$\begin{aligned}&\left(\mathbf{M}_{i,\bullet}\otimes\mathbf{Q}_{i,\bullet}^{(j-1)}\right)\left((\mathbf{M}_{l,\bullet})'\otimes\left(\mathbf{Q}_{l,\bullet}^{(j-1)}\right)'\right)\\ &= \underbrace{\left(\mathbf{M}_{i,\bullet}(\mathbf{M}_{l,\bullet})'\right)}_{=G_{i,l}}\otimes\underbrace{\left(\mathbf{Q}_{i,\bullet}^{(j-1)}\left(\mathbf{Q}_{l,\bullet}^{(j-1)}\right)'\right)}_{=G_{i,l}^{j-1}} = G_{i,l}^{j},\end{aligned}$$

according to the calculation rules for the Kronecker product and matrix multiplication and the induction hypothesis.

Proof of Eq. (17)

Assuming one interaction between marker positions k and j leads to the equation

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{M}\boldsymbol{\beta} + \underbrace{\begin{pmatrix}M_{1,k}M_{1,j}\\ M_{2,k}M_{2,j}\\ \vdots\\ M_{n,k}M_{n,j}\end{pmatrix}}_{=\mathbf{R}}h_{k,j} + \boldsymbol{\epsilon},$$

where we call the vector $\mathbf{R}$. According to the calculation rules for multivariate normal distributions, the covariance matrix of the corresponding interaction effect $\mathbf{g_2}$ is $\sigma_h^2\mathbf{R}\mathbf{R}'$. We have to show that this expression is equal to Eq. (17). As before, we consider the entry (l, i) of $\sigma_h^2\mathbf{R}\mathbf{R}'$ which is given by $M_{l,k}M_{l,j}M_{i,k}M_{i,j}\sigma_h^2$. Moreover the (l, i)-th entry of $\left(\mathbf{M}_{\bullet,k}\mathbf{M}_{\bullet,k}'\right)$ is $M_{l,k}M_{i,k}$ and the (l, i)-th entry of $\left(\mathbf{M}_{\bullet,j}\mathbf{M}_{\bullet,j}'\right)$ is $M_{l,j}M_{i,j}$ which proves the statement.

References

Abdollahi-Arpanahi R, Pakdel A, Nejati-Javaremi A, Moradi Shahrbabak M, Morota G, Valente BD, Kranis A, Rosa GJM, Gianola D (2014) Dissection of additive genetic variability for quantitative traits in chickens using SNP markers. J Anim Breed Genet 131(3):183–193

Clifford D, McCullagh P (2006) The regress function. R News 6(2):10

Clifford D, McCullagh P (2014) The regress package. R package version 1.3-14

Cockerham CC (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39(6):859–882

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

de los Campos G, Perez-Rodriguez P (2014) BGLR: Bayesian Generalized Linear Regression. R package version 1.0.3. http://CRAN.R-project.org/package=BGLR

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. England, Benjamin Cummings

Gianola D, Rosa GJM (2015) One hundred years of statistical developments in animal breeding. Annu Rev Anim Biosci 3:19–56

Gianola D, Morota G, Crossa J (2014) Genome-enabled prediction of complex traits with kernel methods: What have we learned?. In: Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397

Hallgrímsdóttir IB, Yuster DS (2008) A complete classification of epistatic two-locus models. BMC Genet 9(1):17

Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91(1):47–60

He D, Wang Z, Parida L (2015) Data-driven encoding for quantitative genetic trait prediction. BMC Bioinform 16(Suppl 1):S10

Henderson CR (1984) Application of linear models in animal breeding. University of Guelph, Guelph

Henderson CR (1985) Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. J Anim Sci 60(1):111–117

Henderson CR, Quaas RL (1976) Multiple trait evaluation using relatives records. J Anim Sci 43:1188

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31(2):423–447

Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4(2):e1000008

Hu Z, Li Y, Song X, Han Y, Cai X, Xu S, Li W (2011) Genomic value prediction for quantitative traits under the epistatic model. BMC Genet 12:15

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

Kempthorne O (1954) The correlation between relatives in a random mating population. In: Proceedings of the Royal Society of London. Series B-Biological Sciences 143, vol 910, pp 103–113

Mackay TFC (2013) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet 15:22–33. doi:10.1038/nrg3627

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Morota G, Gianola D (2014) Kernel-based whole-genome prediction of complex traits: a review. Front Genet 5:363. doi:10.3389/fgene.2014.00363

Morota G, Koyama M, Rosa GJM, Weigel KA, Gianola D (2013) Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. Genet Sel Evol 45:17

Muñoz PR, Resende MFR, Gezan SA, Resende MDV, de los Campos G, Kirst M, Huber D, Peter GF (2014) Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics 198(4):1759–1768

Ober U, Huang W, Magwire M, Schlather M, Simianer H, Mackay TFC (2015) Accounting for genetic architecture improves sequence based genomic prediction for a Drosophila fitness trait. PLoS One 10(5):e0126880. doi:10.1371/journal.pone.0126880

Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228

R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge

Strandén I, Christensen OF (2011) Allele coding in genomic evaluation. Genet Sel Evol 43:25

Su G, Christensen OF, Ostersen T, Henryon M, Lund MS (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLOS One 7(9):e45293

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197(4):1343–1355

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423

Varona L, Vitezica ZG, Munilla S, Legarra A (2014) A general approach for calculation of genomic relationship matrices for epistatic effects. In: Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Wang D, El-Basyoni IS, Baenziger PS, Crossa J, Eskridge KM, Dweikat I (2012) Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. Heredity 109(5):313–319

Wittenburg D, Melzer N, Reinsch N (2011) Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. BMC Genet 12:74

Zeng Z, Wang T, Zou W (2005) Modeling quantitative trait loci and interpretation of models. Genetics 169(3):1711–1725

On the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE)

Martini *et al. BMC Bioinformatics* (2017) 18:3
DOI 10.1186/s12859-016-1439-1

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE)

Johannes W. R. Martini[1*], Ning Gao[1,2], Diercles F. Cardoso[1,3], Valentin Wimmer[4], Malena Erbe[1,5], Rodolfo J. C. Cantet[6] and Henner Simianer[1]

Abstract

Background: Epistasis marker effect models incorporating products of marker values as predictor variables in a linear regression approach (extended GBLUP, EGBLUP) have been assessed as potentially beneficial for genomic prediction, but their performance depends on marker coding. Although this fact has been recognized in literature, the nature of the problem has not been thoroughly investigated so far.

Results: We illustrate how the choice of marker coding implicitly specifies the model of how effects of certain allele combinations at different loci contribute to the phenotype, and investigate coding-dependent properties of EGBLUP. Moreover, we discuss an alternative categorical epistasis model (CE) eliminating undesired properties of EGBLUP and show that the CE model can improve predictive ability. Finally, we demonstrate that the coding-dependent performance of EGBLUP offers the possibility to incorporate prior experimental information into the prediction method by adapting the coding to already available phenotypic records on other traits.

Conclusion: Based on our results, for EGBLUP, a symmetric coding $\{-1, 1\}$ or $\{-1, 0, 1\}$ should be preferred, whereas a standardization using allele frequencies should be avoided. Moreover, CE can be a valuable alternative since it does not possess the undesired theoretical properties of EGBLUP. However, which model performs best will depend on characteristics of the data and available prior information. Data from previous experiments can for instance be incorporated into the marker coding of EGBLUP.

Keywords: Genomic prediction, Epistasis model, Interaction

Background

Genomic prediction aims at forecasting qualitative or quantitative properties of individuals based on known genetic information. The genetic information can for instance be given by single-nucleotide-polymorphisms (SNPs) or other kinds of genetic data of individual animals, plant lines or humans. Applied to animals and plants, genomic prediction is of central importance for breeding within the concept of *genomic selection* [1, 2]. Moreover, genomic prediction can also be used in medicine or epidemiology for risk assessment or prevalence studies of (partially) genetically determined diseases (e.g. [3]). One of the standard approaches for genomic prediction of quantitative traits is based on a linear regression model in which the phenotype is described by a linear function of the genotypic markers. In more detail, the standard additive linear model is defined by the equation

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1)$$

*Correspondence: jmartin2@gwdg.de
[1]Department of Animal Sciences, Georg-August University, Albrecht Thaer-Weg 3, Göttingen, Germany
Full list of author information is available at the end of the article

where $\mathbf{y}$ is the $n \times 1$ vector of phenotypes of the n individuals, $\mathbf{1}$ the $n \times 1$ vector with each entry equal to 1, μ the fixed effect and $\mathbf{M}$ the $n \times p$ matrix giving the p marker values of the n individuals. Moreover, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown marker effects and $\boldsymbol{\epsilon}$ a random $n \times 1$ error vector with $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. Since the number of markers p is typically much larger than the number of individuals n, the additional assumption that $\beta_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\beta^2)$ is usually made (and all random terms together are considered as stochastically independent). In particular, using an approach of maximizing the density of a certain distribution [4], this assumption allows us to determine the penalizing weight in a Ridge Regression approach which is known as *ridge regression best linear unbiased prediction* (RRBLUP) and which is fully equivalent to its relationship matrix-based counterpart *genomic best linear unbiased prediction* (GBLUP)[1] [5, 6]. The answer to the question which type of marker coding is appropriate in $\mathbf{M}$ depends on the combination of the type of genotypic marker and ploidy of the organism dealt with. For instance, if haploid organisms are considered or presence/absence markers are used, a possible coding for the j-th marker value of the i-th individual $M_{i,j}$ is the set $\{0, 1\}$. Counting the occurrence of an allele of a diploid organism, the sets $\{0, 1, 2\}$ or $\{-1, 0, 1\}$, or rescaled variants can be used. If the marker effects $\boldsymbol{\beta}$ and the fixed effect μ are predicted/estimated as $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}$ on the basis of a training set, the expected phenotypes of individuals from a test set, which were not used to determine $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}$, can be predicted by using their marker information in Eq. (1) with $\hat{\mu}, \hat{\boldsymbol{\beta}}$. We will call the difference between the predicted expected phenotype and the estimated fixed effect the predicted *genetic value*. For the purely additive model of Eq. (1) and a diploid organism with possible genotypes *aa*, *aA* and *AA* for locus j, the choice of how to translate these possibilities into numbers was reported not to affect the predictive ability notably, as long as the difference between the coding of *aa* and *aA* is the same as between *aA* and *AA* and equal for all markers [5, 7–9]. However, an extension of the additive model, which we call the *extended* GBLUP model (EGBLUP) [10, 11]

$$y_i = \mu + \sum_{j=1}^{p} M_{i,j}\beta_j + \sum_{k=1}^{p}\sum_{j=k}^{p} M_{i,j}M_{i,k}h_{j,k} + \epsilon_i, \qquad (2)$$

has been shown to exhibit strong coding dependent performance [12, 13]. Here, $h_{j,k} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_h^2\right)$ is the pairwise interaction effect of markers j and k and all other variables as previously defined (all terms stochastically independent). Compared to Eq. (1), this model additionally incorporates pairwise products of marker values as predictor variables and thus allows us to model interactions between markers. Moreover, the interaction of a marker with itself gives a possibility to model dominance effects (see e.g. [11, 14–16]). The epistasis model of Eq. (2) and some variations with restrictions on which markers can interact have been the main object of investigation in several publications and models incorporating epistasis have been viewed as potentially beneficial for the prediction of complex traits [10, 11, 17–19], but a marker coding dependent performance was observed [12, 13].

In this work, we investigate how the marker coding specifies the effect model for markers with two or three possible values and show how we can find the marker coding for an a priori specified model. We discuss advantages and disadvantages of different coding methods and investigate properties of alternative linear models based on categorical instead of numerical dosage variables. In particular, we show how to represent these models as genomic relationship matrices. Finally, we compare the predictive abilities of different epistasis models on simulated and publicly available data sets and demonstrate a way of using the coding-dependent performance of EGBLUP to incorporate prior information.

Methods

Data sets used for assessing predictive ability

Simulated data

A population with 10 000 bi-allelic markers spread across five chromosomes was simulated, using the QMSim software [20]. The size of the first chromosome was 140 centimorgan (cM) with 3 500 markers. Chromosomes 2 to 5 had a size of 110 cM (2 750 markers), 80 cM (2 000 markers), 50 cM (1 250 markers) and 20 cM (500 markers), receptively. In order to allow mutations and linkage disequilibrium establishment, a historical population was simulated with 5 000 individuals (2 500 males and 2 500 females) with random mating for 1 000 generations with constant population size and with a replacement rate of 0.2 for males and females. Then the population size was reduced to 1 000 individuals for 20 additional generations (generation 1 001 to 1 020). The simulated mutation rate was $2.5 \cdot 10^{-5}$.

We used this simulated genotypes as basis and modeled three different types of genetic architecture (purely additive, purely dominant and purely epistatic), each with a varying number of quantitative trait loci (QTL) on top. We chose these types of genetic architecture, without additive effects in the dominance and epistasis scenarios, to make the three scenarios as different as possible. To model the phenotype, out of the 10 000 markers, 200 were drawn randomly from each of the five chromosomes to define in total 1 000 QTL for additive or dominance effects. For the purely additive scenario, the 1 000 additive effects were drawn independently from a $\mathcal{N}(0, 1)$ distribution. For the first additive trait A1, 10 out of the 1 000 QTL were drawn and the genetic values of all individuals were calculated

according to the effects of these 10 loci. To define a broad sense heritability of 0.8, the genetic values were standardized to mean 0 and variance 1 and individual errors were drawn from a $\mathcal{N}(0, 0.25)$ distribution. Having added these individual errors to the genetic values, these phenotypes were again standardized to mean 0 and variance 1. For the second trait A2, additional 90 QTL were drawn from the initial 1 000 to give in total 100 QTL for this trait including the QTL of trait A1 with their corresponding effects. Analogously, for A3, all initially drawn 1 000 QTL were used. The standardization procedure was identical to the one previously described for A1. For the comparison of genomic prediction with different relationship models, these 1 000 markers were removed. The relationship matrices were based on the remaining 9 000 markers.

For the dominance scenario D1 (10 QTL), D2 (100 QTL) and D3 (1 000 QTL), we used the same QTL positions as for A1, A2, and A3, respectively, but simulated $\mathcal{N}(0, 1)$-distributed dominance effects. The standardization procedure to a broad sense heritability of 0.8 was carried out as described before.

For the epistasis traits E1, E2 and E3, 1 000, 10 000 or 100 000 pairs of markers were drawn randomly and for each draw, one of the nine possible configurations of the pair was randomly chosen to have an $\mathcal{N}(0, 1)$-distributed effect. For instance, having drawn the marker pair j, k, only the configuration $(M_{i,j}, M_{i,k}) = (0, 2)$ was chosen to have an effect, which again was drawn randomly. This was done independently for each trait, which means trait E2 does not necessarily share causal combinations of markers with trait E1. The phenotypes were standardized as described above. Note, that the markers involved in causal combinations were not removed here, since in expectation, every marker is somehow involved in the phenotype of trait E2 and E3.

We repeated this whole procedure, including the simulation of the genotypes, 20 times and compared the different models by their average predictive ability across the 20 repetitions. The simulated data can be found in Additional file 1 of this publication.

Wheat data

The wheat data which we used to compare different methods was published by Crossa et al. [21]. The 1279 DArT markers of 599 CIMMYT inbred wheat lines indicate whether a certain allele is present (1) or not (0). The phenotypic data describes standardized records of grain yield under four environmental conditions.

Mouse data

The mouse data set we used was published and described by Solberg et al. [22] and Valdar et al. [23], and was downloaded from the corresponding website of the Wellcome Trust Centre for Human Genetics. The physical map of single nucleotide polymorphisms (SNPs) was updated to the latest version of the mouse genome (*Mus musculus*, assembly GRCm38.p4) with the `biomaRt R` package [24, 25]. Only SNPs mapped to the GRCm38.p4 were used for further analysis. For the remaining markers, the ratio of missing marker values was rather low (0.33%) and we performed a random imputation. The nucleotide coded genotypes were translated to a {0,1,2} coding, where 0 and 2 denote the two homozygous and 1 the heterozygous genotype. SNPs with minor allele frequency (MAF) smaller than 0.01 were excluded from the dataset. Imputation, recoding, and quality control of genotypes were carried out with the `synbreed R` package simultaneously [26]. A number of 9265 SNPs remained in the dataset for further analysis. We only used individuals with available records for all considered traits for further analysis, which reduced the number of individuals to 1 298. We focused on the provided pre-corrected residuals of 13 traits from which fixed effects of trait-specific relevant covariates such as sex, season, month, have already been subtracted. A detailed description of the traits can be found on the corresponding sites of the UCL. Moreover, the data resulting from quality control and filtering as well as the corrected phenotypes of the traits we used can be found in Additional file 1.

Genomic relationship based prediction and assessment of predictive ability

We used an approach based on relationship matrices for genomic prediction. The underlying concept of this approach is the equivalence of marker effect-based and genomic relationship-based prediction ([5, 10, 11]). Given the respective relationship matrix, the prediction is performed by Eq. (3) (for a derivation of this equation see the supporting information of [11]):

$$\begin{pmatrix} \hat{\mathbf{g}}_{train} \\ \hat{\mathbf{g}}_{test} \end{pmatrix} = \left[\mathbf{T}_{train} - s^{-1} \begin{pmatrix} \mathbf{J}_{s\times s} & 0 \\ 0 & 0 \end{pmatrix} + \sigma_\epsilon^2 \left(\frac{1}{\sigma_\beta^2} \mathbf{G}^{-1} \right) \right]^{-1} \left(\begin{pmatrix} \mathbf{y}_{train} \\ 0 \end{pmatrix} - \begin{pmatrix} \mathbf{1}_s \bar{y}_{train} \\ 0 \end{pmatrix} \right) \tag{3}$$

The matrix $\mathbf{G}$ is the central object denoting the genomic relationship matrix of the respective model. The variables $\hat{\mathbf{g}}_i$ are the predicted genetic values (expected phenotype minus the fixed effect $\hat{\mu}$) of the respective set (training or test set). Moreover, s is the number of genotypes in the training set, $\mathbf{1}_s$ is the vector of length s with each entry equal to 1, $\mathbf{J}_{s\times s}$ is the analogous $s \times s$ matrix with each entry equal to 1 and $\bar{y}_{train}$ is the empirical mean of the training set. Here, $\mathbf{T}_{train}$ denotes the diagonal matrix of dimension n with 0 on the diagonal at the positions of the test set genotypes, and 1 for the training set individuals.

To assess the predictive ability of different models, we chose a test set consisting of ∼10% of the total number of individuals (100, 60, or 130 for the simulated, the wheat and the mouse data, respectively). We then used the remaining individuals as a training set and predicted the genetic values for all individuals using Eq. (3). The variance components σ^2_ϵ and σ^2_β were estimated from the training set using version 3.1 of the `R` package `EMMREML` [27]. The relationship matrix relating the genotypes of the training set was used to estimate the variance components based on the phenotypes of the training set only. The variance components were then used with the complete relationship matrix for the prediction of the genetic values of all individuals in Eq. (3). This procedure was repeated 200 times, with independently drawn test sets. The average correlation r between observed and predicted mean phenotypes of the test set was used as a measure of predictive ability. A description of how the different effect models can be translated into relationship matrices is given in the results. For the Gaussian kernel, we used the bandwidth parameter $b = 2q^{-1}_{0.5}$, with $q_{0.5}$ the median of all squared Euclidean distances between the individuals of the respective data. For the simulated data which consisted of 20 independent data sets, we present the average predictive ability and the average standard error of the mean. For the wheat and the mouse data, we used Tukey's 'Honest Significant Difference' test to contrast the performance of the different prediction methods (`TukeyHSD()` and `lm()` of `R` [28]).

Incorporation of prior information by marker coding

As described above, the data we used offers records of different traits or trait×environment combinations of the same individuals. We will illustrate that the coding-dependent performance of EGBLUP can also be used to incorporate a priori information into the model by choosing the coding for each interaction with already provided data and by using the corresponding relationship matrix for prediction under altered environmental conditions or for a correlated trait. We used for the wheat data the following procedure:

1) We predicted all the interactions $\hat{h}_{k,l}$ for a given trait in a given environment, under the use of the $\{0, 1\}$ coding originally provided by Crossa et al. [21] (as described by Martini et al. [11]).
2) We changed the "orientation" of all markers at once by substituting 0 by 1, and 1 by 0 and predicted all interactions $\tilde{h}_{k,l}$ under the use of the altered coding.
3) If the ratio of $\left|\frac{\hat{h}_{k,l}}{\tilde{h}_{k,l}}\right|$ was greater than or equal to 1, we assumed that the original orientation provided by the data set describes the respective interaction better than the alternative coding.
4) We then calculated a relationship matrix for each interaction individually by
$$\mathbf{G}_{k,l} = \left(\mathbf{M}_{\bullet,\mathbf{k}}\mathbf{M}'_{\bullet,\mathbf{k}}\right) \circ \left(\mathbf{M}_{\bullet,\mathbf{l}}\mathbf{M}'_{\bullet,\mathbf{l}}\right)$$
with $\mathbf{M}_{\bullet,k}$ denoting the $n \times 1$ vector of marker data of locus k for all individuals in the respective coding which seems to fit the interaction better according to 3) (see [11, 29]). Here, ∘ denotes the Hadamard product.
5) The overall relationship matrix was then defined by
$$\mathbf{G} = \sum_{k=1}^{p}\sum_{l\geq k}^{p}\mathbf{G}_{k,l}.$$

We used the data of each environment to calculate an optimally coded relationship matrix for this environment, which was used afterwards for predicting phenotypes in the other environments. The underlying heuristic of step 3) is that a small effect means that the interaction is less important in the respective coding. If the underlying effect model defined by the coding does not capture the data structure, the estimated effect should be close to zero. However, if the effect of a combination is important to describe the phenotype distribution, a larger effect should be assigned (see also Example 1, where the estimated effect is 0, if the underlying parameterization cannot describe the present effect distribution).

For the mouse data, we used the 13 considered traits to construct a relationship matrix for each of them. Each relationship matrix was afterwards used for prediction within the data of the twelve other traits. The two different codings which were compared here, were the $\{0, 1, 2\}$ coding based on the imputed originally provided data and its inverted version with 0 and 2 permuted.

Results

In the following, we will highlight aspects of the behavior of the additive effect model of Eq. (1) when the marker coding is altered. These properties of the additive model will afterwards be compared to those of the epistasis model of Eq. (2).

All relationship matrices will be assumed to be positive definite and thus invertible. Mathematical derivations of the illustrated properties can be found in Additional file 2.

Properties of GBLUP

We start with the effect of translations of the coding, that is the addition of a number p_j to the initially chosen marker coding of marker j.

Property 1 (Translation-invariance of GBLUP) *Let* $\mathbf{P}$ *denote a vector whose entries give the arbitrary translations* p_j *of the coding of the locus j. Moreover, let the ratio of* σ^2_ϵ *and* σ^2_β *be known and unchanged if the marker*

coding is translated. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}$ denote the predicted / estimated quantities if the initial coding **M** *is used in the Mixed Model Equation approach of Eq.* (1) *and let $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mu}$ denote the corresponding quantities if the translation $\tilde{\mathbf{M}} := \mathbf{M} - \mathbf{1P}'$ is used instead of* **M***. Then the following statements hold:*

a) $\tilde{\mu} = \hat{\mu} + \mathbf{P}'\hat{\boldsymbol{\beta}}$
b) $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$
c) *The prediction of the expected phenotype of each genotype is independent of whether* **M** *or* $\tilde{\mathbf{M}}$ *is used.*

The statement of Property 1 has already been discussed in literature [5, 7–9], and we will present a mathematical derivation based on the Mixed Model Equations in Additional file 2. The proof will be a blueprint for the derivation of other properties based on the Mixed Model Equations which can also be found in Additional file 2. Descriptively, we can see the presented invariance with respect to translations the following way: If we change the coding to $\tilde{\mathbf{M}} := \mathbf{M} - \mathbf{1P}'$, then $\tilde{\mathbf{M}}$, $\tilde{\mu} := \hat{\mu} + \mathbf{P}'\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}$ will fit the phenotypes the same way as **M**, $\hat{\mu}$ and $\hat{\boldsymbol{\beta}}$ do. Thus, the prediction of the marker effects and consequently the prediction of the expected phenotypes of individuals will not be affected by the change of coding as long as the method of evaluating the "goodness of fit", that is the penalizing weight in a Ridge Regression approach remains unchanged. For this reason, it is important to note here that we made the precondition that the ratio of the variance components, which defines the penalty for effect size, will not be changed. This guarantees that the method of how to quantify the "goodness of fit" remains the same. In practice this may not exactly be the case if the vector **P** has non-identical entries, that is if the translation of the coding is not equal for all loci, since the variance components are usually estimated from the same data and the translation may have an effect on this estimation. However, this effect has been assessed as being negligible in practice [9]. To assess this problem from a theoretical point of view, without preconditions on the changes of σ_i^2, the method for determining the variance components has to be taken into account to see whether a change in the marker coding has an influence on the ratio of the determined variance components. The next property considers the effect of rescaling the given marker coding.

Property 2 (Scaling invariance of GBLUP) *Let $\hat{\boldsymbol{\beta}}$, $\hat{\mu}$, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mu}$ denote the quantities as defined in Property 1 with $\tilde{\mathbf{M}} := c\mathbf{M}$ for a $c \neq 0$. Moreover, let σ_ϵ^2 and σ_β^2 for* **M** *be known and let the variance components used for the Ridge Regression approach based on $\tilde{\mathbf{M}}$ fulfill $\frac{\tilde{\sigma}_\epsilon^2}{\tilde{\sigma}_\beta^2} = c^2\frac{\sigma_\epsilon^2}{\sigma_\beta^2}$. Then the following statements hold:*

a) $\tilde{\mu} = \hat{\mu}$
b) $\tilde{\boldsymbol{\beta}} = c^{-1}\hat{\boldsymbol{\beta}}$
c) *The prediction of the expected phenotype of each genotype is independent of whether* **M** *or* $\tilde{\mathbf{M}}$ *is used.*

An important aspect of Property 2 is the precondition that the ratio of the variance components is adapted. In practice, when σ_β^2 is estimated, we can assume that this circumstance will approximately be given, however, we have to highlight again that this also depends on the method of how the variance components are determined.

Epistasis models of shape of Eq. (2)

The full EGBLUP model of Eq. (2) adds interaction terms of shape $h_{j,k}M_{i,j}M_{i,k}$ to the additive model of Eq. (1). We will focus on the properties of these additional terms in the following. Evidently, the product structure of the additional covariates generates a dependence of the underlying effect model on the marker coding. In particular, the genotype coded as zero has a special role. If $M_{i,j}$ equals zero, the whole term $h_{j,k}M_{i,j}M_{i,k}$ will be equal to zero, independently of the values of $h_{j,k}$ and $M_{i,k}$. Thus, the model has the implicit assumption that a certain set of combinations do not interact. The marker coding decides which interactions are different from zero a priori and which combinations are clustered. For instance, for the coding $\{-1, 0, 1\}$ for the genotypes $\{aa, aA, AA\}$ of a diploid organism, any interaction with a heterozygous locus will be zero, whereas the interactions with the homozygous locus *aa* will be zero if the coding $\{0, 1, 2\}$ is used. Table 1 illustrates the differences of the two different standard codings ($\{-1, 0, 1\}$ vs. $\{0, 1, 2\}$). Here we see that the marker coding $\{0, 1, 2\}$ implies that the effect is monotonously increasing (or decreasing if $h_{j,k}$ is negative) with the distance from the origin, whereas the coding $\{-1, 0, 1\}$ gives a different topology by only giving weight to the double homozygous. It is not obvious which coding is to be preferred and which reasonable assumptions on the effect of pairs can be made. In the following, we will discuss theoretical properties of the model induced by the marker coding.

As a first important observation, we note that the codings $\{-1, 0, 1\}$ and $\{0, 1, 2\}$ are translations of each other. Their very different interaction effect topologies illustrate that the epistasis model is not invariant with respect to

Table 1 Comparison of the interaction effects which are given implicitly by the marker coding $\{-1, 0, 1\}$ (left) and $\{0, 1, 2\}$ (right) in the interaction terms of EGBLUP. Each entry has to be multiplied with the interaction effect $h_{j,k}$

	aa	aA	AA
bb	1	0	-1
bB	0	0	0
BB	-1	0	1

	aa	aA	AA
bb	0	0	0
bB	0	1	2
BB	0	2	4

translations. This fact that translations modify the model also makes obvious that by subtracting the matrix $\mathbf{1P}'$ with $\mathbf{P}$ containing the allele frequencies of the respective marker, which is the standard normalization in the additive model [6], we will change the coding for the markers according to their frequencies and thus implicitly use different effect models for each pair of loci. We do not see a theoretical basis for this discrimination in an infinitesimal model without additional prior knowledge and therefore will consider mainly models which treat markers equally. Moreover, as gene frequencies are sometimes poorly estimated and very influential, avoiding their use seems to be appealing.

As illustrated, the epistasis model is not invariant with respect to translations, but we show now that the previously described invariance with respect to rescaling persists also for the epistatis model.

Property 3 (Scaling invariance of EGBLUP) *Let $\hat{\boldsymbol{\beta}}$, $\hat{\mu}$, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mu}$ denote the quantities as defined in Property 1 with $\tilde{\mathbf{M}} := c\mathbf{M}$ for a $c \neq 0$. Moreover, let $\hat{\boldsymbol{h}}$ and $\tilde{\boldsymbol{h}}$ denote the corresponding predictions for the interaction effects. Let σ_ϵ^2, σ_β^2, σ_h^2 for $\mathbf{M}$ be known and let the variance components used for the Ridge Regression approach based on $\tilde{\mathbf{M}}$ fulfill $\frac{\tilde{\sigma}_\epsilon^2}{\tilde{\sigma}_\beta^2} = c^2\frac{\sigma_\epsilon^2}{\sigma_\beta^2}$ and $\frac{\tilde{\sigma}_\epsilon^2}{\tilde{\sigma}_h^2} = c^4\frac{\sigma_\epsilon^2}{\sigma_h^2}$. Then the following statements hold:*

a) $\tilde{\mu} = \hat{\mu}$
b) $\tilde{\boldsymbol{\beta}} = c^{-1}\hat{\boldsymbol{\beta}}$
c) $\tilde{\boldsymbol{h}} = c^{-2}\hat{\boldsymbol{h}}$
d) The prediction of the expected phenotype of each genotype is independent of whether $\mathbf{M}$ or $\tilde{\mathbf{M}}$ is used.

A formal derivation of this property based on the Mixed Model Equations can be found in the Additional file 2, but the statements are also plausible if we follow the descriptive argumentation for the invariance of the additive model: If $\hat{\mu}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{h}}$ fit the phenotypic data best when marker matrix $\mathbf{M}$ is used, $c^{-1}\hat{\boldsymbol{\beta}}$ and $c^{-2}\hat{\boldsymbol{h}}$ will fit the phenotypic data the same way if $\mathbf{M}$ is substituted by $\tilde{\mathbf{M}}$ in Eq. (2) (for any constant $c \neq 0$). The important precondition is that the penalizing weight, which defines which fit is "best", is adapted. A question that might come up in the context of Properties 2 and 3 is whether we could also multiply each coding for locus j with its own constant $c_j \neq 0$, similar to what we had for Property 1 and vector $\mathbf{P}$. A problem that will appear here is that the variance of the marker effects will not be changed uniformly and thus, we cannot simply adapt the variance components to cancel the impact of rescaling. An individual rescaling and thus weighting of each marker [30], as well as a completely individual coding of each genotype of each locus, without the side conditions that the differences in the coding of the heterozygous and the two homozygous genotypes are identical across all loci or at least symmetric for each locus [12, 13], indeed has an impact on the predictive ability of the models, in particular also on that of GBLUP. However, the variance components σ_i^2 can be globally adapted to cancel the impact of a non-uniform rescaling of the marker coding, in case that some columns of $\mathbf{M}$ are multiplied with c and the others with $-c$ (due to the assumption of all effects being symmetrically distributed around mean zero). An adapted sign of the effects also allows the predicted effect model to remain unchanged.

