BMC Genomics

RESEARCH ARTICLE Open Access

Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines

Christian Riedelsheimer, Frank Technow and Albrecht E Melchinger*

Abstract

Background: There is increasing empirical evidence that whole-genome prediction (WGP) is a powerful tool for predicting line and hybrid performance in maize. However, there is a lack of knowledge about the sensitivity of WGP models towards the genetic architecture of the trait. Whereas previous studies exclusively focused on highly polygenic traits, important agronomic traits such as disease resistances, nutrifunctional or climate adaptational traits have a genetic architecture which is either much less complex or unknown. For such cases, information about model robustness and guidelines for model selection are lacking. Here, we compared five WGP models with different assumptions about the distribution of the underlying genetic effects. As contrasting model traits, we chose three highly polygenic agronomic traits and three metabolites each with a major QTL explaining 22 to 30% of the genetic variance in a panel of 289 diverse maize inbred lines genotyped with 56,110 SNPs.

Results: We found the five WGP models to be remarkable robust towards trait architecture with the largest differences in prediction accuracies ranging between 0.05 and 0.14 for the same trait, most likely as the result of the high level of linkage disequilibrium prevailing in elite maize germplasm. Whereas RR-BLUP performed best for the agronomic traits, it was inferior to LASSO or elastic net for the three metabolites. We found the approach of genome partitioning of genetic variance, first applied in human genetics, as useful in guiding the breeder which model to choose, if prior knowledge of the trait architecture is lacking.

Conclusions: Our results suggest that in diverse germplasm of elite maize inbred lines with a high level of LD, WGP models differ only slightly in their accuracies, irrespective of the number and effects of QTL found in previous linkage or association mapping studies. However, small gains in prediction accuracies can be achieved if the WGP model is selected according to the genetic architecture of the trait. If the trait architecture is unknown *e.g.* for novel traits which only recently received attention in breeding, we suggest to inspect the distribution of the genetic variance explained by each chromosome for guiding model selection in WGP.

Keywords: Genomic selection, Whole-genome prediction, Genetic architecture, Complex traits, *Zea mays*

Background

Whole-genome prediction (WGP) is expected to reshape plant breeding fundametally in the near future [1-3]. Whereas the approach has been initially proposed [4] and rapidly implemented in animal breeding [5], recent empirical studies demonstrated also its potential in hybrid maize breeding [6-8]. Recently, we showed that WGP allows a reliable screening of large germplasm collections of diverse maize inbred lines for their potential to create superior hybrids [9]. However, these studies exclusively focused on predicting highly polygenic traits such as grain yield or biomass accumulation with genetic architectures close to the infinitesimal genetic model [10].

In maize, several economically important traits are genetically less complex with few quantitative trait loci

*Correspondence: melchinger@uni-hohenheim.de
Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany

(QTL) explaining a large proportion of the genetic variance. Examples include pest and disease resistances or nutrifunctional compounds such as bioavailable minerals [11] or β-carotene [12]. In addition, disease resistances are often found to be controlled by a combination of race-specific resistance loci with large effects involved in pathogen recognition, and a large number of loci with small effects involved in basal resistance. Such a mixed QTL effect distribution can be found in maize *e.g.* for rust [13], Giberella ear rot [14,15] or to a lesser extent for Northern corn leaf blight [16,17].

For such traits, the assumption of normally distributed SNP effects underlying ridge regression, the most commonly applied WGP model, is severely violated. Heslot *et al.* [18] found for polygenic traits in several plant species only minor differences between ridge regression and models with different assumptions of the underlying distribution of SNP effects. However, these differences are expected to be much larger for traits controlled by only a few QTL. Recently, Clark *et al.* [19] simulated this situation under the assumption of the historical population structure of Holstein cattle. They found that under the assumption of either few common or few rare quantitative trait loci, a Bayesian variable selection model (BayesB) outperforms ridge regression by far. For Holstein cattle, Hayes *et al.* [20] found also the BayesA model to be superior to ridge regression in the case of coat color or milk-fat percentage.

Cattle differs greatly in its population structure and LD level from elite maize germplasm, which has faced severe genetic bottlenecks during domestication and the creation of genetically distinct heterotic pools to maximize exploitation of heterosis in hybrid breeding [21,22]. Hence, results from cattle might not be directly transferable to maize, for which little is known about WGP for traits with a simpler genetic architecture. Moreover, the genetic architecture of a trait is often unclear in crops. Especially if the trait has not yet been extensively dissected by linkage or association mapping, which might be the case for traits which gained only recently in importance such as nutritional properties, nutrient acquisition traits or traits related to climate change adaptation.

To fill this apparent gap of knowledge, we investigated WGP in a diverse collection of 289 maize inbred lines with traits which largely vary in their genetic architecture. To let the genetic basis differ as much as possible, we chose three highly polgygenic agronomic traits and three metabolites, each one controlled by a different major QTL explaining about 22% to 30% of genetic variance. With this empirical set-up, we asked the following questions:

- To what extent do distinct WGP models differ in their prediction accuracies for a diversity panel of maize inbred lines if the genetic architecture of the trait changes dramatically?
- Are there guidelines for plants concerning the choice of the most promising WGP model?

Methods

Genetic material

The genetic material consisted of 289 maize inbred lines which were previously described in great detail [9,23-25]. The population constituted a global sample of elite breeding material from worldwide sources with a focus on North America and Europe and encompassed 285 lines from the Dent heterotic pool (Stiff-Stalk and non-Stiff-Stalk) and 4 from the European Flint pool, which served as check genotypes.

