doi:10.1093/plphys/kiab273 PLANT PHYSIOLOGY 2021: 187: 1011–1025

Plant Physiology®

Research Article

The metabolomic landscape of rice heterosis highlights pathway biomarkers for predicting complex phenotypes

Zhiwu Dan ,[1] Yunping Chen ,[1] Hui Li,[1] Yafei Zeng,[1] Wuwu Xu,[1] Weibo Zhao,[1] Ruifeng He[2] and Wenchao Huang[1,*,†]

1 State Key Laboratory of Hybrid Rice, Key Laboratory for Research and Utilization of Heterosis in Indica Rice, the Ministry of Agriculture, College of Life Sciences, Wuhan University, Wuhan 430072, China

2 Institute of Biological Chemistry, Washington State University, Pullman, Washington 99164-6414, USA

*Author for communication: wenchaoh@whu.edu.cn

†Senior author.

Z.D. designed the research; Z.D. and W.H. collected phenotypic data; Z.D. and Y.C. performed most of the metabolomics experiments; H.L., Y.Z., W.X., and W.Z. participated in material preparations, experiments, and data analyses in metabolomics; Z.D. managed comprehensive data collection and analyses; W.H. supervised the experiments; Z.D., Y.C., R.H., and W.H. wrote the manuscript.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys/pages/general-instructions) is: Wenchao Huang (wenchaoh@whu.edu.cn).

Abstract

Understanding the molecular mechanisms underlying complex phenotypes requires systematic analyses of complicated metabolic networks and contributes to improvements in the breeding efficiency of staple cereal crops and diagnostic accuracy for human diseases. Here, we selected rice (*Oryza sativa*) heterosis as a complex phenotype and investigated the mechanisms of both vegetative and reproductive traits using an untargeted metabolomics strategy. Heterosis-associated analytes were identified, and the overlapping analytes were shown to underlie the association patterns for six agronomic traits. The heterosis-associated analytes of four yield components and plant height collectively contributed to yield heterosis, and the degree of contribution differed among the five traits. We performed dysregulated network analyses of the high- and low-better parent heterosis hybrids and found multiple types of metabolic pathways involved in heterosis. The metabolite levels of the significantly enriched pathways (especially those from amino acid and carbohydrate metabolism) were predictive of yield heterosis (area under the curve = 0.907 with 10 features), and the predictability of these pathway biomarkers was validated with hybrids across environments and populations. Our findings elucidate the metabolomic landscape of rice heterosis and highlight the potential application of pathway biomarkers in achieving accurate predictions of complex phenotypes.

Introduction

Variations in the levels of specific metabolites are closely related to the quantitative changes in complex phenotypes. For example, in a previous study in tomato (*Solanum lycopersicum*), most of the identified metabolites that belong to central metabolic pathways were significantly correlated with whole-plant phenotypic traits (Schauer et al., 2006). Recently, 40 plasma metabolites explained the variance in gut

Received April 22, 2021. Accepted May 27, 2021. Advance access publication June 14, 2021

microbiome α-diversity in humans (Wilmanski et al., 2019). Although the combination of metabolites has potential for predicting multiple polygenic phenotypes (Wen et al., 2014; Dan et al., 2019, 2020), the prediction of individuals with the same performance is hampered by molecular heterogeneity (Chen et al., 2014; Menche et al., 2017; Guo et al., 2019). Moreover, the contribution of statistically insignificant metabolites to phenotypic variances under one condition was ignored in the other conditions. With the rapid advancements in dysregulated network analysis of metabolomics (Chong et al., 2018; Shen et al., 2019), the development of metabolomic biomarkers at the pathway level after discrete metabolites provides approaches to increase the predictability of complex phenotypes.

Heterosis, which has been widely used for improving global food production, has complex characteristics, and the metabolomic mechanisms have yet to be elucidated (Darwin, 1876; Williams, 1959). With continuously growing populations and dramatic climatic changes, the breeding of new heterotic and adaptive hybrids are a major challenge for traditional breeding programs (Varshney et al., 2018; Hickey et al., 2019). Previous studies conducted on hybrid crops (including maize, wheat, and rice) have demonstrated that the screened metabolites detected from leaves or roots have predictive power for biomass (Lisec et al., 2011), grain weight and production (Zhao et al., 2015; Xu et al., 2016; Dan et al., 2019), and yield heterosis (Dan et al., 2020). Obstacles such as feature selection and cross-validation procedures still exist (Crossa et al., 2017; Dan et al., 2019), and the metabolomic connections between components (e.g. grain number and grain weight) and complex traits (e.g. yield and biomass) are largely unknown. Therefore, metabolome-based precision designs require optimization to achieve accurate predictions across populations and environments.

To understand the metabolomic mechanisms of heterosis and identify robust pathway biomarkers for yield heterosis in rice, we identified heterosis-associated analytes and revealed their contribution to six agronomic traits. The metabolic pathways involved in heterosis were identified through dysregulated network analysis of the high- and low-better parent heterosis hybrids, and the finding of overlapping pathways revealed the metabolomic landscape of heterosis for both vegetative and reproductive traits. Quantitative changes in the significantly enriched pathways were predictive of yield heterosis, and the pathway biomarkers at a small number were further validated with hybrids across environments and a separate hybrid population, suggesting a wide application potential for predicting complex phenotypes.

Results

Identifying heterosis-associated analytes for six agronomic traits

To identify metabolic analytes associated with rice (*Oryza sativa*) heterosis, we phenotyped grain yield; four yield components (seed setting rate, grain weight, grain number, and tiller number); and plant height (PH, a yield-related trait) for a hybrid population (complete diallel crosses with 18 parents) and collected untargeted metabolite profiles from 15-d-old parental seedlings (Supplemental Table S1). Previous results have demonstrated that the calculated average parental metabolite levels are appropriate for representing the hybrid metabolite profiles (Dan et al., 2020). We performed a Pearson correlation analysis on the transformed parental metabolite levels and better-parent heterosis (BPH), which estimates the degree of hybrid performance outperforming the better parent, with the high values always pursued by the breeders, of the six investigated traits (Supplemental Figure S1). Although the degree of heterosis largely varied across traits at both individual and population levels (Figure 1A), closer links between the average parental metabolite levels and heterosis were observed based on the number of significant correlations, compared to those of the differences in and ratios of the values (Figure 1B).

Next, we performed partial least squares (PLS) regression analysis (Wold, 1975), which handles high-dimensional megavariate relationships, on the average parental metabolite levels to identify predictive analytes for heterosis, namely, heterosis-associated analytes. The number of latent factors that are proxies for blocks of directly observed variables ranged from 1 to 17, and 3 or 4 latent factors, at which the *r* value was the highest, were chosen for each trait in building predictive models (Supplemental Figure S2). In addition, both 10-fold cross-validation and a permutation test were performed for the six predictive models to estimate the issue of overfitting (Supplemental Figure S3). The optimal number of predictive analytes ranging from 100 to 300 was chosen for each trait after removing redundant feature information (Figure 1C). The correlation coefficients between the observed and predicted values of BPH for PH and grain yield at the maturation stage were 0.68 and 0.60, respectively (Figure 1, D and E), showing a higher predictability for the vegetative trait than those for reproductive traits (Supplemental Figure S4). For PH, 100 heterosis-associated analytes were identified, and an analyte (peak tag: M163T337_NEG) was annotated as 4-hydroxycinnamic acid with the corresponding standard (Figure 1F). The metabolite levels of 4-hydroxycinnamic acid, whose positive relationship with PH has been confirmed in diverse plants (Gui et al., 2011; Riedelsheimer et al., 2012b; Li et al., 2015), had significant positive correlations with PH heterosis (Figure 1G). None of the heterosis-associated analytes overlapped with the five reproductive traits (Figure 1H). In yield heterosis, more weight was observed for seed setting rate and tiller number, compared to grain number and grain weight, based on the number of overlapping heterosis-associated analytes.

Connections of heterosis-associated analytes among traits

To investigate the connections of heterosis-associated analytes among the traits, we performed both partial and Pearson correlation analyses on heterosis of the five

