








Can biochemical traits bridge the gap between genomics and plant performance? A study in rice under drought

Giovanni Melandri ^{1,2,†}, Eliana Monteverde ^{2,3}, David Riewe ^{4,5}, Hamada AbdElgawad ^{6,7}, Susan R. McCouch ^{2,*} and Harro Bouwmeester ^{1,8,*‡}

- 1 Laboratory of Plant Physiology, Wageningen University and Research, Wageningen, the Netherlands
- 2 School of Integrative Plant Sciences, Plant Breeding and Genetics Section, Cornell University, Ithaca, New York, USA
- 3 Departamento de Biología Vegetal, Facultad de Agronomía, Laboratorio de Evolución y Domesticación de las Plantas, Universidad de La República, Montevideo, Uruguay
- 4 Julius Kühn-Institute (JKI), Federal Research Centre for Cultivated Plants, Institute for Ecological Chemistry, Plant Analysis and Stored Product Protection, Berlin, Germany
- 5 Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany
- 6 Laboratory for Integrated Molecular Plant Physiology Research, University of Antwerp, Antwerp, Belgium
- 7 Department of Botany, Faculty of Science, Beni-Suef University, Beni Suef, Egypt
- 8 Plant Hormone Biology group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, the Netherlands

*Author for correspondence: srm4@cornell.edu (S.R.M.), h.j.bouwmeester@uva.nl (H.B.)

†Present address: INRAE, University of Bordeaux, UMR BFP, Villenave d'Ornon, France.

‡Senior author.

These authors contributed equally (G.M. and E.M.).

G.M., H.B., and S.R.M. designed the research plan. G.M. collected the leaf samples from the field trial; D.R. and G.M. performed the GC–MS analysis of the samples. H.A. performed the oxidative stress status analysis of the samples. G.M. and E.M. performed all the statistical analyses of the data. G.M., E.M., S.R.M., and H.B. wrote the manuscript. S.R.M. and H.B. agree to serve as the authors responsible for contact and ensure communication. 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: Giovanni Melandri (giovannimelandri@gmail.com).

Abstract

The possibility of introducing metabolic/biochemical phenotyping to complement genomics-based predictions in breeding pipelines has been considered for years. Here we examine to what extent and under what environmental conditions metabolic/biochemical traits can effectively contribute to understanding and predicting plant performance. In this study, multivariable statistical models based on flag leaf central metabolism and oxidative stress status were used to predict grain yield (GY) performance for 271 *indica* rice (*Oryza sativa*) accessions grown in the field under well-watered and reproductive stage drought conditions. The resulting models displayed significantly higher predictability than multivariable models based on genomic data for the prediction of GY under drought ($Q^2 = 0.54–0.56$ versus 0.35) and for stress-induced GY loss ($Q^2 = 0.59–0.64$ versus 0.03–0.06). Models based on the combined datasets showed predictabilities similar to metabolic/biochemical-based models alone. In contrast to genetic markers, models with enzyme activities and metabolite values also quantitatively integrated the effect of physiological differences such as plant height on GY. The models highlighted antioxidant enzymes of the ascorbate–glutathione cycle and a lipid oxidation stress marker as important predictors of rice GY stability under drought at the reproductive stage, and these stress-related variables were more predictive than leaf central

metabolites. These findings provide evidence that metabolic/biochemical traits can integrate dynamic cellular and physiological responses to the environment and can help bridge the gap between the genome and the phenome of crops as predictors of GY performance under drought.

Introduction

In rice (*Oryza sativa*), as in most crops, grain yield (GY) is a highly complex trait. It is controlled by many genes of small effect, and these genes operate in coordinated networks that are influenced by pleiotropic and epistatic effects as well as by genotype-by-environment-by-management interactions (Xing and Zhang, 2010).

The intrinsic complexity and polygenic nature of GY makes it a difficult trait to improve using marker-assisted selection. On the other hand, genomic selection (GS) overcomes the limits that are associated with the absence of major effect genes by simultaneously estimating the effect of many markers (and underlying genes) distributed over the whole genome. However, the need to account for different sources of nongenetic variability, and nonadditive modes of gene action has made model choice and implementation of GS challenging for improving complex traits (Rice and Lipka, 2021).

One limitation of genomics for the prediction of complex phenotypes, such as GY, lies in the fact that the information encoded in genetic markers is a poor predictor of an organism's ability to dynamically respond to environmental stimuli at the physiological level (Yin et al., 2004). For these reasons, in other important cereal crops, such as maize (*Zea mays*) and wheat (*Triticum aestivum*), physiological traits have been studied in connection with genetics to improve crop performance (Cooper et al., 2014; Reynolds and Langridge, 2016). Cellular physiology provides a key interface between genotype and phenotype. It represents an internal phenotype (endophenotype) that integrates transcriptomic, proteomic, and metabolomic networks of regulation that are interconnected and continuously respond to environmental factors (Großkinsky et al., 2015). Among these multiple cellular layers of information, metabolite levels are more directly linked to the phenotype than are gene transcripts and protein levels (Fernie and Stitt, 2012) and the possibility of introducing metabolic/biochemical phenotyping to complement genomics in breeding pipelines for the improvement of crop performance has been considered for years (Fernandez et al., 2021).

In the last 15 years, metabolome-based models have been used to predict complex traits, such as biomass, in large *Arabidopsis* (*Arabidopsis thaliana*) and maize populations of recombinant inbred lines (Meyer et al., 2007; Sulpice et al., 2009; Steinfath et al., 2010; Riedelsheimer et al., 2012). In rice, these models were successfully employed to predict the yield of hybrids by directly using the hybrid's metabolite profiles (Xu et al., 2016) or those of the parents (Dan et al., 2016). Despite the value of these findings for hybrid

breeding programs, the narrow genetic background of the materials used in these studies did not explore the large, qualitative, and quantitative genetic diversity available for rice metabolism (Chen et al., 2014). In addition, most of the metabolomics studies in crop species, including rice, have been conducted under control conditions while in *Arabidopsis* natural variation in metabolic plasticity (i.e. metabolic changes induced by the environmental changes) was shown to be an important factor contributing to phenotypic plasticity (Kleessen et al., 2014). For these reasons, it is still necessary to evaluate the power of metabolic/biochemical-based models for predicting GY under different environmental conditions in large panels of genetically diverse crop accessions. It is also necessary to compare the predictive ability of metabolic/biochemical-based models, genomic-based models, and models based on combined datasets for the same trait to understand when and if metabolic/biochemical traits complement or potentially outperform genomics-based prediction for yield improvement.

It was recently shown that a multivariable model based on levels of flag leaf central metabolites and oxidative stress markers/enzymes was able to efficiently predict drought-induced GY loss in a large panel of genetically diverse *indica* rice accessions grown in the field (Melandri et al., 2020a). Here we use the same dataset to predict GY under both well-watered and drought conditions, in addition to predict stress-induced GY loss. We also evaluate a genomic dataset as the basis for predicting the same GY traits under the same conditions. This allowed us to: (1) analyze the differences between metabolic/biochemical-based models, genomic-based models, and models that integrate the two datasets for the prediction of rice GY traits under well-watered and drought conditions and (2) identify which biochemical pathways/antioxidants are important predictors for GY performance in this crop.

Results

Relationships between GY, plant height (PH), and flowering time (FT)

In this study, the GY performance of 271 tropical and subtropical, traditional, and improved *indica* rice varieties (Supplemental Table S1) was assessed in a field experiment under irrigated (control) and reproductive-stage drought conditions. Drought stress reduced GY (GY_{LOSS}) by an average of 29.3% (paired t test: $P < 0.001$) (Supplemental Table S2). GYs under control (GY_{CON}) and drought (GY_{DRO}) conditions were highly correlated (Pearson correlation, $r = 0.75$, $P < 0.001$), and high estimates of broad-sense heritability were observed under both treatments ($H^2 = 0.89$ and 0.84

for GY_{CON} and GY_{DRO} , respectively) (Figure 1; Supplemental Table S3). Interestingly, GY_{LOSS} was significantly ($P < 0.001$) and negatively correlated ($r = -0.61$) with GY_{DRO} but not with GY_{CON} (Figure 1). This observation indicates that the yield performance of the accessions exhibited genotype-by-treatment interaction and was highly influenced by reproductive-stage drought.

To explore the nature of this interaction, we further assessed the relationships between GY-related traits and two important agronomic phenotypes, plant height (PH), and flowering time (FT). Both of these traits showed high heritability estimates (PH: $H^2 = 0.97$; FT: $H^2 = 0.99$; Supplemental Table S3) and displayed significant variation in the diversity panel (Figure 1; Supplemental Table S2). In this *indica* rice panel, there was a high correlation between PH_{CON} and PH_{DRO} ($r = 0.95$, $P < 0.001$), and the distribution of PH under both treatments was bi-modal, with two, distinct normal distributions around two different peaks (Figure 1). PH differences under both treatments are strongly associated with allelic variation of a single-nucleotide polymorphism (SNP) marker on chromosome 1 (position: 38,286,772 bp; Welch's t test: $P < 0.001$; Supplemental Figure S1). This SNP marker was previously mapped for PH differences in a larger version of this panel (Kadam et al., 2017), and the linkage disequilibrium block (259 kbp) surrounding the marker included the gibberellin 20-oxidase biosynthetic gene, also known as *SEMI-DWARF1* (*OsGa20ox2*, LOC_Os01g66100), introduced during the Green Revolution. The genotypic difference associated with the diagnostic SNP marker is a strong predictor of PH across environments, despite a mean reduction of

8 cm when PH_{DRO} is compared with PH_{CON} (paired t test: $P < 0.001$; Supplemental Table S2). This stature-associated SNP is also strongly associated with yield performance (Welch's t test: $P < 0.001$; Supplemental Figure S1), as evidenced by the significant ($P < 0.001$) negative correlation between PH and GY under both control ($r = -0.31$) and drought ($r = -0.26$) conditions (Figure 1). In contrast, GY_{LOSS} displayed a random distribution among both short and tall accessions of the panel (Supplemental Figure S2).

