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

Xuetong Li^{1,4,#,a}, Hongxia Zhou^{1,4,#,b}, Ning Xiao^{2,c}, Xueting Wu^{1,d}, Yuanhong Shan^{1,e}, Longxian Chen^{1,4,f}, Cuiting Wang^{1,g}, Zixuan Wang^{1,h}, Jirong Huang^{3,*i}, Aihong Li^{2,*j}, and Xuan Li^{1,4,*k}

¹Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai 200032, China

²Lixiahe Agricultural Research Institute of Jiangsu Province, Yangzhou 225007, China

³Department of Biology, College of Life and Environmental Sciences, Shanghai Normal University, Shanghai 200234, China

⁴University of Chinese Academy of Sciences, Beijing 100039, China

Equal contribution.

* Corresponding authors.

Email: lixuan@sippe.ac.cn (Li X), yzlah@126.com (Li A), huangjr@shnu.edu.cn (Huang J)

Running title: Li X et al / Expanding the Coverage of the Metabolic Landscape

27 ^aORCID: 0000-0003-0029-2296

28 ^bORCID: 0000-0001-9206-2580

29 ^cORCID: 0000-0001-6181-2684

30 ^dORCID: 0000-0002-8644-124X

31 ^eORCID: 0000-0002-2169-7308

32 ^fORCID: 0000-0002-1209-1945

33 ^gORCID: 0000-0002-8251-5774

34 ^hORCID: 0000-0002-4198-7230

35 ⁱORCID: 0000-0002-4032-4566

36 ^jORCID: 0000-0001-6161-9796

37 ^kORCID: 0000-0002-7909-7241

38

39

40 Total word counts (from “Introduction” to “Conclusions” or “Materials and methods”):

41 4856

42 Total references: 71

43 Total figures: 5

44 Total tables: 0

45 Total supplementary figures: 10

46 Total supplementary tables: 13

47 Total supplementary files: 1

48 Total letters in the article title: 96

49 Total letters in running title: 51

50 Total word counts in abstract: 205

51

52 **Abstract**

53 Genome-scale metabolomics analysis is increasingly used for pathway and function
 54 discovery in post-genomics era. The great potential offered by developed mass
 55 spectrometry (MS)-based technology has been hindered by the obstacle that only a
 56 small portion of detected metabolites were identifiable so far. To address the critical
 57 issue of low identification coverage in metabolomics, we adopted a deep
 58 metabolomics analysis strategy by integrating advanced algorithms and expanded
 59 reference databases. The experimental reference spectra, and *in silico* reference
 60 spectra were adopted to facilitate the structural annotation. To further characterize the
 61 structure of metabolites, two approaches, structural motif search combined with
 62 neutral loss scanning, and metabolite association network were incorporated into our
 63 strategy. An untargeted metabolomics analysis was performed on 150 rice cultivars
 64 using Ultra Performance Liquid Chromatography (UPLC)-Quadrupole (Q)-Orbitrap
 65 mass spectrometer. 1939 of 4491 metabolite features in MS/MS spectral tag (MS2T)
 66 library were annotated, representing an extension of annotation coverage by an order
 67 of magnitude on rice. The differential accumulation patterns of flavonoids between
 68 *indica* and *japonica* cultivars were revealed, especially O-sulfated flavonoids. A series
 69 of closely-related flavonolignans were characterized, adding further evidence for the
 70 crucial role of tricin-oligolignols in lignification. Our study provides a great template
 71 in the exploration of phytochemical diversity for more plant species.

72

73 **KEYWORDS:** Untargeted metabolomics; MS/MS spectral tag; Structural
 74 characterization; Phytochemical diversity; Flavonoid derivatives