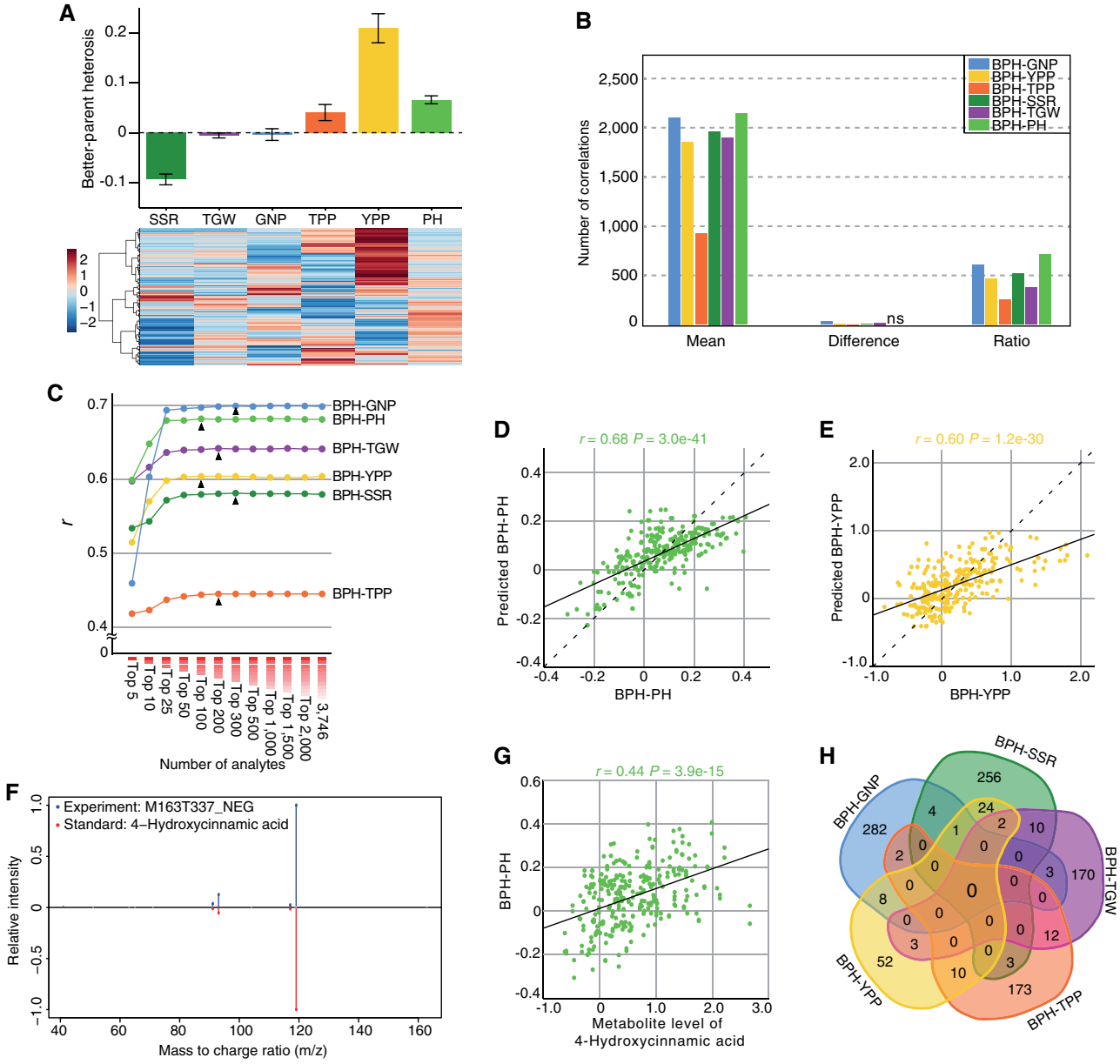

Figure 1 Identification of heterosis-associated analytes for six agronomic traits. A, Heterosis of six agronomic traits at the population and individual levels. Five reproductive traits (including yield and four yield components) and one vegetative trait (PH) were recorded. Bars represent standard errors. B, Number of correlations between transformed parental metabolite levels and heterosis. The means of, differences in, and ratios of parental metabolite levels were calculated to perform Pearson correlations with heterosis of the six traits. Correlations with $P < 0.05$ were considered significant. $N = 3,746$. C, Changes in r values with different numbers of predictive analytes in the PLS regressions. The optimal number of predictive analytes for each trait is marked with a black arrow. D–E, Correlations between the observed and predicted values of heterosis for PH (D) and yield (E) with correspondingly identified heterosis-associated analytes. F, MS/MS spectra of an analyte with peak tag M163T337_NEG and 4-hydroxycinnamic acid standard. G, Correlation between metabolite levels of 4-hydroxycinnamic acid and PH heterosis. H, Venn diagram of heterosis-associated analytes for yield and four yield components. In (A, D, and E) and (G) $N = 287$. SSR, seed setting rate; TGW, thousand-grain weight; GNP, grain number per panicle; TPP, tiller number per plant; YPP, yield per plant.

reproductive traits and PH (Figure 2A). Notably, heterosis of seed setting rate ($R = 0.72$) and tiller number ($R = 0.66$) contributed more than those of grain number ($R = 0.34$) and grain weight ($R = 0.16$) to yield heterosis, based on the correlation coefficients. We then investigated the relationship between the metabolite levels of the 27 overlapping heterosis-associated analytes for yield and seed setting rate (Supplemental Table S2), and found that all analytes had consistent positive or negative correlations with heterosis of the two traits (Figure 2B; Supplemental Table S3). Furthermore, positive and negative correlations were detected among the five reproductive traits, and consistent or opposite relationships were found between the metabolite levels of overlapping heterosis-associated analytes and

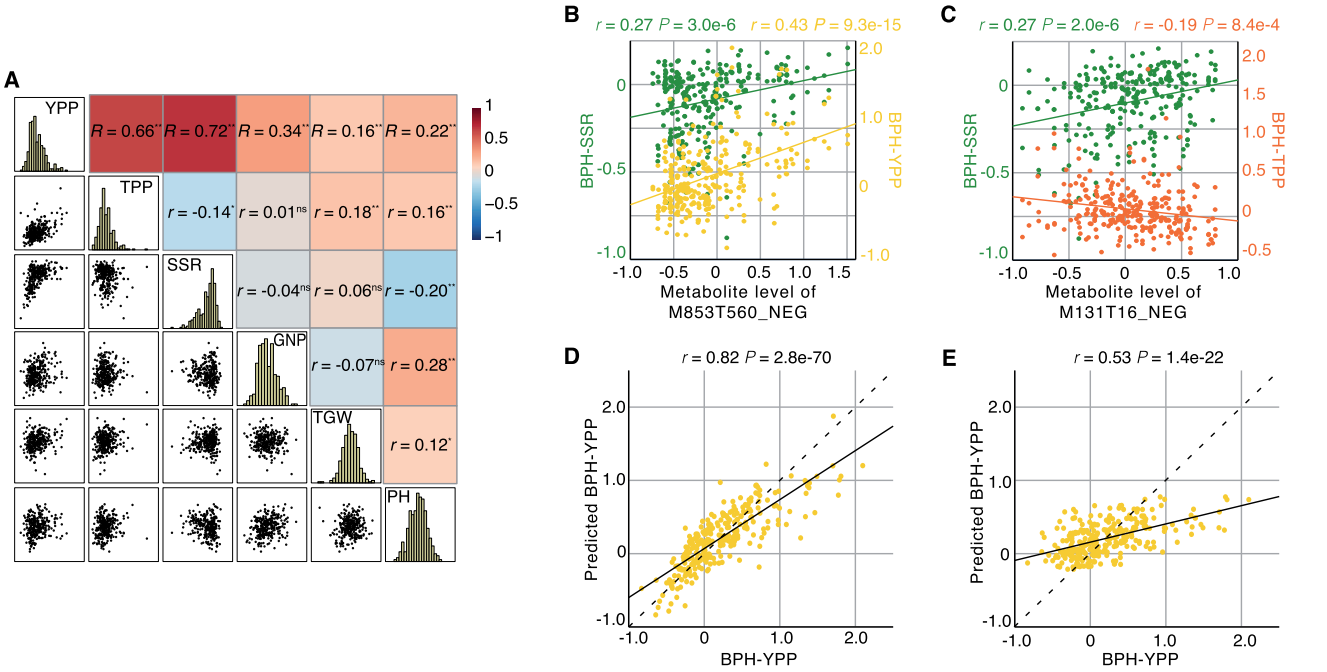

Figure 2 Connections of heterosis-associated analytes among traits. A, Correlations among heterosis of five reproductive traits and PH. Partial correlations were performed to investigate the contribution of four yield components and PH to yield heterosis. Pearson correlations were conducted to analyze the relationship among the four yield components and PH. Correlation coefficients of the partial and Pearson correlations are indicated with *R* and *r*, respectively. *, **, statistically significant at 0.05 and 0.01 levels, respectively; ns, no statistically significant correlation. B, Correlations between metabolite levels of M853T560_NEG and heterosis of seed setting rate and yield. C, Correlations between metabolite levels of M131T16_NEG and heterosis of seed setting rate and tiller number. D, Correlation between the observed and predicted values of yield heterosis based on heterosis of the four yield components and PH. An equation was obtained through stepwise regression analysis: BPH-YPP = BPH-SSR*1.674 + BPH-TPP*0.949 + BPH-TGW*0.571 + BPH-GNP*0.533 + BPH-PH*0.504 + 0.299. E, Correlation between the observed and predicted values of yield heterosis based on heterosis-associated analytes of the four yield components and PH with the equation in Figure 2D. In (A–E), *N* = 287.

heterosis (Figure 2C; Supplemental Figure S5 and Supplemental Table S3), indicating that the overlapping analytes underlie the association patterns for the traits.

We then performed stepwise regression analysis on the heterosis of yield and the four components and found an equation that explained the variance in yield heterosis (r = 0.81; Supplemental Figure S6). Because the degree of heterosis for the four components was predicted with corresponding heterosis-associated analytes (Supplemental Figure S4), we used the predicted values in the equation and calculated new values for yield heterosis. A significant correlation was observed between the observed and predicted values (r = 0.52; Supplemental Figure S6). Furthermore, the percentage of explained variance for yield heterosis slightly increased with the addition of PH to the regression equation (r = 0.82; Figure 2D), and the correlation coefficient increased to 0.53, based on the heterosis-associated analytes of the five traits (Figure 2E). Heterosis of PH was positively correlated with almost all investigated reproductive traits, except seed setting rate (Figure 2A), and the overlapping heterosis-associated analytes were found among these traits with the same correlations as those shown in Figure 2, B and C (Supplemental Figure S7 and Supplemental Table S3). These results indicated that the heterosis-associated analytes of the yield component and yield-related traits collectively contributed to the yield heterosis.

