Genomics Proteomics Bioinformatics 20 (2022) 702–714

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Expanding the Coverage of Metabolic Landscape in Cultivated Rice with Integrated Computational Approaches

Xuetong Li [1,4,#], Hongxia Zhou [1,4,#], Ning Xiao [2], Xueting Wu [1], Yuanhong Shan [1], Longxian Chen [1,4], Cuiting Wang [1], Zixuan Wang [1], Jirong Huang [3,*], Aihong Li [2,*], Xuan Li [1,4,*]

[1] *CAS Key Laboratory of Synthetic Biology / National Center for Gene Research, CAS Center for Excellence in Molecular Plant Sciences / Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai 200032, China*
[2] *Lixiahe Agricultural Research Institute of Jiangsu Province, Yangzhou 225007, China*
[3] *Shanghai Key Laboratory of Plant Molecular Sciences, College of Life Sciences, Shanghai Normal University, Shanghai 200234, China*
[4] *University of Chinese Academy of Sciences, Beijing 100049, China*

Received 27 June 2019; revised 6 May 2020; accepted 8 September 2020
Available online 23 February 2021

Handled by Yu Xue

KEYWORDS

Untargeted metabolomics;
MS/MS spectral tag;
Structural characterization;
Phytochemical diversity;
Flavonoid derivative

Abstract Genome-scale metabolomics analysis is increasingly used for pathway and function discovery in the post-genomics era. The great potential offered by developed mass spectrometry (MS)-based technologies has been hindered, since only a small portion of detected metabolites were identifiable so far. To address the critical issue of low identification coverage in metabolomics, we adopted a deep metabolomics analysis strategy by integrating advanced algorithms and expanded reference databases. The experimental reference spectra and *in silico* reference spectra were adopted to facilitate the structural annotation. To further characterize the structure of metabolites, two approaches were incorporated into our strategy, *i.e.*, structural motif search combined with neutral loss scanning and metabolite association network. **Untargeted metabolomics** analysis was performed on 150 rice cultivars using ultra-performance liquid chromatography coupled with quadrupole-Orbitrap MS. Consequently, a total of 1939 out of 4491 metabolite features in the **MS/MS spectral tag** (MS2T) library were annotated, representing an extension of annotation coverage by an order of magnitude in rice. The differential accumulation patterns of flavonoids between *indica* and *japonica* cultivars were revealed, especially *O*-sulfated flavonoids. A series of closely-related

* Corresponding authors.
E-mail: lixuan@sippe.ac.cn (Li X), yzlah@126.com (Li A), huangjr@shnu.edu.cn (Huang J).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2020.06.018

flavonolignans were characterized, adding further evidence for the crucial role of tricin-oligolignols in lignification. Our study provides an important protocol for exploring **phytochemical diversity** in other plant species.

Introduction

It is estimated that the number of metabolites produced in green plants ranges from 200,000 to 1,000,000, underlying their broad chemical diversity and metabolic complexity [1]. Genome-scale metabolomics analysis has become a powerful tool to elucidate functional genes and pathways for diverse phytochemicals [2–5]. The recent progress in ultra-performance liquid chromatography coupled with high-resolution mass spectrometry (UPLC-HRMS) allows detecting metabolites at an unparalleled sensitivity, resolution, accuracy, and throughput [6]. However, the great power in advanced liquid-phase separation and MS technology has been limited, considering that a vast majority of metabolite features detected from plants remain unidentified currently [7,8]. It is a major challenge to detect and identify the massive amount of heterogeneous phytochemicals with a high dynamic range in concentrations, chemical and physical properties, and structures. The lagging in the identification of metabolites from plant sources can be attributed to various factors, *e.g.*, the insufficient performance of early MS-based platforms, the structural complexity of diverse metabolites, the limited availability of reference mass spectra from standard compounds, and the low throughput for processing and structure elucidating of MS data [9–12]. It is critical to handle and resolve the metabolomic data efficiently, as well as to bridge the gap between technological advances and demands of plant metabolomics research. In recent years, progresses have been made in improving metabolite annotation coverage through collecting reference mass spectra from more standard compounds [13–16] and developing computer-assisted approaches to facilitate the structure elucidation of metabolites [17–20].

Rice (*Oryza sativa* L.) is one of the major staple foods worldwide. Therefore, it is critical to explore chemical compositions and metabolic traits of rice for the enhancement of grain quality and nutritional value [21,22]. The two major subspecies of cultivated rice, *indica* and *japonica*, formed during domestication, display distinct features in morphology and physiology [23–25]. In recent years, a series of studies on rice metabolomics have been performed, which shed a light on the chemical diversity of rice [2,5,26,27]. However, there are still plenty of unknown metabolite features in the aforementioned studies and more efforts are needed to explore the metabolic diversity in rice. Other studies focused on phytochemical genomics to dissect the genetics basis underlying biosynthesis and physiological function of metabolites during the evolution and adaptation of plants [28]. Metabolic quantitative trait loci mapping and metabolic genome-wide association study have been used to reveal the genetic polymorphisms and candidate genes that affect metabolic traits in rice [2,5,27,29].

Our current study was designed to address a key issue in plant metabolomics, that is, the low identification coverage of metabolites. We sought to expand the annotation coverage with computational approaches, by adopting a deep metabolomics analysis strategy that combines experimental and *in silico* reference mass spectral libraries, as well as advanced algorithms. The structural motif search combined with neutral loss scanning and metabolite association network methods were integrated into our strategy to facilitate the characterization of structure and potential function of novel metabolites without reference from the aforementioned libraries. As a proof-of-concept study, using the state-of-the-art ultra-performance liquid chromatography coupled with quadrupole-Orbitrap mass spectrometry (UPLC-Q-Orbitrap-MS) platform, we performed an untargeted metabolomics analysis on a core collection of 150 *indica* or *japonica* cultivars grown in northeastern and southeastern China. An MS/MS spectral tag (MS2T) library for rice grains was constructed containing 4491 metabolite features, of which 1939 were annotated. The annotation coverage of rice metabolome was significantly improved through our strategy. Furthermore, our analyses revealed the systematic difference of metabolomes between *indica* and *japonica* subspecies and major differential accumulation patterns of flavonoid derivatives, especially *O*-sulfated flavonoids. A group of closely-related flavonolignans were newly uncovered in rice, providing further evidence for the crucial role of tricin-oligolignols in the lignification of monocots. Our deep metabolomics analysis strategy expands our understanding of phytochemical diversity and function in rice, which has a profound implication for improving the quality and nutritional value of crops through genetic breeding.

Results and discussion

Integration of computational approaches and performance evaluation

To handle the mass spectral data generated from UPLC-Q-Orbitrap-MS, we adopted a deep metabolomics analysis strategy with integrated computational approaches for sorting tandem mass spectral features and annotating detected metabolites (**Figure 1**). Metabolite annotation mainly contains two complementary approaches by referring to 1) experimental reference mass spectral data collected from public databases, and 2) *in silico* reference mass spectral data generated from structural databases for biologically relevant compounds. We further characterized the structure and potential function of novel metabolites without reference in the aforementioned libraries, using structural motif search combined with neutral loss scanning and metabolite association network (see Materials and methods).

The first annotation approach took advantage of the collections of experimental reference mass spectral data from public databases, including Metlin [16], MassBank [15], and ReSpect [14] (see Materials and methods). We evaluated the performance of two spectral similarity scoring algorithms, normalized dot product (NDP) [30] and INCOS [31], and chose INCOS for subsequent analysis because of its better performance (Figure S1A). Because of the limited availability of experimental reference mass spectra, the second approach was adopted to extend the coverage with *in silico* mass spectral

