Genome Biology

RESEARCH

Open Access

Integrative phenomics, metabolomics and genomics analysis provides new insights for deciphering the genetic basis of metabolism in polished rice

Hui Feng[1†], Yufei Li[2,3†], Guoxin Dai[1†], Zhuang Yang[3†], Jingyan Song[1], Bingjie Lu[1], Yuan Gao[1], Yongqi Chen[1], Jiawei Shi[1], Luis A. J. Mur[6], Lejun Yu[5*], Jie Luo[3,4*] and Wanneng Yang[1*]

[†]Hui Feng, Yufei Li, Guoxin Dai and Zhuang Yang contributed equally.

*Correspondence: yulj@hainanu.edu.cn; jie.luo@hainanu.edu.cn; ywn@mail.hzau.edu.cn

[1] National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China
[2] Hainan Yazhou Bay Seed Laboratory, Sanya 572025, China
[3] School of Breeding and Multiplication (Sanya Institute of Breeding and Multiplication), Hainan University, Sanya 572025, China
[4] Yazhouwan National Laboratory, Sanya 572025, China
[5] State Key Laboratory of Digital Medical Engineering, Key Laboratory of Biomedical Engineering of Hainan Province, School of Biomedical Engineering, Hainan University, Sanya 572025, China
[6] Department of Life Sciences, Aberystwyth University, Aberystwyth, Wales SY23 3DA, UK

Abstract

Background: Metabolomics is one of the most widely used omics tools for deciphering the functional networks of the metabolites for crop improvement. However, it is technically demanding and costly.

Results: We propose a relatively inexpensive approach for metabolomics analysis in which metabolomics is combined with hyperspectral imaging via machine learning. This approach can be used to target important steps in flavonoid and lipid biosynthesis in rice. We extract 1848 hyperspectral indices and 887 metabolites from polished grains of 533 *Oryza sativa* accessions. Hyperspectral indices are then linked to metabolites through correlation analysis and modelling. Based on this, a total of 554 metabolites and 1313 hyperspectral indices are identified for further genome-wide association study (GWAS). By GWAS, we detect 17,509 significant locus-trait associations with 2882 single nucleotide polymorphisms (SNPs). Colocalization analysis links these SNPs to the corresponding metabolites and hyperspectral indices. We detect 6415 pairs of metabolites and hyperspectral indices within a linkage disequilibrium of 300 kb in the *Oryza sativa* genome. We then characterize 1761 candidate genes colocalized to these loci by transcriptomic analysis. We further verify novel candidate genes encoding a novel flavonoid (*LOC_Os09g18450*) and a flavonoid/lipid (*LOC_Os07g11020*) respectively by gene editing and overexpression in rice.

Conclusion: Our findings indicate that hyperspectral imaging combined with machine learning methods could serve as a powerful tool for quickly and inexpensively assessing crop metabolites.

Keywords: Hyperspectral imaging, Metabolism, Machine learning, Genome-wide association study, Polished rice

Background

Cultivated rice is vital to global food security, feeding more than 50% of the global human population [1]. An additional 112 million metric tons of rice is estimated to be necessary for feeding the ever-increasing population by 2035. This is a major challenge given the decreases in arable land and irrigation water as well as the unpredictable climatic conditions [2]. Thus, there is a need to develop new elite rice accessions that are resistant to various biotic and abiotic stresses to secure future rice production [2]. In terms of breeding targets, rice primary metabolites play important roles in maintaining growth and development [3–5], and secondary metabolites can protect the plants from the detrimental effects of stress [5–7]. For example, anthocyanins are involved in pollination, hormone regulation and biotic or abiotic stress responses, which can be linked to their strong antioxidant effects [8, 9]. Moreover, due to these properties, anthocyanins can prevent cardiovascular disease and have anticancer effects in humans [10]. Like anthocyanins, many rice metabolites are beneficial to human health or are able to improve learning and memory [11], either as essential nutrients or via medicinal effects [12]. Phospholipids can inhibit tumour growth and metastasis [13] and can have anti-inflammatory effects [14].

Numerous methods have been developed for the qualitative and quantitative measurement of the > 200,000 diverse metabolites identified in the plant kingdom [15]. Among metabolomic approaches, nuclear magnetic resonance (NMR) spectroscopy- or mass spectrometry (MS)-based methods, such as liquid chromatography–mass spectrometry (LC–MS), gas chromatography–mass spectrometry (GC–MS), ultra-performance LC and high-resolution mass spectrometry (UPLC–MS), electrospray ionization mass spectrometry (ESI–MS) and capillary electrophoresis–mass spectrometry (CE–MS), are the typical methods used to assess large numbers of metabolites rapidly [15]. In addition, for the targeted assessment of particular metabolites, the MALDI–MSI and IMS–MS methods are usually the optimal choices [16]. Among all of the metabolomic platforms, LC–MS and its derivative approaches, such as UPLC–MS, represent the most comprehensive techniques with high metabolite resolving power and sensitivity. For example, the levels of 840 diverse metabolites, including primary metabolites such as nucleotides, amino acids and fatty acids, as well as secondary metabolites such as polyamines, terpenoids and flavonoids, were measured in the leaves of 529 *O. sativa* accessions [4]. UPLC–MS was used to measure the levels of 825 annotated metabolites during the entire rice developmental cycle [17]. By using LC–MS, the levels of 85 flavonoids were measured in 14 plant species, with rice exhibiting the highest accumulation of flavones [18]. These powerful platforms have promoted the vigorous development of crop metabolomic research; nevertheless, the sample preparation process for metabolite measurement is cumbersome and time consuming. Thus, efficient, accurate, and high-throughput metabolite detection technology is needed to improve the application of metabolomics in plant science.

Spectroscopy is a fast and non-destructive technique that can be employed for chemical measurement [19]. The variable absorbance, reflection, or penetrability of light in plants can indicate the intrinsic chemical content as well as the surface structure. Thus, optical sensors can be used to measure the chemicals accumulated in plants during their life cycle [20]. Typically, the levels of macromolecular chemicals (e.g. proteins

and starch) within plants can be predicted precisely and through a limited number of spectral bands because of their higher concentrations than those of small-molecule chemicals, such as metabolites. To date, protein and starch levels have been predicted precisely in many plants, such as rice [21], wheat [22] and maize [23]. As small-molecule metabolite detection is relatively difficult, to date, few studies have focused on the prediction of metabolite levels in plants via spectroscopy. The emergence and development of hyperspectral imaging (HSI) technology has enabled the acquisition of high-resolution continuous spectral information in the visible and infrared bands, which makes it easier to predict the content of low-molecular-weight chemicals in plants. Indeed, two hundred metabolites in the ears and leaves of wheat plants were non-destructively measured through HSI and LASSO regression prediction models, with 32 metabolites being predicted with R^2 values greater than 0.30 [24]. The pelargonidin-3-glucoside (P3G) contents of strawberry plants at different harvest stages were predicted via HSI and partial least square regression (PLSR). The wavelength (1303 nm) related to anthocyanins was found to be the most meaningful wavelength for predicting P3G content [25]. The total anthocyanin content of mulberry fruit was predicted precisely through Vis–NIR HSI and PLSR regression [26]. In addition, HSI can be used to predict the contents of trace elements in wheat grain and flour. By combining the HSI and PLSR prediction methods, calcium, magnesium, molybdenum and zinc levels were predicted precisely with model R^2 values of up to 0.50 [27]. As listed above, HSI can serve as a powerful tool for quantification of small-molecule chemicals, such as metabolites and trace elements, while HSI is used mainly for the prediction of metabolite levels in fruits and vegetables [28]. Few studies have focused on stable crops, especially large populations, and no research on metabolite level measurement by HSI in polished rice has been published thus far.

Genome-wide association study (GWAS) is a powerful tool for analysing complex quantitative traits and screening important candidate genes in plants. The GWAS strategy can be further advanced via combination with metabolomics (mGWAS). In rice, Chen et al. performed a quantitative analysis of 840 biochemical metabolites in 524 natural rice populations and used mGWAS to identify many important genetic loci associated with different metabolites [4]. Chen et al. quantified 805 metabolites in 182 wheat cultivars that could be associated with 1098 mGWAS associations with large effects. The associations were validated via enzymatic assays of the targeted gene products in wheat seeds, which revealed correlations with flavonoid pathways [29]. Another advancement of the GWAS strategy is its combination with phenomics (pGWAS). Yang et al. identified 141 associated loci via GWAS of 15 traits that were acquired non-destructively by high-throughput rice phenotyping, 25 of which included known genes such as the Green Revolution semidwarf gene, *SD1* [30]. Moreover, particular features acquired from HSI and models of macromolecular chemicals, such as proteins [21] and chlorophyll [31], can reportedly be used as indicators for GWAS. Nevertheless, no published research has elucidated whether the features acquired from HSI and models of small-molecule chemicals, such as metabolites, can serve as indicators for GWAS.

In this study, we utilized automatic HSI and a high-performance liquid chromatography (HPLC)-based targeted method to extract hyperspectral indices and metabolites from 533 *O. sativa* accessions. Correlation analysis and eight machine learning methods were used to identify important hyperspectral indices that were associated with the

corresponding metabolites. GWAS was conducted for both metabolites (mGWAS) and hyperspectral indices (hGWAS), and significant SNPs with $p < 1.3E-6$ were selected. To further explore the genetic relationships between metabolites and hyperspectral indices, colocalization analysis was performed on the basis of the screened SNPs from the mGWAS and hGWAS. Gene expression selection and KEGG keyword mapping were subsequently conducted on the basis of the colocalized loci. Furthermore, networks connecting metabolites and hyperspectral indices were established on the basis of the screened candidate genes. The roles of two genes screened from the networks, *LOC_Os07g11020* and *LOC_Os09g18450*, in their respective pathways were validated via gene-editing and overexpression experiments in rice. This work shows how hyperspectral indices that are both phenotypically and genetically associated with corresponding metabolites could serve as indicators of metabolites in GWAS. This strategy can be used to accelerate plant breeding programs.