Metabolic pathways involved in heterosis

The metabolic pathways involved in heterosis need to be elucidated. Of the 3,746 analytes in our study, only 114 had been annotated, making it difficult to perform pathway enrichment analysis based on limited metabolite information. To identify enriched pathways for heterosis of each trait, we first divided the diallel cross population into two distinct regions of high- and low-BPH based on the quartiles (25th and 75th percentiles) at which most of the differential analytes from the empirical Bayesian analysis overlapped with the corresponding heterosis-associated analytes (Figure 3A). We then performed dysregulated network analysis on the two groups with **Met**abolite identification and **D**ysregulated **N**etwork **A**nalysis software (MetDNA; Shen et al., 2019), which annotates metabolites with a recursive algorithm and identifies dysregulated metabolic pathways based on differential metabolic peaks. The results showed that only two pathways were simultaneously enriched for the five reproductive traits (Figure 3B). The enriched pathways for heterosis of the seed setting rate and tiller number had higher percentages of overlapping pathways with yield heterosis

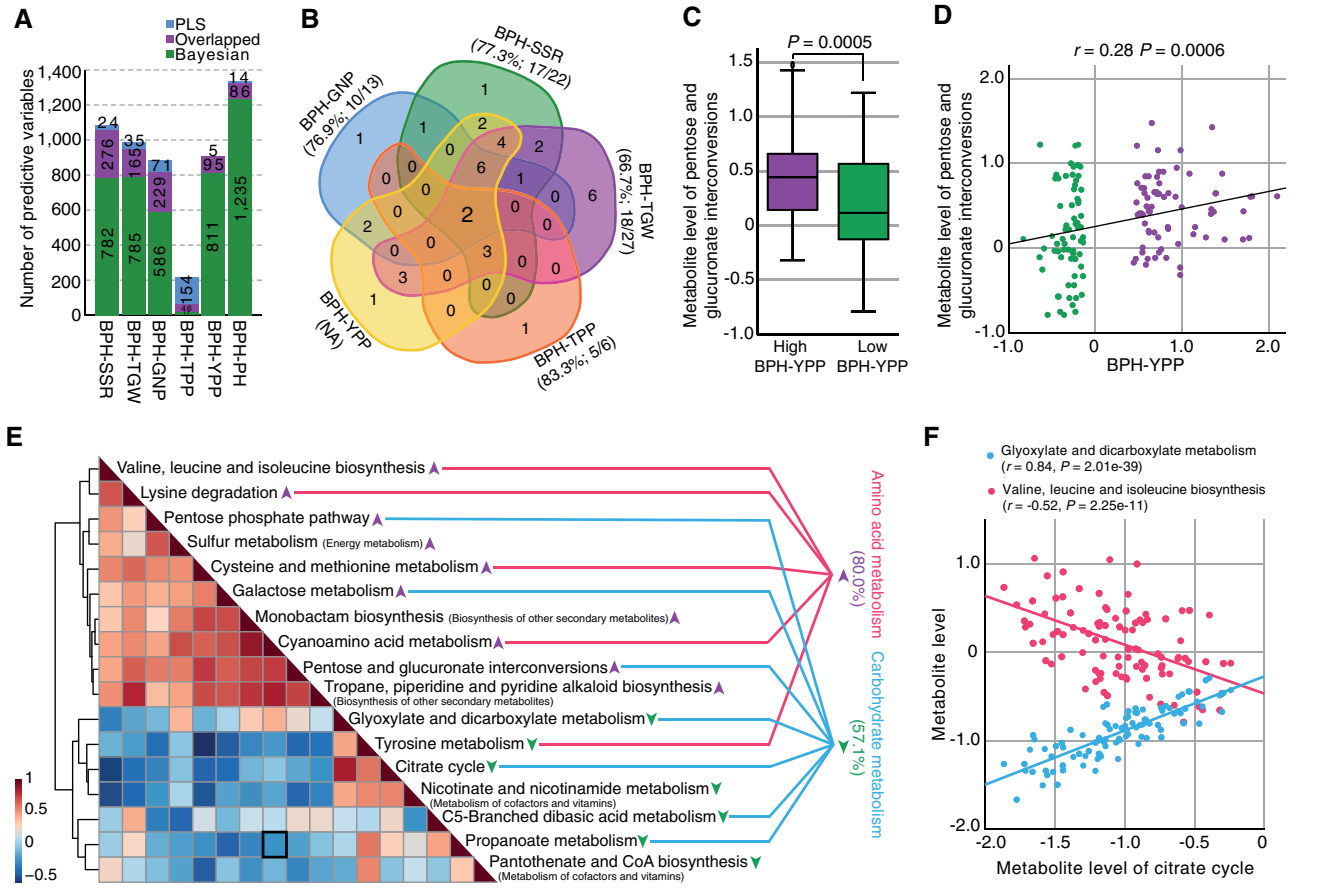

Figure 3 Enriched metabolic pathways for heterosis. A, Overlap of analytes between PLS regression and Bayesian method. B, Venn diagram of enriched pathways for heterosis of the five reproductive traits. The percentages of overlapping pathways for each of the four yield components with yield heterosis are correspondingly shown in brackets. The numbers of overlapping and per se enriched pathways for the four yield components are indicated at the left and right side of the slash, respectively. NA, not applicable. C, Comparison of metabolite levels of pentose and glucuronate interconversions between the high- and low-BPH-YPP hybrids. Independent samples *t* test, two-tailed. *N* = 72. The center line of each boxplot represents the 50th percentile. The bottom and top of each boxplot represent the 25th and 75th percentiles, respectively. The whiskers represent the minimum and maximum values. The circles represent outliers. D, Correlation between metabolite levels of pentose and glucuronate interconversions and yield heterosis. *N* = 144. E, Correlation pattern of significantly enriched pathways for yield heterosis. A total of 17 pathways were significantly enriched for yield heterosis, and Pearson correlations were performed among these pathways based on their quantitative information. The purple and green arrows indicate that the high-BPH-YPP hybrids had high or low metabolite levels, respectively. The percentages of regulated pathways from amino acid metabolism and carbohydrate metabolism are shown in brackets. The correlation between cyanoamino acid metabolism and propanoate metabolism is highlighted with a black square. F, Correlations between metabolite levels of the citrate cycle and two pathways from amino acid and carbohydrate metabolism. *N* = 144.

than those of grain number and grain weight (Figure 3B), which was consistent with the results shown in Figures 1, H and 2, A. With respect to quantitative information on the enriched pathways (the average levels of all metabolites per pathway), 77.3% of the pathways for yield heterosis showed significant differences between the high- and low-BPH hybrids (17 pathways; Figure 3C; Supplemental Tables S4 and S5), and 81.8% of those were significantly correlated with yield heterosis (Figure 3D; Supplemental Table S6). This result confirmed previously reported metabolites that have positive or negative correlations with grain yield or biomass at the pathway level and indicated that the metabolite levels of the enriched pathways were closely related to yield heterosis (Table 1).

We then investigated the correlations of the 17 significantly enriched pathways for yield heterosis, which were mainly from amino acid and carbohydrate metabolism. Two distinct clustering trends were found among the metabolic pathways (Figure 3E), and they were close to the correlation pattern of the 100 yield heterosis-associated analytes (Supplemental Figure S8). Because 114 of the analytes had already been successfully annotated, we converted the compound names of these metabolites into Kyoto Encyclopedia of Genes and Genomes (KEGG) IDs and mapped them to the KEGG metabolic pathways. A total of 18 metabolites were mapped to the pathways listed in Figure 3E, and six metabolites in the cyanoamino acid metabolism (L-phenylalanine, L-aspartate, and L-tyrosine) and propanoate metabolism (dihydroxyacetone phosphate, alpha-