75

Introduction

It is estimated that there are from 200,000 to 1,000,000 metabolites produced in green plants, underlying their broad chemical diversity and metabolic complexity [1]. Genome-scale metabolomics analysis has become a powerful tool in the elucidation of functional gene and pathway for diverse phytochemicals [2-5]. The more recent progresses in UPLC coupled with high-resolution MS, allow detecting metabolites at unparalleled sensitivity, resolution, accuracy, and throughput [6]. However, the great power in advanced liquid-phase separation and mass spectrometry technology has been limited, considering a vast majority of metabolite features detected from plants remain unidentified in current status [7, 8]. It is a major challenge to detect and identify the massive amount of heterogeneous phytochemicals with high dynamic range in concentrations, chemical and physical properties, and structures. The lagging in 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 mass spectral data [9-12]. It is critical to handle and resolve the metabolomics data efficiently, in order to bridge the gap between technological advance and demands of plant metabolomics research. In recent years, progresses have been made in the improvement of 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, and it is critical to explore its chemical compositions and metabolic traits for the enhancement of grains 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 were performed, which provides the foundation for the metabolic components of rice [2, 5, 26, 27]. However, there are plenty of unknown metabolite features in above studies and the metabolic diversity of rice still needs further efforts to explore. Other studies focused on phytochemical genomics to dissect the underlying genetics basis of biosynthesis and physiological function of metabolites

109 during the evolution and adaptation of plants [28]. The metabolic quantitative trait
110 loci (mQTL) mapping and metabolic genome-wide association study (mGWAS) were
111 used to reveal the genetic polymorphisms and candidate genes that affected metabolic
112 traits in rice [2, 5, 27, 29].

113 Our current study was designed to address the key issue in plant metabolomics,
114 that is, the low identification coverage of metabolites. We sought to expand the
115 annotation coverage with computational approaches, by adopting a deep
116 metabolomics analysis strategy that combines experimental and *in silico* reference
117 mass spectral libraries, and advanced algorithms. The structural motif search
118 combined with neutral loss scanning and metabolite association network methods
119 were integrated in our strategy to facilitate the characterization of structure and
120 potential function of novel metabolites without reference from above libraries. As a
121 proof-of-concept study, using state-of-the-art UPLC-Q-Orbitrap mass spectrometer
122 platform, we performed an untargeted metabolomics analysis on a core collection of
123 150 *indica* and *japonica* cultivars grown in northeastern and southeastern China. A
124 MS2T library for rice grains was constructed containing 4491 metabolite features, and
125 of which 1939 were annotated. The annotation coverage of rice metabolome was
126 significantly improved through our strategy. Further, our analyses revealed the
127 systematic difference of metabolomes between *indica* and *japonica* subspecies and
128 major differential accumulation patterns of flavonoid derivatives, especially
129 O-sulfated flavonoids. A group of closely-related flavonolignans were newly uncover
130 in rice, which provided further evidence for the crucial role of tricin-oligolignols in
131 lignification of monocots. Our deep metabolomics analysis strategy expanded our
132 understanding of phytochemical diversity and function in rice, which has profound
133 implication for improving the quality and nutritional value of crops through genetic
134 breeding.

135

136

137 **Results and discussion**

138 **Integrated computational approaches and their evaluations**

139 To handle the mass spectral data generated from UPLC-Q-Orbitrap mass spectrometer,
140 we adopted a deep metabolomics analysis strategy with integrated computational
141 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 above libraries, using structural motif search combined with neutral loss scanning and metabolite association network (*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] (*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 data for annotating those metabolites without a hit in the first approach. The *in silico* mass spectra was generated from in-house structural database (Structural Database of Biologically Relevant Compounds, SDBRC, Table S1) that contains the structural information of over 80,000 biologically relevant compounds collected from KEGG [32], PubChem [33], and KNApSack [34] databases (*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 above approaches, we sampled experimental mass spectra from Metlin and Massbank as query sets (Table S2). In first approach using experimental mass spectra as reference, INCOS had identification rates from 75 to 79% for top 1 match, and 96 to 97% when top 5 matches were included, which are comparatively higher than those of NDP (Figure S1A). In second approach using *in silico* mass spectra as 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 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 mass spectrometry analysis. The identification rate will drop when we search against larger reference database, mainly due to more isomers contained in database [36]. SDBRC contains more biologically relevant compounds

176 and will provide valuable reference structural information for more metabolite
177 features. The combination of this two approaches will greatly expand the annotation
178 coverage of plant metabolomes, and is instrumental in our study on the exploration of
179 phytochemical diversity and function in rice.

