Heredity (2018) 120:342–355
https://doi.org/10.1038/s41437-017-0032-3

the geneticssociety

ARTICLE

Metabolome-wide association studies for agronomic traits of rice

Julong Wei[1,2] · Aiguo Wang[3] · Ruidong Li[1] · Han Qu[1] · Zhenyu Jia[1]

Received: 30 July 2017 / Revised: 6 November 2017 / Accepted: 7 November 2017 / Published online: 11 December 2017

Abstract

Identification of trait-associated metabolites will advance the knowledge and understanding of the biosynthetic and catabolic pathways that are relevant to the complex traits of interest. In the past, the association between metabolites (treated as quantitative traits) and genetic variants (e.g., SNPs) has been extensively studied using metabolomic quantitative trait locus (mQTL) mapping. Nevertheless, the research on the association between metabolites with agronomic traits has been inadequate. In practice, the regular approaches for QTL mapping analysis may be adopted for metabolites-phenotypes association analysis due to the similarity in data structure of these two types of researches. In the study, we compared four regular QTL mapping approaches, i.e., simple linear regression (LR), linear mixed model (LMM), Bayesian analysis with spike-slab priors (Bayes B) and least absolute shrinkage and selection operator (LASSO), by testing their performances on the analysis of metabolome-phenotype associations. Simulation studies showed that LASSO had the higher power and lower false positive rate than the other three methods. We investigated the associations of 839 metobolites with five agronomic traits in a collection of 533 rice varieties. The results implied that a total of 25 metabolites were significantly associated with five agronomic traits. Literature search and bioinformatics analysis indicated that the identified 25 metabolites are significantly involved in some growth and development processes potentially related to agronomic traits. We also explored the predictability of agronomic traits based on the 839 metabolites through cross-validation, which showed that metabolomic prediction was efficient and its application in plant breeding has been justified.

Introduction

Metabolites, collectively called metabolome, span a wide variety of chemical compounds that are intermediates and products of numerous linked biosynthetic and catabolic pathways within a biological system (Cacciatore and Loda 2015). Metabolites play important roles in organisms, for example, in the growth and development of organisms as well as in physiological response to environment. Metabolites may be treated as intermediate traits resulting from the genes that govern the phenotypes of interest (Chan et al. 2010; Fiehn 2002; Fiehn et al. 2000). Tons of intricate and interwoven metabolomic reactions are involved in manifesting the determinant genes to produce the resultant phenotypes. The incidental environmental input, which is independent of DNA, may be observed in both metabolites (intermediate phenotypes) and traits. Thus, a thorough study on metabolome and its relationship with genome, transcriptome and phenotypes will substantially advance the understanding of the genetic architectures and the underlying biological functions for quantitative traits, facilitating breeding programs. With the advanced technology, such as liquid chromatography-mass spectrometry (LC–MS), it becomes routine to obtain high throughput and reproducible quantitative metabolomic data. Recently, there have been many researches on metabolomics in humans (Draisma et al. 2015; Gieger et al. 2008; Rhee et al. 2013) as well as in plants, such as *Arabidopsis thaliana* (Lisec et al. 2008; Meyer et al. 2007; Steinfath et al. 2010), maize (Pace et al. 2015; Riedelsheimer et al. 2012a, 2012b; Wen et al. 2014) and rice (Chen et al. 2014; Gong et al. 2013; Matsuda et al. 2015, 2012; Xu et al. 2016).

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41437-017-0032-3) contains supplementary material, which is available to authorized users.

✉ Zhenyu Jia
zhenyuj@ucr.edu

[1] Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA, USA

[2] College of Animal Science and Technology, Nanjing Agricultural University, Nanjing, Jiangsu, China

[3] College of Animal Science and Technology, China Agricultural University, Beijing, China

As intermediate traits situated between determinant genes and phenotypes of interest, metabolites may well correlate with both genotypes and phenotypes. In the past, the association between metabolome and genome has been well studied using metabolomic quantitative trait locus (mQTL) analyses in which metabolites are treated as quantitative traits for mapping causative loci (Gong et al. 2013; Matsuda et al. 2012). Nevertheless, the association between metabolome and phenotypes has not been adequately addressed, except that simple correlation was used for detection of metabolite-phenotype associations (Meyer et al. 2007). Such simple correlation analysis may suffer a low level of detection power but a high level of false positive rate owing to the fact that these metabolites are analyzed separately. Moreover, there is no control for background effects due to other metabolites because only one metabolite is considered at a time, yielding false detections.

In the current study, we adopted the statistical methods that have been widely used in QTL mapping to perform metabolome-wide association study (MWAS). We compared four regression approaches, i.e., simple linear regression (LR), linear mixed model (LMM), Bayesian analysis with spike-slab priors (Bayes B) and least absolute shrinkage and selection operator (LASSO), by testing their performances on MWAS. Simple linear regression is similar to Pearson's correlation analysis where only one metabolite is considered at a time. An advantage of using the mixed model in MWAS is to provide an efficient control for spurious positives by adding polymetabolomic effects (random effect), which is analogous to polygenic effects in GWAS (Wang et al. 2016a, 2016b; Yu et al. 2006; Zhang et al. 2005). In Bayes B, association analysis is actually translated to a cluster analysis in which trait-associated metabolites and trait-irrelevant metabolites are assigned into different groups (Pérez and de Los Campos 2014). The least absolute shrinkage and selection operator (LASSO) is very efficient for handling the situation where the number of predictors is much larger than the number of observations in regression settings (Tibshirani 1996).

Another type of metabolites-based study is metabolomic prediction, where all metabolites are jointly used for predicting agronomic traits (Riedelsheimer et al. 2012a; Xu et al. 2016). In maize and rice, metabolomic prediction may outperform genomic prediction for the traits with low heritability, for example, grain yield (Xu et al. 2016; Xu et al. 2017). In this study, we also explored the feasibility of predicting phenotypes with metabolomic data using these four methods.

Cultivated rice is one of the most economically important crop that provides a staple food source for more than half of the world population, especially in Asia. In the past decade, understanding the genomes of rice and their functions to important traits has been substantially carried forward (Huang et al. 2010, 2016, 2012; Yano et al. 2016), especially when metabolomics data become available for system-wide analysis. To demonstrate the MWAS methods under consideration, we used a rice dataset in which 840 metabolites, about 3 million SNPs on rice genome, and five quantitative traits (i.e., yield, heading date, plant height, grain length and grain width) were measured for 524 rice varieties using the LC-MS technique (Chen et al. 2014). In the previous study, hundreds of genomic loci have been identified for 840 metabolites as well as agronomic traits (Chen et al. 2014). In the current study, we compared the performances of the four MWAS methods on detection of the association between these metabolites and five agronomic traits, respectively. We also studied the performances of the three methods, i.e., BLUP, Bayes B and LASSO, in predicting the genetic values for these five traits using metabolic data, justifying the application of metabolome in plant breeding. Note that LR is not suitable for prediction study.

Materials and methods

Materials collection

We used a rice data set (Chen et al. 2014) that include 533 rice (*Oryza sativa*) accessions collected from a wide range of geographical locations, including 200 varieties from a core/mini-core collection of Oryza sativa in China (Zhang et al. 2011), 132 parental lines used in the International Rice Molecular Breeding Program (Yu et al. 2003), 148 varieties from a mini-core subset of the US Department of Agricultural Rice Gene Bank (Yan et al. 2009), 18 varieties used for SNP discovery in the OryzaSNP project (McNally et al. 2009), and 35 varieties from the International Rice Research Institute (IRRI). We only selected 524 varieties that have both metabolite and phenotype data. There are two major subpopulations, *indica* (293 varieties) and *japonica* (155 varieties), indicated in Supplementary Fig. S1(A). For each variety, a total of five agronomic traits have been measured, including yield (YD), heading date (HD), plant height (PH), grain length (GL), and grain width (GW).