Genotyping

The population was genotyped with the Illumina SNP chip MaizeSNP50 containing 56,110 SNPs [26]. Quality control preprocessing of SNPs was performed by eliminating SNPs that did not match the following criteria: (i) less than 10% missing values, (ii) minor allele frequency of greater than 2.5%, (iii) no more than three heterozygous genotypes, and (iv) unique allele assignment for the 22 replicated checks of genotype B73. A total of 38,019 SNPs remained and were used for further analysis. Linkage disequilibrium (LD) declined to $r^2 = 0.1$ at approximately 500 kb with a mean LD between adjacent SNPs of 0.34 [9].

Field trials

The population was phenotyped in six environments (three agroecologically diverse locations in the years 2008 and 2009) in Germany [25]. Briefly, the population was split into three maturity groups based on prior knowledge of their flowering time. In the trials of each of the three maturity groups, 100 genotypes, including five common check genotypes, were randomized in a 20×5 α-lattice design with two replications and were planted in 2-row plots. Plots were thinned to a final plant density of 100,000 plants/ha. The common check genotypes were used to adjust for potential differences in the soil fertility among trials in each environment.

Metabolites

Leaf samples were collected in one location 33 d after sowing and processed using an established GC-MS method [27]. Genotypic means of Box-Cox transformed metabolite concentrations were obtained using a linear mixed model analysis including effects for field trial, replication, block, and batch. The whole metabolic profiling procedure including statistical analysis has been described in detail previously [9,23]. From the measured metabolite concentrations, we chose three highly heritable substances as model traits: dopamine, ribitol, and

an unknown metabolite (719700-204). For each metabolite, we found in a genome-wide association (GWA) study a major metabolite QTL (mQTL) on different chromosomes after correcting for population structure and kinship [23]. For dopamine, the major mQTL was found on chromosome 9 and explained 28.9% of the genetic variance. For ribitol, the major mQTL was found on chromosome 10 and explained 22.1% oft the genetic variance. For the unknown metabolite, the major mQTL was found on chromosome 2 and explained 29.8% of the genetic variance. The metabolites were uncorrelated with each other ($|r| \leq 0.10$) and only weakly ($|r| \leq 0.28$) correlated with agronomic traits (Table 1).

Agronomic traits

Dry matter yield of whole-plant biomass (t/ha) and plant height (m) were measured per field plot of the inbred lines. Lignin content was measured as acid detergent lignin (ADL) in the harvested plant material of the inbred lines using calibrated near-infrared spectroscopy (NIRS). The NIRS calibration model was built using phenotypic data from 20 inbred lines, 32 testcrosses and 3 hybrids grown in the same environments as the population of inbred lines analyzed in this study [24]. Heritability estimates and genotypic means were obtained using a one-step linear mixed model analysis as described previously [25]. Using a 1% Bonferroni corrected significance threshold, we could not find any significant SNP-trait association signal using the the same GWA model as for metabolites. Since population size, marker density, and heritabilites were sufficiently high for detecting QTL with large effects, the absence of any significant trait-SNP associations suggest a highly polygenic genetic architecture for the agronomic traits with no major QTL.

Genome partitioning of the genetic variance

To further characterize the genetic architectures of the investigated traits irrespective of the significance thresholds for SNP-trait associations, we compared how the ten chromosomes contributed to the total genetic variance. Later on, we will use these results as a guideline for model selection based on trait architecture.

We adopted the approach of Yang *et al.* [28] to simultaneously estimate the genetic variance explained by each chromosome. In order to derive a guideline which is purely based on trait architecture and not on population structure artefacts, we additionally corrected for population structure by regressing the trait values on the first ten principal components. This linear model can be written as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Q}\boldsymbol{\beta} + \sum_{c=1}^{10}(\mathbf{S}\mathbf{g}_c) + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is a vector with n trait values, $\mathbf{1}$ is a vector of 1's, $\mathbf{Q}$ is a matrix of size $n \times 10$ containing the first 10 principal components calculated from SNP data with $\boldsymbol{\beta}$ containing the corresponding regression coefficients, $\mathbf{S}$ is an incidence matrix allocating components of $\mathbf{y}$ to components of $\mathbf{g}_c$, which is a vector of length n with random genotypic effects attributable to chromosome c with $\mathbf{g}_c \sim N(0, \mathbf{G}_c\sigma^2_{gc})$ and $\mathbf{G}_c = \mathbf{Z}_c\mathbf{Z}_c^T/p_c$ where $\mathbf{Z}_c$ is a matrix of size $n \times p_c$ with standardized levels of SNP alleles on chromosome c. Vector $\mathbf{e}$ contains normally distributed residuals with $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$. The genetic variance contributed by chromosome c was then estimated as $\sigma^2_{gc}/(\sum_{c=1}^{10}(\sigma^2_{gc}) + \sigma^2_e)$.

Variance components were estimated by restricted maximum likelihood (REML) using ASReml-R 3 [29]. Since matrices $\mathbf{G}_c$ were often found to be singular, we used the algorithm of Higham [30] implemented in the function nearPD of the R-package Matrix [31], to approximate the nearest positive definite matrices.

WGP models

We investigated five WGP models that have been recently advocated in the literature for this purpose [4,32-34].

All based upon the classical regression set-up

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (2)$$

where $\mathbf{y}$ is a vector with n trait values, μ is the overall mean, $\mathbf{1}$ is a vector with 1's, $\mathbf{Z}$ is the $n \times p$ matrix of standardized values of SNP alleles, $\mathbf{u}$ is a vector with SNP effects, and $\mathbf{e}$ is a vector of residuals with $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$. Depending on the trait, a combination of genotypic and phenotypic information was available for 276 to 280 genotypes which were used for WGP.

