Accepted Article

Received Date : 19-Apr-2016

Revised Date : 10-Jun-2016

Accepted Date : 15-Jun-2016

Article type : Original Article

Metabolomic Prediction of Yield in Hybrid Rice

Shizhong Xu[a,1], Yang Xu[a], Liang Gong[b] and Qifa Zhang[b,1]

[a]Department of Botany and Plant Sciences, University of California, Riverside, California 92507, United States of America

[b]National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China

[1]Corresponding Authors:

Shizhong Xu

Department of Botany and Plant Sciences

University of California

Riverside, CA 92521

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1111/tpj.13242

This article is protected by copyright. All rights reserved.

Phone: (951) 827-5898

Fax: (951) 827-4437

E-mail: shizhong.xu@ucr.edu

Qifa Zhang

National Key Laboratory of Crop Genetic Improvement

National Centre of Plant Gene Research (Wuhan)

Huazhong Agricultural University

Wuhan 430070, China

E-mail: qifqzh@mail.hzau.edu.cn

Running Title :

Metabolomic Prediction

Key words:

genomic prediction; hybrid breeding; rice; metabolites; transcripts

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

Author Contributions

S.X. and Q.Z. designed the research; S.X. and Q.Z. performed the research; S.X., Y.X. and L.G. analyzed the data; and S.X. and Q.Z. wrote the paper.

Summary

Rice (*Oryza sativa*) provides a staple food source for more than 50% of the world's population. An increase in yield can significantly contribute to global food security. Hybrid breeding can potentially help to meet this goal because hybrid rice often shows a considerable increase in yield when compared with purebred cultivars. We recently developed a marker-guided prediction method for hybrid yield and showed a substantial increase in yield through genomic hybrid breeding. We now have transcriptomic and metabolomic data as potential resources for prediction. Using six prediction methods, including least absolute shrinkage and selection operator (LASSO), best linear unbiased prediction (BLUP), stochastic search variable selection (SSVS), partial least squares (PLS) and support vector machines (SVM-RBF and SVM-POLY), we found that the predictability of hybrid yield can be further increased using these omic data. LASSO and BLUP are the most efficient methods for yield prediction. For high heritability traits, genomic data remain the most efficient predictors. When metabolomic data are used, the predictability of hybrid yield is almost doubled compared with genomic prediction. Of the 21945 potential hybrids derived from 210 recombinant inbred lines, selection of the top 10 hybrids predicted from metabolites would lead to a ~30% increase in yield. We hypothesize that each metabolite represents a biologically built-in genetic network for yield; thus, using metabolites for prediction is equivalent to using information integrated from these hidden genetic networks for yield prediction.

This article is protected by copyright. All rights reserved.

Accepted Article

Introduction

Rice breeders have been struggling to improve yield via selection due to low heritability of the trait. Fortunately, such traits often demonstrate great heterosis (Zhou *et al.* 2012). Therefore, hybrid breeding is key to increasing yield in rice by taking advantage of heterosis. Although superior hybrids have been identified and are being used in rice production, there are about 10,000 rice cultivars available in the world and only a small proportion of all potential crosses have been evaluated in the field. Experimental evaluation of all crosses would allow us to identify the most valuable hybrids. However, any experiment requires temporal and spatial replications and such a large-scale experiment is impractical due to limited resources. Genomic hybrid breeding (Bernardo 1994) provides a solution for predicting all potential hybrids. Theoretically, we can predict all potential hybrids of a given set of parents using a subset of crosses.

Advanced molecular technologies allow us to measure the expression of many metabolites and transcripts. Metabolome and transcriptome provide new sources of data for hybrid prediction. Previously, microarray data were analyzed largely in relation to genomic data, called eQTL mapping (Bing and Hoeschele 2005, Keurentjes *et al.* 2007, Rockman and Kruglyak 2006, Wang *et al.* 2014, West *et al.* 2007). Additionally, metabolites can be detected and quantified by chromatographic mass spectrometry (Dunn and Ellis 2005). These metabolomic profiling data have also been analyzed in relation to genomic data, called mQTL mapping (Carreno-Quintero *et al.* 2013, Chen *et al.* 2014, Gong *et al.* 2013, Keurentjes *et al.* 2006, Keurentjes 2009, Lisec *et al.* 2009, Rowe *et al.* 2008, Schauer and Fernie 2006). An issue with these types of analysis is that the genetic networks inferred from eQTL and mQTL mapping are generic because they are not necessarily associated with any traits.

This article is protected by copyright. All rights reserved.

Accepted Article

For the first time, Frisch *et al.* (2010) predicted hybrid performance of maize using transcriptomic data measured from the parents and showed that transcriptome-based prediction was more accurate than prediction from DNA markers. Fu *et al.* (2012) further compared different methods of hybrid prediction in maize yield using parental transcriptomic. Use of metabolites to predict plant phenotypes has been reported (Feher *et al.* 2014, Gärtner *et al.* 2009, Meyer *et al.* 2007, Riedelsheimer *et al.* 2012, Steinfath *et al.* 2010). Three of these reports addressed hybrid prediction using both genomic and metabolomic data. Gärtner, Steinfath, Andorf, Lisec, Meyer, Altmann, Willmitzer and Selbig (2009) predicted biomass of *Arabidopsis thaliana* using two backcross populations from 359 recombinant inbred lines ($359\times 2=718$ hybrids). They found that the predictability of biomass was 0.17 from genome and 0.16 from metabolome, where the predictability was defined as the squared correlation between the observed and the predicted phenotypes. Riedelsheimer, Czedik-Eysenberg, Grieder, Lisec, Technow, Sulpice, Altmann, Stitt, Willmitzer and Melchinger (2012) predicted traits in hybrid maize using 278 inbred lines crossed with two testers ($285\times 2=570$ hybrids) and found that the average predictability of seven traits was 0.53 from genome and 0.32 from metabolome. We converted their correlations into squared correlations to comply with our definition of predictability. They claimed that the predictabilities were very similar from the two predictors (Appendix S1). The important message from the above studies is that metabolites are useful predictors for quantitative traits. Feher, Lisec, Römisch-Margl, Selbig, Gierl, Piepho, Nikoloski and Willmitzer (2014) also proposed to use metabolites to predict hybrid biomass and they demonstrated the method using a toy example of maize, with only four parents and $\binom{2}{4}=6$ hybrids.

This article is protected by copyright. All rights reserved.

Accepted Article

One characteristic of these experiments is that all hybrids from the current parents have been evaluated, leaving no future crosses for prediction. To predict future crosses, we need a cross experiment with a portion of the hybrids evaluated in the field to predict the remaining portion. The hybrid experiment of Hua *et al.* (2003) set an example of this kind. This experiment involved 210 inbred parents with 21945 potential hybrids. Of all the hybrids, only 278 were evaluated in the field and the remaining 21667 have yet to be tested. We proposed to use the 278 hybrids as a training sample and predict all potential hybrids. This technology is directly transferrable to commercial hybrid breeding. Recently, we predicted yields of all hybrids using SNP data with a predictability of 0.20 (Xu *et al.* 2014). We predicted that if the top 10 crosses were selected for hybrid breeding, the yield would increase by 16%. The predictability for a diverse panel would be higher given the much increased genetic variation. The parents initiating the cross experiment of Riedelsheimer, Czedik-Eysenberg, Grieder, Lisec, Technow, Sulpice, Altmann, Stitt, Willmitzer and Melchinger (2012) represent a diversity panel with a wider inference space. Unfortunately, all potential hybrids were evaluated in that study and thus no future hybrids were predicted.