180

181 Constructing and annotating MS2T library for cultivated rice

182 To construct a MS2T library for metabolomics analysis of rice grains, we used a
183 collection of 150 representative rice accessions (Table S3). Rice grains were harvested
184 from farm lands in southeastern and northeastern China, and were mixed (referred as
185 reference mixture) for subsequent processing. The extracts were subjected to
186 UPLC-Q-Orbitrap mass spectrometer (*Methods*). The raw data from repeated analyses
187 were aligned using Compound Discover software (v2.0, Thermo Scientific). 158,840
188 and 118,077 signals detected from positive and negative modes were grouped to
189 11,263 and 6495 merged compound features, respectively. After the quality control
190 and redundancy filtering steps, 2637 and 2446 metabolite features were retained for
191 positive and negative modes, respectively, in which 2234 and 2123 were tagged with
192 MS2 spectra. Finally, these metabolite features from positive and negative modes
193 were merged, resulting in 4491 metabolite features with 3832 tagged with MS2
194 spectra (Figure S2 and Tables S4, S5). These metabolite features in rice MS2T library
195 were then annotated with our deep metabolomics analysis strategy (*Methods*). 298
196 metabolite features were annotated using experimental mass spectra as reference. For
197 rest 3534 metabolite features, 1641 were annotated using *in silico* mass spectra as
198 reference. Taken together, 1939 metabolite features were annotated in MS2T library
199 for rice grains (Table S5). The MS2T library constructed by our study was reported as
200 recommended [37] (Tables S4 and S5).

201 Benefit from the high-resolution MS and deep metabolomics analysis strategy
202 with integrated computational approaches, we expanded the metabolite annotation
203 coverage of rice cultivars in comparison with previous studies [2, 5, 26, 27].
204 Flavonoids account for a large portion of increase of annotated metabolites in rice
205 grains. The flavonoids annotated in our study display various modifications, such as
206 glycosylation, acetyl-glycosylation, and sulfation. The glycosylation contains
207 monoglycoside, diglycoside, and hexuronide. Examples include RSM04010p
208 (quercetin-3-glucoside), RSM04966p (isovitexin-7-O-xyloside), RSM05128p

(apigenin-7-O-gentiobioside), RSM05322p (demethoxycentaureidin-7-O-rutinoside), and RSM02409n (apigenin 4'-glucuronide) (**Figures 2A,B,C,D,E**). 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) (**Figures 2G,H,I**). For sulfated flavonoids, an uncommon type of flavonoids, we found RSM02011n (ombuin 3-O-sulfate) (**Figure 2F**). 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 benefit to the exploration of molecular mechanism of metabolite modifications in the growth, development and interaction with the environment of plants.

221

222 **Differential metabolic profiles analysis revealing the featured metabolites of** 223 ***indica* and *japonica* cultivars**

224 To characterize the metabolic profiles of grains for diverse rice cultivars and
225 understand their natural variation, we performed the untargeted metabolomics
226 analysis on 59 rice cultivars, including 40 *indica* and 19 *japonica* (*Methods*). The
227 metabolic profiles of rice cultivars contain the relative abundance of 3409 metabolite
228 features (Table S6). The metabolic profiles of 59 rice cultivars were clustered based
229 on the relative abundance of 3409 metabolite features, which displayed the
230 differential patterns between *indica* and *japonica* cultivars (**Figure 3A**). In tree view
231 (**Figure 3B**), the relation between *indica* with *japonica* cultivars is generally
232 consistent to the phylogenetic relationship [40]. Through principal component
233 analysis (PCA), *indica* and *japonica* cultivars were separated by first component (PC1)
234 and second component (PC2), indicating the systematic difference in metabolic
235 profiles between two subspecies (**Figure 3C**). We further performed orthogonal
236 partial least squares discriminate analysis (OPLS-DA) to investigate the featured
237 metabolites that differentiate *indica* and *japonica* cultivars. The *indica* and *japonica*
238 cultivars were separated in two distinct clusters with our OPLS-DA model (**Figure**
239 **4A**). Metabolites with variable importance in projection (VIP) value greater than 2.5,
240 were defined as featured metabolites in our study. Among 58 featured metabolites
241 (Table S7), 11 flavonoids, 3 terpenoids, and 2 phenylpropanoids were annotated.
242 Particularly, three novel triclin derivatives, RSM03724n (triclin-O-sulfatohexoside),

243 RSM04661n (tricin-O-acetylramnoside-O-diacetylramnoside), and RSM05814p
244 (tricin-O-feruloylhexoside-hexoside) (Figures S3A,B,C and Table S7), were
245 characterized using structural motif search combined with neutral loss scanning
246 (*Methods*).

