Theoretical and Applied Genetics (2019) 132:1211–1222
https://doi.org/10.1007/s00122-018-03273-1

ORIGINAL ARTICLE

Efficient genetic value prediction using incomplete omics data

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

Received: 24 September 2018 / Accepted: 21 December 2018 / Published online: 17 January 2019

Abstract

Key message **Covering a subset of individuals with a quantitative predictor, while imputing records for all others using pedigree or genomic data, could improve the precision of predictions while controlling for costs.**

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

Introduction

Genomic prediction, pioneered by Meuwissen et al. (2001), has revolutionized animal and plant breeding by a more efficient selection of promising candidates without phenotypic records (de los Campos et al. 2013; García-Ruiz et al. 2016). Its use is mainly motivated by the high costs of time-intensive phenotyping. Costs can be reduced when all genotypes, e.g., inbred lines in plant breeding and animals in animal breeding, are covered by a predictor while limiting phenotyping to a small number of genotypes with high precision (Kadam et al. 2016). The performance of all other genotypes is then forecast using model parameters estimated in the subset of individuals with full information. In animal breeding, genomic selection has increased the annual selection gain by ≈ 50–100% for yield traits and by 300–400% for traits with low heritability (García-Ruiz et al. 2016).

Breeding programs—particularly in animal breeding—are characterized by a predictor type, which is available for all recorded genotypes with phenotypic information and possibly other predictors which are incomplete in that they cover only a subset of all genotypes (Fragomeni et al. 2015). Usually, the complete predictor is pedigree information, while genomic data are the most prevalent incomplete predictor. To utilize all available phenotypic information for model training, methods were developed that allow the combined use of genomic and pedigree information (Hayes et al. 2009; VanRaden et al. 2009).

Initially, a two-step procedure was the most intuitive choice for accomplishing this goal (VanRaden et al. 2009).

Communicated by Matthias Frisch.

This project was funded by the German Federal Ministry of Education and Research (BMBF) within the projects OPTIMAL (FKZ: 0315958B, 0315958F), SYNBREED (FKZ: 0315528D) and by the German Research Foundation (DFG, Grants No. ME 2260/5-1).

✉ Albrecht E. Melchinger
melchinger@uni-hohenheim.de

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

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

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

[4] Department of Animal Science, Iowa State University, Ames, IA 50011, USA

The first step consists of conventionally estimating breeding values based on pedigree relationships among the candidates and phenotypes of their relatives. Individuals with highly reliable breeding values out of the first step, i.e., those having a high number of offspring, enter the second step in which conventional breeding values are regressed onto SNP genotypes. The estimated model coefficients can then be used to predict breeding values for solely genotyped and potentially very young individuals. However, as the breeding values from step one enter the genomic prediction model as phenotypes, the residual error distribution of such a model must be considered. For instance, the range of the prediction errors will no longer be identically distributed and thus, they should be weighted differently or estimated individually (Aguilar et al. 2010). One popular approach to tackle this problem is deregressing the breeding values and weighting the residuals according to the prediction error variances (Garrick et al. 2009).

Legarra et al. (2009) and Christensen and Lund (2010), independently from each other, developed a single-step BLUP method that blends the numerator relationship matrix **A** with the genomic relationship matrix **G** in a mutual matrix **H** to use all available predictor information simultaneously. An issue with this H-BLUP approach is the compatibility of **A** with **G** (Christensen 2012) and their weighting, which is not trait-independent (Vitezica et al. 2011; Ashraf et al. 2016). Fernando et al. (2014) derived a single-step marker-effect-model that is equivalent to the single-step breeding value model of Legarra et al. (2009) and Christensen and Lund (2010), which does not require weights for **A** and **G** and allows explicit modeling of the imputation errors.

In animal breeding, single-step prediction, particularly H-BLUP, has become a routine procedure (Legarra et al. 2014) for predicting breeding values and is slowly gaining consideration in plants (Ashraf et al. 2016; Ratcliffe et al. 2017). While an extension of the single-step framework to hybrid genotypes was developed for pigs (Xiang et al. 2016), it is yet lacking for the prediction of single-cross hybrids between pure-breeding lines. Technological advances in hybrid breeding of several crop species, such as the rapid generation of completely homozygous lines using the doubled-haploid technology (Wedzony et al. 2009) or ‘Speed breeding’ (Watson et al. 2018), now enable plant breeders to produce $n = 1000$ or more lines per generation and seed or pollen parent, respectively. Assuming that about 90% of these lines were never tested in any hybrid combination, about $n^2 * 0.9^2 = 810{,}000$ putative hybrids per generation would not have a single parent with phenotypically tested progeny. Predicting the performance of these hybrid candidates with low relatedness to tested hybrids using only information from their parents is currently a hot topic in plant breeding (Technow et al. 2014; Kadam et al. 2016; Westhues et al. 2017; Schrag et al. 2018).

Given that the availability of pedigree and genomic data is the most prevalent situation in animal breeding, previous studies have focused exclusively on the application of single-step prediction to these predictors. Recent investigations into the utility of other ‘omics’ data for boosting predictive abilities in plants have suggested that metabolic and transcriptomic data can improve upon predictive abilities based on pedigree or genomic data (Guo et al. 2016; Xu et al. 2016; Dan et al. 2016; Zenke-Philippi et al. 2017; Westhues et al. 2017). These two ‘omics’ technologies are still under development and therefore incur higher costs to breeders than pedigree and genomic data, putting their use into question despite their proven value. A pioneering study by Gamazon et al. (2015) outlined a framework—called PrediXcan—for improving association mapping by harnessing the advantages of gene expression data over genetic markers. PrediXcan infers gene expression data for individuals having only genetic information from individuals having both genetic information and gene expression data to correlate these features with disease traits. As pointed out by the authors of this publication, their method could not account for the attenuation bias associated with the imputation error. Theoretically, multivariate analyses, treating gene expression variables as phenotypes, would be better suited to tackle this problem. Computationally, however, such an approach is impractical in conjunction with appropriate cross-validation methods given that technologies such as RNA-seq now routinely produce thousands of features (Dey et al. 2016).

Hence, our objectives were to (i) explore the single-step prediction framework as a viable improvement over current methods for imputing ‘omics’ predictors, (ii) evaluate the impact of the proportion of imputed genotypes in the training set on the predictive ability and (iii) transfer the single-step framework for breeding value prediction to the prediction of hybrids derived from pure-breeding inbred lines.