Results

Acquisition of metabolites and hyperspectral indices of polished rice

To explore the potential of using hyperspectral indices as indicators of metabolites in polished rice grains for GWAS, 533 *O. sativa* accessions (Additional file 1: Table S1; Additional file 2: Fig. S1) containing 4.3 M SNPs were cultivated at Huazhong Agricultural University, Wuhan city. Seeds of these accessions were harvested in 2019 and used in our analysis (see the "Methods" section). First, 20 seeds of each accession were randomly selected (Fig. 1a), placed in a grinder (Fig. 1b) and ground to powder (Fig. 1c). The powder was used for spectral reflectance acquisition via an automatic hyperspectral image acquisition system (Fig. 1d) and for HPLC-based targeted assessment (Fig. 1e). The matched hyperspectral indices ranging from 400 to 1700 nm (Fig. 1f) and the metabolite traits of 14 groups of each *O. sativa* accession were extracted for further joint analysis (Fig. 1g). For each accession, the levels of a total of 887 diverse metabolites (Fig. 1h) were measured. The average value, standard deviation (SD) and coefficient of variation (CV) of the metabolites in the whole group and the 5 subpopulations, designated Admix, Aus, Indica, Japonica and VI, were calculated. As shown in Additional file 2: Fig. S2, the frequency of the CVs of the metabolites in the 5 subpopulations was consistent with that of the metabolites in the whole group, with a majority of the CVs within 20–40%. Additionally, 1848 hyperspectral indices were targeted through an image analysis pipeline constructed for hyperspectral index extraction (Fig. 1i). Similarly, the average value, SD and CV of the hyperspectral indices were also calculated. As shown in Additional file 2: Fig. S3, the frequency of CVs of the hyperspectral indices in the 5 subpopulations was also consistent with that of the whole group, with a majority of the CVs within 10–30%. Then, correlation analysis (Fig. 1j) and machine learning methods (Fig. 1k) were subsequently performed to screen for important hyperspectral indices of the corresponding metabolites. The screened matched pairs of important hyperspectral indices and metabolites (Fig. 1l) were related to particular phenotypes. GWAS was performed, and then SNPs colocalized with metabolites and hyperspectral indices were screened (Fig. 1m). Gene expression selection and KEGG mapping were conducted on the basis of the colocalized SNPs (Fig. 1n). Finally, several candidate genes

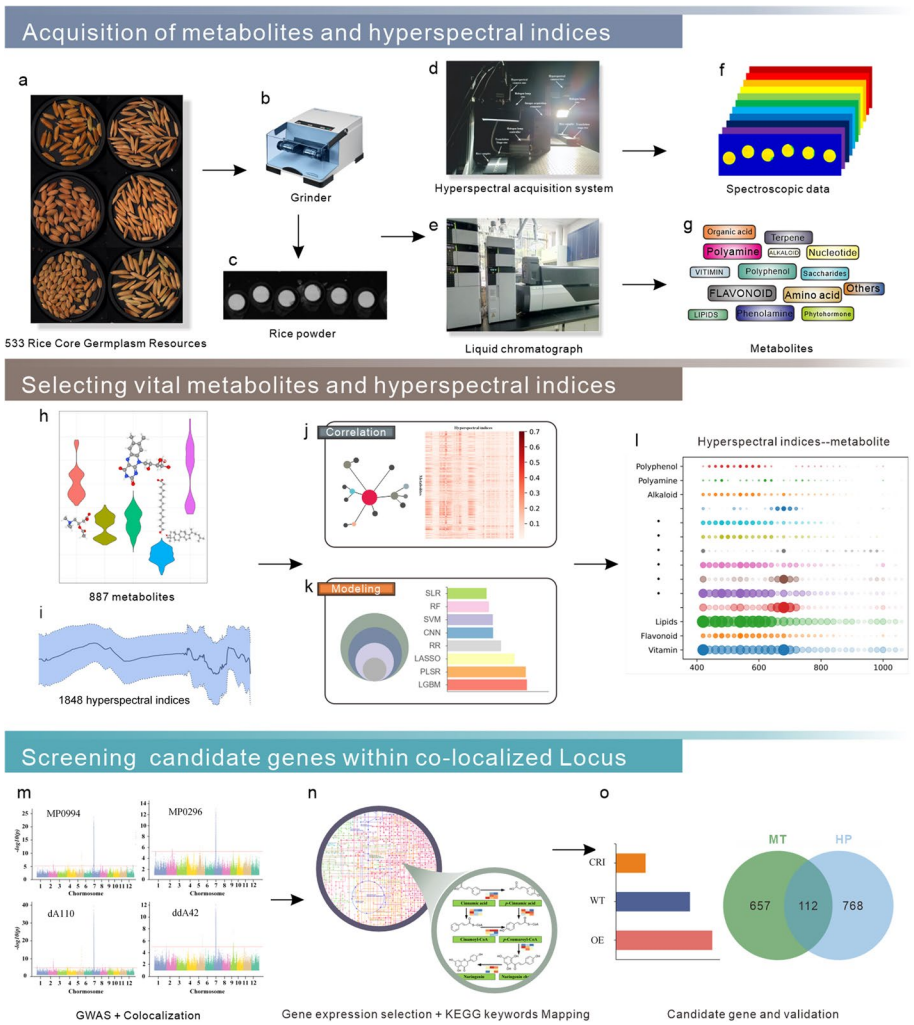

Fig. 1 Combination of hyperspectral imaging and machine learning to decipher the genetic architecture of rice grain metabolism. For the acquisition of metabolites and hyperspectral indices, 533 rice core germplasm resources were collected (**a**); a grinder (**b**) was used to grind them into powder (**c**); and a hyperspectral imaging system (**d**) and a metabolite measurement system (**e**) to acquire the corresponding spectroscopic data (**f**) and metabolic data (**g**). To select vital metabolites and hyperspectral indices, several procedures were performed on the acquired 887 metabolites (**h**) and 1848 hyperspectral indices (**i**), including correlation analysis (**j**) and machine learning-based modelling (**k**); the phenotypically related metabolites and hyperspectral indices thus identified were screened (**l**). For further screening of genetically related metabolites and hyperspectral indices, GWASs and colocalization analyses were performed (**m**), and colocalized loci functioning in corresponding metabolic pathways were identified (**n**), after which several candidate genes were screened and validated (**o**)

that were both phenotypically and genetically related to metabolites and hyperspectral indices were targeted for validation (Fig. 1o). Descriptions of the process and details of these spectral traits are shown in Additional file 1: Table S2 and Table S3 and in Additional file 2: Fig. S4 and Fig. S5. The original hyperspectral images obtained at 400–1700 nm for 533 *O. sativa* accessions were available at http://plantphenomics.hzau.edu.cn/usercrop/Rice/image/2024-HSI.

Screening important hyperspectral indices of metabolites through correlation analysis and machine learning methods

A total of 1848 hyperspectral indices and 887 metabolites (Fig. 2a and Additional file 1: Table S4 and Table S5) were used in this analysis. Pearson correlation coefficients of the hyperspectral indices and metabolites were calculated, and matching pairs with values greater than 0.30 were identified. This method targeted 551 metabolites and 1284 hyperspectral indices (Fig. 2b and Additional file 1: Table S6). Moreover, eight machine learning methods, namely, partial least squares regression (PLSR), light gradient boosting machine (LGBM) regression, least absolute shrinkage and selection operator (LASSO) regression, ridge regression (RR), convolutional neural network (CNN) regression, support vector machine (SVM) regression, random forest (RF) regression and stepwise linear regression (SLR), were used for metabolite prediction and vital hyperspectral index selection. Of the eight models, LGBM performed the best because there were 235 metabolites with an R_p greater than 0.50. Whereas, for the other methods listed above,

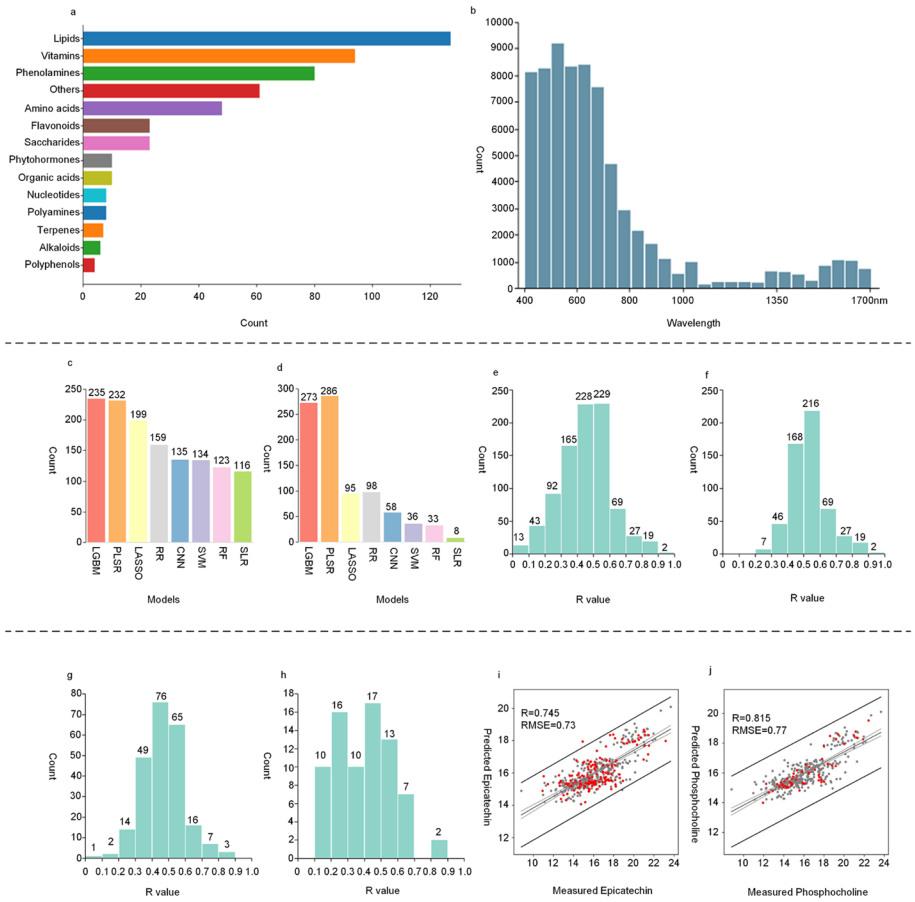

Fig. 2 Information on phenotypically related metabolites and hyperspectral indices. **a** The number of metabolites in each group. **b** Distribution of hyperspectral indices from 400 to 1700 nm that are phenotypically related to metabolites. **c** The number of metabolites with R_p values greater than 0.5 identified through modelling. **d** The number of metabolites with the best R_p values among the eight machine learning models. **e** The R_p distribution of all 887 measured metabolites. **f** The R_p distribution of the 554 screened metabolites. **g** The R_p distribution of lipid groups. **h** The R_p distribution of the flavonoid groups. **i** Scatter plot of epicatechin. **j** Scatter plot of phosphocholine

the number of metabolites with R_p greater than 0.50 was 232, 199, 159, 135, 134, 123 and 116, respectively (Fig. 2c and Additional file 1: Table S7). For each metabolite, when considering the highest R_p value of the eight machine learning models as the optimal R_p value, PLSR and LGBM performed well in predicting the greatest number of metabolites (286 and 273 metabolites, respectively), with optimal R_p values being derived from these two models. The other models did not perform so well, as the number of metabolites with optimal R_p values ranged from 8 to 98 (Fig. 2d). Using the highest R_p from the eight models for each metabolite, the R_p of all 887 targeted metabolites followed a normal distribution, with an average value of 0.45 (Fig. 2e). However, an R_p of over 0.50 was set as the threshold for highly predictive metabolite screening, as there was at least one metabolite with R_p over 0.50 in each metabolite group. This led to 346 metabolites with high R_p (> 0.50) values being identified in each of the eight derived models, and 859 corresponding vital hyperspectral indices were screened (Additional file 1: Table S8). Ultimately, by combining the selected results from correlation analysis and machine learning models, a total of 554 diverse metabolites and 1313 different hyperspectral indices were screened (Additional file 1: Table S9 and Table S10).