A total of 840 metabolites in leaves were measured using a LC-MS/MS platform. Here we only used 839 metabolites, with one invariable metabolite removed. For each variety of rice, two biological replicates were collected from two different experimental sites at Huazhong Agricultural University, Wuhan, China. Leaves were collected at the five-leaf stage from three different plants per line for each replicate. For each metabolite, the expression values from two biological replicates were averaged and then followed by a log transformation to generate a single expression value for that metabolite. We treated each metabolite as a regular trait and calculated the heritability for each metabolite using ANOVA method (Chen et al. 2014). The detailed estimation of heritability for a metabolite trait is provided in Supplementary Note S1. The Detailed information about the collection of metabolites and data management is available from the original study (Chen et al. 2014).

We also compared the result of metabolomic prediction with that of genomic prediction through cross validation. The raw genomic data are a total of about 6.7 billion 90-bp paired reads that were sequenced from the genomes of the rice varieties using the Illumina HiSeq 2000 platform. In combination with the sequences of 950 rice varieties from Huang et al. (2012), these raw reads were aligned to the rice reference genome (Nipponbare, MSU version 6.1) to obtain about 6.5 million high-quality SNPs. After removal of the missing genotypes, a total of 3,616,597 SNPs were eventually included in the study. The genetic relationship between the rice varieties may be well estimated using such highly saturated genomic data. All data, including agronomic traits, genotypes for SNP, and metabolites, were scaled to have mean of zero and unit variance.

Methods of association

Four linear regression models have been investigated in the study. The first model is a simple linear regression (LR) where the phenotypic value of an agronomic trait is described by

$$y = X\beta + Z_k\gamma_k + \varepsilon \tag{1}$$

where X is a design matrix for some systematic effects irrelevant to metabolites (e.g., locations, years, and so on), β is a vector of such systemic effects. When there are no systematic effects, this term will vanish. Independent variable Z_k is a vector of expression values of the kth metabolite and γ_k is the effect of the kth metabolite on the agronomic trait under study. The residual error ε is assumed to be normally distributed with $N(0, I\sigma^2)$. The Wald test, defined as follows, is used for testing the γ_k:

$$W_k = \frac{\hat{\gamma}_k^2}{\mathrm{var}(\hat{\gamma}_k^2)}. \tag{2}$$

Under the null model $H_0 : \gamma_k = 0$, the Wald test follows approximately a Chi-square distribution with one degree of freedom.

The second model is the linear mixed model (LMM), where a random effect is added to each observation to control the poly-metabolite effects. This is similar to the use of polygene in genome-wide association studies (GWAS). The model is described as

$$y = X\beta + Z_k\gamma_k + \xi + \varepsilon, \tag{3}$$

in which ξ is an $n \times 1$ vector of random effects with an assumed $N(0, K\phi^2)$ distribution where ϕ^2 is the variance of this random effect and K is covariance structure inferred from all metabolomic data as follows,

$$K = \frac{1}{d}\sum_{k=1}^{m} Z_k Z_k^T \tag{4}$$

where d is the average value of the diagonal elements of matrix K (a normalization factor). The K is analogous to the kinship matrix that is used in GWAS for capture of genetic structure among individuals. Under the mixed model, the expectation of y is $\mathrm{E}(y) = X\beta + Z_k\gamma_k$ and the variance is

$$\mathrm{var}(y) = V = K\phi^2 + I\sigma^2 \tag{5}$$

The restricted maximum likelihood (REML) method was used to estimate the variance components (ϕ^2 and σ^2). The eigen-decomposition algorithm was implemented when the restricted likelihood function was evaluated to reduce the computational complexity (Kang et al. 2008). The best linear unbiased estimates (BLUE) for the fixed effects are obtained using

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma}_k \end{bmatrix} = \begin{bmatrix} X^T V^{-1} X & X^T V^{-1} Z_k \\ Z_k^T V^{-1} X & Z_k^T V^{-1} Z_k \end{bmatrix}^{-1} \begin{bmatrix} X^T V^{-1} y \\ Z_k^T V^{-1} y \end{bmatrix} \tag{6}$$

Similar to the LR analysis, the Wald test was conducted to test $H_0 : \gamma_k = 0$.

The first two linear models are "single-effect" models because the associations between metabolites and traits are tested one metabolite at a time. The next two approaches, the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996) and Bayes B (Pérez and de Los Campos 2014), are multiple-effect models, in which all metabolites are jointly analyzed. These two approaches are particularly efficient for variable selection or dimensionality reduction.

When using the LASSO method, we first remove the fixed effects from the model by centering the phenotypic values, yielding the following model.

$$y = X\beta + \sum_{k=1}^{m} Z_k\gamma_k + \varepsilon \tag{7}$$

All γ_k's are estimated by minimizing the sum of squares of residuals with a restriction that the sum of the absolute values of all estimated γ_k is less than a predetermined constant, which is mathematically equivalent to

$$\gamma^{LASSO} = \arg\min_{\gamma}\left\{(y - Z\gamma)^T(y - Z\gamma) + \lambda\sum\nolimits_{k=1}^{m}|\gamma_k|\right\} \tag{8}$$

where λ is a shrinkage factor obtained via cross-validation. Note that with the special L_1 restriction, the values for majority of γ_k's are in shrinkage towards zero if they contribute little to the trait. Therefore, in the LASSO method, variable selection is accomplished by eliminating metabolites with trivial effect from the regression model. After variable selection, the number of non-zero effects is always smaller than the sample size; thereafter, the regular least squares methods may apply. We used the GLMNET/R package for LASSO computation (Tibshirani 1996); however, GLMNET/R package does not provide the variance for an estimated effect (γ_k). To facilitate a Wald test aforementioned, we calculated an empirical variance for each estimated γ_k using bootstrap of size 1000. Thus, the variance of an estimated effect is simply the variance of the 1000 estimated values for γ_k through bootstrap analysis.

For the Bayes B method, similar regression model is used to describe the relationship between a trait and all metabolites (Equation 7). In Bayes B, a mixture distribution is assigned to each γ_k as follows,

$$\gamma_k \sim \pi N(0, \sigma_k^2) + (1-\pi)N(0,0) \tag{9}$$

where π is a prior probability that a metabolite is selected (included in the model), σ_k^2 is the variance of the normal distribution from which a nontrivial γ_k is sampled, whereas $N(0,0)$ is a probability mass at zero (equivalent to a normal distribution with zero mean and zero variance). The binary variable δ_k takes 1 if $\gamma_k \sim N(0,\sigma_k^2)$ or 0 if $\gamma_k \sim N(0,0)$. Similar to variable selection in LASSO, Bayes B analysis eliminates metabolites with trivial effects by placing them in the normal distribution with zero variance in the mixture distribution. Posterior probability of $\delta_k = 1$ is calculated for each metabolite effect (γ_k) using Bayesian setting, and a γ_k is declared as significant if its posterior probability is greater than 0.95, a nominal cutoff that has been widely accepted. In the study, the BGLR/R package was used to perform the Bayes B analysis (Pérez and de Los Campos 2014).

As in GWAS, population structure (*indica* vs. *japonica*) effects can be incorporated into the MWAS linear models. Let Q and K be population structure and metabolites covariance structure (equivalent to kinship in GWAS), respectively. We used LR(Q), LMM(Q + K), LASSO(Q) and Bayes B(Q) to denote the ancillary structures that have been utilized in four linear models. Note that only LMM can take advantage of using metabolites covariance structure (K).

Determinant of critical value for a test

Since a metabolite is tested at a time in the two single-effect models (LR and LMM), we identified the empirical critical value for the test statistic ($-\log_{10}(p)$) using 1000 permuted samples to deal with the inflated type I errors due to multiple comparisons. Specifically, we randomly shuffled the phenotypic values of the 524 rice accessions for each permuted sample such that the associations between the phenotype and metabolites were completely wiped out. The permuted dataset was then analyzed by LR and LMM to obtain a test statistic for each of the metabolites, and the maximum of the test statistics was saved for this permuted sample. Such permutation was repeated by 1000 times and the 1000 maximum test statistics from the 1000 permuted samples form a *null* distribution, of which the 95 percentile was used as the empirical critical value for testing each metabolite in the original data. The critical values of the test statistic ranged from 3.95 to 4.23 across the five traits for LR and LMM. Therefore, we chose to use 4.0 as an approximately empirical critical value for both LR and LMM across all 5 traits. Note that the Bonferroni corrected critical value is $-\log_{10}(0.05/839) \approx 4.22$, which is very close to the selected empirical critical value. Since all the metabolites are tested simultaneously in the two multiple-effect models (LASSO and Bayes B), no multiple-test correction is needed and thus the nominal critical value, i.e., $-\log_{10}(0.05) = 1.30$ was used for metabolite detection.