Permuting the role of the alleles at locus *j*. Let locus j have the possible allele configurations aa, aA and AA. The prediction performance of GBLUP is unaffected by the choice of whether the allele variant a or A is counted, since we can express a permutation of the initial coding $\{0, 1, 2\}$ by a translation by -2 and a multiplication of the coding by -1.

Obviously, this argumentation cannot be used for the epistasis model, since we do not have the possibility to translate the marker coding. This fact raises the question under which circumstances the epistasis EGBLUP model is unaffected by a permutation of the role of the allele variants.

Property 4 (Symmetric role of the alleles in EGBLUP) *Let us consider locus j with alleles a and A and locus k with alleles b and B (of a diploid organism). Let us use the same coding for both loci and let the three variants of aa, aA and AA be coded by three different numbers $M_{aa} < M_{aA} < M_{AA}$ (or $M_{aa} > M_{aA} > M_{AA}$). The only coding for the epistasis terms, whose corresponding effect model on the tuples*

$$\{(j,k)|j \in \{aa, aA, AA\}, k \in \{bb, bB, BB\}\}$$

is invariant with respect to a permutation of the role of allele a and A satisfies $-M_{aa} = M_{AA}$ and $M_{aA} = 0$. Analogously, for markers with only two possible values, the coding has to satisfy $-M_a = M_A$.

Property 4 is of central theoretical importance since it implies that the only coding for $\{0, 1\}$ marker in EGBLUP, which is invariant with respect to a permutation of the meaning of 0 and 1 is the coding $\{-c, c\}$ ($c \neq 0$). Moreover, if EGBLUP shall possess this reasonable property for markers with three possible values, we have to use the coding $\{-c, 0, c\}$. We will give an example to illustrate why this property is important for determining marker effects and thus why it may also be important for the overall predictive ability of the model.

Example 1 (Marker effects and quadratic loss) *Let us consider markers with two possible variants and let us*

assume that for each pair of markers, the correct underlying weights of the combinations is given by a coding as $\{0,1\}$. *We use a* $\{0,1\}$ *coding, but we do not know which variants of the two loci have to be coded as 1 to capture the real effect distribution. We assume that we decide which allele is coded as zero, by drawing independently from a Bernoulli-distribution with* $p = 0.5$ *for each marker. To see how good the real underlying weight distribution is captured, we measure the quadratic loss between the best possible fit and the real underlying weights. Let the coding*

$$\begin{array}{c|c|c} & a & A \\ b & 0 & 0 \\ B & 0 & 1 \end{array} \tag{4}$$

be the correct underlying effect distribution, with the corresponding underlying interaction effect equal to 1 (the problem remains the same if the underlying interaction effect is multiplied with any number $c \neq 0$*). With a probability of 0.25, we will code both markers j and k correctly and minimize the distance to zero by predicting* $\hat{h}_{j,k} = 1$. *However, with a probability of 0.75, we will make a mistake and choose an incorrect orientation, which means an incorrect underlying parametric model, such as*

$$\begin{array}{c|c|c} & a & A \\ b & 1 \cdot h_{j,k} & 0 \\ B & 0 & 0 \end{array} \tag{5}$$

In this situation, we can determine the optimally fitting interaction $\hat{h}_{j,k}$, *which describes the distribution of Eq.* (4) *best, when model Eq.* (5) *is used, by minimizing the quadratic Euclidean distance between both effect distributions. In more detail, using a minimal quadratic loss means we have to find an* $\hat{h}_{j,k}$ *which minimizes the quadratic distance between the matrices of Eq.* (4) *and Eq.* (5)*:*

$$(1h_{j,k} - 0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 \tag{6}$$

which is equal to

$$h_{j,k}^2 + 1.$$

Thus, the optimal $\hat{h}_{j,k}$ *minimizing Eq.* (6) *is 0 and the expected quadratic loss when the right coding with unknown orientation is used, is* $0.25 \cdot 0 + 0.75 \cdot 1 = 0.75$.

Analogously, if we use the coding $\{-1,1\}$ *instead of Eq.* (5)*, we will obtain the quadratic distance*

$$3(h_{j,k}-0)^2+(h_{j,k}-1)^2 \qquad or \qquad 3(h_{j,k}-0)^2+(h_{j,k}+1)^2$$

each with probability 0.5, depending on whether -1 *or* $+1$ *coincides with the 1 of the real underlying effects. Consequently, the minimum quadratic distance is 0.75 with probability 1, for* $\hat{h}_{j,k} = \pm 0.25$. *Thus, in this example, even though the coding* $\{-1,1\}$ *specifies a model which is surely wrong, the average quadratic loss is equal to the situation in which we know the exact shape of the effect distribution but not its orientation. If the real underlying effect distribution deviates from the* $\{0,1\}$ *coding of Eq.* (4)*, the possibility to adapt the orientation might be even more important.*

Example 1 illustrated that the expected quadratic loss of the estimated marker-pair weights is equal for the codings $\{-1,1\}$ and $\{0,1\}$ even in the case that the underlying effects are a version of the latter one but with unknown orientation. Moreover, we can observe the following: Let us assume that the real underlying interactions (j,k), (j,l) and (k,l) of the three loci j,k,l are described by certain $\{0,1\}$-codings, meaning that one certain configuration has an interaction effect but the others do not. Given the underlying effects, we can adapt the coding of j,k and l by considering the effects of the pairs (j,k), (j,l). However, then the effect distribution within the model is also determined for the pair (k,l), because the marker coding has already been fixed. This configuration does not necessarily describe the interaction of (k,l) well. This fact illustrates that due to the way of how interactions are incorporated into the model in EGBLUP, the model with an asymmetric coding lacks a full flexibility to adapt to any situation. This problem does not appear with the symmetric coding, since the model is independent of the decision which allele is coded as ± 1. However, there are also good reasons for choosing other types of coding. Firstly, it is not clear whether the effect that we have illustrated on the level of marker effects and quadratic loss, also translates to the level of prediction of genetic values. In the latter approach, all effects are predicted simultaneously and thus errors of individual effects can cancel out in the sum. Secondly, from a biological point of view, the symmetric coding seems inadequate: Let us consider markers with two variants and let the two loci j and k have the possible variants a,A and b,B, respectively. The symmetric coding $\{-1,1\}$ assigns the weight $1h_{j,k}$ to the combinations (a,b) and (A,B), meaning that the most distant genotypes, which do not share any allele, are treated as being equal in the model. Thus, overall, it is not clear which coding will be most appropriate in general. Especially in situations in which additional information on the nature of the marker or the biology of the trait is available, this information may be used to specify the effect model. In the next paragraph, we illustrate how much freedom the marker coding gives to specify the model.

Finding the marker coding for an a priori specified model. Let us consider a model with identical marker coding M_{aa}, M_{aA} and M_{AA} for each locus. Then the weights in the model are given by

$$\begin{aligned} a_{1,1} &= M_{aa}^2 \;\; a_{1,2} = M_{aa}M_{aA} \;\; a_{1,3} = M_{aa}M_{AA} \\ a_{2,2} &= M_{aA}^2 \;\; a_{2,3} = M_{aA}M_{AA} \;\; a_{3,3} = M_{AA}^2. \end{aligned} \tag{7}$$

If we want to predefine the weights $a_{r,s}$ and calculate a corresponding coding, we see that not all choices of weights can be translated into a coding for the epistasis model of Eq. (2) since contradictions can arise. However, the following statement holds:

Property 5 *Let three weights $a_{r,s}$ of Eq.* (7) *which include the three variables M_{aa}, M_{aA}, M_{AA} in at least one weight $a_{r,s}$ be given by arbitrary nonzero numbers. Then the marker codings as well as the other weights are determined up to their signs.*

Categorical effect models

In the following, we discuss categorical effect models in which we do not treat the marker data as numerical dosage, but as categorical variables. The goal is to build an epistasis model without the undesired properties of EGBLUP which have been described previously. We model the effects of allele *combinations* as being independently drawn from a Gaussian distribution with mean zero. For instance, for an additive marker effect model, the effects of *aa*, *aA* and *AA* are independently originating from the same distribution. For the analogous epistasis model, the effect of each combination of the alleles of two loci is drawn independently from the same distribution. We will introduce dummy $\{0, 1\}$ variables to indicate which allele configuration is present and thus inflate the number of variables in our model. The important fact to notice in this context is that we can use a relationship matrix approach for genomic prediction (see "Methods") and thus do not need to handle the high number of variables. This procedure also reduces computation time compared to the effect based approach. All considered effects β_j of the variables are assumed to come from the same distribution: $\beta_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_\beta)$.

A categorical marker effect model (CM) The underlying concept of this model is to code the configurations *aa*, *aA*, *AA* of locus *j* as three different variables. The effect of each genotype is estimated on its own. The assumption of a constant allele substitution effect, that is that the effect of *AA* equals twice the effect of *A*, which is made in the additive numerical GBLUP model, is not made here (see Fig. 1). We translate the genotypes (*aa*, *aA*, *AA*) which can be found at locus *j* to $((0,0,1),(0,1,0),(1,0,0))$. The latter triples indicate which of the three states is present. A genotype of three loci described by $(2, 0, 1)$ in the numerical GBLUP coding, will here be coded by the nine-tuple $(1,0,0,0,0,1,0,1,0)$ (a triple for each locus, describing its state). We then simply use model Eq. (1) with the new coding. Advantages of this model are that it is also invariant to an exchange of the role of *a* and *A* (as GBLUP of Eq. (1) is as well), since we will only permute the meaning of the positions in the triple but change their entries accordingly. Moreover, we can account for dominance by estimating each effect on its own. A disadvantage is the increased number of variables but this can be overcome easily by the use of relationship matrices for genomic prediction. Property 6 describes the relation between the CM model and GBLUP for markers with only two possible values:

Property 6 (GBLUP and CM for markers with two possible states) *For markers with only two possible states, let* $\mathbf{M}$ *denote the $n \times p$ marker matrix in the $\{-1, 1\}$ coding. The relationship matrix of GBLUP is given by (a rescaled version of) $\mathbf{MM}'$. Moreover, let $\mathbf{C}$ be the relationship matrix of the CM model. Then*

$$\mathbf{C} = 0.5(\mathbf{MM}' + \mathbf{J}_{n\times n}p) \tag{8}$$

where p is the number of markers and $\mathbf{J}_{n\times n}$ the $n \times n$ matrix with each entry equal to 1.

The linear relationship of the covariance matrices demonstrated in Property 6 implies that the prediction performances of GBLUP and CM are identical for markers with only two possible values.

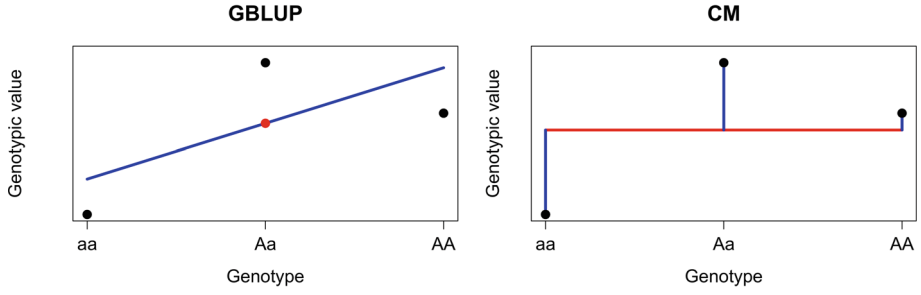

Fig. 1 Comparison of the parametrization of the genotypic values in GBLUP and the categorical marker effect model CM: Black dots: genotypic values of the corresponding genotype of a certain locus. GBLUP parameterizes the genotypic values by a fixed effect (red dot) and a random effect determining the slope (*blue* line), whereas CM parameterizes by the fixed effect (*red* line) and independent random effects (*blue* lines) for each genotype

Property 7 (Equivalence of GBLUP and CM for markers with two possible states) *Let us assume that the ratio of the variance components is fixed such that Property 1 holds for the CM model. Then GBLUP and the CM model are identical for markers with only two possible values.*

A categorical epistasis model (CE) Analogously to the CM model, we translate the genotype of pairs of loci, e.g. (aA, bb) into $\{0,1\}$-tuples. Here, a nine-tuple indicates which *combination* of alleles of two loci is present. To translate the genotype $(2,0,1)$ of the numerical $\{0,1,2\}$ coding into the CE coding, we have to translate each marker pair. Each pair is coded by a nine-tuple with only one entry equal to 1 which indicates the configuration:

$$\Big(\underbrace{\bullet}_{(2,2)},\underbrace{\bullet}_{(2,1)},\underbrace{\bullet}_{(2,0)},\underbrace{\bullet}_{(1,2)},\underbrace{\bullet}_{(1,1)},\underbrace{\bullet}_{(1,0)},\underbrace{\bullet}_{(0,2)},\underbrace{\bullet}_{(0,1)},\underbrace{\bullet}_{(0,0)}\Big). \tag{9}$$

The assignment of the configuration of the respective marker pair to the position of the nine-tuple can be chosen arbitrarily but has of course to be used consistently for all individuals. Let us assume that we have three subsequent loci with genotypes $(2,0,1)$ in the ordinary numerical coding. Then, there are three possible interactions: the first two loci have the combination $(2,0)$ which will be coded as $(0,0,1,0,0,0,0,0,0)$. Additionally, the second pair is $(2,1)$ which will be coded as $(0,1,0,0,0,0,0,0,0)$, whereas the last pair $(0,1)$ is translated to $(0,0,0,0,0,0,0,1,0)$. As already mentioned, an obvious disadvantage of the model is the high number of variables, but we do not have to solve the system for these variables to perform genomic prediction, since we can use equivalent genomic relationship matrices. Moreover, this model eliminates several disadvantages of EGBLUP: i) The model is invariant with respect to the decision which allele is used as reference ("orientation"), since it is based on categorical variables indicating which genotype is present, ii) the effects the model can assign to different pairs of loci are not connected between pairs by their respective codings (as described for the asymmetrically coded EGBLUP after Example 1), and iii) compared to the symmetric $\{-1,0,1\}$ coding of EGBLUP, CE does not generally assign the same effects to the most different allele combinations.

Relationship matrices for the respective marker models

Let $\mathbf{M}$ be the marker matrix of the respective numerical coding (0,1,2 or $-1,0,1$). In the following, we will present the corresponding relationship matrices for each model.

GBLUP. The relationship matrix for the GBLUP model is given by $\mathbf{MM}'$ (the $n \times p$ genotype matrix multiplied with its transposed version).

Epistasis models based on Eq. (2). The relationship matrix corresponding to the interactions of Eq. (2) where $j \geq k$ is given by

$$\mathbf{H} = 0.5\left(\mathbf{MM}' \circ \mathbf{MM}'\right) + 0.5\left(\mathbf{M} \circ \mathbf{M}\right)\left(\mathbf{M} \circ \mathbf{M}\right)'. \tag{10}$$

(for a derivation of this statement see [11]). Note here again that the GBLUP model is not affected by a translation of the coding in $\mathbf{M}$, but the performance of EGBLUP is affected.

The categorical marker (CM) effect model The i,l-th entry of the corresponding relationship matrix $\mathbf{C}$ is given by the inner product of the vectors of the genotypes of individuals i and l in the coding of the CM model. This means that we count the number of loci which have the same configuration. For markers with two possible variants and the marker data in dosage 0,1 coding, we can express the i,l-th entry of $\mathbf{C}$ the following way:

$$C_{i,l} = p - \sum_{j=1}^{p} \left|M_{i,j} - M_{l,j}\right| \tag{11}$$

Analogously, for markers with three different variants, we have to count the number of zeros in the marker vectors $\mathbf{M_{i,\bullet}} - \mathbf{M_{l,\bullet}}$ (For the relation of Eqs. (11) and (8), see the derivation of Eq. (8) in Additional file 2).

The categorical epistasis (CE) model The i,l-th entry of the corresponding relationship matrix $\mathbf{C_E}$ is given by the inner product of the genotypes i, l in the coding of the categorical epistasis model. Thus, the matrix counts the number of pairs which are in identical configuration and we can express the entry $C_{Ei,l}$ in terms of $C_{i,l}$ since we can calculate the number of identical pairs from the number of identical loci:

$$C_{Ei,l} = \sum_{k=1}^{C_{i,l}} k = 0.5 C_{i,l}\left(C_{i,l}+1\right) \tag{12}$$

Here, we also count the "pair" of a locus with itself by allowing $k \in \{1,\ldots,C_{i,l}\}$. Excluding these effects from the matrix would mean, the maximum of k equals $C_{i,l}-1$. In matrix notation Eq. (12) can be written as

$$\mathbf{C}_E = 0.5\mathbf{C} \circ \mathbf{C} + 0.5\mathbf{C} \tag{13}$$

Note here, that the relation between GBLUP and the epistasis terms of EGBLUP is identical to the relation of CM and CE in terms of relationship matrices: For $\mathbf{G} = \mathbf{MM}'$ and $\mathbf{M}$ a matrix with entries only 0 or 1, Eq. (10) gives Eq. (13) with $\mathbf{C} = \mathbf{G}$ and $\mathbf{C}_E = \mathbf{H}$.

Remark 1 (The Gaussian kernel) *Additionally to the previously discussed EGBLUP model, a common approach to incorporate "non-linearities" is based on Reproducing Kernel Hilbert Space regression [21, 31] by modeling*

the covariance matrix as a function of a certain distance between the genotypes. The most prominent variant for genomic prediction is the Gaussian kernel. Here, the covariance $Cov_{i,l}$ of two individuals is described by

$$Cov_{i,l} = \exp(-b \cdot d_{i,l}),$$

with $d_{i,l}$ being the squared Euclidean distance of the genotype vectors of individuals i and l, and b a bandwidth parameter that has to be chosen. This approach is independent of translations of the coding, since the Euclidean distance remains unchanged if both genotypes are translated. Moreover, this approach is also invariant with respect to a scaling factor, if the bandwidth parameter is adapted accordingly (in this context see also [32]*). Thus, EGBLUP and the Gaussian kernel RKHS approach capture both "non-linearities" but they behave differently if the coding is translated.*

Comparison of the performance of the models on different data sets

Results on the simulated data For 20 independently simulated populations of 1 000 individuals, we modeled three scenarios of qualitatively different genetic architecture (purely additive A, purely dominant D and purely epistatic E) with increasing number of involved QTL (see "Methods") and compared the performances of the considered models on these data. In more detail, we compared GBLUP, a model defined by the epistasis terms of EGBLUP with different codings, the categorical models and the Gaussian kernel with each other. All predictions were based on one relationship matrix only, that is in the case of EGBLUP on the interaction effects only. The use of two relationship matrices did not lead to qualitatively different results (data not shown), but can cause numerical problems for the variance component estimation if both matrices are too similar. For each of the 20 independent simulations of population and phenotypes, test sets of 100 individuals were drawn 200 times independently, and Pearson's correlation of phenotype and prediction was calculated for each test set and model. The average predictive abilities of the different models across the 20 simulations are summarized in Table 2 in terms of empirical mean of Pearson's correlation and its average standard error. Comparing GBLUP to EGBLUP with different marker codings, we see that the predictive ability of EGBLUP is very similar to that of GBLUP, if a coding which treats each marker equally is used. Only the EGBLUP version, standardized by subtracting twice the allele frequency as it is done in the commonly used standardization for GBLUP [6], shows a drastically reduced predictive ability for all scenarios (see Table 2, EGBLUP VR). Moreover, considering the categorical models, we see that CE is slightly better than CM and that both categorical models perform better than the other models in the dominance and epistasis scenarios.

Results on the wheat data For EGBLUP, we used here the coding $\{0, 1\}$ which was originally used in the data of the publication, a translation by -1 which leads to $\{-1, 0\}$ representing a coding in which the meaning of 0 and 1 is permuted, and a centered version $\{-1, 1\}$. Moreover, we used the standardization by allele frequencies [6] to calculate EGBLUP. Additionally, we evaluated CM, CE and reevaluated the Gaussian kernel RKHS approach, previously used by Crossa et al. [21] (we used the matrix **K** obtained from the supplementary of the corresponding publication). The results are summarized in Table 3. CM showed exactly identical results to those of GBLUP (which has already been stated theoretically by Property 7) and

Table 2 Predictive abilities of the models on the simulated data. Comparison of the predictive abilities in terms of correlations between the measured phenotypes and the predictions for the individuals of the test sets ("Pearson's correlation"; 100 test set genotypes were drawn randomly from all 1000 genotypes; 200 repeats for each simulated population; 20 independent simulations of population and phenotypes). Traits of different genetic architecture (additive A, dominant D, Epistasis E) and increasing number of QTL. Model abbreviations as introduced in the text. For EGBLUP, only the matrix based on the interactions was considered here

	GBLUP	EGBLUP 0,1,2	EGBLUP -2,-1,0	EGBLUP -1,0,1	EGBLUP VR	CM	CE	K
A1	0.551 ± 0.005	$\mathbf{0.552 \pm 0.005}$	$\mathbf{0.552 \pm 0.005}$	0.550 ± 0.005	0.372 ± 0.006	0.489 ± 0.005	0.494 ± 0.005	0.530 ± 0.005
A2	0.549 ± 0.005	$\mathbf{0.550 \pm 0.005}$	$\mathbf{0.550 \pm 0.005}$	0.548 ± 0.005	0.351 ± 0.006	0.486 ± 0.005	0.490 ± 0.005	0.527 ± 0.005
A3	0.569 ± 0.005	$\mathbf{0.570 \pm 0.005}$	$\mathbf{0.570 \pm 0.005}$	0.568 ± 0.005	0.372 ± 0.006	0.500 ± 0.005	0.504 ± 0.005	0.545 ± 0.005
D1	0.159 ± 0.006	0.160 ± 0.006	0.159 ± 0.006	0.161 ± 0.007	0.111 ± 0.007	0.174 ± 0.006	$\mathbf{0.175 \pm 0.006}$	0.162 ± 0.006
D2	0.172 ± 0.006	0.172 ± 0.006	0.172 ± 0.006	0.171 ± 0.006	0.103 ± 0.006	$\mathbf{0.186 \pm 0.006}$	$\mathbf{0.186 \pm 0.006}$	0.170 ± 0.006
D3	0.156 ± 0.006	0.156 ± 0.006	0.156 ± 0.006	0.158 ± 0.006	0.116 ± 0.006	0.177 ± 0.006	$\mathbf{0.179 \pm 0.006}$	0.160 ± 0.006
E1	0.244 ± 0.006	0.244 ± 0.006	0.244 ± 0.006	0.244 ± 0.006	0.159 ± 0.006	$\mathbf{0.258 \pm 0.006}$	$\mathbf{0.258 \pm 0.006}$	0.243 ± 0.006
E2	0.275 ± 0.006	0.276 ± 0.006	0.276 ± 0.006	0.277 ± 0.006	0.188 ± 0.006	0.301 ± 0.006	$\mathbf{0.302 \pm 0.006}$	0.277 ± 0.006
E3	0.279 ± 0.006	0.278 ± 0.006	0.279 ± 0.006	0.278 ± 0.006	0.176 ± 0.006	$\mathbf{0.304 \pm 0.006}$	$\mathbf{0.304 \pm 0.006}$	0.276 ± 0.006

EGBLUP VR denotes the interaction model based on the by allele frequencies standardized matrix. The given values represent the empirical mean and the corresponding mean standard error across the 20 independently simulated data sets. The highest predictive ability is bold

Table 3 Predictive abilities of the models on the wheat data. Comparison of the predictive abilities as Pearson's correlation of the measured phenotypes and the predictions for the individuals of the test sets (60 test set genotypes, trait: grain yield)

	GBLUP	EGBLUP 0,1	EGBLUP -1,0	EGBLUP -1,1	EGBLUP VR	CE	Gaussian kernel
Environment 1	0.511^{a}	0.554^{bc}	0.561^{bcd}	0.581^{cd}	0.541^{b}	0.558^{bcd}	$\mathbf{0.584}^{d}$
Environment 2	0.499^{a}	0.502^{a}	$\mathbf{0.504}^{a}$	0.495^{a}	0.422^{b}	$\mathbf{0.504}^{a}$	0.500^{a}
Environment 3	0.371^{a}	0.390^{ab}	0.396^{ab}	0.409^{b}	0.365^{a}	0.393^{ab}	$\mathbf{0.422}^{b}$
Environment 4	0.463^{a}	0.498^{b}	0.504^{bc}	0.530^{c}	0.500^{b}	0.502^{b}	$\mathbf{0.531}^{c}$

Letters indicate groups that were not distinguishable at a 5% significance level in a Tukey's 'Honest Significant Difference' test

is therefore not listed separately. Considering the predictive ability of EGBLUP with different codings, a first thing to note is that the variability among the EGBLUP variants is higher than that found on the simulated data. Moreover, with the data sets of environments 1, 3 and 4, EGBLUP tends to outperform GBLUP. Among them, the model with symmetric $\{-1, 1\}$ coding performs best and the VanRaden standardized version of EGBLUP has a significantly reduced predictive ability for the data of environments 1, 2 and 3, which is analogous to what we have already seen on the simulated data. Moreover, the predictive ability of EGBLUP with symmetric coding seems to be closest to that of the Gaussian kernel. For the data of environment 2, no big differences in the performance of the models (except for the allele frequency standardized EGBLUP) can be observed. Overall, the Gaussian kernel RKHS method performs best on this data set and the predictive ability of the CE model is on the level of the asymmetrically coded versions of EGBLUP.

Results on the mouse data We compared the models on 13 traits related to obesity, weight and immunology. Instead of the raw phenotypes, we used pre-corrected residuals which are publicly available (see "Methods"). Again, we compared GBLUP, EGBLUP with 0,1,2 coding as well as with inverted, symmetric and by allele frequencies standardized coding, the categorical models and the Gaussian kernel RKHS approach with each other. The results are summarized in Table 4. The general patterns observed on the previously considered data remain the same: Any EGBLUP version treating the markers equally has at least the same predictive ability as GBLUP for all traits. Among them, the symmetric coding seems to perform best. The allele frequency standardized version of EGBLUP has in three of the 13 traits a higher predictive ability than its other versions (W6W, GrowthSlope, CD8Intensity), but a smaller one in ten cases. Considering only significant differences between CM and GBLUP, CM outperforms GBLUP on the traits %CD4/CD3 and %CD8/CD3 and shows a lower predictive ability only for BMI and BodyLength. Moreover, CE outperforms CM slightly. Overall, two traits are predicted best by EGBLUP VR, three traits by CE, and five by the symmetric version of EGBLUP and the Gaussian kernel, respectively.

Table 4 Predictive abilities of the models on the mouse data. Comparison of the predictive abilities as Pearson's correlation of the measured phenotypes and the predictions for the individuals of the test set (130 test set genotypes). Here, the already for fixed effects pre-corrected residuals of the phenotypes, which are also provided by the publicly available data, were used

	GBLUP	EGBLUP 0,1,2	EGBLUP -2,-1,0	EGBLUP -1,0,1	EGBLUP VR	CM	CE	Gaussian kernel
W6W	0.493^{ab}	0.540^{c}	0.505^{ad}	0.545^{c}	0.553^{ce}	0.486^{b}	0.514^{d}	$\mathbf{0.565}^{e}$
W10W	0.466^{a}	0.491^{bc}	0.474^{ab}	0.495^{bc}	0.461^{a}	0.466^{a}	0.479^{ab}	$\mathbf{0.503}^{c}$
GrowthSlope	0.347^{a}	0.363^{ab}	0.350^{a}	0.364^{ab}	$\mathbf{0.375}^{b}$	0.355^{ab}	0.363^{ab}	0.371^{b}
BMI	0.195^{a}	0.204^{a}	0.200^{a}	$\mathbf{0.210}^{a}$	0.194^{a}	0.153^{b}	0.166^{b}	$\mathbf{0.210}^{a}$
BodyLength	0.271^{a}	0.282^{a}	0.276^{a}	$\mathbf{0.285}^{a}$	0.275^{a}	0.226^{b}	0.240^{b}	0.284^{a}
%B220	0.549^{ab}	0.573^{cde}	0.556^{abc}	0.576^{de}	0.540^{a}	0.547^{ab}	0.561^{bcd}	$\mathbf{0.579}^{e}$
%CD3	0.522^{a}	0.535^{a}	0.527^{a}	$\mathbf{0.536}^{a}$	0.485^{b}	0.521^{a}	0.528^{a}	0.535^{a}
%CD4	0.495^{a}	0.506^{a}	0.499^{a}	$\mathbf{0.508}^{a}$	0.458^{b}	0.495^{a}	0.502^{a}	0.506^{a}
%CD8	0.694^{a}	0.703^{ab}	0.699^{ab}	0.706^{ab}	0.656^{c}	0.706^{ab}	$\mathbf{0.711}^{b}$	0.702^{ab}
%CD4/CD3	0.643^{a}	0.655^{abc}	0.647^{ab}	0.656^{abc}	0.618^{d}	0.660^{bc}	$\mathbf{0.664}^{c}$	0.653^{abc}
%CD8/CD3	0.683^{a}	0.689^{ab}	0.687^{a}	0.690^{ab}	0.638^{c}	0.701^{b}	$\mathbf{0.702}^{b}$	0.686^{a}
CD4Intensity	0.581^{a}	0.601^{b}	0.587^{ab}	$\mathbf{0.603}^{b}$	0.561^{c}	0.578^{ac}	0.586^{ab}	$\mathbf{0.603}^{b}$
CD8Intensity	0.388^{a}	0.442^{b}	0.401^{a}	0.450^{b}	$\mathbf{0.481}^{c}$	0.406^{a}	0.434^{b}	0.475^{c}

Letters indicate groups that were not distinguishable at a 5% significance level in a Tukey's 'Honest Significant Difference' test
For a description of the traits see the corresponding UCL website which is at the moment http://mtweb.cs.ucl.ac.uk/mus/www/mouse/HS/index.shtml

Incorporating prior experimental information by marker coding

The coding-dependent performance of EGBLUP also offers possibilities to incorporate additional information. He et al. [12, 13] have already illustrated the idea of data-driven coding and we have recently shown that information on the performance of genotypes grown under different environmental conditions can be used to select variables within EGBLUP which then can be used for genome assisted prediction within another environment [11]. Here, we will demonstrate that differential coding is also appropriate to incorporate prior experimental information into EGBLUP. For this, we used the different trait ($\times$ environment) combinations and adapted the marker coding of each pair of loci to the data, following the procedure described in the "Methods" section. Important here is that we decided for each pair of markers individually, which orientation the corresponding coding of the particular pair shall have. The "orientation" of the underlying effect model is chosen for each pair. Thus, we cut the connection between the coding of different pairs. The determined relationship matrices are then used to predict within the data of other traits. The results are summarized in Tables 5 and 6 for the wheat and mouse data sets, respectively. We can see here that adapting the coding to data of previous experiments can be beneficial for the predictive ability. In the case of the wheat data set, Table 5 shows that using the data of grain yield of the genotypes grown in environments 3 and 4 to infer the marker coding for each pair of marker, improves the prediction accuracy in environment 2 to a level higher than that of all methods which do not use the data of other experiments (from 0.504 ± 0.007 to 0.544 ± 0.006). The situation is analogue for the predictive ability in environment 3, if the data of environment 2 is used to infer the relationship matrix. However, the gain in predictive ability resulting from this procedure is relatively small compared to the gain by means of variable selection [11]. Adapting the coding to given data also helped to increase predictive ability on the mouse data (see Tables 4 and 6). For instance, improvements from 0.285 ± 0.006 to 0.313 ± 0.005, from 0.536 ± 0.004 to 0.569 ± 0.004, and from 0.664 ± 0.004 to 0.685 ± 0.003 were reached for the traits BodyLength, %CD3 and %CD4/CD3, respectively.