Results

Predictability of yield and its components traits in hybrids. The predictabilities drawn from 10-fold cross validation are illustrated in Figure 1, from which we conclude: (1) The predictability is highly correlated to the heritability of the trait, with KGW having the highest predictability (0.58), average across all methods and omic data) followed by GRAIN (0.31) and YIELD (0.20), and TILLER being the least predictable trait (0.16). The corresponding heritability for each of the four traits (on the plot mean basis) was 0.79, 0.62, 0.43 and 0.31, respectively (Table S1). The correlation between heritability and

This article is protected by copyright. All rights reserved.

predictability is 0.9524 (p = 0.047); (2) For YIELD, metabolome had the highest predictability followed by transcriptome and genome. With the LASSO method, the predictabilities from metabolome, transcriptome and genome were 0.35, 0.23 and 0.20, respectively. Metabolomic prediction for YIELD was almost twice as efficient as genomic prediction; (3) For KGW, genomic prediction was most efficient followed by transcriptomic and metabolomic predictions. KGW is a high heritability trait and genomic prediction remained dominant over other omic predictions.

Analysis of variances for predictabilities. Table 1 is the ANOVA table for predictability. All main effects and two interaction effects are significant. We also performed multiple comparisons for the main effects and the results are depicted in Figure 2. Overall, transcriptomic prediction is better than genomic prediction, while metabolomic prediction is not different from either one (Figure 2a). Predictions of the four traits are significantly different, KGW > GRAIN > YIELD > TILLER (Figure 2b). The six methods are classified into three levels of predictability, BLUP being the best, SSVS the worst and other methods ranging between the two (Figure 2c). Overall, BLUP and LASSO are the best methods for prediction. Detailed information of the ANOVA in terms of main and interaction effects is given in Data S1.

Significance test of predictability. We performed significance tests for the predictability of yield under the LASSO method. The empirical distributions of the predictabilities drawn from 1000 permuted samples are presented in Figure 3. The p-value is zero in every case except for TILLER where the p-value is $2/1000 = 0.002$.

This article is protected by copyright. All rights reserved.

Accepted Article

Comparison of leave and seed metabolites. We also compared predictabilities of 683 metabolites from leaves and 317 metabolites from seeds separately. On average, across all traits and all methods, the predictability was 0.278 from leaves, 0.216 from seeds and 0.308 from all metabolites. To test whether the higher predictability of leaves is due to the larger number of metabolites, we randomly selected 317 metabolites from leaves to match the number from seeds. On average of 20 replicated samples, the predictability of leaves was 0.242, still higher than that of the seeds (Table S2). We concluded that the higher predictability of leaves was not entirely due to the larger number of metabolites. This follows the nominal expectation as flag leaves play a crucial role in yield production. As suggested by a reviewer, we reanalyzed the data for metabolomic prediction using the 100 metaboloites that are measured from both leaves and seeds using all six methods for all four traits. The results are shown in Supplementary Table S2 (last two columns). The general conclusions were that (1) the 100 metabolites predicted poorly compared with the predictions using all 1000 metabolites; (2) leaf metabolites predicted better than seed metabolites for yield, grain number and tiller number but worse for KGW.

Combined prediction using all sources of omic data. The combined prediction rarely outperformed the best single data prediction (Figure S1). This result is consistent with Riedelsheimer et al. (Riedelsheimer, Czedik-Eysenberg, Grieder, Lisec, Technow, Sulpice, Altmann, Stitt, Willmitzer and Melchinger 2012) who did not find any benefit from combining genomic and metabolomic data. However, Gärtner, Steinfath, Andorf, Lisec, Meyer, Altmann, Willmitzer and Selbig (2009) found a considerable increase in predictability by combining the two predictors. A possible reason for the benefit observed by Gärtner et al. may be the small number of SNPs (110) used in that study.

This article is protected by copyright. All rights reserved.

Accepted Article

Predicting untested crosses. The 278 crosses are a small subset of all 21945 crosses. Using parameters estimated from this training sample, we predicted all potential hybrids for YIELD. Data S2 gives the predicted yields of all crosses from all omic data using the LASSO method. This dataset also shows the predicted yields of all crosses sorted by the predicted yield using each of the three omic data. Shanyou 63 (the original hybrid of Zhenshan 97 and Minghui 63) had an average yield of 52.60. The metabolomic data predict that there would be 160 potential hybrids with yield higher than Shanyou 63, with a proportion of 0.7%. According to transcriptomic prediction, there would be 72 potential crosses with yield higher than Shanyou 63. Based on genomic prediction, none of the potential crosses would outperform Shanyou 63. Among the 278 field evaluated IMF2 crosses, 13 of them had yield greater than 52.6 (the yield of Shanyou 63), with a proportion of 4.7%. This proportion is larger than 0.7% in the whole predicted population. Unless the predictability is 100%, the predicted and observed yields have different distributions with the predicted yield having a much smaller variance (as demonstrated in Figure S2). Both the predicted and observed yields had mean of 43.48, but the standard deviations were 3.73 and 5.84, respectively. Because the smaller standard deviation for the predicted yield, the tail above 52.6 covers a much smaller proportion of the sample. For yield, the predictability was 0.35, far less than 100%, therefore, the upper tail above 52.6 was only 0.7%.

Figure 4 shows the average predicted yield and the percent gain when selecting the top crosses for hybrid breeding. For example, if the top 10 crosses predicted from metabolites are used for hybrid breeding, the average predicted yield of the 10 crosses would be 56.38, which represents a 29.6% gain in yield. The 29.6% gain appears to be exaggerated for such a low heritability trait, however, this high gain is largely due to the high selection intensity of the crosses represented by the extremely small proportion selected ($10/21945 = 0.000456$).

This article is protected by copyright. All rights reserved.

Accepted Article

We examined yield prediction using the LASSO method for metabolites and transcripts for each year separately. For metabolomic prediction, the predictabilities are 0.15 and 0.29 for years 1998 and 1999, respectively, and both are lower than the two-year combined prediction (0.35). Similarly, for transcriptomic prediction, the predictabilities are 0.03 and 0.15 for years 1998 and 1999, respectively. Again, both are lower than the combined prediction (0.23). We also predicted yields for 1999 using data from 1998 and vice versa for both metabolomic and transcriptomic predictions using the LASSO method. The average predictabilities from the cross year prediction are (0.27+0.23)/2 = 0.25 for metabolomic prediction and (0.19+0.20)/2 = 0.195 for transcriptomic prediction. These cross year predictions are lower than the two-year combined predictions, reflecting the environmental effects on the predictions.

Metabolites significantly effect on yield. The LASSO method allows us to detect metabolites with significant effects on yield. We choose a $p < 0.01$ criterion to declare significant metabolites. Among the 1000 metabolites, 76 of them are significant and they are listed in Data S3. Among the 76 significant metabolites, 46 are from leaves and 30 from seeds. The top 22 metabolites with the smallest p values are all from leaves. We then used only the 76 metabolites to predict hybrid yield. The predictability drawn from cross-validation is 0.206, which is smaller than 0.35, the predictability using all 1000 metabolites. The conclusion was that the small effect metabolites that failed to reach the significant level do contribute to prediction. In other words, variable selection can decrease the prediction.

This article is protected by copyright. All rights reserved.