Simulation studies

In order to retain the ancillary structures (population structure and metabolites covariance structure) of the original rice data, we adopted the expression values of the 839 metabolites for the 524 rice varieties to generate the pseudo phenotypes as follows. The metabolites were placed in two categories, i.e., trait-relevant metabolites and trait-neutral metabolites, based on the results from the analysis of original rice data using LASSO. In simulation, we randomly selected a metabolite from the category of trait-neutral metabolites and placed a non-zero effect to this metabolite. The pseudo phenotypes are simply the sum of the original phenotype and its deviation due to the manipulated metabolite. Note that the simulation only involved the single selected metabolite; minimum change has been introduced to the original data for generation of the simulated data such that the conclusions made by simulation study can be applicable to the situation when real data are analyzed. In simulation, we only used the heading date (HD) for demonstration. We examined 14 scenarios in which we let the heritability of the selected metabolite (ratio of the variance due the selected metabolite and the total phenotypic variance) be 0.00, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, and 0.20, respectively. For each scenario, the simulation was replicated 500 times, from which, the average power and Type 1 error rate were computed. The power is defined as the proportion of trait-relevant metabolites that have been successfully detected, whereas the Type 1 error rate is defined as the proportion of the trait-neutral metabolites that have been incorrectly claimed.

Methods of prediction

Three of the four methods, i.e., LMM, LASSO, and Bayes B, can be used for prediction of the trait using metabolites. Note that LR can only be used for phenotype-metabolite association detection. To make prediction, we partitioned the sample into a training sample and a test sample. For LMM, the model parameters were first estimated from the training sample, and then they were used to predict the phenotypic values of the test sample. Let y_1 and y_2 be the trait values of the training and test samples, respectively. The polymetabolic effect is similarly partitioned, and do is the kinship matrix. The predicted trait value (a vector) for individuals in the test sample is expressed as

$$\hat{y}_2 = X_2\hat{\beta} + \hat{\phi}^2 K_{21}(K_{11}\hat{\phi}^2 + I\hat{\sigma}^2)^{-1}(y_1 - X_1\hat{\beta}). \quad (10)$$

Predictability, a measure of the accuracy level of the prediction, is defined as the correlation coefficient between the predicted phenotypic values and the observed phenotypic values:

$$r_{y\hat{y}} = \frac{\text{cov}(y_2, \hat{y}_2)}{\sqrt{\text{var}(y_2)\text{var}(\hat{y}_2)}} \quad (11)$$

Since there is no independent test set, we performed a 10-fold cross-validation (CV) on the rice data. The rice data were partitioned into 10 portions. In each fold cross-validation, we used 9 portions for training the model and then used the resultant model to predict the phenotypic values or the remaining 1 portion. Eventually, each portion of the data has been used once for test, and each individual in the dataset has a predicted phenotypic value. The predictability depends on how the sample was partitioned. Therefore, we replicated the 10-fold cross-validation for 50 times and the average of the predictability was reported for both metabolomics predictors and genomic predictors. Similar 10-fold cross-validations were also conducted for the LASSO and Bayes B methods.

We developed an R pipeline to implement the MWAS and metabolic prediction. The source code of the pipeline is available from github (https://github.com/JulongWei/MWAS).

Results

Analysis of rice data

We first performed principal component analysis (PCA) for the 524 rice varieties using the 839 metabolites. The PC plot based on the first two principal components is shown in Supplementary Fig. S1(A), with two subpopulations (*indica* and *japonica*) well separated, which is similar to the PC plot using genomic data (Supplementary Fig. S2). The undefined accessions, which represent the genetic admixture of *indica* and *japonica*, are distributed between the two subpopulations (Chen et al. 2014). In the subsequent MWAS, we included the first three principal components in the regression models to capture the population structure. We also compared the pairwise correlations between metabolites using 524 rice accessions. The results are shown in the heat map (Supplementary Fig. S1(B)). We scrutinized a total of 351,541 pairs of metabolites of which about 20% pairs are negative correlation and the remaining pairs are positive correlation. Only 0.2% pairs (689) had strong correlation ($|r| \geq 0.8$), and a large proportion (57%) are considered to have weak correlation ($|r| \leq 0.2$), where r is the Pearson's correlation coefficient.

We first analyzed the five traits using the LR model with and without population structure, where LR(Q) denotes the LR analysis with population structure and LR denotes the LR analysis without population structure. The results are illustrated in Fig. 1, with results of LR in left panels and the results of LR(Q) in the right panels. In regard to yield (YD), if population structure was not considered (panel **a**), there are about 300 significant metabolites many of which appear to be false positives. However, with the correction of population structure (panel **b**), the number of significant metabolites dropped to 20. The similar results have been observed for GW too (panel **i** vs. panel **j**). For the other three traits (HD, PH and GL), the analyses with population structure (panels **d**, **f**, and **h**) or without population structure (panels **c**, **e**, and **g**) gave similar results, i.e., both analyses claimed too many significant metabolites, indicating a lack of control for false discoveries.

The results from analyzing five traits (YD, HD, PH, GL, and GW) using the other three methods (LMM, LASSO, and Bayes B) are given in Supplementary Fig. S3, Supplementary Fig. 2 and Figure S4, respectively. In each figure, left panels showed the results from the analysis without ancillary structures (Q or K) and right panels showed the results from the analysis with the ancillary structure. In LMM analysis, the test statistics ($-\log_{10}(p)$) were substantially reduced compared to LR analyses, with only 2 metabolites detected to be associated with HD and GW traits. In LASSO and Bayes B, the estimated effects for the majority of the metabolites were close to zero, reflecting strong shrinkage property of these two approaches. A total of 20 metabolites have been identified to be associated with five traits by LASSO, whereas only 2 metabolites were detected for the two traits (HD and GW) by Bayes B. Compared to the LR or LR(Q), the other three methods (LMM, LASSO, and Bayes B) provide a good control for false positives. For these three methods, there is no significant difference between the analyses with and without considering population structure. LASSO seemed to be a better method in