Table 1 The enriched metabolic pathways for yield heterosis

Pathway name	*P*-value of enrichment analysis	*P*-value of *t* test	Metabolite level	Previously known metabolites	Species
Tyrosine Metabolism	0.046378	3.47E-10	Low	Succinic acid, tyrosine, maleic acid, dopamine, fumarate	*Arabidopsis* (Meyer et al., 2007; Sulpice et al., 2013), maize (Riedelsheimer et al., 2012b; Obata et al., 2015)
Pantothenate and CoA Biosynthesis	0.001271	3.54E-04	Low	Aspartate, valine	*Arabidopsis* (Meyer et al., 2007; Sulpice et al., 2010), tomato (Schauer et al., 2006), maize (Obata et al., 2015; de Abreu et al., 2017)
Propanoate Metabolism	0.001338	1.81E-04	Low	Succinic acid	*Arabidopsis* (Meyer et al., 2007), tomato (Schauer et al., 2006)
Nicotinate and Nicotinamide Metabolism	0.014907	2.60E-04	Low	Succinic acid, aspartate, fumarate, nicotinate, gamma-aminobutyric acid	*Arabidopsis* (Sulpice et al., 2013), tomato (Schauer et al., 2006), maize (Obata et al., 2015; de Abreu et al., 2017)
C5-Branched Dibasic Acid Metabolism	0.017663	1.31E-05	Low	Glutamate, 2-oxoglutarate, itaconate	*Arabidopsis* (Sulpice et al., 2010), tomato (Schauer et al., 2006), maize (Obata et al., 2015)
Citrate Cycle	0.00471	2.95E-08	Low	Succinic acid, citric acid, fumarate, malate	*Arabidopsis* (Meyer et al., 2007; Sulpice et al., 2013), tomato (Schauer et al., 2006), maize (Obata et al., 2015)
Glyoxylate and Dicarboxylate Metabolism	0.021109	1.72E-03	Low	Succinic acid, glutamine, citric acid, serine, glycine, 2-oxoglutarate, malate, glyceric acid, glutamate	*Arabidopsis* (Meyer et al., 2007; Sulpice et al., 2009, 2010, 2013), tomato (Schauer et al., 2006), maize (Obata et al., 2015)
Butanoate Metabolism	0.00072	0.17	Low	Succinic acid, maleic acid, glutamate, 2-oxoglutarate, fumarate, gamma-aminobutyric acid	*Arabidopsis* (Meyer et al., 2007; Sulpice et al., 2010, 2013), tomato (Schauer et al., 2006), maize (Obata et al., 2015)
Galactose Metabolism	0.008015	5.09E-05	High	Glycerol, raffinose, galactinol, glucose	Maize (Obata et al., 2015), *Miscanthus* (Maddison et al., 2017)
Pentose and Glucuronate Interconversions	0.014907	5.06E-04	High	Glycerol, xylose, xylitol	Maize (Obata et al., 2015)
Sulfur Metabolism	0.017663	4.67E-04	High	Succinic acid	Maize (Obata et al., 2015)
Cysteine and Methionine Metabolism	0.022276	2.28E-05	High	Aspartate	Maize (Obata et al., 2015)
Pentose Phosphate Pathway	0.022462	1.16E-06	High	Glycerate, glucose	Maize (Obata et al., 2015), *Miscanthus* (Maddison et al., 2017)
Monobactam Biosynthesis	0.029861	2.33E-03	High	Aspartate, threonine	Maize (Obata et al., 2015)
Tropane, Piperidine and Pyridine alkaloid Biosynthesis	0.030102	5.58E-04	High	Putrescine, nicotinate, nicotinate	*Arabidopsis* (Meyer et al., 2007), maize (de Abreu et al., 2017; Obata et al., 2015)
Lysine Degradation	0.001925	9.41E-03	High	Succinic acid	Maize (Obata et al., 2015)
Valine, Leucine and Isoleucine Biosynthesis	0.00705	2.89E-03	High	Valine, threonine	Maize (Obata et al., 2015)
Cyanoamino acid Metabolism	0.043081	2.85E-02	High	Glycine, tyrosine, asparagine	*Arabidopsis* (Gärtner et al., 2009; Sulpice et al., 2013)
	0.022462	0.08	High	–	–

(continued)

Table 1 Continued

Pathway name	*P*-value of enrichment analysis	*P*-value of *t* test	Metabolite level	Previously known metabolites	Species
Phenylalanine, Tyrosine and Tryptophan Biosynthesis					
Glycine, Serine and Threonine Metabolism	0.01074	0.33	High	Glycerate, threonine, aspartate,	Maize (Obata et al., 2015)
Pyruvate Metabolism	0.001879	0.32	High	Succinic acid, fumarate	Maize (Obata et al., 2015)
Phenylalanine Metabolism	0.036123	0.67	High	Benzoic acid, succinic acid, fumarate	*Arabidopsis* (Sulpice et al., 2013), maize (Obata et al., 2015)
Synthesis and Degradation of Ketone Bodies	0.004325	–	–	–	–

The pathway name, *P*-value, metabolite level, previously known metabolites, and corresponding species are provided. Since two pathways have no reported metabolites and one pathway's quantitative information is not available, corresponding areas are marked with horizontal lines.

hydroxybutyric acid, and pyruvaldehyde) pathways were selected for further correlation analysis. Metabolites in the same pathways had significant positive correlations, and metabolites in different pathways had significant negative or no correlations (Supplemental Table S7). As shown in Figure 3E, the average levels of the six metabolites in the two pathways were significantly negatively correlated (Supplemental Figure S9). After the metabolite levels of the enriched pathways were compared between the high- and low-BPH hybrids, we found that all pathways involved in amino acid metabolism, except for tyrosine metabolism, had high metabolite levels in high-BPH hybrids, and 57.1% of the pathways from carbohydrate metabolism had low metabolite levels in high-BPH hybrids (Supplemental Table S5). Because negative correlations existed between the metabolite levels of amino acid and carbohydrate metabolism (Figure 3F; Supplemental Table S6), we speculated that higher metabolite levels of amino acid metabolism and lower metabolite levels of carbohydrate metabolism were closely related to a higher degree of yield heterosis.

With respect to the four yield components, the significantly enriched pathways showed different correlation manners across traits, and most of the manners were similar to those of corresponding heterosis-associated analytes (Supplemental Figures S10–S12). Accordingly, we constructed a metabolomic landscape for heterosis of both reproductive and vegetative traits through overlapping pathways (Figure 4). In concordance with the yield heterosis—as shown in Figure 3E—most of the significantly enriched pathways from amino acid metabolism demonstrated positive correlations with heterosis of grain weight (100%) and seed setting rate (66.7%), and the pathways from carbohydrate metabolism were negatively correlated (100% and 25%, respectively). In contrast to the reproductive traits, 83.3% of the enriched pathways from amino acid metabolism were negatively correlated with PH heterosis, and 75% of those from carbohydrate metabolism were positively correlated. Thus, the metabolite levels of the significantly enriched pathways (especially those in amino acid and carbohydrate metabolism) for the four yield components always had consistent correlation patterns with the degree of yield heterosis, whereas those for vegetative trait (PH) manifested opposite relationships with the five reproductive traits (yield and yield components).

The enriched pathways are predictive of yield heterosis

Based on the metabolite levels of the significantly enriched pathways for yield heterosis, we performed biomarker analysis by calculating the ratios of all pathway pairs, which can increase the chance of identifying individual biomarkers (Chong et al., 2019). The univariate receiver operating characteristic (ROC) curve analysis showed that a cutoff of 0.551 for ratios of tyrosine metabolism and sulfur metabolism could distinguish between the high- and low-BPH hybrids, with an area under the curve (AUC) equal to 0.836 (Figure 5, A and B; Supplemental Table S8). When multivariate ROC curve analysis was performed to identify biomarkers, the AUC increased to 0.907, and the predictive accuracy was 0.827 (Figure 5, C and D). The best model contained only 10 features; tyrosine metabolism was highly important and was frequently selected (Figure 5E; Supplemental Figure S13 and Supplemental Table S9), demonstrating the critical role of tyrosine metabolism in yield heterosis.

We investigated the relationship between the metabolite levels of L-tyrosine and yield heterosis in the whole hybrid population and found no significant correlation (Figure 5F). However, the average levels of the five annotated metabolites that participate in tyrosine metabolism (some of which had significant negative correlations with yield heterosis; Supplemental Table S10), namely, L-tyrosine, maleic acid, atrolactic acid, 4-hydroxycinnamic acid, and 1,4-dihydroxybenzene, were significantly negatively correlated with yield heterosis ($r = -0.23$; Figure 5G). Furthermore, we evaluated the impact of changes in pathway information on