Materials and methods

Experiment 1

To address objectives (i) and (ii), a published dataset of maize inbred lines (Yang et al. 2014) was used. It originally comprised 513 maize lines representing the global maize diversity and was reduced to the set of tropical and subtropical lines ($n = 211$), subsequently referred to as Exp1, which is the largest of the four subgroups classified by Guo et al. (2016). All inbred lines were evaluated in five different environments in China described in detail by Yang et al. (2014). Best linear unbiased predictors (BLUPs) were calculated by Yang et al. (2014) for all 211 inbred lines and for 17 traits of which six are analyzed in this study. These traits were selected based on our prediction results so that two traits were predicted with greater precision by transcriptomic data and four further traits were predicted with greater precision by genomic data. After filtering for minor allele frequency (≥ 5%), heterozygosity rate (≤ 5%) and call frequency (≥ 95%), missing genotypes were imputed using the Beagle software (Browning and Browning 2009) resulting in 37,760 SNPs for these 211 genotypes. Transcriptomic data on 28,850 annotated genes were available for 149 out of the 211 genotypes (Fu et al. 2013) and provided the basis for further analyses.

To mitigate the possibility of inflated predictive abilities, which can arise from large differences in the average performance of the different subpopulations (Windhausen et al. 2012), we employed a STRUCTURE analysis (Pritchard et al. 2000) on the SNP data, exploring eight ($K \in [2..9]$) putative ancestral populations. Based on the estimated likelihood of the data for each K, we chose $K = 4$ and carried out a principal component analysis (PCA) on the scaled and centered genomic feature matrix. By taking into account the primary assignment of genotypes to any of the four putative ancestral populations, we reduced the panel of inbred lines to 164 genotypes that were clustered together in the PCA plot.

In summary, genomic and phenotypic information was available for 164 inbred lines, which we denote as the 'Full Set,' whereas transcriptomic data were available only for a subset of 110 inbred lines, which we denote as the 'Reduced Set.'

Experiment 2

To address objectives (i) and (iii), a maize hybrid dataset, subsequently denoted as Exp2, was used. The material comprised 1,521 hybrids produced in 16 factorial mating designs between 142 Dent and 103 Flint parent lines described in detail by Westhues et al. (2017). Best linear unbiased estimates (BLUEs) were computed across all factorials and three or more agro-ecologically diverse environments across Germany for seven agronomically important traits in silage maize by accounting for year, location, field replication, block and genotype effects as well as their interactions. All 245 parent lines have pedigrees reaching back at least to their grandparents (Westhues et al. 2017) and were genotyped using the Illumina BeadChip MaizeSNP50 (Ganal et al. 2011). The same SNP quality checks as for the material in Exp1 were applied here, yielding 7013 polymorphic marker loci for the Dent and 6212 for the Flint lines, respectively. Gene expression data were obtained for a subset of 60 Dent and 43 Flint lines, which are parents of 685 hybrid progeny. Similarly to Exp1, gene expression data were sampled at a very early developmental stage (7 days after sowing from seedlings) under highly standardized conditions to ensure that genotype by environment interactions are negligible. These lines were allocated to six microarrays in a partially replicated design, and their gene expression data were obtained through two-color hybridizations using the same custom 2K microarray (GPL22267) for each factorial (Westhues et al. 2017). After applying established normalization procedures (Smyth and Speed 2003; Ritchie et al. 2007) separately to each factorial using the R-package *limma* (Ritchie et al. 2015), gene expression-derived BLUEs were computed for the 103 parents using 1323 transcripts while accounting for the different microarrays by modeling them as a fixed effect (Westhues et al. 2017).

Kernels

Depending on the dataset, up to three predictors were available for agronomic trait predictions, namely pedigree data (P), genomic data (G) and transcriptomic data (T). The corresponding feature matrix ($\mathbf{M}$) for the inbred lines has dimensions $n \times p$, where n denotes the number of genotypes and p the number of features, i.e., number of transcripts for T and number of SNPs for G. For the sake of simplicity, the derivations are detailed first for predicting genetic values of inbred lines. Later, we briefly describe the extension to models required for predicting the genetic value of hybrids. For the complete predictor, we generated kernels by centering and standardizing all features in $\mathbf{M}$ to unit variance (VanRaden 2008) as

$$\mathbf{K} = \frac{1}{p}\mathbf{M}\mathbf{M}^{\top}. \tag{1}$$

In the case of pedigree data (P), the numerator relationship matrix was used directly for $\mathbf{K}$.

Genetic model

The general model for genetic values of the lines was:

$$\mathbf{y} = \mu + \mathbf{Z}\mathbf{g} + \mathbf{e}, \tag{2}$$

where $\mathbf{y}$ is the vector of observed inbred line performance, μ is the fixed model intercept, $\mathbf{Z}$ is a design matrix that relates the vector $\mathbf{g}$ of random genetic effects, which has null mean and covariance equal to $\mathbf{K}\sigma^2_g$, to the vector of performance $\mathbf{y}$ and $\mathbf{e}$ is a vector of residual errors with null mean and covariance $\mathbf{I}\sigma^2_e$. Prediction models were implemented using the R-package *BGLR* (Pérez and de Los Campos 2014) and additionally the R-package *sspredr* https://github.com/mwesthues/sspredr when one predictor was imputed. All variance components were estimated by Gibbs sampling. For each model, the sampler was run for 30,000 iterations, half of which were discarded as burn-in and default hyper-parameters of the *BGLR* package, as described in Pérez and de Los Campos (2014), were used for the prior distributions.

Single-step prediction

If the genetic values are fully captured by the features in $\mathbf{M}$, Eq. (2) can be expressed as

$$\mathbf{y} = \mu + \mathbf{Z}\mathbf{M}\boldsymbol{\alpha} + \boldsymbol{e}, \tag{3}$$

(Fernando et al. 2014), with $\boldsymbol{\alpha}$ being the vector of random, partial-regression coefficients of the feature covariates, because the ridge regression model is equivalent to the BLUP model (Ruppert et al. 2003). Hence, the genetic values from the prediction model are given by:

$$\hat{\mathbf{g}} = \mathbf{M}\hat{\boldsymbol{\alpha}}, \tag{4}$$

where $\hat{\boldsymbol{\alpha}}$ is the solution to the ridge regression model. The relationship between the variance of marker effects σ^2_α and the variance σ^2_g of genetic values can be computed either as $\sigma^2_g/2\sum_j p_j(1-p_j)$, where p_j is the frequency of SNP j (Habier et al. 2007), or as $\sigma_g = \sigma_a/p$ (Mrode 2014, p. 183).

Consider now the situation in which one predictor is complete in the sense that the whole population is covered by features for that predictor but only a subset of that population has information for a second predictor. We can take a similar route as in Fernando et al. (2014) and impute covariates of the incomplete predictor, denoted as 'predictor' and indicated through the asterisk symbol '*', by using covariates of the complete predictor, denoted as 'imputor.' Let the subscript 1 denote the subset of individuals covered only by the imputor, whereas the subset of individuals covered by both the imputor and the predictor is indicated by the subscript 2.