The R_p values of the 554 screened metabolites also followed a normal distribution, and their average value was 0.53, which was higher than the average R_p value of 0.45 for all 887 measured metabolites (Fig. 2f and Additional file 1: Table S7). Moreover, the average R_p of the 13 metabolite groups was greater than 0.50, with only the amino acid group having an average R_p lower than 0.50. Notably, the average R_p values of the terpene, polyphenol and alkaloid groups were greater than 0.60, with values of 0.64, 0.62 and 0.60, respectively. Most of the 554 metabolites were located in the lipid, vitamin, organic acid, amino acid and flavonoid metabolite groups (Fig. 2g, h; Additional file 2: Fig. S6), and many of these metabolites have medical value or play important roles in rice growth and development. For example, the metabolite epicatechin in the flavonoid group can enhance metabolism [32] and regulate immunity and defence against tumours [33]. The optimal R_p of epicatechin was 0.75 (Fig. 2i), indicating good predictability. The metabolite phosphocholine in the lipid group is an important growth regulator of rice; it can reduce the photorespiration of plants [34], promote crop growth [35], and promote the formation of rice seedling roots [36]. The measured phosphocholine level also showed high predictability, as the optimal R_p was 0.82 (Fig. 2j). Most of the 1313 important hyperspectral indices screened were first-order (dA) or second-order (ddA) derivatives of the average reflectance. For example, the distributions of the epicatechin and choline contents and the corresponding top 5 vital hyperspectral indices screened were all dA- or ddA-related traits (Additional file 2: Fig. S7). Moreover, the level distribution of some vital spectral indices is consistent with the corresponding metabolite level distribution, even in different rice subgroups (Additional file 2: Fig. S8). Ultimately, the 554 metabolites and 1313 hyperspectral indices screened via correlation analysis and machine learning models were used for further GWAS.

GWAS of screened metabolites and important hyperspectral indices

GWAS coupled with metabolites (mGWAS) and important hyperspectral indices (hGWAS) were performed through EMMAX (see the "Methods" section). Overall, a total of 17,509 significant locus–trait associations were revealed within 2882 screened

lead SNPs ($P < 1.3E-6$) for 554 metabolites and 1313 hyperspectral indices. Specifically, 3421 significant loci–metabolite associations were identified for 514 metabolites, with 1715 lead SNPs, whereas 14,088 significant loci–hyperspectral index associations were found for 1181 hyperspectral indices, with 1300 lead SNPs (Additional file 1: Table S11).

Colocalization analysis between selected metabolites and hyperspectral indices

Chromosomal colocalization analysis was performed with 1715 lead SNPs linked to metabolites and 1300 lead SNPs linked to hyperspectral indices. Considering the linkage disequilibrium (LD) decay of rice (Additional file 2: Fig. S9), a colocalization region was defined as being 300 kb in size. To decrease the number of possible false positives, only co-detected loci underlying metabolic traits highly correlated with the spectral traits ($r > 0.3$, $P < 1.3E-6$, Pearson's correlation coefficient) were considered. This led to 6415 locus pairs of metabolites and hyperspectral indices being extracted, among which there were 14,128 significant loci–trait associations with 2152 lead SNPs. This represented 2415 significant locus–metabolite associations of 496 metabolites with 1184 lead SNPs and 11,713 significant loci–hyperspectral index associations of 1122 hyperspectral indices with 1101 lead SNPs (Additional file 1: Table S12).

High-throughput screening of candidate genes regulating hyperspectral and metabolite levels

We next sought to select phenotypically and genetically related matching pairs of metabolites and hyperspectral indices. This involved relating gene expression selection and KEGG keyword mapping with these colocalized lead SNPs. The detailed keyword descriptions concerning each metabolite group are listed in Additional file 1: Table S13. A total of 12,512 locus-trait associations with high gene expression, which could be related to 399 metabolites and 1105 hyperspectral indices within KEGG pathways, were identified (Additional file 1: Table S14). The associations of the hyperspectral indices were more than ninefold greater than those of the metabolites because of the redundancy of the spectral bands. However, the average associations of the screened hyperspectral indices were more than twofold greater than those of the metabolites, indicating their potential applicability in novel candidate gene selection. Among these significant locus–metabolite associations, lipid-related associations ranked first, the number of which was 566. Each lipid-related metabolite had an average of 4.5 significant associations, and their standard deviation was 4.2. The vitamin- and organic acid-related associations ranked second and third, the numbers of which were 364 and 233, respectively. For each metabolite in these two groups, the average numbers were 4.4 and 4.3, and the standard deviations were 5.3 and 7.0, respectively. However, the polyphenol-related associations were the fewest in number with only two. The details of all the statistical information regarding the associations and corresponding loci are available in Additional file 1: Table S15 and Table S16.

For each metabolite group, the matching hyperspectral indices were located mainly within visible-light bands (400–760 nm). The lipid groups had the most matched pairs with hyperspectral indices, whereas the flavonoid groups ranked fifth (Fig. 3a). The 399 screened metabolites had an average R_p of 0.53, with an overall normal R_p distribution. The R_p values of 60.4% of the metabolites were greater than 0.50, which was better than

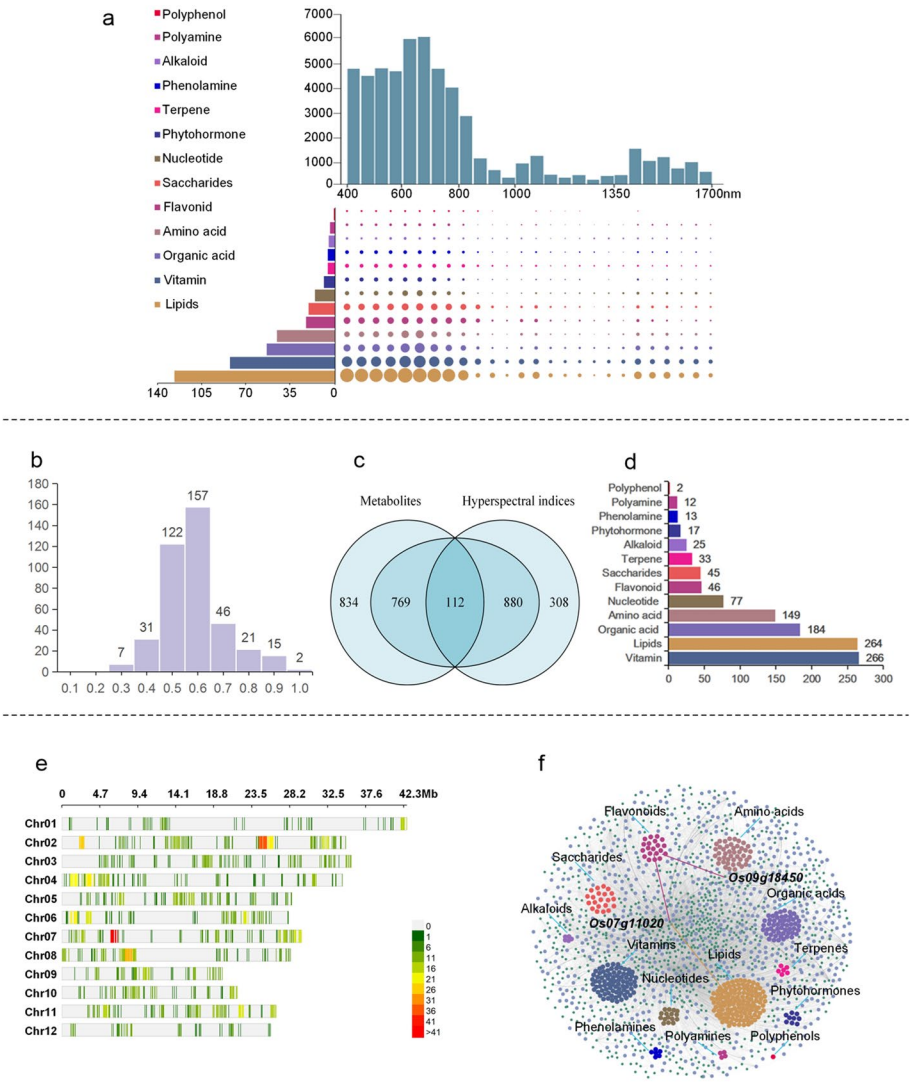

Fig. 3 Information on the screened hyperspectral indices and metabolites that are both phenotypically and genetically related. **a** UpSet plot of the screened metabolites of 13 groups and hyperspectral indices that are phenotypically and genetically related. **b** R_p distribution of the screened metabolites after gene expression selection and KEGG key word mapping. **c** The colocalized loci of metabolites and hyperspectral indices. **d** Distribution of the screened loci among different metabolite groups. **e** The distribution of the colocalized SNPs of metabolites and hyperspectral indices among chromosomes. **f** A relationship network constructed on the basis of the screened metabolites and hyperspectral indices and colocalized loci

those of all 887 metabolites (52.5%) without selection (Fig. 3b). When 1761 loci/candidate genes were screened with the 399 reserved metabolites and 1105 hyperspectral indices, 881 were related to metabolites, 992 were associated with hyperspectral indices, and 112 loci were related to both (Fig. 3c). Among the 1761 loci/candidate genes, most were linked to vitamin-related roles, with those involved in lipid metabolism ranked second (respectively 266 and 264). A total of 46 loci/candidate genes were linked to flavonoid biochemistry, which was the 6th ranked group (Fig. 3d). In terms of chromosomal colocalization, while there was some variation, most metabolite/hyperspectral index-selected genes colocalized to chromosomes 2 and 7 (Fig. 3e).

Finally, a relationship network between the 1105 hyperspectral indices screened and 399 metabolites identified on the basis of colocalized loci was constructed as a resource for polished rice metabolism gene selection (Fig. 3f). The 399 metabolites were clustered and highlighted according to the group information listed in the network, which makes it convenient for selection of candidate locus regulating the levels of metabolites with similar structures and corresponding hyperspectral indices. Another correlation network of the screened metabolites, the hyperspectral indices and the colocalized loci are shown in Additional file 2: Fig. S10. The 1761 candidate locus were highlighted and separated into three groups listed as same candidate locus screened from hGWAS and mGWAS, candidate locus screened from hGWAS only and candidate locus screened from mGWAS only. There were 880 candidate loci screened from hGWAS only, indicating the great potential of using important hyperspectral indices as indicators of metabolites to unravel the genetic mechanisms that regulate metabolite levels in polished rice. All the metabolites and hyperspectral indices were filtered from the original data; they were phenotypically related and may have a genetic relationship, as they were colocalized within 300 kb. Thus, the metabolites and hyperspectral indices in the networks were worthy of attention, and the candidate loci in the networks were valuable genetic resources for subsequent functional gene analysis on regulating the levels of metabolites and corresponding hyperspectral indices.