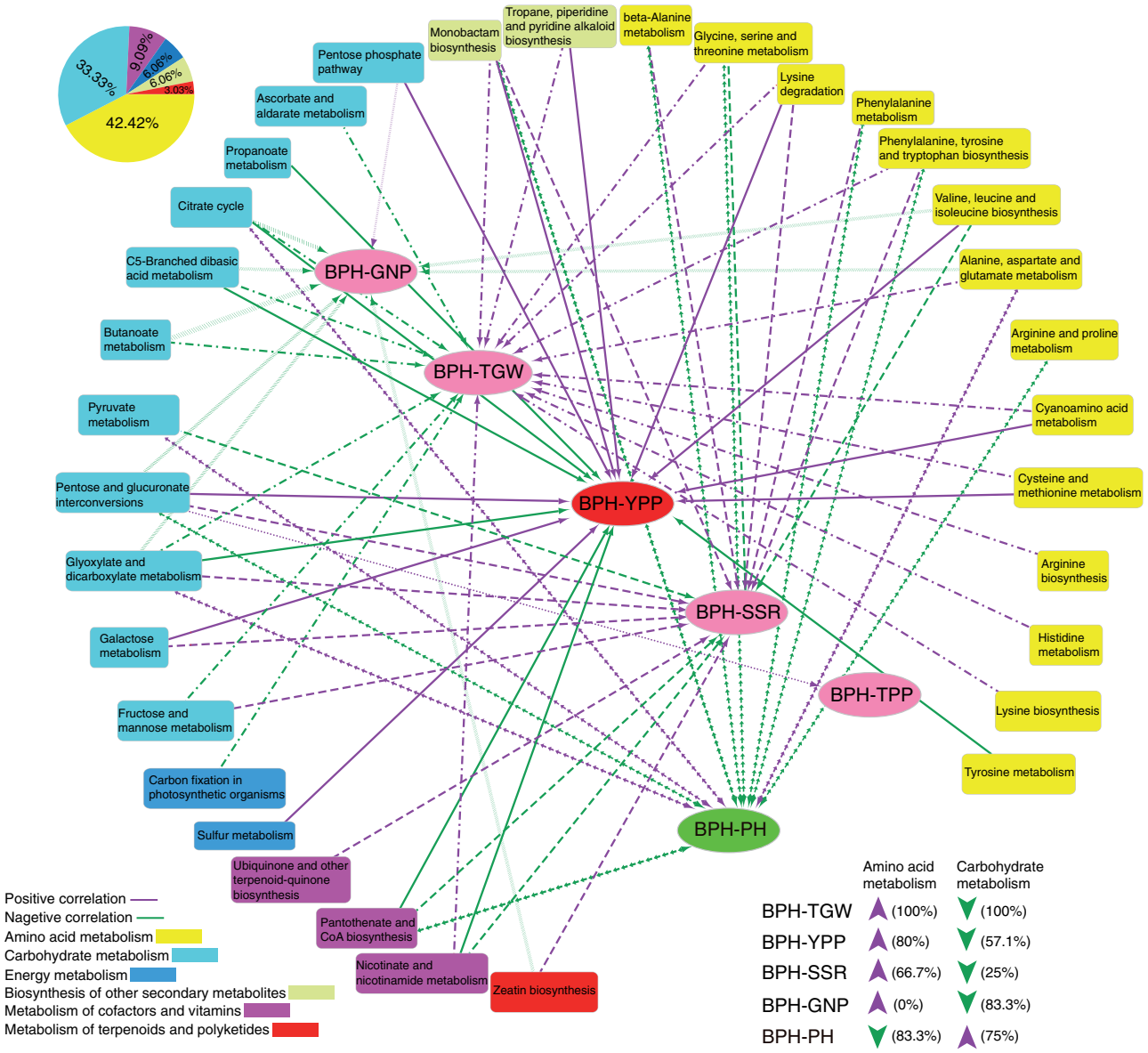

Figure 4 Metabolomic landscape of heterosis for six agronomic traits. The landscape of heterosis was created by the overlapping metabolic pathways between traits. All the significantly enriched pathways from amino acid metabolism were positively correlated with heterosis of grain weight, and all the pathways from carbohydrate metabolism were negatively correlated. Similarly, four of six significantly enriched pathways from amino acid metabolism displayed positive correlations with heterosis of seed setting rate, and one out of four pathways from carbohydrate metabolism displayed a negative correlation. Eight significantly enriched pathways for grain number (namely, zeatin biosynthesis, two pathways in amino acid metabolism, and five in carbohydrate metabolism) showed negative relationships, and the pentose phosphate pathway showed a positive correlation. Only one pathway was significantly enriched for tiller heterosis, and the metabolite levels of pentose and glucuronate interconversions were positively correlated with tiller heterosis. In contrast to the above-mentioned correlation patterns, five out of six significantly enriched pathways in amino acid metabolism showed negative correlations with heterosis of PH, and three out of four pathways in carbohydrate metabolism showed positive correlations. Pearson correlation analysis was performed based on the metabolite levels of the significantly enriched pathways, and a correlation was significant when the $P < 0.05$. Positive and negative correlations are indicated in different colors. The metabolic pathways from different types are marked correspondingly. Purple and green arrows indicate high-BPH hybrids with high or low metabolite levels, respectively. Numbers in brackets represent percentages of regulated pathways from amino acid and carbohydrate metabolism.

predictions by adding new metabolites to tyrosine metabolism, given that KEGG or other databases are dynamic and more metabolites can be identified and added to a metabolic pathway. We first included two putatively annotated metabolites (succinate and acetoacetate) when calculating the metabolite levels of tyrosine metabolism. The correlation coefficient increased to 0.28 when succinate was added (P = 1.0e-6), and it further changed to 0.34 after the two metabolites were used (P = 2.0e-9; Supplemental Figure S14). However, the correlation coefficients decreased when using other metabolites (uracil and L-phenylalanine) that are not involved in tyrosine metabolism (Supplemental Figure S14).

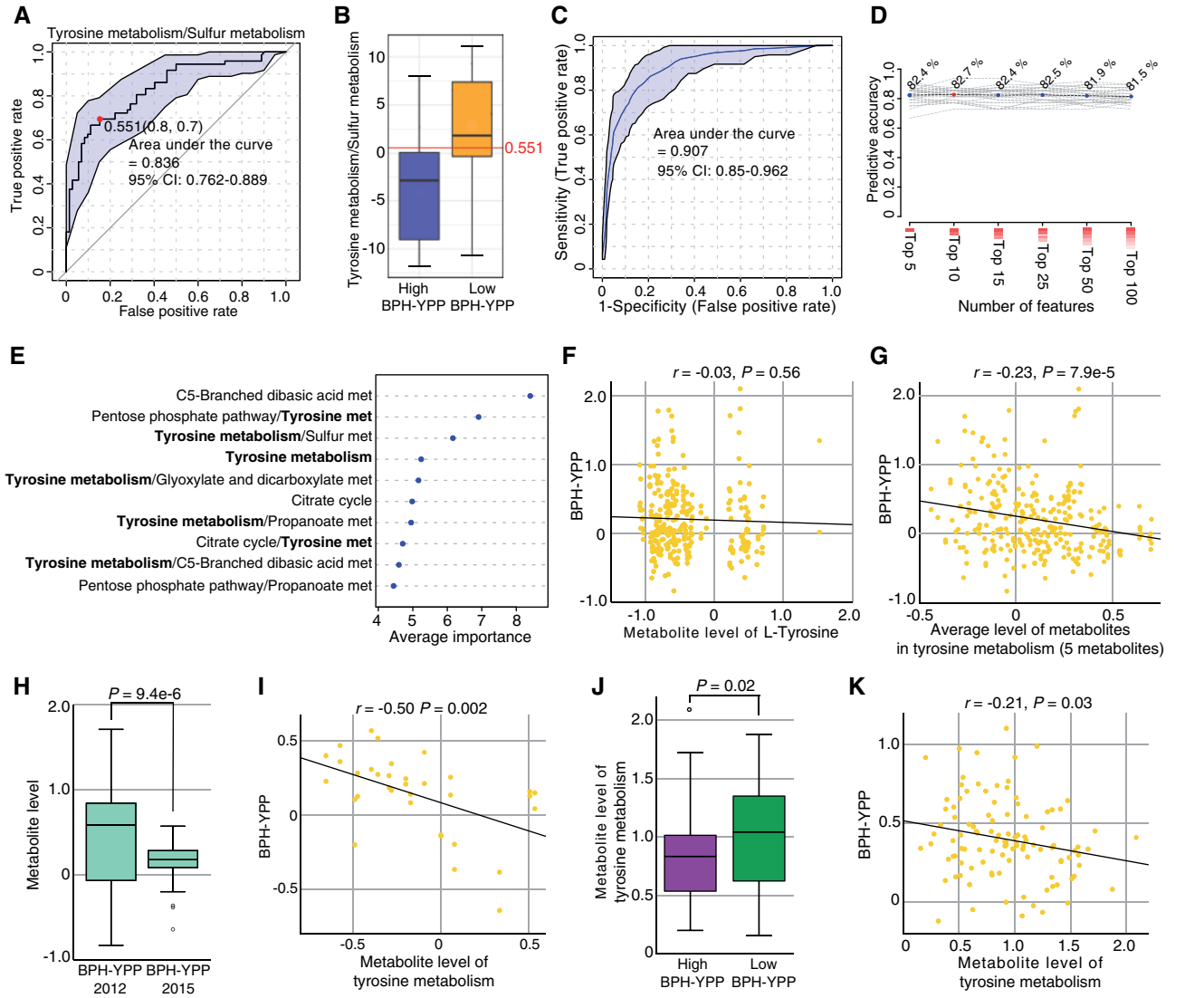

Figure 5 The enriched pathways are predictive of yield heterosis. A, AUC for the ratio of tyrosine metabolism to sulfur metabolism. Univariate ROC curve analysis was performed on high- and low-BPH-YPP hybrids from the diallel cross population to identify biomarkers. The shadow is the computed 95% confidence band. B, Box plot of ratios of tyrosine metabolism to sulfur metabolism. The red line indicates the optimal cutoff value. $N = 72$. C, AUC for the top 10 features based on the multivariate ROC curve analysis. D, Predictive accuracies with different numbers of features. E, Average importance of the top 10 features. Met = metabolism. F, Correlation between the metabolite levels of L-tyrosine and yield heterosis. $N = 287$. G, Correlation between the average metabolite levels of the five annotated metabolites in tyrosine metabolism and yield heterosis. $N = 287$. H, Comparison of yield heterosis for 34 hybrids across growth conditions. Paired samples t test, two-tailed. $N = 33$. I, Correlation between the metabolite levels of tyrosine metabolism and yield heterosis of the 34 hybrids grown under different conditions. $N = 34$. J, Comparison of the metabolite levels of tyrosine metabolism between the high- and low-BPH-YPP hybrids ($N = 53$ and 54, respectively) from a testcross population. K, Correlation between the metabolite levels of tyrosine metabolism and yield heterosis of the testcross population ($N = 107$). The center line of each boxplot represents the 50th percentile. The bottom and top of each boxplot represent the 25th and 75th percentiles, respectively. The whiskers represent the minimum and maximum values. The circles represent outliers.

Thus, the metabolite levels of tyrosine metabolism, rather than those of L-tyrosine alone, were predictive of yield heterosis, and the performance of pathway biomarkers was determined by the completeness and accuracy of the pathway information.