In this study, FT was synchronized by sowing and transplanting accessions on different dates to ensure that drought stress was imposed on all genotypes at the flowering stage (Kadam et al., 2018). FT synchronization was largely, but not entirely achieved (Supplemental Table S1). Reproductive-stage drought stress resulted in an almost uniform delay of three days in FT for all accessions (Supplemental Table S2) consistent with the high correlation ($r = 0.97$, $P < 0.001$) between FT under control (FT_{CON}) and drought (FT_{DRO}) conditions (Figure 1). Despite the synchronization of FT for stress application, FT_{DRO} was negatively and weakly correlated with GY_{LOSS} ($r = -0.16$, $P < 0.01$) suggesting that drought-induced yield loss was partially mitigated by late FT.

Variation in GY is better predicted by metabolome/oxidative stress status-based models than by genomic-based models

Our previous work (Melandri et al., 2020a) showed that cross-validated (CV) partial least squares regression (PLSR) modeling based on 111 flag leaf metabolites, oxidative stress

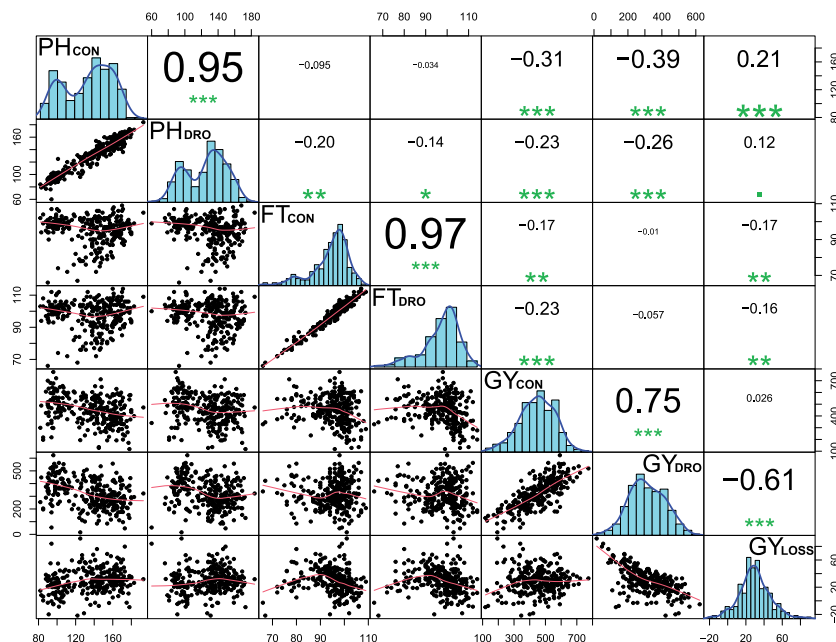


Figure 1 Correlation matrix between values (BLUEs) of PH, FT, and GY—under control (CON) and drought (DRO) conditions—and GY loss (GY_{LOSS}) of the 271 *indica* rice accessions. PH units are expressed in centimeters, FT in days, GY in grams/m², and GY_{LOSS} in percentage. Pearson correlations (r , stronger correlations are represented by larger numbers) and levels of significance (in green, *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$) are reported in the upper-right portion of the matrix. Scatterplots of the pairwise combinations between traits (trendline in red) are reported in the bottom-left portion of the matrix. Trait distributions are represented along the diagonal of the matrix (trendline in blue).

markers, and enzyme activities (hereafter MetabOxi) measured under drought efficiently predicted stress-induced GY_{LOSS} in 292 genetically diverse *indica* rice genotypes.

In this study, we expanded the PLSR modeling approach to predict GY_{CON} and GY_{DRO} , in addition to GY_{LOSS} , in a subset of 271 accessions from the same experiment using the MetabOxi dataset (Supplemental Table S4). Control values of the MetabOxi dataset were used to predict GY_{CON} while drought values of the same dataset were used to predict GY_{DRO} and GY_{LOSS} . To compare the strength of modeling using these biochemical markers with genomic prediction, a genomic dataset consisting of 81,347 SNP markers on the 271 accessions (Supplemental Data Set 1) was also used to build PLSR models for prediction of GY_{CON} , GY_{DRO} , and GY_{LOSS} . In addition, Ridge-Regression Best Linear Unbiased Prediction (RR-BLUP) and BayesB models, more commonly employed in GS studies, were used to run MetabOxi- and genomic-based models as the basis for comparing the predictive ability of all three models for the same traits. Goodness of prediction (Q^2) was used to quantify the predictability of the models (for more details on the calculation of Q^2 , see “Materials and methods”) and used throughout the manuscript to describe and discuss the results. Prediction accuracies were also calculated as Pearson correlation (Pearson’s r) coefficients between observed and predicted GY values.

For each GY trait, Q^2 values were similar between the 10-fold CV PLSR, RR-BLUP, and BayesB models and always higher for the MetabOxi than the genomic dataset (Figure 2, A and C). Q^2 values were most similar for GY_{CON} when MetabOxi- ($Q^2 = 0.32–0.40$) and genomic-based ($Q^2 = 0.31–0.32$) models were compared. Differences in predictability were greater for GY_{DRO} where genomic-based models displayed similar values ($Q^2 = 0.35$) as for GY_{CON} but MetabOxi-based models showed markedly better values ($Q^2 = 0.54–0.56$). The gap between MetabOxi- and genomic-based models further increased when predicting GY_{LOSS} , with MetabOxi-based models showing good predictability ($Q^2 = 0.59–0.64$) and genomic-based models showing almost null values ($Q^2 = 0.03–0.06$). In all cases, Pearson correlation coefficients showed the same trends as Q^2 values (Figure 2, A and C). Overall, the higher predictive power of the MetabOxi dataset compared to the genomic dataset suggests that metabolite levels and enzyme activities are more closely aligned to the ability of a plant to dynamically respond to stress than are fixed genetic determinants. This is especially true for drought-induced GY_{LOSS} , followed by GY_{DRO} and, to a lesser extent, for GY_{CON} .

Models based on combined MetabOxi and genomic data have a similar predictive power as MetabOxi-based models alone

We compared the prediction accuracies of the 10-fold CV PLSR, RR-BLUP, and BayesB models based on the combined MetabOxi and genomic datasets (MetabOxi + Genomic in Figure 2B) with the same models based on a single dataset

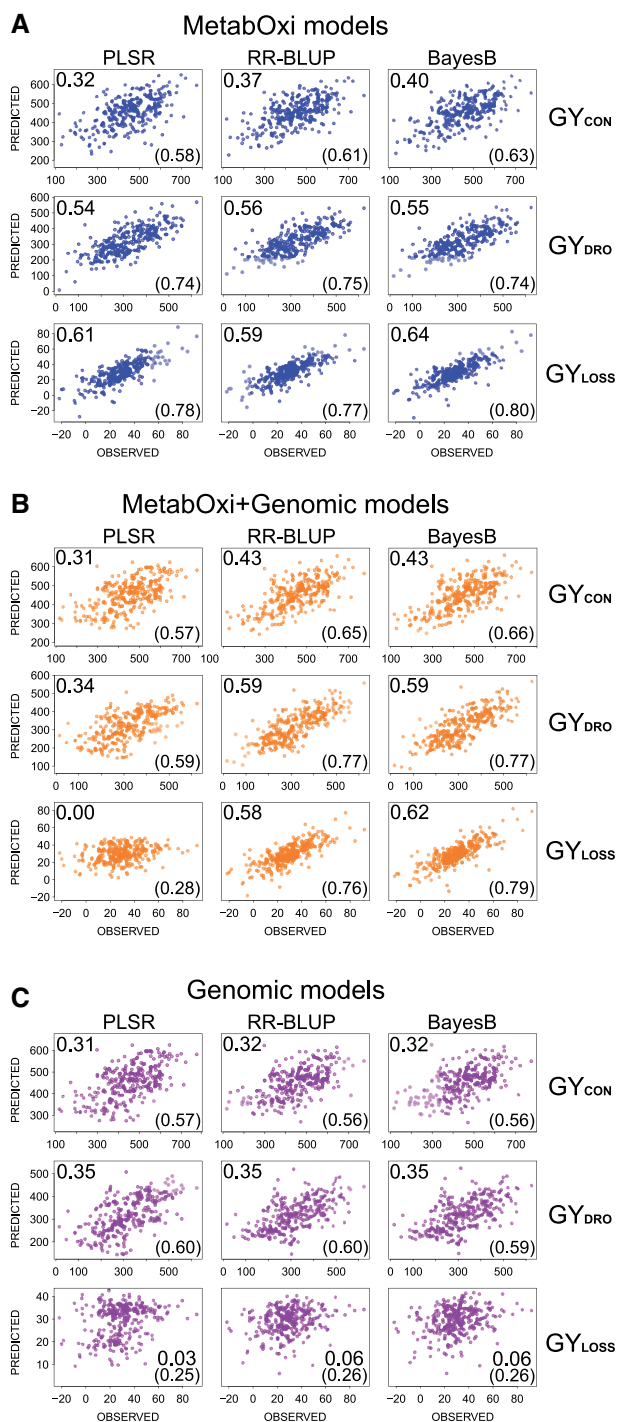


Figure 2 Multivariate models for the prediction of GY performance in the 271 *indica* rice accessions of the panel. Scatterplots of observed (BLUES) versus predicted values of the 10-fold CV MetabOxi-based (A, in blue), genomic-based (B, in purple), and MetabOxi + Genomic-based (C, in orange) PLSR, RR-BLUP, and BayesB models for the prediction of GY—under control (CON) and drought (DRO) conditions—and GY_{LOSS} . GY units are expressed in grams/m² and GY_{LOSS} in percentage. Predictability values (Q^2 and Pearson’s r) of the models are displayed in each scatterplot (Pearson’s r values in brackets).