Table 1 Phenotypic correlations among traits

-	Plant height	Lignin content	Dopamine	Ribitol	719700-204
Dry matter yield	0.62	0.32	-0.28	0.02	-0.24
Plant height	-	0.50	-0.17	0.07	-0.12
Lignin content	-	-	-0.20	-0.11	-0.08
Dopamine	-	-	-	-0.10	0.08
Ribitol	-	-	-	-	-0.03

RR-BLUP

Ridge regression (RR) tackles the $p \gg n$ problem in WGP by minimizing the residual sum of squares (RSS $= (\mathbf{y} - \mathbf{1}\mu - \mathbf{Zu})^T(\mathbf{y} - \mathbf{1}\mu - \mathbf{Zu})$) by bounding the Euclidean (L_2) norm of $\mathbf{u}$ to a constraint: $||\mathbf{u}||_2 = \sqrt{\sum_{i=1}^{p} u_i^2} < c_{\mathrm{RR}}$ which leads to a homogeneous shrinkage of all SNP effects towards zero (Figure 1). The RR estimator is given by

$$\hat{\mathbf{u}}_{\mathrm{RR}} = \arg\min_{\mathbf{u}}(\mathrm{RSS} + \lambda_{\mathrm{RR}}||\mathbf{u}||_2). \quad (3)$$

The Lagrangian multiplier λ_{RR} is a regularization parameter which controls the amount of shrinkage. It can be estimated as $\lambda_{\mathrm{RR}} = \sigma_e^2/\sigma_u^2$ by regarding $\mathbf{u}$ as random effects with $\mathbf{u} \sim N(0, \mathbf{I}\sigma_u^2)$ with σ_u^2 being the SNP effect variance estimated by REML. In this setting, $\hat{\mathbf{u}}_{\mathrm{RR}}$ is equivalent to the best linear unbiased predictor (BLUP) of $\mathbf{u}$ [35,36].

For computational convenience, RR-BLUP can be transformed to a mathematically equivalent model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Sg} + \mathbf{e} \quad (4)$$

with $\mathbf{g}$ being a vector of random genotype effects with $\mathrm{var}(\mathbf{g}) = \mathbf{G}\sigma_g^2$ and whole-genome relationship matrix $\mathbf{G} = \mathbf{ZZ}^T/p$. The solution vector of SNP effects can then be obtained as $\hat{\mathbf{u}}_{\mathrm{RR}} = \mathbf{Z}^T\mathbf{G}^{-1}\hat{\mathbf{g}}$ [37]. Here, $\mathbf{G}$ is an inner-product kernel which allows to perform all computations in the space of n genotypes instead of p SNPs, a shortcut which is well established in the field of kernel-based machine learning [38].

LASSO

As an alternative to ridge regression, it was suggested to use an L_1 penalty to tackle the $p \gg n$ problem [39]. This estimator was termed least absolute shrinkage and selection operator (LASSO) and has recently been suggested for whole-genome prediction [32,40,41]. The estimator is given by

$$\hat{\mathbf{u}}_{\mathrm{L}} = \arg\min_{\mathbf{u}}(\mathrm{RSS} + \lambda_{\mathrm{L}}||\mathbf{u}||_1) \quad (5)$$

which bounds the Manhattan (L_1) norm of $\mathbf{u}$ to a constraint: $||\mathbf{u}||_1 = \sum_{i=1}^{p} |u_i| < c_{\mathrm{L}}$. The LASSO penalty is a diamond shaped constraint which allows not only to shrink coefficients towards zero but to set some coefficients to exactly zero (Figure 1). Unlike RR, LASSO, cannot be 'kernelized', *i.e.*, it is not possible to transform the LASSO estimator into an equivalent kernel regression problem in the space of n genotypes [38]. Hence, LASSO regression has to be carried out with the full set of SNPs. Here, we used the R package glmnet, a fast implementation using cyclic coordinate descent to compute the complete LASSO path solution [42].