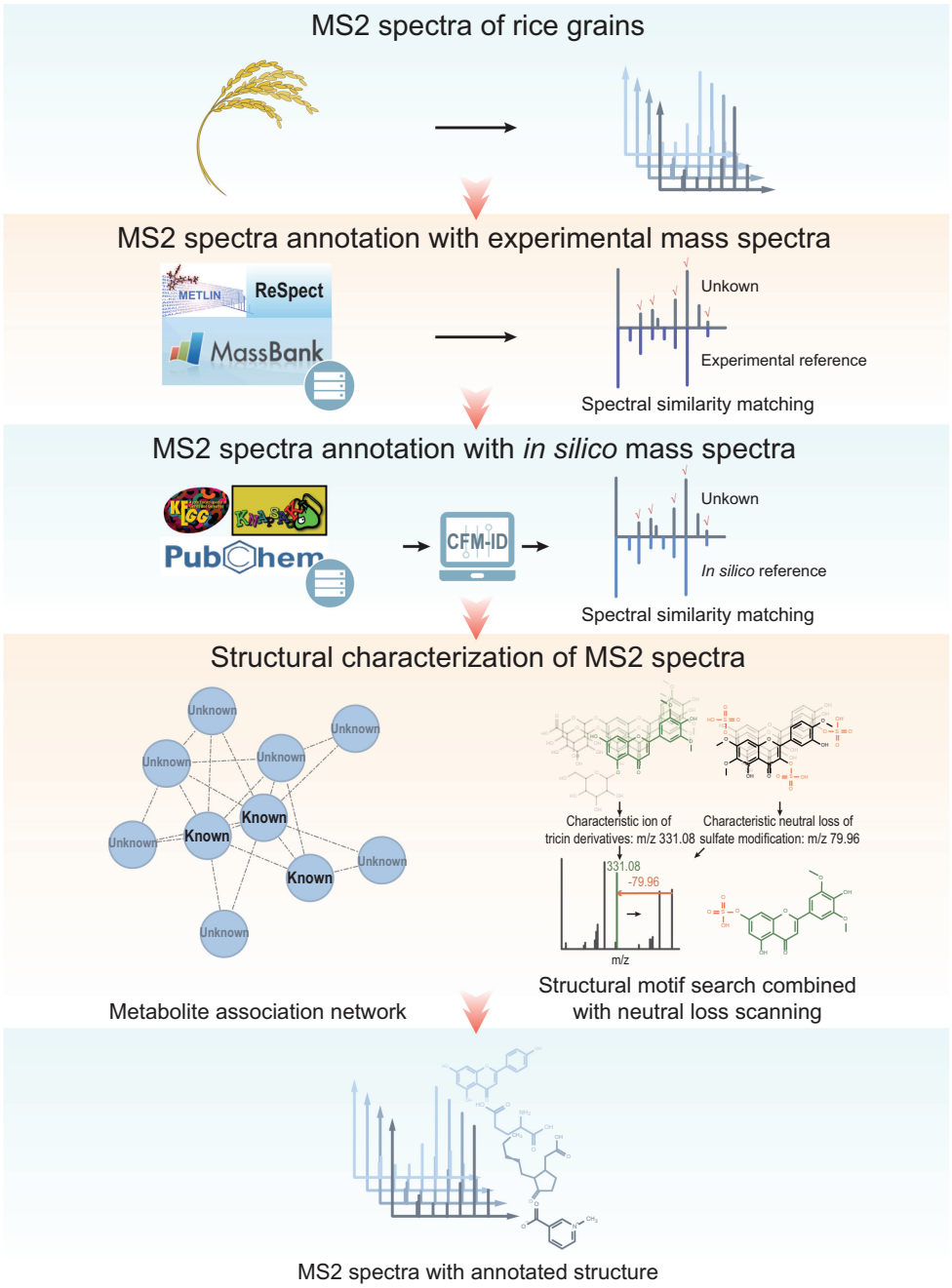

Figure 1 The deep metabolomics analysis strategy for large-scale structural annotation

The first approach adopted experimental reference mass spectra collected from public databases, Metlin, MassBank, and ReSpect, to annotate detected metabolites. The second approach adopted *in silico* reference mass spectra predicted from biologically relevant structure databases, KEGG, PubChem, and KNApSAcK, to annotate detected metabolites with improved coverage. CFM-ID software was used for *in silico* MS/MS spectra prediction. Two advanced methods were performed to characterize novel metabolites without reference in the aforementioned spectral and structural databases. The metabolite association network was constructed to infer the structure of unknowns based on known compounds within a common cluster of related metabolites. The structural motif search combined with neutral loss scanning method was implemented to characterize the substructure of novel metabolites, by matching unknown mass spectra with characteristic fragment ion and neutral loss of specific skeletons and modifications.

data for annotating those metabolites without a hit in the first approach. The *in silico* mass spectra were generated from an in-house structural database, structural database of biologically relevant compounds (SDBRC; Table S1), which contains the structural information of over 80,000 biologically relevant compounds collected from KEGG [32], PubChem [33], and KNApSAck [34] databases (see Materials and methods). The program, CFM-ID [18,35], was used for *in silico* fragmentation of compounds from SDBRC and similarity scoring of query and reference mass spectra.

To evaluate the performance of the aforementioned approaches, we sampled experimental mass spectra from Metlin and Massbank as query sets (Table S2). In the first approach using experimental mass spectra as a reference, INCOS had identification rates from 75% to 79% for top 1 match, and from 96% to 97% when the top 5 matches were included, which are comparatively higher than those of NDP (Figure S1A). In the second approach using *in silico* mass spectra as a reference, its performance was evaluated against KEGG and SDBRC libraries, respectively. The identification rates from 52% to 73% were observed for top 1 match, and 86% to 96% when the top 5 matches were included (Figure S1B). Searching against SDBRC results in lower identification rates than KEGG. The ubiquitous isomeric compounds generally have highly similar mass spectra and are difficult to distinguish through MS analyses. The identification rate will drop when we search against a larger reference database, mainly due to more isomers contained in the database [36]. SDBRC contains more biologically relevant compounds and will provide valuable reference structural information for more metabolite features. The combination of these two approaches will greatly expand the annotation coverage of plant metabolome and is instrumental in our study on the exploration of phytochemical diversity and function in rice.

The annotated MS2T library defines the metabolic diversity of rice grains

To construct an MS2T library for metabolomics analysis of rice grains, we used a collection of 150 representative rice accessions (Table S3). Rice grains were harvested from farmlands in southeastern and northeastern China and were mixed (referred to as reference mixture) for subsequent processing. The extracts were subjected to UPLC-Q-Orbitrap-MS (see Materials and methods). The raw data from repeated analyses were aligned using Compound Discover software (v2.0, ThermoFisher Scientific). Firstly, 158,840 and 118,077 signals detected from positive and negative modes were grouped to 11,263 and 6495 merged compound features, respectively. After the quality control and redundancy filtering steps, 2637 and 2446 metabolite features were retained for positive and negative modes, respectively, in which 2234 and 2123 were tagged with MS2 spectra. Finally, these metabolite features from positive and negative modes were merged, resulting in 4491 metabolite features with 3832 tagged with MS2 spectra (Figure S2; Tables S4 and S5). These metabolite features in the rice MS2T library were then annotated with our deep metabolomics analysis strategy (see Materials and methods). As a result, 298 metabolite features were annotated using experimental mass spectra as reference. For rest 3534 metabolite features, 1641 were annotated using *in silico* mass spectra as reference. Taken together, 1939 metabolite features were annotated in the MS2T library for rice grains (Table S5). The MS2T library constructed by our study was reported as recommended [37] (Tables S4 and S5).

Benefit from the HRMS and the deep metabolomics analysis strategy with integrated computational approaches, we expanded the metabolite annotation coverage of rice cultivars in comparison with previous studies [2,5,26,27]. Flavonoids account for a large portion of the increase of annotated metabolites in rice grains. The flavonoids annotated in our study display various modifications, such as glycosylation, acetyl-glycosylation, and sulfation. The glycosylation contains monoglycoside, diglycoside, and hexuronide. Examples include RSM04010p (quercetin-3-glucoside), RSM04966p (isovitexin-7-*O*-xyloside), RSM05128p (apigenin-7-*O*-gentiobioside), RSM05322p (demethoxycentaureidin-7-*O*-rutinoside), and RSM02409n (apigenin 4′-glucuronide) (**Figure 2**A–E). For sulfated flavonoids, an uncommon type of flavonoids, we found RSM02011n (ombuin 3-*O*-sulfate) (Figure 2F). The acetyl-glycosylation contains aliphatic and aromatic acylated glycoside. Examples include RSM05065p [tricin 7-(6-malonylglucoside)], RSM05648p [isovitexin 7-*O*-(6‴-*O*-E-p-coumaroyl)glucoside], and RSM05758p [7-*O*-(6-feruoylglucosyl)isoorientin] (Figure 2G–I). These modifications make flavonoids diverse in solubility, reactivity, stability, and function [38,39]. The flavonoids annotated in our study contribute to deepening our understanding of the diversity of enzymatic modifications in rice, which is beneficial to the exploration of the molecular mechanism of metabolite modifications in the growth, development, and interaction with the environment of plants.

Differential metabolic profile analysis reveals the featured metabolites of *indica* and *japonica* cultivars

To characterize the metabolic profiles of grains for diverse rice cultivars and understand their natural variation, we performed the untargeted metabolomics analysis on 59 rice cultivars, including 40 *indica* and 19 *japonica* (see Materials and methods). The metabolic profiles of rice cultivars contain the relative abundance of 3409 metabolite features (Table S6). The metabolic profiles of 59 rice cultivars were clustered based on the relative abundance of 3409 metabolite features, which displayed the differential patterns between *indica* and *japonica* cultivars (**Figure 3**A). In tree view (Figure 3B), the relation between *indica* with *japonica* cultivars is generally consistent with the phylogenetic relationship [40]. Through principal component analysis (PCA), *indica* and *japonica* cultivars were separated by the first component (PC1) and the second component (PC2), indicating the systematic difference in metabolic profiles between two subspecies (Figure 3C). We further performed orthogonal partial least squares discriminate analysis (OPLS-DA) to investigate the featured metabolites that differentiate *indica* and *japonica* cultivars. The *indica* and *japonica* cultivars were separated into two distinct clusters with our OPLS-DA model (**Figure 4**A). Metabolites with variable importance in projection value greater than 2.5, were defined as featured metabolites in our study. Among 58 featured metabolites (Table S7), 11 flavonoids, 3 terpenoids, and 2 phenylpropanoids were annotated. Particularly, three novel tricin derivatives, RSM03724n (tricin-*O*-sulfatohexoside), RSM04661n (tricin-*O*-acetylrhamnoside-*O*-diacetylrhamnoside), and RSM05814p (tricin-*O*-feruloyhexoside-hexoside) (Figure S3A–C; Table S7), were characterized using structural motif search combined with neutral loss scanning (see Materials and methods).