(Figure 2, A and C) for GY traits. The MetabOxi + Genomic-based RR-BLUP and BayesB models always predicted GY traits better than genomic-based

models alone, while the PLSR models showed no difference (Figure 2, B and C). Compared with MetabOxi-based models, MetabOxi + Genomic-based RR-BLUP and BayesB models showed a virtually identical predictability for GY_{CON} , GY_{DRO} , and GY_{LOSS} (Figure 2, A and B). The MetabOxi + Genomic-based PLSR models displayed lower predictability values for GY-related traits, particularly for GY_{DRO} and GY_{LOSS} (GY_{CON} showed the same predictability). Overall, these results suggest that combining MetabOxi and genomic information into a single model did not improve the prediction of GY-related traits compared with the use of MetabOxi-based models alone. In the case of PLSR, the integration of the two datasets reduced the predictability of GY under stress compared with the MetabOxi-based models.

Adjusting GY for PH and FT consistently improves the predictive power of genomic-based models only

Given the influence of PH and FT variation on GY performance (as discussed above), we next tested if accounting for differences in PH and FT in the context of GY performance could improve the predictability of MetabOxi- and/or genomic-based models. To address this question, the 10-fold CV PLSR, RR-BLUP, and BayesB models were re-run using re-estimated values of GY_{CON} , GY_{DRO} , and GY_{LOSS} calculated using PH, FT, or both (PH + FT) as trait covariates.

The MetabOxi-based models showed virtually no improvement in predicting GY_{CON} when the GY values were adjusted using PH and/or FT as covariates, while prediction of GY_{LOSS} and GY_{DRO} was slightly improved (a max Q^2 increase of ~ 0.12 was observed using PH + FT corrected values) (Table 1). In contrast, the genomic-based models displayed a larger increase in predictability using the covariate-adjusted GY traits. The increase in predictability was again minimal for GY_{CON} (max Q^2 increase of 0.05) and larger for GY_{DRO} and GY_{LOSS} (max Q^2 increase of ~ 0.25 , for both traits). Predictability values for the genomic-based models were most improved for GY_{DRO} and GY_{LOSS} when the data were adjusted using either PH alone or PH + FT as covariates, while improvement was minimal (max Q^2 increase of 0.08) when FT alone was used as a covariate (Table 1). This suggests that variation in PH exerts a stronger influence on GY performance under drought than variation in FT, consistent with the correlations among agronomic traits described above (Figure 1). Despite the increased predictive ability of the genomic-based and the MetabOxi-based models when covariate-adjusted GY traits were used as input data, it is noteworthy that the MetabOxi-based models always outperformed the genomic models in terms of predictability (Table 1). In all cases, Pearson correlation coefficients showed the same trends as Q^2 values (Table 1).

We also considered the effect of PH and FT (individually and together) as secondary traits to predict GY by running multi-trait PLSR and RR-BLUP models (for more details on the calculation of the models see “Materials and methods”) and report the results in Supplemental Table S5.

Interestingly, the results (compare Supplemental Table S5 with Table 1) show that better GY predictabilities were determined when PH, FT, or both were used as covariates for calculating the BLUEs rather than incorporating them into the multi-trait models.

Taken together, these results suggest that metabolite values and enzyme activities provide a way to quantitatively estimate dynamic physiological responses to stress that differentiate individual plants in a population, and that these MetabOxi-variables already integrate inherent differences in PH and FT known to impact GY performance.

Rankings of MetabOxi-based model predictors reveal the importance of biochemical pathways and antioxidants for GY performance

Each of the 10 MetabOxi-based sub-models (generated by the CV procedure) for the prediction of GY traits provided a rank of importance for the 111 MetabOxi-variables. By multiplying the 10 ranks derived from single sub-models, the overall ranking of each MetabOxi-variable was calculated (Supplemental Tables S6–S8). We next determined the correlation between the MetabOxi-variables and GY traits, PH, and FT (Supplemental Tables S9 and S10) to gain insight into the nature (positive or negative) and strength (r) of their associations. The top three MetabOxi-variables from each GY model (Table 2), that is, those with the lowest rank-products (lower rank-product implies higher importance), indicate which biochemical and antioxidant pathways are important for GY prediction in our dataset (Figure 3). We additionally tested if these top-ranked predictors had a significant effect on GY by fitting them in linear models as single explanatory variables for the trait (Supplemental Table S11).

Among the 111 MetabOxi-variables evaluated as top-ranked predictors for GY_{CON} , organic acids consistently ranked high (Table 2). Chlorogenic acid (3-caffeoyl-quinic acid), a compound with antioxidant and pathogen defense activity in plants, is present among the top-ranked variables for all three models (ranked first, second, and third in the PLSR, RR-BLUP, and BayesB models, respectively). It correlates negatively with GY_{CON} ($r = -0.36$, $P < 0.001$), positively with PH_{CON} ($r = 0.39$, $P < 0.001$) and negatively with FT_{CON} ($r = -0.33$, $P < 0.001$) (Table 2; Supplemental Table S9). These correlations suggest that the variation in concentration of this compound among rice accessions fully integrates the complex interconnections between GY, PH, and FT observed in the rice panel (Figure 1). A second organic acid, α -ketoglutaric acid (2-oxo-glutaric acid) ranked first in both the RR-BLUP and BayesB models. It is an intermediate of the tricarboxylic acid (TCA) cycle, like isocitric and citric acid, which are the second and third top-ranked variables of the PLSR model. All three of these TCA cycle intermediates are positively correlated with GY_{CON} ($P < 0.001$, $r = 0.27$, 0.39, and 0.37 for α -ketoglutaric, isocitric, and citric acid, respectively) and negatively with PH_{CON} ($P < 0.001$, $r = -0.31$, -0.38 , and -0.45 for α -ketoglutaric, isocitric, and citric acid, respectively) while they show no correlation with FT_{CON}

Table 1 Predictability of MetabOxi- and genomic-based models for GY traits nonadjusted and adjusted by PH and FT

GY traits		MetabOxi-based models						Genomic-based models					
		PLSR		RR-BLUP		BayesB		PLSR		RR-BLUP		BayesB	
		Q ²	r	Q ²	r	Q ²	r	Q ²	r	Q ²	r	Q ²	r
BLUEs no cov	GY _{CON}	0.32	0.58	0.37	0.61	0.40	0.63	0.31	0.57	0.32	0.56	0.32	0.56
	GY _{DRO}	0.54	0.74	0.56	0.75	0.55	0.74	0.35	0.60	0.35	0.60	0.35	0.59
	GY _{LOSS}	0.61	0.78	0.59	0.77	0.64	0.80	0.03	0.25	0.06	0.26	0.06	0.26
BLUEs cov PH	GY _{CON}	0.34	0.59	0.37	0.61	0.40	0.63	0.31	0.57	0.31	0.56	0.31	0.56
	GY _{DRO}	0.60	0.78	0.64	0.80	0.64	0.80	0.54	0.73	0.53	0.73	0.53	0.73
	GY _{LOSS}	0.58	0.76	0.60	0.77	0.60	0.77	0.24	0.51	0.32	0.57	0.32	0.57
BLUEs cov FT	GY _{CON}	0.30	0.56	0.35	0.59	0.38	0.61	0.35	0.60	0.35	0.60	0.35	0.59
	GY _{DRO}	0.58	0.77	0.62	0.79	0.62	0.79	0.41	0.64	0.42	0.65	0.42	0.64
	GY _{LOSS}	0.66	0.81	0.66	0.81	0.67	0.82	0.03	0.30	0.14	0.37	0.14	0.38
BLUEs cov PH and FT	GY _{CON}	0.32	0.57	0.36	0.60	0.38	0.62	0.36	0.61	0.36	0.60	0.36	0.60
	GY _{DRO}	0.65	0.81	0.68	0.83	0.69	0.83	0.59	0.77	0.58	0.76	0.58	0.76
	GY _{LOSS}	0.67	0.82	0.66	0.82	0.67	0.82	0.24	0.51	0.30	0.55	0.30	0.55

Predictability (Q² and Pearson's *r*) values of the PLSR, RR-BLUP, and BayesB models for the best linear unbiased estimators (BLUEs) of GY—under control (CON) and drought (DRO) conditions—and GY loss (GY_{LOSS}) calculated considering PH and FT as covariates (cov PH, cov FT, cov PH&FT) or without (no cov, same values as in Figure 2, A and C).