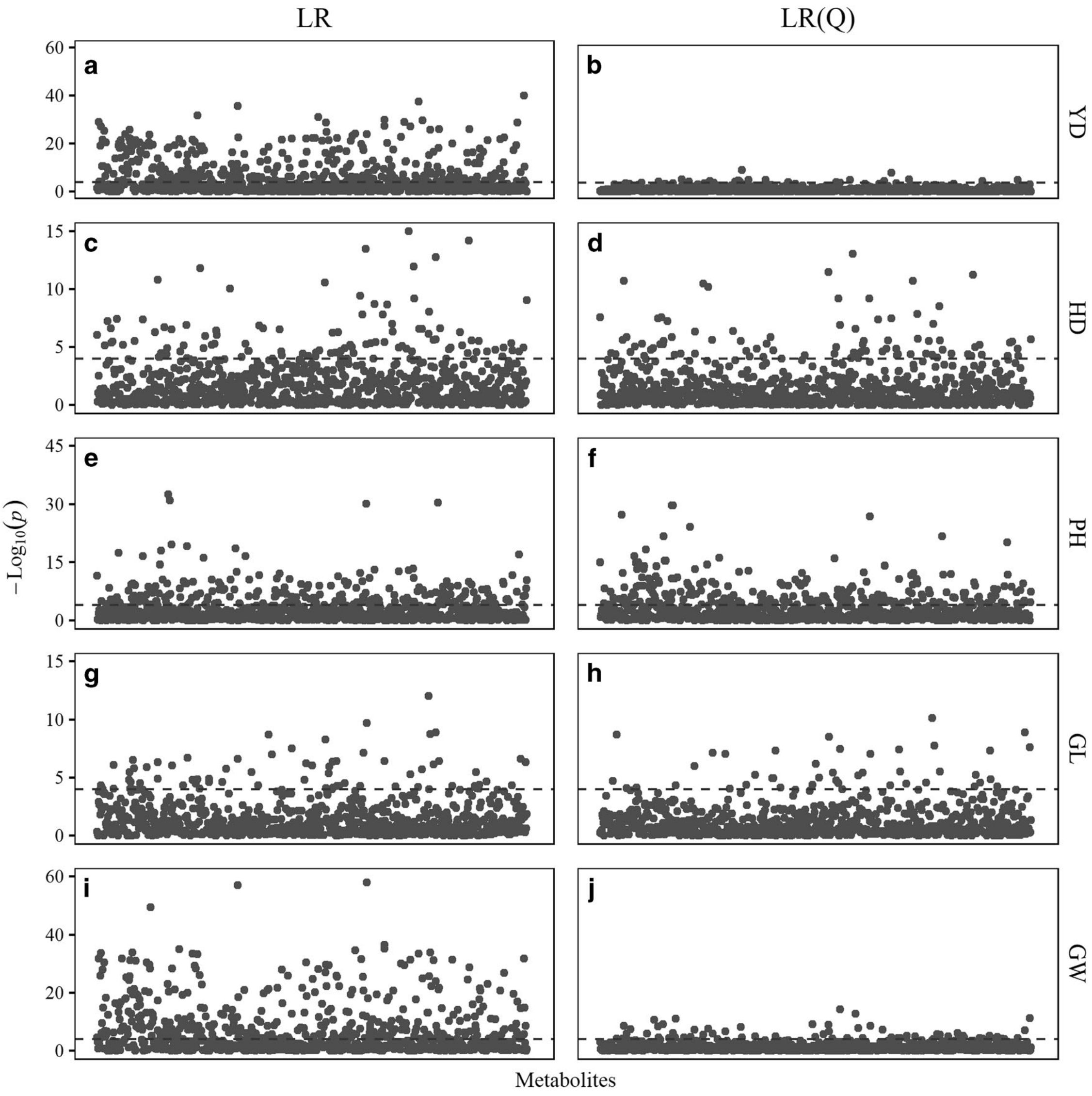

Fig. 1 Metabolome-wide association studies for the five traits from LR method. Left panels (**a, c, e, g,** and **i**) show results for LR without population structure and right panels (**b, d, f, h** and **j**) show results for LR with population structure, denoting LR(Q). The horizontal dashed lines are the permutation generated critical values for the $-\log_{10}(p)$ test statistics

detection of the phenotype-metabolite association, with a higher power and a lower false positive rate.

For these four methods, we also checked the Q-Q plots (the estimated $-\log(p)$ vs. the expected $-\log(p)$) as described in literatures (Huang et al. 2010, 2012). Supplementary Fig. S5 shows that the Q-Q plots for LR are far removed from the diagonal line and the estimated $-\log(p)$ values are consistently larger than the expected $-\log(p)$ values, indicating serious false detections. The Q-Q plots for LMM (Supplementary Fig. S6) suggest that the inclusion of polymetabolomic term can effectively control false positives. We also show the Q-Q plots of the two multiple-effects models, LASSO (Supplementary Fig. S7) and Bayes B (Supplementary Fig. S8). Supplementary Fig. S7 shows most of points are distributed around the expected line except the points with large logarithm p value, suggesting LASSO can control false positive as effectively as LMM. Supplementary Fig. S8 shows that Bayes B has lower estimated $-\log(p)$ values than expected ones, probably indicating Bayes B having low detection power. Our results showed that Bayes B has the same detection power as LMM because we have used permutation to further control family-wise type I errors (the critical value is for LMM), whereas the critical value used for Bayes B is $-\log_{10}(0.05) = 1.30$.

Comparisons based on simulated data

We compared the statistical properties of the four methods (LR, LMM, LASSO, and Bayes B) through intensive simulation experiments, in which the simulated heritability

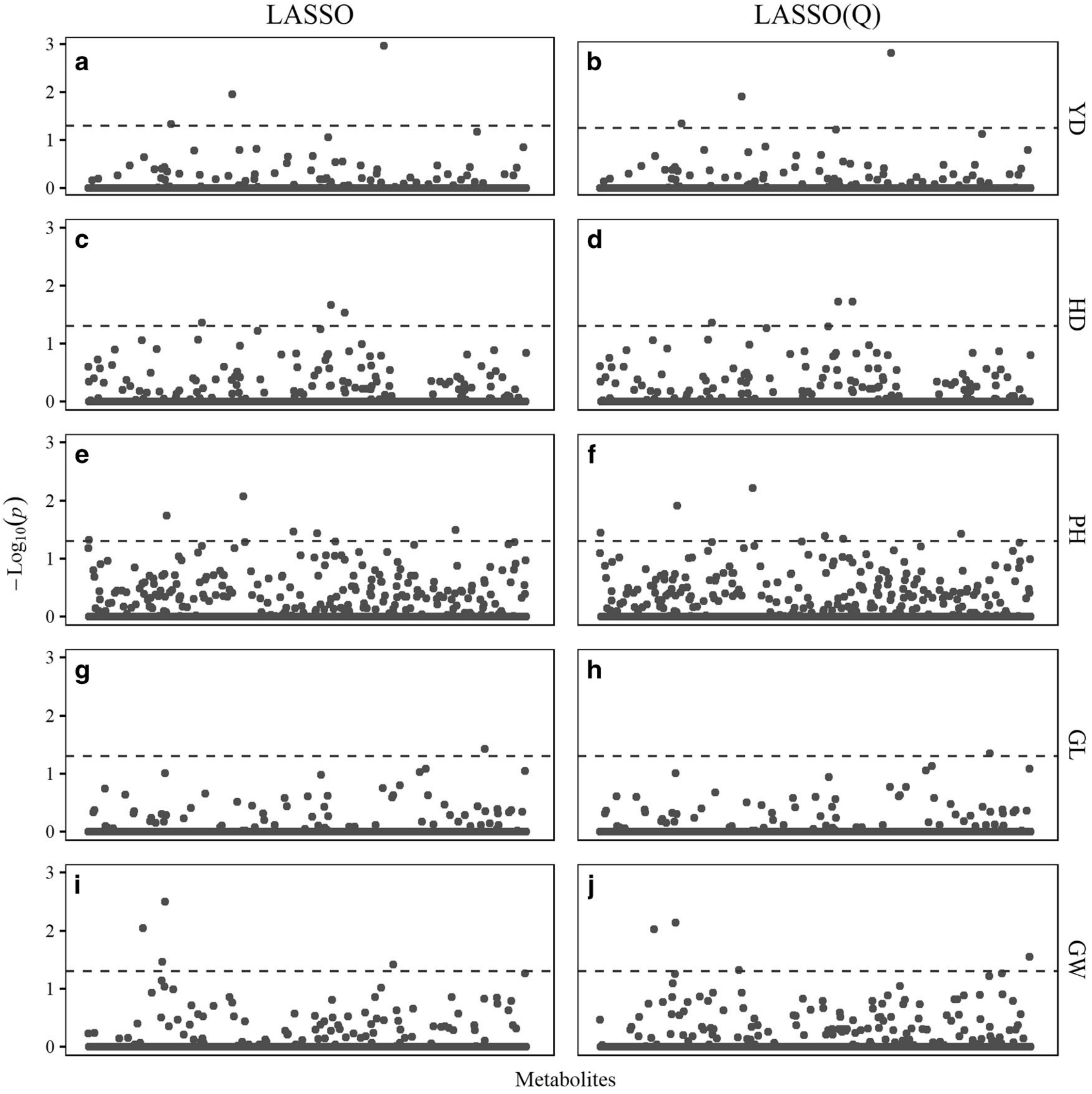

Fig. 2 Metabolome-wide association studies for the five traits from LASSO method. Left panels (**a, c, e, g,** and **i**) show results for LASSO without population structure and right panels (**b, d, f, h,** and **j**) show results for LASSO with population structure, denoting LASSO(Q). The horizontal dashed lines are the critical values at the 0.05 level, $-\log_{10}(0.05) = 1.30$

for the selected metabolites ranged from 0 to 0.20 and each scenario has been replicated for 500 times. The conclusion of simulated studies were based on the average of the 500 replicated experiments.