Accepted Article

Discussion

We found that metabolomic prediction for hybrid yield is more effective than genomic prediction. Yield is a trait with low heritability but is the most important trait in rice breeding. Even a slight improvement in prediction will translate into a significant increase in yield. We performed similar prediction for the 210 inbred lines (RIL) and observed the same trend of metabolomic prediction being consistently better than genomic prediction (Figure S3). We also observed that the predictability in RILs was higher than that in hybrids, which may be explained by the fact that the metabolites were directly measured in the parents. Metabolites and transcripts are intermediate phenotypes correlated to yield. They are biologically and developmentally closer to yield than SNPs and this may partially explain why their prediction is higher than genomic prediction.

Looking back to hybrid rice, on average across all traits and methods, the predictability was 0.31 from metabolomic data and 0.30 from genomic data. In the maize hybrid prediction of Riedelsheimer, Czedik-Eysenberg, Grieder, Lisec, Technow, Sulpice, Altmann, Stitt, Willmitzer and Melchinger (2012), the average predictability across traits was 0.53 from genomic data and 0.32 from metabolomic data (Appendix S1). This discovery came as a surprise to the authors. The fact that metabolomic prediction is better than genomic prediction for hybrid rice is even more surprising, considering (i) the small number of metabolites relative to the large number of SNPs and (ii) the metabolic profile being a snapshot at a specific moment in time (Riedelsheimer, Czedik-Eysenberg, Grieder, Lisec, Technow, Sulpice, Altmann, Stitt, Willmitzer and Melchinger 2012). Heritability describes the linear relationship between trait and genotype (Falconer and Mackay 1996) and genomic prediction is able to capture this relationship. The expression level of each metabolite may be a complicated non-linear function of the SNP genotypes, but this complicated function is captured by the metabolite biologically, not mathematically (Gärtner,

This article is protected by copyright. All rights reserved.

Steinfath, Andorf, Lisec, Meyer, Altmann, Willmitzer and Selbig 2009). The unknown complex functions are biologically built-in and using information integrated from these hidden functions for prediction is much more powerful than using any mathematical functions inferred by our models. Although it is biologically interesting to find the functional relationships, breeders are perhaps more interested in the results: as long as it helps prediction, the technology can be adopted. Breeders may not want to wait years to find the biological reasons before applying the technology.

There are about 200,000 different metabolites in the plant kingdom (Bino *et al.* 2004), but in each experiment, only a few hundred of them can be measured. This small proportion already predicts the yield better than genomic data. If we can increase the number of metabolites to a few thousands, the predictability may be further improved. In addition to metabolites, we found that transcripts also improved yield prediction, although not as efficient as metabolites. Proteomic data have been analyzed in a hybrid rice and its parents (Xiang *et al.* 2013). Such data may also be used to predict hybrid yield, albeit no such a study has been reported. For the same token, small RNA data or any other molecular data can be used for prediction of yield (Zhang *et al.* 2014). Advanced technologies are available to measure large array of phenotypes and the data are called phenomes (Houle *et al.* 2010). In general, metabolome, transcriptome and proteome also belong to phenomes, which are quantitative traits and can be used as secondary traits for indirect selection of yield. These data are further closer to yield and may be significantly better predictors of hybrid yield. One obvious question is that "yield" itself is just one of the phenomic data and using yields of the parents to predict the yield of hybrids is no better than any other phenotypic traits. We emphasize the multivariate nature of the phenomes. If we can measure thousands of phenotypes from a single plant and these phenotypes may be included in a single model

This article is protected by copyright. All rights reserved.

Accepted Article

for prediction. Such as phenomic prediction can be performed the same way as the metabolomic prediction.

The narrow genetic background of the materials in this study may limit the application of the result to hybrid breeding. The parameters reported can only be used to predict hybrids from the same 210 lines. However, this study provides a proof of concept for prediction of hybrid yield. In practical hybrid breeding, a balanced random partial rectangle cross design (BRPRCD) may be implemented using the majority of available rice cultivars. A toy example of crosses using 8 male and 14 female lines is demonstrated in Table S3, which can be extended to any numbers of male and female lines. Imagine that if we choose 500 male and 1,000 female lines of all rice cultivars in the world to create a BRPRCD experiment. Although it is impossible to evaluate all 500,000 crosses in the field, using 800 crosses for field evaluation, for example, would allow us to predict all potential crosses. The predicted top crosses, not yet evaluated, would then be field evaluated. This omic-data guided hybrid breeding can more effectively identify the best hybrids in the world, leading to improved yields and global food security.

Experimental procedures

Material collection. We analyzed a hybrid population of rice (*Oryza sativa*) derived from the cross between Zhenshan 97 and Minghui 63 (Hua, Xing, Wu, Xu, Sun, Yu and Zhang 2003, Hua *et al.* 2002, Xing *et al.* 2002). This hybrid (Shanyou 63) is the most widely grown hybrid in China. A total of 210 RILs were derived by single-seed descent from this hybrid. We created 278 crosses by randomly pairing the 210 RILs. This population is called an immortalized F_2 (IMF2) because it mimics an F_2 population with a 1:2:1 ratio of the three genotypes (Hua, Xing, Wu, Xu, Sun, Yu and Zhang 2003). We analyzed four traits to

This article is protected by copyright. All rights reserved.

Accepted Article

evaluate the efficacy of hybrid prediction: (1) yield (YIELD), (2) 1000-grain weight (KGW), (3) grain number per plant (GRAIN) and (4) tiller number per plant (TILLER). For the RIL population, each trait was measured from four replicated experiments (1997 and 1998 from one location, 1998 and 1999 from another location). In each replicate, eight plants from each line were sampled and the average phenotype represented the phenotypic value of the line (Xing, Tan, Hua, Sun, Xu and Zhang 2002, Yu *et al.* 2011). For the IMF2 population, each trait was measured in two consecutive years (1998 and 1999). Each year, eight plants from each cross were measured and the average yield of the eight plants was treated as the original data point. Both experiments were conducted under a randomized complete block design with replicates (years and locations) as the blocks.

Three omic (genomic, transcriptomic and metabolomic) data collected from the 210 RILs were used for prediction. The genomic data are represented by 1619 bins inferred from ~270,000 SNPs of the rice genome (Yu, Xie, Wang, Xing, Xu, Li, Xiao and Zhang 2011). All SNPs within a bin have exactly the same segregation pattern (perfect LD). The bin genotypes of the 210 RILs were coded as 1 for the Zhenshan 97 genotype and 0 for the Minghui 63 genotype. Genotypes of the hybrids were deduced from genotypes of the two parents. The transcriptomic data consisted of 24994 gene expression traits measured in tissues sampled from flag leaves for all the 210 RILs in 2008 (Wang, Yu, Weng, Xie, Xu, Li, Xiao and Zhang 2014). Each line had two biological replicates, but RNA extracted from the two replicates was mixed in a 1:1 ratio before microarray expression profiling was conducted. The original expression levels were $\log_2$ transformed before analysis. The metabolomic data consisted of 683 metabolites measured from flag leaves and 317 metabolites measured from germinated seeds (Gong, Chen, Gao, Liu, Zhang, Xu, Yua, Zhang and Luo 2013). The metabolomic data were collected in 2009 and 2010 (two replicates). For metabolic profiling, germinated seeds were sampled in one biological replicate in 2009 and one in 2010,

This article is protected by copyright. All rights reserved.