Let us further denote the matrix of centered covariates of the predictor as

$$\mathbf{W}^* = \begin{bmatrix} \hat{\mathbf{W}}^*_1 \\ \mathbf{W}^*_2 \end{bmatrix}, \tag{5}$$

where the elements of $\mathbf{W}^*_2$ are the observed feature covariates while the feature covariates for genotypes not covered by the predictor can be imputed with the aid of the covariance matrix of the imputor $\mathbf{K}$ and the feature matrix of the predictor $\mathbf{W}^*_2$ by calculating ($\hat{\mathbf{W}}^*_1 = \mathbf{K}_{12}\ \mathbf{K}^{-1}_{22}\ \mathbf{W}^*_2$). In this case, we need to partition the vector of genetic values $\mathbf{g}$ into a component $\mathbf{g}_1$ (i.e., genetic values of genotypes covered only by the imputor) and a component $\mathbf{g}_2 = \mathbf{W}^*_2\ \boldsymbol{\alpha}^*$ (i.e., genetic values of genotypes covered by both the predictor and the imputor where $\boldsymbol{\alpha}^*$ pertains to the random feature effects of the predictor). Genetic values ($\mathbf{g}_1$ and $\mathbf{g}_2$) are distributed according to a multivariate normal distribution:

$$\begin{aligned}\mathbf{g_1} &= E(\mathbf{g}_1|\mathbf{g}_2) + \boldsymbol{\epsilon} \\ &= \mathbf{K}_{12}\ \mathbf{K}^{-1}_{22}\ \mathbf{W}^*_2\ \hat{\boldsymbol{\alpha}}^* + (\mathbf{g_1} - \mathbf{K}_{12}\ \mathbf{K}^{-1}_{22}\ \mathbf{W}^*_2\ \hat{\boldsymbol{\alpha}}^*) \\ &= \hat{\mathbf{g}}_1 + \boldsymbol{\epsilon},\end{aligned} \tag{6}$$

where the first term pertains to $\hat{\mathbf{g}}_1$, i.e., the genetic values estimated by imputing the predictor, the second term pertains to the residual imputation error $\boldsymbol{\epsilon}$ and where the partitions of the matrix

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}^*_{22} \end{bmatrix}, \tag{7}$$

correspond to $\mathbf{g}_1$ and $\mathbf{g}_2$, respectively (Fernando et al. 2014).

The covariance matrix of $\boldsymbol{\epsilon}$ is $(\mathbf{K}_{11} - \mathbf{K}_{12}\ \mathbf{K}^{-1}_{22}\ \mathbf{K}_{21})\sigma^2_g = (\mathbf{K}^{11})^{-1}\sigma^2_g$ (Legarra et al. 2009), and the covariance of $\hat{\mathbf{g}}_1$ is $\mathbf{K}_{12}\ \mathbf{K}^{-1}_{22}\ \mathbf{W}^*_2\ \mathbf{W}^{*\top}_2\ \mathbf{K}^{-1}_{22}\ \mathbf{K}_{21}\ \sigma^2_{\alpha^*}$, (Fernando et al. 2014), where $\sigma^2_{\alpha^*}$ was treated as an unknown with a scaled-inverse Chi-square prior (Pérez and de Los Campos 2014).

The general model for the single-step procedure can then be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{W}^*\ \boldsymbol{\alpha}^* + \mathbf{U}\boldsymbol{\epsilon} + \mathbf{e}, \tag{8}$$

with

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix}, \mathbf{U} = \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{0} \end{bmatrix}. \tag{9}$$

Using the notation in Fernando et al. (2014), the phenotypes can be untangled as:

$$\begin{aligned}\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix}\begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} + \mathbf{e} \\ &= \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix}\begin{bmatrix} \mathbf{K}_{12}\ \mathbf{K}^{-1}_{22}\ \mathbf{W}^*_2\ \boldsymbol{\alpha}^* + \boldsymbol{\epsilon} \\ \mathbf{W}^*_2\ \boldsymbol{\alpha}^* \end{bmatrix} + \mathbf{e} \\ &= \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix}\begin{bmatrix} \hat{\mathbf{W}}^*_1\ \boldsymbol{\alpha}^* + \boldsymbol{\epsilon} \\ \mathbf{W}^*_2\ \boldsymbol{\alpha}^* \end{bmatrix} + \mathbf{e}\end{aligned} \tag{10}$$

The model above indicates that single-step prediction models ultimately use the matrix of known and imputed feature covariates $\mathbf{W}^*$ and the imputation error $\boldsymbol{\epsilon}$ for the estimation of genetic values.

Genetic values can be estimated as:

$$\hat{\mathbf{g}} = \begin{bmatrix} \mathbf{Z}_1\hat{\mathbf{W}}^*_1 \\ \mathbf{Z}_2\mathbf{W}^*_2 \end{bmatrix}\hat{\boldsymbol{\alpha}}^* + \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{0} \end{bmatrix}\hat{\boldsymbol{\epsilon}} \tag{11}$$

The previous models can simply be extended to hybrids by defining separate feature matrices for each predictor and each heterotic group, here $\mathbf{M}_D$ for Dent and $\mathbf{M}_F$ for Flint parent lines in Exp2. The general model for the calculation of genetic values of hybrids in Exp2 is then

$$\mathbf{y} = \mu + \mathbf{Z}_D\mathbf{g}_D + \mathbf{Z}_F\mathbf{g}_F + \mathbf{e}, \tag{12}$$

corresponding to Eq. 2, ignoring effects for specific combining ability.

Here, we did not consider the inclusion of dominance effects for the hybrid material because previous investigations into the predictive ability of various ‘omics’ predictors in the same material concluded no considerable gains (Westhues et al. 2017; Schrag et al. 2018).