247 We further observed the differential accumulation patterns of C-glycosylated,
248 O-glycosylated, and O-sulfated flavonoid derivatives among featured metabolites.
249 The levels of four C-glycosylated flavonoids (flavone C-hexoside and flavone
250 C-pentoside), RSM03824p (cytisoside), RSM04142p (precatorin I), RSM03991p
251 (trihydroxy-methoxyflavone C-hexoside) (Figure S3D), and RSM04767p
252 (di-C,C-pentosyl-apigenin) (Figure S3E) are significantly higher in *indica* than
253 *japonica* cultivars (**Figure 4B** and Table S7). In contrast, the levels of four
254 O-glycosylated flavonoids with guaiacylglyceryl or acyl modification, RSM05526p
255 (tricin 4'-O-(guaiacylglyceryl) ether 7"-O-glucopyranoside), RSM05648p (isovitexin
256 7-O-(6"-O-E-p-coumaroyl)glucoside), RSM04661n(tricin
257 O-acetylramnoside-O-diacetylramnoside), and RSM05814p (tricin
258 O-feruloylhexosyl-O-hexoside) are significantly higher in *japonica* than *indica*
259 cultivars (**Figure 4C** and Table S7). Furthermore, differences in two O-sulfated
260 flavonoids, RSM02011n (ombuin 3-O-sulfate) and RSM03724n (tricin
261 O-sulfatohexoside), were observed between *indica* and *japonica* cultivars (**Figure 4D**
262 and Table S7). The differential accumulation patterns of C-glycosylated and
263 O-glycosylated flavonoids in rice grains were consistent with previous studies in rice
264 leaves [26, 41]. Additionally, we expanded our findings to O-sulfated flavonoids, an
265 uncommon variety of flavonoid derivatives catalyzed by sulfotransferases [39, 42]. It
266 has been revealed that the natural variation of salicylic acid sulfotransferase encoding
267 gene cause the differentiation of resistance to rice stripe virus between *indica* and
268 *japonica* subspecies [43], highlighting the significant role of sulfation in pathogen
269 resistance of rice. However, rare study has been performed to characterize flavonoid
270 sulfotransferases in rice. The differential accumulation patterns of O-sulfated
271 flavonoids revealed by our study provided new insight to the natural variation of
272 flavonoid sulfotransferase activity, which is benefit to the exploration of biosynthesis
273 genes of flavonoid sulfotransferases and their potential functions in pathogen
274 resistance of rice.

275

276 **Constructing metabolite association network and uncovering diverse** 277 **flavonolignans from rice grains**

278 Network-based analysis is widely used in metabolomics studies for understanding of
279 metabolite interaction, structural characterization, and pathway elucidation [20,
280 44-47]. Previous studies suggested that metabolites with similar structure generally
281 display correlation in their abundance, so the structure of unknown metabolites can be
282 inferred by knowns through metabolite association network [2, 4, 27, 48].

283 We constructed the metabolite association network with Gaussian graphical
284 model (GGM) [49], using the metabolic profiles of 59 rice cultivars (*Methods*, Figure
285 S4A and Table S8). This network contains 2874 nodes (metabolites) with 42,147
286 significant edges (metabolite pairs). The 64 clusters were isolated (Table S9) from
287 GGM network using Molecular Complex Detection (MCODE) program [50]
288 (*Methods*). A subgroup of the first-ranked cluster mainly contains flavonoids. Besides,
289 within the second-ranked cluster, a large number of nodes were annotated as
290 terpenoids, most of which are triterpenoids (Figure S4B and Table S10).