Evidence of metabolism-hyperspectral linkages

To experimentally validate the direct metabolite–hyperspectral index association, the metabolites with the highest correlation and colocalization with the hyperspectral indices were selected. For example, the correlation network and hierarchical cluster analyses showed a strong correlation between the hyperspectral index ddA42 and the accumulation patterns of multiple flavonoids (including catechins, epicatechins, and their glycosylated (e.g. epicatechin O-hexoside) and polymerized form compounds (e.g. procyanidin B2, procyanidin B3), and a few unmodified flavonols) in the rice population, whereas dA270 showed a high correlation with multiple lipids (mainly diacylglycerophospholipids) at the locus (Additional file 2: Fig. S11A and S11B; Fig. 4h). In addition, a prominent locus, at the 6.08 Mb on chromosome 7 was located by the above metabolites and hyperspectral traits via mGWAS and mQTL analyses (LMM, n = 533). The levels of these metabolites and hyperspectral traits were significantly associated with lead single nucleotide polymorphism (SNP) sf0706085999 (Fig. 4a–d). These results imply that there may be genetic factor(s) at this locus that regulates multiple metabolites and hyperspectral indices simultaneously. In searching for candidate genes, we noted that *Rc* (*LOC_Os07g11020*) in this region has previously been reported to affect seed coat colour, and the functional variant SNPs of *Rc* were strongly associated with the levels of hyperspectral index ddA42 and dA270 and the above metabolites (Fig. 4e–g). Furthermore, we conducted haplotype analysis based on four SNPs and two InDels (Insert or delete fragments) in the gene coding region and promoter region. The one SNP and one InDel in the fifth exon of the *LOC_Os07g11020* gene result in early stopping of the gene (SNP4) and a frameshift variant (InDel2), respectively. And haplotypes with disrupted gene function of *LOC_Os07g11020* (type3, type5 and type6) contain significantly lower

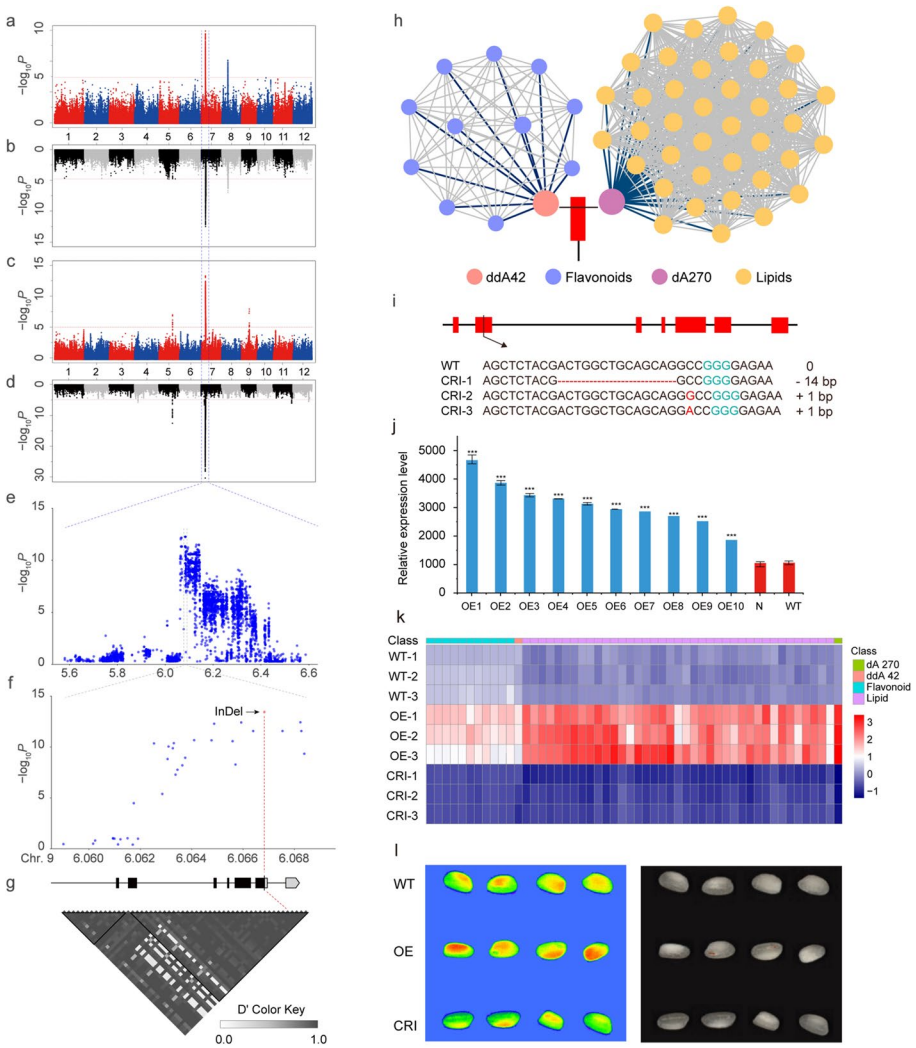

Fig. 4 Role of the candidate gene *Rc* in the regulation of hyperspectral traits and the accumulation of lipids and flavonoids. **a**–**b** Manhattan plot displaying the GWAS results for the level of the hyperspectral trait dA270 and the metabolite trait PC 18:0/18:0. **c**–**d** Manhattan plot displaying the GWAS results for the level of the hyperspectral trait ddA42 and the metabolite trait epicatechin. **e** Manhattan plot of the chromosomal 7 region 5.6 ~ 6.6 Mb, where SNPs were significantly associated with the hyperspectral traits dA270 and ddA42 and the metabolite traits PC 18:0/18:0 and epicatechin. Overview (**f**) and closer view (**g**) of the associated SNPs and InDels in the region Chr7 6.05989–60.68818 Mb where the *Rc* locus was located. **h** Triple relationships centred on sf0706085999. A total of 40 metabolite traits (11 flavonoids and 39 lipids) and 2 hyperspectral traits (dA270 and ddA42) were identified via mGWAS and hGWAS. **i** Generation of *Rc* mutations via CRISPR/Cas9. The sequences of the *Rc* mutant CRI-1/2/3 are shown. Protospacer-adjacent motifs (PAMs) are indicated in green. Deletions and insertions are indicated by dashes and in red, respectively. **j** Bar plots showing the relative expression levels of *Rc* in the wild-type (WT), ten *Rc*-OE (OE1–10) and one negative (N) transgenic line. **k** The levels of 40 metabolites and 2 hyperspectral traits in polished rice powder from the WT, CRI and OE lines. For metabolomic and hyperspectral data, the data per row are *Z*-score standardized. **l** Inversion results of the procyanidin content distributed in polished rice grains. The vital coefficient selected from the PLSR model of the corresponding rice power data was used to generate a pseudocolour image

levels of hyperspectral ddA42, dA270, flavonoids, and lipids than normal functioning haplotypes (type1, type2, type4) (Additional file 2: Fig. S12A), suggesting that *Rc* may affect these traits.

To confirm the regulatory function of *Rc* and hyperspectrum is specifically associated with metabolites, we constructed *Rc*-CRISPR (CRI) and overexpression (OE) lines (Fig. 4i–j). The transgenic progeny (T2 generation) overexpressing *Rc* exhibited increased levels of catechins, epicatechins, and their glycosylated (e.g. epicatechin O-hexoside) and polymerized form compounds (e.g. procyanidin B2, procyanidin B3), and multiple diacylglycerophospholipids. In contrast, T2 CRI lines presented the opposite phenotype (Fig. 4k). This result shows that ddA42 has the potential to specifically refer to unmodified flavonols, catechins and catechin derivatives in the flavonoid family, whereas dA270 can specifically refer to diacylglycerophospholipids in the lipid family.

To further confirm the role(s) of *Rc* in regulating flavonoids, we first determined the nuclear localization of *Rc* (Additional file 2: Fig. S13) and performed RNA-seq analysis of seeds at the filling stage for the wild-type ZH11, OE and CRI lines. A total of 729 differentially expressed genes (DEGs) were detected (|log2fold change| > 1 and false discovery rate $P < 0.01$), of which 84.6% (617 genes) were upregulated in the OE lines compared with the wild-type ZH11 plants. GO and KEGG analyses of the upregulated genes in the OE lines revealed that the terms flavonoid biosynthesis, phenylpropanoid biosynthesis and fatty acid biosynthesis process were enriched (Additional file 2: Fig. S14A and Fig. S14C). Overall, 309 DEGs were detected (|log2-fold change| > 1, $P < 0.01$ for false discovery rate), of which 78.6% (242 genes) were downregulated in the CRI lines compared with the wild-type lines. Similarly, the downregulated genes in the CRI lines were enriched in flavonoid biosynthesis, anthocyanin-containing compound biosynthesis and fatty acid biosynthesis by GO and KEGG analyses (Additional file 2: Fig. S14B and Fig. S14D). These results suggest that *Rc* may control the accumulation of flavonoids and lipids in rice grains by regulating the expression of genes in these metabolite biosynthesis pathways. In addition, the distribution of the flavonoid procyanidin in polished rice was inverted pixel by pixel according to the vital coefficients acquired from the PLSR model constructed from the rice powder data. In Fig. 4L, the procyanidin content is indicated by the pseudocolor shade. Compared to the WT lines, the OE lines presented greater procyanidin contents; however, in the CRI lines, lower procyanidin levels were observed. This result indicates that hyperspectroscopy can accurately reflect small molecule metabolite content.