We further observed the differential accumulation patterns of *C*-glycosylated, *O*-glycosylated, and *O*-sulfated flavonoid derivatives among featured metabolites. The levels of four *C*-glycosylated flavonoids (flavone *C*-hexoside and flavone *C*-pentoside), RSM03824p (cytisoside), RSM04142p (precatorin

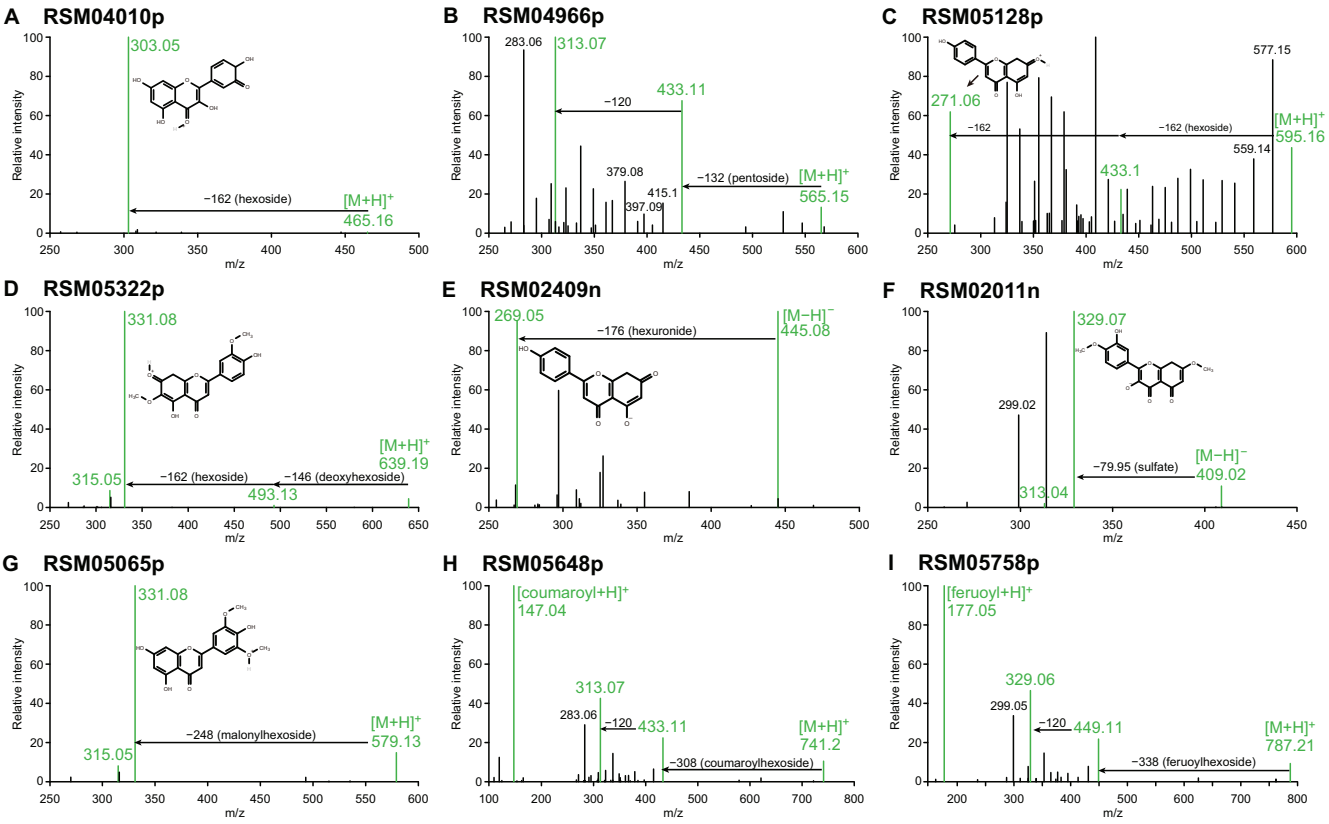

Figure 2 The mass spectra of annotated flavonoids with diverse modifications

A. RSM04010p (isoquercitrin). The *m*/*z* 303.04922 is the featured protonated ion of quercetin, and the neutral loss of *m*/*z* 162.1087 corresponds to a hexoside group. **B.** RSM04966p (isovitexin-7-*O*-xyloside). The *m*/*z* 313.07016 and *m*/*z* 433.11319 are the featured protonated ions of isovitexin (the neutral loss of *m*/*z* 120.043 is the characteristic of *C*-hexosyl flavonoids), and the neutral loss of *m*/*z* 132.0404 corresponds to a pentoside group. **C.** RSM05128p (apigenin-7-*O*-gentiobioside). The *m*/*z* 271.05936 is the featured protonated ion of apigenin, and the neutral loss of *m*/*z* 324.1032 corresponds to two hexoside groups. **D.** RSM05322p (demethoxycentaureidin-7-*O*-rutinoside). The *m*/*z* 315.04944 and *m*/*z* 331.08078 are the featured protonated ions of demethoxycentaureidin, and the neutral loss of *m*/*z* 146.0587 corresponds to a deoxyhexoside (rhamnoside) group. **E.** RSM02409n (apigenin 4′-glucuronide). The *m*/*z* 269.04575 is the featured deprotonated ion of apigenin, and the neutral loss of *m*/*z* 176.0317 corresponds to a hexuronide group. **F.** RSM02011n (ombuin 3-*O*-sulfate). The *m*/*z* 313.03574 and *m*/*z* 329.06674 are the featured deprotonated ions of ombuin, and the neutral loss of *m*/*z* 79.95658 corresponds to a sulfate group. **G.** RSM05065p [tricin 7-(6-malonylglucoside)]. The *m*/*z* 315.04868 and *m*/*z* 331.08017 are the featured protonated ions of tricin, and the neutral loss of *m*/*z* 248.0524 corresponds to a malonylhexoside group. **H.** RSM05648p [isovitexin 7-*O*-(6‴-*O*-E-p-coumaroyl)glucoside]. The neutral loss of *m*/*z* 308.0898 corresponds to a coumaroylhexoside group, and the *m*/*z* 147.04376 is the featured protonated ion of p-coumaroyl unit. **I.** RSM05758p [7-*O*-(6-feruoylglucosyl)isoorientin]. The *m*/*z* 449.10651 and *m*/*z* 329.06485 are the featured protonated ions of isoorientin, the neutral loss of *m*/*z* 338.0989 corresponds to a feruoylhexoside group, and the *m*/*z* 177.05418 is the featured protonated ion of feruoyl unit. $[M+H]^+$ and $[M-H]^-$ indicate the protonated and deprotonated precursor ions of flavonoids, respectively; RSM*****p/n indicates the serial number of rice's screening mass spectra acquired in positive or negative ion mode.

I), RSM03991p (trihydroxy-methoxyflavone *C*-hexoside) (Figure S3D), and RSM04767p (di-*C*,*C*-pentosyl-apigenin) (Figure S3E) are significantly higher in *indica* than *japonica* cultivars (Figure 4B; Table S7). In contrast, the levels of four *O*-glycosylated flavonoids with guaiacylglyceryl or acyl modification, RSM05526p [tricin 4′-*O*-(guaiacylglyceryl) ether 7″-*O*-glucopyranoside], RSM05648p [isovitexin 7-*O*-(6‴-*O*-E-p-coumaroyl)glucoside], RSM04661n (tricin *O*-acetylrhamnoside-*O*-diacetylrhamnoside), and RSM05814p (tricin *O*-feruloylhexosyl-*O*-hexoside) are significantly higher in *japonica* than *indica* cultivars (Figure 4C; Table S7). Furthermore, differences in two *O*-sulfated flavonoids, RSM02011n (ombuin 3-*O*-sulfate) and RSM03724n (tricin *O*-sulfatohexoside), were observed between *indica* and *japonica* cultivars (Figure 4D; Table S7). The differential accumulation patterns of *C*-glycosylated and *O*-glycosylated flavonoids in rice grains were consistent with previous studies in rice leaves [26,41]. Additionally, we expanded our findings to *O*-sulfated flavonoids, an uncommon variety of flavonoid derivatives catalyzed by sulfotransferases [39,42]. It has been revealed that the natural variation of salicylic acid sulfotransferase encoding gene causes the differentiation of resistance to rice stripe virus between *indica* and *japonica* subspecies [43], highlighting the significant role of sulfation in pathogen resistance of rice. However, a rare study has been performed to characterize flavonoid sulfotransferases in rice. The differential accumulation