(Table 2). This indicates that higher abundance of TCA cycle intermediates is associated with shorter PH in the panel, and higher GY_{CON} performance, independent of differences in FT (Figure 3). A third variable, galactinol, is a sugar alcohol with osmoprotective/antioxidant activity. It ranked second in the BayesB model for GY_{CON} (fourth in the RR-BLUP model, Supplemental Table S6) and, like chlorogenic acid, is negatively correlated with GY_{CON} ($r = -0.23$, $P < 0.05$) but positively correlated with FT_{CON} ($r = 0.50$, $P < 0.001$); it shows no association with differences in PH_{CON} (Table 2). This may indicate that the variation in galactinol levels impacts GY performance and is mainly associated with the imperfect FT synchronization of the panel. Less clear is the biochemical contribution of uridine, the third best predictor of the RR-BLUP model (fourth in the BayesB model, Supplemental Table S6) where variation is not associated with differences in GY_{CON} though it is positively correlated with PH_{CON} ($r = 0.22$, $P < 0.05$) and negatively with FT_{CON} ($r = -0.22$, $P < 0.05$) (Table 2). The presence of uridine among the top BayesB model predictors hints at the importance of both PH and FT differences on GY performance of the accessions under control conditions.

In contrast to the predictors identified for GY_{CON}, the top-ranked predictors in the MetabOxi-based GY_{DRO} and GY_{LOSS} models are mostly antioxidant enzymes or oxidative stress markers that are not significantly correlated with variation in either PH_{DRO} or FT_{DRO} (Table 2). The antioxidant enzyme dehydroascorbate reductase (DHAR) is the highest-ranking model predictor (rank-prod = 1 in all models) for both GY_{DRO} and GY_{LOSS}. It is positively ($r = 0.58$, $P < 0.001$) and negatively ($r = -0.62$, $P < 0.001$) correlated with the two traits, respectively, and these correlations are the most significant among the 111 MetabOxi-variables (Supplemental Table S10). The lipid peroxidation product malondialdehyde (MDA) ranked as the second-best variable in the PLSR and RR-BLUP models, and third in BayesB, for

the prediction of both GY_{DRO} and GY_{LOSS} (Table 2). In contrast to DHAR, MDA is negatively correlated with GY_{DRO} ($r = -0.41$, $P < 0.001$) and positively with GY_{LOSS} ($r = 0.61$, $P < 0.001$). The fact that DHAR and MDA are the top-ranked predictors of GY_{DRO} and GY_{LOSS} suggests that, during drought imposition, the oxidative stress status of the flag leaf is more predictive of GY performance than flag leaf central metabolism (Figure 3). This is underscored by the fact that another antioxidant enzyme, monodehydroascorbate reductase (MDHAR) was identified as the second (BayesB model) and third (PLSR and RR-BLUP) most important predictor of GY_{LOSS}. MDHAR is only marginally correlated with GY_{DRO} ($r = 0.21$, $P = 0.05$) (Supplemental Table S10) and not correlated with GY_{LOSS} (Table 2). In addition, MDHAR was the only top-ranked predictor with a nonsignificant effect on GY (GY_{LOSS}) when fit as a single explanatory variable in a linear model with the trait as response variable (Supplemental Table S11). Like DHAR, MDHAR regenerates oxidized ascorbate to its reduced form and is involved in the ascorbate–glutathione antioxidant cycle. Its presence as a top-ranked model variable suggests the importance of this cycle in counteracting drought-induced GY reduction, despite the weak relationship between its activity and GY. Interestingly, α -ketoglutaric acid, which was also positively associated with GY_{CON}, ranked second in the BayesB and third in the RR-BLUP model as a predictor of GY_{DRO} though it was not significantly correlated with GY_{DRO} (Table 2). Differences in α -ketoglutaric acid values were negatively correlated with PH under both control and drought conditions (PH_{DRO}: $r = -0.38$, $P < 0.001$), suggesting that the accumulation of this TCA cycle intermediate in shorter plants contributes positively to both constitutive GY_{CON} performance and to GY_{DRO} performance, even if the latter is also strongly determined by dynamic responses to stress (Figure 1).

Table 2 Best predictive variables of the MetabOxi-based models for the prediction of GY traits

Trait to predict	Rank	MetabOxi-based PLSR model					MetabOxi-based RR-BLUP model					MetabOxi-based BayesB model				
		Variable	Rank-prod	Corr. with traits (rs and P)		FT	Variable	Rank-prod	Corr. with traits (rs and P)		FT	Variable	Rank-prod	Corr. with traits (rs and P)		FT
				GY	PH				GY	PH				GY	PH	
GY _{CON}	1	Chlorogenic acid	1	-0.36***	0.39***	-0.33***	48	0.27***	-0.31***	-0.09 ns	α-ketoglutaric acid	144	0.27***	-0.31***	-0.09 ns	
	2	Isocitric acid	1,024	0.39***	-0.38***	-0.21*	1,152	-0.36***	0.39***	-0.33***	Galactinol	432	-0.23*	-0.01 ns	0.50***	
	3	Citric acid	275,562	0.37***	-0.45***	-0.08 ns	23,328	0.18 ns	0.22*	-0.22*	Chlorogenic acid	4,608	-0.36***	0.39***	-0.33***	
GY _{DRO}	1	DHAR	1	0.58***	-0.10 ns	0.02 ns	1	0.58***	-0.10 ns	0.02 ns	DHAR	1	0.58***	-0.10 ns	0.02 ns	
	2	MDA	1,024	-0.41***	0.15 ns	-0.17 ns	17,280	-0.41***	0.15 ns	-0.17 ns	α-ketoglutaric acid	5,760	0.20 ns	-0.38***	0.05 ns	
	3	MDHAR	233,280	0.21 ns	-0.09 ns	-0.19 ns	20,736	0.20 ns	-0.38***	0.05 ns	MDA	1,244,160	-0.41***	0.15 ns	-0.17 ns	
GY _{LOSS}	1	DHAR	1	-0.62***	-0.10 ns	0.02 ns	1	-0.62***	-0.10 ns	0.02 ns	DHAR	1	-0.62***	-0.10 ns	0.02 ns	
	2	MDA	1,024	0.61***	0.15 ns	-0.17 ns	1,024	0.61***	0.15 ns	-0.17 ns	MDHAR	2,304	-0.05 ns	-0.09 ns	-0.19 ns	
	3	MDHAR	59,049	-0.05 ns	-0.09 ns	-0.19 ns	59,049	-0.05 ns	-0.09 ns	-0.19 ns	MDA	26,244	0.61***	0.15 ns	-0.17 ns	

Top three ranked predictive variables of the 10-fold CV MetabOxi-based PLSR, RR-BLUP, and BayesB models for prediction of GY under control (GY_{CON}) and drought (GY_{DRO}), and for GY_{LOSS}. Variables are ranked based on their rank-product value (Rank-prod.). Correlations between the MetabOxi-variables and the GY traits, PH, and (FT) are reported. R: Pearson correlation coefficient. Bonferroni-corrected significance of the correlation (P): ***P < 0.001, **P < 0.01, *P < 0.05, ns = not significant.

Discussion

The major goal of this study was to compare the potential of flag leaf metabolism/oxidative stress status and genetic markers to predict GY performance in a diversity panel of *indica* rice accessions grown under well-watered and drought conditions in the field. Our results show that MetabOxi-based models predict GY with superior accuracy compared to genomic-based models. This higher accuracy can be explained by the fact that MetabOxi-variables integrate quantitative estimates of complex biological processes summarized as metabolite levels and/or enzyme activities. These measurements incorporate responses from a multi-layered network of regulation (DNA, RNA, and protein) in response to dynamic internal and external stimuli (Keurentjes, 2009; Sulpice and McKeown, 2015). This is especially true in the context of environmental stress-driven perturbations, including those caused by drought. Indeed, the dynamic response to external signals from a changing environment (phenotypic plasticity) is often characterized by changing metabolite levels and enzyme activities as a result of post-translational and/or transcriptional regulation (Stitt et al., 2010). In support of this hypothesis, we observed that MetabOxi-based models were superior at predicting GY_{LOSS} and GY_{DRO}, while genomic-based models predicted GY_{DRO} with lower accuracy and were essentially unable to predict GY_{LOSS} (Figure 2, A and C). Interestingly, under well-watered conditions, the predictive power for GY_{CON} of MetabOxi- and genomic-based models was virtually identical. These findings indicate that under nonstress conditions, genetic determinants are equally predictive of plant performance as basal flag leaf central metabolites and oxidative stress markers/enzymes, but under suboptimal conditions, metabolic/biochemical traits provide valuable endophenotypes that are much more predictive of crop yield stability than genomic information (Kumar et al., 2017; Sulpice, 2020).