Dissecting novel metabolite loci via hGWAS

Our research has revealed that some significant SNPs could be screened by hGWAS only. To expand the possible applicability of our approach, we tested whether a genetic analysis of the hyperspectral data could facilitate the dissection of novel candidate locus regulating metabolite traits. In our hGWAS of ddA42 and mGWAS of flavonoids, the hGWAS revealed a new significant localization site on chromosome 9 compared with the mGWAS (Fig. 4c and d). For instance, epicatechin and ddA42 were related phenotypically and genetically, as their Pearson correlation coefficient was 0.45 and they shared a colocalized SNP sf0706085999. Whereas, an extra SNP on chromosome 9 was screened by the hyperspectral index ddA42 only. A series of significantly linked SNPs were located within the promoter and coding regions of a single gene (*LOC_Os09g18450*) (Fig. 5a–c). The level of ddA42 was significantly associated with the lead SNP sf0911239171 ($P = 2.62 \times 10^{-9}$) (Fig. 5d). *LOC_Os09g18450* is predicted to encode a flavonol synthase

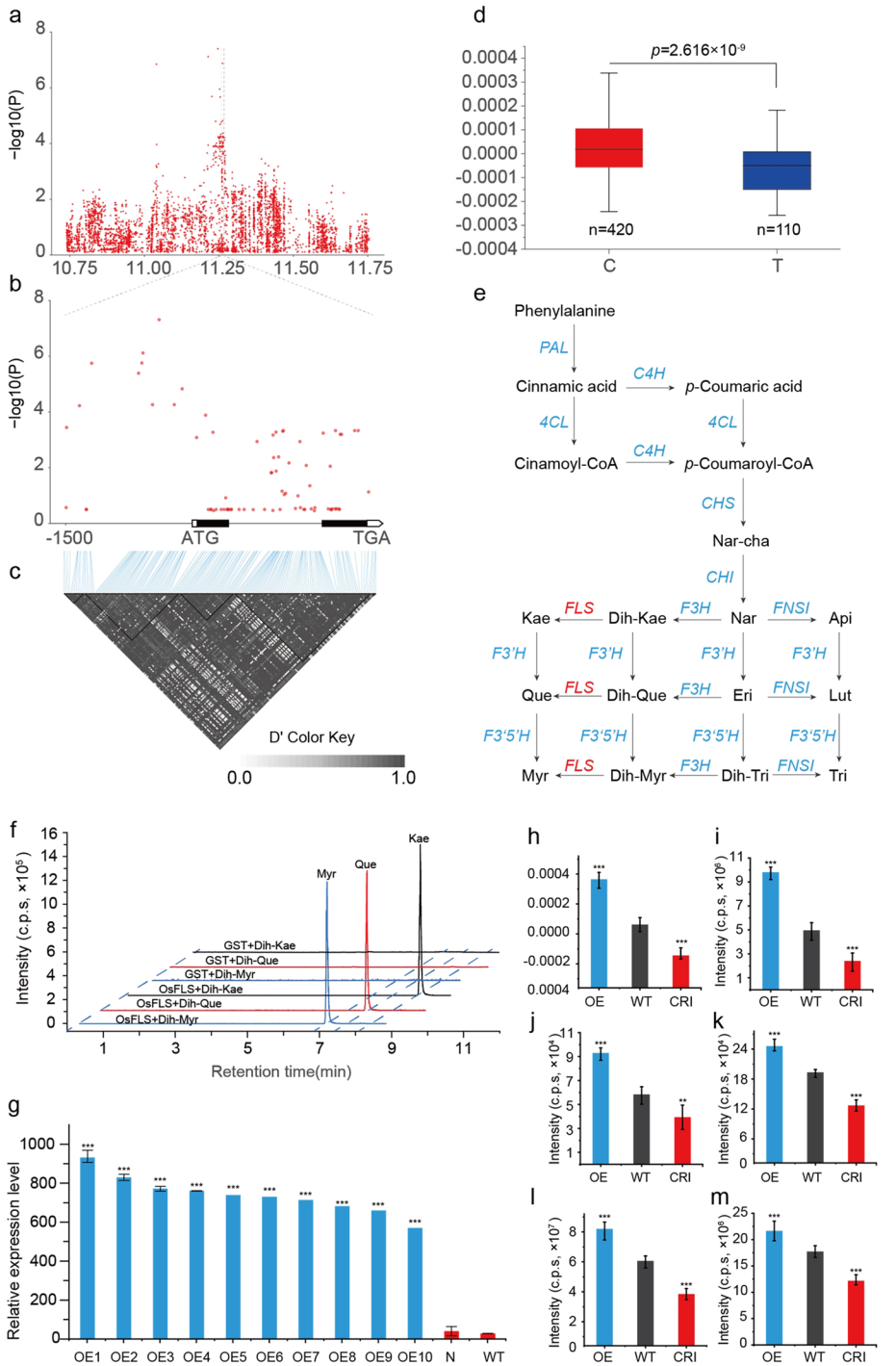

Fig. 5 Novel genetic loci regulating flavonoid accumulation in rice as determined via hGWAS. **a** Regional Manhattan plot of the genomic region on chromosome 9 (10.75 ~ 11.75 Mb). Distribution of SNPs (**b**) in the gene model *Rc* (**c**) and their LD to each other. **d** Plants with the C allele of chr9.sf0911239171 presented significantly higher levels of the hyperspectral trait ddA42 than did the T allele. **e** The biosynthetic pathway of flavonoids. *OsFLS* enzymes are indicated in the pathway. **f** LC–MS chromatograms of in vitro enzyme assays showing the enzyme activity of recombinant *OsFLS*. Protein extract from *E. coli* containing the pDEST17 (GST flag) empty vector was used as a negative control. **g** Bar plots showing the relative expression levels of *OsFLS* in the wild type (WT, Zhonghua11, ZH11), ten *Rc*-OE (OE1–10) and one negative (N) transgenic line. The levels of ddA42 (**h**), kaempferol (**i**), quercetin (**j**), myricetin (**k**), epicatechin (**l**), and catechin (**m**) in polished rice powder from the WT, CRI and OE lines

protein (*OsFLS*) with a molecular mass of 37 kDa, which may be involved in the conversion of dihydroflavonol to flavonol (Fig. 5e). In addition, we conducted haplotype analysis based on eight SNPs in the gene coding region and promoter region. The results showed that SNP5 (CG to C) in the third exon of the *LOC_Os09g18450* resulted in a frameshift variant that disrupted the original biochemical function of the gene. As expected, the levels of flavonoids and hyperspectral ddA42 were significantly lower in haplotypes with loss of gene function (type1, type2, type3, type5) than in haplotypes with normal function (type4) (Additional file 2: Fig. S12B). This result suggests that *LOC_Os09g18450* may function to regulate hyperspectral ddA42 and flavonoid levels.

To characterize the function of this putative gene, its open reading frame was cloned and inserted into expression vectors with glutathione S-transferase (GST) fused to the N-terminus. Then, the recombinant *OsFLS* that was successfully expressed in *E. coli* was purified for enzymatic assays. The enzymatic activity of the recombinant protein was examined by incubating *OsFLS* with Dih-kae (dihydro-kaempferol), Dih-Que (dihydroquercetin), and Dih-Myr (dihydro-myricetin) as substrates. All of these flavanols were accepted by the enzyme, and the reaction products were identified as the flavonols quercetin, kaempferol, and myricetin, respectively, by LC–MS (Fig. 5f). To confirm that *OsFLS* can be linked to ddA42 and flavonoid content, ten transgenic OE lines were obtained from the ZH11 background. The successful construction of OE plants was confirmed by measuring the relative expression of *OsFLS* in T1 seeds at the grain-filling stage (Fig. 5g). We selected the lines with the greatest upregulation of expression levels for metabolomics and hyperspectral detection and found a significant increase in ddA42 and flavonoid contents compared with those in wild-type plants. In contrast, *OsFLS* knockout mutants (CRIs) constructed via CRISPR-Cas9 technology presented a lower ddA42 and flavonoid content than the wild type (Fig. 5h-m). These results confirm that FLS positively regulates the levels of flavonoids and ddA42 in rice grains and that hGWAS can identify 'hidden genetic loci' that are not located by mGWAS. Prior research [30] has demonstrated that 'i-traits' like hyperspectral indices had the capability to identify 'novel' SNPs and candidate loci in GWAS analyses of complex quantitative traits, like the fresh/dry weight and green leaf area of rice. This is because complex quantitative traits can be predicted through the combination of several simpler 'i-traits', and these simpler traits, which are phenotypically and genetically linked to the complex quantitative traits, may offer greater insight into studying them. Moreover, hyperspectral cameras can acquire signals of light reflection in polished rice caused by both atomic electron and molecular vibrational transitions. Metabolites with similar functional groups and structures have similar reflection spectrum, for particular metabolite, the signal may be not strong enough to be identified through GWAS. Whereas the screened hyperspectral indices may reflect the accumulated signals of metabolites with similar functional groups and structures, the signals may be stronger than particular metabolites so as to many novel significant SNPs are screened by hGWAS only.

Discussion

Hyperspectral indices could be indicators of metabolites for GWAS analysis

Metabolomics has become an important tool for screening materials for advanced crop breeding [15, 37–39]. However, the process of quantifying plant metabolites is often

cumbersome and destructive. It also requires significant financial investment. In this study, we have described a pipeline for screening matching pairs of metabolites and important hyperspectral indices that are phenotypically and genetically associated. Furthermore, we showed that the screened important hyperspectral indices could serve as indicators of metabolites for GWAS through colocalization analysis, gene expression pattern selection and the KEGG keyword mapping. GWAS of important hyperspectral indices and corresponding metabolites revealed that candidate genes associated with metabolites could be deciphered through hyperspectral indices, even for metabolites that were not directly measured.

By integrating colocalization analysis, feature screening and modelling results, we identified the characteristic wavelength ranges for each class of metabolites (Additional file 2: Fig. S15). The distribution of the screened hyperspectral indices was similar to that in the visible band range (400–760 nm), while the details of the distribution differed among the 13 metabolite groups (Additional file 2: Fig. S15). For example, characteristic wavelengths for flavonoids were distributed mainly within the ranges 400–900 nm, 938–1138 nm, 1190–1352 nm and 1410–1520 nm. In other studies, absorption peaks were observed at 400–600 nm [40], 700–760 nm [41], 870–900 nm [42], 1100–1140 nm [43], 1150–1400 nm [44] and 1400–1500 nm [45]. These results were highly consistent with the wavelengths associated with the flavonoids screened in this study. Notably, as suggested by other published works, our study excluded the wavelength range of 760–870 nm. However, the associations between these bands and flavonoids need to be further explored. For lipids, the characteristic bands were located primarily within the ranges of 400–850 nm, 938–1138 nm and 1410–1645 nm. Previous research also supported our targeted hyperspectral indices as indicators of corresponding metabolites. For example, a lack of reflectance at 520 nm was associated with anthocyanin accumulation in grape leaves, and reflectance near 520 nm was significantly correlated with carotenoid accumulation in plant leaves. This finding was consistent with our finding identifying dA65 (522 nm) as a vital feature of anthocyanins in polished rice. Furthermore, we observed that wavelengths of 600–650 nm were important for predicting specific lipid contents in plants. Similarly, the phosphatidylcholine content of grape leaves can be measured within the wavelength range of 600–648 nm [46], whereas the screened vital hyperspectral index dA110 (610 nm) was located within this range. This demonstrated the reliability of using vital hyperspectral indices as a replacement for metabolites for GWAS and further candidate gene selection.

To further validate the metabolite prediction ability based on hyperspectral data to target genes, fivefold cross-validation was conducted on each metabolite with the highest Rp of all eight models. The predicted value for each metabolite gained from fivefold cross-validation was reserved for GWAS, and the significant SNPs acquired from the predicted values of metabolites overlapped greatly with the results acquired from real metabolite measurements, especially in the hotspot area of chromosome 7 (Additional file 2: Fig. S16). For example, the true and predictive values of epicatechin and the corresponding hyperspectral index ddA42 both colocalized significant SNPs within *LOC_Os07g11020* (Additional file 2: Fig. S17). The detailed *R* values of the validation set (*Rv* values) are listed in Additional file 1: Table S17.