Alternatively, one could model effects for the hybrids directly instead of partitioning them into general combining effects for each heterotic group plus a specific combining ability term. The corresponding model would be

$$\mathbf{y} = \mathbf{1}\mu + (\mathbf{Z}_D + \mathbf{Z}_F)\mathbf{g} + \mathbf{e}, \tag{13}$$

corresponding to Eq. 2, ignoring effects for specific combining ability. The elements of the incidence matrices $\mathbf{Z}_D$ and $\mathbf{Z}_F$ are either 0.5, if the marker allele was equal to the reference allele, respectively, -0.5, if the marker allele was different from the reference allele (Technow et al. 2012). The matrix $(\mathbf{Z}_D + \mathbf{Z}_F)$ pertains to the genotype of the hybrids. For each hybrid from crosses of lines $i \times j$, the corresponding elements in $(\mathbf{Z}_D + \mathbf{Z}_F)$ can be obtained as the sum of the respective elements $\mathbf{z}_{il}$ in $\mathbf{Z}_D$ and $\mathbf{z}_{jl}$ in $\mathbf{Z}_F$, where l denotes the lth marker locus. We did not consider this model because it is not clear how to define $(\mathbf{Z}_D + \mathbf{Z}_F)$ for quantitative predictors but it might be useful to plant breeders, having merely pedigree information of parent lines as well as genomic information at their disposal. Answering the utility of single-step prediction with shallow pedigrees, however, is out of the scope of this study.

Predictive ability and model validation

For the validation of our predictions, we employed leave-one-out cross-validation (LOOCV), which is free of bias unless the number of observations or features is extremely limited (http://not2hastie.tumblr.com/). In the case of the diversity panel maize inbred lines, LOOCV was performed by using a single genotype as a hold-out sample, which was predicted by using all other 164 minus 1 inbred lines for model training. This process was repeated until all 164 inbred lines had been used once for testing and 163 times for model training. For the single-step prediction models, imputation was performed on the transcriptomic data.

For the hybrid data, LOOCV was carried out as follows: Let D and F denote the set of parental inbred lines from the Dent and the Flint group, respectively. The set H is the subset of 1521 phenotyped hybrids within the set of all possible hybrids $D \times F$, which would comprise $14,626 = 142 \times 103$ hybrids. Let $i \times j$ denote the hybrid to be predicted. We can then define the parents of the training set hybrids as all lines except for the parents of $i \times j$, i.e., $D_i = D \setminus \{i\}$ and $F_j = F \setminus \{j\}$, respectively. The training set is then defined as $H_{TRN}(i \times j) = H \cap (D_i \times F_j)$. Note that H_{TRN} corresponds to T0 hybrids in previous publications (Technow et al. 2014; Kadam et al. 2016; Westhues et al. 2017; Schrag et al. 2018).

Single-step prediction was performed either with (i) transcriptomic data as predictor and genomic or pedigree data as imputor or (ii) genomic data as predictor and pedigree data as imputor.

We judged the performance of each model by looking at its predictive ability, which was calculated as the correlation $\rho(\mathbf{y}, \hat{\mathbf{y}})$, where $\hat{\mathbf{y}}$ is the vector of predicted values for all genotypes. Standard errors for the predictive abilities were calculated by bootstrapping the set of predicted and observed values 1000 times using the *R*-package *boot* (Canty and Ripley 2017).

Core sampling

To address objective (ii), we generated core sets of nine different sizes from Exp1 (Fig. 1). The base of each of these core sets was the set of 110 genotypes that were covered by both genomic and transcriptomic data. A core set was created by declaring a pre-specified fraction $x = 10\%, 20\%, \ldots, 90\%$ of genotypes as lacking transcriptomic information. For each value of x, we created 100 core sets $s \in \{1, 2, 3, \ldots, 100\}$ by repeatedly sampling genotypes at random, whose transcriptomic information was declared as missing. The artificially removed transcriptomic information on these genotypes was then imputed. For each core set, predictive abilities were again computed as described above, using LOOCV and bootstrapping. Comparisons of predictive abilities were made among the core sets, with different fractions of

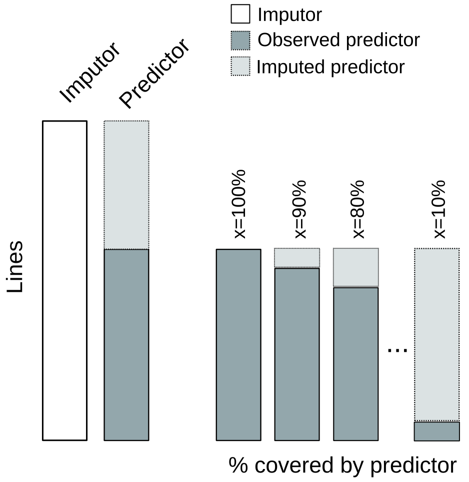

Fig. 1 Scheme for the proportion of genotypes covered by predictors. The two extended bars represent the entire set of genotypes in either Exp1 or Exp2. All genotypes are covered by the imputor (white bar), whereas the predictor (dark gray bar) may cover only a subset of all genotypes. The second bar shows that the fraction of genotypes not covered by the predictor will be imputed (light gray fraction). The four bars to the right outline the principle of generating core sets for Exp1. For nine different values of x, which represent the fraction of genotypes covered by the predictor, core sets are build

incomplete data, as well as between core sets and the full set of genotypes for which features were available.

Data availability

The code and data for the analyses can be accessed from https://github.com/mwesthues/single_step and https://github.com/mwesthues/sspredr. Data from Exp1 were previously published and are linked to in the GitHub repository at https://github.com/mwesthues/single_step#data-preparation.

Results

Genomic information was highly useful for the prediction of all traits, regardless of the experiment. In line with previous studies, predictive abilities achieved with transcriptomic information were as high or higher than those obtained with genomic information when considering yield-related traits or protein (Guo et al. 2016; Westhues et al. 2017; Schrag et al. 2018). Pedigree information displayed competitive predictive abilities for the trait dry matter content and related traits such as 'starch' and 'sugar.'

Experiment 1

At the level of the 110 genotypes for which both genomic (*G*) and transcriptomic (*T*) information was available, the use of *T* yielded higher predictive abilities than the use of *G* for two out of the six traits (Fig. 2). In terms of predictive ability, the absolute advantage of *T* over *G* was as large as 0.15, observed for the trait ear diameter. The predictive ability for predictions based on *G* was always higher for the full set of inbred lines ($n = 164$) than for the reduced set of inbred lines ($n = 110$). Imputing *T* using *G*, denoted as $T^*|G$, for the prediction of all 164 lines, performed at least as well as when predicting using only *G* if *T* was superior over *G*. In the case of kernel width and ear diameter, the use of $T^*|G$ yielded a slight improvement in predictive ability over *G*. For the other four traits, where *T* was a worse predictor for the reduced set of 110 lines than *G*, predictive abilities of $T^*|G$ for the full set of 164 lines were also worse than those based on *G* alone.