The possibility of modeling multi-omics data for a deeper understanding of complex phenotypic traits (e.g. crop yield under stress) and to improve the accuracy of selection in breeding (mainly through GS) is a “hot topic” in plant systems biology and plant breeding (Jamil et al., 2020; Scossa et al., 2021; Tong and Nikoloski, 2021; van Dijk et al., 2021). Our results indicate that the integration of biochemical (MetabOxi) and genomic data in the same statistical model may slightly increase (GY_{CON} and GY_{DRO}) or decrease (GY_{LOSS}) trait predictability compared with the best single omics (MetabOxi-based) models (Figure 2). These results differ from previous studies where the integration of multi-omics data improved the predictability of GY in maize and rice hybrids under nonstressed field conditions (Westhues et al., 2017; Schrag et al., 2018; Wang et al., 2019; Xu et al., 2020). A possible explanation for the difference might be that in our study we brought together two “distant” omics layers which are difficult to connect without the information carried by intermediate omics layers, that is, transcriptome and proteome. In support of this hypothesis, Schrag et al. (2018) found that the combination of two “close”

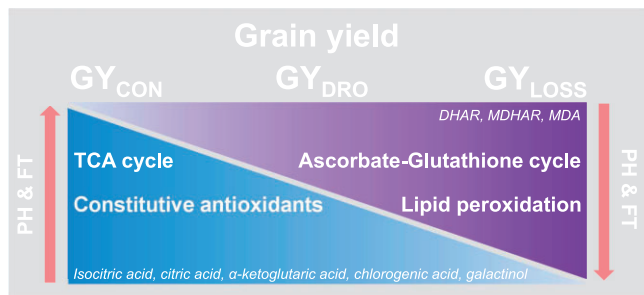


Figure 3 Summary of the main biochemical pathways predictors for GY performance in the *indica* rice panel and their relationships with GY—under control (_{CON}) and drought (_{DRO}) conditions—and GY_{LOSS}. The blue triangle represents the TCA cycle (isocitric, citric, and α-ketoglutaric acids) and constitutive antioxidants (chlorogenic acid and galactinol) which displayed higher prediction importance from left to right (GY_{CON} → GY_{DRO} → GY_{LOSS}). The purple triangle represents the ascorbate–glutathione cycle (DHAR and MDHAR) and lipid peroxidation (MDA) which displayed higher prediction importance from right to left (GY_{CON} ← GY_{DRO} ← GY_{LOSS}). The influence of PH and FT on the pathways of the two triangles is represented by the red arrow (up = high; down = low).

omics layers, genome and transcriptome, was more predictive of GY in maize hybrids than combining genome and metabolome. However, Xu et al. (2020) found that genome and metabolome was the best omics combination for the prediction of GY in rice hybrids, and that a combination of three or four omics layers (adding transcriptome and proteome) provided no improvement. Thus, the value of integrating different types of omics data for the prediction of GY might also depend on factors such as crop species, data quality, and the presence of quantifiable environmental stress at the field site, with the latter having a strong association with metabolic/biochemical data, as clearly demonstrated in this study. This study also underscores the importance of the statistical model used for prediction, as the model itself can impact the predictability of the trait when different omics datasets are integrated. The predictability values of the MetabOxi + Genomic PLSR models were similar to the corresponding MetabOxi-based model for GY_{CON}, but lower for GY_{DRO} and GY_{LOSS} (Figure 2). The fact that the same did not happen for the MetabOxi + Genomic-based RR-BLUP and BayesB models suggest that the PLSR algorithm might not be able to integrate the MetabOxi and genomic datasets in an efficient way. Indeed, the concatenation of two data matrices of vastly different sizes (111 versus 81,347 variables for the MetabOxi and genomic datasets, respectively) into a single matrix in the combined PLSR model might have over-represented the global data structure since the weighting of each variable is governed by the total sum of squares. This resulted in a reduction of the relative contribution of the MetabOxi variables for the prediction (Höskuldsson and Svinning, 2006; Reinke et al., 2018). Overall, these results highlight the need for further studies to better integrate omics data to fully exploit their explanatory power in the context of complex quantitative traits.

Another important finding of our study is that the performance of MetabOxi-based models was little improved when PH and/or FT were introduced as covariates whereas the genomic-based models showed significant improvement, particularly for GY_{DRO} and GY_{LOSS}. The same analysis demonstrated that variation in PH had a stronger influence on GY performance than FT (not entirely surprising, given that FT was synchronized in the study). In rice, variation in the gibberellic acid 20-oxidase gene (*Os20ox2*), also known as the *SEMI-DWARF1* (*SD1*) gene, is significantly associated with both PH and GY. A recessive allele of this gene, *sd1-d* (distinguished by a 382-bp deletion in Exons 1 and 2) was introduced during the “Green Revolution” in the 1960s and has become widely disseminated in modern, high-yielding rice varieties since that time (Asano et al., 2011). Semi-dwarf plants carrying *sd1-d* thrive under favorable conditions (e.g. in paddy fields with availability of water and nitrogen), but yield similarly to the taller, traditional varieties carrying the *SD1* allele under unfavorable conditions (e.g. in upland/rainfed fields where water and/or nitrogen are in short and/or variable supply) (Lafitte et al., 2007). Thus, variation at the *SD1* locus impacts not only PH but has far-reaching repercussions that impact yield performance, and many of the physiological differences associated with PH in the *indica* rice diversity panel are implicitly integrated into the values of metabolites and oxidative stress markers/enzymes in the MetabOxi dataset, but this level of integration is not observed in the genomic data.

An examination of the top-ranked MetabOxi-variables selected as predictors of GY_{CON}, GY_{DRO}, and GY_{LOSS} underscores the integrative nature of the metabolic data.

All three GY_{CON} models identified chlorogenic acid among the top predictors. Chlorogenic acid is strongly and negatively correlated with GY_{CON} and positively correlated with PH_{CON}. This compound is widely described in the literature for its beneficial antioxidant and anti-herbivore activity in plants (Niggeweg et al., 2004; Ferreres et al., 2011; Kundu and Vadassery, 2019), consistent with the high degree of environmental plasticity associated with tall, low-yielding traditional varieties of rice adapted to environmentally variable, low-input production systems (Lempe et al., 2013; Dwivedi et al., 2016). In contrast, the shorter, higher-yielding modern varieties have been bred for relatively stable, high-input systems where high levels of chlorogenic acid provided few advantages. It might be that in this field trial, under irrigated conditions and with the application of fertilizers and weed, insect and disease control, a constitutively higher activity of the chlorogenic acid pathway in the traditional, tall accessions represented a metabolic cost and conferred little or no advantage, as evidenced by the lower GY performance. Chlorogenic acid is also described in the literature as an intermediate compound in lignin biosynthesis (Volpi e Silva et al., 2019). Thus, a higher availability of this metabolite may positively affect growth, resulting in increased PH. Three intermediates of the TCA cycle (citric, isocitric, and α-ketoglutaric acid) were also among the top predicting

MetabOxi-variables of the GY_{CON} models. This fundamental pathway provides energy and carbon skeletons for many plant biosynthetic processes (Sweetlove et al., 2010; Araújo et al., 2012) and these three organic acids are positively correlated with GY_{CON} and negatively with PH_{CON} . The negative correlation with PH_{CON} suggests the presence of an altered TCA/biosynthetic activity between the short, high-yielding varieties of the panel compared to the tall, lower-yielding traditional accessions. This indicates that the translation of increased radiation- and nitrogen-use efficiency into higher yields in modern semi-dwarf rice varieties (Zhu et al., 2016) is also determined by metabolic adaptations of central metabolism (Figure 3).

During drought, leaf oxidative stress status is more predictive of GY performance than central metabolism. This is evidenced by the selection of three antioxidant enzymes/oxidative stress markers—DHAR, MDHAR, and MDA—as top-ranked variables from the MetabOxi-based models for predicting GY_{DRO} and GY_{LOSS} . When drought is imposed at the flowering stage, like in this study, plants are constrained in their ability to make system-wide metabolic adjustments due to the focused export of assimilates from the flag leaf to the developing panicles (Yoshida, 1972; Biswal and Kohli, 2013). This likely increases the importance of antioxidant mechanisms to counteract oxidative damage resulting from enhanced generation of reactive oxygen species (ROS) in response to drought (Melandri et al., 2021). Among the top predictors, MDA, which is negatively correlated with GY_{DRO} and positively with GY_{LOSS} , is a lipid peroxidation product indicative of stress-induced oxidative damage to the cellular lipid membranes (Møller et al., 2007). DHAR and MDHAR are two antioxidant enzymes involved in the ascorbate–glutathione cycle, the central redox-hub *in planta*, where oxidized ascorbate is recycled to its reduced form that, in turn, can be utilized for the scavenging of ROS (Foyer and Shigeoka, 2010; Foyer and Noctor, 2011; Smirnov, 2011). In contrast to MDA, DHAR is positively correlated with GY_{DRO} and negatively with GY_{LOSS} . The opposite relationships of MDA and DHAR with GY traits highlight the importance of the ascorbate–glutathione cycle in preventing drought-induced oxidative damage and its negative impact on rice GY (Melandri et al., 2020a). Surprisingly, MDHAR is not correlated with GY_{DRO} or GY_{LOSS} . The importance of MDHAR in the prediction models may be associated with its ascorbate reducing activity that contributes to increase the efficiency of DHAR in a non-linear synergistic fashion (Shin et al., 2013). Interestingly, the drought levels of DHAR (also of MDHAR and MDA) do not significantly correlate with variation in PH_{DRO} . This suggests that there is abundant genotypic variation for the activity of this enzyme under drought in both the tall, low-yielding traditional landrace varieties and in the shorter, higher-yielding modern varieties of the panel. This combination of findings makes the ascorbate–glutathione cycle, and DHAR in particular, interesting as

breeding targets for improving drought tolerance of rice varieties at the reproductive stage.

Conclusions

This study provides evidence that metabolic/biochemical traits, referred to as endophenotypes, can help bridge the gap between the genome and the visible phenome of plants, and that they outperform the explanatory power of genetic markers when used as variables in models for predicting yield performance under stress. Therefore, breeding pipelines aimed at improving drought resilience in rice could benefit from integrating the information carried by metabolic/biochemical traits representative of the plant endophenome. In particular, our study identified antioxidant enzymes and oxidative stress markers as strong predictors of drought tolerance in rice at the reproductive stage, with higher importance than variables associated with leaf central metabolism. Thus, high activity of leaf antioxidant enzymes and low oxidative damage represent two phenotypes that could guide the development of drought tolerant rice varieties. Despite their value as predictors, using oxidative stress markers and antioxidant enzyme activities as selection tools in breeding is challenging because of their responsiveness to environmental changes, developmental stages, and even diurnal variation. The effort involved in collecting a large number of plant tissue samples in the field, within a limited time window, and synchronizing the developmental stage of many hundreds of accessions, as done in this study, will likely remain a job for the fundamental research community or prebreeding experts, rather than for commercial breeders. Further efforts are needed to translate the information captured by metabolic/biochemical traits into rapid and cost-effective tools for routine breeding application.