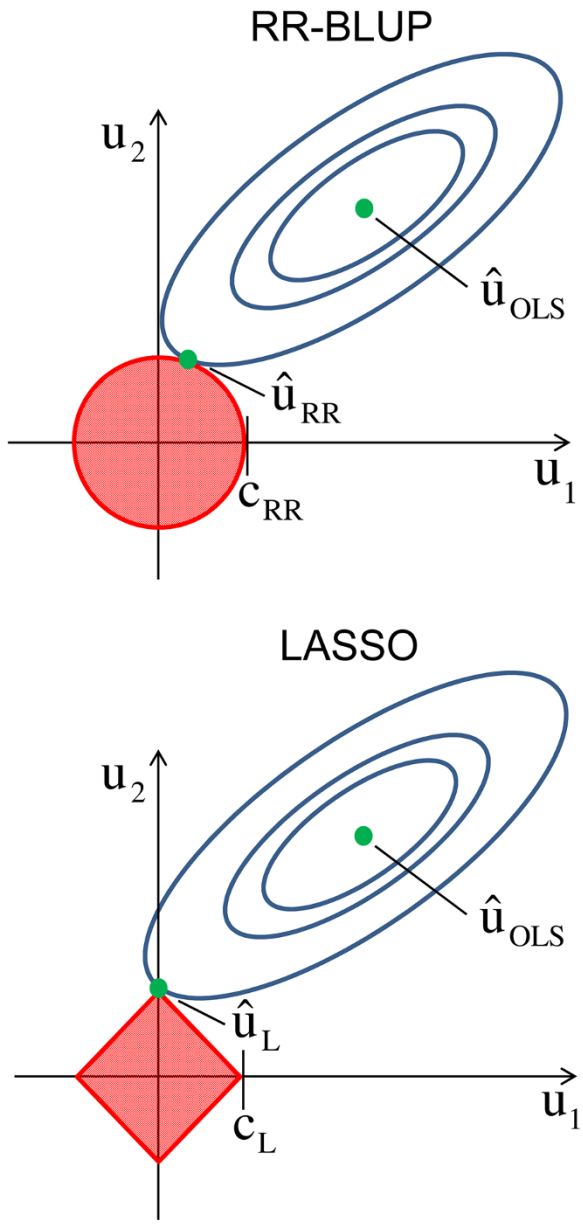

Figure 1 Visualization of the RR-BLUP estimator ($\hat{\mathbf{u}}_{\mathrm{RR}}$) and the LASSO estimator ($\hat{\mathbf{u}}_{\mathrm{L}}$) as solutions to a least-squares problem with different penalization [38,39]. We illustrate a two-dimensional case. The blue ellipses show the contours of the RSS function around the ordinary least-square solution ($\hat{\mathbf{u}}_{\mathrm{OLS}}$). The ridge estimator is the point at which the innermost elliptical contour touches the circular ridge penalty $\sqrt{u_1^2 + u_2^2} < c_{\mathrm{RR}}$. The LASSO estimator is the point at which the innermost elliptical contour touches the diamond shaped LASSO penalty $|u_1| + |u_2| < c_{\mathrm{L}}$. Contrary to the ridge penalty, the LASSO penalty allows estimations to be exactly zero.

Elastic net

The LASSO penalty is known to be somewhat indifferent to the choice among a set of strong but correlated variables. The RR penalty, on the other side, tends to shrink

the coefficients of correlated variables toward each other [38]. The elastic net (EN) estimator is a compromise which can be written as

$$\hat{\mathbf{u}}_{\mathrm{EN}} = \arg\min_{\mathbf{u}}(\mathrm{RSS}+\lambda_{\mathrm{EN}}(\alpha||\mathbf{u}||_1+(1-\alpha)||\mathbf{u}||_2)), \quad (6)$$

and is a weighted mixture between the RR penalty ($\alpha = 0$) and the LASSO penalty ($\alpha = 1$) [33]. While the RR penalty encourages highly correlated variables to be averaged, the LASSO penalty encourages a sparse solution [38]. We again used the implementation in glmnet and performed a grid search to find the combination of α and λ_{EN}, which yielded the lowest mean squared error in the training population.

Reproducing kernel Hilbert space (RKHS) regression

The theory of RKHS regression is rooted in the field of kernel-based machine-learning [38] and has recently been advocated for whole-genome prediction [34]. The approach uses equation 4 but replaces the inner-product matrix **G** with a kernel matrix **K**. The motivation behind RKHS regression lies in the ability to effectively perform non-linear regression in a higher-dimensional feature space so it might capture non-additive genetic effects, if present. Here, we used a Gaussian kernel on genetic distances with $K_{ij} = \exp(-\mathrm{GD}_{ij}/\theta^2)$, where GD_{ij} is the modified Rogers' genetic distance (Euclidean distance scaled to fall between 0 and 1) between genotype i and j, and θ is a smoothing parameter which controls the rate of decay of K_{ij} with increasing genetic distance. The optimum value for θ was chosen from a sequence from 0.1 to 100 at which the maximum likelihood was obtained.

BayesB

As a Bayesian approach, we used a modified version of BayesB, which has a prior assumption that the SNP effects are t-distributed with a point-mass at zero [4]. Following the suggestions of Yang and Tempelman [43], we modeled several hyperparameters as uncertain too. Details of the priors used can be found in Table 2. To fit the model, we ran the Gibbs-sampler for 50,000 iterations. The first 5,000 iterations were discarded as burn-in and only samples from every 10$^{\mathrm{th}}$ post burn-in iteration were stored. For computational convenience, we reduced the number of markers to 5,000 SNPs for which we did not observe any decline in prediction accuracy up to the numerical precision reported in this study.

Validation

A five-fold cross-validation scheme was applied and repeated 20 times. In each repetition, the dataset was divided into 5 disjoint subsets of genotypes whereas one subset served as the validation set and the other four subsets served as the training population to estimate the model parameters for predicting the left-out genotypes in the validation set. In each of the five rounds, the Pearson correlation between the observed and predicted phenotypic values was calculated. The procedure was repeated twenty times to yield 100 cross-validation runs. The predictive ability was then calculated as the Pearson correlation ($r_{(\mathbf{y},\hat{\mathbf{y}})}$) between the observed ($\mathbf{y}$) and predicted ($\hat{\mathbf{y}}$) phenotypic values. The 'prediction accuracy' estimates the correlation ($r_{(\mathbf{g},\hat{\mathbf{g}})}$) between the predicted ($\hat{\mathbf{g}}$) and unobserved true genetic values ($\mathbf{g}$) and was calculated by $r_{(\mathbf{g},\hat{\mathbf{g}})} = r_{(\mathbf{y},\hat{\mathbf{y}})}/h$ where h is the square root of the heritability on a line-mean basis for the agronomic traits. For metabolites, the square root of the estimated repeatability was used.