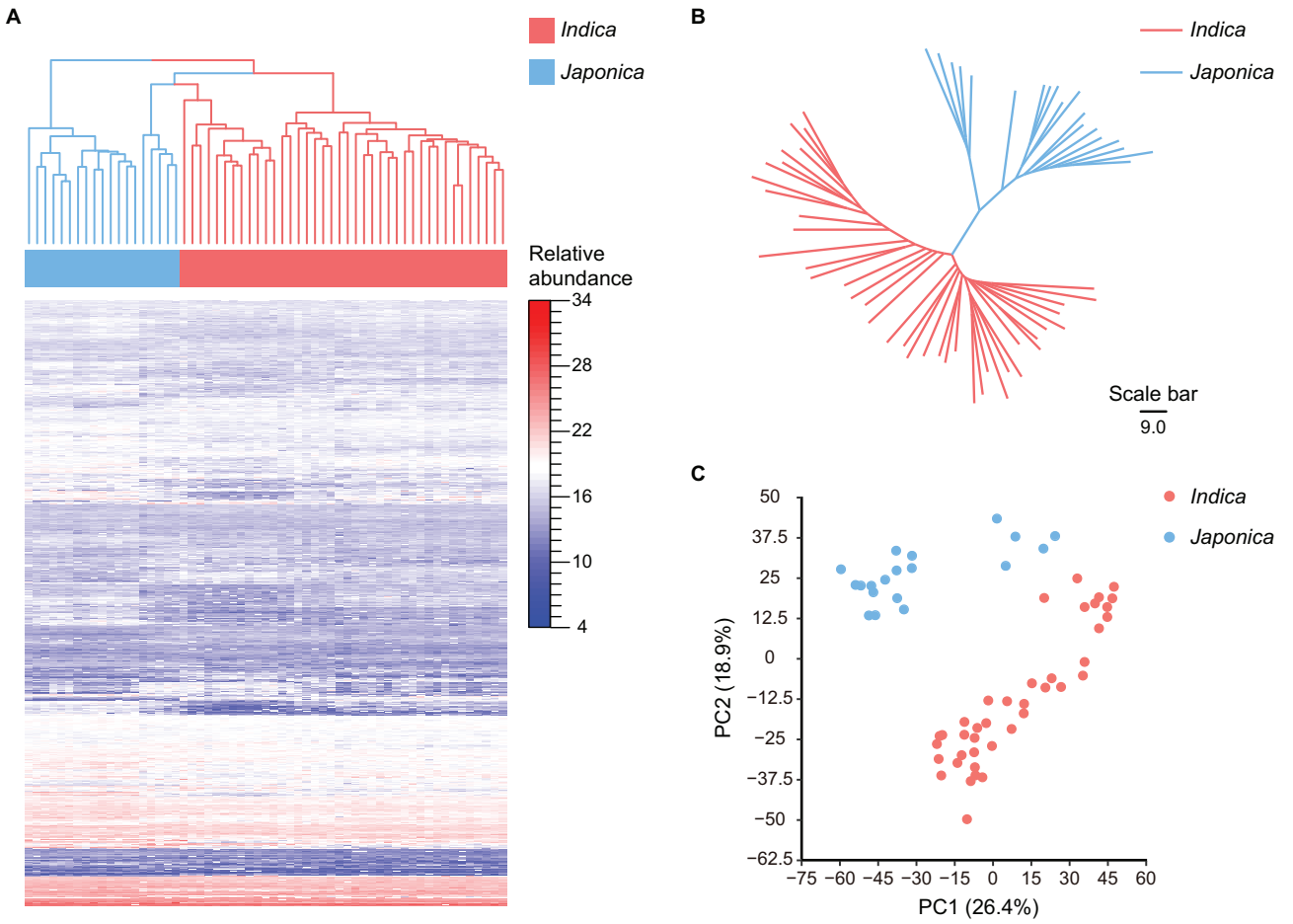

Figure 3 The differential metabolic profiles between *indica* and *japonica* cultivars
A. The heatmap and hierarchical clustering of 59 rice cultivars based on the relative abundances of 3409 metabolites. **B.** The neighbor-joining tree of 59 rice cultivars based on the relative abundances of 3409 metabolites. **C.** The score plot for PCA of 59 rice cultivars based on the relative abundance of 3409 metabolites. The PC1 and PC2 account for 26.4% and 18.9% of the variance, respectively. PCA, principal component analysis; PC1, the first principal component; PC2, the second principal component.

patterns of *O*-sulfated flavonoids revealed by our study provided new insight into the natural variation of flavonoid sulfotransferase activity, which is beneficial to the exploration of biosynthesis genes of flavonoid sulfotransferases and their potential functions in pathogen resistance of rice.

The metabolite association network characterizes diverse flavonolignans involved in lignification

The network-based analysis is widely used in metabolomics studies for the understanding of metabolite interaction, structural characterization, and pathway elucidation [20,44–47]. Previous studies suggested that metabolites with similar structures generally display correlation in their abundance, so the structure of unknown metabolites can be inferred by knowns through the metabolite association network [2,4,27,48].

We constructed the metabolite association network with the Gaussian graphical model (GGM) [49], using the metabolic profiles of 59 rice cultivars (see Materials and methods; Figure S4A and Table S8). This network contains 2874 nodes (metabolites) with 42,147 significant edges (metabolite pairs). The 64 clusters were isolated (Table S9) from the GGM network using molecular complex detection (MCODE) program [50] (see Materials and methods). A subgroup of the first-ranked cluster mainly contains flavonoids. Besides, within the second-ranked cluster, a large number of nodes were annotated as terpenoids, most of which are triterpenoids (Figure S4B; Table S10).

A subgroup of the first-ranked cluster contains 32 metabolites (**Figure 5**A; Table S10). Thirteen of them were annotated as common flavonoids with hydroxy and methoxy groups (Figure S5). Notably, within this cluster, we found some flavonolignans (Figure 5B, Figure S6), which are produced via oxidative coupling between flavonoids with three varieties of monolignols, p-coumaryl, coniferyl, and sinapyl alcohols [51]. RSM04702p (salcolin B) [52] and RSM04355p (5′-methoxyhydnocarpin-D) [53] are guaiacyl flavonolignans, while RSM04382p (aegicin) is p-hydroxyphenyl flavonolignans [54]. Based on the aforementioned findings, we suggested that there are other flavonolignans within this cluster. We then observed the precursor ion and fragmentation pattern of unknowns within this cluster and characterized more flavonolignans. The RSM04691p displays the same fragment ion (*m*/*z* at 315.04895) with RSM04355p in mass spectra, which means