Accepted Article

and flag leaves were sampled in two biological replicates in 2009. In both tissues, the expression level of each metabolite was $\log_2$ transformed. For each line, we took the average of expression levels measured from the two replicates as the measurement of the metabolites.

Methods of prediction. We used six statistical methods for prediction: (1) Least absolute shrinkage selection operator (LASSO) developed by Tibshirani (1996) and implemented in the GlmNet/R program (Friedman *et al.* 2010); (2) Henderson's (1975) best linear unbiased prediction (BLUP) adopted to genomic data analysis (VanRaden 2008) and implemented in our own R program (Xu 2013); (3) Stochastic search variable selection (SSVS) developed by George and McCulloch (1993); (4) Support vector machine using the radial basis function kernel (SVM-RBF); (5) Support vector machine using the polynomial kernel function (SVP-POLY) and (6) Partial least squares (PLS). The SSVS method is also called Bayes B (Meuwissen *et al.* 2001) and was implemented using an R package called BGLR (2014). The two SVM methods were implemented in an R program called kernlab (Karatzoglou *et al.* 2004). The PLS was implemented using an R package called pls (Mevik and Wehrens 2007).

Models of prediction. Let *y* be a $n\times 1$ vector of the phenotypic values for the trait of interest, where *n* is the sample size ($n=210$ in RIL and $n=278$ in IFM2). Let X be a $n\times m$ matrix of predictors used to predict *y*, where *m* is the number of predictors in the model and it depends on the source of data and the model. The first three methods (LASSO, BLUP and SSVS) use a random model, as shown by $y=X\beta+\varepsilon$ where β is a $m\times 1$ vector of model effects and ε is a $n\times 1$ vector of residual errors. The model effect β_k was treated as a random effect with either a normal distribution or a mixture of two normal distributions. The LASSO method can be reformulated as a Bayesian hierarchical model,

This article is protected by copyright. All rights reserved.

$\beta_k \sim N(0, \phi_k^2)$ and $\phi_k^2 \sim \text{Exp}(\frac{1}{2}\lambda^2)$ for all $k = 1, ..., m$, where λ is a shrinkage parameter (Tibshirani 1996), although the original LASSO was not formulated this way. The BLUP method assumes $\beta_k \sim N(0, \frac{1}{m}\phi^2)$ for all $k = 1, ..., m$, where ϕ^2 is called the "polygenic variance". The SSVS method assumes that β_k is sampled from one of two normal distributions with an unknown label of the two distributions. Mathematically, it is described as $\beta_k \sim \eta_k N(0, \Delta) + (1 - \eta_k) N(0, \delta)$ where Δ is the variance of the first normal distribution sampled along with other parameters, $\delta = 10^{-5}$ is the variance of the second normal distribution and η_k is the cluster label having a Bernoulli distribution with probability ρ. The missing cluster label η_k takes 1 if β_k belongs to the first distribution and 0 otherwise. The probability of the missing label ρ was modeled by a Beta(1,1) distribution. All parameters were sampled from their posterior distributions. The above three models (LASSO, BLUP and SSVS) are all linear. The two SVM methods use a non-linear relationship between *y* and *X* described as $y = f(X \mid \beta) + \varepsilon$, where $f(X \mid \beta) = \sum_{j=1}^{n} \beta_j K_h(X, X_j)$ and $K_h(X, X_j)$ is a kernel chosen by the users. We chose the Gaussian kernel (SVM-RBF) and the polynomial kernel (SVM-POLY) functions (Karatzoglou, Smola, Hornik and Zeileis 2004). The PLS method is a hybrid method between principal component analysis (PCA) and multiple regression analysis. It uses the first few latent scores of the *X* matrix as predictors to predict the phenotype. However, it differs from PCA in that the weights of the latent scores are calculated by maximizing the covariance between *y* and the scores (Gelandi and Kowalski 1986). The number of latent components was determined by a 10-fold cross-validation to have a minimum prediction error.

This article is protected by copyright. All rights reserved.

Predictability drawn from cross-validation. The predicted trait is denoted by $\hat{y} = X\hat{\beta}$ for LASSO, PLS, BLUP and SSVS, and $\hat{y} = f(X \mid \hat{\beta})$ for SVM-RBF and SVM-POLY. We used a 10-fold cross-validation to evaluate the predictability of each method, where individuals predicted do not contribute to parameter estimation. The predictability is defined as the squared correlation coefficient between *y* and $\hat{y}$ (Appendix S1). The predictability depends on how the sample is partitioned into the 10 folds. It also depends on the number of folds (Figure S4). Therefore, we replicated the cross-validation analysis 10 times to monitor the variation among the replicates. The predictability increases as the number of folds increases, but often reaches a plateau at fold 10. While further increase of the fold number only improves the predictability slightly. Figure S4 shows the plots of predictability against the number of folds from 10 replicates for YIELD from the metabolomic data using the LASSO method in both the RIL and IMF2 populations.

Defining the *X* matrix. For the RIL population, $X = \{X_{jk}\}$ is a $n \times m$ matrix, where $n = 210$, $m = 1000$ for the metabolomic data, $m = 24994$ for the transcriptomic data and $m = 1619$ for the genomic data. The *j*th row and the *k*th column of matrix *X* is defined as the gene expression level for the transcriptomic data and the level of the *k*th metabolite for the metabolomic data. For the genomic data, $X_{jk} = 1$ for the Zhenshan 97 genotype and $X_{jk} = 0$ for the Minghui 63 genotype.

For the IMF2 population, the $X = \{X_{jk}\}$ matrix has $n = 278$ rows and $m = 2 \times 24994$ columns for the transcriptomic data, $m = 2 \times 1000$ columns for the metabolomic data and $m = 2 \times 1619$ columns for the genomic data. The X_{jk} value for each IMF2 cross is a function of the corresponding predictors of

This article is protected by copyright. All rights reserved.

Accepted Article

their RIL parents. Let π_{jk}^{m} and π_{jk}^{f} be the predictors of the male and female RIL parents, respectively. For the genomic data, $\pi_{jk}^{m}=1$ for the Zhenshan 97 genotype and $\pi_{jk}^{m}=0$ for the Minghui 63 genotype of the RIL parents. For the transcriptomic and metabolomics data, π_{jk}^{m} and π_{jk}^{f} are the corresponding measurements of the expression levels of the two parents. We first defined $Z_{jk}=\pi_{jk}^{m}+\pi_{jk}^{f}$ and $W_{jk}=|\pi_{jk}^{m}-\pi_{jk}^{f}|$ for the IMF2 cross, where Z_{jk} is a predictor for the "additive" effect and W_{jk} is a predictor for the "dominance" effect. Take the genomic data for example, let A_1 be the Zhenshan 97 allele and A_2 be the Minghui 63 allele. The predictors of an IMF2 cross are defined as

$$Z_{jk}=\begin{cases}2=1+1 & \text{for} \quad A_1A_1\\ 1=1+0 & \text{for} \quad A_1A_2\\ 0=0+0 & \text{for} \quad A_2A_2\end{cases} \text{ and } W_{jk}=\begin{cases}0=|1-1| & \text{for} \quad A_1A_1\\ 1=|1-0| & \text{for} \quad A_1A_2\\ 0=|0-0| & \text{for} \quad A_2A_2\end{cases} \qquad (1)$$

Please refer to Table S4 for a summary of the genomic coding system. This particular coding system for transcriptomic and metabolomic data is consistent with the classical coding for hybrid genotypes. The biological justification for the "dominance" is that it may capture information about the difference between the two parents. Since the difference can be positive or negative, the absolute value will make the difference positive in either case.