Table 2 Priors used for BayesB

Parameter	Prior
u_i	$N(0, \sigma^2_{ui})$
$\sigma^2_{ui}\|v_u, S^2_u$	$\begin{cases} 0 & \text{with probability } \pi_u, \\ \chi^{-2}(v_u, S^2_u) & \text{with probability } (1-\pi_u) \end{cases}$
v_u	Gamma($k = 5, \theta = 2$)
S^2_u	Gamma($k = 0.1, \theta = 10$)
π_u	Beta($\alpha = 7, \beta = 3$)
σ^2_e	$\chi^{-2}(v_e, S^2_e = \hat{\sigma^2_e}(v_e - 2)/v_e)$
	$v_e = 4.001$, $\hat{\sigma^2_e}$ estimated with REML

Results

The contribution of the individual chromosomes to the genetic variance differed largely between metabolites and agronomic traits (Figure 2). For the metabolites, the chromosomes containing the major mQTL (chromosome 9 for dopamine, chromosome 10 for ribitol, and chromosome 2 for the unknown metabolite) captured by far the largest portion of the genetic variance, leaving the remaining genetic variance equally distributed over the remaining chromosomes.

For the agronomic traits, the total genetic variance was largely uniformly distributed over all chromosomes. Using the same GWA model as for the metabolites [23], we found that for dry matter yield and plant height, the chromosomes which captured the largest portion of genetic variance contain the strongest GWA signals. However, in no instance was the 1% Bonferroni corrected significance threshold surpassed (dry matter yield: chr. 6, $P = 4.06 \times 10^{-6}$, position 139,284,469, explained genetic variance 8.8%; plant height: chr. 3, $P = 8.6 \times 10^{-6}$, position 163,617,228, explained genetic variance 8.0%).

When excluding chromosomes containing either these two association signals or major mQTL, we observed a tendency that longer chromosomes captured more genetic variance than shorter ones (Figure 2B). This trend was significant ($P < 0.10$) for lignin content ($r = 0.88, P =$

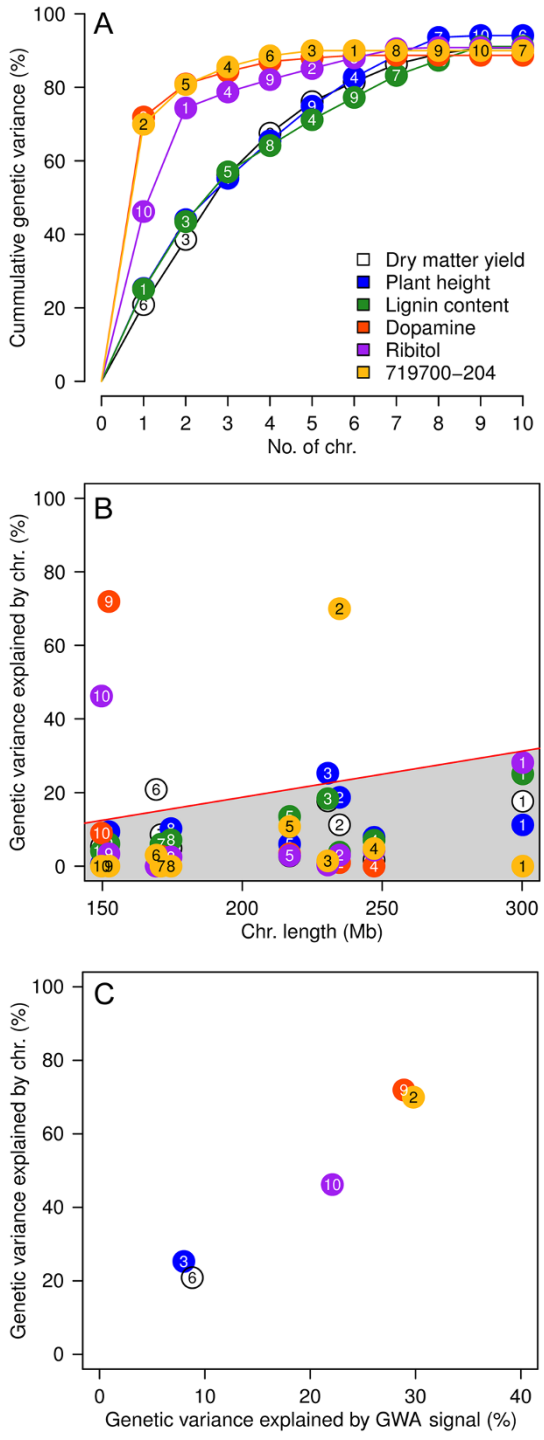

Figure 2 Characterization of the genetic architecture of different traits by genome partitioning of the genetic variance. (**A**) Cummulative genetic variance explained by individual chromosomes. (**B**) Genetic variance explained by each chromosome (number in points). The chromosomes containing either major mQTL for metabolites or putative minor QTL for agronomic traits lie above the red line. (**C**) Genetic variance explained by chromosomes plotted against the genetic variance explained by the GWA signals on these chromosomes.

7.0×10^{-4}) and dry matter yield ($r = 0.60, P = 0.09$). The grey area in Figure 2B was therefore regarded as the range in chromosomal genetic variance explainable by the length of the chromosome. On the other side, the genetic variance contributed by the left-out chromosomes was highly correlated ($r = 0.98, P = 0.003$) with the explained genetic variance of the individual SNPs found by GWA mapping (Figure 2C).

The total genetic variance summed over all chromosomes amounted to 0.91 for dry matter yield, 0.94 for plant height, 0.91 for lignin content, 0.89 for dopamine, 0.91 for ribitol, and 0.90 for the unknown metabolite. These values were close to the heritabilities and repeatabilities obtained from the phenotypic analysis (Table 3).