Discussion

The effect of the choice of marker coding on EGBLUP

We recalled that GBLUP is not sensitive to certain changes of the marker coding if the variance components are adapted accordingly. Analogously, we also proved that the interaction terms of EGBLUP are invariant to factors rescaling the marker coding, but showed that a translation indeed changes the underlying marker effect model drastically. In particular, we demonstrated that the effect model of EGBLUP with the asymmetric 0,1,2 coding is affected by the decision which allele to count. Thus, an important observation concerning EGBLUP is that the only coding allowing a permutation of the roles of the alleles without changing the underlying interaction effect model for the respective marker pair is symmetric around zero. This coding solves the problem of "which allele to count", but we also argued that the symmetric coding appears to be biologically implausible since it assigns the same interaction effect to the most distant genotypes. Concerning the allele frequency adjusted version EGBLUP VR, we illustrated that the different markers are not treated equally and thus that the interaction effect models here depend on the allele frequencies of the involved alleles. On the level of predictive ability, the symmetric coding tends to outperform the asymmetric versions slightly, which can most clearly be seen from the data of environment 1 and 4 of the wheat data set (Table 3). Also with the mouse data set, the symmetric coding had a higher predictive ability than the other codings treating all loci equally for all traits, but the improvements were most often very small. Concerning the allele-frequencies standardized version EGBLUP VR, we observed a drastic reduction in the predictive ability compared to other EGBLUP versions in most of the examples. Illustratively, one reason for the comparatively poor performance can be seen in the following: the relationship matrix corresponding to the interaction effects of EGBLUP in a certain coding is basically the GBLUP relationship matrix, but with each of its entries squared (if all pairwise interactions

Table 5 Predictive abilities on the wheat data when prior information is incorporated in the marker coding of EGBLUP. Predictive abilities when the coding for each interaction is determined based on records under different environmental conditions

	G-Env 1	G-Env 2	G-Env 3	G-Env 4
Environment 1	——	0.555 ± 0.007	0.559 ± 0.007	0.552 ± 0.007
Environment 2	0.503 ± 0.007	——	$\mathbf{0.544 \pm 0.006}$	$\mathbf{0.514 \pm 0.007}$
Environment 3	0.394 ± 0.008	$\mathbf{0.430 \pm 0.008}$	——	0.402 ± 0.008
Environment 4	0.500 ± 0.007	0.511 ± 0.006	0.513 ± 0.006	——

G-Env 1 means that the relationship matrix was constructed under the use of the data of Environment 1 (analogously for other environments; for a description of the construction of the matrices see section "Methods"). Bold numbers indicate predictive abilities higher than that of all previously used methods for this trait

Table 6 Predictive abilities on the mouse data when prior information is incorporated in the marker coding of EGBLUP. Predictive abilities when the coding for each interaction is determined based on the records of other traits

	G-W6W	G-W10W	G-GrowthSlope	G-BMI	G-BodyLength	G-%B220	
W6W	——	0.548 ± 0.004	0.511 ± 0.004	0.507 ± 0.004	0.511 ± 0.004	0.507 ± 0.004	
W10W	**0.519 ± 0.005**	——	0.480 ± 0.005	0.475 ± 0.005	0.475 ± 0.005	0.474 ± 0.005	
GrowthSlope	0.356 ± 0.005	0.355 ± 0.005	——	0.351 ± 0.005	0.355 ± 0.005	0.351 ± 0.005	
BMI	0.202 ± 0.006	0.202 ± 0.006	0.200 ± 0.006	——	**0.243 ± 0.006**	0.200 ± 0.006	
BodyLength	0.283 ± 0.006	0.278 ± 0.006	0.281 ± 0.006	**0.313 ± 0.005**	——	0.276 ± 0.006	
%B220	0.557 ± 0.004	0.557 ± 0.004	0.557 ± 0.004	0.556 ± 0.004	0.556 ± 0.004	——	
%CD3	0.527 ± 0.004	0.527 ± 0.004	0.527 ± 0.004	0.527 ± 0.004	0.527 ± 0.004	**0.562 ± 0.004**	
%CD4	0.500 ± 0.004	0.500 ± 0.004	0.499 ± 0.004	0.499 ± 0.004	0.500 ± 0.004	**0.530 ± 0.004**	
%CD8	0.701 ± 0.003	0.701 ± 0.003	0.700 ± 0.003	0.700 ± 0.003	0.699 ± 0.003	0.708 ± 0.003	
%CD4/CD3	0.649 ± 0.004	0.649 ± 0.004	0.648 ± 0.004	0.648 ± 0.004	0.647 ± 0.004	0.648 ± 0.004	
%CD8/CD3	0.688 ± 0.003	0.688 ± 0.003	0.687 ± 0.003	0.687 ± 0.003	0.686 ± 0.003	0.687 ± 0.003	
CD4Intensity	0.589 ± 0.004	0.588 ± 0.004	0.588 ± 0.004	0.588 ± 0.004	0.588 ± 0.004	0.588 ± 0.004	
CD8Intensity	0.406 ± 0.005	0.405 ± 0.005	0.404 ± 0.005	0.405 ± 0.005	0.405 ± 0.005	0.404 ± 0.005	

	G-%CD3	G-%CD4	G-%CD8	G-%CD4/CD3	G-%CD8/CD3	G-CD4Intensity	G-CD8Intensity
W6W	0.507 ± 0.005	0.507 ± 0.005	0.507 ± 0.005	0.507 ± 0.005	0.507 ± 0.004	0.507 ± 0.005	0.508 ± 0.005
W10W	0.475 ± 0.005	0.475 ± 0.005	0.475 ± 0.005	0.475 ± 0.005	0.475 ± 0.005	0.475 ± 0.005	0.476 ± 0.005
GrowthSlope	0.351 ± 0.005	0.351 ± 0.005	0.351 ± 0.005	0.351 ± 0.005	0.351 ± 0.005	0.351 ± 0.005	0.351 ± 0.005
BMI	0.200 ± 0.006	0.200 ± 0.006	0.201 ± 0.006	0.201 ± 0.006	0.201 ± 0.006	0.200 ± 0.006	0.202 ± 0.006
BodyLength	0.276 ± 0.006	0.276 ± 0.006	0.276 ± 0.006	0.276 ± 0.006	0.276 ± 0.006	0.276 ± 0.006	0.277 ± 0.006
%B220	**0.588 ± 0.004**	**0.582 ± 0.004**	0.570 ± 0.004	0.557 ± 0.004	0.557 ± 0.004	0.556 ± 0.004	0.558 ± 0.004
%CD3	——	**0.569 ± 0.004**	**0.550 ± 0.004**	0.527 ± 0.004	0.527 ± 0.004	0.527 ± 0.004	0.527 ± 0.004
%CD4	**0.545 ± 0.004**	——	0.504 ± 0.004	**0.511 ± 0.004**	**0.510 ± 0.004**	0.500 ± 0.004	0.499 ± 0.004
%CD8	**0.714 ± 0.003**	0.702 ± 0.003	——	**0.722 ± 0.003**	**0.726 ± 0.003**	0.700 ± 0.003	0.7 ± 0.003
%CD4/CD3	0.649 ± 0.004	0.656 ± 0.004	**0.672 ± 0.004**	——	**0.685 ± 0.003**	0.649 ± 0.004	0.649 ± 0.004
%CD8/CD3	0.688 ± 0.003	0.694 ± 0.003	**0.714 ± 0.003**	**0.721 ± 0.003**	——	0.687 ± 0.003	0.687 ± 0.003
CD4Intensity	0.588 ± 0.004	0.589 ± 0.004	0.589 ± 0.004	0.589 ± 0.004	0.588 ± 0.004	——	0.595 ± 0.004
CD8Intensity	0.403 ± 0.005	0.403 ± 0.005	0.403 ± 0.005	0.405 ± 0.005	0.404 ± 0.005	0.414 ± 0.005	——

G-W6W means that the relationship matrix was constructed under the use of the pre-corrected residuals of the trait W6W. Bold numbers indicate predictive abilities higher than that of all previously used methods for this trait

and interactions of a marker with itself are modeled, see [10, 11] and compare to Eq. (10)). The standardization by twice the allele frequencies (and division by a certain factor representing a variance [6]) produces a GBLUP matrix which can possess entries larger than 1 and smaller than 0. In particular, if the GBLUP matrix has negative entries, squaring them changes the order of the relationship between the individuals. For instance, if A has a relation of -0.1 with individual B and -0.3 with individual C, which means that A is more closely related to B than to C, the corresponding EGBLUP matrix states that the relation between A and C is closer than that of A and B. This argumentation is equally true for the symmetric coding, but the portion of negative entries in the corresponding additive relationship matrix was close to zero for the wheat and the mouse data set when the symmetric coding was used in our examples. Overall, in spite of a certain popularity of EGBLUP in recent literature [10, 11, 17] our results suggest that the use of products of marker values as predictor variables is not the best way to incorporate interactions into the GBLUP model. Moreover, contrary to the theoretical findings on the "congruency" of EGBLUP and the Gaussian kernel in a RKHS approach [10], our results show that both methods respond in a different way to a change of marker coding: a translation of the coding has an impact on the predictive ability of EGBLUP, but not on that of the Gaussian kernel. Since the Euclidean distance between two vectors will not change under a translation of both vectors, the corresponding relationship matrix remains identical. A reconsideration

of the limit behavior of EGBLUP when the degree of interaction increases to n-factor interaction (and $n \to \infty$) may therefore be interesting from a theoretical point of view.

Categorical effect models

To develop an alternative to EGBLUP which does not possess the illustrated undesired theoretical properties, but which –unlike the RKHS approaches– allows to interpret the predicted quantities as "effects", we considered the categorical effect models (The effects of the categorical models can be explicitly calculated from phenotypes or genetic values under the use of the well-known Mixed Model formulas for effects with the respective design matrices). As a first step, we constructed the categorical marker effect model CM, which does not use the assumption of a constant allele substitution effect (Fig. 1) and thus gives the possibility to model (over)dominance by modeling an independent effect of each genotype at a locus. The fact that this property can also lead to an increase in predictive ability was illustrated by the simulated dominance scenario. An important result is that this categorical model can be rewritten as a relationship matrix model and thus provides an equivalent to the Ridge Regression/GBLUP duality, but based on a categorical effect model instead of a numerical dosage model. Whether this model increases predictive ability will always depend on the population structure and the influence of dominance effects on a particular trait. For instance, if a population originating from lines from different heterotic pools is considered, the prevalent heterosis effect might be a good reason to use CM instead of GBLUP, since heterosis creates a deviation from the linear dosage model. Moreover, the number of heterozygous and homozygous loci in the data set is important. If most loci are mainly present in only two of the three possible SNP genotypes, CM cannot outperform GBLUP substantially. Interestingly, comparing GBLUP and CM, CM was only significantly outperformed on the traits BMI and BodyLength. Thus, abandoning the assumption of a dosage effect of an allele, which is implemented by counting its occurrence and multiplying it with an additive effect, might not in general be a problem for prediction. Note also that there are other ways of defining marker based dominance matrices as for instance described by Su et al. [33]. Moreover, dominance can implicitly be modeled by an epistatic interaction term of a locus with itsself in Eq. (2) if $j = k$ (see [11]).

Analogously to the relation of GBLUP and EGBLUP, we extended the categorical marker effect model CM to the categorical epistasis model CE. The disadvantage of inflating the model with a huge number of variables is solved for genomic prediction by using an equivalent relationship-matrix-based approach. Interestingly, the analogy of the relation between GBLUP and EGBLUP also translates to the level of relationship matrices, which we illustrated by the theoretical result of Eq. (13). The relationship matrix of CE has the same connection to the relationship matrix of CM as the matrix defined by the interaction terms of EGBLUP has to the genomic relationship matrix of GBLUP. Moreover, CE eliminates undesired theoretical properties of EGBLUP: the question which allele to use as reference is not raised, its structure does not lead to a dependence of the effect models of different pairs of loci, and it does not assign the same effects to the most different allele combinations as the symmetrically coded EGBLUP model does. With the wheat data which consist of markers with only two possible values and for which GBLUP coincides with CM, CE outperformed GBLUP in all environments (Table 3). Moreover, CE slightly improved the predictive ability of CM for all considered traits of the mouse data set. Overall, the CE model is a valuable alternative for modeling epistasis since it eliminates undesired properties of EGBLUP and shows convincing results in practice. However, other more realistic parametric structures of effects in between EGBLUP and CE may be of interest for future research. Important steps into this direction have already been made with the "hybrid" coding according to He et al. [12, 13], in which the marker coding is estimated from the data under the side condition of generating a monotone effect model. Moreover, an interesting approach for future investigation may be the adaption of categorical models to other types of variables, for instance defined by haplotypes.

Incorporating prior experimental information into the coding of EGBLUP

Finally, we demonstrated that marker coding can be used to incorporate prior information. An important property of the procedure we used is that we "decoupled" the effect models for different pairs by allowing to choose the orientation of the parametric model for each pair separately (see "Methods"). In particular, this means that marker j might be coded as 0,1,2 in combination with marker k, but as $-2, -1, 0$ in combination with marker l. The criterion to decide which coding to use, was simple here by comparing the size of the absolute interaction effect of a pair when different "orientations" were used. Note here that the improvement of prediction accuracy was smaller than by means of variable selection on the wheat data set [11]. The relatively small improvement might be a result of only giving the two possibilities of both markers being in the initial coding or both markers with inverted coding, but not choosing from all possible four orientations. We used this simplified procedure, since for other combinations of one marker with original coding and the other marker with inverted coding, the assigned effect will also depend on the orientation of other pairs and thus it is difficult to determine which orientation to choose if we will additionally change the orientation of other pairs. In this regard,

the presented method can be considered as a straightforward ad hoc approach to incorporate prior knowledge into the coding, capturing some part of the covariance structure of the given data and thus improving the predicitve ability on data sets with similar covariance structure.

Conclusion

We illustrated that the EGBLUP model possesses several undesired properties caused by the interactions being modeled by products of marker values. We showed that the symmetrically coded EGBLUP tends to perform best, that the allele frequency standardized version tends to have the lowest predicitve ability and that the CE model can be an attractive alternative to EGBLUP. Prior information from other experiments can be incorporated into the marker coding of EGBLUP, which gives the potential to enhance predictive ability for correlated traits.

Endnote

[1] In literature, the expression GBLUP is used for the reformulated equivalent of Eq. (1) with genetic value $\mathbf{g} := \mathbf{M}\boldsymbol{\beta}$ and thus $\mathbf{g} \sim \mathcal{N}(0, \sigma^2_\beta \mathbf{M}\mathbf{M}')$.

Additional files

Additional file 1: Rdata-file with two lists. The list "Mouse_Data" contains a genotype matrix of 1298 individuals and 9265 markers as well as a matrix with records of 13 traits of the individuals. The list "Simulated_Data" offers the genotypes and phenotypes of the 20 simulations. Each entry of this list is a list of two elements representing genotypes and phenotypes of the respective simulation. Genotypes are given by a matrix of 1000 individuals with 9000 markers. Phenotypes are provided as a data.frame of the 1000 individuals and the 9 different phenotypes described in the Methods section. (RDATA 64512 kb)

Additional file 2: The file presents mathematical arguments for the statements on the properties of the models, which have been made in the main text. (PDF 149 kb)

Abbreviations

CM: Categorical marker effect model; CE: Categorical epistais model; DArT: Diversity Arrays Technology; EGBLUP: Extended genomic best linear unbiased prediction; GBLUP: Genomic best linear unbiased prediction; MAF: Minor allele frequency; SNP: Single nucleotide polymorphism

Acknowledgements

JWRM thanks Maria Emilia Barreyro for helpful discussions.

Funding

We acknowledge support by the Open Access Publication Funds of the Göttingen University. JWRM thanks KWS SAAT SE for financial support. NG thanks the China Scholarship Council (CSC) for financial support. RJCC was supported by grants FONCyT PICT 2013-1661, UBACyT 20020150100230B/2016 and PIP CONICET 833/2013, from Argentina.

Availability of data and materials

The simulated data, the filtered and imputed genotypes of the mouse data and the corrected phenotypes can be found in Additional file 1. The raw mouse data and a detailed description of the data can be found at the corresponding UCL website (at the moment http://mtweb.cs.ucl.ac.uk/mus/www/mouse/HS/index.shtml and http://mtweb.cs.ucl.ac.uk/mus/www/GSCAN/). The wheat data is offered by the corresponding publication. See also the "Methods" section for more details.

Authors' contributions

JWRM: Wrote the manuscript, derived the theoretical proofs of the statements, proposed to consider the topic; proposed and programmed the algorithm to adapt the coding to given data; analyzed the data; NG: supported the data analysis; prepared the mouse data set; parallelized the presented algorithm to adapt the coding to given data; tested the models on different data sets and with different validation methods; DFC: supported the data analysis; reevaluated the results with different prediction pipelines; simulated the genotypes with the QMSim software. VW, ME, RJCC, HS: guided the research. All authors have read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1] Department of Animal Sciences, Georg-August University, Albrecht Thaer-Weg 3, Göttingen, Germany. [2] National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou, China. [3] Departamento de Zootecnia, São Paulo State University, São Paulo, Brazil. [4] KWS SAAT SE, Einbeck, Germany. [5] Institute for Animal Breeding, Bavarian State Research Centre for Agriculture, Grub, Germany. [6] Department of Animal Production, University of Buenos Aires, INPA-CONICET, Buenos Aires, Argentina.

Received: 11 May 2016 Accepted: 17 December 2016
Published online: 03 January 2017

References

1. Meuwissen T, Hayes B, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
2. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res. 2009;91(01):47–60.
3. Abraham G, Tye-Din JA, Bhalala OG, Kowalczyk A, Zobel J, Inouye M. Accurate and robust genomic prediction of celiac disease using statistical learning. PLoS Genet. 2014;10(2):1004137.
4. Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975;31(2):423–47.
5. Habier D, Fernando R, Dekkers J. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177(4):2389–97.
6. VanRaden P. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
7. Piepho HP. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 2009;49(4):1165–76.
8. Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzunova M, Simianer H, Schön CC. Genome-based prediction of testcross values in maize. Theor Appl Genet. 2011;123(2):339–50.
9. Strandén I, Christensen OF. Allele coding in genomic evaluation. Genet Sel Evol. 2011;43(25):1–11. http://www.gsejournal.org/content/43/1/25.
10. Jiang Y, Reif JC. Modeling epistasis in genomic selection. Genetics. 2015;201(2):759–68.
11. Martini JWR, Wimmer V, Erbe M, Simianer H. Epistasis and covariance: How gene interaction translates into genomic relationship. Theor Appl Genet. 2016;129(5):963–76.
12. He D, Wang Z, Parida L. Data-driven encoding for quantitative genetic trait prediction. BMC Bioinformatics. 2015;16(Suppl 1):10.
13. He D, Parida L. Does encoding matter? a novel view on the quantitative genetic trait prediction problem. BMC Bioinformatics. 2016;17(Suppl 9):272.
14. Falconer DS, Mackay TF, Frankham R. Introduction to quantitative genetics.
15. Zeng ZB, Wang T, Zou W. Modeling quantitative trait loci and interpretation of models. Genetics. 2005;169(3):1711–25.

16. Hallgrímsdóttir IB, Yuster DS. A complete classification of epistatic two-locus models. BMC Genet. 2008;9(1):17.
17. Hu Z, Li Y, Song X, Han Y, Cai X, Xu S, Li W. Genomic value prediction for quantitative traits under the epistatic model. BMC Genet. 2011;12(1):15.
18. Mackay TF. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet. 2014;15(1):22–33.
19. Wang D, El-Basyoni IS, Baenziger PS, Crossa J, Eskridge K, Dweikat I. Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. Heredity. 2012;109(5):313–9.
20. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25(5):680–1.
21. Crossa J, de Los Campos G, Pérez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, HJ B. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 2010;186(2):713–24.
22. Solberg LC, Valdar W, Gauguier D, Nunez G, Taylor A, Burnett S, Arboledas-Hita C, Hernandez-Pliego P, Davidson S, Burns P, et al. A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. Mamm Genome. 2006;17(2):129–46.
23. Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JNP, Mott R, Flint J. Genetic and environmental effects on complex traits in mice. Genetics. 2006;174(2):959–84.
24. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. Nat Protoc. 2009;4(8):1184–1191.
25. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–440.
26. Wimmer V, Albrecht T, Auinger HJ, Schoen CC. synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics. 2012;28(15):2086–7.
27. Akdemir D, Godfrey OU. EMMREML: Fitting Mixed Models with Known Covariance Structures. 2015. R package version 3.1. http://CRAN.R-project.org/package=EMMREML.
28. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2014. http://www.R-project.org/.
29. Ober U, Huang W, Magwire M, Schlather M, Simianer H, Mackay TF. Accounting for genetic architecture improves sequence based genomic prediction for a drosophila fitness trait. PloS ONE. 2015;10(5):1–17: e0126880. doi:10.1371/journal. pone.0126880.
30. Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, Li J, Simianer H. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PloS ONE. 2014;9(3):93017.
31. Gianola D, Morota G, Crossa J. Genome-enabled prediction of complex traits with kernel methods: What have we learned? In: Proceedings of the 10th World Congress of Genetics Applied to Livestock Production. Vancouver, BC, Canada; 2014. https://asas.confex.com/asas/WCGALP14/webprogram/Paper10331.html.
32. Long N, Gianola D, Rosa GJ, Weigel KA. Marker-assisted prediction of non-additive genetic values. Genetica. 2011;139(7):843–54.
33. Su G, Christensen OF, Ostersen T, Henryon M, Lund MS. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PloS ONE. 2012;7(9):45293.

Genomic prediction with epistasis models: On the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE)

– Appendix –

In this section, we will give short mathematical proofs for the statements made in the main text.

Proof 1 (Property 1) *The standard approach for the estimation / prediction of the parameters of the mixed model is to maximize the joint density of phenotypes* $\mathbf{y}$ *and the additive effects* $\boldsymbol{\beta}$ *(conditioned on the fixed effect* μ*; multivariate Gaussian, product of the density of* $\boldsymbol{\beta}$ *and the density of the conditional distribution of* $\mathbf{y}$ *for fixed* $\boldsymbol{\beta}$*; the variance components are usually assumed to be known) with respect to* μ *and* $\boldsymbol{\beta}$ *[4]. This approach leads to the linear system*

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{M} \\ \mathbf{M}'\mathbf{1} & \mathbf{M}'\mathbf{M} + \mathbf{I}_p \frac{\sigma_\varepsilon^2}{\sigma_\beta^2} \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{M}'\mathbf{y} \end{pmatrix}. \tag{14}$$

Here, $\mathbf{1}$ *denotes the* $n \times 1$ *vector with all entries equal to* 1*,* $\mathbf{M}$ *is the matrix of genotypes,* $\mathbf{I}_p$ *is the p-dimensional identity matrix and* σ_i^2 *is the respective variance component of the independent Gaussian random terms* $\boldsymbol{\epsilon}$ *or* $\boldsymbol{\beta}$ *(recall Eq. (1) for the model description). What we have to show to prove a) and b) is that* $\hat{\mu}, \hat{\boldsymbol{\beta}}$ *solving system (14) implies that* $\tilde{\mu} := \hat{\mu} + \mathbf{P}'\hat{\boldsymbol{\beta}}$*,* $\hat{\boldsymbol{\beta}}$ *solve the system*

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'(\mathbf{M} - \mathbf{1}\mathbf{P}') \\ (\mathbf{M} - \mathbf{1}\mathbf{P}')'\mathbf{1} & (\mathbf{M} - \mathbf{1}\mathbf{P}')'(\mathbf{M} - \mathbf{1}\mathbf{P}') + \mathbf{I}_p \frac{\sigma_\varepsilon^2}{\sigma_\beta^2} \end{pmatrix} \begin{pmatrix} \tilde{\mu} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \mathbf{1}'\mathbf{y} \\ (\mathbf{M} - \mathbf{1}\mathbf{P}')'\mathbf{y} \end{pmatrix}, \tag{15}$$

which can be verified by a calculation. Statement c) is a consequence of the predicted average phenotype being $\tilde{\mathbf{y}} = \mathbf{1}\tilde{\mu} + \tilde{\mathbf{M}}\tilde{\boldsymbol{\beta}}$*.*

Proof 2 (Property 2) *Analogously to the proof of Property 1, substitute* $\mathbf{M}$*,* σ_β^2 *and* $\hat{\boldsymbol{\beta}}$ *by* $c\mathbf{M}$*,* $c^{-2}\sigma_\beta^2$ *and* $c^{-1}\hat{\boldsymbol{\beta}}$*.*

Proof 3 (Property 3) *Analogously to the proof of Property 1, we maximize the joint density of* $\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{h}$ *(conditioned on the fixed effect* μ*) with respect to* μ*,* $\boldsymbol{\beta}$ *and* $\boldsymbol{h}$*. Thus, we have to find a local extreme of*

$$(\mathbf{y} - \mathbf{1}\mu - \mathbf{M}\boldsymbol{\beta} - \mathbf{N}\mathbf{h})' \frac{1}{\sigma_\varepsilon^2} \mathbf{I}_\mathbf{n} (\mathbf{y} - \mathbf{1}\mu - \mathbf{M}\boldsymbol{\beta} - \mathbf{N}\mathbf{h}) + \boldsymbol{\beta}' \frac{1}{\sigma_\beta^2} \mathbf{I}_\mathbf{p} \boldsymbol{\beta} + \boldsymbol{h}' \frac{1}{\sigma_h^2} \mathbf{I}_{\ell(N)} \boldsymbol{h}. \tag{16}$$

All variables are as previously defined, with $\boldsymbol{h}$ additionally denoting the vector of all interactions and $\mathbf{N}$ *denoting the $n \times \ell(N)$ matrix assigning the respective products of marker values of each of the n individuals to the respective interaction. The length $\ell(N)$ of the rows of matrix N depends on how many interactions are incorporated in the model (e.g. $\ell(N) = p^2$). The important fact is that each entry of* $\mathbf{N}$ *is a product of two marker values. This implies that if we change* $\mathbf{M}$ *to* $c\mathbf{M}$*, we change* $\mathbf{N}$ *to* $c^2\mathbf{N}$*. Calculating the partial derivatives of Eq. (16) with respect to μ, $\boldsymbol{\beta}$ and $\boldsymbol{h}$ gives the linear system*

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{M} & \mathbf{1}'\mathbf{N} \\ \mathbf{M}'\mathbf{1} & \mathbf{M}'\mathbf{M} + \mathbf{I}_{\mathbf{p}} \frac{\sigma_\varepsilon^2}{\sigma_\beta^2} & \mathbf{M}'\mathbf{N} \\ \mathbf{N}'\mathbf{1} & \mathbf{N}'\mathbf{M} & \mathbf{N}'\mathbf{N} + \mathbf{I}_{\ell(N)} \frac{\sigma_\varepsilon^2}{\sigma_h^2} \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{h}} \end{pmatrix} = \begin{pmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{M}'\mathbf{y} \\ \mathbf{N}'\mathbf{y} \end{pmatrix}. \tag{17}$$

If we substitute here $\mathbf{M}$ *by* $c\mathbf{M}$*,* $\mathbf{N}$ *by* $c^2\mathbf{N}$*, σ_β^2 by $c^{-2}\sigma_\beta^2$ and σ_h^2 by $c^{-4}\sigma_h^2$, the new system will be solved by $\tilde{\mu}$, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{h}}$ as stated.*

Proof 4 (Property 4) *Let loci k and j share the same coding. The weights assigned to certain allele combinations are then described by*

$$\begin{array}{cccc} & aa & aA & AA \\ bb & M_{aa}^2 & M_{aa}M_{aA} & M_{aa}M_{AA} \\ bB & M_{aa}M_{aA} & M_{aA}^2 & M_{aA}M_{AA} \\ BB & M_{aa}M_{AA} & M_{aA}M_{AA} & M_{AA}^2 \end{array} \tag{18}$$

If we permute the role of a and A, we mirror the matrix with respect to the middle column. This means, if the model shall be invariant, instead of $\hat{h}_{j,k}$ we should estimate another $\tilde{h}_{j,k}$ such that the former model multiplied with $\hat{h}_{j,k}$ equals the new coding multiplied with $\tilde{h}_{j,k}$. This has to be possible for any $\hat{h}_{j,k}$, in particular for $\hat{h}_{j,k} \neq 0$, and thus $\tilde{h}_{j,k} \neq 0$ (otherwise the effects of the two models cannot be equal). Consequently, a constant $c := \frac{\hat{h}_{j,k}}{\tilde{h}_{j,k}}$ such that the former matrix of weights multiplied by c equals the new weights. In particular this means that the initial weight for $(bb, aa) = M_{aa}^2$ multiplied with c equals the new weight $M_{aa}M_{AA}$ and the initial weight for $(bb, AA) = M_{aa}M_{AA}$ multiplied by c equals the weight M_{aa}^2:

$$cM_{aa}^2 = M_{aa}M_{AA} \qquad \text{and} \qquad cM_{aa}M_{AA} = M_{aa}^2. \tag{19}$$

If $M_{aa} \neq 0$, we have $c^2 = 1$ and thus $c = \pm 1$. If $c = 1$, $M_{aa} = M_{AA}$ which is not allowed, since we are only considering codings with three different values for aa, aA, AA. Then $c = -1$ implies that $M_{aA} = 0$, since $-M_{aa}M_{aA} = M_{aa}M_{aA}$, and that $-M_{aa} = M_{AA}$, since $-M_{aa}^2 = M_{aa}M_{AA}$.
If $M_{aa} = 0$, consider the second row of matrix (18). The reasoning described above gives $cM_{aa}M_{aA} = M_{aA}M_{AA}$, which would imply that $M_{aA} = 0$ or $M_{AA} = 0$, which is not possible since we want to code the three allele combinations differently. Thus, $M_{aa} = 0$ is a contradiction to the model being invariant with respect to the decision which allele to count. Analogously for markers with only two possible values.

Proof 5 (Property 5) *Let us choose three products of Eq. (7) such that variables M_{aa}, M_{aA}, M_{AA} are included as a factor of at least one product. i) If we fix the diagonal $\{a_{m,m}\}_{m=1}^3$, the marker values are given as the square roots (possibly as a complex number with imaginary part nonzero). ii) Let us choose two products on the diagonal and one other product of two different variables (one of them shall not be included in the products on the diagonal). Then the square roots of the elements on the diagonal determine two variables and the remaining variable can be calculated from the last product. iii) Let us choose one element on the diagonal and two elements off-diagonal. Then the corresponding marker value of the diagonal element is determined. One of the other products $a_{r,s}$ is the product of the same variable and another marker value, which determines the other marker value. Analogously for the last variable. iv) Let us choose the three off-diagonal elements. Then we have to solve the system*

$$a_{1,2} = M_{aa}M_{aA} \qquad a_{1,3} = M_{aa}M_{AA} \qquad a_{2,3} = M_{aA}M_{AA},$$

which has a unique solution (up to a sign).

Proof 6 (Property 6) *The (i,l)-th entry of $\mathbf{MM}'$ in the $\{-1,1\}$ coding counts the number of loci in which individual i and l have the same marker value (this is equal to $C_{i,l}$) and subtracts the number of loci with different configuration ($p - C_{i,l}$). Thus $(MM')_{i,l} = 2C_{i,l} - p$.*

Proof 7 (Property 7) *Let $\mathbf{Q}$ denote the $n \times 2p$ matrix giving the coding of the CM model for the n individuals (recall here that we are considering markers with only two variants and $\mathbf{QQ}' = \mathbf{C}$ of Property 6). We know that the marker effect model is equivalent to a model with a corresponding relationship matrix. Moreover, we know from Property 1 that the model is independent of translations of the coding, since it is an additive model. Consequently, it is enough to show that a rescaled version of the GBLUP relationship matrix $\mathbf{MM}'$ is identical to the relationship matrix defined by a translation of $\mathbf{Q}$. This means that we have to*

show that $\alpha \in \mathbb{R}$ *and a* $2p \times 1$ *vector* $\mathbf{P}$ *exist such that*

$$\mathbf{M}\mathbf{M}' = \alpha(\mathbf{Q} - \mathbf{1}\mathbf{P}')(\mathbf{Q} - \mathbf{1}\mathbf{P}')'.$$

Since the rowsum of $\mathbf{Q}$ *equals the number of markers* p *for every row and due to the statement of Propostion 6, this equation is satisfied if* $\alpha = 2$ *and the vector* $\mathbf{P}$ *has the constant entry* 0.5.