There are two matrices, Z and *W*, for the IMF2 crosses. Corresponding to the two matrices, there are two types of effects, α represents the additive effects and δ represents the dominance effects. The *X* matrix for the IMF2 crosses takes the horizontal concatenation of the two matrices, $X=[Z\,||\,W]$. Let $\beta=[\alpha\,//\,\delta]$ be the vertical concatenation of α and δ. The model for prediction is $y=X\beta+\varepsilon$ or $y=f(X\,|\,\beta)+\varepsilon$ depending on the prediction method. We performed centering and scaling for the

This article is protected by copyright. All rights reserved.

predictors by calling the scale() function in R. For the BLUP method, two kinship matrices were fitted to the model, one for the additive variance and one for the dominance variance. The kinship matrices were calculated using the method of Xu (Xu 2013).

Combining three sources of data. The model in the combined analysis was

$$y = X_{SNP}\beta_{SNP} + X_{EXP}\beta_{EXP} + X_{MET}\beta_{MET} + \varepsilon \quad (2)$$

for the linear methods (LASSO, BLUP, SSVS and PLS), where SNP and EXP and MET denote the three omic data sources. For the non-linear methods (SVM-RBF and SVM-POLY), the model was

$$y = f(X \mid \beta) + \varepsilon \quad (3)$$

where $X = [X_{SNP} \,\|\, X_{EXP} \,\|\, X_{MET}]$ is the column concatenation of the three predictor matrices and $\beta = [\beta_{SNP} \,//\, \beta_{EXP} \,//\, \beta_{MET}]$ is the row concatenation of the effects of the three predictors. The BLUP method was based on a linear model in an implicit manor. It explicitly requires three covariance structures (called kinship matrices), one for each data source (dominance covariance structures have been ignored).

Analysis of variance of predictability. We performed an ANOVA under a $3\times4\times6$ factorial design with three predictors (omic data), four traits and six methods. The linear model for the predictability (*P*) is

$$P_{ijk} = \mu + O_i + T_j + M_k + (OT)_{ij} + (OM)_{ik} + (TM)_{jk} + E_{ijk} \quad (4)$$

This article is protected by copyright. All rights reserved.

Accepted Article

where μ is the grand mean, O_i is the effect of the *i*th omic predictor ($i = 1, 2, 3$), T_j is the *j*th trait ($j = 1, ..., 4$), M_k is the *k*th method ($k = 1, ..., 6$) and E_{ijk} is the residual error.

Permutation test for predictability. To test whether or not the observed predictability is significantly different from zero, we randomly shuffled the phenotypes and conducted a prediction under each scenario with the LASSO method. The shuffling process was replicated 1000 times so that the predictabilities formed a null distribution. We then compared the observed predictability against the null distribution to calculate an empirical *p*-value for each predictability.

Accession codes

These data have been deposited in figshare (figshare.com/s/0773080c122d11e58b6306ec4bbcf141).

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

This article is protected by copyright. All rights reserved.

Accepted Article

Acknowledgements

The project was supported by National Science Foundation Collaborative Research Grant DBI-1458515 to SX, Chinese National Natural Science Foundation Grant 31330039 to QZ and China's 111 Project Grant B07041 to QZ.

Supporting Information

Figure S1. Comparison of prediction from combined data with separate analyses of individual omic data in the IMF2 population.

Figure S2. Distributions of the predicted hybrid yield (upper panel) and tested hybrid yield (lower panel)

Figure S3. Predictabilities of four traits from three omic data and six methods in the RIL population.

Figure S4. Cross-validation generated predictabilities of YIELD from metabolomic data using the LASSO method.

Table S1. Heritability of four yield-related traits estimated from both the IMF2 (hybrid) and RIL (parent) populations.

Table S2. Comparison of predictability of traits in hybrids using 683 metabolites of leaves and 317 metabolites of seeds measured from parents.

Table S3. A toy example of a balanced random partial rectangle cross design (BRPRCD) from 8 male (A-H) and 14 female (I-V) lines.

This article is protected by copyright. All rights reserved.

Table S4. The classical coding system of SNP genotypes for a hybrid determined by the two inbred parents.

Data S1. Analyses of variance (ANOVA) of predictability for three predictors, four traits and six methods from the hybrid population.

Data S2. Predicted yields for all 21945 potential crosses derived from 210 inbred lines from three omic data sources.

Data S3. Metabolites with significant effects on yield detected from the LASSO method

Appendix S1

References

Bernardo, R. (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science*, **34**, 20-25.

Bing, N. and Hoeschele, I. (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics*, **170**, 533-542.

Bino, R.J., Hall, R.D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B.J., Mendes, P., Roessner-Tunali, U., Beale, M.H., Trethewey, R.N., Lange, B.M., Wurtele, E.S. and Sumner, L.W. (2004) Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.*, **9**, 418-425.

Carreno-Quintero, N., Bouwmeester, H.J. and Keurentjes, J.J.B. (2013) Genetic analysis of metabolome–phenotype interactions: from model to crop species. *Trends Genet.*, **29**, 41-50.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H., Dong, H., Zhang, W., Zhang, L., Yu, S., Wang, G., Lian, X. and Luo, J. (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.*, **46**, 714-721.

Dunn, W.B. and Ellis, D.I. (2005) Metabolomics: Current analytical platforms and methodologies. *Trends Analyt Chem*, **24**, 285-294.

Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics* 4 edn. Harlow, Essex, UK: Addison Wesley Longman.

Feher, K., Lisec, J., Römisch-Margl, L., Selbig, J., Gierl, A., Piepho, H.-P., Nikoloski, Z. and Willmitzer, L. (2014) Deducing Hybrid Performance from Parental Metabolic Profiles of Young Primary Roots of Maize by Using a Multivariate Diallel Approach. *PLoS One*, **9**, e85435.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, **33**, 1-22.

Frisch, M., Thiemann, A., Fu, J., Schrag, T.A., Scholten, S. and Melchinger, A.E. (2010) Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. *Theor. Appl. Genet.*, **120**, 441-450.

Fu, J., Falke, K.C., Thiemann, A., Schrag, T.A., Melchinger, A.E., Scholten, S. and Frisch, M. (2012) Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. *Theor Appl Genet*, **124**, 825-833.

This article is protected by copyright. All rights reserved.

Accepted Article

Gärtner, T., Steinfath, M., Andorf, S., Lisec, J., Meyer, R.C., Altmann, T., Willmitzer, L. and Selbig, J. (2009) Improved heterosis prediction by combining information on DNA-and metabolic markers. *PLoS One*, **4**, e5220.

Gelandi, P. and Kowalski, B.R. (1986) Partial Least-Squares Regression: A tutorial. *Anal. Chim. Acta*, **185**, 1-17.

George, E.I. and McCulloch, R.E. (1993) Variable selection via Gibbs sampling. *J Am Stat Assoc*, **88**, 881-889.

Gong, L., Chen, W., Gao, Y., Liu, X., Zhang, H., Xu, C., Yua, S., Zhang, Q. and Luo, J. (2013) Genetic analysis of the metabolome exemplified using a rice population. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 20320-20325.

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

Houle, D., Govindaraju, D.R. and Omholt, S. (2010) Phenomics: the next challenge. *Nature Reviews Genetics*, **11**, 855–866.