Experimental validation confirms the reliability of the screened hyperspectral indices as metabolite surrogates

To further confirm the reliability of using vital hyperspectral indices as a replacement for metabolites for GWAS analysis and further candidate gene selection, the pure solutions of epicatechin and procyanidin B1 were measured via a spectrophotometer (the details are provided in Additional file 2: Note S1) to acquire the spectral absorbance curve shown in Additional file 2: Fig. S18 and Fig. S19 and Additional file 1: Table S18, both of which presented a particular absorption valley at approximately 480 nm, which is in accordance with the screened hyperspectral index ddA42 (478 nm). This result further reflects the reliability of the newly screened hyperspectral indices, with ddA42 serving as an indicator of flavone content.

Mapping and identifying loci or genes underlying different metabolite contents is an important additional tool in existing genomics-assisted strategies for crop improvement. In this study, a total of 1761 unique candidate genes were identified and annotated with 769 metabolite-screened genes, and 880 hyperspectral indices were used to select genes. KEGG analysis revealed that, in addition to lipid and flavonoid metabolism, most genes are involved in the synthesis and metabolic pathways of plant primary and secondary metabolites. Many of these metabolites play important roles in the interaction between plants and their environment, and the candidates assigned in our study may provide new resources for further functional validation. For example, lignin is an important polymer in plant resistance to pests and diseases. Although phenylalanine ammonia-lyase (PAL) genes linked to lignin production have been reported, few of these genes have been genetically mapped in rice. Here, we used colocalization analysis to map multiple PAL genes and found that the *LOC_Os02g41650* genes were significantly associated with lignin. This gene has been previously reported as *OsPAL3* and is involved in lignin synthesis, thereby enhancing disease resistance in rice. Moreover, numerous mGWAS-targeted metabolites, such as amino acids, vitamins, lipids and polyphenols, determine the nutritional quality of rice and can be mapped to possible assigned candidate genes (Additional file 1: Tabel S15). In addition, several genetic loci that regulate the accumulation of nutrients associated with the maintenance of health in humans have been identified via mGWAS and hGWAS. The thiamine (vitamin B1) level is correlated with the hyperspectral dA212 in rice. The dA212 colocalized with thiamine according to both the mGWAS and hGWAS at 1.76 Mb on chromosome 6, and the *Wx* gene at this locus was shown to regulate thiamine accumulation in rice grains.

Taken together, the extensive data generated in this study represent a valuable resource for further studies on the biosynthetic pathways and regulatory circuits of plant metabolites. Moreover, our approach can help overcome the limitations of single-data-type approaches. By combining multiple datasets, it is possible to compensate for missing or unreliable information within any single data type. We have demonstrated the power of our approach by using this multidimensional approach to encompass genomic, hyperspectromic, and metabolomic data to identify important genes contributing to metabolic pathways such as lipid and flavonoid metabolism. In terms of practicality, these examples illustrate how hyperspectroscopic groups will greatly aid in improving metabolomic research for a wide range of natural genetic variations.

hGWAS has the potential to explore genes involved in broad-spectrum regulation of metabolites

The process of light propagation within plants is intricate, involving the scattering and absorption of photons as they interact with plant tissues. These interactions lead to various optical phenomena, including the reflection, transmission, and absorption of light by the plant's cellular structures. To quantify the degree of light attenuation in plant materials, optical sensors like hyperspectral cameras can be employed to measure the amount of light that is reflected from or transmitted through the plant tissues caused by both atomic electron and molecular vibrational transitions [47]. Physically, the compressing and stretching process of covalently bonded atoms in heteronuclear molecules will result in 'overtones' of the fundamental vibration frequency when light interacts with plants. Overtones, akin to harmonics, consist of a series of frequencies that are multiples of the fundamental vibration's frequency. However, 'combinations' arise when two or more fundamental frequencies converge to possess the same energy level in the spectrum. It is normal that a diverse array of combinations can emerge within any specific molecule, leading to spectra that typically exhibit broad and frequently overlapping peaks. Thus, high-resolution spectroscopy is indispensable for identifying and distinguishing all the diverse structural components. For simple molecules like water, it is easy to be quantified through spectra as typical absorption peaks can be observed around 1450 nm and 1950 nm. Whereas, for most molecules in the plant kingdom, their characteristic spectrum peaks could be broad, weak and overlapping. To address this issue, chemometrics is usually used for the analysis of complex spectral data [48].

Plants harbour an extraordinarily diverse array of metabolites, with estimates suggesting that their numbers could span from 200,000 to over 1 million. A significant proportion of these metabolites exhibit structural similarities; for instance, all flavonoids feature three analogous organic rings, while fatty acids universally display elongated carbon chain structures. These shared structural characteristics can lead to spectral bands that are close to each other when analysed using hyperspectral imaging. To enhance the precision of hyperspectral bands for the detection of specific metabolites, a hyperspectral imaging system with a spectral resolution of approximately 2 nm was applied in the research. Furthermore, we constructed a pipeline to screen metabolites and hyperspectral indices that were phenotypically and genetically related. Based on the result, correlation networks between the screened hyperspectral indices acquired by hyperspectral cameras and metabolites collected through mass spectrometry were constructed, aiming to pinpoint specific hyperspectral bands indicative of key metabolites and their closely related compounds. In addition to constructing correlation networks to identify hyperspectral indicators for metabolites, we also sought to broaden the application scope of this methodology to explore if the screened hyperspectral indices could be effective in GWAS analysis. Intriguingly, we observed that metabolites and their corresponding hyperspectral indicators share the same statistically significant SNP loci. This finding underscores the potential of hyperspectral imaging not only as an alternative to mass spectrometry in detecting metabolite content but also as a potent, cost-effective tool for exploring metabolite regulatory genes and loci.

Upon further analysis of hGWAS and mGWAS results, we observed that hGWAS tended to identify more association sites than mGWAS. Notably, hGWAS enabled the

successful identification of crucial genes regulating the specific metabolite levels. Taking into account the vast disparity between the extensive number of plant metabolites (ranging from 200,000 to 1 million) and the relatively limited number of hyperspectral bands (1944), we hypothesize that certain hyperspectral signatures may possess the capability to concurrently signify particular metabolite and its closely related substances. We further speculate that the new loci identified in hGWAS might be linked to the overall content of metabolites and their high analogues. To validate this finding, we aggregated the content of 11 flavonoids associated with ddA42 as presented in Figs. 4 and 5, and conducted mGWAS based on the total content. The lead SNP (chr9. sf0911239171) locus on chromosome 9 was successfully pinpointed in the newly obtained mGWAS results (Additional file 2: Fig. S20). Moreover, the elevated levels of flavonoids such as epicatechin, kaempferol, and myricetin in *OsFLS* overexpression material indicate that this gene has the capacity to regulate multiple aforementioned flavonoids simultaneously (Fig. 5i–m). These findings not only confirm why hGWAS is capable of uncovering novel metabolite regulatory locus but also offer novel insights into the mining of broad-spectrum metabolite regulatory genes.

Conclusions

Our research underscores the potential of combining hyperspectral imaging with machine learning methods to rapidly and cost-effectively quantify crop metabolites. We proposed a pipeline to screen crucial hyperspectral indices that exhibit both phenotypic and genetic correlations with specific metabolites. Additionally, our findings suggest that these selected hyperspectral indices can serve as indicators for corresponding metabolites in GWAS for candidate gene identification. Notably, hyperspectral indicators have the potential to uniquely uncover novel candidate loci, as they can concurrently signify metabolites with similar structures. The correlation networks established in our research, which connect hyperspectral indices with metabolites, encompass a wealth of genetic resources that will facilitate further candidate gene selection in polished rice. The methodologies introduced in this study exhibit promising potential for application in a variety of crops beyond rice, ultimately facilitating the exploration of metabolic regulation mechanisms in crops and accelerating the breeding process.

Methods

Plant materials

A diverse global collection of 533 *O. sativa* accessions, including both landraces and elite varieties, was collected [49]. Rice plants examined under field conditions were grown during the normal rice-growing season at the Experimental Station of Huazhong Agricultural University (Wuhan, China). All the seeds were planted in a seedbed in mid-May and transplanted to the field in mid-June. The plants withing a row were 16.5 cm apart, and the rows were 26 cm apart. Mature seeds of each accession were randomly collected and pooled for metabolic profiling.

Hyperspectral data acquisition for polished rice

An HSI system was built to collect hyperspectral data from 533 polished rice samples. This system contained translation stages, halogen lamps and two hyperspectral cameras

covering the spectral range of 400–1700 nm. Twenty polished rice grains of each accession were ground to powder and then poured into 8-mm-wide plastic lids. To allow simultaneous data acquisition, a dozen samples were placed on black metal trays. Dark current and gain calibrations were necessary to correct the dark current of the charge-coupled devices (CCDs) in the cameras and eliminate the influence of uneven light among the wavelengths. The metal trays moved with translation stages under the control of the programmable logic controller until all the samples were captured by fixed cameras. The corresponding acquired data were stored on two computers synchronously. The whole data collection process was carried out in a dark room with only halogen lamp lighting to provide an accurate light source [23]. Finally, for each accession, the spectral information of 486 bands within 400–1700 nm was acquired from the HSI system.

Metabolite measurement for polished rice

For metabolome analysis, samples were analysed via HPLC. The HPLC analytical conditions were as follows: column, Shim-pack GISS C18 (pore size 1.9 μm, length 2.1 × 100 mm); solvent system, water (0.04% acetic acid): acetonitrile (0.04% acetic acid); flow rate, 0.4 mL/min; temperature, 40 °C; and injection volume, 2 μL. The gradient program was as follows: 0 min, 5% B; 12.0 min, 95% B; 13.2 min, 95% B; 13.3 min, 5% B; 15.0 min, 5% B. The targeted metabolic profiling analysis was performed via scheduled multiple reaction monitoring (MRM) via an LC-ESI-QQQ-MS/MS system (LCMS-8060, SHIMADZU, Japan). The ESI source operation parameters were as follows: nebulizing gas flow, 3 L min^{-1}; heating gas flow, 10 L min^{-1}; interface temperature, 500 °C; DL temperature, 250 °C; heat block temperature, 400 °C; and drying gas flow, 10 L min^{-1}. The data recorded were processed with LabSolutions 5.91 software. A total of 837 metabolites were detected via this method. These metabolites belonged to a total of 13 classes, namely, lipids, vitamins, amino acids and their derivatives, sugars, flavonoids, alkaloids, terpenoids, nucleic acids and their derivatives, organic acids, hormones, polyamines, polyphenols and phenolamines (Additional file 1: Table S4).