Prediction accuracies of WGP ranged between 0.45 and 0.82 with standard deviations between 0.05 to 0.12 across traits and models (Table 3). The largest differences in accuracies between models ranged from 0.05 to 0.14 for the same trait. Between RR-BLUP and RKHS, we found no difference in the prediction accuracies above 0.01 for any trait.

For agronomic traits, prediction accuracies were highest for RR-BLUP with a drop of 0.09 to 0.12 if LASSO or elastic net was used and with a drop of 0.01 to 0.11 if BayesB was used.

The ranking of the prediction accuracies for the WGP models was reverse for metabolites. Here, prediction accuracies were highest for LASSO or elastic net with a drop of 0.05 to 0.14 when using RR-BLUP. For metabolites, no differences in the prediction accuracies above 0.01 were observed between RR-BLUP and BayesB. For dopamine and the unknown metabolites, the mQTL were precisely found with LASSO, elastic net and also RR-BLUP (Figure 3). For all three models, their largest absolute SNP effect matched exactly with the SNP identified by GWA mapping. However, the three models differed drastically in their sparsity in SNP effects, and the distance over which the mQTL effect was distributed. Whereas the mQTL effects declined sharply with LASSO or elastic net, they were diluted over a much longer distance with RR-BLUP.

Discussion

We found in a diverse panel of elite maize inbred lines that prediction accuracies obtained with five different WGP models were remarkable similar, even for traits with drastically deviating genetic architecture. Our results suggest that small gains in accuracies (up to 0.14) can be gained if the WGP model is selected according to the genetic architecture underlying the trait.

Recently, Heslot *et al.* [18] reported similar small differences for seven parametric WGP models when comparing them for different presumable highly polygenic agronomic traits over eight datasets of barley, *Arabidopsis*

Table 3 Prediction accuracies ($r_{(g,\hat{g})}$) and their standard deviations (s.d.) for different WGP models

Trait	h^2	RR-BLUP		LASSO		Elastic net		RKHS		BayesB	
		$r_{(g,\hat{g})}$	s.d.	$r_{(g,\hat{g})}$	s.d.	$r_{(g,\hat{g})}$	s.d.	$r_{(g,\hat{g})}$	s.d.	$r_{(g,\hat{g})}$	s.d.
Dry matter yield	0.93	0.61	0.07	0.51	0.11	0.56	0.08	0.61	0.07	0.59	0.08
Plant height	0.97	0.57	0.09	0.45	0.11	0.48	0.11	0.57	0.09	0.56	0.08
Lignin content	0.88	0.69	0.07	0.60	0.08	0.60	0.10	0.68	0.07	0.58	0.09
Dopamine	0.97	0.74	0.06	0.79	0.06	0.79	0.06	0.74	0.07	0.75	0.06
Ribitol	0.95	0.49	0.12	0.61	0.10	0.63	0.10	0.50	0.10	0.50	0.11
719700-204	0.96	0.79	0.06	0.82	0.05	0.82	0.05	0.80	0.05	0.80	0.08

Results are averaged over all 100 cross-validation runs. For the agronomic traits, h^2 is the heritability on a line-mean basis and for the metabolites, the repeatability is shown.

thaliana, maize, and wheat. For the metabolites, however, our results differ from those obtained from Clark *et al.* [19], who investigated the influence of genetic architecture on prediction accuracies achieved by RR-BLUP or BayesB. Whereas these authors found only slight differences for simulated traits with a genetic architecture close to the infinitesimal genetic model, BayesB outperformed RR-BLUP by an increase in prediction accuracy of ≈ 0.4 if the trait is controlled by either a few common or a few rare QTL. Simulation also predicted a drop in prediction accuracy in case of RR-BLUP for traits controlled by a small number of QTL [44]. Although LASSO, elastic net, and BayesB showed higher accuracies compared to RR-BLUP for metabolites, we found the differences to be remarkable small in case of LASSO or elastic net and negliable in the case of BayesB.

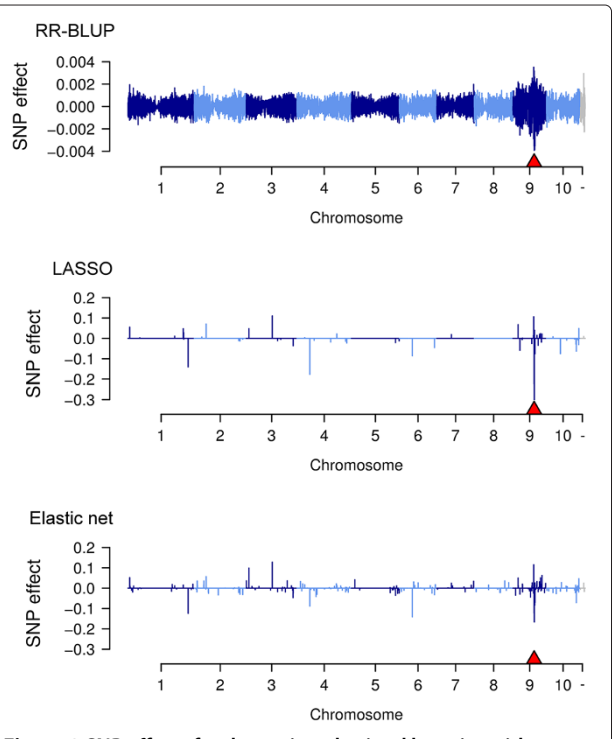

Figure 3 SNP effects for dopamine obtained by using either RR-BLUP, LASSO, or the elastic net model. The position of the mQTL is indicated as a red triangle.