291 A subgroup of the first-ranked cluster contains 32 metabolites (**Figure 5A** and
292 Table S10). 13 of them were annotated as common flavonoids with hydroxy and
293 methoxy groups (Figure S5). Notably, within this cluster, we found some
294 flavonolignans (**Figure 5B** and Figure S6), which are produced via oxidative coupling
295 between flavonoids with three varieties of monolignols, p-coumaryl, coniferyl, and
296 sinapyl alcohols [51]. RSM04702p (Salcolin B) [52] and RSM04355p
297 (5'-Methoxyhydnicarpin-D) [53] are guaiacyl flavonolignans, and RSM04382p
298 (aegicin) is p-hydroxyphenyl flavonolignans [54]. Based on above findings, we
299 suggested that there are other flavonolignans within this cluster. We then observed the
300 precursor ion and fragmentation pattern of unknowns within this cluster and
301 characterized more flavonolignans. The RSM04691p displays same fragment ion (m/z
302 at 315.04895) with RSM04355p in mass spectra, which means they have same
303 flavonoid moiety in structures. The mass difference between their precursor ions is
304 30.01031, corresponding to a methoxy group. Thus the RSM04691p has an additional
305 methoxy group at the coniferyl alcohol moiety of RSM04355p, which was
306 characterized as palstatin [55], a syringyl flavonolignan. With same method, the
307 structures of RSM05474p, RSM05479p, RSM05574p, and RSM04546n were
308 characterized. The RSM05474p was characterized as triclin
309 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-oligolignols trimers, which are formed by further chain extension through oxidative coupling between tricin-oligolignols dimer (RSM04702p) and p-coumaryl alcohol or coniferyl alcohol via ether or furan bridge (Figure 5B and Figures S6E,F,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 (Figures S6I,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 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-oligolignols 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 and Table S10).

337

338 **Conclusions**

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 was 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 were uncovered. Our strategy can be applied to the metabolomics researches of other agronomically important plants, with great potential in the enhancement of crops quality and nutrition value through genetic breeding.

356

357 **Materials and methods**

358 **Plant materials**

359 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 360 in 2015 and 2016 from two locations of China: farm lands in Jiangsu (Yangzhou, E 361 119°53', N 32°42', southeastern China) and in Heilongjiang (Harbin, E 126°53', N 362 45°69', northeastern China). Rice cultivars were planted in the field, ten plants for 363 each row, and three rows for each accession.

364 For each accession, grains were harvested for two biological replicates, each 365 containing grains from three individual plants. Grains were packed in gauze bag and 366 air-dried in shade. Two grams of dried grains were ground using tissue grinder 367 (catalog No. 05010997; Shanghai BiHeng Biotechnology Company Limited; 368 Shanghai; China) at 55Hz for 40 seconds. The fine powder for each accession was 369 stored at -80°C for subsequent processing.

370

371 **Chemicals**

372 HPLC grade methanol, acetonitrile and acetic acid were obtained from Merck 373 Company (catalog No. 1.06007.4008, 1.00030.4008, 5.43808.0250; Merck KGaA; 374 Darmstadt; Germany). Ultra-pure water was produced using Millipore water purifier 375

(Milli-Q; Millipore; Billerica; MA; USA). The lidocaine and lincomycin (CAS No. 137-58-6, 859-18-7; Dr. Ehrenstorfer, GmbH; Augsburg; Germany) were purchased from ANPEL Laboratory Technologies (Shanghai) Inc.. Other chemicals were purchased from Sigma-Aldrich (Shanghai) Trading Co., Ltd. (Sigma-Aldrich; Merck KGaA; Darmstadt; Germany), if not otherwise specified.

381

382 **Metabolite extraction**

383 150 mg of powder of rice grains was mixed with 1.5 mL 70% aqueous methanol
384 solution A (containing 1 mg/L vitexin, 1 mg/L p-coumaric acid, and 1 mg/L lidocaine
385 as internal standards). The mixture was vortexed every 10 min for 3 times and placed
386 in 4°C refrigerator overnight. The mixture was then centrifuged at 12,000g for 10 min
387 in 4°C. The supernatant of mixture was dried with concentrator under vacuum and
388 re-dissolved with 150 uL 70% aqueous methanol solution B (containing 1 mg/L
389 capsaicin and 1 mg/L lincomycin as internal standards). Then the extract was filtered
390 with 0.22µm filter (catalog No. SCAA-104; ANPEL Laboratory technologies Inc;
391 Shanghai; China) and transferred into sample bottle for the subsequent UPLC-MS/MS
392 analysis.