Incorporating gene annotation into genomic prediction of complex phenotypes

Genetics: Early Online, published on August 24, 2017 as 10.1534/genetics.117.300198

Incorporating gene annotation into genomic prediction of complex phenotypes

Ning Gao[*,†,‡], **Johannes W.R. Martini**[†,‡], **Zhe Zhang**[*], **Xiaolong Yuan**[*], **Hao Zhang**[*], **Henner Simianer**[†,1] **and Jiaqi Li**[*,1]
[*]National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China, [†]University of Göttingen, Animal Breeding and Genetics Group, Göttingen 37075, Germany, [‡]Contributed equally, [1]Corresponding authors

ABSTRACT Today, genomic prediction (GP) is an established technology in plant and animal breeding programs. Current standard methods are purely based on statistical considerations but do not make use of the abundant biological knowledge, which is easily available from public data bases. Major questions that have to be answered before biological prior information can be used routinely in GP approaches are which types of information can be used, and at which points they can be incorporated into prediction methods. In this study, we propose a novel strategy to incorporate gene annotation into genomic prediction of complex phenotypes by defining haploblocks according to gene positions. Haplotype effects are then modeled as categorical or as numerical allele dosage variables. The underlying concept of this approach is to build the statistical model on variables representing the biologically functional units. We evaluate the new methods with data from a heterogeneous stock mouse population, the *Drosophila Genetic Reference Panel* (*DGRP*), and a rice breeding population from the Rice Diversity Panel. Our results show that using gene annotation to define haploblocks often leads to a comparable, but for some traits to a higher predictive ability compared to SNP based models or to haplotype models that do not use gene annotation information. Modeling gene interaction effects can further improve predictive ability. We also illustrate that the additional use of markers which have not been mapped to any gene in a second separate relatedness matrix does in many cases not lead to a relevant additional increase in predictive ability when the first matrix is based on haploblocks defined with gene annotation data, suggesting that intergenic markers only provide redundant information on the considered data sets. Therefore, gene annotation information seems to be appropriate to perceive the importance of DNA segments. Finally, we discuss the effects of gene annotation quality, marker density, and linkage disequilibrium on the performance of the new methods. To our knowledge, this is the first work which incorporates epistatic interaction or gene annotation into haplotype based prediction approaches.

KEYWORDS genomic selection; genomic prediction; gene annotation; categorical model; haplotype

In the last years, the superiority of genomic prediction (GP) (Meuwissen *et al.* 2001) over pedigree based best linear unbiased prediction (Henderson 1984) and marker assisted selection has been demonstrated (Albrecht *et al.* 2011; Crossa *et al.* 2010). GP has been applied to many different organisms, including humans (de los Campos *et al.* 2013), model species such as *Drosophila melanogaster* (Ober *et al.* 2012), plants (Jannink *et al.* 2010; Hayes *et al.* 2013), domestic animals (Hayes and Goddard 2010), and aquaculture species (Sonesson and Meuwissen 2009). Accompanied by the fast development of genotyping and sequencing technologies in the last decades, a huge number of different methods for GP have been established (Gianola 2013; de Vlaming and Groenen 2015; Misztal and Legarra 2016). Among these methods, the current standard method is ridge regression best linear unbiased prediction (*rrBLUP*) which uses single nucleotide polymorphisms (SNPs) as predictor variables. It has been shown that this marker effect ridge regression model can be translated into a relationship matrix based approach

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Tuesday 22nd August, 2017
[1]Henner Simianer: hsimian@gwdg.de. Albrecht-Thaer-Weg 3, Göttingen, 37075, Germany
[1]Jiaqi Li: jqli@scau.edu.cn. Wushan road 483, College of Animal Science, South China Agricultural University, Guangzhou, 510642, P.R.China

(GBLUP) (Habier *et al.* 2007), and this correspondence between marker effect and relationship matrix models allows us to use the classical methodology which has been developed for the pedigree BLUP for GP.

Most of the established GP methods are based on purely statistical considerations and disregard existing biological knowledge. A remarkable difference exists between the – often mechanistically simplistic – structure of statistical models describing the phenotype and the complexity of the biological processes underlying the phenotypic expression. Only recently, researchers started to work on bridging the gap between mathematical models and underlying biological mechanisms. Encouragingly, several recent studies have shown that integrating biological information in proper ways improves predictive ability under certain circumstances. For instance, it has been shown that GP accuracies can be improved by incorporating results from genome wide association studies (GWAS), either from databases (Zhang *et al.* 2014) or from the data set on hand (de los Campos *et al.* 2013; Gao *et al.* 2015; Ramstein *et al.* 2016). Other types of biological information, which are easily available from public databases, include gene annotation, information on biochemical interactions, and gene expression networks. In some of the latest publications, different types of biological knowledge were incorporated by partitioning markers into classes based on their functional annotation (Morota *et al.* 2014; Do *et al.* 2015; Abdollahi-Arpanahi *et al.* 2016; MacLeod *et al.* 2016) or gene ontology categories (Edwards *et al.* 2016). After the partitioning, one approach is to assign different prior distributions to the different classes of SNPs and then to use all markers for prediction (MacLeod *et al.* 2016). Another way is performing GP with each class separately and then selecting classes that give the best predictive ability for further predictions (Morota *et al.* 2014; Do *et al.* 2015; Abdollahi-Arpanahi *et al.* 2016; Edwards *et al.* 2016). It has been demonstrated that these approaches for incorporating biological knowledge improve the predictive ability in some cases.

At the same time, it is suggested to alter the structure of the standard models using alternative predictor variables, for instance haplotypes or interactions terms (Su *et al.* 2012; Jiang and Reif 2015; Martini *et al.* 2016). Whereas standard models are based on individual SNP markers, several new approaches are built on haplotypes (Calus *et al.* 2008; Cuyabano *et al.* 2014, 2015; Meuwissen *et al.* 2014; Yang 2015), that is on tuples of SNPs. The basic underlying assumption for models based on individual markers is that, at a sufficiently high density, at least one marker is in linkage disequilibrium (LD) with each quantitative trait locus (QTL). However, if more than two alleles of a gene exist in a population, multi-allelic haplotypes are expected to capture the state of a QTL better than single markers (Calus *et al.* 2008; Meuwissen *et al.* 2014). For this reason, haplotypes instead of single markers were used as predictor variables in several recent publications (Cuyabano *et al.* 2014, 2015; Meuwissen *et al.* 2014; Yang 2015). In these studies, for each haploblock, pseudo-markers were created by counting the number of copies of the respective allele carried by a certain individual (Meuwissen *et al.* 2014). Thus, the pseudo-marker matrix had the entries {0,1,2} and the haplotype based relatedness matrix was constructed as the dot products of the rows of this pseudo-marker matrix. The relatedness matrix was further scaled by the number of haploblocks.

Here we propose several new approaches of using gene annotation to define haplotypes in both numerical dosage and categorical effect models. To bridge the gap between the mathematical models and biology, the first step is to describe the biological system using a mathematical model on its biologically functioning units. As a first attempt, we consider the protein coding genes (and thus the corresponding proteins) including their regulatory regions as biologically acting units, hoping to capture some characteristics of the biology of complex phenotypes. In addition, we extend the haplotype based categorical effect models to epistasis models and show how all these approaches can be translated into relatedness matrices. We then test the prediction performance of our approaches with several data sets with different genetic background and discuss the similarities and relatedness of the different approaches.

Material and Methods

In order to incorporate gene annotation into GP, we firstly mapped SNPs to genes according to their relative positions, and defined haploblocks using the phased SNP data (detailed description below). Gene based haplotypes were coded using both numerical and categorical approaches. Numeric coding refers to a dosage model in which the assumption of intra-locus additive allele effects is made (Calus *et al.* 2008; Cuyabano *et al.* 2014, 2015; Meuwissen *et al.* 2014; Yang 2015). With A denoting the reference allele in a diploid population, intra-locus additivity means for instance for the SNP marker based *GBLUP* that the marker state AA ($\hat{=}2$) at locus i, has twice the effect of AB ($\hat{=}1$). The categorical coding does not assume this intra-locus additivity, but models the effect of a haplotype allele being present twice, independent of the effect when being present once. For instance, the effect of configuration AA in Table 1 is assumed to be independent from AB. Thus, the categorical model can capture dominance (Martini *et al.* 2017). We then constructed relatedness matrices for both types of models. The following sections give a detailed description of these steps.

Table 1 Categorical and numerical codings of a haploblock with four alleles. A, B, C, and D are four alleles of the same haploblock.

Allele 1	Allele 2	Haplotype categories	Allele dosage			
			A	B	C	D
A	A	AA	2	0	0	0
A	B	AB	1	1	0	0
A	C	AC	1	0	1	0
A	D	AD	1	0	0	1
B	B	BB	0	2	0	0
B	C	BC	0	1	1	0
B	D	BD	0	1	0	1
C	C	CC	0	0	2	0
C	D	CD	0	0	1	1
D	D	DD	0	0	0	2

SNP mapping and gene based haploblock derivation

The latest version of the gene annotation of each considered species was downloaded from Ensemble (http://www.ensembl.org)

using the *biomaRt* package (Durinck *et al.* 2005, 2009) of the statistical platform R (R Core Team 2016) (Table 3). Only genes indicated as "protein_coding" by the "gene_biotype" attribute were considered. Gene boundaries were extended by 5kb in both upstream and downstream flanking regions to include possible regulatory elements. Then SNPs were mapped to these genic regions based on their corresponding genomic positions. After the SNP mapping step, SNP sets were formed for genes with at least one mapped marker. For genes with only one mapped SNP, the corresponding haploblock existed of only this marker. For genes with more than one mapped SNP, phased alleles of the corresponding SNPs were combined into haplotypes with the approach described by Meuwissen *et al.* (2014). Briefly, haplotypes were built via the following steps:

- Initialization: for each gene, start with the first SNP $j = 1$.
- Step 1: include SNP $j + 1$ into the haploblock.
- Step 2: determine the number of alleles of the haploblock defined by these $j + 1$ markers across the whole population.
- Step 3: repeat Step 1 and Step 2 if the number of alleles remains below a previously chosen threshold restricting the number of alleles of a haploblock (we used 10 as proposed by Meuwissen *et al.* (2014)). Otherwise, if the number of alleles exceeds this threshold, the lastly added SNP is excluded from the current haploblock and is used as the starting position of the next haploblock. Return the alleles of the current haploblock and go to the initialization step with the lastly added SNP to define the next haploblock. Repeat this procedure until all SNPs of the currently considered gene are processed.

This approach produces one or more haploblocks with at least two haplotype alleles per block for each gene. The effects of haplotypes were then coded with two different ways:

1) Numerical (allele dosage) coding: For each haploblock, artificial SNPs are created for each haplotype allele, and these "SNPs" are coded as the number of copies ({0,1,2}) present in the respective individual. The sum over all alleles of a certain autosomal haploblock must be two for each individual when diploid species are considered.
2) Categorical coding: Haplotype variants are coded by the haplotype allele configurations (genotypes). Each allele combination has its own independent effect in the categorical coding strategy.

Table 1 contrasts the different codings of a haploblock with four alleles A, B, C, and D.

The genomic prediction models

We compared the predictive ability of the proposed approaches to the standard *GBLUP* (VanRaden 2008). The genomic prediction model can be expressed as

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{g} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is the vector of pre-corrected phenotypes; $\mathbf{1}_n$ is an $n \times 1$ vector with entries equal to one; μ is the overall mean; $\mathbf{g} \sim \mathcal{N}(0, \mathbf{K}\sigma_g^2)$ is a vector of genetic values and $\mathbf{K}$ is the relatedness matrix of the respective models (Table 2); σ_g^2 is the genetic variance; $\mathbf{e} \sim \mathcal{N}(0, \mathbf{I}\sigma_e^2)$ is a vector of residuals and σ_e^2 is the model residual variance.

For *GBLUP*, the relatedness matrix was calculated according to VanRaden (2008). Briefly, let p_k denote the minor allele frequency (MAF) of marker k, $\mathbf{M}$ denote the {0,1,2} coded genotypes, and $\mathbf{Z}$ denote the MAF adjusted marker matrix with entries $(0 - 2p_k)$, $(1 - 2p_k)$ and $(2 - 2p_k)$ for genotypes AA, AB, and BB, respectively. The relatedness matrix is calculated as $\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_{k=1}^{m} p_k(1-p_k)}$. The "extended *GBLUP*" (*EGBLUP*) (Su *et al.* 2012; Jiang and Reif 2015; Martini *et al.* 2016), whose epistasis relatedness matrix is $\mathbf{EG} = \mathbf{G}\#\mathbf{G}$, was also calculated for comparison. Here, # denotes the Hadamard product. In *EGBLUP*, we only modeled the interaction effect and ignored the additive SNP effects, since additive effects can be expressed as the sum of their interactions. Moreover, we saw in previous studies that the predictive ability of the model including both matrices –the additive and the pairwise interaction matrix– will usually tend to the predictive ability of the model with only the matrix with higher predictive ability. Thus, a small potential gain faces the disadvantage of potentially causing numerical problems in the estimation of the variance components, due to the very similar structure of the matrices $\mathbf{G}$ and $\mathbf{EG}$.

For the SNP-based categorical model (*CM*; Martini *et al.* (2017)), the relatedness matrix $\mathbf{S}$ has the entries $S_{ij} = \frac{\sum_{k=1}^{m} \phi_{ijk}}{m}$, where ϕ_{ijk} is scored 1 if individual i and j share the same genotype at maker k, otherwise ϕ_{ijk} is scored 0, m is the number of SNPs. For data sets of completely inbred lines without heterozygous markers, the *CM* model has been shown to be equivalent to *GBLUP* (Martini *et al.* 2017). The first order epistasis among markers can be modeled by extending *CM* to the *CE* (categorical epistasis) model, where the genotype combinations of each pair of loci are treated as categorical variables and the relatedness of two individuals is measured by counting the number of pairs of markers in the same state. The relatedness matrix of *CE* can be expressed as $\mathbf{E} = 0.5 \times m\mathbf{S}\#(m\mathbf{S} + \mathbf{1}_{n\times n})/m^2$ (Martini *et al.* 2017).

Analogously, we also used these two types of models for gene annotation based variables (see above). In the numerical allele dosage coding, pseudo-markers are created and the haplotype based, intra-locus additive genetic relatedness matrix is constructed as the dot product of the haplotype allele matrix ($\mathbf{M}_{H_{GA}}$). The intra-locus additive relatedness matrix is expressed as $\mathbf{G}_{H_{GA}} = \frac{\mathbf{M}_{H_{GA}}\mathbf{M}'_{H_{GA}}}{Q}$, where $\mathbf{M}_{H_{GA}}$ is a matrix of pseudo-markers with values 0, 1, and 2 representing the number of copies of each haplotype allele being present and where Q is the number of haploblocks. We call this model haplotype based genomic best linear unbiased prediction given gene annotation ($G_HBLUP|GA$). For comparison, the haplotype-based model without gene annotation (G_HBLUP) was also calculated. Here the haplotype based relatedness matrix is $\mathbf{G}_H = \frac{\mathbf{M}_H\mathbf{M}'_H}{Q}$ (Meuwissen *et al.* 2014). Haplotypes are built here for each chromosome separately (starting with the first marker and following their physical order).

In the categorical coding, we count the number of haploblocks which are in the same state between pairs of individuals, and the relatedness is measured as the ratio between the number of haploblocks with identical state and the total number of haploblocks. In an equation form, the relatedness matrix can be expressed as $\tilde{\mathbf{S}}$ with entries $\tilde{S}_{ij} = \frac{\sum_{q=1}^{Q} \phi_{ijq}}{Q}$ representing the relatedness between individuals i and j. Moreover, ϕ_{ijq} is scored 1 if individual i and j have the same state on haploblock q, otherwise ϕ_{ijq} is scored 0. We call this model haplotype based

Table 2 Relatedness matrices in corresponding models (see text for definition of the variables). # means "Hadamard product".

Models	Relatedness matrices (K)	Description
$GBLUP$	$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_{k=1}^{m} p_k(1-p_k)}$	Genomic best linear unbiased prediction
$EGBLUP$	$\mathbf{EG} = \mathbf{G}\#\mathbf{G}$	Extended (epistatic) $GBLUP$
G_HBLUP	$\mathbf{G}_H = \frac{\mathbf{M}_H\mathbf{M}_H'}{Q_H}$	Haplotype based $GBLUP$
$G_HBLUP\|GA$	$\mathbf{G}_{H_{GA}} = \frac{\mathbf{M}_{H_{GA}}\mathbf{M}_{H_{GA}}'}{Q}$	Haplotype based $GBLUP$ given gene annotation
CM	$\mathbf{S} = \left(\frac{\sum_{k=1}^{m}\phi_{ijk}}{m}\right)_{ij}$	Categorical marker effect model
CE	$\mathbf{E} = \frac{0.5\times m\mathbf{S}\#(m\mathbf{S}+\mathbf{1}_{n\times n})}{m^2}$	Categorical epistasis model
C_HM	$\mathbf{S}_H = \left(\frac{\sum_{q=1}^{Q_H}\phi_{ijq}}{Q_H}\right)_{ij}$	Haplotype based CM
C_HE	$\mathbf{E}_H = \frac{0.5\times Q_H\mathbf{S}_H\#(Q_H\mathbf{S}_H+\mathbf{1}_{n\times n})}{{Q_H}^2}$	Haplotype based CE
$C_HM\|GA$	$\tilde{\mathbf{S}} = \left(\frac{\sum_{q=1}^{Q}\phi_{ijq}}{Q}\right)_{ij}$	Haplotype based CM given gene annotation
$C_HE\|GA$	$\tilde{\mathbf{E}} = \frac{0.5\times Q\tilde{\mathbf{S}}\#(Q\tilde{\mathbf{S}}+\mathbf{1}_{n\times n})}{Q^2}$	Haplotype based CE given gene annotation

categorical model given gene annotation ($C_HM|GA$). Similar to the SNP version of the categorical model, we can build a relatedness matrix for modeling the first order epistasis among haploblocks in the form $\tilde{\mathbf{E}} = 0.5 \times Q\tilde{\mathbf{S}}\#(Q\tilde{\mathbf{S}} + \mathbf{1}_{n\times n})/Q^2$. We call this model the haplotype based categorical epistasis model given gene annotation ($C_HE|GA$). For comparison, a categorical haplotype model based on the haploblocks suggested by Meuwissen *et al.* (2014) (without the use of gene annotation) was constructed as well. We denote the categorical version of this haplotype model as C_HM. A corresponding epistatic version which models the first order epistasis among haploblocks was developed and denoted as C_HE.

In the $G_HBLUP|GA$, $C_HM|GA$, and $C_HE|GA$ models, only SNPs which have been mapped to genes are included. Therefore, we evaluated a broadened model

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{g} + \mathbf{g}_u + \mathbf{e} \quad (2)$$

including unmapped markers as well. The model terms here are the same as those defined in model 1, except for the additional term $\mathbf{g}_u \sim \mathcal{N}(0, \mathbf{K}_u\sigma^2_{g_u})$ which models the effects captured by unmapped SNPs. Here $\mathbf{K}_u$ and $\sigma^2_{g_u}$ denote the relatedness matrix calculated with unmapped SNPs and the corresponding variance component. We introduced the notation $G_HBLUP|GA*$, $C_HM|GA*$, and $C_HE|GA*$, for the broadened versions, respectively. In $G_HBLUP|GA*$, $\mathbf{K}_u = \frac{\mathbf{Z}_u\mathbf{Z}_u'}{2\sum_{k=1}^{m'} p_k(1-p_k)}$, where $\mathbf{Z}_u$ is the matrix containing the MAF adjusted genotypes of unmapped SNPs and where m' is the number of unmapped SNPs. In $C_HM|GA*$, $\mathbf{K}_u = \mathbf{S}_u = \left(\frac{\sum_{k=1}^{m'}\phi_{ijk}}{m'}\right)_{i,j}$. In $C_HE|GA*$, $\mathbf{K}_u = \mathbf{E}_u = 0.5 \times m'\mathbf{S}_u\#(m'\mathbf{S}_u + 1_{n\times n})/{m'}^2$.

In both models 1 and 2, variance components were estimated using average information restricted maximum likelihood (AI-REML) (Jensen *et al.* 1997) via the *regress* (Clifford and McCullagh 2014) package for the R statistical platform (R Core Team 2016). Given the dispersion matrices and the variance components, predictions of genetic values were obtained by solving the mixed model equations (Henderson 1984).

Data

For all data sets used for model evaluation, SNPs with a call rate of less than 95% or minor allele frequency (MAF) smaller than 0.01 and individuals with a call rate of less than 95% were excluded. Missing genotypes were imputed and phased simultaneously using *Beagle* (version 4.1) (Browning and Browning 2008) - which was embedded in the *synbreed* R package (version 0.11; Wimmer *et al.* (2012)) - using the default parameter settings. Important characteristics of the data sets after quality control are described in Table 3.

Mouse data The heterogeneous stock (HS) mice data was generated by the Wellcome Trust Centre for Human Genetics (WTCHG) (Valdar *et al.* 2006a). Genotypes and phenotype records were available at http://mtweb.cs.ucl.ac.uk/mus/www/mouse/HS/index.shtml. In total, 9,266 SNPs and 1,940 individuals remained after quality control steps. For computational simplicity, we used the pre-corrected phenotypes provided by Valdar *et al.* (2006b). Physical positions of single nucleotide polymorphisms (SNPs) were mapped to the latest version of the mouse genome (*Mus musculus, assembly GRCm38.p4*) with the *biomaRt* (Durinck *et al.* 2005, 2009) R package. Only SNPs mapped to the *GRCm38.p4* were used for further analysis. Gene boundaries were downloaded from Ensemble with the *biomaRt* (Durinck *et al.* 2005, 2009) R package. Sixteen phenotypic traits related to growth, obesity, and immunology were used in this study to compare the performance of our models.

Drosophila melanogaster data The *Drosophila Genetic Reference Panel* (*DGRP*) is a population consisting of 205 inbred lines derived from the Raleigh, USA population (Mackay *et al.* 2012). Genetic variants called from whole genome sequencing data were downloaded from the *DGRP2* website (http://dgrp2.gnets.ncsu.edu/). In total, 2,863,909 SNPs remained after quality control steps. The gene annotation information of the latest version of the *D. melanogaster* genome (*Drosophila melanogaster, assembly Release 6*) was downloaded from *Ensemble* via the *biomaRt* (Durinck *et al.* 2005, 2009) R package (Table 3). We used two adaptive

Table 3 Data sets description. # means "the number".

Data sets	# of individuals	# of markers	Reference genome	# of mapped SNPs	# of represented genes	# of haploblocks
Mice	1,940	9,266	*Mus musculus (GRCm38.p4)*	5,036	4,100	4,119
DGRP	205	2,863,909	*Drosophila melanogaster (assembly Release 6)*	2,467,249	12,586	725,520
Rice	315	58,227	*Oryza sativa Japonica Group (Build 4.0)*	44,831	22,509	25,453

traits (Mackay *et al.* 2012), one food intake trait (Garlapow *et al.* 2015), two alcohol sensitivity traits (Morozova *et al.* 2015), and twelve olfactory behavior traits (Arya *et al.* 2015) to evaluate the models. The line means (males and females independently) of all traits were adjusted for the effects of a *Wolbachia* infection and five major inversions (*In(2L)t*, *In(2R)NS*, *In(3R)K*, *In(3R)P*, and *In(3R)Mo*) using a mixed model $\bar{\mathbf{Y}} = \mathbf{Xb} + \mathbf{u} + \mathbf{e}$. $\bar{\mathbf{Y}}$ is a vector of line means; $\mathbf{X}$ is a design matrix assigning the fixed effects $\mathbf{b}$ to the lines. The random line effects were modeled $\mathbf{u} \sim \mathcal{N}(0, \mathbf{G}\sigma_u^2)$, where $\mathbf{G}$ is the marker derived genomic relationship matrix according to VanRaden (2008); $\mathbf{e} \sim \mathcal{N}(0, \mathbf{I}\sigma_e^2)$ is a vector of model residuals. Variance components were estimated using the *regress* (Clifford and McCullagh 2014) R package. The adjusted phenotypes $\bar{\mathbf{Y}} - \mathbf{X}\tilde{\mathbf{b}}$ - without any weight - were used for model evaluation.

Rice data The genotypes and phenotypes of the rice breeding population were downloaded from the rice diversity panel (https://ricediversity.org; Begum *et al.* (2015); Spindel *et al.* (2015)). In total, 315 elite rice breeding lines from the International Rice Research Institute (IRRI) irrigated rice breeding program were included in this data set. Several traits such as plant height (PH), flowering time (FLW), and grain yield (YLD) were recorded in both, the dry (DS) and the wet season (WS) for the years 2009–2012. The means of the phenotypes across years for DS or WS for each line were used as response variable (provided by Spindel *et al.* (2015)). In total, 58,227 SNPs passed the quality control steps and remained for further analysis. The gene annotation information of the latest version of the rice genome (*Oryza sativa Japonica Group, Build 4.0*) was downloaded from Ensemble via the *biomaRt* (Durinck *et al.* 2005, 2009) R package.

Predictive ability evaluation

We used 20 replicates of a 5-fold random cross-validation to assess the predictive ability of the different approaches. The variance components were estimated within the training set. Phenotypes of the validation set were treated as unknown and genetic values were predicted based on models 1 and 2, respectively. The predictive ability was calculated as Pearson's correlation between the predicted genetic values and the (pre-corrected) phenotypes of the validation population. Predictive abilities of other models were compared to $GBLUP$ (allele dosage models) or CM (categorical models) via a two-sided t-test. Moreover, for Figure 1, the relative predictive abilities were calculated as the ratio between the mean predictive ability of the alternative models and that of $GBLUP$. The models were clustered based on these relative predictive abilities using the *pheatmap* R package, where the hierarchical clustering is performed according to the euclidean distance of the vectors of relative predictive abilities for all traits.

Data availability

The mouse data used in this study is available at http://mtweb.cs.ucl.ac.uk/mus/www/mouse/HS/index.shtml. The Drosophila melanogaster data is available at http://dgrp2.gnets.ncsu.edu/. The rice breeding population data is available at https://ricediversity.org.

Results

Predictive abilities on the considered data sets

In this work, we considered marker based and (gene annotation guided) haplotype based models. We built the models on numerical allele dosage or on categorical variables, and incorporated epistasis. In the following, we will compare the predictive ability of the different models on three data sets. The results are summarized in Figure 1 and Tables 4 and 5. Additional results for the *Drosophila* data set, which are not included in these tables, can be found in Supplementary Table 1.

Mouse data Let us consider the predictive abilities of the different models for the growth related traits body weight at six to ten weeks (W6W, W7W, W8W, W10W) and the growth slope from six to ten weeks (GSL). Here, we observe consistent patterns for certain changes of numerical dosage and categorical models (Figure 1). The step from $GBLUP$ to G_HBLUP improves predictive ability by $4-6\%$, which can similarly be observed from CM to C_HM. The improvement from marker based models to the gene annotation guided haplotype based models is less than from marker based models to the ordinary haplotype models without the use of gene annotation. Moreover, the incorporation of epistasis improves the predictive ability consistently from $GBLUP$ to $EGBLUP$, from CM to CE, from C_HM to C_HE, and from $C_HM|GA$ to $C_HE|GA$. Overall, C_HE shows the highest predictive ability for these traits, and the differences between $|GA$ models and those incorporating the unmapped markers in a second matrix ($|GA*$) are small.

For the obesity related traits, body mass index (BMI) and body length (BL), all categorical models and $EGBLUP$ are outperformed by $GBLUP$ (Tables 4 and 5). For the numerical dosage models, we see that the predictive ability of $GBLUP$ is increased by the step to G_HBLUP, which again is improved by using gene annotation in $G_HBLUP|GA$. Analogously, the predictive ability of CM is similar to that of C_HM, which is improved by incorporating gene annotation information in $C_HM|GA$. The same stepwise improvement is true for CE, C_HE, and $C_HE|GA$.

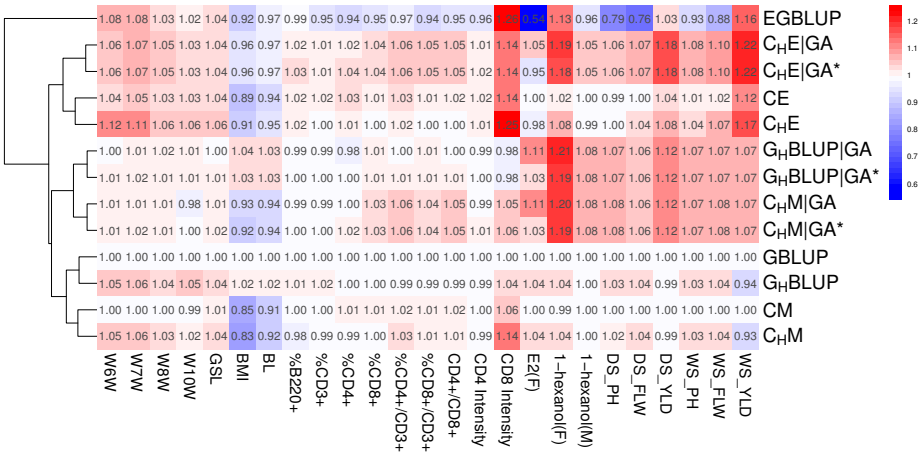

Figure 1 Comparison of the predictive ability of different models. Rows are different models and columns are traits from three data sets. For each trait, relative predictive ability is calculated by setting *GBLUP* as reference (mean accuracies divided by that of *GBLUP*). For the *DGRP*, only traits where gene annotation based models give extra predictive accuracy are presented. Trait "E2" of male lines in the *DGRP* data was also removed due to the extremely low predictive ability. W6W-W10W: body weight at 6, 7, 8, and 10 weeks; GSL: growth slope between 6 and 10 weeks of age; BMI: body mass index; BL: body length; %B220+: percentage of B220 cells; %CD3+: percentage of CD3 cells; %CD4+: percentage of CD4 cells; %CD8+: percentage of CD8 cells; %CD4+/CD3+: percentage of CD4 and CD3 cells; %CD8+/CD3+: percentage of CD8 and CD3 cells; CD4+/CD8+: ratio of CD4 to CD8 cells; CD4Intensity: CD4inCD3XGeoMean; CD8Intensity: CD8inCD3YGeoMean. F: female; M: male. DS: dry season; WS: wet season; PH: plant height; FLW: flower time; YLD: grain yield.

Comparing the epistasis models to the additive effect models, we observe an increase in predictive ability for all categorical models. The predictive ability of CE is higher than that of CM, which can analogously be observed comparing C_HM to C_HE, and $C_HM|GA$ to $C_HE|GA$. The use of a second relatedness matrix constructed with unmapped markers does not lead to a relevant increase in predictive ability (Figure 1, Tables 4 and 5). Overall, due to the relative low performance of the categorical models, $G_HBLUP|GA$, and $G_HBLUP|GA*$ perform best for BMI and BL, respectively.

For the immunology traits except CD8Intensity, we observe a relatively homogeneous predictive ability across all models (Tables 4 and 5). The performance of $EGBLUP$ is constantly low on these traits. For the traits CD8+, CD4+/CD3+, CD8+/CD3+, and CD4+/CD8+, we see that the categorical gene annotation based haplotype models $C_HM|GA$ and $C_HE|GA$ perform notably better than the other models. The epistasis variant $C_HE|GA$ improves the predictive ability slightly, compared to $C_HM|GA$.