One major reason of the minor differences in prediction accuracies among the different models lies in the high level of LD found in elite breeding germplasm of maize. Our results suggest that with this level of LD ($r^2 = 0.1$ at $\approx$ 500 kb), accuracies are quite similar irrespective whether the effect of large QTL are precisely captured (as in the case of LASSO, elastic net, or BayesB) or spread over a larger region (as in the case of RR-BLUP and RKHS). Since our population was highly diverse for elite maize germplasm in Europe, it is unlikely that breeders are confronted with lower levels of LD unless they work with highly exotic germplasm for which LD has been reported to decline within 5-10 kb [45].

Moreover, the high similarity of RKHS and RR-BLUP suggest that either (i) non-additive, epistatic genetic effects are not present, (ii) these are so small that they are negligible in WGP for the investigated traits, or (iii) RKHS regression is unable to capture them. In either case, for prediction purposes RKHS does not seem to yield any advancements over RR-BLUP for situations comparable to our germplasm and traits. Dominance, as another source of non-additive genetic, effects cannot be present in the inbred lines investigated in this study. For predicting heterozygeous F_1 maize hybrids, however, it has been shown that modeling dominance effects can result in higher prediction accuracies [8].

Although BayesB reached for 5 of the 6 traits a higher prediction accuracy than the worst model, we cannot recommend it because of the excessively larger computation time and the negliable differences in prediction accuracies compared with RR-BLUP in case of the metabolites as the result of probably only sampling error.

We found the approach to partition genetic variance over chromosomes useful for guiding the breeder which WGP model to prefer in the case of little or no prior knowledge on the genetic architecture. Whereas for the agronomic traits an approximately linear increase of cumulative explained genetic variance matched with a

superiority of the L_2 penalty (RR-BLUP), the L_1 penalty (LASSO) or a mixture of both penalties (elastic net) performed better in the case of the metabolites with a strong convex curve curvature (Figure 2A). Although for dry matter yield and plant height, barely significant association signals with a proportion of explained genetic variance $< 9\%$ led to a chromosomal genetic variance slightly above the range expected from length of the chromosome (Figure 2B), these effects were too small to justify the use of the elastic net or LASSO.

As an alternative to this approach, Hayes *et al.* [20] estimated successively the genetic variance explained by each chromosome segment and compared it with the genetic variance captured by the remaining part of the genome. To correct for the non-independence of neighbouring segments, they applied a bias correction using an expectation maximization (EM) algorithm. Such a correction is not necessary if the variance components for all chromosomes are estimated simultaneously as applied in this study; this is a further advantage besides its straightforward implementation using standard mixed model software packages such as ASReml.

Conclusions

Our empirical data of WGP in a large panel of diverse maize inbred lines suggest that (i) different WGP models differ only slightly in their prediction accuracies, irrespective of the number and effects of QTL found in association analysis, (ii) small gains in prediction accuracies can be obtained if the WGP model is selected according to the genetic architecture of the trait, (iii) genome partitioning of genetic variance offers a straightforward approach for model selection if the genetic architecture is unknown. The question of which WGP model to choose is therefore not expected to hamper implementation of WGP in maize breeding.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CR analyzed the data and wrote the manuscript. AEM supervised the research. FT wrote the C code for implementing BayesB. All authors read and approved the final manuscript.

Acknowledgements

We thank the staff of the experimental stations of the University of Hohenheim for conducting the field experiments. We thank the groups of Mark Stitt and Lothar Willmitzer of the Max Planck Institute of Molecular Plant Physiology for performing the metabolic profiling. This research was funded by the German Federal Ministry of Education and Research (BMBF) within the project GABI-Energy (FKZ: 0315045) and the AgroClustEr 'Synbreed - Synergistic plant and animal breeding' (FKZ: 0315528D).