Hua, J., Xing, Y., Wu, W., Xu, C., Sun, X., Yu, S. and Zhang, Q. (2003) Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc Natl Acad Sci USA*, **100**, 2574-2579.

Hua, J.P., Xing, Y.Z., Xu, C.G., Sun, X.L., Yu, S.B. and Zhang, Q. (2002) Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. *Genetics*, **162**, 1885-1895.

This article is protected by copyright. All rights reserved.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004) Kernlab - an S4 package for kernel Methods in R. *Journal of Statistical Software*, **11**, 1-20.

Keurentjes, J.J., Fu, J., de Vos, C.H., Lommen, A., Hall, R.D., Bino, R.J., van der Plas, L.H., Jansen, R.C., Vreugdenhil, D. and Koornneef, M. (2006) The genetics of plant metabolism. *Nat. Genet.*, **38**, 842-849.

Keurentjes, J.J.B. (2009) Genetical metabolomics: closing in on phenotypes. *Curr. Opin. Plant Biol.*, **12**, 223-230.

Keurentjes, J.J.B., Fu, J., Terpstra, I.R., Garcia, J.M., van den Ackerveken, G., Snoek, L.B., Peeters, A.J.M., Vreugdenhil, D., Koornneef, M. and Jansen, R.C. (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 1708-1713.

Lisec, J., Steinfath, M., Meyer, R.C., Selbig, J., Melchinger, A.E., Willmitzer, L. and Altmann, T. (2009) Identification of heterotic metabolite QTL in Arabidopsis thaliana RIL and IL populations. *Plant J.*, **59**, 777-788.

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

Mevik, B.-H. and Wehrens, R. (2007) The pls Package: Principal Component and Partial Least Squares Regression in R. *J Stat Softw*, **18**, 1-24.

Meyer, R.C., Steinfath, M., Lisec, J., Becher, M., Witucka-Wall, H., Törjék, O., Fiehn, O., Eckardt, Ä., Willmitzer, L., Selbig, J. and Altmann, T. (2007) The metabolic signature related to high plant growth rate in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, **104**, 4759-4764.

This article is protected by copyright. All rights reserved.

Accepted Article

Perez, P. and de los Campos, G. (2014) Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics*, **198**, 483-495.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L. and Melchinger, A.E. (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature genetics*, **44**, 217-220.

Rockman, M.V. and Kruglyak, L. (2006) Genetics of global gene expression. *Nat. Rev. Genet.*, **7**, 862-872.

Rowe, H.C., Hansen, B.G., Halkier, B.A. and Kliebenstein, D.J. (2008) Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. *Plant Cell*, **20**, 1199-1216.

Schauer, N. and Fernie, A.R. (2006) Plant metabolomics: towards biological function and mechanism. *Trends Plant Sci.*, **11**, 508-516.

Steinfath, M., Strehmel, N., Peters, R., Schauer, N., Groth, D., Hummel, J., Steup, M., Selbig, J., Kopka, J., Geigenberger, P. and Van Dongen, J.T. (2010) Discovering plant metabolic biomarkers for phenotype prediction using an untargeted approach. *Plant Biotechnol. J.*, **8**, 900-911.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol*, **58**, 267-288.

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

Wang, J., Yu, H., Weng, X., Xie, W., Xu, C., Li, X., Xiao, J. and Zhang, Q. (2014) An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. *J. Exp. Bot.*, **65**, 1069-1079.

This article is protected by copyright. All rights reserved.

Accepted Article

West, M.A.L., Kim, K., Kliebenstein, D.J., van Leeuwen, H., Michelmore, R.W., Doerge, R.W. and St Clair, D.A. (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics*, **175**, 1441-1450.

Xiang, X., Ning, S. and Wei, D. (2013) Proteomic profiling of rice roots from a super hybrid rice cultivar and its parental lines. *Plant Omics Journal*, **6**, 318-324.

Xing, Y., Tan, F., Hua, J., Sun, L., Xu, G. and Zhang, Q. (2002) Characterization of the main effects, epistatic effects and their environmental interactions of QTLs on the genetic basis of yield traits in rice. *Theor. Appl. Genet.*, **105**, 248-257.

Xu, S. (2013) Mapping quantitative trait loci by controlling polygenic background effects. *Genetics*, **195**, 1209-1222.

Xu, S., Zhu, D. and Zhang, Q. (2014) Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proceedings of the National Academy of Sciences*, **111**, 12456-12461.

Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., Xiao, J. and Zhang, Q. (2011) Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS One*, **6**, e17595. doi:17510.11371/journal.pone.0017595.

Zhang, L., Peng, Y., Wei, X., Dai, Y., Yuan, D., Lu, Y., Pan, Y. and Zhu, Z. (2014) Small RNAs as important regulators for the hybrid vigour of super-hybrid rice. *J. Exp. Bot.*, **65**, 5989-6002.

Zhou, G., Chen, Y., Yao, W., Zhang, C., Xie, W., Hua, J., Xing, Y., Xiao, J. and Zhang, Q. (2012) Genetic composition of yield heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 15847-15852.

This article is protected by copyright. All rights reserved.

Accepted Article

Figure legends

Figure 1. Predictabilities of four traits from three omic data and six methods in the IMF2 population. The four traits are labeled as YIELD, KGW, GRAIN and TILLER. The three omic data are genomic, transcriptomic and metabolomic data. The six statistical methods are LASSO, BLUP, SSVS, SVM-RBF, SVM-POLY and PLS.

Figure 2. Multiple comparisons illustrated by boxplots. In each boxplot, the line in the middle of the box represents the median denoted by Q2 (50%). The open diamond in each box indicates the mean. The upper and lower ages of the box represent Q3 (75%) and Q1 (25%) of the sample, respectively. The whiskers define Q3 + 1.5×IQR and Q1 - 1.5×IQR, where IQR is the interquartile range. The small open circles represent outliers. The top panel (a) compares the means of predictability for the three omic data on average over all four traits and six methods. The capital letters above the group labels represent the test results, with different letters representing significant difference between groups. For example, transcriptomic prediction (A) is significantly better than genomic prediction (B), but metabolomic prediction (AB) is not significantly different from either of the other two predictions. The panel in the middle (b) compares the mean predictabilities of the four traits on average across all three omic data and six methods. The panel at the bottom (c) compares the mean predictabilities of all six methods on average over all three omic data and four traits.

Figure 3. The null distributions of predictabilities obtained from the LASSO prediction. The dark triangle in each panel represents the observed predictability, which is far beyond the null distribution. The null distribution was obtained from 1000 permuted samples by randomly shuffling the phenotype. Each column of the figure represents a trait (Yield, KGW, Grain or Tiller) and each row of the figure represents a data source (Genome, Transcriptome or metabolome).

This article is protected by copyright. All rights reserved.

Accepted Article

Figure 4. Average predicted yield of top crosses selected for breeding (left *y*-axis) and percent gain in yield of hybrid selection (right *y*-axis) from metabolomic, transcriptomic and genomic data using the LASSO method. The average yield of all 21945 potential crosses is 43.6 (the starting value of the left *y*-axis).