Experiment 2

In the 'Reduced' set of Exp2, transcriptomic data were the best single predictor for the traits lignin, protein and particularly for dry matter yield, where the predictive ability

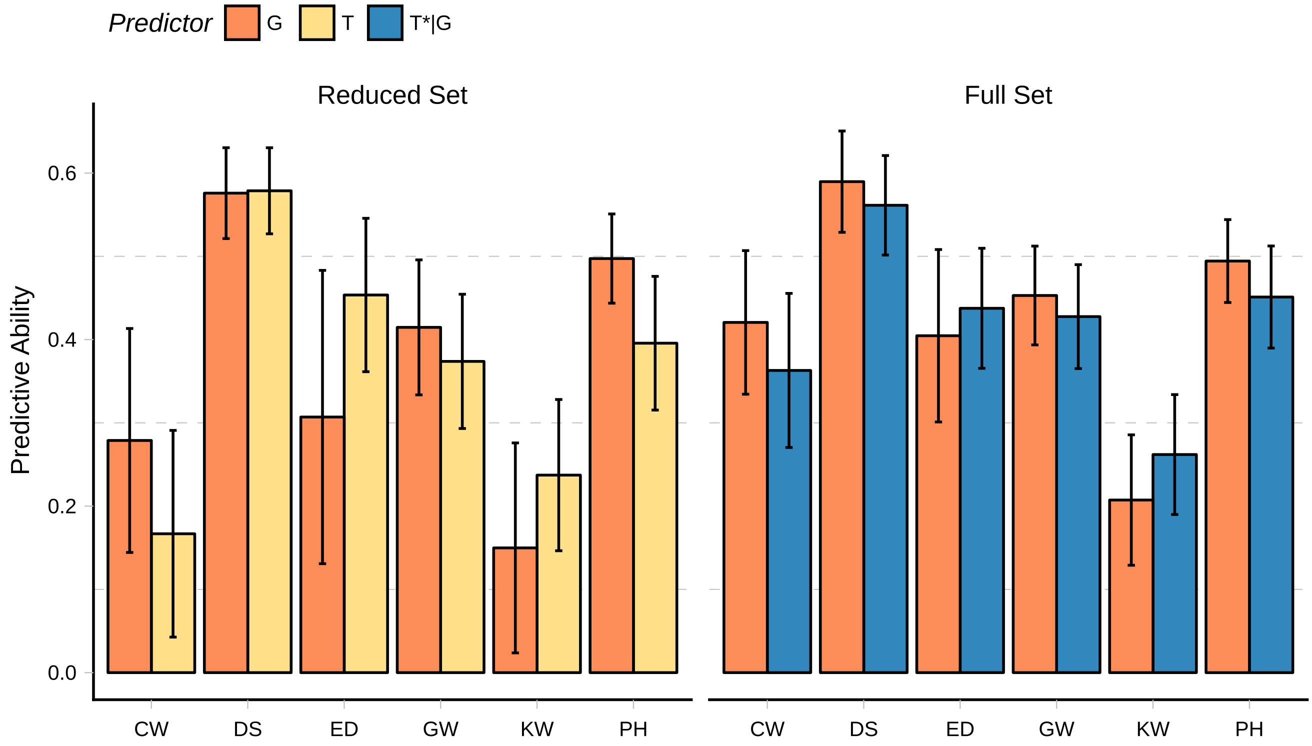

Fig. 2 Predictive abilities and bootstrapped standard errors for tropical/subtropical inbred lines from the maize diversity panel (Exp1) and six agronomic traits. As predictors, genomic (*G*) data, transcriptomic (*T*) data and their combination ($T^*|G$) were used. The 'Reduced Set' includes a subset of 110 lines covered by genomic and transcriptomic information, whereas the 'Full Set' comprises all 164 genotypes. In the 'Full Set' of genotypes, transcriptomic records were imputed for 54 lines

obtained with T was 0.08 points higher than for G as the second best predictor for this trait (Fig. 3). In the reduced set, P had the highest predictive ability for starch and dry matter content. For fat and sugar, G was the superior predictor in the reduced set. In the full set of 1521 hybrids, predictive abilities obtained with G increased markedly for all traits with the exception of fat. The single-step combination of genomic and transcriptomic information ($T^*|G$) yielded a small improvement over G alone for dry matter yield and protein. Compared to using only G, trait prediction did not benefit from the single-step approach when T was not superior to G in the reduced set. The only exception was protein. Predictive abilities resulting from combinations of pedigree with genomic ($G^*|P$) information were higher than P alone for all traits, and $T^*|P$ improved upon P in the 'full' set for four traits.

Impact of coverage by the incomplete predictor

Predictive abilities for the six traits in the different core sets of Exp1 varied widely within most traits (Fig. 4). For the majority of traits, an initial increase in predictive abilities was observed when moving from a core set size of 10% to 20%/30%, regardless of whether G or T individually was the superior predictor. For the traits cob weight, grain weight and plant height, which were predicted with higher precision via G than via T in the reduced set of 110 lines, predictive abilities were higher when only a small number of genotypes was included in the core set. This corresponds to a situation where a large number of genotypes is being imputed and the majority of the information is derived from genomic information. A hike in predictive abilities, when moving from a core fraction of 10% to 90%, was observed for ear diameter and kernel width, which were predicted with greater precision when using T individually compared to using G individually. The only trait that was barely affected by a change in the core fraction was days until silking for which the predictive abilities using either G or T individually were almost identical (Fig. 2).

Discussion

Addition of new predictors Pedigree records are the most ubiquitous predictor in animal and plant breeding programs, because they are easy and inexpensive to collect. Compared to pedigree information, which represent the expected relationship among individuals, genomic information captures Mendelian sampling as well as the effects of selection and thereby provides an improved proxy of the realized