To validate the contribution of quantitative changes in tyrosine metabolism in predicting yield heterosis, both univariate and multivariate ROC curve analyses were performed on the metabolite levels of 34 hybrids with different performances across growth conditions (Figure 5H). Tyrosine metabolism functioned as a critical feature in both analyses (Supplemental Figures S15 and 16; Supplemental Tables S11 and 12), and a significant negative correlation was found between tyrosine metabolism and yield heterosis (Figure 5I). Subsequently, we obtained the metabolite levels of tyrosine metabolism from another testcross population containing 107 hybrids (Supplemental Table S13). As shown in Figure 3E, the metabolite levels of tyrosine metabolism in the high-

BPH group were significantly lower than those in the low-BPH group (Figure 5J). Furthermore, the metabolite levels of tyrosine metabolism showed a significant negative correlation with yield heterosis (Figure 5K). Thus, the metabolite levels of the significantly enriched pathways were predictive of yield heterosis across environments and populations.

Discussion

With the rapid developments in systems biology, the elucidation of molecular mechanisms and exploration of biomarkers based on metabolic pathways for complex phenotypes can accelerate the establishment of precision design programs, such as precision breeding or precision medicine. In this study, untargeted metabolite profiles and computational analyses were combined to explore the metabolomic mechanisms underlying heterosis of six agronomic traits in rice. Consistent with previous findings (Dan et al., 2019, 2020), we found that the average parental metabolite levels, which are additive metabolite profiles, are appropriate predictors for diverse over-dominant phenotypes (better parent heterosis). The changes from metabolomic additive effects to phenotypic over-dominance effects may be partially explained by the combination of hierarchical structure and multiplicative interactions of complex traits (Dan et al., 2015). Additional systematic analyses—incorporating both hybrid individuals and populations—can be performed in the near future. We determined the optimal number of heterosis-associated analytes for each trait by performing the PLS regression multiple times. This strategy makes possible the optimal selection of features for diverse phenotypes (Sprenger et al., 2018; Dan et al., 2019; Hu et al., 2019). In evaluating the performance of PLS or random forest models, changes in the number of predictive variables (top 50–3,746 predictive analytes in Figure 1C and top 5–100 predictive features in Figure 5D) yielded slight variations in predictive models, which are similar to the finding of predicting potato drought tolerance using the random forest method (Sprenger et al., 2018). We speculate that this phenomenon may arise from the inclusion of the most contributed predictive variables, namely, the top 50 analytes in Figure 1C and top 5 features in Figure 5D, in predictive models. We also analyzed the connections between metabolite levels of specific analytes and heterosis of multiple traits, which are rarely reported in previous studies (Dan et al., 2016; Xu et al., 2016; Wilmanski et al., 2019). The overlapping heterosis-associated analytes were found to underlie the association patterns among traits. The metabolic pathways involved in heterosis were finally identified through dysregulated network analysis of the high- and low-BPH hybrids, among which the high-performance hybrids are usually selected by plant breeders, and the correlation patterns of the significantly enriched pathways were similar to those of the corresponding heterosis-associated analytes. However, we were unable to pair the analytes and metabolic pathways because the number of annotated metabolites was rather low (3% of all detected analytes), and the functions of the lipids (which account for about 50% of the annotated metabolites) were mostly unknown. The annotation of new metabolites and functional analyses are urgently required to obtain more details about the connections between predictive analytes and enriched metabolic pathways.

Pathway biomarkers were developed for yield heterosis based on quantitative information on significantly enriched metabolic pathways, and the performance of these biomarkers was validated with hybrids across environments and populations. Because all metabolites per pathway, rather than a single metabolite, were used for the calculation of metabolite levels, the pathway biomarkers may overcome the negative effects of molecular heterogeneity in predicting individuals with the same performance (Menche et al., 2017; Guo et al., 2019). In addition, the changes in molecular levels that are triggered by environmental discrepancies can also be "buffered" by the pathway biomarkers with the inclusion of both significant and "insignificant" variables in predictive models, which may contribute to the breeding of adaptive varieties (Varshney et al., 2018; Hickey et al., 2019). The robust predictive power of the pathway biomarkers was unexpected, given that the predictability of grain weight and yield heterosis with sets of metabolites was <0.8 in previous studies (Dan et al., 2019, 2020). The metabolite levels of tyrosine metabolism were stable biomarkers for both the training and validation sets, and the average levels of the five metabolites involved in tyrosine metabolism also displayed a significant negative correlation with yield heterosis. However, the metabolite levels of L-tyrosine showed no significant correlation with yield heterosis. We believe that the metabolomic biomarkers identified in this study emphasize quantitative changes in enriched metabolic pathways rather than differences between metabolites. The metabolite levels of L-tyrosine may have significant negative correlations with yield heterosis, and the remaining metabolites involved in tyrosine metabolism (which had significant negative correlations with yield heterosis) in this study can have no correlation with yield heterosis in other hybrid populations. This contradiction can be understood as metabolomic heterogeneity among populations, similar to the expressional heterogeneity of complex diseases among patients (Menche et al., 2017; Guo et al., 2019). Furthermore, the latest findings demonstrate that changes in metabolite levels of steroid hormone biosynthesis are precisely timed to gestation in pregnant women (Liang et al., 2020). Thus, we anticipate that refined pathway biomarkers based on omics analyses, including genomics (Riedelsheimer et al., 2012a; Millet et al., 2019), transcriptomics (Sprenger et al., 2018; Azodi et al., 2020), proteomics (Zhang et al., 2016; Dou et al., 2020), and lipidomics (Aviram et al., 2016; de Abreu et al., 2018), may provide better predictions than the traditional sets of predictive variables.

The prevailing negative correlations between metabolite levels of amino acid metabolism and carbohydrate metabolism suggest that focusing on the regulation of specific metabolic pathways may facilitate the conformation of yield

heterosis. With respect to the metabolomic connections of heterosis among traits, the significantly enriched pathways for the yield components always had similar correlation patterns with yield heterosis, whereas that for PH showed an opposite relationship with yield heterosis. Thus, we speculate that there is a rough balance between amino acid metabolism and carbohydrate metabolism in yield heterosis (Dan et al., 2015, 2020), and this balance may originate from metabolomic connections of the remaining reproductive traits (yield components) and vegetative traits (yield-related traits) with different degrees of contribution. The strategy of investigating metabolomic connections between the component and complex traits through overlapping pathways may be used to analyze molecular connections among different complex human diseases—with the knowledge that patients with different diseases share sets of disease-associated genes (Barabasi et al., 2011; Menche et al., 2015, 2017).

Our results provide a metabolomic landscape of heterosis in rice, as well as an evaluation of the application potential of biomarkers based on enriched pathways for yield heterosis. Optimal balances among specific metabolic pathways and reproductive and vegetative traits are critical for yield heterosis. Quantitative changes in pathway biomarkers predict yield heterosis without considering discrepancies in growth conditions and hybrid populations, indicating the wide application potential of pathway biomarkers for predicting complex phenotypes and thus achieving precision design programs.

Materials and methods

Plant materials and phenotyping

Eighteen traditional rice (*O. sativa*) cultivars that include both *indica* and *japonica* were parents of one hybrid population, with a complete diallel cross design (Dan et al., 2020). Phenotypic data of five reproductive traits, namely, seed setting rate, thousand-grain weight (Dan et al., 2019), grain number per panicle, tiller number per plant, and yield per plant (YPP, Dan et al., 2020), were collected at the maturation stage. Plant height was also measured at the maturation stage. Trait values of the 18 parents and 287 hybrids were collected and used for the analyses. Another testcross population consisted of a Honglian-type cytoplasmic male-sterile line (Yuetai A) and recombinant inbred lines (F_5). The YPP of the maintainer line (Yuetai B), 107 pairs of parent hybrids, was measured at the maturation stage. A total of 34 hybrids that were reciprocals from the diallel cross population were replanted with the testcross population, and their yield performance was recorded for analysis. Details such as locations, planting time, and plant densities of the two hybrid populations were described in a previous study (Dan et al., 2019).

Metabolomics

Metabolite profiling analysis of the parental seedlings was performed as described previously (Dan et al., 2020). Briefly, untargeted metabolite profiles of 15-d-old seedlings were collected with a 1290 Infinity liquid chromatography system (Agilent Technologies, Santa Clara, CA, USA), Agilent quadrupole time-of-flight mass spectrometer (Agilent 6550 iFunnel QTOF; Agilent Technologies, Santa Clara, CA, USA), and Triple TOF 6600 mass spectrometer (AB SCIEX, Foster City, CA, USA). The metabolites were annotated using an in-house standard spectral library, and the lipids were annotated through matching with an in-house tandem mass spectrometry (MS/MS) spectral library. Data reliability was checked using a quality control sample, and the metabolite levels of a total of 3,746 detected analytes, among which 114 metabolites were annotated using the in-house spectral libraries, were normalized (sum, log, and none) for the statistical analyses.