Firstly, we compared the empirical power of the four methods. The detection power increased with the size of simulated effect for all four methods (Fig. 3, left panel). When the simulated metabolite effect accounted for 20% of phenotypic variance (heritability = 0.2), the empirical power of all four methods are close to 1, suggesting a high level of efficiency for these methods. When the simulated metabolite effect was small, the four methods possessed the distinct powers, with LMM showing the lowest power and LR showing the highest power.

On the other hand, LR was subject to highest level of Type I error rate compared to the other three methods (Fig. 3, right panel). The sum of the results from the simulated study indicated that LASSO analysis is optimal in MWAS regarding both detection power and control for false discoveries.

The annotation of the identified metabolites associated with agronomic traits

In the subsequent analyses, we only consider three methods LMM, LASSO, Bayes B because the previous results indicated that LR suffered high level of false detection rate. For each of the three methods, we combined the results of analyses with and without population structure. We detected 2, 25, and 4 metabolites to be significantly associated with five traits by using LMM, LASSO, Bayes B, respectively.

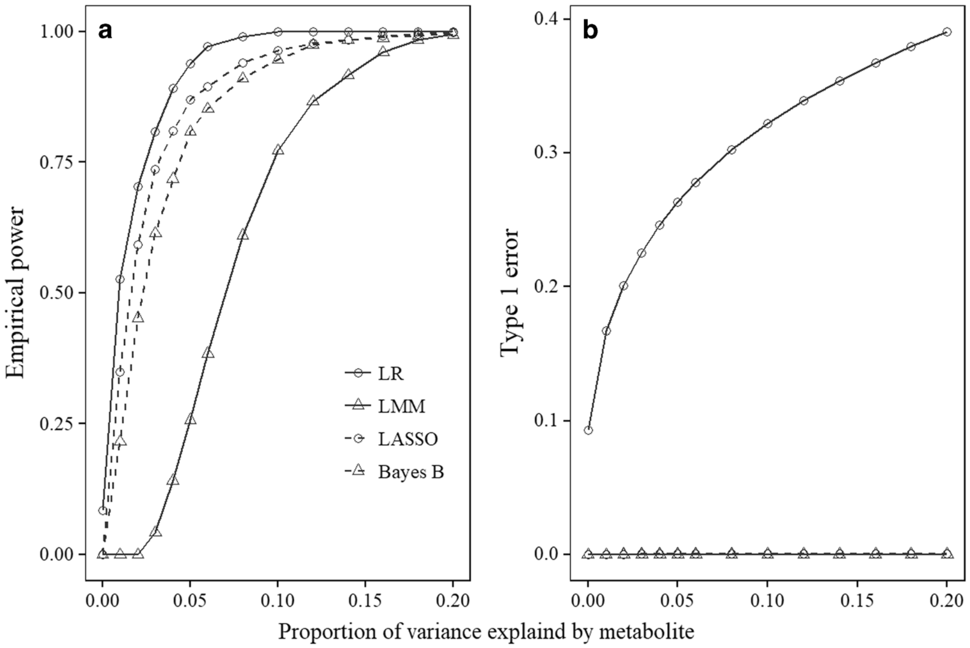

Fig. 3 The results of the 14 scenarios with simulation effects ranging from 0 to 0.20 from the four methods, LR, LMM, LASSO and Bayes B, each scenario with 500 replications: **a** The empirical power and **b** The type 1 error

The metabolites identified by the LMM and Bayes B were also detected by LASSO (see the Supplementary Fig. S9), which is consistent with the simulate study that showed LASSO is the most powerful method. Therefore, the annotation analysis was focused on the 25 metabolites detected by LASSO. Table 1 summaries the information on all the 25 metabolites, including p values in two scenarios from LASSO, heritability estimated from two biological replicates, and the name and category that each metabolite falls in. Annotation of metabolites are also available in the websites (http://rice.hzau.edu.cn/rice/clone/list_8.php). The PH is the trait for which we detected the most associated metabolites (10 significant metabolites), which is consistent with the result that the PH has the highest level of heritability (about 0.75 in Table 2). Most of the identified metabolites are associated with a single trait; whereas Mr1270 (N-Feruloylserotonin) appeared to be responsible for two traits, i.e., HD and PH.

We also implemented the association analysis using two subspecies (*Indica* and *Japonica*), separately. The results of the two subspecies (*indica* and *Japonica*) are shown in Supplementary Fig. S10 and S11, respectively. The results were similar to the results when the two subspecies were combined in the analysis, i.e., LASSO outperformed the other three methods. A total of 30 and 12 metabolites assumed to be association with agronomical traits are found in the *Indica* and *Japonica* populations, respectively (Supplementary Table S1 and S2), with no common metabolites identified in both subspecies (Supplementary Fig. S12). In addition, very few common metabolites were found when we compared the result from analysis using entire sample with the results from the analyses where subspecies were used separately.

Prediction of traits using metabolites

The identified trait-associated metabolites may be used for the prediction of the traits of interest as the SNPs in Genomic Prediction. Table 2 shows that metabolome can explain a large fraction of phenotypic variance (ranging from 0.6147 to 0.7453), but lower than that can be explained by using genomic data. This may be due to the fact that much less metabolites (839 metabolites) than DNA variants (usually a few millions of SNPs) have been used for prediction.

We adopted three prediction models, i.e., BLUP (equivalent to LMM), Bayes B and LASSO, to explore how well metabolome can be used to predict the phenotypes. Ten-fold cross validation has been repeated for 50 times for each method, and the average and standard deviation of predictability were calculated for comparison. The predictability of GBLUP, a trait-prediction method based on genomic data (VanRaden 2008), was also included in the comparison where five agronomic traits were analyzed, separately. In order to test if the number of predictive variables affects the performance of prediction models, we randomly selected a subset of 839 SNPs and used them in G-BLUP analysis, called G-BLUP (839), to compare with the metabolome-based prediction analyses in which 839 metabolites have been used in predicting the genetic values of the traits. All five analyses showed low standard deviation of predictability, indicating the stable property for

Table 1 The identified metabolites for the five traits from LASSO method in the whole population

Trait	Metabolite	LASSO	LASSO (Q)[a]	Heritability[b]	Name	Category
YD	Mr1201	0.05	0.04	0.84	Cyanidin 3-O-glucoside	Anthocyanin
	Mr1336	0.01	0.01	0.59	L-Arginine	Amino acid
	Mr1679	0.001	0.001	0.68	—[c]	—
HD	Mr1270	0.04	0.04	0.34	N-Feruloylserotonin	Phoyphenol
	Mr1389	0.06	0.05	0.80	—	—
	Mr1531	0.06	0.05	0.48	Phellodenol H	Phoyphenol
	Mr1554	0.02	0.02	0.41	—	—
	Mr1584	0.03	0.02	0.52	—	—
PH	Mr1004	0.05	0.04	0.27	N-Caffeoylputrescine	Phenolamine
	Mr1192	0.02	0.01	0.27	Thiamin	Vitamine
	Mr1270	0.06	0.05	0.34	N-Feruloylserotonin	Phoyphenol
	Mr1360	0.01	0.01	0.00	—	—
	Mr1363	0.05	0.06	0.30	—	—
	Mr1472	0.03	0.05	0.02	—	—
	Mr1525	0.04	0.04	0.59	4,6-Dihydroxyquinoline *O*-hexoside	Amino acid derivative
	Mr1565	0.05	0.05	0.63	—	—
	Mr1838	0.03	0.04	0.28	—	—
	Mr1968	0.05	0.05	0.55	—	—
GL	Mr1904	0.04	0.04	0.42	—	—
GW	Mr1136	0.01	0.01	0.73	—	—
	Mr1183	0.03	0.08	0.80	4-O-p-Coumaroylqionic acid	Ploypheno1
	Mr1188	0.003	0.01	0.47	Trigonelline	Alkaloid
	Mr1330	0.14	0.05	0.41	L-Lysine	Amino acid
	Mr1702	0.04	0.09	0.37	—	—
	Mr1930	0.14	0.05	0.26	—	—
	Mr1993	0.05	0.03	0.45	—	—