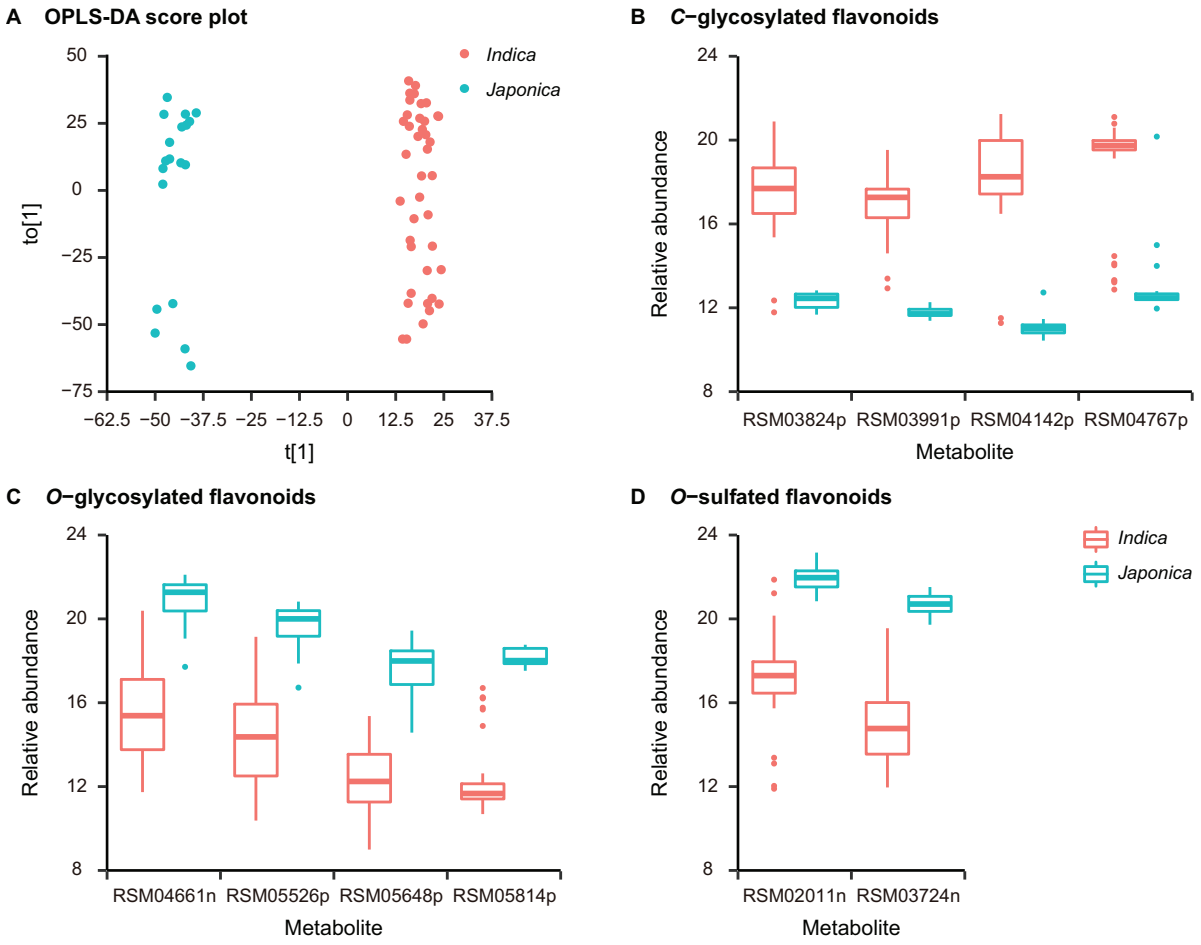

Figure 4 The featured metabolites of *indica* and *japonica* cultivars

A. The score plot for OPLS-DA of 59 rice cultivars based on the relative abundances of 3409 metabolites. The R^2X, R^2Y (goodness-of-fit parameter), and Q^2 (predictive ability parameter) of the OPLS-DA model are 0.555, 0.99, and 0.98, respectively. **B.** The boxplot of the relative abundances of four *C*-glycosylated flavonoids among featured metabolites. RSM03824p, cytisoside; RSM03991p, trihydroxy-methoxyflavone *C*-hexoside; RSM04142p, precatorin I; RSM04767p, di-*C*,*C*-pentosyl-apigenin. **C.** The boxplot of the relative abundances of four *O*-glycosylated flavonoids among featured metabolites. RSM04661n, tricin *O*-acetylrhamnoside-*O*-diacetylrhamnoside; RSM05526p, tricin 4′-*O*-(guaiacylglyceryl) ether 7″-*O*-glucopyranoside; RSM05648p, isovitexin 7-*O*-(6‴-*O*-E-p-coumaroyl)glucoside; RSM05814p, tricin *O*-feruloylhexoside-*O*-hexoside. **D.** The boxplot of the relative abundances of two *O*-sulfated flavonoids among featured metabolites. RSM02011n, ombuin 3-*O*-sulfate; RSM03724n, tricin *O*-sulfatohexoside. OPLS-DA, orthogonal partial least squares discriminate analysis; t[1], the predictive component of the OPLS-DA model; to[1], the orthogonal component of the OPLS-DA model.

they have the same flavonoid moiety in structures. The mass difference between their precursor ions is 30.01031, corresponding to a methoxy group. Thus the RSM04691p has an additional methoxy group at the coniferyl alcohol moiety of RSM04355p, which was characterized as palstatin [55], a syringyl flavonolignan. With the same method, the structures of RSM05474p, RSM05479p, RSM05574p, and RSM04546n were characterized. The RSM05474p was characterized as tricin *O*-[guaiacyl-(*O*-p-coumaroyl)-glyceryl] ether [56], which has an additional coumaroyl unit, a featured modification of lignins [57], at the guaiacylglyceryl group of RSM04702p. The RSM05479p, RSM05574p, and RSM04546n were characterized as tricin-oligolignol trimers, which are formed by further chain extension through oxidative coupling between tricin-oligolignol dimer (RSM04702p) and p-coumaryl alcohol or coniferyl alcohol via ether or furan bridge (Figure 5B; Figure S6E–G) [58]. In previous studies, the presence of unusual catechyl lignins derived from caffeyl alcohol had been revealed in plants [59]. Unexpectedly, in our study, we found that RSM04164p and RSM04201p show the spectrum features of catechyl flavonolignans, which both have the characteristic fragment ion of flavone moiety and neutral loss of caffeyl alcohol unit (*m*/*z* 166.0626). Thus we inferred that the structures of RSM04164p and RSM04201p are dihydroxy-dimethoxyflavone and tetrahydroxy-methoxyflavone moiety linked with a caffeyl alcohol unit by dioxane bridge, respectively (Figure S6I and J).

In addition to RSM04382p and RSM04702p found in rice leaves and grains previously [4,26,60], the rest eight flavonolignans were characterized by our study in rice grains, which greatly expanded the diversity of flavonolignans in rice. Previously, the occurrence of tricin in lignins has been reported in a

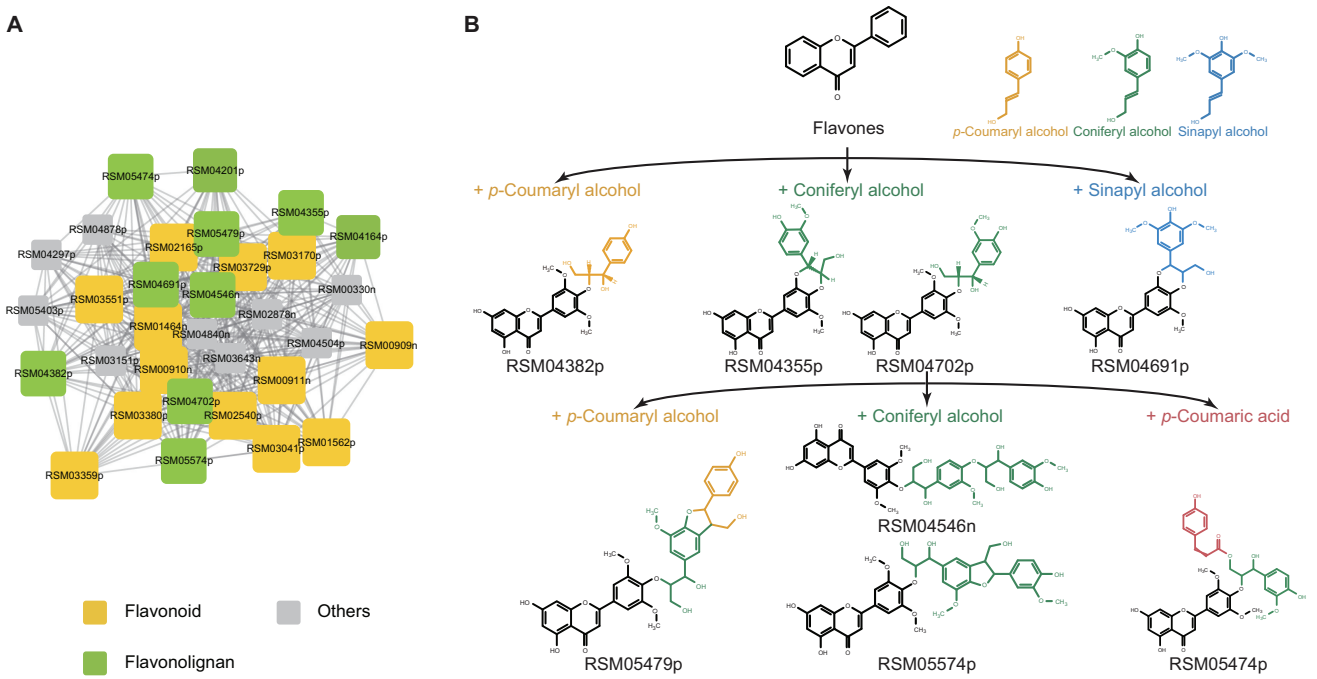

Figure 5 The subgroup of the first-ranked cluster containing flavonoids and flavonolignans

A. Components and their partial correlation relationships within the subgroup of the first-ranked cluster. **B.** The structure and relationship of characterized flavonolignans. Colors in yellow, green, and blue denote p-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol or their derived moieties, respectively. Color in red denotes the coumaroyl unit. RSM04382p, RSM04355p, RSM04702p, and RSM04691p are flavonolignan dimers derived from the oxidative coupling between flavonoids with monolignols. RSM05474p has an additional coumaroyl unit on the guaiacylglyceryl group of RSM04702p. RSM05479p, RSM05574p, and RSM04546n are flavonolignan trimers derived from the further chain extension between RSM04702p with monolignols.

series of monocots [61]. Tricin was found to be incorporated into lignins as tricin-oligolignols and acts as a nucleation site in the initiation of lignin polymers in maize [51,58]. A group of closely-related tricin-oligolignol dimers and trimers found in our study further supported the crucial role of tricin-oligolignols in lignification of rice. Additionally, the characterization of non-tricin flavonolignans, such as RSM04355p and RSM04691p, provided evidence for the presence of more diverse flavonoids in lignification. Within this cluster, six additional metabolites were found to contain featured ions of tricin in their mass spectra, which may be putative tricin derivatives. Two of them show the neutral loss of guaiacylglyceryl or p-hydroxyphenylglyceryl unit, although their entire structures remain unknown (Figure S7; Table S10).