***Drosophila* data** In the *DGRP* population, we analyzed 17 phenotypic traits (34 trait-sex combinations) related to adaptation, food intake, alcohol sensitivity, and olfactory behavior (Supplementary Table 1). Overall, gene annotation based models improve or maintain the predictive ability in 13 out of 34 scenarios compared to SNPs based models (Supplementary Table 1). $GBLUP$ performs best in 15 scenarios. Predictive ability of CM is omitted since it is similar to $GBLUP$ (identical in 21 scenarios) due to the extremely rare occurrence of heterozygotes (0.39%) in the *DGRP* population. Tables 4 and 5 show the two traits for which gene annotation based models show a considerable improvement. In one of the alcohol sensitivity traits, which was measured as alcohol knockdown time (Mean Elution Time, MET) in an "inebriometer" after a second exposure (E2) following a 2 hours recovery period (Morozova *et al.* 2015), $G_HBLUP|GA$ improves the predictive ability in females from 0.202 to 0.225 compared to $GBLUP$. However, the predictive ability for E2 in males is close to zero. In the olfactory behavior trait "1-hexanol", predictive ability is improved by $G_HBLUP|GA$ from 0.185 (0.235) in $GBLUP$ to 0.223 (0.254) for females (males). For both traits E2 and 1-hexanol, for which $G_HBLUP|GA$ and $C_HM|GA$ have the same performance, neither modeling epistasis nor including unmapped SNPs in a second relatedness matrix leads to an additional improvement.

Rice data With the rice data, we observe a systematic improvement by the use of models built on gene annotation based haplotypes. Whereas the performance of G_HBLUP is on average very similar to that of $GBLUP$ across traits, $G_HBLUP|GA$ systematically outperforms other numerical dosage models on five out of six traits (Table 4). The categorical models CM, C_HM, and $C_HM|GA$ (Table 5) perform very similar to their numerical allele dosage counterparts, which meets our expectations on the similarity of $GBLUP$ and CM on data with a low heterozygocity rate. For the categorical espistasis models, we observe a systematic improvement of predictive ability from CE to C_HE and to $C_HE|GA$. For the incorporation of epistasis, we see a consistent tendency across traits. Thus, CE tends to perform better than CM, C_HE better than C_HM, and $C_HE|GA$ better than $C_HM|GA$. However, the transition from the additive to the epistasis model does not improve predictive ability of numerical allele dosage models on the traits plant height and flowering time (from $GBLUP$ to $EGBLUP$). Overall, for plant height, flowering time, and grain yield, predictive abilities were improved by $C_HE|GA$ by 6.4% (8.1%), 6.7% (9.9%), and 17.6% (21.7%), respectively, in dry season (wet season) compared to $GBLUP$. An inclusion of unmapped SNPs in a second relatedness matrix did not improve predictive ability for any trait / model combination for the rice data.

Table 4 Predictive ability in allele dosage models (mean ± SE). * indicates models including gene based haplotypes and unmapped SNPs simultaneously. For the *DGRP* data set, two traits for which the gene annotation based models show improved predictive ability are presented. W6W-W10W: body weight at 6, 7, 8, and 10 weeks; GSL: growth slope between 6 and 10 weeks of age; BMI: body mass index; BL: body length; %B220+: percentage of B220 cells; %CD3+: percentage of CD3 cells; %CD4+: percentage of CD4 cells; %CD8+: percentage of CD8 cells; %CD4+/CD3+: percentage of CD4 and CD3 cells; %CD8+/CD3+: percentage of CD8 and CD3 cells; CD4+/CD8+: ratio of CD4 to CD8 cells; CD4Intensity: CD4inCD3XGeoMean; CD8Intensity: CD8inCD3YGeoMean. F: female; M: male. DS: dry season; WS: wet season; PH: plant height; FLW: flower time; YLD: grain yield. For each trait (row), the values in bold face indicate the best prediction among all models and values in italic are those significantly higher than $GBLUP$ ($p < 0.05$, pairwise t-test).

Data sets	Traits	$GBLUP$	$EGBLUP$	G_HBLUP	$G_HBLUP\|GA$	$G_HBLUP\|GA*$
Mouse	W6W	0.494±0.001	***0.534±0.002***	*0.521±0.001*	0.496±0.002	*0.498±0.001*
	W7W	0.495±0.002	***0.537±0.002***	*0.527±0.002*	*0.502±0.002*	*0.503±0.002*
	W8W	0.510±0.001	*0.523±0.001*	***0.531±0.001***	*0.518±0.001*	*0.517±0.001*
	W10W	0.481±0.001	*0.491±0.002*	***0.507±0.001***	*0.487±0.001*	*0.486±0.001*
	GSL	0.389±0.001	***0.405±0.002***	***0.405±0.001***	0.388±0.001	*0.392±0.001*
	BMI	0.224±0.002	0.206±0.002	*0.228±0.002*	***0.234±0.002***	*0.231±0.002*
	BL	0.264±0.002	0.255±0.002	*0.268±0.002*	*0.272±0.002*	***0.273±0.002***
	%B220+	0.546±0.002	0.541±0.001	***0.549±0.002***	0.543±0.002	*0.547±0.002*
	%CD3+	0.522±0.002	0.495±0.002	***0.531±0.002***	0.517±0.002	0.523±0.002
	%CD4+	0.481±0.002	0.454±0.001	0.481±0.001	0.473±0.002	***0.482±0.002***
	%CD8+	0.702±0.001	0.668±0.001	0.701±0.001	*0.706±0.001*	***0.707±0.001***
	%CD4+/CD3+	0.638±0.001	0.617±0.001	0.633±0.001	*0.641±0.001*	***0.642±0.001***
	%CD8+/CD3+	0.676±0.001	0.636±0.002	0.670±0.002	***0.680±0.001***	***0.680±0.001***
	CD4+/CD8+	0.671±0.001	0.636±0.001	0.665±0.001	*0.674±0.001*	***0.675±0.001***
	CD4Intensity	0.573±0.002	0.550±0.002	0.569±0.002	0.570±0.002	***0.574±0.002***
	CD8Intensity	0.388±0.002	***0.489±0.002***	*0.404±0.002*	0.379±0.002	0.382±0.002
DGRP	E2 (F)	0.202±0.010	0.110±0.012	*0.210±0.010*	***0.225±0.010***	0.208±0.010
	E2 (M)	0.026±0.010	0.038±0.008	*0.039±0.010*	***0.045±0.009***	*0.041±0.011*
	1-hexanol (F)	0.185±0.010	*0.209±0.010*	*0.193±0.009*	***0.223±0.009***	*0.220±0.010*
	1-hexanol (M)	0.235±0.009	0.225±0.009	0.236±0.009	***0.254±0.008***	***0.254±0.008***
Rice	DS_PH	0.486±0.007	0.383±0.006	*0.499±0.007*	***0.522±0.007***	***0.522±0.007***
	DS_FLW	0.534±0.005	0.405±0.006	*0.556±0.005*	***0.568±0.005***	***0.568±0.005***
	DS_YLD	0.289±0.006	0.298±0.008	0.285±0.006	***0.323±0.005***	***0.323±0.005***
	WS_PH	0.482±0.006	0.448±0.007	*0.496±0.005*	***0.516±0.005***	***0.516±0.005***
	WS_FLW	0.467±0.007	0.412±0.008	*0.487±0.006*	***0.502±0.006***	*0.501±0.006*
	WS_YLD	0.258±0.007	***0.299±0.008***	0.242±0.007	*0.276±0.008*	*0.276±0.008*
	Mean accuracy	0.431	0.418	0.440	**0.444**	**0.444**

Predictive ability vs. unexplained variance

To highlight the difference between explained variance and predictive ability, we plotted the unexplained error variance for each model and trait against the predictive ability (Figure 2). Here, we excluded the C_HE model, because its relatedness matrix has very small off-diagonal elements for the mouse data set. This leads to a situation in which the covariance matrix is more similar to the identity matrix than usual. Consequently, a certain part of the variance can be assigned to either the error or to the relatedness matrix, which causes extreme estimates for the variance components for some traits on the mouse data. Considering Figure 2, we see that there is a negative correlation between the error variance and predictive ability for most of the traits, which indicates that a model explaining the variance better also gives a higher predictive ability. However, this correlation is not -1 and has a high variation across traits. For some traits, it is even positive for the considered models. Moreover, we see also that $EGBLUP$ has the tendency to be perceived as an "outlier" in several traits, which has already been seen with the results on predictive ability alone.

Discussion

The concept of gene annotation based haplotype models

The prediction methods used in this work are all built on the classical standard assumption of the genetic values (and the error terms) being multivariate Gaussian distributed. Different concepts of defining matrices reflecting genomic relatedness were applied and the well-known mixed model equations (Henderson 1984) were used for the prediction of genetic values. Implicitly, each protocol of constructing a relatedness matrix is based on prior assumptions on how the multivariate Gaussian distributed genetic values are generated. For instance for the $GBLUP$ model these assumptions are that each marker has an intra-locus additive dosage effect, and that all these marker effects are independent realizations from the same 1-dimensional Gaussian distribution. Clearly, in a situation in which the number of markers (predictor variables) is much higher than the

Table 5 Predictive ability in categorical models (mean ± SE). * indicates models including gene based haplotypes and unmapped SNPs simultaneously. For the *DGRP* data set, two traits for which the gene annotation based models show improved predictive ability are presented. W6W-W10W: body weight at 6, 7, 8, and 10 weeks; GSL: growth slope between 6 and 10 weeks of age; BMI: body mass index; BL: body length; %B220+: percentage of B220 cells; %CD3+: percentage of CD3 cells; %CD4+: percentage of CD4 cells; %CD8+: percentage of CD8 cells; %CD4+/CD3+: percentage of CD4 and CD3 cells; %CD8+/CD3+: percentage of CD8 and CD3 cells; CD4+/CD8+: ratio of CD4 to CD8 cells; CD4Intensity: CD4inCD3XGeoMean; CD8Intensity: CD8inCD3YGeoMean. F: female; M: male. DS: dry season; WS: wet season; PH: plant height; FLW: flower time; YLD: grain yield. For each trait (row), the values in bold face indicate the best prediction among all models and values in italic are those significantly higher than CM ($p < 0.05$, pairwise t-test).

Data sets	Traits	CM	CE	C_HM	C_HE	$C_HM\|GA$	$C_HE\|GA$	$C_HM\|GA*$	$C_HE\|GA*$
Mouse	W6W	0.493±0.002	*0.516±0.002*	*0.519±0.002*	***0.551±0.002***	*0.498±0.002*	*0.524±0.002*	*0.501±0.002*	*0.525±0.002*
	W7W	0.497±0.002	*0.519±0.002*	*0.527±0.002*	***0.550±0.002***	*0.502±0.002*	*0.528±0.002*	*0.504±0.002*	*0.528±0.002*
	W8W	0.512±0.002	*0.527±0.002*	*0.525±0.001*	***0.543±0.001***	*0.515±0.002*	*0.533±0.002*	*0.517±0.002*	*0.533±0.002*
	W10W	0.477±0.002	*0.494±0.002*	*0.491±0.002*	***0.511±0.002***	0.473±0.002	*0.494±0.002*	*0.479±0.002*	*0.497±0.002*
	GSL	0.394±0.001	*0.404±0.001*	*0.403±0.001*	***0.414±0.001***	0.394±0.001	*0.404±0.001*	*0.396±0.001*	*0.406±0.001*
	BMI	0.190±0.003	*0.199±0.003*	0.186±0.003	*0.203±0.003*	*0.208±0.002*	***0.216±0.002***	*0.207±0.002*	*0.215±0.002*
	BL	0.239±0.002	*0.248±0.002*	*0.244±0.002*	*0.252±0.002*	*0.249±0.002*	***0.257±0.002***	*0.248±0.002*	*0.255±0.002*
	%B220+	0.544±0.002	*0.559±0.002*	0.534±0.002	*0.556±0.001*	0.541±0.002	*0.558±0.002*	*0.545±0.002*	***0.560±0.002***
	%CD3+	0.523±0.003	***0.530±0.003***	0.518±0.003	0.523±0.003	0.518±0.003	*0.527±0.003*	0.523±0.003	*0.529±0.003*
	%CD4+	0.486±0.001	*0.494±0.001*	0.478±0.001	0.488±0.001	0.482±0.001	*0.492±0.001*	*0.491±0.001*	***0.498±0.001***
	%CD8+	0.707±0.001	*0.712±0.001*	0.700±0.001	0.699±0.001	*0.722±0.001*	***0.728±0.001***	*0.721±0.001*	***0.728±0.001***
	%CD4+/CD3+	0.651±0.001	*0.655±0.001*	*0.654±0.001*	0.650±0.001	*0.674±0.001*	***0.678±0.001***	*0.674±0.001*	***0.678±0.001***
	%CD8+/CD3+	0.684±0.001	*0.686±0.001*	0.680±0.001	0.674±0.001	*0.706±0.001*	***0.709±0.001***	*0.706±0.001*	***0.709±0.001***
	CD4+/CD8+	0.682±0.001	*0.685±0.001*	0.678±0.001	0.674±0.001	*0.703±0.001*	***0.707±0.001***	*0.703±0.001*	***0.707±0.001***
	CD4Intensity	0.574±0.002	*0.582±0.002*	0.569±0.002	*0.580±0.002*	0.569±0.002	*0.579±0.002*	*0.578±0.002*	***0.586±0.002***
	CD8Intensity	0.413±0.002	*0.443±0.002*	*0.442±0.002*	***0.485±0.002***	0.409±0.003	*0.444±0.002*	0.412±0.002	*0.443±0.002*
DGRP	E2 (F)	0.201±0.010	0.201±0.010	*0.211±0.010*	0.198±0.011	***0.225±0.010***	*0.212±0.011*	0.208±0.010	0.192±0.011
	E2 (M)	0.026±0.010	0.028±0.010	*0.040±0.010*	*0.038±0.009*	***0.045±0.009***	*0.043±0.009*	*0.039±0.011*	*0.039±0.012*
	1-hexanol (F)	0.184±0.010	*0.188±0.010*	*0.193±0.009*	*0.199±0.009*	***0.222±0.009***	*0.221±0.009*	*0.220±0.010*	*0.219±0.010*
	1-hexanol (M)	0.234±0.009	0.235±0.009	0.235±0.009	0.233±0.009	***0.254±0.008***	*0.247±0.008*	***0.254±0.008***	*0.246±0.009*
Rice	DS_PH	0.486±0.007	0.483±0.007	*0.497±0.007*	0.484±0.006	***0.523±0.007***	*0.517±0.007*	***0.523±0.007***	*0.517±0.007*
	DS_FLW	0.534±0.005	*0.536±0.005*	*0.556±0.005*	*0.556±0.005*	*0.567±0.005*	***0.570±0.005***	*0.567±0.005*	*0.569±0.005*
	DS_YLD	0.289±0.006	*0.301±0.006*	0.285±0.006	*0.313±0.006*	*0.325±0.005*	***0.340±0.005***	*0.324±0.005*	***0.340±0.005***
	WS_PH	0.482±0.006	*0.487±0.006*	*0.496±0.005*	*0.500±0.005*	*0.518±0.005*	***0.521±0.005***	*0.518±0.005*	***0.521±0.005***
	WS_FLW	0.467±0.007	*0.476±0.007*	*0.487±0.006*	*0.500±0.006*	*0.504±0.006*	***0.513±0.006***	*0.503±0.006*	***0.513±0.006***
	WS_YLD	0.258±0.007	*0.288±0.007*	0.241±0.007	*0.302±0.007*	*0.275±0.008*	***0.314±0.008***	*0.275±0.008*	***0.314±0.008***
	Mean accuracy	0.432	0.441	0.438	0.449	0.447	**0.457**	0.448	0.456

number of individuals, and without penalization of effect sizes, any fit of the data which is generated by one of the presented models can also be obtained by an intra-locus additive marker model. However, the regularization implemented by the shrinkage of effect sizes in the ridge regression approach pushes the estimated effects towards the framework defined by the prior assumptions. Thus, prior assumptions reflecting underlying biological processes may improve the estimation of the effects of the predictor variables. In this work, these prior assumptions were set by building the model on predictor variables defined by protein coding genes. Not each marker has an effect, but rather the biological unit "gene". More specific knowledge, for instance on the biology of the respective trait, has not been used. With this conceptually simple modification, the epistatic $C_HE|GA$ model had a higher predictive ability than *GBLUP* for all traits of the rice and the mouse data, except for BMI and BL (Tables 4 and 5; Figure 1). For the *Drosophila* data set, *GBLUP* remained the best model on average (Supplementary Table 1).

Predictive abilities and model clusters

The predictive abilities of the different models are shown relative to the predictive ability of *GBLUP* in Figure 1. This relative performance gives four main clusters (based on the predictive abilities for the data presented in Figure 1; an extended pattern based on the data including all traits of the *Drosophila* data set can be found in Supplementary Figure 1).

The first cluster consists of *EGBLUP* only, whose relative predictive ability varies substantially across traits. The reason for being distinct from all other models can be seen in the centering by allele frequencies, which had been applied to the additive *GBLUP* matrix, before the Hadamard square was calculated. Since the epistatic effects are modeled as products of the centered matrix entries, this *EGBLUP* version is built on allele-frequency-dependent parametric models for the interaction effects, which means that each pair of marker has its own interaction model, which may lead to the strong variation of the performance across traits (Martini *et al.* 2017).

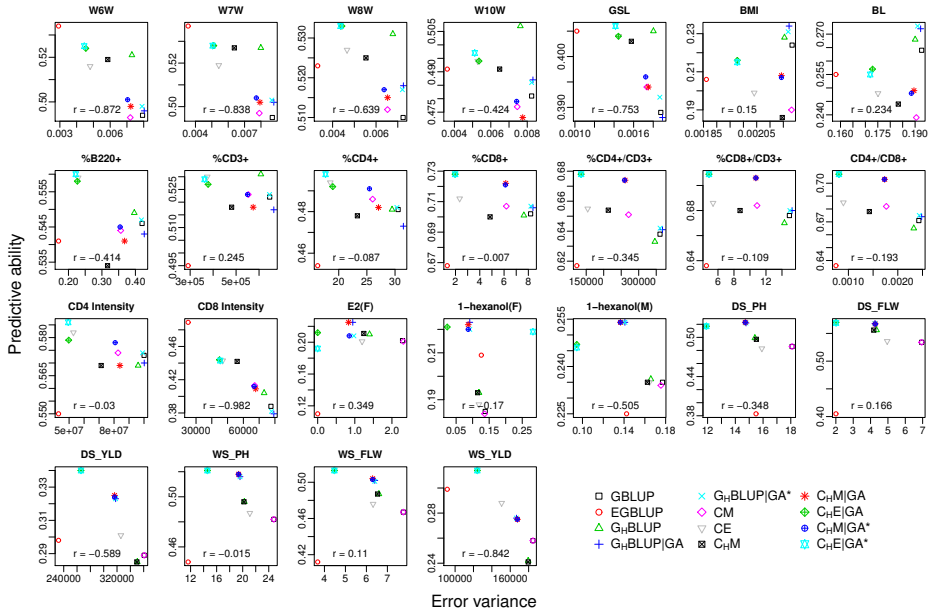

Figure 2 Error variance *VS*. predictive ability. Description of traits and models: see text and Figure 1.

The second cluster consists of the four categorical epistasis models, of which $C_HE|GA(*)$ shows the highest average predictive ability across traits. CE is more similar to C_HE than to $C_HE|GA(*)$, which is in line with the conceptual structure of the models. In C_HE, consecutive SNPs are combined into haploblocks but no external information is used to define them. $C_HE|GA(*)$ uses the gene annotation information additionally. In particular on the rice data, this conceptual construction steps also translate into predictive ability, where CE is outperformed by C_HE, whose predictive ability is further improved by $C_HE|GA(*)$ for all traits.

The third cluster contains $G_HBLUP|GA$ and $C_HM|GA$, both of which are built upon gene-annotation based haplotypes. Even though the underlying variables are more complex than single markers, their behavior relative to each other is very similar to the comparison of the marker based numerical dosage model $GBLUP$ and the categorical marker model CM (Figure 1).

The fourth cluster consists of $GBLUP$, G_HBLUP, CM, and C_HM. Except for the traits BMI, BL, and CD8Intensity, the performance of CM is very similar to that of $GBLUP$. Indeed, both methods are also theoretically identical in the case that each predictor variable has only two possible states, for instance due to complete homozygosity (Martini *et al.* 2017). However, their performances on the mouse data set illustrate that the mean predictive ability of CM and $GBLUP$ can also be very similar for data in which the two homozygous and the heterozygous states are well represented (56.06%, 34.40%, and 9.53% of 0, 1, and 2, respectively). The two models perform very similarly for the majority of the considered traits, and their difference is only visible for BMI, BL, and CD8Intensity. The fact that $GBLUP$ is more similar to its haplotype analogue G_HBLUP than to the categorical marker model CM is most probably a result of the difference in predictive ability for these traits. Indeed, if the additional traits of the *Drosophila* data set are included, $GBLUP$ and CM are closest (Supplementary Figure 1), which may be a result of the high frequency of homozygous markers in the *DGRP* data set (84.10%, 0.39%, and 15.51% of 0, 1, and 2, respectively) and of the two models consequently being almost identical for all additional traits which have not been included in Figure 1.

Overall, the clusters based on predictive abilities are in line with the conceptual construction of the models. Our results show that accounting for gene locations when defining haploblocks can improve the predictive ability, using intra-locus additive or categorical models. Across the traits of Figure 1, the categorical epistasis model $C_HE|GA$ shows the highest predictive ability on average. For the rice data, $C_HE|GA$ has the highest predictive ability for five of six traits. The trait plant height in dry season is predicted best by $C_HM|GA$. Adding a second relatedness matrix defined by SNPs that have not been mapped to genes (indicated by an *) does not systematically improve the predictive ability for most of the considered traits, indicating that unmapped SNPs do not contain sufficient additional information.

Factors affecting the performance of the gene annotation based haplotype models

As previously argued, the $|GA$ approaches are based on the concept of defining biologically functional units as predictor variables and by this constructing a statistical framework which reflects the underlying biological processes. In addition to general factors affecting the performance of GP such as the training set size, the number of markers, the genetic distance between training and test set, and the genetic architecture of the trait of interest (Shengqiang *et al.* 2009; Daetwyler *et al.* 2010), there are other important factors influencing the performance of gene annotation based prediction methods.

Evidently, a reference genome and the annotation information must be available for the target species. The quality of the annotation information will have an important impact on the number of predictor variables, on the set and the number of

markers which is mapped to genes, and on how the markers are clustered. Generally, with a decreasing number of markers, the average predictive ability will decrease (Ober *et al.* 2012). However, in our results the addition of a second relatedness matrix based on unmapped markers did overall not relevantly improve predictive ability. Thus, the marker reduction does not seem to be a critical point for the data sets used in this work.

Addressing the percentage of genes represented by haploblocks, in the mouse data set only 18.4% (4,100 out of 22,225) of all genes were represented by SNPs (Table 3). For the rice data set, for which the $|GA$ models improved the predictive ability strongly, 63.1% (22,509 out of 35,679) of the genes were modeled by at least one haploblock, whereas for the *Drospohila* data 90.4% (12,586 out of 13,918) of the genes were included in the model. Even though the latter had the highest percentage of represented genes, the use of gene annotation did not lead to a systematic improvement, but $GBLUP$ outperformed the other models for the majority of the traits (Supplementary Table 1). Besides other factors, this may in part be a result of the small population size, and of the way that the phenotypes were corrected. The correction already included the G matrix and may have slightly adapted the remaining variance to this matrix. Nevertheless, we used this approach of correction since a correction for fixed effects was necessary and this type of correction has already been used previously (Edwards *et al.* 2016).

Concerning this genotype-phenotype mapping, the results on the mouse data, where all categorical models are outperformed by $GBLUP$ for the traits BMI and BL, illustrate again that a crucial point is the trait specific architecture. The fact that the CM model, which has an advantage when dominance structures are present (Martini *et al.* 2017), is significantly outperformed by $GBLUP$ can be seen as an indicator for the absence of statistical dominance. However, the observation of a reduced predictive ability of categorical models, which incorporate dominance, should be interpreted with caution since such global quantities may not be directly linked to a biological genetic architecture of the trait (Huang and Mackay 2016).

Another important characteristic may be the average number of markers included in a haploblock which is influenced by the number of markers mapped to a gene, but also by the LD pattern of the data. It is clear that in a data set for which each haploblock consists of only one marker, a haplotype model is identical to the corresponding marker model. For the mouse data with 5,036 mapped SNPs and 4,119 haploblocks (Table 3), the majority of the haploblocks consists of not more than two markers (on average 1.22 markers per haploblock). This explains partially why the increase in predictive ability with gene annotation based haplotypes is not on the same scale as for the rice data (1.76 markers per haploblock). However, our results also show that an increasing average number of markers per haploblock does not necessarily make a model more different from a marker based model. This becomes clear by considering the fact that all haplotype models without the use of gene annotation have a higher average number of markers per haploblock than the $|GA$ models, but are still clustered closer to their respective marker model than the $|GA$ models. The average number of markers per haplotype was 7.28, 7.32, and 8.08 for the mouse, the *DGRP*, and rice data for the models without gene annotation, which was reduced to 1.22, 3.4, and 1.76 respectively for the $|GA$ models. For data sets with a rapid LD decay, adding markers to a haplotype block will rapidly increase the number of haplotype alleles. With the "maximum number of alleles" method, which we used for the construction of haplotypes, a lower LD leads to fewer markers per haploblock, which may make the haplotype based models more similar to the corresponding SNP based models. The *DGRP* population exhibits a rapid LD decay (Mackay *et al.* 2012), which is also reflected by the fact that the haploblocks in models without gene annotation on average have a comparable number of markers for the three data sets, even though the marker density of the *DGRP* data is much higher. For the *DGRP* data, the average number of markers per haploblock is the highest of the three data sets (3.4) for the $|GA$ models, which is a consequence of the high number of markers mapped to genes. This illustrates again that the interplay of multiple factors makes pure and simple statements on the causes of differences in the predictive ability difficult.

Conclusions

In this study, we proposed different ways to incorporate gene annotation information into different haplotype based genomic prediction approaches, including categorical and epistasis models. We used gene annotation information to point at the DNA segments which are more likely to play an important role in the biology of the trait and to define the model on the biologically functional unit "gene". We validated the new methods with several data sets representing different data structures (with respect to marker density, extent of LD and diversity) and a wide range of traits. Our results show that gene annotation can be beneficial in the construction of haplotype based models, if some pre-requirements, such as the availability of a reference genome and sufficiently accurate gene annotation information, are fulfilled. The suggested strategy allows us to measure the pairwise individual similarity on the gene level and provides a novel option for incorporating gene annotation into GP.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NG, JWRM, and HS conceived the study. NG and JWRM performed the analysis and wrote the manuscript. ZZ,XLY,HZ,HS, and JQL contributed to the manuscript. All authors have read and approved the final manuscript.

Acknowledgments

Ning Gao thanks China Scholarship Council (CSC) for the financial support of his study in Germany. Johannes W.R. Martini thanks KWS SAAT SE for financial support. This work is partly founded by the earmarked fund for China Agriculture Research System (CARS-35), National Natural Science Foundation of China (31772556, 31371258), Basic Work of Science and Technology Project (2014FY120800), Guangdong Sailing Program (2014YT02H042), and Guangdong Natural Science Foundation (2014A030313453). We thank the colleagues who have generated the data used in the present study for making them openly accessible.

Supplementary materials

Supplementary Figure 1. Comparison of the predictive ability of different models. In the *DGRP*, traits with extremely low predictive abilities among models are excluded from the figure.

Supplementary Table 1. Predictive ability in the *DGRP* (mean ± SE).

Literature Cited

Abdollahi-Arpanahi, R., G. Morota, B. D. Valente, A. Kranis, G. J. Rosa, *et al.*, 2016 Differential contribution of genomic regions to marked genetic variation and prediction of quantitative traits in broiler chickens. Genetics, Selection, Evolution: GSE **48**: 10.

Albrecht, T., V. Wimmer, H.-J. Auinger, M. Erbe, C. Knaak, *et al.*, 2011 Genome-based prediction of testcross values in maize. Theoretical and Applied Genetics **123**: 339–350.

Arya, G. H., M. M. Magwire, W. Huang, Y. L. Serrano-Negron, T. F. C. Mackay, *et al.*, 2015 The genetic basis for variation in olfactory behavior in Drosophila melanogaster. Chemical Senses **40**: 233–243.

Begum, H., J. E. Spindel, A. Lalusin, T. Borromeo, G. Gregorio, *et al.*, 2015 Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). PLoS ONE **10**: e0119873.

Browning, B. L. and S. R. Browning, 2008 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. American Journal of Human Genetics **84**: 210–223.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. De Roos, and R. F. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics **178**: 553–561.

Clifford, D. and P. McCullagh, 2014 The regress package R package version 1.3-14.

Crossa, J., G. d. l. Campos, P. Pérez, D. Gianola, J. Burgueño, *et al.*, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics **186**: 713–724.

Cuyabano, B. C., G. Su, and M. S. Lund, 2014 Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics **15**: 1171.

Cuyabano, B. C., G. Su, and M. S. Lund, 2015 Selection of haplotype variables from a high-density marker map for genomic prediction. Genetics, Selection, Evolution: GSE **47**: 61.

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

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen, 2013 Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genetics **9**: e1003608.

de Vlaming, R. and P. J. F. Groenen, 2015 The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. BioMed Research International **2015**: doi: 10.1155/2015/143712.

Do, D. N., L. L. G. Janss, J. Jensen, and H. N. Kadarmideen, 2015 SNP annotation-based whole genomic prediction and selection : An application to feed efficiency and its component traits in pigs. Journal of Animal Science **93**: 2056–2063.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, *et al.*, 2005 BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics **21**: 3439–3440.

Durinck, S., P. T. Spellman, E. Birney, and W. Huber, 2009 Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols **4**: 1184–91.

Edwards, S. M., I. F. Sørensen, P. Sarup, T. F. C. Mackay, and P. Sørensen, 2016 Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster. Genetics **203**: 1871–1883.

Gao, N., J. Li, J. He, G. Xiao, Y. Luo, *et al.*, 2015 Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model. BMC Genetics **16**: 120.

Garlapow, M. E., W. Huang, M. T. Yarboro, K. R. Peterson, and T. F. C. Mackay, 2015 Quantitative Genetics of Food Intake in Drosophila melanogaster. PLoS ONE **10**: e0138129.

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

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

Hayes, B. and M. Goddard, 2010 Genome-wide association and genomic selection in animal breeding. Genome **53**: 876–883.

Hayes, B. J., N. O. I. Cogan, L. W. Pembleton, M. E. Goddard, J. Wang, *et al.*, 2013 Prospects for genomic selection in forage plant species. Plant Breeding **132**: 133–143.

Henderson, C. R., 1984 *Applications of linear models in animal breeding*. University of Guelph Press, Guelph, Canada.

Huang, W. and T. F. Mackay, 2016 The genetic architecture of quantitative traits cannot be inferred from variance component analysis. PLoS genetics **12**: e1006421.

Jannink, J.-L., A. J. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics **9**: 166–77.

Jensen, J., E. A. Mantysaari, P. Madsen, and R. Thompson, 1997 Residual Maximum Likelihood Estimation of (Co) Variance Components in Multivariate Mixed Linear Models using Average Information. Journal of Indian Society of Agricultural Statistics **49**: 215–236.

Jiang, Y. and J. C. Reif, 2015 Modeling Epistasis in Genomic Selection. Genetics **201**: 759–768.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, *et al.*, 2012 The Drosophila melanogaster Genetic Reference Panel. Nature **482**: 173–178.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, *et al.*, 2016 Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics **17**: 144.

Martini, J. W. R., N. Gao, D. F. Cardoso, V. Wimmer, M. Erbe, *et al.*, 2017 Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended gblup and properties of the categorical epistasis model (ce). BMC Bioinformatics **18**: 3.

Martini, J. W. R., V. Wimmer, M. Erbe, and H. Simianer, 2016 Epistasis and covariance: how gene interaction translates into genomic relationship. Theoretical and Applied Genetics **129**: 963–976.

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

Meuwissen, T. H. E., J. Odegard, I. Andersen-Ranberg, and E. Grindflek, 2014 On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. Genetics, Selection, Evolution: GSE **46**: 49.

Misztal, I. and A. Legarra, 2016 Invited review: efficient computation strategies in genomic selection. animal **93**: doi: 10.1017/S1751731116002366.