[a] The sign “Q” represents the LASSO method with population structure

[b] Here heritability of metabolites are estimated based on ANOVA method from the two replicates of measurement

[c] “—” denotes the metabolites we don’t know their details, including name and category

Table 2 Variance parameters of the five agronomic traits from genomic and metabolic information

	Genome			Metabolome		
	Genetic variance	Residual variance	Heritability	Genetic variance	Residual variance	Heritability
YD	0.6086	0.3076	0.6643	0.6262	0.3924	0.6147
HD	1.2791	0.0525	0.9606	0.7970	0.4048	0.6632
PH	0.7289	0.1004	0.8789	0.6804	0.2325	0.7453
GL	0.9986	0.1136	0.8979	0.7480	0.4734	0.6124
GW	0.6241	0.0983	0.8639	0.5656	0.2879	0.6627

prediction (Fig. 4). The average predictability of the five traits are 0.590, 0.503, 0.460, 0.458, and 0.422, respectively, across five analyses including two GBLUP methods and three metabolite-based prediction methods. It appeared that the predictability for Genomic Prediction was higher than that for metabolite-based prediction methods, which agreed with the previous studies (Riedelsheimer et al. 2012a; Xu et al. 2016). This can be explained by the fact that metabolic data are not as accurately and completely measured as genomic data.

We randomly selected one out of 50 cycles of ten-fold cross validation for demonstration of the predictability as

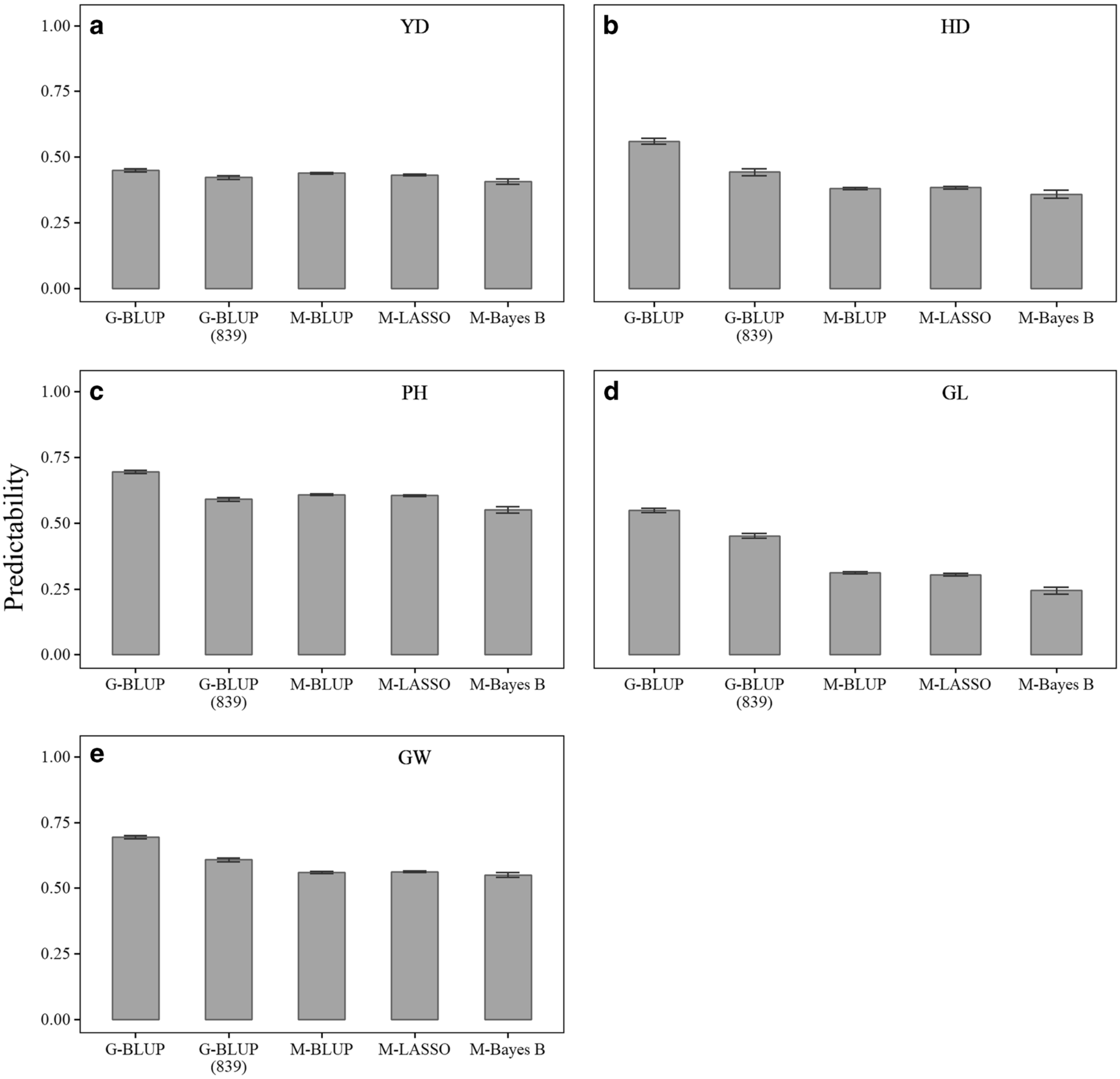

Fig. 4 Predictability for the five traits from genomic and metabolic information using BLUP, Bayes B, and LASSO methods, where the bars are their standard error estimated from 50 replicates cross validation. The prefix of methods represent what kind of information are used, G for genome and M for metabolome. And G-BLUP (839) represents we randomly selected 839 SNPs to construct the kinship matrix, which is equal to the number of metabolites

shown in Supplementary Fig. S13, where the observed phenotypic values were plotted against the predicted phenotypic values based on four analyses and five traits. For both training set (gray points) and test set (black triangle points), the data points are all concentrated around the diagonal line, with training data tighter than the test data. GBLUP are more compact than the other three methods, indicating a better prediction than that for metabolite-based prediction.

Discussion

Due to the similarity in the data structure for GWAS and MWAS, the statistical methods that have been widely employed in GWAS may be adopted in MWAS. In the study, we explore this possibility by applying four regression models, i.e., simple linear regression (LR), linear mixed model (LMM), Bayesian analysis with spike-slab priors (Bayes B) and least absolute shrinkage and selection operator (LASSO), to a rice data set for which 839 metabolites and five traits have been measured. This is the first comprehensive investigation of the association between the metabolites and the agronomic traits in the rice. A total of 25 trait-associated metabolites have been identified. Literature search showed that some of the identified metabolites play important role in several growth and development process in plants, directly or indirectly related to agronomic traits. For example, both Mr1004 and Mr1270 belong to family of hydroxycinnamic acid amides (HCAAs), which are accumulated in various tissues and organs of plants and are involved in plant growth and development process (Facchini et al. 2002; Martin-Tanguy 1985). Studies have proved that Mr1004, which is also called N-caffeoylputrescine (CP), is involved to the development of flower (Martin-Tanguy 1985) and the defense against disease or herbivores (Kaur et al. 2010); therefore, it is not surprising to find association between Mr1004 (or CP) and PH. The previous research indicated that CP can be regulated by jasmonic acid through its control on the expression of biosynthetic enzymes of CP (Onkokesung et al. 2012). Indeed, some studies have collected evidences showing that jasmonic acid are involved in stem elongation (Heinrich et al. 2013; Hyun et al. 2008). Metabolite Mr1270, which is also named N-Feruloylserotonin(FS), is synthesized on the catalysis of serotonin N-hydroxycinnamoyltransferase (SHT) (Jang et al. 2004). It has been reported that FS acts as phytoalexins in defense against plant pathogens (Kumarasamy et al. 2003; Tanaka et al. 2003). To our best knowledge, it is the first time to report that FS is significantly associated with both HD and PH. Another metabolite significantly associated with PH is thiamin (Mr1192), which participates in many important biology processes that are potentially relevant to PH, such as glycolysis, the pentose phosphate pathway and the thricarboxylic acide cycle (Krampitz 1969). Mr1188, named trigonelline, was found to be associated with GW in our analysis. It has been reported that trigonelline, as one of alkaloid, plays an important role in the regulation of cell growth and development (Mazzuca et al. 2000) and in an osmoregulatory mechanism against water deficiency (Cho et al. 2003). In addition, a study in peanut showed that reduced trigonelline is associated with increased grain yield (Cho et al. 2011). We also identified trait-associated metabolites the functions of which have not been well established. Follow-up researches are needed to advance our understanding of the biology between these metabolites and traits.