Conclusion

The technical and analytical obstacles in the identification of metabolites hindered the further research of phytochemical diversity and function in plants. To address the issue of low identification coverage in plant metabolomics, we adopted a deep metabolomics analysis strategy for large-scale metabolite structural annotation. The experimental and *in silico* mass spectra were used to facilitate the metabolite annotation with high coverage. The structural motif search combined with neutral loss scanning and metabolite association network methods were further adopted to characterize the structure and function of metabolites in rice. The untargeted metabolomics study on rice grains was performed, and the coverage of annotated metabolites was significantly improved. Benefited from the rice metabolome with expanded annotation coverage, the systematic differences in metabolic profiles between *indica* and *japonica* cultivars were further defined, including the differential accumulation patterns of *C*-glycosylated, *O*-glycosylated, and *O*-sulfated flavonoids, and a series of closely-related flavonolignans with key roles in lignification was uncovered. Our strategy can be applied to the metabolomics researches of other agronomically important plants, with great potential in the enhancement of crop quality and nutrition value through genetic breeding.

Materials and methods

Plant materials

Rice cultivars, including 93 *japonica* and 85 *indica* accessions, were used in this study (Table S3). Rice cultivars were planted and harvested during the summer season in 2015 and 2016 from two locations of China: farmlands in Yangzhou, Jiangsu Province (E 119°53′, N 32°42′; in southeastern China) and farmlands in Harbin, Heilongjiang Province (E 126°53′, N 45°69′; in northeastern China). Rice cultivars were planted in the field, ten plants for each row, and three rows for each accession.

For each accession, grains were harvested for two biological replicates, each containing grains from three individual plants.

Grains were packed in the gauze bag and air-dried in shade. Two grams of dried grains were ground using tissue grinder (Catalog No. 05010997, Shanghai BiHeng Biotechnology Company Limited, Shanghai, China) at 55 Hz for 40 s. The fine powder for each accession was stored at −80 °C for subsequent processing.

Chemicals

HPLC-grade methanol (Catalog No. 1.06007.4008), acetonitrile (Catalog No. 1.00030.4008), and acetic acid (Catalog No. 5.43808.0250) were obtained from Merck (Darmstadt, Germany). Ultra-pure water was produced using Millipore water purifier (Milli-Q, Millipore, Billerica, MA). The lidocaine (Catalog No. 137-58-6, Dr. Ehrenstorfer GmbH, Augsburg, Germany) and lincomycin (Catalog No. 859-18-7, Dr. Ehrenstorfer GmbH) were purchased from ANPEL Laboratory Technologies (Shanghai) Inc. Other chemicals were purchased from Sigma-Aldrich (Shanghai) Trading Co., Ltd. (Sigma-Aldrich, Merck KGaA), if not otherwise specified.

Metabolite extraction

Briefly, 150 mg of powder of rice grains was mixed with 1.5 ml 70% aqueous methanol solution A (containing 1 mg/l vitexin, 1 mg/l p-coumaric acid, and 1 mg/l lidocaine as internal standards). The mixture was vortexed every 10 min for three times and placed in 4 °C refrigerator overnight. The mixture was then centrifuged at 12,000 *g* for 10 min in 4 °C. The supernatant of the mixture was dried with a concentrator under vacuum and re-dissolved with 150 µl 70% aqueous methanol solution B (containing 1 mg/l capsaicin and 1 mg/l lincomycin as internal standards). Then the extract was filtered with 0.22 µm filter (Catalog No. SCAA-104, ANPEL Laboratory technologies Inc.) and transferred into a sample bottle for the subsequent UPLC-MS/MS analysis.

UPLC-MS/MS analysis

Chromatographic separation of extract samples was performed on Waters Acquity Ultra Performance Liquid Chromatography using an ACQUITY UPLC BEH C18 column (1.7 µm, 2.1 mm × 100 mm) (Waters Corporation, Milford, MA). The mobile phase consisted of (A) water with 0.04% acetic acid and (B) acetonitrile with 0.04% acetic acid. The gradient program was as follows: 95:5 A/B at 0 min, 5:95 A/B at 20.0 min, 5:95 A/B at 24.0 min, 95:5 A/B at 24.1 min, and 95:5 A/B at 30.0 min. The flow rate was 0.25 ml/min and the injection volume was 5 µl. The column temperature was 40 °C.

The UPLC system was coupled with Q Exactive hybrid Q-Orbitrap-high resolution mass spectrometer (Q-Orbitrap-HRMS) (ThermoFisher Scientific, Waltham, MA). The MS acquisition was performed in positive and negative ionization with FullScan/dd-MS2 (top 8) mode, in which the MS/MS spectra of most abundant ions (top 8) within each scanning window was automatically obtained. MS full scan mass resolution was set to 70,000 at m/z 200 and data-dependent MS/MS with full scan mass resolution was reduced to 17,500 at m/z 200. The m/z range of MS full scan was 100–1000.

Heated electrospray ionization (HESI) parameters were as follows: spray voltage (+), 4000 V; spray voltage (−), 3500 V; capillary temperature, 320 °C; sheath gas, 35 arb; aux gas, 8 arb; probe heater temperature, 350 °C; S-Lens RF level, 50. Higher energy collisional dissociation (HCD) energies were 15 eV and 40 eV, and the average MS/MS spectrum was obtained. The mass spectrometer was calibrated using Pierce LTQ Velos ESI positive ion calibration solution and Pierce ESI negative ion calibration solution (ThermoFisher Scientific).

The sequence of injections for extract samples was randomized to reduce bias. The grains mixture of 150 randomly selected rice accessions was used to build a reference MS2T library. The reference mixture was submitted to UPLC-MS/MS system once every 10 samples. In total, injections of reference mixture were repeated 43 times in positive and negative modes.

Mass spectrum data processing

The raw data generated from HESI-Q-Orbitrap-HRMS were processed with Compound Discoverer software (v2.0, ThermoFisher Scientific) using its automatic workflow. The retention time aligning parameters were as follows: mass tolerance, 5 ppm; maximum shift, 0.5 min. The unknown compounds detecting parameters were as follows: min peak intensity, 2×10^6; S/N threshold, 5.

Raw metabolite features were further filtered by: 1) removing signals that are of poor quality or non-biological origin [62], *i.e.*, features with reproducibility < 90%, sample to blank ratio < 10%, relative standard deviation > 50%, or peak area < 1×10^5; and 2) removing redundancy from multi-ion adducts (Na^+, K^+, $NH4^+$, Cl^-), isotopes, in-source fragmentation, or dimerization. Metabolite features in positive ($[M+H]^+$) and negative ($[M-H]^-$) modes were merged with the following parameters: exact mass tolerance, 5 ppm; retention time tolerance, 0.5 min. The in-house script based on Xcalibur development kit in Xcalibur software (v2.2, ThermoFisher Scientific) was used to automatically extract the MS2 spectra of metabolite features.

Metabolite annotation

Metabolite annotation mainly adopted two complementary approaches with experimental/*in-silico* mass spectra as reference. The first approach used the experimental reference mass spectra library collected from public databases, such as Metlin [16], MassBank [15], and ReSpect [14]. This library contained a total of 98,658 mass spectra for about 24,385 compounds. Two algorithms, NDP and INCOS, were implemented as described [30,31] using Perl scripts to score the similarity between query and reference mass spectra. INCOS algorithm was selected for further analysis because of its better performance. We respectively searched against the Metlin, MassBank, and ReSpect libraries and merged the annotation results subsequently. The experimental reference mass spectra that have similar precursor m/z (mass tolerance: 10 ppm) with query mass spectra were retrieved and compared for similarity using INCOS. Reference mass spectra with a similarity score > 0.75 were retained for the annotation of query mass spectra. The reference spectra with the highest similarity score in the annotation results was selected as the putative annotation for the query spectra. In the evaluation of NDP and INCOS algorithms

using query spectra sampled from Metlin or MassBank, the query spectra themselves were excluded from the matching results to rule out bias. The performance of the INCOS with similarity score cutoff (0.75) was further evaluated with the test set for standard MS/MS spectra of Fiehn HILIC library from MassBank of North America (Figure S8A; Table S11).

The second approach was adopted to extend the annotation coverage with *in silico* mass spectra. First, the structure data were collected from three biologically relevant structure databases, including KEGG, 'BioChem' (the manually selected subset of biologically relevant compounds in PubChem), and KNApSAck. For 'BioChem' database, compounds in PubChem with NCBI BioSystems annotation, biological role classification of ChEBI Ontology, Flavonoids or Prenol Lipids classification of LIPID MAPS [63] were selected. The OpenBabel software [64] was used to convert the raw structural data to machine-readable structural information, including formula, exact mass, simplified molecular-input line-entry system, and the IUPAC international chemical identifier. Through merging compounds from different databases and removing redundancy, we constructed a structural database of biologically relevant compounds (SDBRC), which was used as a reference to retrieve and generate *in silico* mass spectra. The program, CFM-ID, was used for *in silico* fragmentation of compounds from SDBRC, and similarity scoring as described [18,35]. CFM-ID adopts a machine learning technique with a probabilistic generative model for the compound fragmentation process. The source code for CFM-ID software (v2.0) was obtained from the SourceForge platform (https://sourceforge.net/projects/cfm-id/) and compiled on the Linux system (CentOS release 6.2). The *in silico* mass spectra for reference compounds that have similar mass (mass tolerance: 5 ppm) with query mass spectra were generated using CFM-ID, and similarity scores between the query and *in silico* mass spectra were calculated. Reference compounds with a similarity score >0.3 were retained for the annotation of query mass spectra. The performance of the annotation through CFM-ID with similarity score cutoff (0.3) was further evaluated with the test set for standard MS/MS spectra of Fiehn HILIC library from MassBank of North America (Figure S8B; Table S11).