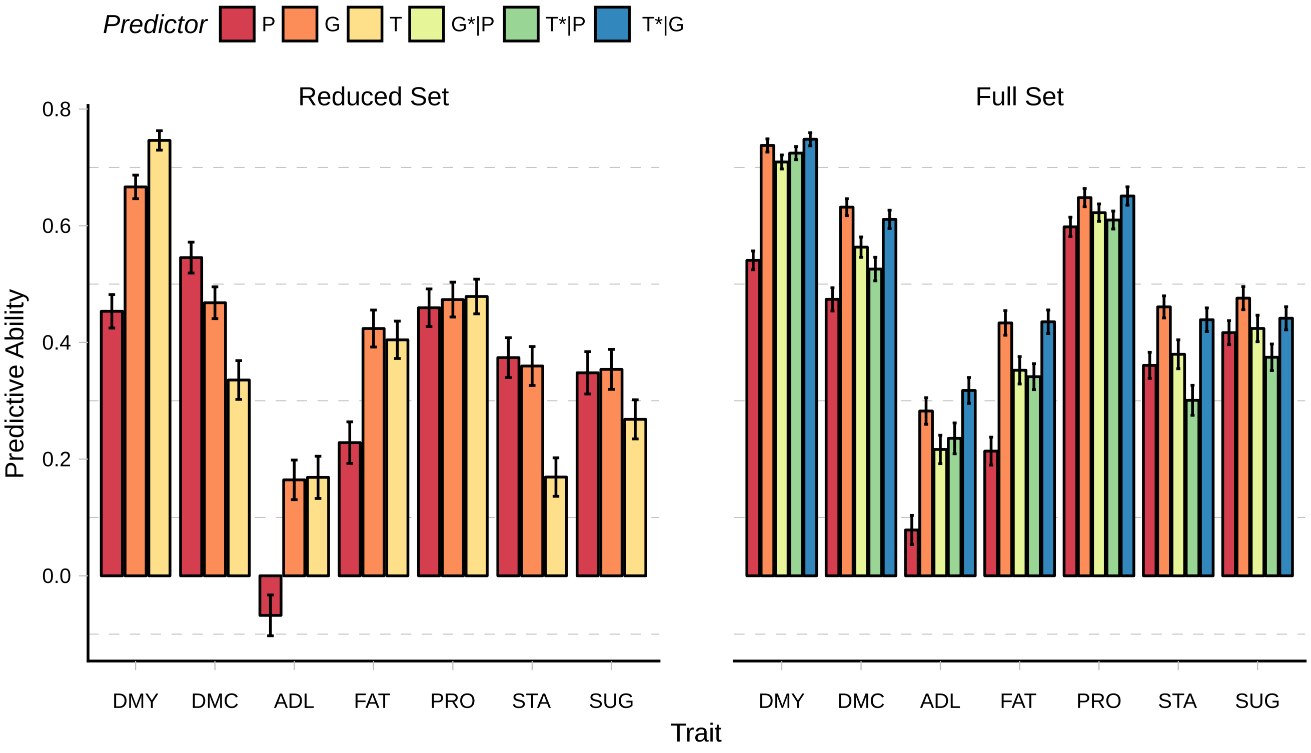

Fig. 3 Predictive abilities and bootstrapped standard errors for the set of maize hybrids (Exp2) and seven agronomic traits. As predictors, pedigree (*P*), genomic (*G*), transcriptomic (*T*) data and combinations thereof were used. The 'Full Set' of genotypes comprises 1521 hybrids based on 245 parental inbred lines, whereas the 'Reduced Set' comprises 685 hybrids based on 103 parental inbred lines

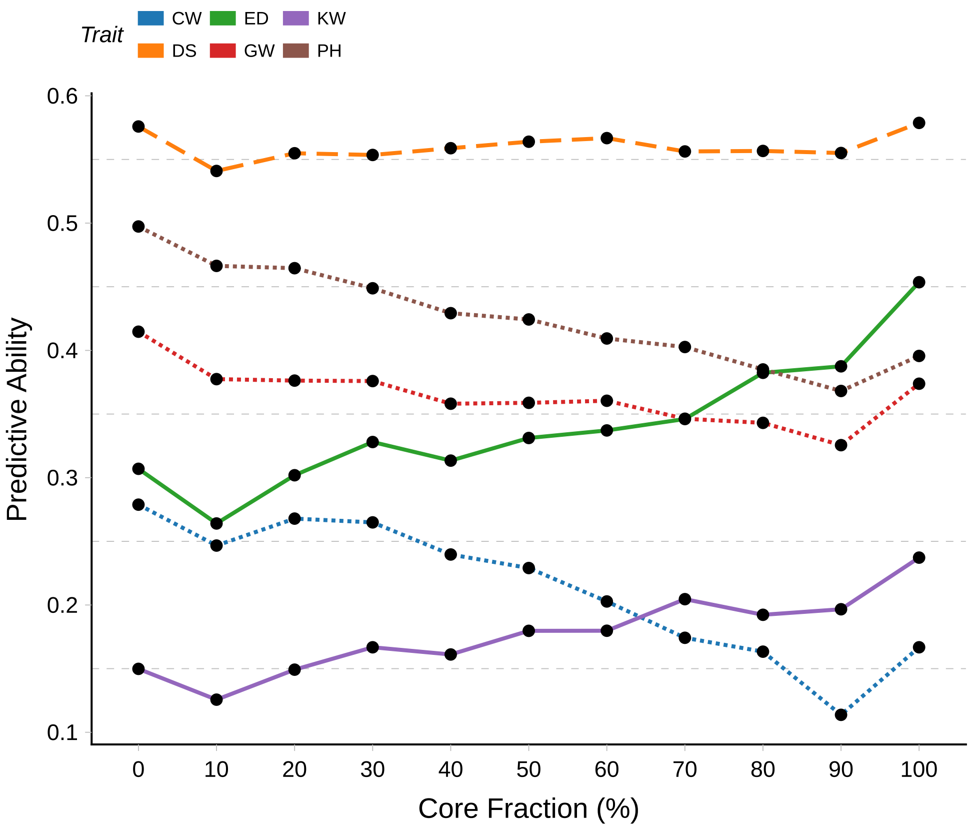

Fig. 4 Predictive ability under a single-step prediction model as a function of the fraction (x) of genotypes covered by the predictor T. Results are shown for the core set of 110 inbred lines from Exp1 for which both G and T are available for any genotype and six agronomic traits. The predictive ability at a core fraction of 0% is based only on G, and at a core fraction of 100%, it is based only on T. Solid lines indicate that $T > G$, dashed lines indicate that $T \approx G$ and dotted lines indicate that $G > T$ when running predictions using single predictors on the same dataset

relationship among individuals. Nevertheless, genomic data do not exhaustively capture physiological epistasis (Jiang and Reif 2015; Guo et al. 2016; Vazquez et al. 2016) attributable to interactions within and between different biological strata (Sackton and Hartl 2016). Such interactions were found to be pervasive throughout the genomes of yeast (Brem et al. 2005) and humans (Brown et al. 2014) and have motivated studies on the utility of downstream 'omics' predictors for integrating such interactions. Recently, encouraging results with regard to predicting complex traits using 'omics' data were found for humans (Vazquez et al. 2016), maize inbred lines (Guo et al. 2016) and maize hybrids (Westhues et al. 2017; Schrag et al. 2018).

Applicability of the single-step framework

Several studies described the usefulness of gene expression data for the prediction of complex traits but higher costs compared to genetic markers have so far limited increased usage of this valuable data source. A pioneering study by Gamazon et al. (2015) showed a way for dealing with the problem of missing data in 'omics' features by imputing transcriptome information for some individuals using other individuals having both genetic marker and gene expression data as a training set. Unfortunately, their approach cannot account for the incurred imputation error. Single-step prediction has become an established method for addressing this shortcoming when complete pedigree but only incomplete genomic records are available. Its theoretical framework has been developed with the premise of the imputed features having heritabilities equal to one. With gene expression data, this condition may not necessarily hold true but it should be noted that all gene expression data used in our study were sampled under highly standardized conditions and at a very early development stage to ensure negligible experimental error. Likewise, genomic data have limitations in that it can be affected by ascertainment bias, sampling errors and imputation errors (Pérez-Enciso et al. 2015).