Data processing and hyperspectral index extraction

An image analysis pipeline was compiled for image processing and hyperspectral index extraction via LabVIEW and C++programming [31]. The acquired data were normalized via dark current and whiteboard data to obtain the rectified binary data stream. Image segmentation and masking and data extraction processes were carried out to obtain equivalent hyperspectral indices among the wavelengths. The details of these steps were described in a previous study [31]. A total of 1944 hyperspectral indices were ultimately acquired, which could be divided into four types: average reflectance (A), first derivative of average reflectance (dA), second derivative of average reflectance (ddA), and the logarithm of average reflectance (lgA). The acquisition range of the two hyperspectral cameras overlapped within 900–1000 nm, resulting in some redundancy in the hyperspectral indices. The redundancy was removed, and the outliers of each feature were eliminated according to the 3σ criteria, leaving a total of 1848 hyperspectral indices.

Hyperspectral feature screening and prediction model establishment

Pearson correlation coefficients of 887 metabolites and 1848 hyperspectral indices were computed in R 4.1.2, and those with values greater than 0.30 were considered highly related matching pairs. Meanwhile, eight models, including the SLR, PLSR, RF, RR, LASSO, SVM, LGBM and CNN models, were established for predicting metabolite content through hyperspectral indices (Additional file 2: Fig. S21). Automatic parameterization was used in the models to reduce artificial error. The training and test sets were derived by dividing the data into a ratio of 1:4. Regression models were built in Python 3.6 via the sklearn machine learning package, statsmodels library, etc. (Additional file 3: Note S2). The details were as follows: (1) for the SLR, the AIC was adopted as the evaluation index, and the difference between the two adjacent $R_p{}^2$ values (the R^2 values of the test set) was no less than 0.01. (2) PLSR was performed with the PLSR regression function, and the scale parameter set was True. (3) RF, RR, LASSO and SVM regressions were built with gridSearchCV and tenfold cross-validation. (4) The CNN model was implemented via a sequential function and adjusted by observing the loss image to prevent overfitting, and the number of epochs was thirty. Finally, important features and metabolites of those models with R values greater than 0.50 were selected and combined with the above highly related matching pairs for subsequent GWAS. (5) The LGBM was developed via the Python library of the same name, automatic tuning functionalities were incorporated to ensure optimal regression fitting for each metabolite.

Genome-wide association analysis

GWAS was performed to test the statistical associations between genotype and phenotype via efficient mixed model association expedited (EMMAX, v20120210). The genotype data (SNPs) were obtained from the RiceVarMap website (http://ricevarmap.ncpgr.cn), and a total of 4,300,150 high-quality SNPs with a minor allele frequency (MAF) > 0.05 and deletion rate < 0.1 in the 533 rice cultivars were selected for hGWAS and mGWAS. The Emmax-kin program was used to construct the kinship (K) matrix. The Manhattan plot was drawn through the R package (qqman) and in-house R scripts. The calculated genome-wide suggestive threshold, which was based on the original Bonferroni calculation of 1/Me, was $P = 1.3 \times 10^{-6}$ for the whole population.

Candidate gene screening and identification

The colocalization analysis was conducted based on the screened lead SNPs of hGWAS and mGWAS, and those loci within the 300 Kb in the chromosome served as colocalized loci. After obtaining the colocalized locus of selected metabolites and hyperspectral indices, the gene information upstream and downstream of the site is screened based on the degree of LD linkage disequilibrium. According to the rice gene expression profile data downloaded online (https://ricexpro.dna.affrc.go.jp), those genes with expression levels higher than 200 RPKM (reads per kilobase of transcript per million reads mapped) at the heading stage of rice panicles or 7–21 days after pollination of rice endosperm were selected for further gene function annotation and KEGG pathway analysis using eggNOG-mapper software. The selected loci were divided into 14 groups, and several metabolite-related keywords were attached to each group. For example, the keywords

'phenylpropanoid', 'flavonoid', 'stilbenoid', 'caffeine', 'flavone', 'flavonol', 'anthocyanin', and 'isoflavonoid' were linked to the flavonoid-related loci, whereas the keywords 'folate', 'nicotinate', 'nicotinamide', 'thiamine', 'ascorbate', 'vitamin', 'pantothenate', 'riboflavin', 'retinol', 'biotin', and 'novobiocin' were used as labels for vitamin-related loci. Then, the locus of each group was mapped to the enrichment pathway information for rice. These mapping processes were automatically performed via scripts compiled with Python 3.6 (Additional file 3: Note S2).

Metabolite and hyperspectral index colocalization network construction

Relationship networks between screened metabolites and hyperspectral indices were constructed on the basis of the location of significant SNPs screened separately via the hGWAS and mGWAS approaches, and the Pearson correlation coefficient between them was examined. Considering the LD decay of rice, the matching pairs of metabolites and hyperspectral indices that had significant SNPs within 300 kb in chromosomes and a Pearson correlation coefficient over 0.30 were connected in the constructed networks. The networks were constructed through the software Gephi 0.10.

Construction of transgenic lines

The *LOC_Os07g11020*, *LOC_Os09g18450* and *LOC_Os06g04200* overexpression constructs were generated by directionally inserting the full complementary DNA (cDNA) from Nipponbare, first, into the entry vector DONR207 and then into the destination vector pJC034 with the maize ubiquitin promoter via the Gateway recombination reaction (Invitrogen). The construct vector was subsequently introduced into Agrobacterium strain EHA105, which was subsequently transformed into ZH11 (Zhonghua11). The mutant constructs for the above three genes were generated via CRISPR-Cas9 technology.

Quantitative RT-PCR (RT-qPCR) analysis

Total RNA was extracted via TRIzol reagent (Life Technologies) and reverse transcribed into cDNA according to the manufacturer's instructions for ReverTra Ace qPCR RT Master Mix with gDNA Remover (Toyobo). RT-qPCR analysis was performed using SYBR Green Real-time PCR Master Mix (Toyobo). Rice UBQ5 was used as an internal standard to normalize the transcription of the examined genes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03513-w.

Additional file 1. A master file for Table S1-18.

Additional file 2. Note S1. Figure S1-S21.

Additional file 3. Note S2. Data processing process of this research.

Acknowledgements

We thank doctors Wei Chen, Lingfeng Duan, Chenglong Huang (Huazhong Agricultural University) for discussion of the results and suggestions on data analyses, Luis A.J. Mur (Aberystwyth University) for editing and polishing the grammar of the article.

Authors' contributions

W.Y., J.L. and L.Y. designed the research, H.F. conceived the project and supervised the study. Y.L., G.D. and Z.Y. performed the experiments. J.S., B.L., Y.G and Y.C. also performed experiments or analyzed the data. H.F., Y.L. and G.D. analyzed the

date and wrote the manuscript. J.L. and Y.L. provided the materials and support the data. W.Y., LAJM and J.L. provided constructive suggestions and revised the manuscript. All authors read and approved the final manuscript.

Peer view information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Funding
This work was supported by the grants from the Biological Breeding-National Science and Technology Major Project (STI2030-Major Projects), the National key research and development program (2022YFD1900701-4), General Program of National Natural Science Foundation of China (32470432), National Natural Science Foundation of China (U21A20205), Key projects of Natural Science Foundation of Hubei Province (2021CFA059), Major Science and Technology Project of Hubei Province (2021AFB002), Fundamental Research Funds for the Central Universities (2662024SZ001, 2021ZKPY006, 2662022JC006), HZAU-AGIS Cooperation Fund (SZYJY2022014), 111 Project Crop genomics and Molecular Breeding (B20051), Hubei Provincial Department of Education Science and Technology Plant Project (2023DJC153), and Wuhan Science and Technology Plan Project (2023020402010780).

Data availability
The hyperspectral images of 533 O. sativa accessions can be viewed and downloaded via http://plantphenomics.hzau.edu.cn/usercrop/Rice/image/2024-HSI by the following steps: (i) select 'Rice (Image)'; (ii) select '2024-HSI'; (iii) select one of the 'Bands' and 'Accession ID' and then press 'Search images'; (iv) twenty-one images from 400–1000 nm and eighteen images from 1000–1700 nm of the O. sativa accessions are included under 'Accession ID'. Each image contains the name of the material and the corresponding wavelength. For example, the image 'C001-C014-400.jpg' represents a grayscale image collected at 400 nm, featuring 14 materials from C001 to C014 arranged from top to bottom. These images can also be acquired via https://doi.org/10.6084/m9.figshare.28426328.v1 [50]. The scripts about correlation analysis, machine learning regression, co-localization analysis, gene expression analysis and KEGG analysis could be downloaded via the link https://github.com/dittlespark/Rice_HSI_MT_Programme [51]. and https://doi.org/10.5281/zenodo.14880312 [52]. All the figures and supplementary tables could be downed via the link https://doi.org/https://doi.org/10.6084/m9.figshare.28091447 [53]. All other reasonable requests for data and research materials can be accommodated by reaching out to the corresponding authors.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 May 2024 Accepted: 24 February 2025
Published online: 12 March 2025

References

1. Seck PA, Diagne A, Mohanty S, Wopereis MC. Crops that feed the world 7: Rice. Food security. 2012;4:7–24.
2. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557:43–9.
3. Sulpice R, McKeown PC. Moving toward a comprehensive map of central plant metabolism. Annu Rev Plant Biol. 2015;66:187–210.
4. Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, Li Y, Liu X, Zhang H, Dong H, et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet. 2014;46:714–21.
5. Fernie AR, Tohge T. The Genetics of Plant Metabolism. Annu Rev Genet. 2017;51:287–310.
6. Carreno-Quintero N, Acharjee A, Maliepaard C, Bachem CW, Mumm R, Bouwmeester H, Visser RG, Keurentjes JJ. Untargeted metabolic quantitative trait loci analyses reveal a relationship between primary metabolism and potato tuber quality. Plant Physiol. 2012;158:1306–18.
7. Zaynab M, Fatima M, Abbas S, Sharif Y, Umair M, Zafar MH, Bahadar K. Role of secondary metabolites in plant defense against pathogens. Microb Pathog. 2018;124:198–202.
8. Yu X, Yang T, Qi Q, Du Y, Shi J, Liu X, Liu Y, Zhang H, Zhang Z, Yan N. Comparison of the contents of phenolic compounds including flavonoids and antioxidant activity of rice (Oryza sativa) and Chinese wild rice (Zizania latifolia). Food Chem. 2021;344:128600.
9. Xia D, Zhou H, Wang Y, Li P, Fu P, Wu B, He Y. How rice organs are colored: The genetic basis of anthocyanin biosynthesis in rice. The Crop Journal. 2021;9:598–608.
10. Singh PK, Rawal HC, Panda AK, Roy J, Mondal TK, Sharma TR. Pan-genomic, transcriptomic, and miRNA analyses to decipher genetic diversity and anthocyanin pathway genes among the traditional rice landraces. Genomics. 2022;114:110436.
11. Joensuu M, Wallis TP, Saber SH, Meunier FA. Phospholipases in neuronal function: A role in learning and memory? J Neurochem. 2020;153:300–33.
12. Saito K, Matsuda F. Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol. 2010;61:463–89.