The annotation results for 17 metabolite features were further identified through the comparison of retention time (RT) and MS/MS spectra with standard compounds (Figure S9; Table S12).

Structural motif search combined with neutral loss scanning

The structural motif search combined with neutral loss scanning was further developed from previous studies [14,26,60,65]. It is based on the theory that compounds with similar structures (*i.e.*, same skeletons or modifications) would generate featured fragment ions or neutral losses in mass spectral analysis. These compounds often belong to a certain phytochemical class. Flavonoids have a core diphenylpropane backbone (C6-C3-C6) with diverse modifications and display regular fragmentation patterns in their mass spectra. To mine their fragmentation regularities systematically and facilitate the characterization of novel flavonoids, 3145 MS/MS spectra of two major classes of flavonoids, flavones and flavonols, were generated *in silico* by CFM-ID software from structure data in LIPID MAPS Structure Database [63]. Through the statistical analysis, we obtained a series of structural motifs (characteristic fragment ions) frequently found in mass spectra of flavones and flavonols, such as m/z at 287.0550145 (featured ion of kaempferol derivatives), m/z at 303.0499291 (featured ion of quercetin derivatives), m/z at 271.0600999 (featured ion of apigenin derivatives), and m/z at 301.0706646 (featured ion of chrysoeriol derivatives). Besides, a set of frequently found neutral losses were observed, such as the neutral losses of hexoside (m/z 162.0530308), pentoside (m/z 132.0423309), rhamnoside (m/z 146.0576808), hexuronide (m/z 176.0322455), sulfate (m/z 79.9568149), and coumaroylhexoside (m/z 308.0892455) groups. We searched for the presence of structural motifs and neutral losses in unknown MS/MS spectra to characterize its putative structure. The detailed steps of structural motif search combined with neutral loss scanning are listed in Figure S10.

The metabolic profiles of rice cultivars

The metabolic profiles of 59 rice cultivars grown in Yangzhou in 2016 were obtained from the corresponding peak areas of raw mass spectrometric data using Compound Discoverer software (v2.0, ThermoFisher Scientific). The metabolic profiles of rice cultivars were defined according to our reference MS2T library. The metabolite features in metabolic profiles were aligned with reference MS2T library to determine corresponding structural information as described [65]. The parameters used to determine corresponding structural information were as follows: the tolerance of retention time, 0.35 min; the tolerance of mass, 5 ppm. To ensure the consistency among samples during UPLC-MS/MS analysis, the reference control mixtures were inserted into the analytical sequence once every 10 samples. The data of metabolite abundance was normalized based on internal standard and reference control mixtures as described [66]. Two biological repeats for each rice accession were performed, and the normalized data was averaged and $\log_2$-transformed for further analysis. The detailed steps of the acquisition and processing of relative abundance data of metabolites are listed in File S1.

Construction of GGM-based network

For the construction of the GGM network, a data matrix containing the relative abundance of 3409 metabolite features for 59 rice cultivars was first generated. GeneNet package [67] was used to calculate the partial correlation coefficients and test the significance of the partial correlation of each metabolite pair. The metabolite pairs with probability greater than 0.99 (local FDR <0.01) were defined as significant edges and included in the GGM network. The Cytoscape software [68] was used for the visualization of the GGM network. The MCODE application was used to find clusters from the GGM network with parameters as defaulted [50].

Statistical analysis

R software (v3.2.3, https://www.R-project.org/) [69] was mainly used for statistical analyses, if not specifically indicated otherwise. The metabolic profiles of rice cultivars were clustered using hierarchical clustering. The method of hierarchical clustering is the unweighted pair group method with arithmetic

mean. The heatmap was constructed by heatmap.2 function in the gplots package (https://CRAN.R-project.org/package=gplots) [70]. The relatedness distance between the metabolic profiles of rice cultivars was calculated by Euclidean distance function. The neighbor-joining tree was constructed by MEGA7 software [71] using the matrix of Euclidean distance between metabolic profiles of rice cultivars. PCA and OPLS-DA were carried out by SIMCA-P software (v14.0, Umetrics, Umea, Sweden).

Data availability

The MS2 spectral data and MS2T library have been deposited in the National Omics Data Encyclopedia database at the Bio-Med Big Data Center, CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institute of Nutrition and Health (SINH), Chinese Academy of Sciences (NODE: OEP001184), which are publicly accessible at https://www.biosino.org/node/project/detail/OEP001184.

CRediT author statement

Xuetong Li: Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Hongxia Zhou:** Methodology, Validation, Formal analysis, Investigation. **Ning Xiao:** Investigation, Resources. **Xueting Wu:** Formal analysis. **Yuanhong Shan:** Formal analysis, Investigation. **Longxian Chen:** Investigation. **Cuiting Wang:** Investigation. **Zixuan Wang:** Conceptualization. **Jirong Huang:** Conceptualization, Resources. **Aihong Li:** Investigation, Resources. **Xuan Li:** Conceptualization, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing financial interests.

Acknowledgments

We thank Prof. Jie Luo for assistance in sample preparation in LC-MS analysis, and Dr. Ping Chen for help with data submission to the National Omics Data Encyclopedia database. This work was supported by grants from the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA24010400), the National Key R&D Program of China (Grant Nos. 2018YFA0900700 and 2019YFA0904601), the Major Project of Jiangsu Province of China for Significant New Varieties Development (Grant No. PZCZ201702), and the National Natural Science Foundation of China (Grant Nos. 31900470, 31701137, and 31972881).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2020.06.018.

ORCID

ORCID 0000-0003-0029-2296 (Xuetong Li)
ORCID 0000-0001-9206-2580 (Hongxia Zhou)
ORCID 0000-0001-6181-2684 (Ning Xiao)
ORCID 0000-0002-8644-124X (Xueting Wu)
ORCID 0000-0002-2169-7308 (Yuanhong Shan)
ORCID 0000-0002-1209-1945 (Longxian Chen)
ORCID 0000-0002-8251-5774 (Cuiting Wang)
ORCID 0000-0002-4198-7230 (Zixuan Wang)
ORCID 0000-0002-4032-4566 (Jirong Huang)
ORCID 0000-0001-6161-9796 (Aihong Li)
ORCID 0000-0002-7909-7241 (Xuan Li)

References

[1] Saito K, Matsuda F. Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol 2010;61:463–89.
[2] Matsuda F, Nakabayashi R, Yang Z, Okazaki Y, Yonemaru J, Ebana K, et al. Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. Plant J 2015;81:13–23.
[3] Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C, et al. Rewiring of the fruit metabolome in tomato breeding. Cell 2018;172:249–61.e12.
[4] Chen W, Wang W, Peng M, Gong L, Gao Y, Wan J, et al. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. Nat Commun 2016;7:12767.
[5] Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet 2014;46:714–21.
[6] Alvarez Rivera G, Ballesteros Vivas D, Parada Alfonso F, Ibanez E, Cifuentes A. Recent applications of high resolution mass spectrometry for the characterization of plant natural products. Trends Analyt Chem 2019;112:87–101.
[7] Alseekh S, Fernie AR. Metabolomics 20 years on: what have we learned and what hurdles remain? Plant J 2018;94:933–42.
[8] da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. Proc Natl Acad Sci U S A 2015;112:12549–50.
[9] Vinaixa M, Schymanski EL, Neumann S, Navarro M, Salek RM, Yanes O. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: state of the field and future prospects. Trends Analyt Chem 2016;78:23–35.
[10] van der Hooft JJJ, de Vos RCH, Ridder L, Vervoort J, Bino RJ. Structural elucidation of low abundant metabolites in complex sample matrices. Metabolomics 2013;9:1009–18.
[11] Wolfender JL, Nuzillard JM, van der Hooft JJJ, Renault JH, Bertrand S. Accelerating metabolite identification in natural product research: toward an ideal combination of liquid chromatography-high-resolution tandem mass spectrometry and NMR profiling, *in silico* databases, and chemometrics. Anal Chem 2019;91:704–42.
[12] Allard PM, Genta Jouve G, Wolfender JL. Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification. Curr Opin Chem Biol 2017;36:40–9.
[13] Zhao X, Zeng Z, Chen A, Lu X, Zhao C, Hu C, et al. Comprehensive strategy to construct in-house database for accurate and batch identification of small molecular metabolites. Anal Chem 2018;90:7635–43.
[14] Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, Sakata A, et al. RIKEN tandem mass spectral database (ReSpect)