It is the first time to apply four regression methods (LR, LMM, Bayes B, and LASSO) to MWAS. Our results showed that LR is powerful but suffers high level of false detection rate, which is consistent with the previous report (Meyer et al. et al. 2007). Complex traits, such as YD, are determined by many QTLs through many trait-relevant pathways that involve representative metabolites. In LR analysis, only one metabolite is included in the regression model at a time. Without controlling population structure, the LR is simply a single-regression analysis, leaving the data of other metabolites completely ignored. By including the metabolome-inferred population structure in LR, this model is somehow equivalent to multiple-regression approaches, such as LASSO or BLUP. Therefore, there is significant difference between the LR models with and without controlling population structure (the first three principal components of 839 metabolites). However, inclusion of metabolome-inferred population structure in LASSO or BLUP will not improve modeling because these two multiple-regression methods analyze all metabolites simultaneously and such population structure has already been used.

LMM provides a good control for false discoveries using a kinship matrix which is calculated using genomic data in GWAS (Kang et al. 2010, 2008; Lippert et al. 2011; Zhang et al. 2005, 2010; Zhou and Stephens 2012); in the current MWAS study, the kinship matrix was calculated using metabolic data. In LMM, all the random effects, called polygenic effects in GWAS (Yang et al. 2010), follow a multivariate normal distribution with the kinship matrix dictating their variance-covariance structure. The use of the kinship is to capture the relatedness of individuals such that only the common variance shared by all loci or metabolites needs to be estimated, leading to an efficient information sharing among predictors and a substantial dimensionality reduction for the parameters that need to be estimated. As a result, the false discovery rate has been effectively controlled and the detection power has been greatly increased. Like the composite interval mapping (CIM) method in the QTL mapping (Zeng 1994), all genomic loci or metabolites are treated as the random effects and included in kinship matrix to provide a control for background noise in LMM.

However, the polygenic or polymetabolomic background has already absorbed the single metabolites under the test in LMM, which may reduce the detection power (Listgarten et al. 2012). Alternative strategies have been proposed and discussed for handling this issue (Bernardo 2013; Listgarten et al. 2012; Rincent et al. 2014; Speed et al. 2012; Wei and Xu 2016). LASSO can avoid the repetition through shrinkage of nonsense regressors (Tibshirani 1996), which provides another strategy to increase the detection power. The analyses of both the real data and simulated data demonstrated that LASSO outperformed the other methods. Because the variance of a parameter or effect is difficult to calculate with a close form in LASSO, bootstrap (a computationally expensive approach) may be used for estimation of the empirical variance to accomplish a Wald test. Moreover, a calculation involving all the independent variables simultaneously in LASSO will significantly add to computational burden, especially when big data including millions predictors and thousands of individuals are analyzed. With the aid of parallel computations in computer cluster, one can easily analyze the GWAS or MWAS data by implementing LASSO analysis, for example, the LASSO algorithm incorporated in 'plink' software (Chang et al. 2015).

Finally, we explored the prediction of traits in rice populations using metabolite data. One of the important missions for crop breeding is to select the elite lines. Compared with the traditional selection based on phenotypes, selection based on genomes and metabolomes will be much less expensive and more efficient (Riedelsheimer et al. 2012a; VanRaden et al. 2009). Genomic selection has been proven to be an effective method for estimation breeding values in dairy cattle (Hayes et al. 2009). Our study demonstrated that both genome and metabolome can achieve high level of the accuracy for trait prediction, justifying their potential use in crop breeding. The results showed that genomic prediction was more accurate than metabolic prediction based on cross-validation. This may be attributed to the fact that only a small fraction of the metabolites were considered whereas the whole-genome loci were included in the study. There are about 200,000 metabolites in plant kingdom (Saito and Matsuda 2010); however, only 839 metabolic profiles were collected as a snapshot at a specific moment in a specific organ. Including more useful metabolic profiles in metabolic prediction model will boost the predictability. Statistical methods for trait prediction can be classified into two categories, *i.e.*, (1) all predictors share a common spread/variance, for example, BLUP (RRBLUP or GBLUP) (VanRaden 2008), and (2) a small portion of the predictors are relevant to the trait, such as Bayes series methods (Habier et al. 2011; Meuwissen et al. 2001) and LASSO (Tibshirani 1996).The first category works well when there is few large-effect predictors, whereas the second category is ideal when major predictors exist (Zhou et al. 2013). In the study, both categories performed well on metabolic prediction.

The average predictability of the five agronomic traits based on BLUP is 0.590 and 0.460 for genomic prediction and metabolic prediction, respectively, which are much higher than what were previously reported (Riedelsheimer et al. 2012a; Xu et al. 2016). Note that the previous study was based on the analysis of 210 recombinant inbred lines, and the predictabilities for KGW, grain, yield and tiller were 0.58, 0.31, 0.20 and 0.16, respectively. The high predictability in our study is mainly due to the high quality population with abundant genetic diversities. The rice population in the original study contains 533 lines from five research projects around the world (Chen et al. 2014). Most individuals are from three programs, the core/mini-core collection of rice in China (200 varieties) (Zhang et al. 2011), the International Rice Molecular Breeding Program (132 varieties) (Yu et al. 2003) and USDA core collections (148 varieties) (Yan et al. 2009), which can be regarded as the representative of thousands rice accessions. It has been indicated in literature that the prediction accuracy of genomic selection (GS) are largely affected by the composition of the reference population, and its size and its relatedness to breeding population (Bassi et al. 2016; Bentley et al. 2014; Hayes et al. 2009; Isidro et al. 2015). This is also true for metabolic selection (MS). Therefore, it is crucial to construct a useful reference population for the use of GS and/or MS.

We conclude that the four methods (LR, LMM, Bayes B, and LASSO), which have been commonly used in QTL mapping, can be adopted and applied to MWAS, and LASSO outperformed the other three for this purpose. Moreover, we demonstrated the feasibility of predicting genetic values for traits using metabolome data. Our study also uncovered important metabolites that are relevant to agronomically important traits. Our future research will involve the development of optimal prediction strategies when genomic, metabolomic and transcriptomic data are all available.

Acknowledgements We appreciate Dr. Shizhong Xu at University of California, Riverside for providing insightful suggestions. We also appreciate Drs. Qifa Zhang and Weibo Xie at Huazhong Agricultural University for providing the rice data.

Funding This work was supported for start-up funding of UCR to Zhenyu Jia.

Author contribution Z.J. and A.G.W. conceived and designed the experiments. J.L.W., R.D.L., and H.Q. conducted the experiments and analyzed data. J.L.W. wrote the program. J.L.W. and Z.J. wrote the manuscript. All authors have read and approved the final manuscript.

Compliance with ethical standards

Competing interests The authors declare that they have no competing interests.