Morota, G., R. Abdollahi-Arpanahi, A. Kranis, and D. Gianola,

2014 Genome-enabled prediction of quantitative traits in chickens using genomic annotation. BMC Genomics **15**: 109.

Morozova, T. V., W. Huang, V. A. Pray, T. Whitham, R. R. H. Anholt, *et al.*, 2015 Polymorphisms in early neurodevelopmental genes affect natural variation in alcohol sensitivity in adult drosophila. BMC Genomics **16**: 865.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, *et al.*, 2012 Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genetics **8**: e1002685.

R Core Team, 2016 R: A language and environment for statistical computing.

Ramstein, G. P., J. Evans, S. M. Kappler, R. B. Mitchell, K. P. Vogel, *et al.*, 2016 Accuracy of Genomic Prediction in Switchgrass (Panicum virgatum L.) Improved by Accounting for Linkage Disequilibrium. G3 (Bethesda, Md.) **6**: 1049–1062.

Shengqiang, Z., J. C. M. Dekkers, R. L. Fernando, and J. L. Jannink, 2009 Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics **182**: 355–364.

Sonesson, A. K. and T. H. E. Meuwissen, 2009 Testing strategies for genomic selection in aquaculture breeding programs. Genetics, Selection, Evolution: GSE **41**: 37.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, *et al.*, 2015 Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. PLoS Genetics **11**: e1004982.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund, 2012 Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PloS ONE **7**: e45293.

Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, *et al.*, 2006a Genome-wide genetic association of complex traits in heterogeneous stock mice. Nature Genetics **38**: 879–887.

Valdar, W., L. C. Solberg, D. Gauguier, W. O. Cookson, J. N. P. Rawlins, *et al.*, 2006b Genetic and environmental effects on complex traits in mice. Genetics **174**: 959–984.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414–4423.

Wimmer, V., T. Albrecht, H.-J. Auinger, and C.-C. Schön, 2012 synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics (Oxford, England) **28**: 2086–2087.

Yang, D., 2015 Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. BMC Genetics **16**: 144.

Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao, *et al.*, 2014 Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS ONE **9**: e93017.

Supplementary material

Supplementary Table 1 summarizes the predictive ability of the models across *all Drosophila* traits. Moreover, Fig. 1 illustrates the model clusters when the predictive abilities on *all* traits are used to define their similarity.

Supplementary table 1. Genomic prediction accuracy in the DGRP.

Traits	GBLUP	EGBLUP	G_HBLUP	G_HBLUP|GA	G_HBLUP|GA*	CM	CE	C_HM	C_HE	C_HM|GA	C_HE|GA	C_HM|GA*	C_HE|GA*
startle.Female	**0.364±0.009**	0.227±0.010	0.359±0.009	0.358±0.009	0.343±0.010	**0.364±0.009**	0.361±0.009	0.360±0.009	0.340±0.009	0.358±0.009	0.340±0.009	0.343±0.010	0.333±0.010
startle.Male	0.331±0.010	0.207±0.011	0.323±0.010	0.322±0.010	0.309±0.010	**0.332±0.010**	0.329±0.010	0.324±0.010	0.304±0.010	0.322±0.010	0.305±0.010	0.308±0.010	0.309±0.010
starvation.Female	**0.374±0.008**	0.171±0.008	0.359±0.008	**0.374±0.007**	0.371±0.007	**0.374±0.008**	0.369±0.008	0.360±0.008	0.328±0.008	**0.374±0.007**	0.342±0.007	0.373±0.007	0.303±0.009
starvation.Male	**0.404±0.007**	0.192±0.008	0.393±0.007	0.401±0.006	0.400±0.006	**0.404±0.007**	0.399±0.007	0.393±0.006	0.362±0.006	0.401±0.006	0.369±0.006	0.397±0.007	0.321±0.008
foodintake.Female	0.282±0.010	0.203±0.008	0.283±0.010	0.290±0.010	0.311±0.009	0.283±0.010	0.284±0.010	0.283±0.010	0.282±0.009	0.291±0.010	0.288±0.009	0.310±0.009	**0.322±0.008**
foodintake.Male	0.377±0.007	0.292±0.010	0.371±0.007	**0.380±0.007**	0.365±0.007	0.377±0.007	0.378±0.007	0.371±0.007	0.366±0.007	**0.380±0.007**	0.376±0.007	0.367±0.007	0.366±0.008
Alcohol.sensitivity.E2.Female	0.202±0.010	0.110±0.012	0.210±0.010	**0.225±0.010**	0.208±0.010	0.201±0.010	0.201±0.010	0.211±0.010	0.198±0.011	**0.225±0.010**	0.212±0.011	0.208±0.010	0.192±0.011
Alcohol.sensitivity.E2.Male	0.026±0.010	0.038±0.008	0.039±0.010	**0.045±0.009**	0.041±0.011	0.026±0.010	0.028±0.010	0.040±0.010	0.038±0.009	**0.045±0.009**	0.043±0.009	0.039±0.011	0.039±0.012
Alcohol.sensitivity.tolerance.Female	0.054±0.011	**0.101±0.008**	0.058±0.011	0.047±0.012	0.042±0.011	0.054±0.011	0.058±0.011	0.059±0.011	0.077±0.011	0.047±0.012	0.068±0.011	0.041±0.010	0.052±0.010
Alcohol.sensitivity.tolerance.Male	0.018±0.010	**0.072±0.005**	0.025±0.010	0.016±0.011	0.017±0.011	0.016±0.010	0.019±0.010	0.025±0.010	0.040±0.010	0.015±0.011	0.032±0.010	0.017±0.012	0.025±0.011
Olfactory.behavior.hexanal.Female	**0.328±0.008**	0.221±0.010	0.321±0.008	0.323±0.008	0.318±0.009	0.327±0.008	0.327±0.008	0.321±0.008	0.313±0.008	0.324±0.008	0.315±0.008	0.320±0.009	0.291±0.008
Olfactory.behavior.hexanal.Male	0.215±0.006	0.154±0.011	0.209±0.007	**0.220±0.007**	0.212±0.008	0.215±0.006	0.216±0.007	0.209±0.007	0.211±0.008	**0.220±0.007**	0.218±0.008	0.211±0.008	0.203±0.008
Olfactory.behavior.two_heptanone.Female	0.285±0.007	0.247±0.006	0.288±0.007	0.295±0.007	0.292±0.006	0.285±0.007	0.287±0.006	0.288±0.007	0.29±0.006	**0.296±0.007**	0.295±0.006	0.293±0.006	0.266±0.006
Olfactory.behavior.two_heptanone.Male	0.149±0.009	**0.159±0.011**	0.145±0.009	0.146±0.010	0.140±0.010	0.149±0.009	0.151±0.009	0.144±0.009	0.152±0.009	0.146±0.010	0.153±0.010	0.141±0.010	0.145±0.009
Olfactory.behavior.two_phenylEthylAlcohol.Female	**0.338±0.005**	0.241±0.010	0.333±0.005	0.332±0.005	0.321±0.005	0.338±0.005	0.336±0.005	0.333±0.005	0.321±0.005	0.333±0.005	0.322±0.005	0.321±0.005	0.302±0.006
Olfactory.behavior.two_phenylEthylAlcohol.Male	**0.164±0.008**	0.082±0.011	0.153±0.008	0.153±0.008	0.130±0.010	0.163±0.008	0.162±0.009	0.153±0.008	0.143±0.009	0.154±0.008	0.146±0.009	0.130±0.010	0.126±0.011
Olfactory.behavior.methylSalicylate.Female	0.352±0.010	0.266±0.010	0.344±0.009	**0.359±0.009**	0.355±0.009	0.352±0.010	0.353±0.009	0.344±0.009	0.341±0.009	**0.359±0.009**	0.351±0.009	0.353±0.010	0.324±0.009
Olfactory.behavior.methylSalicylate.Male	**0.337±0.007**	0.213±0.008	0.323±0.008	0.325±0.008	0.314±0.010	**0.337±0.007**	**0.337±0.007**	0.322±0.008	0.315±0.007	0.325±0.008	0.319±0.007	0.315±0.009	0.295±0.009
Olfactory.behavior.benzaldehyde.Female	**0.421±0.007**	0.238±0.009	0.410±0.007	0.405±0.007	0.399±0.007	**0.421±0.007**	**0.421±0.007**	0.410±0.007	0.398±0.007	0.405±0.007	0.396±0.007	0.399±0.007	0.390±0.008
Olfactory.behavior.benzaldehyde.Male	**0.324±0.008**	0.119±0.011	0.312±0.008	0.310±0.008	0.294±0.009	0.323±0.008	0.322±0.008	0.312±0.008	0.296±0.007	0.309±0.008	0.296±0.007	0.298±0.009	0.284±0.009
Olfactory.behavior.eugenol.Female	0.243±0.011	0.191±0.016	0.234±0.012	0.229±0.012	0.215±0.012	0.243±0.011	**0.245±0.012**	0.233±0.012	0.236±0.013	0.229±0.012	0.235±0.013	0.212±0.013	0.219±0.013
Olfactory.behavior.eugenol.Male	**0.194±0.009**	0.046±0.014	0.174±0.009	0.185±0.008	0.167±0.009	**0.194±0.009**	0.190±0.009	0.175±0.009	0.151±0.011	0.186±0.008	0.164±0.010	0.163±0.009	0.136±0.009
Olfactory.behavior.helional.Female	0.128±0.014	0.106±0.012	0.119±0.014	0.130±0.014	0.118±0.014	0.127±0.014	0.130±0.014	0.119±0.014	0.127±0.013	0.130±0.014	**0.137±0.013**	0.118±0.014	0.128±0.014
Olfactory.behavior.helional.Male	0.100±0.013	0.063±0.014	0.104±0.013	0.062±0.014	0.125±0.014	0.099±0.013	0.101±0.013	0.103±0.013	0.104±0.013	0.060±0.014	0.076±0.013	0.126±0.014	**0.127±0.014**
Olfactory.behavior.lCarvone.Female	0.283±0.010	0.218±0.012	0.275±0.010	0.280±0.010	0.264±0.011	0.283±0.010	**0.285±0.010**	0.274±0.010	0.278±0.010	0.280±0.010	0.281±0.010	0.266±0.010	0.261±0.011
Olfactory.behavior.lCarvone.Male	0.343±0.007	0.229±0.010	0.341±0.007	0.324±0.006	0.314±0.007	0.343±0.007	**0.345±0.007**	0.340±0.007	0.334±0.007	0.323±0.006	0.321±0.007	0.315±0.007	0.327±0.007
Olfactory.behavior.dCarvone.Female	0.220±0.011	0.210±0.012	0.205±0.012	0.213±0.011	0.206±0.011	0.220±0.011	0.223±0.011	0.205±0.012	0.219±0.012	0.212±0.011	**0.225±0.011**	0.205±0.011	0.213±0.011
Olfactory.behavior.dCarvone.Male	**0.257±0.011**	0.206±0.010	0.254±0.011	0.239±0.010	0.217±0.011	**0.257±0.011**	**0.257±0.011**	0.254±0.011	0.252±0.010	0.240±0.010	0.242±0.010	0.217±0.011	0.217±0.012
Olfactory.behavior.one_hexanol.Female	0.185±0.010	0.209±0.010	0.193±0.009	**0.223±0.009**	0.220±0.010	0.184±0.010	0.188±0.010	0.193±0.009	0.199±0.009	0.222±0.009	0.221±0.009	0.220±0.010	0.219±0.010
Olfactory.behavior.one_hexanol.Male	0.235±0.009	0.225±0.009	0.236±0.009	**0.254±0.008**	**0.254±0.008**	0.234±0.009	0.235±0.009	0.235±0.009	0.233±0.009	**0.254±0.008**	0.247±0.008	**0.254±0.008**	0.246±0.009
Olfactory.behavior.ethylAcetate.Female	**0.448±0.007**	0.321±0.011	0.437±0.007	0.436±0.006	0.423±0.007	**0.448±0.007**	0.449±0.007	0.438±0.007	0.432±0.007	0.437±0.006	0.434±0.006	0.424±0.007	0.423±0.008
Olfactory.behavior.ethylAcetate.Male	**0.338±0.015**	0.189±0.019	0.329±0.015	0.327±0.014	0.306±0.016	0.337±0.015	0.332±0.015	0.329±0.015	0.303±0.015	0.327±0.014	0.305±0.014	0.313±0.015	0.297±0.016
Olfactory.behavior.ethylButyrate.Female	**0.404±0.008**	0.258±0.010	0.391±0.008	0.393±0.008	0.381±0.007	0.403±0.007	0.403±0.007	0.391±0.008	0.384±0.007	0.393±0.008	0.388±0.007	0.384±0.007	0.387±0.007
Olfactory.behavior.ethylButyrate.Male	**0.469±0.007**	0.292±0.008	0.454±0.007	0.449±0.007	0.435±0.009	0.469±0.007	0.467±0.007	0.455±0.007	0.435±0.007	0.450±0.007	0.436±0.007	0.436±0.008	0.431±0.007

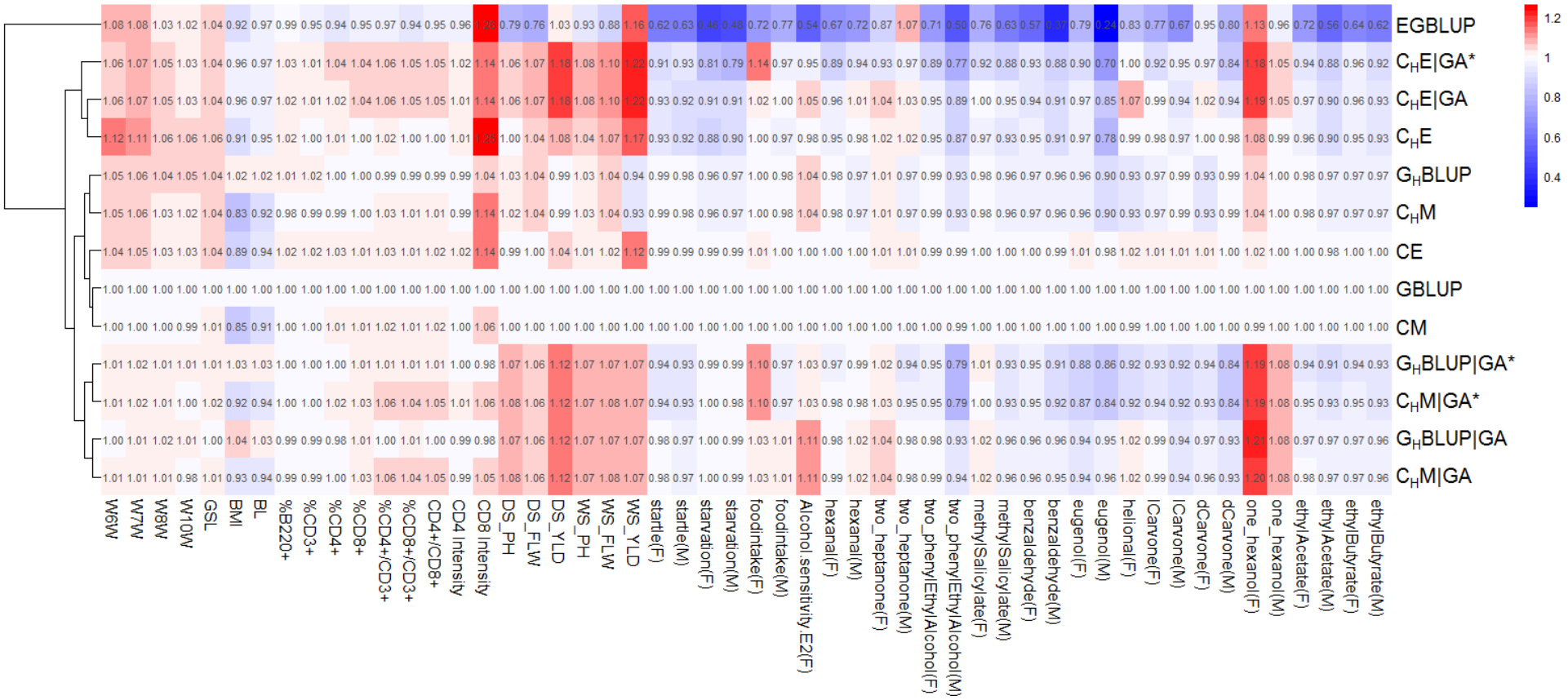

Figure 1: Cluster of the different models when all traits of the *Drosophila* data set are included. (Compare to Figure 1 of "Incorporating gene annotation into genomic prediction of complex phenotypes")

Discussion and Outlook

We will review the coding-dependence of EGBLUP, discuss the results of a simulation study on the usefulness of the total genetic value in line breeding, and give a short summarizing discussion on the importance of epistasis and an outlook. The truncated selection scheme with random mating has been simulated twice with two independently programmed scripts in `R`. The targeted mating simulation has not completely been replicated due to the computational demands. The results presented were obtained in collaboration with Torsten Pook using his breeding scheme simulation package.

Reviewing the coding dependence of EGBLUP

Let us reconsider the epistasis model modeling interactions by monomials in the marker values. For pairwise interactions, we extend the linear model of Eq. (1) to the polynomial of degree two which has been the central object in the section "Epistasis and covariance: How gene interaction translates into genomic relationship" (Eq.(2)):

$$y_i = \mu + \mathbf{M}_{i,\bullet}\boldsymbol{\beta} + \sum_{k=1,\ldots,p;l>k} M_{i,k}M_{i,l}h_{k,l} + \epsilon_i.$$

For an ordinary least squares approach (provided that a solution exists), the predictions $\hat{\mathbf{y}}$ are invariant to translations of the marker coding, but the estimates of the effects $\hat{\mu}$ and $\hat{\boldsymbol{\beta}}$ may change.

For the mixed model approach, which can be considered as a ridge regression with penalty on effect sizes, this change of the estimates $\hat{\boldsymbol{\beta}}$ of OLS induces a loss of the translation invariance of $\hat{\mathbf{y}}$. This is a result of the effect sizes being penalized in the corresponding extension of Eq. (5).

We will give an example and discuss the effect of translations of the marker coding in a more general way afterwards.

Example 1 (Translations of the marker coding)**.** *Let the marker data of five individuals with two markers be given:*

$$\mathbf{y} = (-0.72, 2.34, 0.08, -0.89, 0.86) \qquad \mathbf{M} = \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ 2 & 0 \\ 2 & 1 \\ 1 & 0 \end{pmatrix}$$

Moreover, let us use the original matrix $\mathbf{M}$ *and the by allele frequencies centered matrix*

$\tilde{\mathbf{M}} := \mathbf{M} - \mathbf{1}\underbrace{(1.6, 1)}_{=:\mathbf{P}'}$. *Then, for the corresponding OLS estimates, we receive*

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{h}_{1,2} \end{pmatrix}_{OLS} = \begin{pmatrix} 1.830 \\ -0.970 \\ 1.880 \\ -1.140 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} \tilde{\mu} \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{h}_{1,2} \end{pmatrix}_{OLS} = \begin{pmatrix} 0.334 \\ -2.110 \\ 0.056 \\ -1.140 \end{pmatrix}$$

Note here that the estimated effects $\hat{\boldsymbol{\beta}}$ *change. However, the estimated interaction* $\hat{h}_{1,2}$ *as well as* $\hat{\mathbf{y}}$ *remain unchanged.*

Contrarily, if we apply the mixed model RRBLUP of Eq. (6) with $\lambda = 1$ *as penalty factor for additive effects and the interaction, we receive*

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{h}_{1,2} \end{pmatrix}_{RR_1} = \begin{pmatrix} 1.812 \\ -0.889 \\ 0.712 \\ -0.480 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} \tilde{\mu} \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{h}_{1,2} \end{pmatrix}_{RR_1} = \begin{pmatrix} 0.334 \\ -1.151 \\ 0.090 \\ -0.575 \end{pmatrix}.$$

Both solutions produce different predictions $\hat{\mathbf{y}}$ *(each with their respective marker matrix* $\mathbf{M}$ *or* $\tilde{\mathbf{M}}$*).*

However, if we only penalize the effect size of the interaction term, both methods give different estimates for the fixed effect and the additive effects, but the same predictions $\hat{\mathbf{y}}$ *- independent of the translation. To distinguish the different approaches, we use the notation* $RR_{\lambda_h=1}$ *for latter regression, which only penalizes the interaction size.*

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{h}_{1,2} \end{pmatrix}_{RR_{\lambda_h=1}} = \begin{pmatrix} 2.685 \\ -1.540 \\ 1.025 \\ -0.570 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} \tilde{\mu} \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{h}_{1,2} \end{pmatrix}_{RR_{\lambda_h=1}} = \begin{pmatrix} 0.334 \\ -2.110 \\ 0.113 \\ -0.570 \end{pmatrix}$$

Note again that here $\hat{\mathbf{y}} = \tilde{\mathbf{y}}$.

The different cases presented in Example 1 have a certain systematic which we will

discuss in the following. In particular the circumstance that the predictions of $\mathbf{y}$ are again coinciding for $RR_{\lambda_h=1}$ –independently of the coding– is a result of following simple proposition which has several interesting implications.

Proposition 1. *Let $\mathbf{M}_{i,\bullet}$ be the p vector of the marker values of individual i and let $f(\mathbf{M}_{i,\bullet}) : \mathbb{R}^p \to \mathbb{R}$ be a polynomial in the marker data of (total) degree D. Moreover, let $\tilde{\mathbf{M}} := \mathbf{M} - \mathbf{1}\mathbf{P}'$ be a translation of the marker coding (as in Example 1) and let us define a polynomial $\tilde{f}$ in the translated variables $\tilde{\mathbf{M}}$ by $\tilde{f}(\tilde{\mathbf{M}}_{i,\bullet}) := f(\tilde{\mathbf{M}}_{i,\bullet} + \mathbf{P}') = f(\mathbf{M}_{i,\bullet})$. Then for any data $\mathbf{y}$ the goodness of fit will be identical*

$$\sum_{i=1,\ldots,n} (y_i - f(\mathbf{M}_{i,\bullet}))^2 = \sum_{i=1,\ldots,n} (y_i - \tilde{f}(\tilde{\mathbf{M}}_{i,\bullet}))^2$$

and for any monomial m of highest (total) degree D, the corresponding coefficients a_m of $f(\mathbf{M}_{i,\bullet})$ and $\tilde{a}_m$ of $\tilde{f}(\tilde{\mathbf{M}}_{i,\bullet})$ will be identical:

$$a_m = \tilde{a}_m.$$

Proof. The fact that the goodness of fit remains the same results from the definition of the polynomials. To see that the coefficients of monomials of highest (total) degree are identical, choose a monomial $m(M_{l_1}, M_{l_2}, ..., M_{l_d})$ of the loci $l_1, ..., l_d$ of (total) degree D of f. Multiplying the factors of $m(\tilde{M}_{l_1} + P_{l_1}, \tilde{M}_{l_2} + P_{l_2}, ..., \tilde{M}_{l_d} + P_{l_d})$ gives the same monomial $m(\tilde{M}_{l_1}, \tilde{M}_{l_2}, ..., \tilde{M}_{l_d})$ as a summand of highest (total) degree, plus additional monomials of lower (total) degree. Thus, the coefficients of monomials of (total) degree D remain the same. □

Proposition 1 implies that if we change the marker coding from $\mathbf{M}$ to $\tilde{\mathbf{M}}$, we can simply adapt the polynomial from f to $\tilde{f}$ to have the same goodness of fit. If f and $\tilde{f}$ are valid fits, this also means that the OLS estimates $\hat{\mathbf{y}}$ will not change when the marker coding changes. However, note here that Proposition 1 demands a certain flexibility on the model in terms of having the possibility to adapt any coefficient of monomials of lower (total) degree. We cannot adapt the regression completely if certain coefficients are forced to zero by the model structure. We will illustrate this with an example.

Example 2 (Models without certain terms of intermediate degree)**.** *Let us consider the data $\mathbf{M}$ and $\mathbf{y}$ of Example 1 but with the assumption that marker 2 does not have*

an additive effect.

Then

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{h}_{1,2} \end{pmatrix}_{OLS} = \begin{pmatrix} 3.710 \\ -2.098 \\ -0.012 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} \tilde{\mu} \\ \tilde{\beta}_1 \\ \tilde{h}_{1,2} \end{pmatrix}_{OLS} = \begin{pmatrix} 0.334 \\ -2.110 \\ -1.162 \end{pmatrix}$$

and also the estimates $\hat{\mathbf{y}}$ *and* $\tilde{\mathbf{y}}$ *are different.*

Example 2 illustrates that the model requires a certain completeness of the different variables to allow the adaption to translations of the coding. In more detail, for any monomial, the model has to include all "smaller" monomials:

Definition 1 (Completeness of a polynomial model)**.** *Let* $\mathbf{M}_{i,\bullet}$ *be the* p *vector of the marker values of individual* i *and let* $f(\mathbf{M}_{i,\bullet}) : \mathbb{R}^p \to \mathbb{R}$ *be a polynomial of total degree* D *in the marker data. The polynomial model* f *is called complete if for any monomial* $\mathbf{M}_{i,j_1}^{d_1}\mathbf{M}_{i,j_2}^{d_2}\cdots\mathbf{M}_{i,j_m}^{d_m}$ *of* f*, all monomials*

$$\mathbf{M}_{i,j_1}^{\delta_1}\mathbf{M}_{i,j_2}^{\delta_2}\cdots\mathbf{M}_{i,j_m}^{\delta_m} \; \forall\, 0 \leq \delta_1 \leq d_1, \; \forall\, 0 \leq \delta_2 \leq d_2, \; \ldots, \forall\, 0 \leq \delta_m \leq d_m$$

are included with an coefficient to be estimated.

Given that the model is "complete", Proposition 1 has various implications. In the following, we will present two corollaries which explain the results observed in our examples and theoretical properties of the considered methods.

Corollary 1. *Let an OLS estimate of a complete polynomial model* $f(\mathbf{M}_{i,\bullet})$ *exist. Then the estimates of the coefficients of highest (total) degree as well as the predictions* $\hat{\mathbf{y}}$ *are invariant with respect to translations of the marker coding.*

Corollary 1 is a result of the OLS method being defined only by the goodness of fit and explains why the OLS estimates $\hat{h}_{1,2}$ and $\tilde{h}_{1,2}$ of Example 1 are identical. Moreover, it also states that the estimates of additive effects will be unaffected by translations of the marker coding if a model without interactions in considered.

For penalized regressions, we receive the following result:

Corollary 2. *Let us consider a model given by a complete polynomial of (total) degree D and a mixed model which only penalizes the coefficients of monomials of highest (total) degree D. Then the prediction $\hat{\mathbf{y}}$ is independent of translations of the marker coding.*

Corollary 2 gives the result that RRBLUP with a constant fixed effect $\mathbf{1}\mu$ is invariant to translations of the marker coding which has for instance previously been proven using the mixed model equations (which is slightly more complicated and less general than here; Martini et al. (2017)). Moreover, the argumentation on hand also illustrates that the crucial point of the invariance of RRBLUP is the lack of a penalty factor for the intercept, that is the monomial of degree zero. Since in EGBLUP, the size of coefficients of monomials of degree one and two are both penalized, EGBLUP loses its invariance with respect to translations of the marker coding. An invariance would be given in the case that only the interactions have a penalty, but neither the additive effects, nor the intercept. This is also the reason why the predictions of $\mathbf{y}$ obtained by $RR_{\lambda_h=1}$ of Example 1 are invariant to translations.

Remark 1. *Proposition 1 stated that the coefficients of monomials of highest total degree D of f and $\tilde{f}$ will be identical. This statement can even be generalized for some situations. Consider for instance the model*

$$y_i = f(M_{i,1}, M_{i,2}, M_{i,3}) + \epsilon_i =$$

$$= \mu + \beta_1 M_{i,1} + \beta_2 M_{i,2} + \beta_3 M_{i,3} + h_{2,3} M_{i,2} M_{i,3} + \epsilon_i$$

The model is a polynomial f of total degree two. Thus, Proposition 1 states that the coefficient of monomial $M_{i,2}M_{i,3}$ will be identical for f and $\tilde{f}$. However, since $M_{i,1}$ is not included in any other monomial, its coefficient will also be identical for both polynomials. Proposition 1 was not generalized into this direction to make the manuscript not more technical than necessary. The statement made in Proposition 1 is sufficient to explain the observations related to genomic prediction models.

Total genetic values in breeding programs

How to use total genetic values including epistatic effects for line breeding has long been and is still being discussed in scientific literature. A recent publication has for instance confirmed that selecting for the phenotype instead of the additive breeding value, can produce a higher long-term gain in truncated selection programs, which the authors relate to a slower reduction of effective population size and of additive genetic variance (Esfandyari et al. 2017). In particular this means, a similar long-term response may be obtained when the selection intensity is reduced in a program selecting for the additive genetic value. Moreover, it has been pointed out that the long-term response will also depend on details on the genetic architecture of the trait under consideration, for instance on whether the sign of a marker effect can change when the genetic background changes (Paixão and Barton 2016).

We simulated a truncated selection program with traits of different genetic architecture, focusing here on the aspect of whether an allele substitution effect will change its sign when the genetic background changes (*qualitative epistasis*) or whether it will only change its magnitude (*quantitative epistasis*).

Overall, we considered i) an additive genetic architecture, ii) a quantitative and iii) a qualitative pair epistasis scenario, and iv) a qualitative epistasis scenario of three way interaction. The details of the genetic architecture will be given below.

For each of these genetic architectures, we selected for the i) regressed additive breeding value, or ii) for the epistatic genetic value regressed by the categorical epistasis model, or iii) for the phenotype in simulated "truncated selection with random mating" programs. The results show, that the long-term response can indeed be improved by using alternatives to the additive breeding value as selection criteria. However, the selection for the breeding value will only be outperformed when the genetic value has already been driven close to its maximum, that is when the genetic variance has already been reduced drastically. This circumstance causes doubts that the improvement will have a practical relevance, since in practice, genetic variance is usually introduced into

breeding programs to prevent an excessive fixation of alleles in the population.

Since the crucial point for the short and mid-term superiority of the program selecting for the additive genetic value may be the implemented random mating, we also simulated a program with targeted mating. The results show that incorporating epistasis can increase the selection gain in early generations and also maintain this advantage over time. However, this improvement could only be observed when we used a very detailed knowledge about the genetic architecture of the trait. Such details on the biology of the phenotype may in practice not be available for complex traits. Thus, on the one hand we demonstrated that the total genetic value may indeed be useful in non-random mating scenarios, but on the other hand, our results suggest that a practical relevance may not materialize in the near future. In the following, we explain the details of the simulations.

General setup of the simulations

The "genome" of our simulated diploid organism consisted of 10 chromosomes, each with 500 markers. We started from an F_2 population coming from two homozygous parents and selected with different selection criteria. The breeding objective was to combine the alleles of both parental lines optimally. The markers were coded as 0,1,2 counting the alleles of parental line 1. Each generation consisted of 1000 individuals.

Genetic values

Out of the 5000 markers, 120 were randomly (uniformly) chosen to define quantitative trait loci (QTL). For an epistatic genetic architecture, pairs (or triplets) were randomly formed from the set of QTL. The effects of (pairs or triplets of) markers were arranged according to the respective genetic architecture (see below). The QTL and their effects remained constant over generations, but were randomly drawn at the beginning of each repetition of the whole simulation. Each combination of genetic architecture and selection criterion was simulated in 100 independent repetitions.

The genetic values of the individuals were calculated as the sum of the effects of the configuration of each unit (marker, pair of markers or triplet of markers). In more

detail, in the additive model

$$\mathbf{g} := \mathbf{M}_{QTL}\boldsymbol{\beta} \tag{7}$$

with $\mathbf{M}_{QTL}$ the 1000×120 marker states $(0, 1, 2)$ of the 1000 individuals at the 120 QTL and $\boldsymbol{\beta}$ the corresponding effects.