393

394 **UPLC-MS/MS analysis**

395 Chromatographic separation of extract samples was performed on Waters Acquity
396 Ultra Performance Liquid Chromatography using an ACQUITY UPLC BEH C18
397 column (pore size: 1.7µm, length: 2.1*100mm) (Waters Corporation; Milford; MA;
398 USA). The mobile phase consisted of (A) water with 0.04% acetic acid and (B)
399 acetonitrile with 0.04% acetic acid. The gradient program was as follows: 95:5 A/B at
400 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
401 A/B at 30.0 min. The flow rate was 0.25 mL/min and the injection volume was 5 µL.
402 The column temperature was 40°C.

403 The UPLC system was coupled with Q Exactive™ Hybrid Quadrupole-Orbitrap
404 High Resolution Mass Spectrometer (Q-Orbitrap-HRMS) (Thermo Fisher Scientific;
405 Waltham; MA; USA). The MS acquisition was performed in positive and negative
406 ionization with FullScan/dd-MS2 (top 8) mode, in which the MS/MS spectra of most
407 abundant ions (top 8) within each scanning window was automatically obtained. MS
408 full scan mass resolution was set to 70,000 at m/z 200 and data-dependent MS/MS

409 with full scan mass resolution was reduced to 17,500 at m/z 200. The m/z range of
410 MS full scan was 100-1000.

411 Heated electrospray ionization (HESI) parameters were as follows: Spray voltage
412 (+), 4000 V; Spray voltage (-), 3500 V; Capillary temperature, 320°C; Sheath gas, 35
413 arb; Aux gas, 8 arb; Probe heater temperature, 350°C; S-Lens RF level, 50. Higher
414 energy Collisional Dissociation (HCD) energies were 15eV and 40eV, and average
415 MS/MS spectrum was obtained. The mass spectrometer was calibrated using Pierce™
416 LTQ Velos ESI positive ion Calibration Solution and Pierce™ ESI negative ion
417 Calibration Solution (Thermo Fisher Scientific; Waltham; MA; USA).

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

423

424 **Mass spectrum data processing**

425 The raw data generated from HESI-Q-Orbitrap-HRMS were processed with
426 Compound Discoverer software (version 2.0, Thermo Scientific) using its automatic
427 workflow. The retention time aligning parameters were as follows: mass tolerance,
428 5ppm; maximum Shift, 0.5min. The unknown compounds detecting parameters were
429 as follows: min peak intensity, 2E6; S/N threshold, 5.

430 Raw metabolite features were further filtered by: 1) removing signals that are of
431 poor quality or non-biological origin [62], i.e. features with reproducibility <90%,
432 sample to blank ratio <10%, relative standard deviation (RSD) >50%, or peak area
433 less than 1E5; 2) removing redundancy from multi-ion adducts (Na^+ , K^+ , NH_4^+ , Cl^-),
434 isotopes, in-source fragmentation, or dimerization. Metabolite features in positive
435 ($[\text{M}+\text{H}]^+$) and negative ($[\text{M}-\text{H}]^-$) modes were merged with following parameters:
436 exact mass tolerance, 5ppm; and retention time tolerance, 0.5min. The in-house script
437 based on Xcalibur Development Kit (XDK) in Xcalibur software (version 2.2,
438 Thermo Scientific) was used to automatically extract MS2 spectra of metabolite
439 features.

440

441 **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 ReSpec [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 ReSpec 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 were compared for similarity using INCOS. Reference mass spectra with similarity score >0.75 was 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 first annotation approach (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 and 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 manual 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 (SMILES), and the IUPAC International Chemical Identifier (InChI). Through merging compounds from different databases and removing redundancy, we constructed a structural database that contains 85,342 non-redundant compounds (SDBRC), which was used as 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 machine learning technique with probabilistic generative model for compound fragmentation process. The source code for CFM-ID software (version 2.0) was obtained from SourceForge platform (<https://sourceforge.net/projects/cfm-id/>), and compiled on 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 query and *in silico* mass spectra were calculated. Reference compounds with 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 and 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 and 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. In order 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 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). In addition, 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 were listed in Figure S10.