Prospects of using incomplete gene expression data

The great prospects of imputing transcriptomic via genomic information in a single-step approach arise from two factors: (i) high correlations between the off-diagonal elements of the kernels for G and T, suggesting that genomic and transcriptomic information largely overlaps (Westhues et al. 2017) and (ii) transcripts potentially improving predictive ability by integrating abundant interactions between distal SNPs and transcripts as observed in hybrid maize (Westhues et al. 2017). Imputing missing transcriptomic information via genomic information in a single-step approach can thus be regarded a promising method for cost-effective improvements in predictive ability and to handle missing data in genetic evaluations. Here, we considered single-step approaches for a maize inbred line diversity panel as well as for a collection of maize hybrids using pedigree, genomic and transcriptomic data as predictors. Unlike genomic markers, quantitative 'omics' predictors will not have repeatabilities equal to one, and thus, negatively affect the precision of the imputation when using them as imputors, which was not considered here.

Superiority of single-step models in line prediction When using a single-step approach based on complete genomic and incomplete transcriptomic information, gains in predictive ability for the inbred lines were at best just slightly higher than when using complete genomic information alone. One of the reasons might be that the subset of genotypes that was covered by transcriptomic data was not a good representation of the full genetic space of all available genotypes. A larger dataset was available to Ashraf et al. (2016), who imputed genotypes for about 10,000 wheat lines using pedigree information and observed greater prediction accuracy of the single-step method over genomic BLUP for all four evaluated traits. Predictive abilities reported for the maize diversity panel in this study were slightly different from those reported by Guo et al. (2016), for three possible reasons: First, we reduced the dataset to 164 tropical/subtropical lines with little evident population structure and excluded all other genotypes from our analyses. In contrast, Guo et al. (2016) used genotypes from four different subpopulations with a model using a fixed effect to account for the differences between subpopulations. Second, we applied quality checks for the predictor data after generating the subset of 164 inbred lines, whereas Guo et al. (2016) applied the quality checks to 368 genotypes. Third, while Guo et al. (2016) used fivefold cross-validation with 500 repetitions, we employed a leave-one-out cross-validation scheme (LOOCV). Notwithstanding, the relative differences between predictive abilities for G and T, respectively, were similar in both studies.

Superiority of single-step models in hybrid prediction Hybrid breeding is a particularly challenging field for prediction tools (Kadam et al. 2016). Here, $2n$ parent individuals from two genetically distinct heterotic groups are crossed to each other, yielding n^2 potential hybrid progeny that would require intensive field testing. In medium-sized plant breeding programs, the advent of the doubled-haploid (DH) technology (Wedzony et al. 2009) allows for an annual production of thousands of parent lines in each heterotic group, amounting to millions of putative hybrid progeny. Westhues et al. (2017) showed that in hybrid breeding, the probability of successfully selecting the best observed genotypes based on the best predicted candidates is a strongly convex function of the predictive ability. Thus, even minor gains, when using transcriptomic and genomic data in a single step, might justify additional investments in RNA-seq at least for a subset of genotypes that covers the genetic space of the breeding material well. For the hybrid data, the predictive abilities of G increased considerably in the full set of genotypes compared to the reduced set for several agronomic traits. Hence, the inclusion of transcriptomic data could not further improve upon genomic information alone. Predictive abilities obtained when using pedigree data could be improved considerably when, depending on the agronomic trait, combined with either transcriptomic or genomic information. This suggests that single-step prediction should always be considered when the incomplete predictor offers an appreciably higher predictive ability compared to the complete predictor. A previous study using the same data as in Exp2 concluded that modeling of specific combining ability (SCA) effects did not appreciably improve predictive abilities over that of a purely additive model for any of the investigated traits (Westhues et al. 2017). Nevertheless, the inclusion of SCA effects in a single-step prediction model could be achieved by computing the element-wise product between the feature matrices $\mathbf{M}_D$ and $\mathbf{M}_F$ of the two corresponding heterotic groups (Martini et al. 2016). While different studies have found that the low SCA/GCA-ratio in hybrid maize breeding programs typically does not warrant its inclusion, this situation might be different in crops where heterotic patterns are not yet clearly defined (Zhao et al. 2015).

Coverage of the genetic space Whereas animal breeding populations are oftentimes so large that the assembly of a core set of individuals with records on multiple predictors can effectively be done at random, provided that more than 10,000 animals have predictor information (Fragomeni et al. 2015; Lourenco et al. 2015; Masuda et al. 2016), population sizes in plant breeding programs are typically much smaller. Single-step prediction offers the promise of leveraging the predictive ability for all genotypes by borrowing the superior performance of an auxiliary predictor available for only a subset of genotypes, thereby reducing predictor costs. Hence, we were interested in determining to what extent individuals should be complemented with information on another predictor. By randomly and repeatedly declaring transcriptomic information missing from genotypes in Exp1, we examined how reliably single-step prediction works depending on the fraction of genotypes covered by both the complete and the incomplete predictor. We observed that about $x = 20 - 30\%$ of genotypes need to be covered by both predictors to match the performance achieved with the complete predictor individually. Beyond this fraction of covered genotypes, the trend in predictive ability changes was stable in either direction. Predictive abilities increased simultaneously with x when the incomplete predictor outperformed the complete predictor in the reduced set of genotypes and decreased with increasing x when the complete predictor was the better predictor. This pattern clearly elucidates the influence of the difference between the two predictors on predictive ability in a single-step approach. Further research is warranted to investigate how the minimum value for *x* depends on the sample size of genotypes having complete information for the imputor but incomplete information for the predictor.

Conclusions

We successfully applied single-step prediction to inbred and hybrid datasets while imputing with a quantitative downstream 'omics' predictor. By declaring different subsets of individuals as covered by one or two predictors, respectively, we could elucidate the influence of differences between two predictors on predictive ability when using a single-step approach. When the discrepancy in predictive ability between the complete and the incomplete predictor was skewed profoundly in favor of the incomplete predictor, such as for pedigree and transcriptomic data, single-step prediction was shown to be highly beneficial. Extensions to more than two predictors were outside the scope of this study but, with mounting interest in systems genetics, should be considered in the future.