Materials and methods

Genetic resources and experimental design

The 271 accessions of rice (*O. sativa* subsp. *indica*) (Supplemental Table S1) were part of a larger panel (approximately 300) used in a field experiment at the International Rice Research Institute, Los Baños, Philippines during the 2013 dry season. The panel includes traditional and improved *indica* rice varieties originating from rice-growing countries in tropical and sub-tropical regions around the world. The panel was evaluated for a number of diverse traits as the basis for GWA mapping (Rebolledo et al., 2016; Kadam et al., 2017, 2018; Melandri et al., 2020b). The experiment comprised a control field and a drought stress field, with three replicates (experimental blocks) of the panel arranged in a serpentine design for each treatment. To synchronize flowering, the accessions were divided into six groups according to the number of days to flowering (previously collected data), and progressively sown and transplanted, with intervals of 10 days between each group. Drought stress consisted of 14 consecutive days of water withholding applied only to the stress field at the reproductive stage (targeting 50% flowering). At the end of the stress

period, the field was re-watered until all the accessions reached maturity for harvest. Further details on the field experiment can be found in [Kadam et al. \(2018\)](#).

Statistical analysis of agronomic traits

Best linear unbiased predictors (BLUEs) of PH, FT, and GY for individual accessions in the same treatment were calculated considering only field replicates (two for control and three for drought) used for the metabolomics and oxidative stress status analyses ([Supplemental Table S1](#)). The BLUEs for each line under the two experimental conditions were calculated by the following general mixed model:

$$y_{ij} = \mu + G_i + \gamma_j + e_{ij}$$

where y_{ij} is the response variable for the i th genotype at the j th block, μ is the intercept, G_i is the effect of the i th genotype, γ_j is the random effect of the j th block with $\gamma_j \sim N(0, \sigma_B^2)$, and e_{ij} is the experimental error. PH under control (PH_{CON}) and drought (PH_{DRO}) conditions was expressed in centimeter. FT under control (FT_{CON}) and drought (FT_{DRO}) conditions were expressed as number of days (calendar days in 2013) required for 50% flowering. GY under control (GY_{CON}) and drought (GY_{DRO}) conditions was expressed in grams/meter square. Percentage of GY loss (GY_{LOSS}) of each accession was calculated as $100 \cdot (GY_{CON} - GY_{DRO}) / (GY_{CON})$. BLUEs of GY_{CON}, GY_{DRO}, and GY_{LOSS} were also calculated considering PH, FT, and PH&FT as covariates by the following general mixed model:

$$y_{ij} = \mu + G_i + \gamma_j + \beta_1 X_{1ij} + \beta_2 X_{2ij} + e_{ij}$$

where y_{ij} is the response variable for the i th genotype at the j th block, μ is the intercept, G_i is the effect of the i th genotype, γ_j is the random effect of the j th block with $\gamma_j \sim N(0, \sigma_B^2)$, X_{1ij} and X_{2ij} are the covariates in the i th genotype and the j th block, and β_1 and β_2 are the regression slopes of the covariates (PH and/or FT) for the corrected BLUEs, and e_{ij} is the experimental error.

For each agronomic trait under the same condition, broad-sense heritability (H^2), which captures the proportion of phenotypic variance explained by genetic factors ([Supplemental Table S3](#)), was calculated by the following formula:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{r}}$$

where σ_g^2 is the genotypic variance, σ_e^2 is the environmental variance, and r is the number of replications.

Leaf tissue sampling, metabolite profiling, oxidative stress status analysis, and data pre-processing

Flag/top leaves of the 271 rice accessions were sampled from control and drought field replicates (two for control and three for drought) and immediately frozen in liquid nitrogen as described in [Melandri et al. \(2020a\)](#). Drought field replicates were collected (09.30–11.00 h) on Day 14 of the

stress treatment. Control field replicates were collected 2 days later, during the same time window. For each accession and condition, equal amounts of leaf tissue from each field replicate were pooled together before performing biochemical analyses. Leaf tissues were analyzed by untargeted GC–MS-based metabolite profiling to assess the variation in polar metabolites as described by [Riewe et al. \(2012\)](#) and [Riewe et al. \(2016\)](#). A total of 88 metabolites were identified, predominantly primary metabolites (amino acids, sugars, and organic acids). Glucose, fructose, and sucrose were quantified spectrophotometrically ([Riewe et al., 2008](#)). The same leaf materials were analyzed for the oxidative stress status. For this, the level of molecular antioxidants (2), oxidative stress markers (2), and the activity of enzymes (16) involved in the antioxidant system and in photorespiration were quantified using high-throughput colorimetric assays ([Zinta et al., 2014](#); [AbdElgawad et al., 2016](#)). Further details on metabolite profiling and analysis of oxidative stress status of the samples can be found in [Melandri et al. \(2020a\)](#). The values of metabolites and oxidative stress markers/enzymes activities were \log_{10} transformed to improve normality before being used for statistical analyses. Imputation of missing values of metabolites and oxidative stress markers/enzyme activities was performed by the function *knnImputation* in the R package “DMwR” ([Torgo, 2010](#)). The list of the 111 metabolites and oxidative stress markers/enzyme activities considered in this study, and their variation among accessions and treatments, are shown in [Supplemental Table S4](#).

Genotypic data

The 271 accessions of this study represent a subgroup of a larger panel of 329 *indica* accessions that were genotyped using genotyping-by-sequencing. The genotypic dataset consisted of 91,591 SNP markers (with 22.8% missing data imputed by the Fast Phase Hidden Markov Model, [Scheet and Stephens, 2006](#)) with minor allele frequency (MAF) ≥ 0.05 ([Rebolledo et al., 2016](#)). The subset of accessions (271) used in this study altered the MAF threshold and therefore the 91,591 SNPs were re-filtered for MAF ≥ 0.05 to exclude rare alleles. The resulting 81,347 SNP map is available in a nucleotide-based hapmap format (hmp) as [Supplemental Data Set 1](#) (.rds). Before being used for modeling, the SNP map was transformed from the hapmap format to a numeric “0, 1, 2” format using an R script, where “0” and “2” denote the major and minor homozygous alleles, respectively, and “1” denotes the heterozygote.

Multivariable models for the prediction of GY traits

Metabolome/oxidative stress status-based and genomic-based multivariable models were used to predict GY performance of the rice accessions. Three different methods were used to generate the prediction models: PLSR, RR-BLUP, and BayesB. For real prediction estimates based on independent data, the multivariable models were built employing a 10-fold CV procedure for which the 271 accessions were randomly subdivided into 10 groups without replacement. These 10 groups were kept the same in generating PLSR, RR-

BLUP, and BayesB prediction models thus allowing for full comparability of their results. Each multivariable model was fit with data from 9th of the groups (training set), while data from the 10th group was used for model testing (test set), and the process was iterated such that each group of samples was used for model testing one time. The predictability (Q^2) for the 10-fold CV models was calculated as follows:

$$Q^2 = 1 - \text{PRESS}/\text{TSS}$$

where PRESS is the predictive residual error sum of squares, and TSS the total sum of squares. PRESS and TSS were calculated as follows:

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2$$

where y_i is the observed GY (GY_{CON} and GY_{DRO} , or GY_{LOSS}) value of the i th individual, \hat{y}_i is the predicted GY value of the i th individual, and \bar{y} is the mean of the predicted GY values of the n (271) individuals.

PLSR models

Let \mathbf{Y} be an $n \times 1$ vector of GY responses (BLUEs of GY_{CON} and GY_{DRO} , or GY_{LOSS}) and \mathbf{X} is an n -observation by p -variable matrix of predictors (the set of 111 metabolites/oxidative stress markers and enzymes or the 81,347 SNP markers), PLSR aims to decompose \mathbf{X} into a set of A orthogonal scores such that the covariance with corresponding \mathbf{Y} scores is maximized. The X -weight and Y -loading vectors that result from the decomposition are used to estimate the vector of regression coefficients, β_{PLS} , such that $\mathbf{Y} = \mathbf{X} \beta_{\text{PLS}} + \varepsilon$ where ε is an $n \times 1$ vector of error terms. The R package “pls” (Mevik and Wehrens, 2007) was used for PLSR in this study. Each variable was centered (mean subtraction) and scaled (standard deviation division) before analysis. In the 10-fold CV procedure, for each training set, a PLSR model was constructed with the GY trait as a single dependent variable (\mathbf{Y}) and the set of metabolites/oxidative stress markers and enzymes or the SNP markers as the independent variables (\mathbf{X}). To choose the appropriate number of factors for each training model (A from above), leave-one-out cross validation was used to estimate root mean squared error (RMSECV) for models fit with zero through 10 factors (linear combinations of the metabolites/oxidative stress markers and enzymes or of the SNP markers), and the model that produced the smallest RMSECV was selected for prediction of the GY trait in the test set.