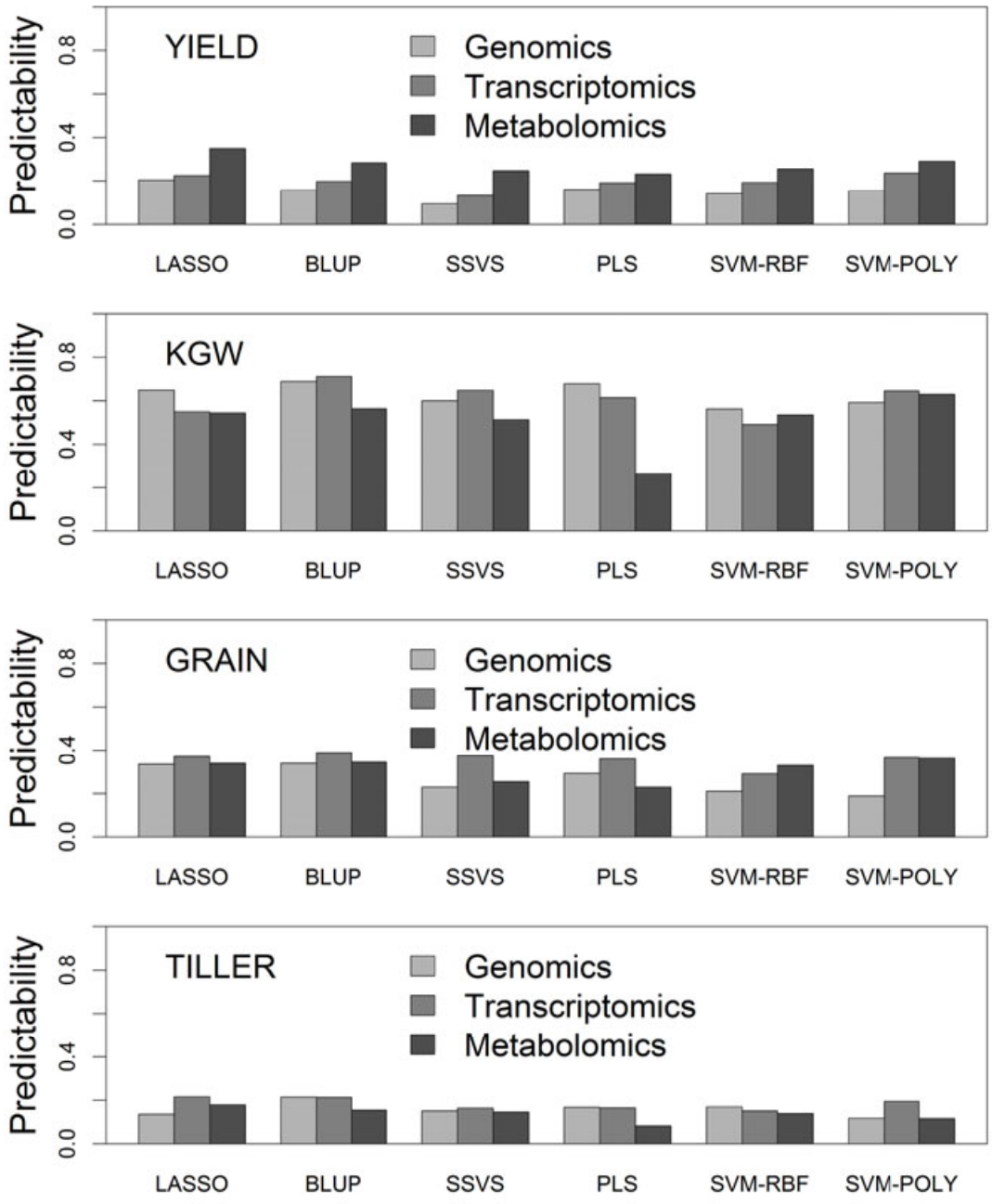
Table 1 Analyses of variances of predictability from a $3\times4\times6$ factorial design with three predictors (omic data), four traits and six prediction methods

Source	DF	Sum of Square	Mean Square	F-test	p-value
Predictor	2	0.0174	0.0087	4.81	0.0155
Trait	3	1.9265	0.6421	354.64	<0.0001
Method	5	0.0508	0.0102	5.62	0.0009
Method×Predictor	10	0.0648	0.0065	3.58	0.0032
Method×Trait	15	0.0321	0.0021	1.18	0.3350
Predictor×Trait	6	0.1130	0.0188	10.41	<0.0001
Residual[a]	30	0.0543	0.0018		

[a] Residual is the predictor×method×trait interaction, whose mean square is the denominator for all the F tests.

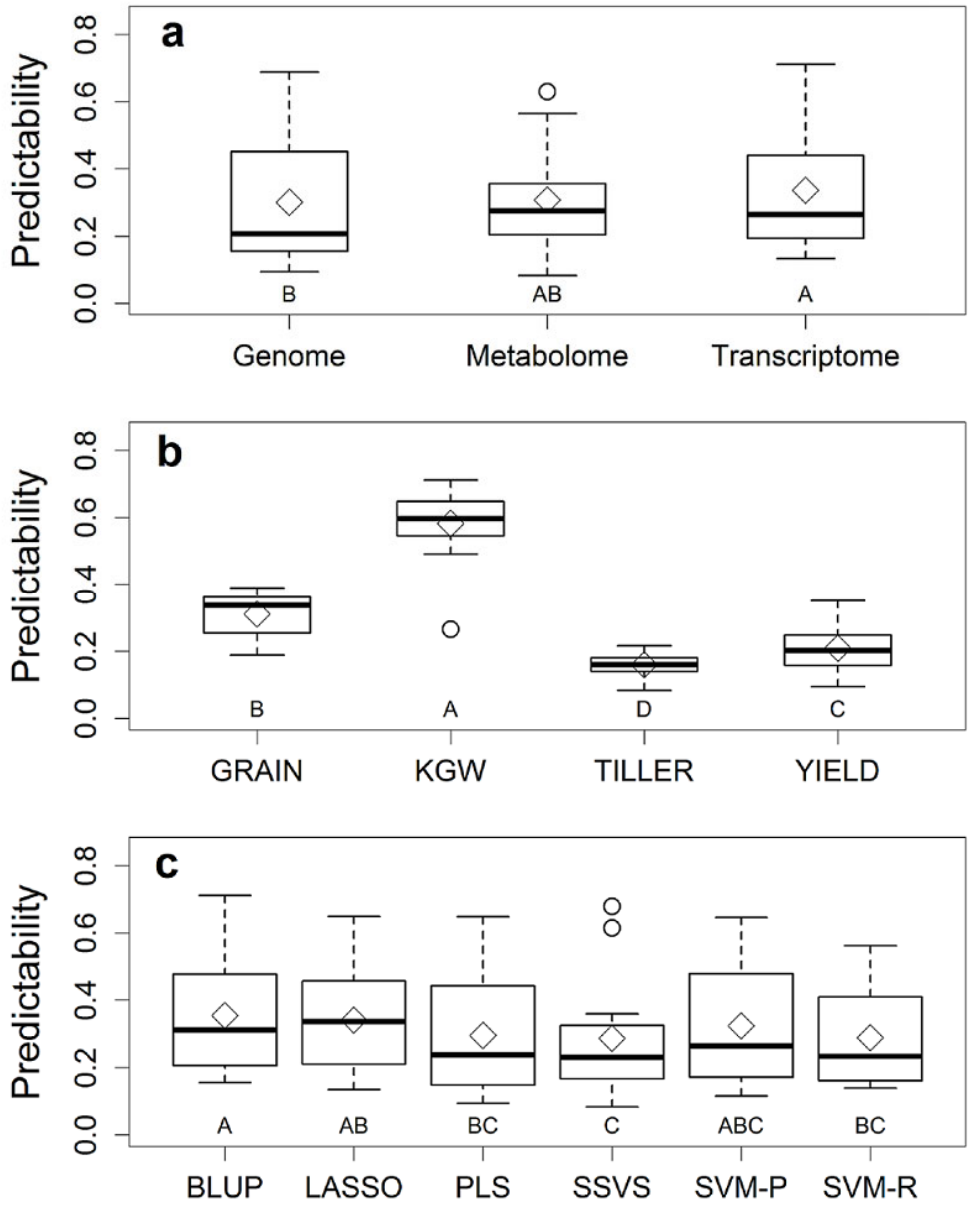
This article is protected by copyright. All rights reserved.