Acknowledgements We thank T. A. Schrag from the University of Hohenheim for providing the phenotypic data as well as S. Scholten, A. Thiemann and F. Seifert from the University of Hamburg for providing the gene expression data for Experiment 2, respectively. Furthermore, we would like to thank researchers and institutions who contributed to the development of the maize diversity panel and associated data from Experiment 1, in particular Jianbing Yan and Haijun Liu from Huazhong Agricultural University Wuhan in China. We thank T. A. Schrag and W. Molenaar for valuable suggestions for improving the content of this manuscript. The authors acknowledge support by the state of Baden-Württemberg through bwHPC. Financial support for M. W. was provided by the Fiat Panis foundation, Ulm, Germany.

Author contribution statement MW and CH conceived the study. AEM, RF and GT guided the structure of the research and checked the methodology and results for validity. MW and CH drafted the manuscript. MW and CH implemented the prediction models and developed software. MW analyzed the data. All authors interpreted the results, read and approved the final version of the manuscript.

Compliance with ethical standards

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

References

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

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

Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436(7051):701–3. https://doi.org/10.1038/nature03865

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

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84(2):210–223. https://doi.org/10.1016/j.ajhg.2009.01.005

de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet 9(7):1–15. https://doi.org/10.1371/journal.pgen.1003608

Canty A, Ripley BD (2017) Boot: bootstrap R (S-Plus) function

Christensen OF (2012) Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet Sel Evol 44:37. https://doi.org/10.1186/1297-9686-44-37

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

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

Dey KK, Hsiao CJ, Stephens M (2016) Clustering RNA-seq expression data using grade of membership models. https://doi.org/10.1101/051631

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

Fragomeni BO, Lourenco DAL, Tsuruta S, Masuda Y, Aguilar I, Legarra A, Lawlor TJ, Misztal I (2015) Hot topic: use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. J Dairy Sci 98(6):4090–4094. https://doi.org/10.3168/jds.2014-9125

Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, Zhang J, He C, Du X, Peng Z, Wang B, Zhai L, Dai C, Xu J, Wang W, Li X, Zheng J, Chen L, Luo L, Liu J, Qian X, Yan J, Wang J, Wang G (2013) RNA sequencing reveals the complex regulatory network in the maize kernel. Nat Commun 4:2832. https://doi.org/10.1038/ncomms3832

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

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent P, Rose M, Schön CC, Sun Q, Walter H, Martin OC, Falque M (2011) A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PloS ONE 6(12):e28-334. https://doi.org/10.1371/journal.pone.0028334

García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Tassell CPV (2016) Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci USA 113(33):201519,061. https://doi.org/10.1073/PNAS.1519061113

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

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

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–97. https://doi.org/10.1534/genetics.107.081190

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

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

Kadam D, Potts S, Bohn MO, Lipka AE, Lorenz A (2016) Genomic prediction of hybrid combinations in the early stages of a maize hybrid breeding pipeline. G3 6:3443–3453. https://doi.org/10.1101/054015

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

Legarra A, Christensen OF, Aguilar I, Misztal I (2014) Single step, a general approach for genomic selection. Livestock Sci 166(1):54–65. https://doi.org/10.1016/j.livsci.2014.04.029

Lourenco DAL, Tsuruta S, Fragomeni B, Masuda Y, Aguilar I, Legarra A, Bertrand J, Amen T, Wang L, Moser D, Misztal I (2015) Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J Anim Sci 93:2653–2662. https://doi.org/10.2527/jas2014-8836

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

Masuda Y, Misztal I, Tsuruta S, Legarra A, Aguilar I, Lourenco D, Fragomeni B, Lawlor T (2016) Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. J Dairy Sci 99(3):1968–1974. https://doi.org/10.3168/jds.2015-10540

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

Mrode RA (2014) Linear Models for the Prediction of Animal Breeding Values, 3rd edn. CABI, Oxfordshire, https://doi.org/10.1017/CBO9781107415324.004

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

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

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959. https://doi.org/10.1111/j.1471-8286.2007.01758.x

Ratcliffe B, Gamal El-Dien O, Cappa EP, Porth I, Klapste J, Chen C, El-Kassaby Y (2017) Single-step BLUP with varying genotyping effort in open-pollinated picea glauca. G3 7:935–942. https://doi.org/10.1534/g3.116.037895

Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK (2007) A comparison of background correction methods for two-colour microarrays. Bioinformatics 23(20):2700–2707. https://doi.org/10.1093/bioinformatics/btm412

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43(7):e47. https://doi.org/10.1093/nar/gkv007

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge

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

Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE (2018) Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. Genetics https://doi.org/10.1534/genetics.117.300374

Smyth GK, Speed T (2003) Normalization of cDNA microarray data. Methods 31(4):265–273. https://doi.org/10.1016/S1046-2023(03)00155-5

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

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

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423. https://doi.org/10.3168/jds.2007-0980

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

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

Vitezica ZG, Aguilar I, Misztal I, Legarra A (2011) Bias in genomic predictions for populations under selection. Genetics Res 93(5):357–66. https://doi.org/10.1017/S001667231100022X

Watson A, Ghosh S, Williams MJ, Cuddy W, Simmonds J, Rey MD, Md Hatta MA, Hinchliffe A, Steed A, Reynolds D, Adamski N, Breakspear A, Korolev A, Rayner T, Dixon LE, Riaz A, Martin W, Ryan M, Edwards D, Hickey L (2018) Speed breeding is a powerful tool to accelerate crop research and breeding. Nat Plants 4:23–29

Wedzony M, Forster B, Zur I, Golemiec E, Scechynska-Hebda M, Dubas E, Gotebiowska G (2009) Progress in doubled haploid technology in higher plants. In: Touarev A, Forster BP, Mohan JS (eds) Advances in haploid production in higher plants, chap 1. Springer, New York

Westhues M, Schrag TA, Heuer C, Utz HF, Schipprack W, Seifert F, Ehret A, Schlereth A, Stitt M, Nikoloski Z, Willmitzer L, Schön CC, Melchinger AE (2017) Omics-based hybrid prediction in maize. Theor Appl Genet 130:1927–1939. https://doi.org/10.1101/134668

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink JL, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, Melchinger

AE (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 (Bethesda, Md) 2(11):1427–36. https://doi.org/10.1534/g3.112.003699

Xiang T, Nielsen B, Su G, Legarra A, Christensen OF (2016) Application of single-step genomic evaluation for crossbred performance in pig. J Anim Sci 94(3):936–948. https://doi.org/10.2527/jas2015-9930

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

Yang N, Lu Y, Yang X, Huang J, Zhou Y, Ali F, Wen W, Liu J, Li J, Yan J (2014) Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. PLoS Genet 10(9):1–2. https://doi.org/10.1371/journal.pgen.1004573

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

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