References

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J (2016) Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci 242:23–36

Bentley AR, Scutari M, Gosman N, Faure S, Bedford F, Howell P et al. (2014) Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. Theor Appl Genet 127(12):2619–2633

Bernardo R (2013) Genomewide markers for controlling background variation in association mapping. Plant Genome **6**(1)

Cacciatore S, Loda M (2015) Innovation in metabolomics to improve personalized healthcare. Ann N Y Acad Sci 1346(1):57–62

Chan EK, Rowe HC, Hansen BG, Kliebenstein DJ (2010) The complex genetic architecture of the metabolome. PLoS Genet 6(11): e1001198

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4(1):7

Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W et al. (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet 46(7):714–721

Cho Y, Kodjoe E, Puppala N, Wood A (2011) Reduced trigonelline accumulation due to rhizobial activity improves grain yield in peanut (*Arachis hypogaea* L.). Acta Agric Scand, Sect B-Soil Plant Sci 61(5):395–403

Cho Y, Njiti V, Chen X, Lightfoot D, Wood A (2003) Trigonelline concentration in field-grown soybean in response to irrigation. Biol Plant 46(3):405–410

Draisma HH, Pool R, Kobl M, Jansen R, Petersen A-K, Vaarhorst AA et al. (2015) Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. Nat Commun 6:7208

Facchini PJ, Hagel J, Zulak KG (2002) Hydroxycinnamic acid amide metabolism: physiology and biochemistry. Can J Bot 80 (6):577–589

Fiehn O (2002) Metabolomics–the link between genotypes and phenotypes. Plant Mol Biol 48(1-2):155–171

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. Nat Biotechnol 18(11):1157–1161

Gieger C, Geistlinger L, Altmaier E, De Angelis MH, Kronenberg F, Meitinger T et al. (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 4(11):e1000282

Gong L, Chen W, Gao Y, Liu X, Zhang H, Xu C et al. (2013) Genetic analysis of the metabolome exemplified using a rice population. Proc Natl Acad Sci 110(50):20320–20325

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinforma 12(1):186

Hayes BJ, Bowman PJ, Chamberlain A, Goddard M (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92(2):433–443

Heinrich M, Hettenhausen C, Lange T, Wünsche H, Fang J, Baldwin IT et al. (2013) High levels of jasmonic acid antagonize the biosynthesis of gibberellins and inhibit the growth of *Nicotiana attenuata* stems. Plant J 73(4):591–606

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42(11):961–967

Huang X, Yang S, Gong J, Zhao Q, Feng Q, Zhan Q et al. (2016) Genomic architecture of heterosis for yield traits in rice. Nature 537(7622):629–633

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q et al. (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44 (1):32–39

Hyun Y, Choi S, Hwang H-J, Yu J, Nam S-J, Ko J et al. (2008) Cooperation and functional diversification of two closely related galactolipase genes for jasmonate biosynthesis. Dev Cell 14 (2):183–192

Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128(1):145–158

Jang S-M, Ishihara A, Back K (2004) Production of coumaroylserotonin and feruloylserotonin in transgenic rice expressing pepper hydroxycinnamoyl-coenzyme A: serotonin N-(hydroxycinnamoyl) transferase. Plant Physiol 135(1):346–356

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42 (4):348–354

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178(3):1709–1723

Kaur H, Heinzel N, Schöttner M, Baldwin IT, Gális I (2010) R2R3-NaMYB8 regulates the accumulation of phenylpropanoid-polyamine conjugates, which are essential for local and systemic defense against insect herbivores in *Nicotiana attenuata*. Plant Physiol 152(3):1731–1747

Krampitz L (1969) Catalytic functions of thiamin diphosphate. Annu Rev Biochem 38(1):213–240

Kumarasamy Y, Middleton M, Reid R, Nahar L, Sarker S (2003) Biological activity of serotonin conjugates from the seeds of *Centaurea nigra*. Fitoterapia 74(6):609–612

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. Nat Methods 8(10):833–835

Lisec J, Meyer RC, Steinfath M, Redestig H, Becher M, Witucka-Wall H et al. (2008) Identification of metabolic and biomass QTL in *Arabidopsis thaliana* in a parallel analysis of RIL and IL populations. Plant J 53(6):960–972

Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D (2012) Improved linear mixed models for genome-wide association studies. Nat Methods 9(6):525–526

Martin-Tanguy J (1985) The occurrence and possible function of hydroxycinnamoyl acid amides in plants. Plant Growth Regul 3 (3):381–399

Matsuda F, Nakabayashi R, Yang Z, Okazaki Y, Yonemaru J, Ebana K et al. (2015) Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. Plant J 81(1):13–23

Matsuda F, Okazaki Y, Oikawa A, Kusano M, Nakabayashi R, Kikuchi J et al. (2012) Dissection of genotype–phenotype associations in rice grains using metabolome quantitative trait loci analysis. Plant J 70(4):624–636

Mazzuca S, Bitonti M, Innocenti A, Francis D (2000) Inactivation of DNA replication origins by the cell cycle regulator, trigonelline, in root meristems of *Lactuca sativa*. Planta 211(1):127–132

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ et al. (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci USA 106(30):12273–12278

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

Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Törjék O et al. (2007) The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 104 (11):4759–4764

Onkokesung N, Gaquerel E, Kotkar H, Kaur H, Baldwin IT, Galis I (2012) MYB8 controls inducible phenolamide levels by activating three novel hydroxycinnamoyl-coenzyme A: polyamine transferases in *Nicotiana attenuata*. Plant Physiol 158(1):389–407

Pace J, Yu X, Lübberstedt T (2015) Genomic prediction of seedling root length in maize (*Zea mays* L.). Plant J 83(5):903–912

Pérez P, de Los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 114:164442

Rhee EP, Ho JE, Chen M-H, Shen D, Cheng S, Larson MG et al. (2013) A genome-wide association study of the human metabolome in a community-based cohort. Cell Metab 18(1):130–143

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R et al. (2012a) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44 (2):217–220

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C et al. (2012b) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci USA 109(23):8872–8877

Rincent R, Moreau L, Monod H, Kuhn E, Melchinger AE, Malvar RA et al. (2014) Recovering power in association mapping panels with variable levels of linkage disequilibrium. Genetics 197 (1):375–387

Saito K, Matsuda F (2010) Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol 61:463–489

Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. Am J Human Genet 91(6):1011–1021

Steinfath M, Gärtner T, Lisec J, Meyer RC, Altmann T, Willmitzer L et al. (2010) Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. Theor Appl Genet 120(2):239–247

Tanaka E, Tanaka C, Mori N, Kuwahara Y, Tsuda M (2003) Phenylpropanoid amides of serotonin accumulate in witches' broom diseased bamboo. Phytochemistry 64(5):965–969

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Statis Soc Series 58 (1):B 267–288.

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

VanRaden P, Van Tassell C, Wiggans G, Sonstegard T, Schnabel R, Taylor J et al. (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92(1):16–24

Wang S-B, Feng J-Y, Ren W-L, Huang B, Zhou L, Wen Y-J et al. (2016a) Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep 6:19444

Wang S-B, Wen Y-J, Ren W-L, Ni Y-L, Zhang J, Feng J-Y et al. (2016b) Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. Sci Rep 6:29951

Wei J, Xu S (2016) A random-model approach to QTL mapping in multiparent advanced generation intercross (MAGIC) populations. Genetics 202(2):471–486

Wen W, Li D, Li X, Gao Y, Li W, Li H et al. (2014) Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun 5:19444

Xu S, Xu Y, Gong L, Zhang Q (2016) Metabolomic prediction of yield in hybrid rice. Plant J 88(2):219–227

Xu Y, Xu C, Xu S (2017) Prediction and association mapping of agronomic traits in maize using multiple omic data. Heredity

Yan WG, Li Y, Agrama HA, Luo D, Gao F, Lu X et al. (2009) Association mapping of stigma and spikelet characteristics in rice (*Oryza sativa* L.). Mol Breed 24(3):277–292

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565–569

Yano K, Yamamoto E, Aya K, Takeuchi H, Lo P-c, Hu L et al. (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet 48(8):927–934

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38 (2):203

Yu S, Xu W, Vijayakumar C, Ali J, Fu B, Xu J et al. (2003) Molecular diversity and multilocus organization of the parental lines used in the International Rice Molecular Breeding Program. Theor Appl Genet 108(1):131–140

Zeng Z-B (1994) Precision mapping of quantitative trait loci. Genetics 136(4):1457–1468

Zhang H, Zhang D, Wang M, Sun J, Qi Y, Li J et al. (2011) A core collection and mini core collection of *Oryza sativa* L. in China. Theor Appl Genet 122(1):49–61

Zhang Y-M, Mao Y, Xie C, Smith H, Luo L, Xu S (2005) Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). Genetics 169(4):2267–2275

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42(4):355–360

Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet 9(2):e1003264

Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44(7):821–824