RR-BLUP models

The RR-BLUP model is described as follows:

$$y_i = \mu + \sum_{k=1}^p x_{ik} \beta_k + \varepsilon_i$$

where y_i is the GY response (BLUE of GY_{CON} and GY_{DRO} , or GY_{LOSS}) of the i th individual, μ is the intercept, x_{ik} is the

genotype at the k th predictor of the i th individual, p is the total number of predictors (the set of 111 metabolites/oxidative stress markers and enzymes or the 81,347 SNP markers), β_k is the estimated random additive effect of the k th predictor with $\beta_k \sim N(0, \sigma_g^2)$, and ε_i is the residual error term with $\varepsilon_i \sim N(0, \sigma_e^2)$. The BLUP of each β_k received the following penalty:

$$J(\beta) = \sum_{k=1}^p \beta_k^2$$

where all the terms are the same as those described above. This model was implemented using the R package “rrBLUP” (Endelman, 2011).

BayesB models

The basic model of BayesB is the same as RR-BLUP, but in this case all parameters are treated as random variables in a Bayesian framework, and we do not assume the same variance for all predictor effects. The prior distributions were defined as $g \sim N(0, D)$ and $e \sim N(0, \sigma_e^2 I)$, where $D = \text{diag}(\sigma_{g1}^2, \dots, \sigma_{gp}^2)$, for the intercept μ we assumed a flat prior. For each l , the prior distribution of σ_{gi}^2 is assumed to be zero with probability π and a scaled inverse chi-squared distribution with probability $(1 - \pi)$. The prior of π is a beta distribution. The prior of σ_e^2 is also a scaled inverse chi-squared distribution. A Gibbs sampler algorithm was then applied to infer all the parameters in the model. The BayesB model was implemented using the R package “BGLR” (Pérez and De Los Campos, 2014).

Predictor importance for the MetabOxi-based models

For the metabolome/oxidative stress status-based models, the relative importance of the predictors was summarized using rank-products (Smit et al., 2007; Mumm et al., 2016). To this purpose, for each of the 10 different single sub-models (generated by the cross-validating procedure), the predictors were ranked (from 1 to 111) based on their absolute regression coefficient (with rank 1 for the predictors with the highest absolute value). Then, for each predictor, the rank numbers from the 10 sub-models were multiplied together, giving a final rank-product for the overall model. A low rank-product implies that a predictor is of high importance in the model.

Combined MetabOxi and genomic-based models

For the PLSR models based on the combined MetabOxi and genomic data, a single data matrix of predictors (P -variables) was generated concatenating the two datasets based on the samples (n -observations). Then, the PLSR models for the prediction of GY responses were built as for the single datasets, described above. The combined RR-BLUP and BayesB models were built as follows:

$$y = X_{\text{SNP}} \beta_{\text{SNP}} + X_{\text{MET}} \beta_{\text{MET}} + \varepsilon$$

where SNP and MET denote the genomic and MetabOxi data, respectively. Parameter assumptions were the same as

for the prediction models based on a single dataset described above.

Multi-trait MetabOxi and genomic-based models

For the multi-trait PLSR and RR-BLUP models based on MetabOxi and genomic data, the values of the secondary traits (PH, FT, and both) were included in the training sets of the CV procedure, but not in the test sets. Then the models were run as described above. It was not possible to run a multi-trait BayesB model because of the specific R package (“BGLR”; Pérez and De Los Campos, 2014) we used.

Supplemental data

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

Supplemental Figure S1. Boxplots representing the PH performance under control (PH_{CON}) and drought (PH_{DRO}) conditions of the rice accessions carrying the minor (AA) or major (GG) alleles at the locus (Chr1 pos: 38,286,772 bp; Supplemental Data Set 1) associated with the gibberellin 20-oxidase biosynthetic gene (*SEMI-DWARF1*; recessive *sd1* and functional wild-type *SD1* allele).

Supplemental Figure S2. Distributions of GY traits in the 271 *indica* rice accessions sorted by increasing PH.

Supplemental Table S1. PH, FT, and GY of the 271 *indica* rice accessions.

Supplemental Table S2. Agronomic trait performance of the 271 *indica* rice accessions.

Supplemental Table S3. Heritabilities and variances for PH, FT, and GY of the 271 *indica* rice accessions.

Supplemental Table S4. Flag leaf values of the 111 MetabOxi variables in the 271 *indica* rice accessions under control and drought conditions.

Supplemental Table S5. Predictability of multi-trait MetabOxi- and genomic-based models for the prediction of GY traits.

Supplemental Table S6. Ranking of the MetabOxi-variables in the PLSR, RR-BLUP, and BayesB models for the prediction of GY under control conditions.

Supplemental Table S7. Ranking of the MetabOxi-variables in the PLSR, RR-BLUP, and BayesB models for the prediction of GY under drought conditions.

Supplemental Table S8. Ranking of the MetabOxi-variables in the PLSR, RR-BLUP, and BayesB models for the prediction of drought-induced GY loss.

Supplementary Table S9. Correlations between control values of the 111 MetabOxi-variables and GY, FT, and PH under control conditions.

Supplementary Table S10. Correlations between drought values of the 111 MetabOxi-variables and GY, GY loss, FT, and PH under drought conditions.

Supplemental Table S11. Results of the linear models created by fitting GY traits (response) and the top-ranked predictors (single explanatory variable) identified by the MetabOxi-based models.

Supplemental Data Set 1. A total of 81,347 SNP map in hapmap (hmp) format.

Acknowledgments

The authors wish to thank Dr Willem Kruijer for his critical reading and suggestions to improve the draft versions of the manuscript.

Funding

This work is part of the ‘Growing Rice like Wheat’ research programme financially supported by an anonymous private donor, via Wageningen University Fund, for the first author’s PhD fellowship (Giovanni Melandri). The Department of Plant Biology at Facultad de Agronomía, Universidad de la República, financially supported Eliana Monteverde while performing this research. GC–MS analysis was enabled by the Transnational Access capacities of the European Plant Phenotyping Network (EPPN, grant agreement no. 284443) funded by the FP7 Research Infrastructures Programme of the European Union. We also acknowledge financial support from the Bill and Melinda Gates Foundation from the ‘Rapid Mobilization of Alleles for Rice Cultivar Improvement in Sub-Saharan Africa’ project at Cornell University.

Conflict of interest statement. Authors have no conflict of interest to declare.

References

- Abdelgawad H, Zinta G, Hegab MM, Pandey R, Asard H, Abuelsoud W (2016) High salinity induces different oxidative stress and antioxidant responses in maize seedlings organs. *Front Plant Sci* 7: 1–11
- Araújo WL, Nunes-Nesi A, Nikoloski Z, Sweetlove LJ, Fernie AR (2012) Metabolic control and regulation of the tricarboxylic acid cycle in photosynthetic and heterotrophic plant tissues. *Plant, Cell Environ* 35: 1–21
- Asano K, Yamasaki M, Takuno S, Miura K, Katagiri S, Ito T, Doi K, Wu J, Ebana K, Matsumoto T, et al. (2011) Artificial selection for a green revolution gene during japonica rice domestication. *Proc Natl Acad Sci USA* 108: 11034–11039
- Biswal AK, Kohli A (2013) Cereal flag leaf adaptations for grain yield under drought: Knowledge status and gaps. *Mol Breed* 31: 749–766
- 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
- Cooper M, Gho C, Leafgren R, Tang T, Messina C (2014) Breeding drought-tolerant maize hybrids for the US corn-belt: discovery to product. *J Exp Bot* 65: 6191–6194
- 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: 1–9
- Dwivedi SL, Ceccarelli S, Blair MW, Upadhyaya HD, Are AK, Ortiz R (2016) Landrace germplasm for improving yield and abiotic stress adaptation. *Trends Plant Sci* 21: 31–42
- Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4: 250–255
- Fernandez O, Millet EJ, Rincenc R, Prigent S, Pétriacq P, Gibon Y (2021) Plant metabolomics and breeding. *Adv Bot Res* 98: 207–235