for phytochemicals: a plant-specific MS/MS-based data resource and database. Phytochemistry 2012;82:38–45.
[15] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 2010;45:703–14.
[16] Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN: a metabolite mass spectral database. Ther Drug Monit 2005;27:747–51.
[17] Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y, et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. Nat Methods 2018;15:53–6.
[18] Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics 2014;11:98–110.
[19] Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation. J Cheminform 2016;8:3.
[20] Shen X, Wang R, Xiong X, Yin Y, Cai Y, Ma Z, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nat Commun 2019;10:1516.
[21] Kusano M, Yang Z, Okazaki Y, Nakabayashi R, Fukushima A, Saito K. Using metabolomic approaches to explore chemical diversity in rice. Mol Plant 2015;8:58–67.
[22] Okazaki Y, Saito K. Integrated metabolomics and phytochemical genomics approaches for studies on rice. Gigascience 2016;5:11.
[23] Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature 2012;490:497–501.
[24] Zhang J, Luo W, Zhao Y, Xu Y, Song S, Chong K. Comparative metabolomic analysis reveals a reactive oxygen species-dominated dynamic model underlying chilling environment adaptation and tolerance in rice. New Phytol 2016;211:1295–310.
[25] Zhou Q, Fu H, Yang D, Ye C, Zhu S, Lin J, et al. Differential alternative polyadenylation contributes to the developmental divergence between two rice subspecies *japonica* and *indica*. Plant J 2019;98:260–76.
[26] Chen W, Gong L, Guo Z, Wang W, Zhang H, Liu X, et al. A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. Mol Plant 2013;6:1769–80.
[27] Matsuda F, Okazaki Y, Oikawa A, Kusano M, Nakabayashi R, Kikuchi J, et al. Dissection of genotype-phenotype associations in rice grains using metabolome quantitative trait loci analysis. Plant J 2012;70:624–36.
[28] Saito K. Phytochemical genomics–a new trend. Curr Opin Plant Biol 2013;16:373–80.
[29] Gong L, Chen W, Gao Y, Liu X, Zhang H, Xu C, et al. Genetic analysis of the metabolome exemplified using a rice population. Proc Natl Acad Sci U S A 2013;110:20320–5.
[30] Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. J Am Soc Mass Spectrom 1994;5:859–66.
[31] Sokolow S, Karnofsky J, Gustafson P. The Finnigan library search program: Finnigan application report 2. San Jose: Finnigan Corp; 1978.
[32] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
[33] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res 2016;44:D1202–13.
[34] Afendi FM, Okada T, Yamazaki M, Hirai Morita A, Nakamura Y, Nakamura K, et al. KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. Plant Cell Physiol 2012;53:e1.
[35] Allen F, Pon A, Wilson M, Greiner R, Wishart D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res 2014;42:W94–9.
[36] Bocker S. Searching molecular structure databases using tandem MS data: are we there yet? Curr Opin Chem Biol 2017;36:1–6.
[37] Fernie AR, Aharoni A, Willmitzer L, Stitt M, Tohge T, Kopka J, et al. Recommendations for reporting metabolite data. Plant Cell 2011;23:2477–82.
[38] Zhao CL, Yu YQ, Chen ZJ, Wen GS, Wei FG, Zheng Q, et al. Stability-increasing effects of anthocyanin glycosyl acylation. Food Chem 2017;214:119–28.
[39] Teles YCF, Souza MSR, Souza MFV. Sulphated flavonoids: biosynthesis, structures, and biological activities. Molecules 2018;23:480.
[40] Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 2010;42:961–7.
[41] Dong X, Chen W, Wang W, Zhang H, Liu X, Luo J. Comprehensive profiling and natural variation of flavonoids in rice. J Integr Plant Biol 2014;56:876–86.
[42] Galland M, Boutet Mercey S, Lounifi I, Godin B, Balzergue S, Grandjean O, et al. Compartmentation and dynamics of flavone metabolism in dry and germinated rice seeds. Plant Cell Physiol 2014;55:1646–59.
[43] Wang Q, Liu Y, He J, Zheng X, Hu J, Liu Y, et al. *STV11* encodes a sulphotransferase and confers durable resistance to rice stripe virus. Nat Commun 2014;5:4768.
[44] Morreel K, Saeys Y, Dima O, Lu F, Van de Peer Y, Vanholme R, et al. Systematic structural characterization of metabolites in *Arabidopsis* via candidate substrate-product pair networks. Plant Cell 2014;26:929–45.
[45] Nguyen TK, Jamali A, Lanoue A, Gontier E, Dauwe R. Unravelling the architecture and dynamics of tropane alkaloid biosynthesis pathways using metabolite correlation networks. Phytochemistry 2015;116:94–103.
[46] Ding Y, Chang J, Ma Q, Chen L, Liu S, Jin S, et al. Network analysis of postharvest senescence process in citrus fruits revealed by transcriptomic and metabolomic profiling. Plant Physiol 2015;168:357–76.
[47] Li D, Heiling S, Baldwin IT, Gaquerel E. Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. Proc Natl Acad Sci U S A 2016;113:E7610–8.
[48] Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohney RP, Milburn MV, et al. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. PLoS Genet 2012;8:e1003005.
[49] Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst Biol 2011;5:21.
[50] Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003;4:2.
[51] Lan W, Lu F, Regner M, Zhu Y, Rencoret J, Ralph SA, et al. Tricin, a flavonoid monomer in monocot lignification. Plant Physiol 2015;167:1284–95.
[52] Syrchina A, Gorshkov A, Shcherbakov V, Zinchenko S, Vereshchagin A, Zaikov K, et al. Flavonolignans of *Salsola collina*. Chem Nat Compd 1992;28:155–8.
[53] Stermitz FR, Tawara Matsuda J, Lorenz P, Mueller P, Zenewicz L, Lewis K. 5′-Methoxyhydnocarpin-D and pheophorbide A: *Berberis* species components that potentiate berberine growth inhibition of resistant *Staphylococcus aureus*. J Nat Prod 2000;63:1146–9.
[54] Cooper R, Gottlieb HE, Lavie D. A new flavolignan of biogenetic interest from *Aegilops ovata* L.—part I. Isr J Chem 1977;16:12–5.
[55] Pettit GR, Meng Y, Stevenson CA, Doubek DL, Knight JC, Cichacz Z, et al. Isolation and structure of palstatin from the amazon tree *Hymeneae palustris*. J Nat Prod 2003;66:259–62.

[56] Nakajima Y, Yun YS, Kunugi A. Six new flavonolignans from *Sasa veitchii* (Carr.) Rehder. Tetrahedron 2003;59:8011–5.
[57] Ralph J. Hydroxycinnamates in lignification. Phytochem Rev 2010;9:65–83.
[58] Lan W, Morreel K, Lu F, Rencoret J, del Río JC, Voorend W, et al. Maize tricin-oligolignol metabolites and their implications for monocot lignification. Plant Physiol 2016;171:810–20.
[59] Chen F, Tobimatsu Y, Havkin Frenkel D, Dixon RA, Ralph J. A polymer of caffeyl alcohol in plant seeds. Proc Natl Acad Sci U S A 2012;109:1772–7.
[60] Yang Z, Nakabayashi R, Okazaki Y, Mori T, Takamatsu S, Kitanaka S, et al. Toward better annotation in plant metabolomics: isolation and structure elucidation of 36 specialized metabolites from *Oryza sativa* (rice) by using MS/MS and NMR analyses. Metabolomics 2014;10:543–55.
[61] Lan W, Rencoret J, Lu F, Karlen SD, Smith BG, Harris PJ, et al. Tricin-lignins: occurrence and quantitation of tricin in relation to phylogeny. Plant J 2016;88:1046–57.
[62] Duan L, Molnár I, Snyder JH, Ga S, Qi X. Discrimination and quantification of true biological signals in metabolomics analysis based on liquid chromatography-mass spectrometry. Mol Plant 2016;9:1217–20.
[63] Sud M, Fahy E, Cotter D, Brown AH, Dennis EA, Glass CK, et al. LMSD: LIPID MAPS structure database. Nucleic Acids Res 2007;35:527–32.
[64] Oboyle NM, Banck M, James CAJ, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. J Cheminform 2011;3:33.
[65] Matsuda F, Yonekura Sakakibara K, Niida R, Kuromori T, Shinozaki K, Saito K. MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites. Plant J 2009;57:555–77.
[66] van der Kloet FM, Bobeldijk I, Verheij ER, Jellema RH. Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. J Proteome Res 2009;8:5132–41.
[67] Juliane S, Rainer OR, Korbinian S. GeneNet: modeling and inferring gene networks. R package version 2015;1.2.13. Available from: https://CRAN.R-project.org/package = GeneNet.
[68] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.
[69] R Core Team. R: a language and environment for statistical computing. Vienna: R foundation for statistical computing 2018. Available from: https://www.R-project.org/.
[70] Gregory RW, Ben B, Lodewijk B, Robert G, Wolfgang H, Andy L, et al. gplots: various R programming tools for plotting data. R package version 3.0.1, 2016. Available from: https://CRAN.R-project.org/package = gplots.
[71] Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 2016;33:1870–4.