For a trait with pairwise epistatic genetic architecture, the equation is analogous, but with 120/2 units, each unit defined by a pair of two QTL (each with 9 states). Thus, the epistatic architecture can be expressed by Eq. (7) with an $1000 \times 9 \cdot 120/2$ matrix $\mathbf{M}_{QTL}^{E_2}$ with entries 0 or 1 indicating in which state the respective marker pair is (Martini et al. 2017). For each individual i, any 9-tuple describing the state of a pair has only one entry equal to 1 and all others 0. As previously, $\boldsymbol{\beta}$ denotes the effects but is here a $9 \cdot 120/2 \times 1$ vector.

Analogously, for epistasis of degree 3, $\mathbf{M}_{QTL}^{E_3}$ is a $1000 \times 27 \cdot 120/3$ matrix and $\boldsymbol{\beta}$ a $27 \cdot 120/3 \times 1$ vector of effects.

Genetic architectures

We considered an additive genetic architecture described by Eq. (7), quantitative pairwise epistasis, qualitative pairwise epistasis and (qualitative) epistasis of degree three.

Quantitative pair epistasis Quantitative epistasis refers to a genetic architecture in which the effect signs will not change when the genetic background changes, but only their size will. Here, for a pair epistasis model, the genetic background is defined by the corresponding other marker of the interacting pair. We implemented this concept the following way: The 120 QTL were randomly partitioned into 60 pairs. For each pair j, k, three effects were drawn independently from an $\mathcal{N}(0, 1)$ distribution of which the absolute values were used and which were ordered: $0 \leq \beta_j^{(2)} \leq \beta_j^{(1)} \leq \beta_j^{(0)}$. The effect of the genetic background k was drawn analogously but with different order $0 \leq \beta_k^{(0)} \leq \beta_k^{(1)} \leq \beta_k^{(2)}$. The effect of a configuration $(M_{i,j}, M_{i,k}) = (m_1, m_2)$ was then defined by $\beta_j^{(m_1)} \cdot \beta_k^{(m_2)}$. With this implementation of quantitative epistasis, one of the corners $(0, 2)$ or $(2, 0)$ had the highest effect and the opposite corner the lowest. This guaranteed that this interaction could benefit from combining the genes of the original

parental lines and had not already been optimal in one of the parental configurations $(2, 2)$ or $(0, 0)$.

Qualitative pair epistasis To model qualitative epistasis meaning that a marker changes the sign of its effect if the genetic background changes, we drew the effects of the nine configurations randomly from a normal distribution $\mathcal{N}(0, 1)$ but adapted their signs in such a way that the corners $(0, 2)$ and $(2, 0)$ had positive effects and all others had negative effects (with absolute size as initially drawn). The result is a qualitative epistasis scenario in which whether state 0 or 2 of marker j has a positive influence on the phenotype depends on the state of marker k.

Cubic qualitative epistasis Analogously to the qualitative epistasis, we partitioned the 120 QTL into 40 sets of 3 markers which built an interaction unit. The effects of the 27 possible configurations were independently drawn from an $\mathcal{N}(0, 1)$ distribution and the signs were adapted such that the configurations $(0, 0, 2)$, $(0, 2, 0)$, $(2, 0, 0)$, $(0, 2, 2)$, $(2, 0, 2)$, and $(2, 2, 0)$ had a positive sign and all other configurations had negative effects.

Broad sense heritability, formation of the phenotype and recombination

The phenotypes of the individuals were simulated by

$$\mathbf{y} := \mathbf{g} + \boldsymbol{\epsilon} \tag{8}$$

with $\mathbf{g}$ the 1000×1 genetic values generated by the respective underlying genetic architecture and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\boldsymbol{\epsilon}} \mathbf{I}_n)$. We started with a broad sense heritability h^2

$$h^2 := \frac{\mathrm{Var}(\mathbf{g})}{\mathrm{Var}(\mathbf{g}) + \sigma^2_{\boldsymbol{\epsilon}}} \tag{9}$$

of 0.8 and kept $\sigma^2_{\boldsymbol{\epsilon}}$ constant as it has been defined in the first generation. A consequence is a reduction of broad sense heritability over generations due to the reduction of genetic variance. For each new generation, the genetic values $\mathbf{g}$ were calculated, and the realized phenotypes were simulated with a realization of $\boldsymbol{\epsilon}$ and Eq. (8).

Modeling recombination To model recombination, we drew for each chromosome a number $k \sim Pois(1)$ following a Poisson distribution with mean 1 to define the number of recombinations for this chromosome. The k recombination points were then drawn uniformly from the 500 markers.

Truncated selection with random mating

For truncated selection with random mating, the current generation was evaluated according to the respective selection criterion. The best 200 individuals were selected from which 1000 random crosses are derived (with replacement of the parents). In the following, we explain the different selection criteria.

Phenotype The 200 individuals with the highest phenotype $\mathbf{y}$ were selected.

Predicted additive breeding value The additive genetic values $\hat{\mathbf{g}}_a$ were regressed based on the model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{g}_a + \boldsymbol{\epsilon}. \tag{10}$$

Here, $\mathbf{1}$ denotes the 1000×1 vector with each entry equal to 1. Moreover, μ is a fixed effect and $\mathbf{g}_a \sim \mathcal{N}(0, \sigma_g^2\mathbf{G})$ is assumed to come from a multivariate Gaussian distribution, where $\mathbf{G}$ is the additive genomic relationship matrix ($\mathbf{G} = \mathbf{M}'\mathbf{M}$), and where again $\mathbf{M}$ is the marker matrix of all SNPs, but without the markers defining the QTL. Moreover, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2\mathbf{I})$ with $\mathbf{I}$ denoting the identity of dimension 1000. The variance components σ_g^2, σ_ϵ^2 were estimated using the `R` package `EMMREML` (Akdemir and Godfrey 2015). The 200 individuals with the highest values of $\hat{\mathbf{g}}_a$ were selected and randomly crossed.

Predicted epistatic breeding value To select for an epistatic genetic value, we used the categorical epistasis model (Martini et al. 2017). For this, we predicted $\hat{\mathbf{g}}_{CE}$ based on model (10) but with $\mathbf{g}_{CE} \sim \mathcal{N}(0, \sigma_{CE}^2\mathbf{CE})$ instead of $\mathbf{g}_a$. Here, $\mathbf{CE}$ denotes the genomic relationship matrix of a model in which each combination of the alleles of two loci has its own effect independently coming from the same Gaussian distribution (Martini et al. 2017). The markers defining the QTL were excluded. The prediction of

genetic values was carried out analogously to the additive breeding value, but with the alternative genomic relationship matrix.

Results Figs. 2 and 3 illustrate the response of the population to selection with the different selection criteria. In more detail, three characteristics are considered. The first column on the left-hand side illustrates the development of the mean total genetic value of the population over time, divided by the maximal value which is possible with the effects of the respective simulation. The development of the correlation of the selection criteria and the known total genetic value is shown in the middle column, and the variance of the genetic value is summarized in the right-hand column. Note that the results shown represent the mean of 100 independent simulations.

Comparing the additive to the quantitative pair epistasis scenario presented in Fig. 2, we make the following main observations: The short-term response is very similar between the three selection criteria for both genetic architectures. The fact that the response to phenotypic selection is not significantly lower than the response to selection for the additive breeding value is a result of the high starting heritability of 0.8. A lower heritability would make the programs more different in early generations. Moreover, we see that the selection for the phenotype produces the highest long-term response, which is followed by the selection for the epistatic genetic value. Selecting for the additive breeding value shows the lowest maximal value. This relation has a duality in the genetic variance plots indicating that additive selection reduces the variance $\mathrm{Var}(\mathbf{g})$ faster than the other selection criteria. This observation also suggests that the reason for a higher long-term response is a reduced selection pressure on single markers. Since we have on average 12 QTL on each chromosome, QTL with positive effects will also be in linkage with QTL of negative effects. A strong selection pressure on QTL with high effects will necessarily also lead to the fixation of linked QTL with small negative effects. This circumstance reduces the maximal value which is reached. In case that the selection is more fuzzy and thus fixes single markers with less pressure, desirable recombinations have more time to occur. In these aspects, the two scenarios of genetic architecture are very similar, only the maximal value which is reached is slightly reduced for the quantitative pair epistasis scenario. A difference we observe is in the ranking

of the correlations between the selection criteria and the true total genetic value. For the additive genetic architecture, the correlation between true total genetic value and estimated breeding value is higher than between true total genetic value and estimated epistatic breeding value. Approximately from generation 20 on, both quantities seem to have a very similar prediction accuracy for the total genetic value. For quantitative pair epistasis, the correlation of the regressed epistatic genetic value with the true total genetic value is for many generations higher than the correlation of the estimated additive genetic value with the true total genetic value. However, this higher accuracy of the prediction of the total genetic value does not lead to an improved response to selection. Note here that the correlation functions tend to become rougher when the genotypes tend to be fixed. This circumstance reflects the strong variation when the variance Var($\mathbf{g}$) approaches zero.

Considering Fig. 3, which illustrates the same characteristics for the qualitative pair and the qualitative cubic scenario, we see that the regression with an epistasis matrix gives a more accurate prediction of the total genetic value than the estimated additive genetic value. Moreover, the difference in accuracy of the two predictions is increased, compared to the quantitative pair epistasis scenario. However, again the improved prediction accuracy does not lead to an increased selection gain, but contrarily the selection for the additive value benefits from the more complex genetic architecture on the short-term. For the long-term response, the order of the different programs remains as for the other genetic architectures, but the maximal values which are reached are stepwise reduced when the genetic architecture becomes more complicated.

Discussion In this simulation, we observed indeed an improvement of the long-term response to selection when alternatives to the (additive) breeding value were used as selection criteria. However, the improvement has only been realized when the program selecting for the additive breeding value has already been close to its maximum. Thus, this characteristic does not seem to have a general practical relevance, since in real breeding programs additional genetic variance is usually introduced to prevent complete fixation, and mutations occur. Moreover, the improvement was also observed in the scenario of additive genetic architecture, suggesting that it may mainly be a result of

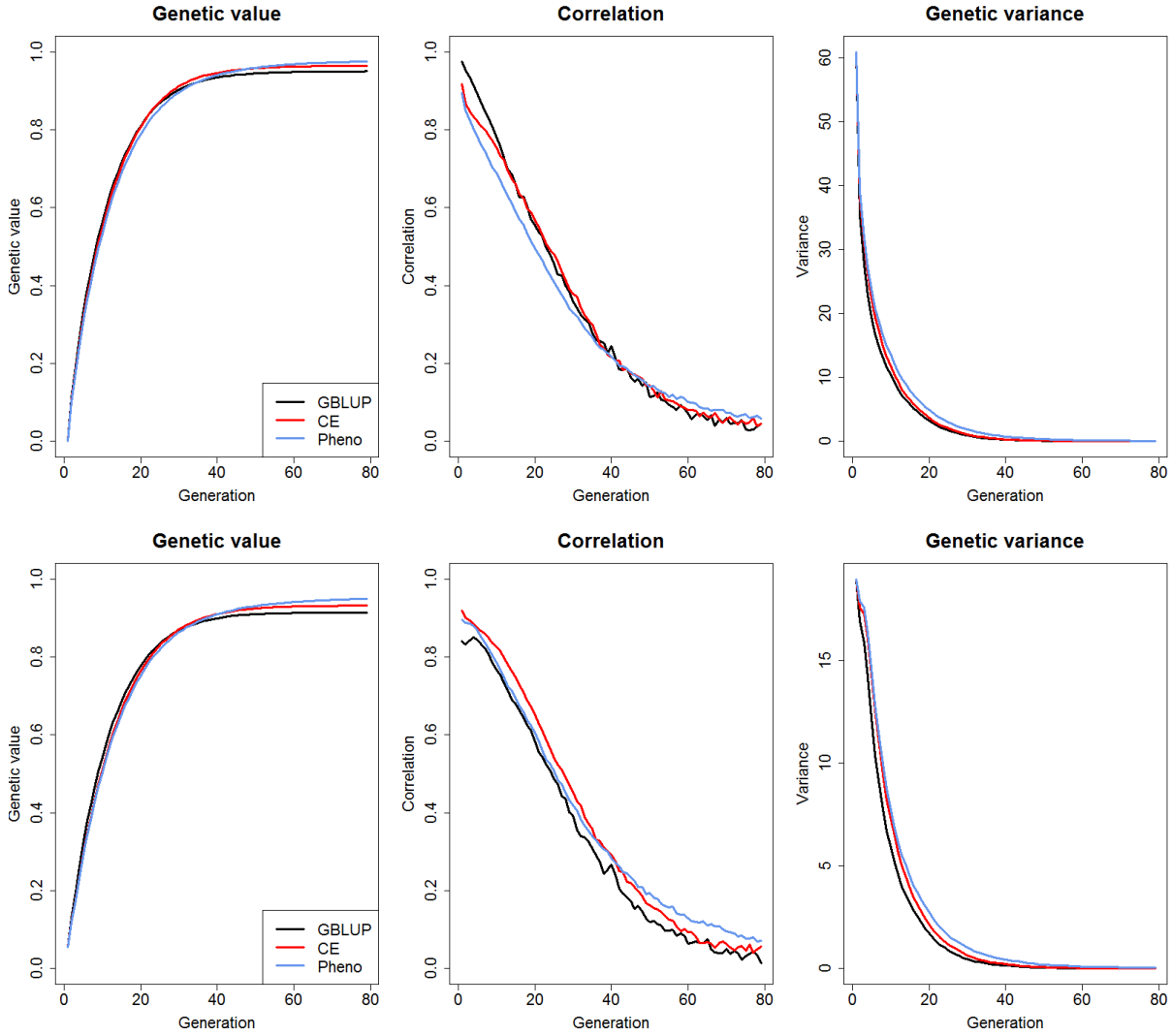

Figure 2: Truncated selection with random mating under additive genetic architecture or quantitative pair epistasis. Response to selection with different selection criteria (black: additive breeding value regressed by GBLUP; red: epistatic breeding value regressed with CE; blue: phenotype). First row: Additive genetic architecture. Second row: Quantitative pair epistasis. First column: Development of the mean phenotype of the population over generations. Here, the response is standardized by dividing by the maximal genetic value which is possible with the respective marker effects of the simulation. The graphs show the mean of 100 independent simulations with randomly drawn QTL and corresponding effects. Second column: Correlation of the real, known total genetic value and the respective selection criterion. Third column: The variance of the total genetic value Var($\mathbf{g}$) in the respective generation.

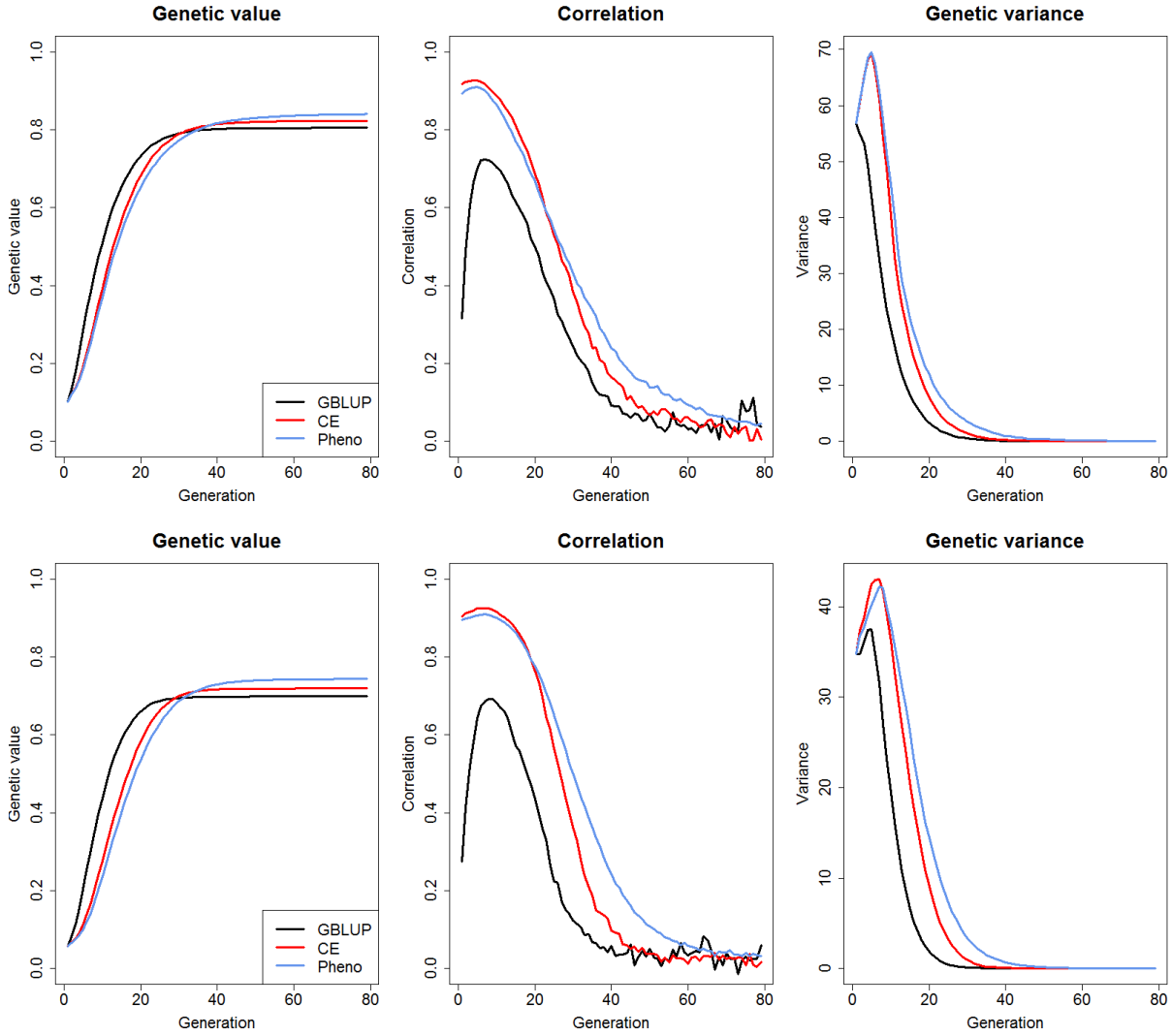

Figure 3: Truncated selection with random mating under qualitative pair epistasis and cubic qualitative epistasis. Upper row: qualitative pair epistasis; lower row: cubic qualitative epistasis; Organization of the plots as described in Fig. 2.

fixing markers with less pressure and giving the population more time to recombine beneficial alleles. Thus, our results are in accordance with other simulations addressing this topic with different models of genetic architecture and other simulation settings (Esfandyari et al. 2017; Forneris et al. 2017). In the next section, we investigate whether an improved prediction accuracy for the total genetic value may be of advantage in targeted mating programs.

Targeted mating

To implement a breeding program with targeted mating, we used the following approach: Each pair of individuals was evaluated by a prediction of the expected performance of their offspring according to the respective model (details given below). The results were summarized in a matrix $\mathbf{E}$. To generate the following generation, the 1000 pairs with the highest expected performance of their offspring were chosen, and each selected pair was mated once to generate one individual for the next generation (without the diagonal of $\mathbf{E}$, that is without selfing). The underlying genetic architectures were the quantitative or the qualitative pair epistasis scenario, respectively. We used a mixed model approach and the package `EMMREML` by Akdemir and Godfrey (2015) to estimate all required parameters. In the following, we explain the different criteria used for cross evaluation.

Mean of the parental additive breeding values The additive genetic values $\hat{\mathbf{g}}_a$ were regressed (Eq. (10) with all markers, also including the QTL) from the phenotypes of the current generation and $E_{i,j} = 0.5 \cdot (\hat{g}_{a,i} + \hat{g}_{a,j})$.

Estimated additive breeding value with information which markers are QTL The additive genetic values were estimated, but restricted only to the causative markers. The predicted cross-performance was calculated as the mean of these genetic values of the potential parental lines.

Predicted epistatic breeding value with information which markers are QTL We restricted the CE model to the causative markers, which means we used a model in which all causative markers have pairwise interactions. Since each pair of 120 QTLs can have nine configurations, we predicted $9 \cdot \frac{120 \cdot 119}{2}$ effects based on the 1000 phenotypes of the individuals of the current generation. Having these effects, we calculated for each pair of potential parents the probabilities of each combination of their QTL in the offspring. The model of recombination is important to calculate these probabilities, and we used the knowledge on the simulation described in "Modeling recombination". The pair of individuals was then evaluated by the expected genetic value of its offspring.

Predicted epistatic breeding value with information which markers are QTL and which pairs of markers interact Having the additional information on which markers interact, we restricted the model to the $9 \cdot 60$ interaction effects. Having estimated these effects, we calcualted for each pair of potential parents the probabilities of obtaining certain combinations of interacting QTL in the offspring to predict the expected total genetic value of the latter.

Results We evaluated the different selection criteria in the targeted mating approach for the quantitative and qualitative pair epistasis scenarios. In the quantitative pair epistasis scenario, the selection for the additive value performs on average identical to the epistasis model which is based on knowledge of which markers are QTL and which pairs interact (black line "GBLUP" vs. roughly dashed red line "CE|QTL + Interactions" in the first row and first column of Fig. 4). The selection based on the epistasis model which is restricted to the causative markers, but which models all pairwise interactions between them exhibits the lowest performance (CE|QTL). In the qualitative pair epistasis scenario, we see indeed that the selection based on an epistasis model can increase selection gain. However, this is only the case when we know the causative mutations *and* which pairs interact (CE|QTL + Interactions). Without the additional knowledge on which pairs interact and thus using a model incorporating all pairwise interactions of the QTL (CE|QTL), we observe an immense reduction of selection gain compared to a selection for the additive genetic value.

Discussion Our results in the targeted mating scenario with qualitative pair epistasis as genetic architecture give a proof of concept that epistasis models can theoretically be used to improve selection gain over time (qualitative pair epistasis and "CE|QTL + Interactions"). However, we also see that incomplete information can reduce the performance of the breeding program drastically (compare the dashed red lines in Fig. 4). Since in practice, the information on the biology of the trait will not be that detailed for complex traits, possible knowledge on the locations of QTL should rather be incorporated in an additive marker model. An immediate practical relevance of using the total genetic value has not been demonstrated by our simulated targeted mating program.

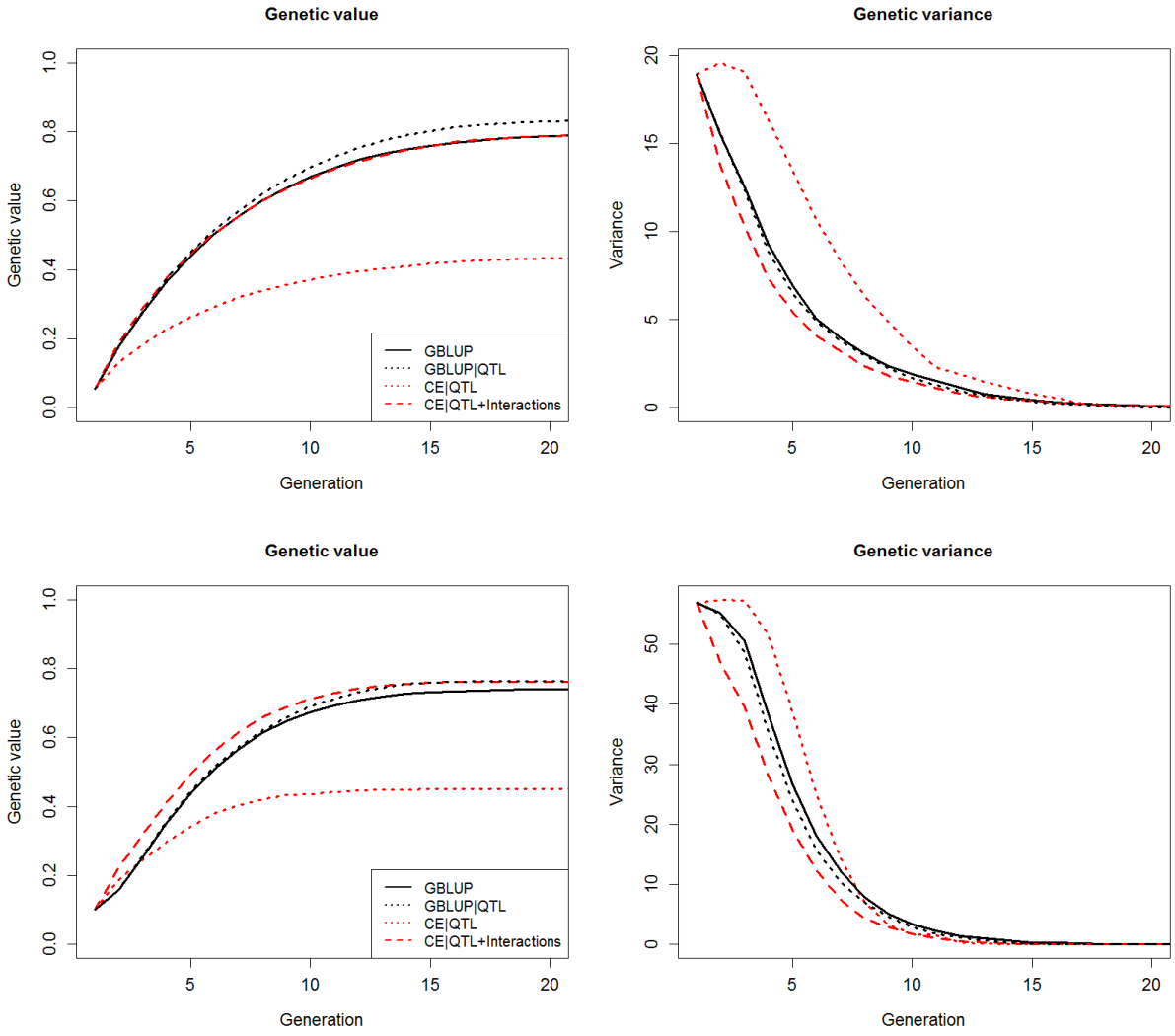

Figure 4: Targeted mating in the quantitative and qualitative pair epistasis scenarios. Upper row: Development of the mean total genetic value and the genetic variance Var(**g**) over time in the quantitative pair epistasis scenario. Lower row: Qualitative epistasis. Black: the additive selection models GBLUP or GBLUP restricted to only causative markers (GBLUP|QTL). Red: the epistasis models CE restricted to the causative markers, but with all pairwise interactions between them (CE|QTL) and a model that only incorporates the causative interactions, but estimates the effects of the nine configurations of each pair (CE|QTL + Interactions).

Reviewing the main results of this work

In this work, we investigated theoretical and practical aspects of different epistasis models. We showed that the large number of variables on which pair epistasis models are built on is not an obstacle to use them for genomic prediction, since interaction effect models can be translated into genomic relationship matrix approaches. The corresponding epistatic relationships can be easily calculated as Hadamard products of the additive relationship matrices. This is true for GBLUP and EGBLUP (Jiang and Reif 2015; Martini et al. 2016), but also for CM and CE (Martini et al. 2017).

To illustrate the potential of variable selection in epistasis models, we implemented an approach of selecting interactions with data from prior experiments. In more detail, we used the data of wheat lines grown in a certain environment to predict the interaction effects with a ridge regression, and discarded the interactions with smallest absolute effect sizes. The remaining interactions were then used to define the statistical model for prediction within other environmental conditions. Conceptually, this is a relatively simple approach and may have advantages over other methods when the selected interactions are supposed to be used for genomic prediction afterwards. Since we perform the variable selection already in the same framework in which we use them afterwards, this criterion may be more appropriate than for instance a method in which each pair is tested isolatedly without considering the structure of the remaining data. Thus, our approach is a conceptually simple option out of the many different methods proposed to identify statistically important interactions (Wang et al. 2016; Li et al. 2016; Xu et al. 2016; Frost et al. 2016; Hung et al. 2016; Sung et al. 2016). This topic may in particular be of special interest for plant breeding in which the prediction of the performance of the same lines in different environmental situations is of importance and where also a lot of data from previous experiments may be available.

We illustrated that EGBLUP has the disadvantage of the marker coding having an impact on the predictive ability, and proposed the categorical models CM and its epistatic extension CE as alternatives to GBLUP and EGBLUP. We showed that the predictive abilities of these models not only remain on a reasonable level but that they

also exceed the predictive ability of GBLUP on many traits of the considered data sets (Martini et al. 2017). In particular this illustrated again that the intra-locus additivity of GBLUP, implemented by multiplying the marker value $0, 1$, or 2 with the marker effect is not essential for genomic prediction of phenotypes. The CM model gives a comparable predictive ability, and does not assume this intra-locus additivity, but instead models classical additive and dominance effects jointly.

Moreover, we used external gene annotation data to define haploblocks and thus to build the statistical model on biologically functional units instead of on single markers (Gao et al. 2017). The topic of incorporating gene annotation data has been addressed in several publications in the last years and has often been approached on the marker level by building different annotation classes of markers and treating them differently in the prediction (Morota et al. 2014; Do et al. 2015; MacLeod et al. 2016). We fused here allele-dosage models or categorical models (Martini et al. 2017) with a haploblock approach (Meuwissen et al. 2014) and with the information on where genes are located, to create a model that defines a relationship on the level "protein coding gene". In particular on the rice data, our methods exhibited a relevant increase in predictive ability compared to marker based models or models based on haplotypes which do not use gene annotation information. We focused here on the haploblock characteristic and did not compare our models to single marker approaches using gene annotation. This should be done in future work to compare the improvement when identical gene annotation information is incorporated in different ways. Overall, our model illustrated that using external biological information can be beneficial and that it can be relatively easily incorporated in different ways into the prediction model.

Our main results are in line with the conclusions of many other publications of the last years which illustrated that epistasis models (Crossa et al. 2010; Ober et al. 2011; Zhang et al. 2015) and the incorporation of external information (Zhang et al. 2014) can improve the prediction of phenotypes. This circumstance may be interpreted as another hint for the importance of epistasis for the formation of phenotypes, also on the statistical effect level (Tyler et al. 2016; Sohail et al. 2017). Yet, since the additive effects also tend to "obscure" epistatic effects (Sackton and Hartl 2016), it is not clear

whether an improved predictive ability of an epistasis model remains when the marker density increases. It may be the case that more of the variance caused by epistatic effects is taken up by additive effects when more markers with a gradual decrease in linkage disequilibrium are added. However, this topic needs a more systematic investigation to give a clear answer.

It is also an open question how an improved predictive ability can be used beneficially in breeding programs. In our simulations of truncated selection with random mating, a higher long-term response was observed when alternatives to the breeding value were used as selection criterion. This observation is in line with literature reporting a potential increase in long-term response and a maintenance of additive variance over time resulting from an epistatic genetic architecture (Carlborg et al. 2006; Paixão and Barton 2016; Esfandyari et al. 2017; Forneris et al. 2017). However, a major part of the gain in long-term response when alternatives to the breeding value are used as selection criterion seems to be simply caused by the reduced fixation speed, thus providing more time to combine positive alleles by recombination (Esfandyari et al. 2017). This view is supported by our result of observing an improved long-term response also with an additive genetic architecture. The differences in long-term gain of the considered selection criteria were in our examples similar across the different scenarios of genetic architecture, but a slight tendency of the long-term gap becoming bigger with a more complex genetic architecture may be observed (Figs. 2 and 3). However, the additional gain has only been realized when the program selecting for the breeding value has already been close to its maximum. In real breeding programs in which additional variance may be added for instance from pre-breeding programs, and where mutations occur, this plateau is usually not reached. Thus, the practical relevance of this improved long-term gain seems low.