The metabolic profiles of rice cultivars

The metabolic profiles of 59 rice cultivars grown in Yangzhou in 2016 was obtained from the corresponding peak areas of raw mass spectrometric data using Compound Discoverer software (v2.0, Thermo 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, 0.35min; and the tolerance of mass, 5ppm. 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 log₂-transformed for further analysis. The detailed steps of the acquisition and processing of relative abundance data of metabolites were listed in Appendix S1.

Construction of GGM-based network

For construction of 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 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 GGM network. The Cytoscape software [68] was used for the visualization of GGM network. The MCODE application was used to find clusters from GGM network with parameters as defaulted [50].

Statistical analysis

R software (version 3.2.3; <https://www.R-project.org/>) [69] was mainly used for

statistical analysis, if not specifically indicated otherwise. The metabolic profiles of rice cultivars were clustered using hierarchical clustering. The method of hierarchical clustering is Unweighted Pair Group Method with Arithmetic mean (UPGMA). The heatmap was constructed by heatmap.2 function in gplots package (<https://CRAN.R-project.org/package=gplots>)[70]. The relatedness distance between 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 (version 14.0; Umetrics, Sweden).

Data availability

The data supporting the findings of this study are available from the supplementary materials.

Authors' contributions

X (Xuan) L conceived and designed the project. X (Xuetong) L, HZ, NX,YS, LC, CW and AL prepared samples and performed LC-MS analysis experiments. X (Xuetong) L, HZ, XW, and YS performed bioinformatics analyses. JH and ZW advised on rice metabolomics experiment and data analysis. X (Xuan) L and X (Xuetong) L wrote and edited the manuscript. All authors approved the manuscript.

Competing interests

The authors have declared no competing financial interests.

Acknowledgements

We thank Prof. Jie Luo for assistance in sample preparation in LC-MS analysis, and Ms. Ping Chen for help with data submission. This work was supported by grants from the Special Fund for Strategic Pilot Technology of Chinese Academy of Sciences [grant number XDA24010403], the National Key Research and Development Program of China [grant number 2018YFA0900700], the Strategic Project for Biological Resources and Service Network of Chinese Academy of Sciences [grant number ZSYS-014], the Major Project of Jiangsu Province of China for Significant

576 New Varieties Development [grant number PZCZ201702], the National Natural
577 Science Foundation of China [grant numbers 31900470, 31701137, 31771412].
578

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. *TrAC Trends in Analytical Chemistry* 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. *TrAC Trends in Analytical Chemistry* 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. *Therapeutic Drug Monitoring* 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* 2018.

[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. *Proceedings of the National Academy of Sciences of the United States of America* 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. Finnigan Corp. San Jose, CA, March 1978.

[32] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 2000;28:27–30.

[33] Bryant S. PubChem: An information resource linking chemistry and biology. *Abstracts of Papers of the American Chemical Society* 2006;231.

[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 and Cell Physiology* 2011;53:e1–e.

[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.

[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. *Journal of Integrative Plant Biology* 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. *Proceedings of the National Academy of Sciences of the United States of America* 2016;113:E7610–E8.

[48] Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohny 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*. *Chemistry of Natural Compounds* 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*. *Journal of natural products* 2000;63:1146–9.

[54] Cooper R, Gottlieb HE, Lavie D. A new flavolignan of biogenetic interest from *Aegilops ovata* L.—Part I. *Israel Journal of Chemistry* 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 *Hymenaea palustris*. *Journal of natural products*

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. *Phytochemistry Reviews* 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 Physiology* 2016;pp.02012.2016.

[59] Chen F, Tobimatsu Y, Havkin Frenkel D, Dixon RA, Ralph J. A polymer of caffeoyl 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, Shen Ga, 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 Research* 2007;35:527–32.

[64] Oboyle NM, Banck M, James CAJ, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* 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 1.2.13. <https://CRAN.R-project.org/package=GeneNet> 2015.

[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 Research* 2003;13:2498–504.