Received: 27 February 2012 Accepted: 14 August 2012
Published: 4 September 2012

References

1. Morrell PL, Buckler ES, Ross-Ibarra J: **Crop genomics: advances and applications.** *Nat Rev Genet* 2011, **13:**85–96.
2. Heffner E, Sorrells M, Jannink JL: **Genomic Selection for Crop Improvement.** *Crop Sci* 2009, **49:**1–12.
3. Jannink JL, Lorenz AJ, Iwata H: **Genomic selection in plant breeding: from theory to practice.** *Brief Funct Genomics* 2010, **9:**166–177.
4. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819–1829.
5. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: Genomic selection in dairy cattle: progress and challenges.** *J. Dairy Sci* 2009, **92:**433–443.
6. 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:**339–350.
7. Zhao Y, Gowda M, Liu W, Würschrum T, Maurer HP, Longin FH, Ranc N, Reif JC: **Accuracy of genomic selection in European maize elite breeding populations.** *Theor Appl Genet* 2012, **124:**769–776.
8. Technow F, Riedelsheimer C, Schrag TA, Melchinger AE: **Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects.** *Theor Appl Genet* 2012, **125:**1181-1194.
9. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE: **Genomic and metabolic prediction of complex heterotic traits in hybrid maize.** *Nat Genet* 2012, **44:**217–220.
10. Fisher R: **The correlation between relatives on the supposition of Mendelian inheritance.** *Trans Roy Soc Edinburgh* 1918, **52:**399–433.
11. Simic D, Mladenovic Drinic S, Zdunic Z, Jambrovic A, Ledencan T, Brkic J, Brkic A, Brkic I: **Quantitative trait Loci for biofortification traits in maize grain.** *J Hered* 2012, **103:**47–54.
12. Wong JC, Lambert RJ, Wurtzel ET, Rocheford TR: **QTL and candidate genes phytoene synthase and zeta-carotene desaturase associated with the accumulation of carotenoids in maize.** *Theor Appl Genet* 2004, **108:**349–359.
13. Wisser RJ, Balint-Kurti PJ, Nelson RJ: **The genetic architecture of disease resistance in maize: a synthesis of published studies.** *Phytopathology* 2006, **96:**120–129.
14. Martin M, Miedaner T, Dhillon BS, Ufermann U, Kessel B, Ouzunova M, Schipprack W, Melchinger AE: **Colocalization of QTL for Gibberella Ear Rot Resistance and Low Mycotoxin Contamination in Early European Maize.** *Crop Sci* 2011, **51:**1935–1945.
15. Martin M, Miedaner T, Schwegler DD, Kessel B, Ouzunova M, Dhillon BS, Schipprack W, Utz HF, Melchinger AE: **Comparative Quantitative Trait Loci Mapping for Gibberella Ear Rot Resistance and Reduced Deoxynivalenol Contamination across Connected Maize Populations.** *Crop Sci* 2012, **52:**32–43.
16. Welz H, Geiger H: **Genes for resistance to northern corn leaf blight in diverse maize populations.** *Plant Breeding* 2000, **119:**1–14.
17. Poland JA, Bradbury PJ, Buckler E, Nelson RJ: **Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize.** *Proc Natl Acad Sci USA* 2011, **108:**6893–6898.
18. Heslot N, Yang HP, Sorrells ME, Jannink JL: **Genomic Selection in Plant Breeding: A Comparison of Models.** *Crop Sci* 2012, **52:**146–160.
19. Clark SA, Hickey JM, van der Werf JH: **Different models of genetic variation and their effect on genomic evaluation.** *Genet Sel Evol* 2011, **43:**18.
20. Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME: **Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits.** *PLoS Genet* 2010, **6:**e1001139.
21. Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS: **Investigation of the bottleneck leading to the domestication of maize.** *Proc Natl Acad Sci USA* 1998, **95:**4441–4446.
22. Duvick D, Smith J, Cooper M: **Long-Term Selection in a Commercial Hybrid Maize Breeding Program.** *Plant Breed Rev* 2004, **24:**109–151.
23. Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE: **Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize.** *Proc Natl Acad Sci USA* 2012, **109:**8872–8877.
24. Grieder C, Mittweg G, Dhillon BS, Montes JM, Orsini E, Melchinger AE: **Determination of methane fermentation yield and its kinetics by**

near infrared spectroscopy and chemical composition in maize. *JNIRS* 2011, **19:**463–477.
25. Grieder C, Dhillon B, Schipprack W, Melchinger A: **Breeding maize as biogas substrate in Central Europe: II. Quantitative-genetic parameters for inbred lines and correlations with testcross performance.** *Theor Appl Genet* 2012, **124:**981–988.
26. 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: **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* 2011, **6:**e28334.
27. Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR: **Gas chromatography mass spectrometry-based metabolite profiling in plants.** *Nat Prot* 2006, **1:**387–396.
28. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM: **Genome partitioning of genetic variation for complex traits using common SNPs.** *Nat Genet* 2011, **43:**519–525.
29. Butler D, Cullis B, Gilmour A, Gogel B: *ASReml-R reference manual, version 3*. Queensland, Australia: The Department of Primary Industries and Fisheries; 2009.
30. Higham N: **Computing the nearest correlation matrix — a problem from finance.** *IMA J Numer Anal* 2002, **22:**329–343.
31. Bates D, Maechler M: **Matrix: Sparse and Dense Matrix Classes and Methods** 2012. [http://cran.r-project.org/package=Matrix].
32. de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM: **Predicting quantitative traits with regression models for dense molecular markers and pedigree.** *Genetics* 2009, **182:**375–385.
33. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J R Statist Soc B* 2005, **67:**301–320.
34. Gianola D, van Kaam JBCHM: **Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits.** *Genetics* 2008, **178:**2289–2303.
35. Ruppert D, Wand M, Carroll R: *Semiparametric Regression*. 1st edition. New York, NY, USA: Cambridge University Press; 2003.
36. Piepho H: **Ridge regression and extensions for genomewide selection in maize.** *Crop Sci* 2009, **49:**1165–1176.
37. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: a tool for genome-wide complex trait analysis.** *Am J Hum Genet* 2011, **88:**76–82.
38. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning*. 2nd edition. New York, NY, USA: Springer Science+Business Media; 2009.
39. Tibshirani R: **Regression shrinkage and selection via the lasso.** *J R Statist Soc B* 1996, **58:**267–288.
40. Li Z, Sillanpää MJ: **Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection.** *Theor Appl Genet* 2012, **125:**419-435.
41. Ogutu JO, Schulz-Streeck T, Piepho HP: **Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions.** *BMC Proceedings* 2012, **6**(Suppl 2):S10.
42. Friedman J, Hastie T, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent.** *J Stat Softw* 2010, **33:**1–22.
43. Yang W, Tempelman RJ: **A Bayesian Antedependence Model for Whole Genome Prediction.** *Genetics* 2011, **190:**1491–1501.
44. Habier D, Fernando RL, Dekkers JCM: **The impact of genetic relationship information on genome-assisted breeding values.** *Genetics* 2007, **177:**2389–2397.
45. Yan J, Shah T, Warburton M, Buckler E, McMullen M, Crouch J: **Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers.** *PLoS ONE* 2009, **4:**e8451.

doi:10.1186/1471-2164-13-452
Cite this article as: Riedelsheimer *et al.*: Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics* 2012 **13**:452.