Since the short and mid-term superiority of a selection for the breeding value may be enhanced by the implemented random mating, we simulated a targeted mating program, for which an incorporation of epistasis may be of advantage. Here, we gave a proof of concept that the consideration of epistasis can theoretically improve the selection gain in this breeding scheme. Considering the response to selection with the "CE|QTL +

Interactions" criterion in the qualitative epistasis scenario of Fig. 4, we see a very similar maximal level compared to the additive selection with knowledge on which markers are QTL, but a relevantly faster short-term improvement. To achieve this improvement, we needed detailed information on the location of the QTL and also on which pairs of them interact. In real breeding programs, this level of detailed information will hardly be given. Considering the response to selection in the breeding program using an epistasis model with the knowledge which markers are QTL, but not which pairs interact, we see that the response is drastically reduced, which illustrates the pitfalls of applying epistasis models in this situation of targeted mating. The more complex statistical epistasis model does not show the robustness that the selection for the additive genetic value shows. Using the more detailed model but with incomplete information reduces the selection gain strongly (CE|QTL vs. CE|QTL + Interactions of Fig. 4). Since real underlying biological processes are much more complicated in the sense that they are not only based on disjunct pairwise interactions but on higher order interaction, there is the permanent threat of having insufficient information. Comparing this behavior to the responses of selection for the additive value(s), we see that the GBLUP model has a relative high performance in all genetic architectures, and that the knowledge of which markers are QTL constantly improves the mid and long-term response across different genetic architecture scenarios (Fig. 4). Thus, we gave a proof of concept for the theoretical usefulness of a more accurate prediction of total genetic values, but also illustrated that in practice the required conditions will usually not be satisfied. For these reasons, epistasis models may be interesting for predicting crossbreed performance or for the prediction of the phenotype of a plant line under different environmental conditions, but a practical usefulness in line breeding has not been demonstrated. However, this topic should also be addressed in more detail. The fact that the knowledge on the locations of the QTL does improve the long-term response in our simulation when selecting with an additive model, also supports the hope that predictions across different populations may be more robust when additional knowledge on the biology of the trait is incorporated (Snelling et al. 2013).

Potential future research topics

The field of statistical epistasis offers many open questions to be addressed. Extending this work, of special interest would be the variable selection problem in the context of epistasis models. A conceptually relatively simple approach could be the comparison of different variable selection methods or tests for interactions (for instance according to Wang et al. (2016); Li et al. (2016); Xu et al. (2016); Frost et al. (2016); Hung et al. (2016); Sung et al. (2016)) and the incorporation of the corresponding interactions in a prediction model, as it has been done based on the estimated absolute effect sizes in a ridge regression approach on the wheat data set in this work. The prediction of the performance of lines under different environmental conditions, which is often addressed by adapting the covariance structure of measurements to the covariance of environmental conditions (Cuevas et al. 2017) or gene×environment interaction, is an important topic in plant breeding and may also be addressed by means of selecting interactions.

Concerning our gene annotation based haploblock model, a key characteristic is which annotation class is used. We used haploblocks defined according to the location of protein coding genes to create a model built upon these biological functional units. Here, considering other types of units or several different types simultaneously with their own variance components may be interesting. Moreover, we used in our work marker based models and haploblock models as references, but not marker models using gene annotation (Morota et al. 2014; Do et al. 2015; MacLeod et al. 2016). A systematic comparison of these approaches on different data sets may be interesting. Addressing our gene annotation based haplotype models with epistasis, the question of whether the information on which variables interact could also be derived from external sources such as data bases on biochemical pathways, is also of extraordinary interest.

The previously mentioned question of how the marker density affects the predictive ability of marker based epistasis models should be investigated. An approach can be to follow the work by Ober et al. (2012) who reduced the marker density of a large data set step-wise and considered the predictive ability of GBLUP at different marker densities.

Finally, the question of whether and how an improved prediction of the total genetic value can be used in a breeding scheme to increase selection gain is still open and relevant. For instance, different crossbred breeding schemes could be checked for points at which a higher predictive ability of the phenotype could be of advantage. However, the results of this work on the targeted mating do not provide an optimistic perspective on this topic.

References

Abdollahi-Arpanahi, R., Morota, G., Valente, B. D., Kranis, A., Rosa, G. J., and Gianola, D. Differential contribution of genomic regions to marked genetic variation and prediction of quantitative traits in broiler chickens. *Genet Sel Evol*, 48:10, 2016. doi: 10.1186/s12711-016-0187-z.

Akdemir, D. and Godfrey, O. U. EMMREML: Fitting mixed models with known covariance structures. *R package version 3.1*, 2015. URL https://CRAN.R-project.org/package=EMMREML.

Albrecht, T., Auinger, H.-J., Wimmer, V., Ogutu, J. O., Knaak, C., Ouzunova, M., Piepho, H.-P., and Schön, C.-C. Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theor Appl Genet*, 127: 1375–1386, 2014. doi: 10.1007/s00122-014-2305-z.

Aschard, H. A perspective on interaction effects in genetic association studies. *Genet Epidemiol*, 40:678–688, 2016. doi: 10.1002/gepi.21989.

Bateson, W. *Mendel's Principles of heredity*. Cambridge University Press, 1909. doi: 10.5962/bhl.title.44575.

Bernardo, R. and Yu, J. Prospects for genomewide selection for quantitative traits in maize. *Crop Sci*, 47:1082–1090, 2007. doi: 10.2135/cropsci2006.11.0690.

Bian, Y. and Holland, J. B. Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. *Heredity*, 118:585–593, 2017. doi: 10.1038/hdy.2017.4.

Carlborg, Ö., Kerje, S., Schütz, K., Jacobsson, L., Jensen, P., and Andersson, L.

A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res*, 13:413–421, 2003. doi: 10.1101/gr.528003.

Carlborg, Ö., Jacobsson, L., Åhgren, P., Siegel, P., and Andersson, L. Epistasis and the release of genetic variation during long-term selection. *Nat Genet*, 38:418–420, 2006. doi: 10.1038/ng1761.

Chen, Q., Mao, X., Zhang, Z., Zhu, R., Yin, Z., Leng, Y., Yu, H., Jia, H., Jiang, S., Ni, Z., Jiang, H., Han, X., Liu, C., Hu, Z., Wu, X., Hu, G., Xin, D., and Qi, Z. SNP-SNP interaction analysis on soybean oil content under multi-environments. *PLOS ONE*, 11:e0163692, 2016. doi: 10.1371/journal.pone.0163692.

Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*, 11:2463–2468, 2002. doi: 10.1093/hmg/11.20.2463.

Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 10:392–404, 2009. doi: 10.1038/nrg2579.

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., and Braun, H. J. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186:713–724, 2010. doi: 10.1534/genetics.110.118521.

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P., and de los Campos, G. Bayesian genomic prediction with genotype× environment interaction kernel models. *G3-Genes Genom Genet*, 7:41–53, 2017. doi: 10.1534/g3.116.035584.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 185:1021–1031, 2010. doi: 10.1534/genetics.110.116855.

de los Campos, G., Gianola, D., and Rosa, G. J. M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J Anim Sci*, 87:1883–1887, 2009. doi: 10.2527/jas.2008-1259.

de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., and Crossa, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet Res*, 92:295–308, 2010. doi: 10.1017/S0016672310000285.

Do, D. N., Janss, L. L. G., Jensen, J., and Kadarmideen, H. N. SNP annotation-based whole genomic prediction and selection: an application to feed efficiency and its component traits in pigs. *J Anim Sci*, 93:2056–2063, 2015. doi: 10.2527/jas.2014-8640.

Edwards, S. M., Thomsen, B., Madsen, P., and Sørensen, P. Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. *Genet Sel Evol*, 47:60, 2015. doi: 10.1186/s12711-015-0132-6.

Ehrenreich, I. M. Epistasis: Searching for interacting genetic variants using crosses. *Genetics*, 206:531–535, 2017. doi: 10.1534/genetics.117.203059.

Esfandyari, H., Henryon, M., Berg, P., Thomasen, J. R., Bijma, P., and Sorensen, A. C. Response to selection in finite locus models with nonadditive effects. *J Hered*, 108:318–327, 2017. doi: 10.1093/jhered/esw123.

Falconer, D. S. and Mackay, T. F. C. *Introduction to Quantitative Genetics*. Pearson Education, London, 1996. ISBN 978-0582243026.

Fang, L., Sahana, G., Ma, P., Su, G., Yu, Y., Zhang, S., Lund, M. S., and Sørensen, P. Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. *BMC Genomics*, 18:604, 2017a. doi: 10.1186/s12864-017-4004-z.

Fang, L., Sahana, G., Ma, P., Su, G., Yu, Y., Zhang, S., Lund, M. S., and Sørensen, P. Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet Sel Evol*, 49:44, 2017b. doi: 10.1186/s12711-017-0319-0.

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

Forneris, N. S., Vitezica, Z. G., Legarra, A., and Pérez-Enciso, M. Influence of epistasis on response to genomic selection using complete sequence data. *Genet Sel Evol*, 49: 66, 2017. doi: 10.1186/s12711-017-0340-3.

Forsberg, S. K. G., Bloom, J. S., Sadhu, M. J., Kruglyak, L., and Carlborg, Ö. Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nat Genet*, 49:497–503, 2017. doi: 10.1038/ng.3800.

Frost, H. R., Amos, C. I., and Moore, J. H. A global test for gene-gene interactions based on random matrix theory. *Genet Epidemiol*, 40:689–701, 2016. doi: 10.1002/gepi.21990.

Galarza-Muñoz, G., Briggs, F. B. S., Evsyukova, I., Schott-Lerner, G., Kennedy, E. M., Nyanhete, T., Wang, L., Bergamaschi, L., Widen, S. G., Tomaras, G. D., Ko, D. C., Bradrick, S. S., Barcellos, L. F., Gregory, S. G., and Garcia-Blanco, M. A. Human epistatic interaction controls IL7R splicing and increases multiple sclerosis risk. *Cell*, 169:72–84, 2017. doi: 10.1016/j.cell.2017.03.007.

Gao, N., Martini, J. W. R., Zhang, Z., Yuan, X., Zhang, H., Simianer, H., and Li, J. Incorporating gene annotation into genomic prediction of complex phenotypes. *Genetics*, 2017. doi: 10.1534/genetics.117.300198.

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

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. Additive genetic variability and the Bayesian alphabet. *Genetics*, 183:347–363, 2009. doi: 10.1534/genetics.109.103952.

Gianola, D., Morota, G., and Crossa, J. Genome-enabled prediction of complex traits with kernel methods: What have we learned? *Proceedings, 10th World Congress of Genetics Applied to Livestock Production*, 2014.

Gianola, D. and Rosa, G. J. M. One hundred years of statistical developments in animal breeding. *Annu Rev Anim Biosci*, 3:19–56, 2015. doi: 10.1146/annurev-animal-022114-110733.

Gianola, D. and Van Kaam, J. B. C. H. M. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178:2289–2303, 2008. doi: 10.1534/genetics.107.084285.

González-Domínguez, J., Ramos, S., Touriño, J., and Schmidt, B. Parallel pairwise epistasis detection on heterogeneous computing architectures. *IEEE T Parall Distr*, 27:2329–2340, 2016. doi: 10.1109/TPDS.2015.2460247.

Grattapaglia, D. and Resende, M. D. V. Genomic selection in forest tree breeding. *Tree Genet Genomes*, 7:241–255, 2011. doi: 10.1007/s11295-010-0328-4.

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

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12:186, 2011. doi: 10.1186/1471-2105-12-186.

Haley, C. S. and Visscher, P. M. Strategies to utilize marker-quantitative trait loci associations. *J Dairy Sci*, 81:85–97, 1998. doi: 10.3168/jds.S0022-0302(98)70157-2.

Hallander, J. and Waldmann, P. The effect of non-additive genetic interactions on selection in multi-locus genetic models. *Heredity*, 98:349, 2007. doi: 10.1038/sj.hdy.6800946.

Harris, B. L. and Johnson, D. L. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J Dairy Sci*, 93:1243–1252, 2010. doi: 10.3168/jds.2009-2619.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J Dairy Sci*, 92:433–443, 2009a. doi: 10.3168/jds.2008-1646.

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

Hayes, B. J. and Goddard, M. E. Genome-wide association and genomic selection in animal breeding. *Genome*, 53:876–883, 2010. doi: 10.1139/G10-076.

Hayes, B. J., Cogan, N. O. I., Pembleton, L. W., Goddard, M. E., Wang, J., Spangenberg, G. C., and Forster, J. W. Prospects for genomic selection in forage plant species. *Plant Breeding*, 132:133–143, 2013. doi: 10.1111/pbr.12037.

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. Genomic selection for crop improvement. *Crop Sci*, 49:1–12, 2009. doi: 10.2135/cropsci2008.08.0512.

Henderson, C. R. Best linear unbiased prediction of breeding values not in the model for records. *J Dairy Sci*, 60:783–787, 1977. doi: 10.3168/jds.S0022-0302(77)83935-0.

Henderson, C. R. and Quaas, R. L. Multiple trait evaluation using relatives' records. *J Anim Sci*, 43:1188–1197, 1976. doi: 10.2527/jas1976.4361188x.

Henderson, C. R. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447, 1975. doi: 10.2307/2529430.

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. Genomic selection in plant breeding: a comparison of models. *Crop Sci*, 52:146–160, 2012. doi: 10.2135/cropsci2011.06.0297.

Heslot, N., Jannink, J.-L., and Sorrells, M. E. Perspectives for genomic selection applications and research in plants. *Crop Sci*, 55:1–12, 2015. doi: 10.2135/cropsci2014.03.0249.

Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., and "Implementing Genomic Selection in CGIAR Breeding Programs Workshop Participants". Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat Genet*, 49:1297–1303, 2017. doi: doi:10.1038/ng.3920.

Hill, W. G. "Conversion" of epistatic into additive genetic variance in finite populations and possible impact on long-term selection response. *J Anim Breed Genet*, 134:196–201, 2017. doi: 10.1111/jbg.12270.

Hill, W. G., Goddard, M. E., and Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genet*, 4:e1000008, 2008. doi: 10.1371/journal.pgen.1000008.

Huang, W. and Mackay, T. F. C. The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLOS Genet*, 12:e1006421, 2016. doi: 10.1371/journal.pgen.1006421.

Hung, H., Lin, Y.-T., Chen, P., Wang, C.-C., Huang, S.-Y., and Tzeng, J.-Y. Detection of gene–gene interactions using multistage sparse and low-rank regression. *Biometrics*, 72:85–94, 2016. doi: 10.1111/biom.12374.

Isik, F. Genomic selection in forest tree breeding: the concept and an outlook to the future. *New Forest*, 45:379–401, 2014. doi: 10.1007/s11056-014-9422-z.

Jannink, J.-L., Lorenz, A. J., and Iwata, H. Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics*, 9:166–177, 2010. doi: 10.1093/bfgp/elq001.

Jiang, Y. and Reif, J. C. Modeling epistasis in genomic selection. *Genetics*, 201:759–768, 2015. doi: 10.1534/genetics.115.177907.

Jing, J.-J., Lu, Y.-Z., Sun, L.-P., Liu, J.-W., Gong, Y.-H., Xu, Q., Dong, N.-N., and Yuan, Y. Epistatic SNP interaction of ERCC6 with ERCC8 and their joint protein expression contribute to gastric cancer/atrophic gastritis risk. *Oncotarget*, 8:43140, 2017. doi: 10.18632/oncotarget.17814.

Jünger, D., Hundt, C., González-Domínguez, J., and Schmidt, B. Ultra-fast detection of higher-order epistatic interactions on GPUs. *Parallel Processing Workshops*, Euro-Par 2016:421–432, 2017. doi: 10.1007/978-3-319-58943-5_34.

Kässens, J. C., Wienbrandt, L., Schimmler, M., González-Domínguez, J., and Schmidt, B. Combining GPU and FPGA technology for efficient exhaustive interaction analysis in GWAS. *IEEE 27th International Conference on Application-specific Systems,*

Architectures and Processors (ASAP), 2016:170–175, 2016. doi: 10.1109/ASAP.2016.7760788.

Li, J., Malley, J. D., Andrew, A. S., Karagas, M. R., and Moore, J. H. Detecting gene-gene interactions using a permutation-based random forest method. *BioData Min*, 9:14, 2016. doi: 10.1186/s13040-016-0093-5.

Lopes, M., Bovenhuis, H., van Son, M., Nordbø, Ø., Grindflek, E., Knol, E., and Bastiaansen, J. Using markers with large effect in genetic and genomic predictions. *J Anim Sci*, 95:59–71, 2017. doi: 10.2527/jas.2016.0754.

MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., Schrooten, C., Hayes, B. J., and Goddard, M. E. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*, 17:144, 2016. doi: 10.1186/s12864-016-2443-6.

Martini, J. W. R., Wimmer, V., Erbe, M., and Simianer, H. Epistasis and covariance: how gene interaction translates into genomic relationship. *Theor Appl Genet*, 129: 963–976, 2016. doi: 10.1007/s00122-016-2675-5.

Martini, J. W. R., Gao, N., Cardoso, D. F., Wimmer, V., Erbe, M., Cantet, R. J. C., and Simianer, H. Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). *BMC Bioinformatics*, 18:3, 2017. doi: 10.1186/s12859-016-1439-1.

Martini, J. W. R., Schrauf, M. F., Garcia-Baccino, C. A., Pimentel, E. C. G., Munilla, S., Rogberg-Muñoz, A., Cantet, R. J. C., Reimer, C., Gao, N., Wimmer, V., and Simianer, H. The effect of the H^{-1} scaling factors τ and ω on the structure of H in the single-step procedure. *Genet Sel Evol*, 50(1):16, 2018a. doi: 10.1186/s12711-018-0386-x.

Martini, J. W., Rosales, F., Ha, N.-T., Kneib, T., Heise, J., and Wimmer, V. Lost in translation: On the impact of data coding on penalized regression with interactions. *arXiv preprint arXiv:1806.03729*, 2018b.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829, 2001. URL http://www.genetics.org/content/157/4/1819.

Meuwissen, T. H. E., Odegard, J., Andersen-Ranberg, I., and Grindflek, E. On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet Sel Evol*, 46:49, 2014. doi: 10.1186/1297-9686-46-49.

Morota, G. and Gianola, D. Kernel-based whole-genome prediction of complex traits: a review. *Front Genet*, 5:363, 2014. doi: 10.3389/fgene.2014.00363.

Morota, G., Abdollahi-Arpanahi, R., Kranis, A., and Gianola, D. Genome-enabled prediction of quantitative traits in chickens using genomic annotation. *BMC Genomics*, 15:109, 2014. doi: 10.1186/1471-2164-15-109.

Mrode, R. A. *Linear models for the prediction of animal breeding values*. CABI Publishing, Wallingford, 2014. ISBN 978-1780643915.

Nakaya, A. and Isobe, S. N. Will genomic selection be a practical method for plant breeding? *Ann Bot*, 110:1303–1316, 2012. doi: 10.1093/aob/mcs109.

Newell, M. A. and Jannink, J.-L. Genomic selection in plant breeding. *Crop Breeding: Methods and Protocols*, pages 117–130, 2014. doi: 10.1007/978-1-4939-0446-4_10.

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., Simianer, H., and Hoeschele, I. Predicting genetic values: A kernel-based best linear unbiased prediction with genomic data. *Genetics*, 188:695–708, 2011. doi: 10.1534/genetics.111.128694.

Ober, U., Ayroles, J. F., Stone, E. A., Richards, S., Zhu, D., Gibbs, R. A., Stricker, C., Gianola, D., Schlather, M., Mackay, T. F. C., and Simianer, H. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLOS Genet*, 8:e1002685, 2012. doi: 10.1371/journal.pgen.1002685.

Ober, U., Huang, W., Magwire, M., Schlather, M., Simianer, H., and Mackay, T. F. C. Accounting for genetic architecture improves sequence based genomic prediction for a *Drosophila* fitness trait. *PLOS ONE*, 10:e0126880, 2015. doi: 10.1371/journal.pone.0126880.

Paixão, T. and Barton, N. H. The effect of gene interactions on the long-term response to selection. *P Natl Acad Sci USA*, 113:4422–4427, 2016. doi: 10.1073/pnas.1518830113.

Reinhardt, F., Liu, Z., Seefried, F., and Thaller, G. Implementation of genomic evaluation in german holsteins. *Interbull Bulletin*, 40:219–226, 2009. URL https://journal.interbull.org/index.php/ib/article/viewFile/1116/1107.

Resende, M. F. R., Munoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., Resende, M. D. V., and Kirst, M. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol*, 193:617–624, 2012. doi: 10.1111/j.1469-8137.2011.03895.x.

Sackton, T. B. and Hartl, D. L. Genotypic context and epistasis in individuals and populations. *Cell*, 166:279–287, 2016. doi: 10.1016/j.cell.2016.06.047.

Sarup, P., Jensen, J., Ostersen, T., Henryon, M., and Sørensen, P. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred danish duroc pigs. *BMC Genet*, 17:11, 2016. doi: 10.1186/s12863-015-0322-9.

Schaeffer, L. Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet*, 123:218–223, 2006. doi: 10.1111/j.1439-0388.2006.00595.x.

Shengqiang, Z., Dekkers, J. C. M., Fernando, R. L., and Jannink, J.-L. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics*, 182:355–364, 2009. doi: 10.1534/genetics.108.098277.

Snelling, W. M., Cushman, R. A., Keele, J. W., Maltecca, C., Thomas, M. G., Fortes, M. R. S., and Reverter, A. Breeding and genetics symposium: Networks and pathways to guide genomic selection. *J Anim Sci*, 91:537–552, 2013. doi: 10.2527/jas.2012-5784.

Sohail, M., Vakhrusheva, O. A., Sul, J. H., Pulit, S. L., Francioli, L. C., Genome of the Netherlands Consortium, Alzheimer's Disease Neuroimaging Initiative, van den Berg, L. H., Veldink, J. H., de Bakker, P. I. W., Bazykin, G. A., Kondrashov, A. S., and Sunyaev, S. R. Negative selection in humans and fruit flies involves synergistic epistasis. *Science*, 356:539–542, 2017. doi: 10.1126/science.aah5238.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. E. Genomic selection using different marker types and densities. *J Anim Sci*, 86:2447–2454, 2008. doi: 10.2527/jas.2007-0010.

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redona, E., Jannink, J.-L., and McCouch, S. Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity*, 116:395–408, 2016. doi: 10.1038/hdy.2015.113.

Su, G., Christensen, O. F., Janss, L., and Lund, M. S. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. *J Dairy Sci*, 97:6547–6559, 2014. doi: 10.3168/jds.2014-8210.

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., and Lund, M. S. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLOS ONE*, 7:e45293, 2012. doi: 10.1371/journal.pone.0045293.

Su, L., Meng, X., Ma, Q., Bai, T., and Liu, G. LPRP: A gene–gene interaction network construction algorithm and its application in breast cancer data analysis. *Interdiscip Sci Comput Life Sci*, 2016. doi: 10.1007/s12539-016-0185-4.

Sung, P.-Y., Wang, Y.-T., Yu, Y.-W., and Chung, R.-H. An efficient gene–gene interaction test for genome-wide association studies in trio families. *Bioinformatics*, 32: 1848–1855, 2016. doi: 10.1093/bioinformatics/btw077.

Technow, F., Riedelsheimer, C., Schrag, T. A., and Melchinger, A. E. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor Appl Genet*, 125(6):1181–1194, 2012. doi: 10.1007/s00122-012-1905-8.

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

Tyler, A. L., Donahue, L. R., Churchill, G. A., and Carter, G. W. Weak epistasis generally stabilizes phenotypes in a mouse intercross. *PLOS Genet*, 12:e1005805, 2016. doi: 10.1371/journal.pgen.1005805.

Uppu, S. and Krishna, A. Improving strategy for discovering interacting genetic variants in association studies. *International Conference on Neural Information Processing*, 2016:461–469, 2016. doi: 10.1007/978-3-319-46687-3_51.

Uppu, S., Krishna, A., and Gopalan, R. P. Towards deep learning in genome-wide association interaction studies. *Pacific Asia Conference on Information Systems*, 2016:1–11, 2016. URL http://aisel.aisnet.org/pacis2016/20.

Valdar, W., Solberg, L. C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W. O., Taylor, M. S., Rawlins, J. N. P., Mott, R., and Flint, J. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet*, 38:879–887, 2006a. doi: 10.1038/ng1840.

Valdar, W., Solberg, L. C., Gauguier, D., Cookson, W. O., Rawlins, J. N. P., Mott, R., and Flint, J. Genetic and environmental effects on complex traits in mice. *Genetics*, 174:959–984, 2006b. doi: 10.1534/genetics.106.060004.

VanRaden, P. M. Efficient methods to compute genomic predictions. *J Dairy Sci*, 91: 4414–4423, 2008. doi: 10.3168/jds.2007-0980.

Veroneze, R., Lopes, P. S., Lopes, M. S., Hidalgo, A. M., Guimarães, S. E. F., Harlizius, B., Knol, E. F., van Arendonk, J. A. M., Silva, F. F., and Bastiaansen, J. W. M. Accounting for genetic architecture in single-and multipopulation genomic prediction using weights from genomewide association studies in pigs. *J Anim Breed Genet*, 133: 187–196, 2016. doi: 10.1111/jbg.12202.

Wang, M. H., Sun, R., Guo, J., Weng, H., Lee, J., Hu, I., Sham, P. C., and Zee, B. C.-Y. A fast and powerful W-test for pairwise epistasis testing. *Nucleic Acids Res*, 44:e115–e115, 2016. doi: 10.1093/nar/gkw347.

Wong, A. S. L., Choi, G. C. G., and Lu, T. K. Deciphering combinatorial genetics. *Annu Rev Genet*, 50:515–538, 2016. doi: 10.1146/annurev-genet-120215-034902.

Xu, J., Yuan, Z., Ji, J., Zhang, X., Li, H., Wu, X., Xue, F., and Liu, Y. A powerful score-based test statistic for detecting gene-gene co-association. *BMC Genet*, 17:31, 2016. doi: 10.1186/s12863-016-0331-3.

Xu, S., Zhu, D., and Zhang, Q. Predicting hybrid performance in rice using genomic best linear unbiased prediction. *P Natl Acad Sci USA*, 111:12456–12461, 2014. doi: 10.1073/pnas.1413750111.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*, 46: 100–106, 2014. doi: 10.1038/ng.2876.

Zhang, Z., Ober, M., Uand Erbe, Zhang, H., Gao, N., He, J., Li, J., and Simianer, H. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLOS ONE*, 9:e93017, 2014. doi: 10.1371/journal.pone.0093017.

Zhang, Z., Erbe, M., He, J., Ober, U., Gao, N., Zhang, H., Simianer, H., and Li, J. Accuracy of Whole Genome Prediction Using a Genetic Architecture Enhanced Variance-Covariance Matrix. *G3-Genes Genom Genet*, 5:615–627, 2015. doi: 10.1534/g3.114.016261.

Zhao, Y., Mette, M. F., and Reif, J. C. Genomic selection in hybrid breeding. *Plant Breeding*, 134:1–10, 2015. doi: 10.1111/pbr.12231.

Appendix

Short Curriculum Vitae

Personal details

Name:	Johannes Wolfgang Robert Martini
Date of birth:	5th of April 1983 in Kronach, Germany

Education and Career

10/2014 - 09/2017	**Scientific Assistant** Animal Breeding and Genetics, University of Goettingen
2015 & 2016	**Scientific exchange** (twice 4 months) Group of Quantitative Genetics, Prof. Dr. Rodolfo JC Cantet Universidad de Buenos Aires, Argentina
04/2014 - 09/2014	**Postdoc** Max-Planck-Institute for Developmental Biology, Tuebingen
03/2014	**Doctorate Dr. rer. nat.** University of Mannheim („magna cum laude“) *Thesis:* „Excursions in the Theory of Ligand Binding“
04/2011 - 03/2014	**Scientific Assistant** Institute for Mathematical Stochastics, University of Goettingen
10/2005 - 03/2011	**Studies of Mathematics** University of Bayreuth, Dipl.-Math. (final grade: 1.2) *Thesis:* „Robust Estimation Of Time Series Models With Regression-Type Score Function“
10/2003 - 01/2009	**Studies of Biology** University of Bayreuth, Dipl.-Biol. (final grade: 1.3) *Major: Molecular and Cell biology* *Thesis:* „Analysis Of SRP-Dependent Secretion In *Bacillus Subtilis*“

Languages

German (native), English (fluently), Spanish (fluently), French (basic knowledge)

Programming Skills

Linux, R (advanced), Matlab/Scilab, Maxima, Magma (basic)

Other work related to the PhD project

Posters, presentations (selected) and projects

07-2014 **Project Presentation**
KWS SAAT SE

02-2015 **Incorporating Epistasis Effects into Genomic Prediction**
Presentation, KWS SAAT SE

03-2015 **Incorporating Gene Interaction into Genomic Prediction**
Presentation, KWS SAAT SE

09-2015 **A Framework to Incorporate Knowledge on Gene Interaction into Genomic Relationship**
Presentation (coauthor), 66th EAAP Annual Meeting, Warsaw, Poland

03-2016 **Genomic Prediction with Epistasis Models: Properties, Problems and Perspectives**
Poster, DAGStat 2016, Göttingen

06-2016 **Genomic Prediction Based on Interaction Networks of Markers: How to incorporate Prior Experimental Information**
Poster and selected poster presentation, ICQG 5, Madison, Wisconsin, USA

09-2016 **Nicht-additive Verwandtschaftsmodelle und deren Nutzen für die genomische Vorhersage**
Presentation, DGfZ Jahrestagung 2016, Hannover

10-2016 **Epistasie: Allgegenwärtig aber entbehrlich?**
Presentation, "Genetisch-Statistischer Ausschuss" of the DGfZ

02-2017 **Genomic Selection and Measures of Kinship**
Granted DAAD project, PPP Argentina

03-2017 **On Non-additive Marker Effect Models for Genomic Prediction and Their Potential Usefulness in Breeding**
Presentation, Young Excellence Seminar of KWS SAAT SE

05-2017 **Integration of Biological Knowledge into Genomic Prediction**
Invited talk, Workshop Biometrische Aspekte der Genomanalyse 2017, Heidelberg

09-2017 **On the Usefulness of the Prediction of Total Genetic Values in Livestock Breeding Programs**
Presentation, 68th EAAP Annual Meeting, Tallinn, Estonia

Research stays and additional courses attended

08-2015	**Research stay**
until	Group of Animal Breeding (Prof. Cantet)
12-2015	University of Buenos Aires, Argentina
06-2016	**Short courses** "Applied Plant Genomics and Bioinformatics" and "Computational Approaches for Inference and Analysis of Molecular Networks" Madison, Wisconsin, USA
10-2016	**Research stay**
until	Group of Animal Breeding (Prof. Cantet)
02-2017	University of Buenos Aires, Argentina

Additional publications related to the PhD project

Martini, J. W. R., Gao, N., Wimmer, V., and Simianer, H. Nicht-additive Verwandtschaftsmodelle und deren Nutzen für die genomische Vorhersage. Vortragstagung der DGfZ und GfT am 20./21. September 2016 in Hannover

Gao, N., Martini, J. W. R., Zhang, Z., Li, J., and Simianer, H. Approaches to incorporate genome annotation into genomic prediction. Vortragstagung der DGfZ und GfT am 20./21. September 2016 in Hannover

Martini, J. W. R., Schrauf, M. F., Garcia-Baccino, C. A., Pimentel, E. C. G., Munilla, S., Rogberg-Muñoz, A., Cantet, R. J. C., Reimer, C., Gao, N., Wimmer, V., and Simianer, H. The effect of the $\mathbf{H}^{-1}$ scaling factors τ and ω on the structure of $\mathbf{H}$ in the single-step procedure. *Genet Sel Evol*, 50:16, 2018. doi: 10.1186/s12711-018-0386-x.

Anlage 3: E r k l ä r u n g e n

1. Hiermit erkläre ich, dass diese Arbeit weder in gleicher noch in ähnlicher Form bereits anderen Prüfungsbehörden vorgelegen hat.

   Weiter erkläre ich, dass ich mich an keiner anderen Hochschule um einen Doktorgrad beworben habe.

   Göttingen, den .....

   .....
   (Unterschrift)

2. Hiermit erkläre ich eidesstattlich, dass diese Dissertation selbständig und ohne unerlaubte Hilfe angefertigt wurde.

   Göttingen, den .....

   .....
   (Unterschrift)

3. Der Eigenanteil an den Publikationen ist in den jeweiligen “Author contribution statements” festgehalten. Hiermit bestätige ich, dass diese dem wirklichen Eigenanteil entsprechen.

Göttingen, den 30.09.17