Accepted Article



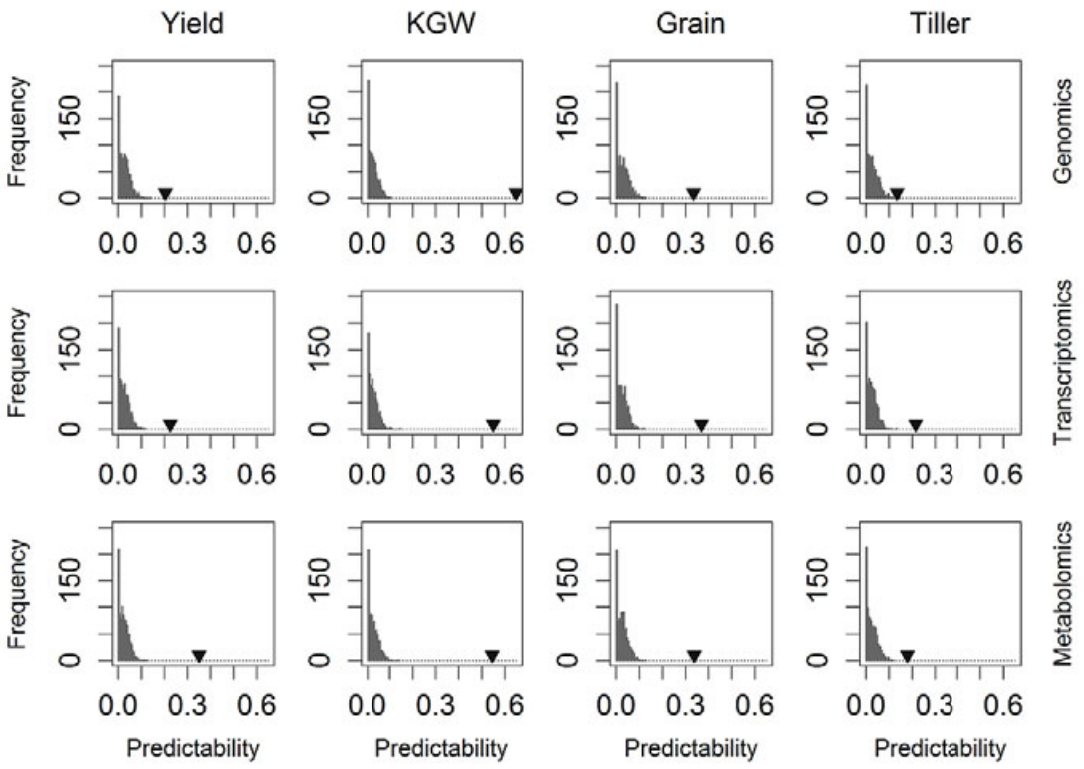
This article is protected by copyright. All rights reserved.

Accepted Article



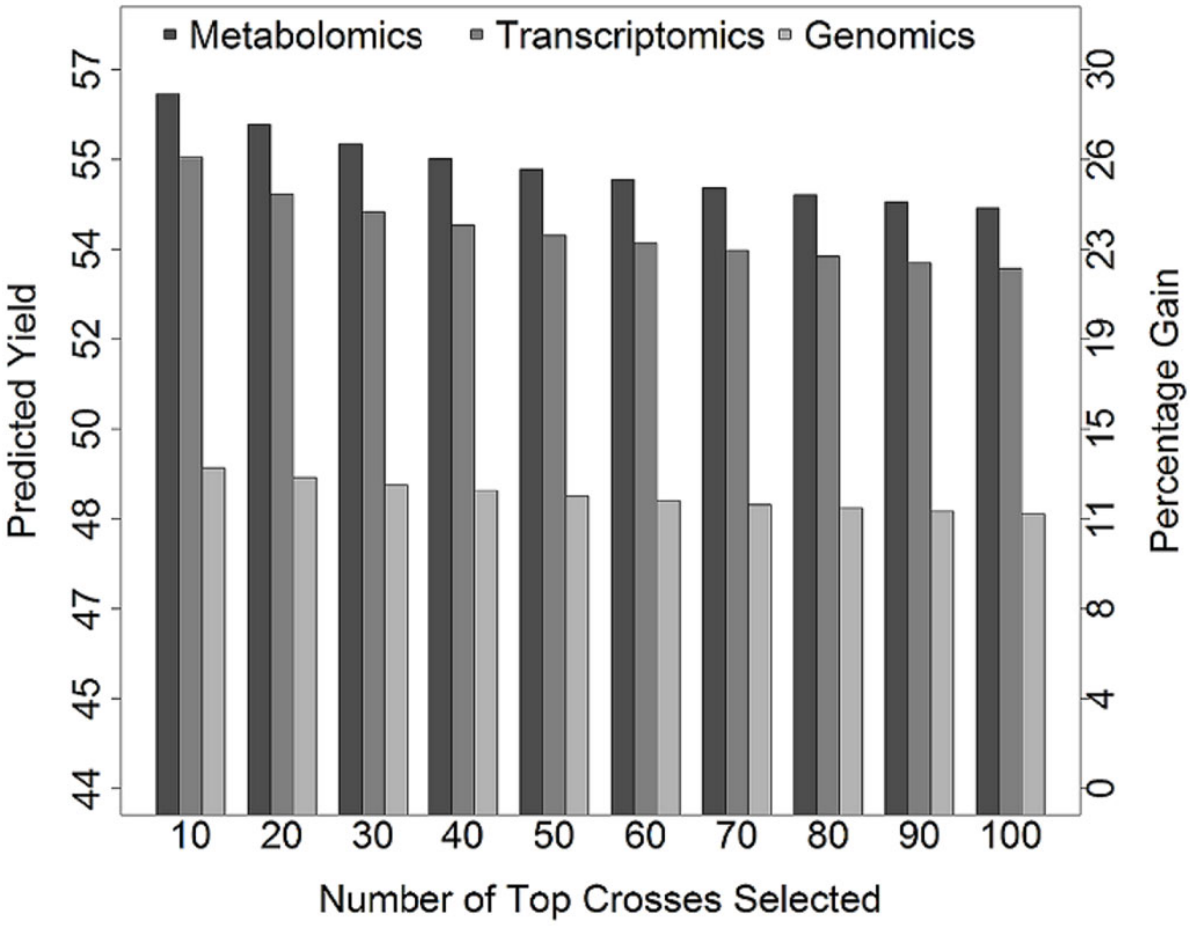
This article is protected by copyright. All rights reserved.

Accepted Article



This article is protected by copyright. All rights reserved.

Accepted Article



This article is protected by copyright. All rights reserved.