Identification of heterosis-associated analytes

To identify analytes that were closely associated with heterosis of each trait, we used the PLS regression method (Wold, 1975). PLS is an iterative algorithm with the involvement of latent factors and is suitable for conducting multivariate analysis when the number of predictor variables (*X* variables) significantly exceeds that of response variables (*Y* variables). The latent factors or latent variables, which can be numerically assessed and provide consistent information for further development of predictive models (Wold, 1975), are formed to not only maximize the explained variance of predictive variables, but also to maximize the covariance of observations (Bijlsma et al., 2006). Values of BPH and the means of parental metabolite levels were *X* and *Y* variables, respectively. The number of latent factors was first set to 50, and the largest number of extracted latent factors was 17. The number of latent factors was then set to three or four, at which the *r* value was the highest among predictive models with different numbers of latent factors, to perform the second regression. To evaluate the performance of the PLS-based models, both cross-validation and permutation test were performed to check whether the models were overfitted. Hybrids from the diallel cross population were divided into high- and low-BPH groups according to the 75th and 25th percentiles of heterosis of each trait. The PLS-discriminant analysis was then performed with the module "Statistical Analysis" on MetaboAnalyst (www.metaboanalyst.ca; Xia and Wishart, 2011). The 10-fold cross-validation method was used, and three parameters were provided to describe the model performance: prediction accuracy, sum of squares of the model (R^2), and cross-validated R^2 (i.e. Q^2; Wold et al., 2001). The separation distance (B/W), which is the ratio of the between-group sum of squares (B) and the within-group sum of squares (W; Bijlsma et al., 2006), was selected for the permutation test (2,000 permutations). The relationship of the B/W distribution between the original and permutated data is indicated by the observed statistical *P*-value. Subsequently, the values of variable importance in the projection, which are the weighted sums of squares of the model's weights (Wold et al., 2001), of the three or four latent factors were averaged to evaluate the importance of each analyte. To remove redundant feature information, the

top 2,000, 1,500, 1,000, 500, 300, 200, 100, 50, 25, 10, and 5 analytes from the 3,746 predictive analytes were selected for multiple PLS regressions. The optimal number of predictive analytes for each trait was determined when r plateaued. The predictive analytes chosen for multiple traits were treated as overlapping heterosis-associated analytes. The parameters for heterosis-associated analytes and constants were used to describe the connections between metabolite levels and heterosis.

Dysregulated network analysis

To identify the metabolic pathways involved in heterosis of the six traits, pathway enrichment analysis was performed on the diallel cross population. Because of the fact that only 114 metabolites (3% of all detected analytes) had been annotated using the in-house standard spectral libraries, it was difficult to conduct pathway enrichment analysis using traditional strategies. Thus, we utilized the metabolic reaction network-based recursive algorithm (MetDNA; Shen et al., 2019), which can achieve large-scale metabolite annotations for untargeted metabolomics without the dependence of comprehensive standard spectral libraries. The principle of MetDNA is that metabolites in a reaction pair with similar structures tend to have similar MS2 spectra. With the availability of a small library of MS2 spectra, MetDNA significantly and progressively expanded the number of annotated metabolites through the recursive algorithm. The dysregulated metabolic peaks were first discovered using a univariate test (Student's t test or Mann–Whitney–Wilcoxon test), and the dysregulated peaks with annotations were then mapped to the KEGG metabolic pathways. The metabolite level of one dysregulated pathway was the average level of all annotated metabolites in the pathway. To ensure the sensitivity and specificity of the pathway biomarkers, the diallel cross population was divided into high and low parts based on the 75th and 25th percentiles of the heterosis of each trait. When performing dysregulated network analysis with the MetDNA web server (http://metdna.zhulab.cn), the high- and low-BPH hybrids (hybrids with heterosis $\geqslant$75th and $\leqslant$25th percentiles, respectively) were the control and case groups, respectively. Analytes with m/z, retention time, and average parental metabolite levels constituted the MS1 peak table, and the raw MS/MS files (mgf format) of a quality control sample (two injections) were the MS2 data files. The corresponding parameters were as follows: ionization polarity, negative; liquid chromatograph, RP; MS instrument, Sciex TripleTOF; collision energy, 35 ± 15; univariate statistics, Student's t test; species: *Arabidopsis thaliana* (Thale Cress); cutoff P-value, 0.05; P-value adjustment, yes. For the testcross population, the hybrids were divided into two parts (54 hybrids and 53 hybrids) in the dysregulated network analysis, according to the values of yield heterosis. Metabolic pathways were grouped according to the KEGG pathway database (https://www.genome.jp/kegg/pathway.html; Kanehisa et al., 2014).

ROC curve analysis

Quantitative information on the significantly enriched pathways for yield heterosis was used for the ROC curve analysis with the module "Biomarker Analysis" on MetaboAnalyst (Xia and Wishart, 2011). In the normalization procedures for both univariate and multivariate ROC curve analyses, none was performed for sample normalization and data scaling. The top 100 metabolite ratios (viz. pathway ratios) were computed and included to facilitate the identification of individual biomarkers (Chong et al., 2019). The top 20 metabolite ratios were computed and included in the ROC curve analyses of the 34 hybrids. Random forest (Breiman, 2001) was selected as the classification and the feature ranking method in the multivariate ROC curve analysis. To ensure the performance of random forest models, the "Biomarker Analysis" module performs Monte Carlo cross-validation through balanced subsampling. In each cross-validation, two-thirds of the hybrids were used to evaluate feature rankings, and the top 2, 3, 5, 10, etc., important analytes were selected to build classification models, which were then validated with one-third of the hybrids. The cross-validation procedures were repeated 500 times to calculate the performance and 95% confidence interval (95% confidence band) for each model.

Statistical analyses

Pearson correlations between heterosis and transformed parental metabolite levels, among heterosis of the investigated traits (pairwise) and among heterosis-associated analytes (pairwise), were performed using the analysis path of "Correlation Heatmaps" in the module "Statistical Analysis" on MetaboAnalyst (Xia and Wishart, 2011). Correlations with $P < 0.05$, were considered significant. Empirical Bayesian analysis of differential analytes for the high- and low-BPH groups was performed with the analysis path of "Empirical Bayesian Analysis of Metabolites." An equal group variance was assumed, and 0.9 was set as the fudge factor (a0) and posterior delta. Unpaired t tests (adjusted P-value cutoff: 0.05) with equal group variance were performed between the high- and low-BPH groups with the analysis path of "T tests." Compound names of the annotated metabolites were converted into KEGG IDs with the analysis path of "Compound ID Conversion" in the module "Other Utilities." PLS regressions of BPH and metabolite levels were performed using SPSS (IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.). Partial correlations (two-tailed) between the four yield components/PH and yield were performed to investigate the contribution of the four yield components and PH to yield heterosis using SPSS. The two analyzed traits were variables, and the remaining four traits were treated as control variables in partial correlations. Pearson correlations (two-tailed) between the observed and predicted BPH, or between metabolite levels and BPH, were implemented using SPSS, with the correlation coefficient as predictability. Stepwise regression was used to describe yield heterosis (dependent variable) with the four components and PH (independent variables) using SPSS.

Independent samples *t* test (two-tailed) and paired samples *t* test (two-tailed) were used to compare the differences in pathway levels between the high- and low-BPH hybrids and phenotypic differences in the 34 hybrids across growth conditions using SPSS. Venn diagrams were drawn using a webtool from http://bioinformatics.psb.ugent.be/webtools/Venn.

Accession numbers

All phenotypic data were provided in supporting information and the raw metabolite profiles were deposited in the metabolomic database: MetaboLights (MTBLS742; Dan et al., 2020).

Supplemental data

The following materials are available in the online version of this article.

Supplemental Figure S1. Heatmap for correlations between heterosis and transformed parental metabolite levels.

Supplemental Figure S2. Determining the number of latent factors for heterosis of each trait.

Supplemental Figure S3. Cross-validation and permutation test of the PLS-based models.

Supplemental Figure S4. Scatter plots for observed and metabolome-predicted heterosis of the four yield components.

Supplemental Figure S5. Correlations between the metabolite levels of overlapping heterosis-associated analytes and heterosis of yield and yield components.

Supplemental Figure S6. Scatter plots for observed and yield components-predicted yield heterosis.

Supplemental Figure S7. Correlations between the metabolite levels of overlapping heterosis-associated analytes and heterosis of PH and yield and yield components.

Supplemental Figure S8. Heatmap for correlations among the 100 yield heterosis-associated analytes.

Supplemental Figure S9. Correlation between average levels of metabolites in cyanoamino acid metabolism and propanoate metabolism.

Supplemental Figure S10. Heatmaps for correlations among the screened heterosis-associated analytes and enriched metabolic pathways for SSR.

Supplemental Figure S11. Heatmaps for correlations among the screened heterosis-associated analytes and enriched metabolic pathways for TGW.

Supplemental Figure S12. Heatmaps for correlations among the screened heterosis-associated analytes and enriched metabolic pathways for grain number per plant.

Supplemental Figure S13. Multivariate ROC curve analysis of high- and low-BPH hybrids from the diallel cross population.

Supplemental Figure S14. Correlations between the average levels of metabolites and yield heterosis.

Supplemental Figure S15. AUC for tyrosine metabolism based on the univariate ROC curve analysis of 34 hybrids.

Supplemental Figure S16. Multivariate ROC curve analysis of the 34 hybrids.

Supplemental Table S1. Phenotypic data of parents and hybrids.

Supplemental Table S2. Overlapping heterosis-associated analytes among traits.

Supplemental Table S3. Correlations between metabolite levels of overlapping heterosis-associated analytes and BPH.

Supplemental Table S4. Metabolite levels of the enriched pathways for yield heterosis of the high- and low-BPH hybrids from the diallel cross population.

Supplemental Table S5. *T* test of metabolite levels of the enriched pathways for yield heterosis.

Supplemental Table S6. Correlations between metabolite levels of the enriched pathways and yield heterosis.

Supplemental Table S7. Correlations of six metabolites in cyanoamino acid metabolism and propanoate metabolism.

Supplemental Table S8. Univariate ROC curve analysis of the 17 significantly enriched pathways for yield heterosis.

Supplemental Table S9. Multivariate ROC curve analysis of the 17 significantly enriched pathways for yield heterosis.

Supplemental Table S10. Correlations between yield heterosis and the five annotated metabolites in tyrosine metabolism.