- Fernie AR, Stitt M** (2012) On the discordance of metabolomics with proteomics and transcriptomics: Coping with increasing complexity in logic, chemistry, and network interactions. *Plant Physiol* **158**: 1139–1145
- Ferreres F, Figueiredo R, Bettencourt S, Carqueijeiro I, Oliveira J, Gil-Izquierdo A, Pereira DM, Valentão P, Andrade PB, Duarte P, et al.** (2011) Identification of phenolic compounds in isolated vacuoles of the medicinal plant *Catharanthus roseus* and their interaction with vacuolar class III peroxidase: An H₂O₂ affair? *J Exp Bot* **62**: 2841–2854
- Foyer CH, Noctor G** (2011) Ascorbate and glutathione: the heart of the redox hub. *Plant Physiol* **155**: 2–18
- Foyer CH, Shigeoka S** (2010) Understanding oxidative stress and antioxidant functions to enhance photosynthesis. *Plant Physiol* **155**: 93–100
- Großkinsky DK, Svendsgaard J, Christensen S, Roitsch T** (2015) Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. *J Exp Bot* **66**: 5429–5440
- Höskuldsson A, Svinning K** (2006) Modelling of multi-block data. *J Chemom* **20**: 376–385
- Jamali IN, Remali J, Azizan KA, Nor Muhammad NA, Arita M, Goh H-H, Aizat WM** (2020) Systematic multi-omics integration (MOI) approach in plant systems biology. *Front Plant Sci* **11**: 944
- Kadam NN, Struik PC, Rebolledo MC, Yin X, Jagadish SVK** (2018) Genome-wide association reveals novel genomic loci controlling rice grain yield and its component traits under water-deficit stress during the reproductive stage. *J Exp Bot* **69**: 4017–4032
- Kadam NN, Tamilselvan A, Lawas LMF, Quinones C, Bahuguna RN, Thomson MJ, Dingkuhn M, Raveendran M, Struik PC, Yin X, et al.** (2017) Genetic control of plasticity in root morphology and anatomy of rice in response to water deficit. *Plant Physiol* **174**: 2302–2315
- Keurentjes JJ** (2009) Genetical metabolomics: closing in on phenotypes. *Curr Opin Plant Biol* **12**: 223–230
- Kleessen S, Laitinen R, Fusari CM, Antonio C, Sulpice R, Fernie AR, Stitt M, Nikoloski Z** (2014) Metabolic efficiency underpins performance trade-offs in growth of *Arabidopsis thaliana*. *Nat Commun* **5**: 1–10
- Kumar R, Bohra A, Pandey AK, Pandey MK, Kumar A** (2017) Metabolomics for plant improvement: Status and prospects. *Front Plant Sci* **8**: 1–27
- Kundu A, Vadassery J** (2019) Chlorogenic acid-mediated chemical defence of plants against insect herbivores. *Plant Biol* **21**: 185–189
- Lafitte HR, Yongsheng G, Yan S, Li ZK** (2007) Whole plant responses, key processes, and adaptation to drought stress: the case of rice. *J Exp Bot* **58**: 169–175
- Lempe J, Lachowicz J, Sullivan AM, Queitsch C** (2013) Molecular mechanisms of robustness in plants. *Curr Opin Plant Biol* **16**: 62–69
- Melandri G, AbdElgawad H, Floková K, Jamar DC, Asard H, Beemster GTS, Ruyter-Spira C, Bouwmeester HJ** (2021) Drought tolerance in selected aerobic and upland rice varieties is driven by different metabolic and antioxidative responses. *Planta* **254**: 13
- Melandri G, AbdElgawad H, Riewe D, Hageman JA, Asard H, Beemster GTS, Kadam N, Jagadish K, Altmann T, Ruyter-Spira C, et al.** (2020a) Biomarkers for grain yield stability in rice under drought stress. *J Exp Bot* **71**: 669–683
- Melandri G, Prashar A, McCouch SR, Van Der Linden G, Jones HG, Kadam N, Jagadish K, Bouwmeester H, Ruyter-Spira C** (2020b) Association mapping and genetic dissection of drought-induced canopy temperature differences in rice. *J Exp Bot* **71**: 1614–1627
- Mevik BH, Wehrens R** (2007) The pls package: principal component and partial least squares regression in R. *J Stat Softw* **18**, <https://doi.org/10.18637/jss.v018.i02>
- Meyer RC, Steinfath M, Lisek J, Becher M, Witucka-Wall H, Törjék O, Fiehn O, Eckardt Ä, 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
- Møller IM, Jensen PE, Hansson A** (2007) Oxidative modifications to cellular components in plants. *Annu Rev Plant Biol* **58**: 459–481
- Mumm R, Hageman JA, Calingacion MN, de Vos RCH, Jonker HH, Erban A, Kopka J, Hansen TH, Laursen KH, Schjoerring JK, et al.** (2016) Multi-platform metabolomics analyses of a broad collection of fragrant and non-fragrant rice varieties reveals the high complexity of grain quality characteristics. *Metabolomics* **12**: 1–19
- Niggeweg R, Michael AJ, Martin C** (2004) Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nat Biotechnol* **22**: 746–754
- Pérez P, De Los Campos G** (2014) Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**: 483–495
- Rebolledo MC, Peña AL, Duitama J, Cruz DF, Dingkuhn M, Grenier C, Tohme J** (2016) Combining image analysis, genome wide association studies and different field trials to reveal stable genetic regions related to panicle architecture and the number of spikelets per panicle in rice. *Front Plant Sci* **7**: 1–12
- Reinke SN, Galindo-Prieto B, Skotare T, Broadhurst DI, Singhanian A, Horowitz D, Djukanović R, Hinks TSC, Geladi P, Trygg J, et al.** (2018) OnPLS-based multi-block data integration: a multivariate approach to interrogating biological interactions in asthma. *Anal Chem* **90**: 13400–13408
- Reynolds M, Langridge P** (2016) Physiological breeding. *Curr Opin Plant Biol* **31**: 162–171
- Rice BR, Lipka AE** (2021) Diversifying maize genomic selection models. *Mol Breed* **41**: 33
- Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisek J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE** (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat Genet* **44**: 217–220
- Riewe D, Grosman L, Zauber H, Wucke C, Fernie AR, Geigenberger P** (2008) Metabolic and developmental adaptations of growing potato tubers in response to specific manipulations of the adenylate energy status. *Plant Physiol* **146**: 1579–1598
- Riewe D, Jeon HJ, Lisek J, Heuermann MC, Schmeichel J, Seyfarth M, Meyer RC, Willmitzer L, Altmann T** (2016) A naturally occurring promoter polymorphism of the *Arabidopsis* FUM2 gene causes expression variation, and is associated with metabolic and growth traits. *Plant J* **88**: 826–838
- Riewe D, Koohi M, Lisek J, Pfeiffer M, Lippmann R, Schmeichel J, Willmitzer L, Altmann T** (2012) A tyrosine aminotransferase involved in tocopherol synthesis in *Arabidopsis*. *Plant J* **71**: 850–859
- Scheet P, Stephens M** (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**: 629–644
- Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE** (2018) Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics* **208**: 1373–1385
- Scossa F, Alseekh S, Fernie AR** (2021) Integrating multi-omics data for crop improvement. *J Plant Physiol* **257**: 153352
- Shin SY, Kim MH, Kim YH, Park HM, Yoon HS** (2013) Co-expression of monodehydroascorbate reductase and dehydroascorbate reductase from *Brassica rapa* effectively confers tolerance to freezing-induced oxidative stress. *Mol Cells* **36**: 304–315
- Smirnoff N** (2011) Vitamin C: the metabolism and functions of ascorbic acid in plants. *Adv Bot Res* **59**: 107–177
- Smit S, van Breemen MJ, Hoefslot HCJ, Smilde AK, Aerts JMFG, de Koster CG** (2007) Assessing the statistical validity of proteomics based biomarkers. *Anal Chim Acta* **592**: 210–217
- Steinfath M, Gärtner T, Lisek J, Meyer RC, Altmann T, Willmitzer L, Selbig J** (2010) Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. *Theor Appl Genet* **120**: 239–247

- Stitt M, Sulpice R, Keurentjes J** (2010) Metabolic networks: how to identify key components in the regulation of metabolism and growth. *Plant Physiol* **152**: 428–444
- Sulpice R** (2020) Closing the yield gap: can metabolomics be of help? *J Exp Bot* **71**: 461–464
- Sulpice R, McKeown PC** (2015) Moving toward a comprehensive map of central plant metabolism. *Annu Rev Plant Biol* **66**: 187–210
- 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
- Sweetlove LJ, Beard KFM, Nunes-Nesi A, Fernie AR, Ratcliffe RG** (2010) Not just a circle: Flux modes in the plant TCA cycle. *Trends Plant Sci* **15**: 462–470
- Tong H, Nikoloski Z** (2021) Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *J Plant Physiol* **257**: 153354
- Torgo L** (2010) *Data Mining with R: Learning with Case Studies*, Chapman & Hall/CRC, Boca Raton, FL
- van Dijk ADJ, Kootstra G, Kruijer W, de Ridder D** (2021) Machine learning in plant science and plant breeding. *iScience* **24**: 101890
- Volpi e Silva N, Mazzafera P, Cesarino I** (2019) Should I stay or should I go: are chlorogenic acids mobilized towards lignin biosynthesis? *Phytochemistry* **166**: 112063
- Wang S, Wei J, Li R, Qu H, Chater JM, Ma R, Li Y, Xie W, Jia Z** (2019) Identification of optimal prediction models using multi-omic data for selecting hybrid rice. *Heredity (Edinb)* **123**: 395–406
- Westhues M, Schrag TA, Heuer C, Thaller G, Utz HF, Schipprack W, Thiemann A, Seifert F, Ehret A, Schlereth A, et al.** (2017) Omics-based hybrid prediction in maize. *Theor Appl Genet* **130**: 1927–1939
- Xing Y, Zhang Q** (2010) Genetic and molecular bases of rice yield. *Annu Rev Plant Biol* **61**: 421–442
- Xu S, Xu Y, Gong L, Zhang Q** (2016) Metabolomic prediction of yield in hybrid rice. *Plant J* **88**: 219–227
- Xu Y, Zhao Y, Wang X, Ma Y, Li P, Yang Z, Zhang X, Xu C, Xu S** (2020) Incorporation of parental phenotypic data into multi-omic models improves prediction of yield-related traits in hybrid rice. *Plant Biotechnol J* **19**: 261–272
- Yin X, Struik PC, Kropff MJ** (2004) Role of crop physiology in predicting gene-to-phenotype relationships. *Trends Plant Sci* **9**: 426–432
- Yoshida S** (1972) Physiological aspects of grain yield. *Annu Rev Plant Physiol* **23**: 437–464
- Zhu G, Peng S, Huang J, Cui K, Nie L, Wang F** (2016) Genetic improvements in rice yield and concomitant increases in radiation- and nitrogen-use efficiency in middle reaches of Yangtze River. *Sci Rep* **6**: 1–12
- Zinta G, Abdelgawad H, Domagalska MA, Vergauwen L, Knapen D, Nijs I, Janssens IA, Beemster GTS, Asard H** (2014) Physiological, biochemical, and genome-wide transcriptional analysis reveals that elevated CO₂ mitigates the impact of combined heat wave and drought stress in *Arabidopsis thaliana* at multiple organizational levels. *Glob Chang Biol* **20**: 3670–3685