13. Jantscheff P, Schlesinger M, Fritzsche J, Taylor LA, Graeser R, Kirfel G, Furst DO, Massing U, Bendas G. Lysophosphatidylcholine pretreatment reduces VLA-4 and P-Selectin-mediated b16.f10 melanoma cell adhesion in vitro and inhibits metastasis-like lung invasion in vivo. Mol Cancer Ther. 2011;10:186–97.
14. Hartmann P, Szabo A, Eros G, Gurabi D, Horvath G, Nemeth I, Ghyczy M, Boros M. Anti-inflammatory effects of phosphatidylcholine in neutrophil leukocyte-dependent acute arthritis in rats. Eur J Pharmacol. 2009;622:58–64.
15. Fernie AR, Schauer N. Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet. 2009;25:39–48.
16. Shen S, Zhan C, Yang C, Fernie AR, Luo J. Metabolomics-centered mining of plant metabolic diversity and function: Past decade and future perspectives. Mol Plant. 2023;16:43–63. https://doi.org/10.1016/j.molp.2022.09.007.
17. Yang C, Shen S, Zhou S, Li Y, Mao Y, Zhou J, Shi Y, An L, Zhou Q, Peng W, et al. Rice metabolic regulatory network spanning the entire life cycle. Mol Plant. 2022;15:258–75.
18. Peng M, Shahzad R, Gul A, Subthain H, Shen S, Lei L, Zheng Z, Zhou J, Lu D, Wang S, et al. Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. Nat Commun. 1975;2017:8.
19. Carvalho S, Macel M, Schlerf M, Moghaddam FE, Mulder PP, Skidmore AK, van der Putten WH. Changes in plant defense chemistry (pyrrolizidine alkaloids) revealed through high-resolution spectroscopy. ISPRS J Photogramm Remote Sens. 2013;80:51–60.
20. Saric R, Nguyen VD, Burge T, Berkowitz O, Trtilek M, Whelan J, Lewsey MG, Custovic E. Applications of hyperspectral imaging in plant phenotyping. Trends Plant Sci. 2022;27:301–15.
21. Sun D, Cen H, Weng H, Wan L, Abdalla A, El-Manawy AI, Zhu Y, Zhao N, Fu H, Tang J, et al. Using hyperspectral analysis as a potential high throughput phenotyping tool in GWAS for protein content of rice quality. Plant Methods. 2019;15:54.
22. Yu K, Anderegg J, Mikaberidze A, Karisto P, Mascher F, McDonald BA, Walter A, Hund A. Hyperspectral Canopy Sensing of Wheat Septoria Tritici Blotch Disease. Front Plant Sci. 2018;9:1195.
23. Feng X, Yu C, Chen Y, Peng J, Ye L, Shen T, Wen H, He Y. Non-destructive Determination of Shikimic Acid Concentration in Transgenic Maize Exhibiting Glyphosate Tolerance Using Chlorophyll Fluorescence and Hyperspectral Imaging. Front Plant Sci. 2018;9:468.
24. Vergara-Diaz O, Vatter T, Carlisle Kefauver S, Obata T, Fernie AR, Luis Araus J. Assessing durum wheat ear and leaf metabolomes in the field through hyperspectral data. Plant J. 2020;102:615–30.
25. Cho J-S, Lim JH, Park KJ, Choi JH, Ok GS. Prediction of pelargonidin-3-glucoside in strawberries according to the postharvest distribution period of two ripening stages using VIS-NIR and SWIR hyperspectral imaging technology. LWT. 2021;141:110875.
26. Huang L, Zhou Y, Meng L, Wu D, He Y. Comparison of different CCD detectors and chemometrics for predicting total anthocyanin content and antioxidant activity of mulberry fruit using visible and near infrared hyperspectral imaging technique. Food Chem. 2017;224:1–10.
27. Hu N, Li W, Du C, Zhang Z, Gao Y, Sun Z, Yang L, Yu K, Zhang Y, Wang Z. Predicting micronutrients of wheat using hyperspectral imaging. Food Chem. 2021;343:128473. https://doi.org/10.1016/j.foodchem.2020.128473.
28. Wieme J, Mollazade K, Malounas I, Zude-Sasse M, Zhao M, Gowen A, Argyropoulos D, Fountas S, Van Beek J. Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: A review. Biosys Eng. 2022;222:156–76.
29. Chen J, Hu X, Shi T, Yin H, Sun D, Hao Y, Xia X, Luo J, Fernie AR, He Z, Chen W. Metabolite-based genome-wide association study enables dissection of the flavonoid decoration pathway of wheat kernels. Plant Biotechnol J. 2020;18:1722–35.
30. Yang W, Guo Z, Huang C, Duan L, Chen G, Jiang N, Fang W, Feng H, Xie W, Lian X, et al. Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. Nat Commun. 2014;5:5087.
31. Feng H, Guo Z, Yang W, Huang C, Chen G, Fang W, Xiong X, Zhang H, Wang G, Xiong L, Liu Q. An integrated hyperspectral imaging and genome-wide association analysis platform provides spectral and genetic insights into the natural variation in rice. Sci Rep. 2017;7:4401.
32. Simos YV, Verginadis II, Toliopoulos IK, Velalopoulou AP, Karagounis IV, Karkabounas SC, Evangelou AM. Effects of catechin and epicatechin on superoxide dismutase and glutathione peroxidase activity, in vivo. Redox Rep. 2012;17(5):181–6. https://doi.org/10.1179/1351000212Y.0000000020.
33. Deb G, Thakur VS, Limaye AM, Gupta S. Epigenetic induction of tissue inhibitor of matrix metalloproteinase-3 by green tea polyphenols in breast cancer cells. Mol Carcinog. 2015;54(6):485-99. https://doi.org/10.1002/mc.22121.
34. Liu S, Xu Z, Essemine J, Liu Y, Liu C, Zhang F, Iqbal Z, Qu M. GWAS unravels acid phosphatase ACP2 as a photosynthesis regulator under phosphate starvation condition through modulating serine metabolism in rice. Plant Commun. 2024;5(7):100885. https://doi.org/10.1016/j.xplc.2024.100885.
35. Qu L, Chu Y-J, Lin W-H, Xue H-W. A secretory phospholipase D hydrolyzes phosphatidylcholine to suppress rice heading time. PLoS Genet. 2021;17:e1009905.
36. Yuan S, Kim SC, Deng X, Hong Y, Wang X. Diacylglycerol kinase and associated lipid mediators modulate rice root architecture. New Phytol. 2019;223:261–76.
37. Zhang F, Guo H, Huang J, Yang C, Li Y, Wang X, Qu L, Liu X, Luo J. A UV-B-responsive glycosyltransferase, OsUGT706C2, modulates flavonoid metabolism in rice. Sci China Life Sci. 2020;63:1037–52.
38. Zhou J, Liu C, Chen Q, Liu L, Niu S, Chen R, Li K, Sun Y, Shi Y, Yang C. Integration of rhythmic metabolome and transcriptome provides insights into the transmission of rhythmic fluctuations and temporal diversity of metabolism in rice. SCIENCE CHINA-LIFE SCIENCES. 2022;65:1764–810.
39. Ma A, Qi X. Mining plant metabolomes: Methods, applications, and perspectives. Plant Communications. 2021;2:100238.
40. Choi J-H, Park SH, Jung D-H, Park YJ, Yang J-S, Park J-E, Lee H, Kim SM. Hyperspectral Imaging-Based Multiple Predicting Models for Functional Component Contents in Brassica juncea. Agriculture. 2022;12(10):1515. https://doi.org/10.3390/agriculture12101515.

41. Sytar O, Zivcak M, Brestic M, Neugart S. Assessment of hyperspectral indicators related to the content of phenolic compounds and multispectral fluorescence records in chicory leaves exposed to various light environments. Plant Physiol Biochem. 2020;154:429–38.
42. Yoon HI, Lee H, Yang J-S, Choi J-H, Jung D-H, Park YJ, Park J-E, Kim SM, Park SH. Predicting Models for Plant Metabolites Based on PLSR, AdaBoost, XGBoost, and LightGBM Algorithms Using Hyperspectral Imaging of Brassica juncea. Agriculture. 2023;13:1–12.
43. Wang Y, Zhang Y, Yuan Y, Zhao Y, Nie J, Nan T, Huang L, Yang J. Nutrient content prediction and geographical origin identification of red raspberry fruits by combining hyperspectral imaging with chemometrics. Front Nutr. 2022;9:980095.
44. Hernández-Hierro JM, Nogales-Bueno J, Rodríguez-Pulido FJ, Heredia FJ. Feasibility study on the use of near-infrared hyperspectral imaging for the screening of anthocyanins in intact grapes during ripening. J Agric Food Chem. 2013;61:9804–9.
45. Xing X, Zhao M, Wang X. TANG Y: Hyperspectral image-based measurement of total flavonoid content of leaf-use Ginkgo biloba L. Food Science and Technology. 2023;43:e100122.
46. Baiano A, Terracone C, Peri G, Romaniello R. Application of hyperspectral imaging for prediction of physico-chemical and sensory characteristics of table grapes. Comput Electron Agric. 2012;87:142–51.
47. Sun D, Robbins K, Morales N, Shu Q, Cen H. Advances in optical phenotyping of cereal crops. Trends Plant Sci. 2022;27:191–208.
48. Bosco GL, James L. Waters Symposium 2009 on near-infrared spectroscopy. TrAC Trends in Analytical Chemistry. 2010;29:197–208.
49. Xie W, Wang G, Yuan M, Yao W, Lyu K, Zhao H, Yang M, Li P, Zhang X, Yuan J. Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. Proc Natl Acad Sci. 2015;112:E5411–9.
50. Feng Hui: Integrative phenomics, metabolomics and genomics analysis provides new insights for deciphering the genetic basis of metabolism in polished rice. Figshare. doi: https://doi.org/10.6084/m9.figshare.28426328.v1.
51. Feng Hui: Integrative phenomics, metabolomics and genomics analysis provides new insights for deciphering the genetic basis of metabolism in polished rice. Github. https://github.com/dittlespark/Rice_HSI_MT_Programme.
52. Feng Hui: Integrative phenomics, metabolomics and genomics analysis provides new insights for deciphering the genetic basis of metabolism in polished rice. Zenodo. doi: https://doi.org/10.5281/zenodo.14880312.
53. Feng Hui: Integrative phenomics, metabolomics and genomics analysis provides new insights for deciphering the genetic basis of metabolism in polished rice. Figshare. doi: https://doi.org/10.6084/m9.figshare.28091447.

Publisher's Note