Supplemental Table S11. Univariate ROC curve analysis of the significantly enriched pathways for yield heterosis of the 34 hybrids.

Supplemental Table S12. Multivariate ROC curve analysis of the significantly enriched pathways for yield heterosis of the 34 hybrids.

Supplemental Table S13. Metabolite levels of the enriched pathways for hybrids from the testcross population.

Acknowledgments

We thank members from the 3134 Laboratory for assistance of collecting phenotypic data and valuable suggestions. We are grateful to David R. Gang for useful advising. And we thank the Shanghai Applied Protein Technology Co., Ltd. for helping untargeted metabolite profiling analysis.

Funding

This research was supported by the National Key R&D Program of China (grant no. 2017YFD0100400), National Natural Science Foundation of China (grant nos. 31771746 and 31801439), National Rice Industry Technology System (grant no. CARS-01-07) and the China Postdoctoral Science Foundation (grant nos. 2018M632910 and 2019M660186).

Conflict of interest statement. The authors declare no conflict of interest statement.

References

Aviram R, Manella G, Kopelman N, Neufeld-Cohen A, Zwighaft Z, Elimelech M, Adamovich Y, Golik M, Wang C, Han X, et al. (2016) Lipidomics analyses reveal temporal and spatial lipid organization and uncover daily oscillations in intracellular organelles. Mol Cell **62:** 636–648

Azodi CB, Pardo J, VanBuren R, de Los Campos G, Shiu SH (2020) Transcriptome-based prediction of complex traits in maize. Plant Cell **32:** 139–151

Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet **12:** 56–68

Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, van Ommen B, Smilde AK (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal Chem **78:** 567–574

Breiman L (2001) Random forests. Mach Learn **45:** 5–32

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

Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, Wishart DS, Xia J (2018) MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res **46:** W486–W494

Chong J, Wishart DS, Xia J (2019) Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. Curr Protoc Bioinform **68:** e86

Crossa J, Perez-Rodriguez P, Cuevas J, Montesinos-Lopez O, Jarquin D, de Los Campos G, Burgueno J, Gonzalez-Camacho JM, Perez-Elizalde S, Beyene Y, et al. (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci **22:** 961–975

Dan Z, Chen Y, Xu Y, Huang JR, Huang JS, Hu J, Yao G, Zhu Y, Huang W (2019) A metabolome-based core hybridisation strategy for the prediction of rice grain weight across environments. Plant Biotechnol J **17:** 906–913

Dan Z, Chen Y, Zhao W, Wang Q, Huang W (2020) Metabolome-based prediction of yield heterosis contributes to the breeding of elite rice. Life Sci Alliance **3:** e201900551

Dan Z, Hu J, Zhou W, Yao G, Zhu R, Huang W, Zhu Y (2015) Hierarchical additive effects on heterosis in rice (Oryza sativa L.). Front Plant Sci **6:** 738

Dan Z, Hu J, Zhou W, Yao G, Zhu R, Zhu Y, Huang W (2016) Metabolic prediction of important agronomic traits in hybrid rice (*Oryza sativa* L.). Sci Rep **6:** 21732

Darwin CR (1876) The Effects of Cross and Self Fertilization in the Vegetable Kingdom. John Murray, London, UK

de Abreu ELF, Li K, Wen W, Yan J, Nikoloski Z, Willmitzer L, Brotman Y (2018) Unraveling lipid metabolism in maize with time-resolved multi-omics data. Plant J **93:** 1102–1115

de Abreu ELF, Westhues M, Cuadros-Inostroza Á, Willmitzer L, Melchinger AE, Nikoloski Z (2017) Metabolic robustness in young roots underpins a predictive model of maize hybrid performance in the field. Plant J **90:** 319–329

Dou Y, Kawaler EA, Cui Zhou D, Gritsenko MA, Huang C, Blumenberg L, Karpova A, Petyuk VA, Savage SR, Satpathy S, et al. (2020) Proteogenomic characterization of endometrial carcinoma. Cell **180:** 729–748

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

Gui J, Shen J, Li L (2011) Functional characterization of evolutionarily divergent 4-coumarate:coenzyme a ligases in rice. Plant Physiol **157:** 574–586

Guo WF, Zhang SW, Zeng T, Li Y, Gao J, Chen L (2019) A novel network control model for identifying personalized driver genes in cancer. PLoS Comput Biol **15:** e1007520

Hickey LT, A NH, Robinson H, Jackson SA, Leal-Bertioli SCM, Tester M, Gao C, Godwin ID, Hayes BJ, Wulff BBH (2019) Breeding crops to feed 10 billion. Nat Biotechnol **37:** 744–754

Hu X, Xie W, Wu C, Xu S (2019) A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol J **17:** 2011–2020

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res **42:** D199–D205

Li Y, Kim JI, Pysh L, Chapple C (2015) Four isoforms of arabidopsis 4-coumarate:CoA ligase have overlapping yet distinct roles in phenylpropanoid metabolism. Plant Physiol **169:** 2409–2421

Liang L, Rasmussen MH, Piening B, Shen X, Chen S, Rost H, Snyder JK, Tibshirani R, Skotte L, Lee NC, et al. (2020) Metabolic dynamics and prediction of gestational age and time to delivery in pregnant women. Cell **181:** 1680–1692

Lisec J, Romisch-Margl L, Nikoloski Z, Piepho HP, Giavalisco P, Selbig J, Gierl A, Willmitzer L (2011) Corn hybrids display lower metabolite variability and complex metabolite inheritance patterns. Plant J **68:** 326–336

Maddison AL, Camargo-Rodriguez A, Scott IM, Jones CM, Elias DMO, Hawkins S, Massey A, Clifton-Brown J, McNamara NP, Donnison IS, et al. (2017) Predicting future biomass yield in *Miscanthus* using the carbohydrate metabolic profile as a biomarker. GCB Bioenergy **9:** 1264–1278

Menche J, Guney E, Sharma A, Branigan PJ, Loza MJ, Baribaud F, Dobrin R, Barabasi AL (2017) Integrating personalized gene expression profiles into predictive disease-associated gene pools. NPJ Syst Biol Appl **3:** 10

Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabasi AL (2015) Uncovering disease-disease relationships through the incomplete interactome. Science **347:** 1257601

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

Millet EJ, Kruijer W, Coupel-Ledru A, Alvarez Prado S, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, van Eeuwijk F, Tardieu F (2019) Genomic prediction of maize yield across European environmental conditions. Nat Genet **51:** 952–956

Obata T, Witt S, Lisec J, Palacios-Rojas N, Florez-Sarasa I, Yousfi S, Araus JL, Cairns JE, Fernie AR (2015) Metabolite profiles of maize leaves in drought, heat, and combined stress field trials reveal the relationship between metabolism and grain yield. Plant Physiol **169:** 2665–2683

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

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012b) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci USA **109:** 8872–8877

Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, et al. (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol **24:** 447–454

Shen X, Wang R, Xiong X, Yin Y, Cai Y, Ma Z, Liu N, Zhu Z-J (2019) Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nat Commun **10:** 1516

Sprenger H, Erban A, Seddig S, Rudack K, Thalhammer A, Le MQ, Walther D, Zuther E, Köhl KI, Kopka J, et al. (2018) Metabolite and transcript markers for the prediction of potato drought tolerance. Plant Biotechnol J **16:** 939–950

Sulpice R, Nikoloski Z, Tschoep H, Antonio C, Kleessen S, Larhlimi A, Selbig J, Ishihara H, Gibon Y, Fernie AR, et al. (2013) Impact of the carbon and nitrogen supply on relationships and connectivity between metabolism and biomass in a broad panel of *Arabidopsis* accessions. Plant Physiol **162:** 347–363

Sulpice R, Pyl ET, Ishihara H, Trenkamp S, Steinfath M, Witucka-Wall H, Gibon Y, Usadel B, Poree F, Piques MC, et al. (2009)

Starch as a major integrator in the regulation of plant growth. Proc Natl Acad Sci USA **106:** 10348–10353

Sulpice R, Trenkamp S, Steinfath M, Usadel B, Gibon Y, Witucka-Wall H, Pyl ET, Tschoep H, Steinhauser MC, Guenther M, et al. (2010) Network analysis of enzyme activities and metabolite levels and their relationship to biomass in a large panel of *Arabidopsis* accessions. Plant Cell **22:** 2872–2893

Varshney RK, Singh VK, Kumar A, Powell W, Sorrells ME (2018) Can genomics deliver climate-change ready crops? Curr Opin Plant Biol **45:** 205–211

Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J, et al. (2014) Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun **5:** 3438

Williams W (1959) Heterosis and the genetics of complex characters. Nature **184:** 527–530

Wilmanski T, Rappaport N, Earls JC, Magis AT, Manor O, Lovejoy J, Omenn GS, Hood L, Gibbons SM, Price ND (2019) Blood metabolome predicts gut microbiome alpha-diversity in humans. Nat Biotechnol **37:** 1217–1228

Wold H (1975) Soft modelling by latent variables: the nonlinear iterative partial least squares approach. J Appl Probab **12:** 117–142

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst **58:** 109–130

Xia J, Wishart DS (2011) Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. Nat Protoc **6:** 743–760

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

Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou JY, Petyuk VA, Chen L, Ray D, et al. (2016) Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell **166:** 755–765

Zhao Y, Li Z, Liu G, Jiang Y, Maurer HP, Würschum T, Mock H-P, Matros A, Ebmeyer E, Schachschneider R, et al. (2015) Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc Natl Acad Sci USA **112:** 15624–15629