[69] R Core Team. R: A Language and Environment for Statistical Computing. R foundation for statistical computing, Vienna, Austria. <https://www.R-project.org/> 2018.

[70] Gregory R. Warnes BB, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables. gplots: Various R programming tools for plotting data. R package version 3.0.1. <https://CRAN.R-project.org/package=gplots> 2016.

[71] Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 2016;33:1870–4.

750 **Figure legends**

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

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

766

767 **Figure 2** The mass spectra of annotated flavonoids with diverse modifications
768 $[M+H]^+$ and $[M-H]^-$ indicate the protonated and deprotonated precursor ion of
769 flavonoids, respectively. RSM*****p/n indicates the serial number of rice's screening
770 mass spectra acquired in positive or negative ion mode. **A.** RSM04010p
771 (isoquercitrin): the m/z 303.04922 is the featured protonated ion of quercetin, and the
772 neutral loss of m/z 162.1087 corresponds to a hexoside group. **B.** RSM04966p
773 (isovitexin-7-O-xyloside): the m/z 313.07016 and 433.11319 are the featured
774 protonated ions of isovitexin (the neutral loss of m/z 120.043 is the characteristic of
775 C-hexosyl flavonoids), and the neutral loss of m/z 132.0404 corresponds to a
776 pentoside group. **C.** RSM05128p (apigenin-7-O-gentiobioside): the m/z 271.05936 is
777 the featured protonated ion of apigenin, and the neutral loss of m/z 324.1032
778 corresponds to two hexoside groups. **D.** RSM05322p
779 (demethoxycentaureidin-7-O-rutinoside): the m/z 315.04944 and 331.08078 are the
780 featured protonated ions of demethoxycentaureidin, and the neutral loss of m/z
781 146.0587 corresponds to a deoxyhexoside (rhamnoside) group. **E.** RSM02409n
782 (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 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 331.08017 are the featured protonated ions of triclin, 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 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.

795

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 abundance of 3409 metabolites. **B.** The Neighbor-joining tree of 59 rice cultivars based on the relative abundance of 3409 metabolites. **C.** The score plot for PCA of 59 rice cultivars based on the relative abundance of 3409 metabolites. The first and second principle components account for 26.4% and 18.9% of variance, respectively.

802

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 abundance of 3409 metabolites. The $t[1]$ and $to[1]$ indicate the predictive component and orthogonal component of OPLS-DA model, respectively. The R^2X , R^2Y (goodness-of-fit parameter), and Q^2 (predictive ability parameter) of OPLS-DA model are 0.555, 0.99 and 0.98. **B.** The boxplot of relative abundance of four C-glycosylated flavonoids among featured metabolites. RSM03824p, cytoside; RSM03991p, trihydroxy-methoxyflavone C-hexoside; RSM04142p, precatorin I; RSM04767p, di-C,C-pentosyl-apigenin. **C.** The boxplot of relative abundance of four O-glycosylated flavonoids among featured metabolites. RSM04661n, triclin O-acetylramnoside-O-diacetylramnoside; RSM05526p, triclin 4'-O-(guaiacylglyceryl) 7''-O-glucopyranoside; RSM05648p, isovitexin 7-O-(6'''-O-E-p-coumaroyl) glucoside; RSM05814p, triclin O-feruloylhexoside-O-hexoside. **D.** The boxplot of relative abundance of two

O-sulfated flavonoids among featured metabolites. RSM02011n, ombuin 3-O-sulfate;
RSM03724n, tricin O-sulfatohexoside.

819

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
flavonolignans 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
flavonolignans trimers derived from the further chain extension between RSM04702p
with monolignols.

832

833 **Supplementary material**

Figure S1 The performance evaluation of two annotation approaches

A. Evaluating the performance of annotation with experimental mass spectra as
reference. Query mass spectra were sampled from Metlin or Massbank, and spectral
similarity was scored with NDP or INCOS algorithm. **B.** Evaluating the performance
of annotation with *in silico* mass spectra as reference. Query mass spectra were
sampled from Metlin or Massbank, and spectral similarity was scored by CFM-ID
software.

Figure S2 The workflow for MS2T library construction

Sample preparation, metabolite extraction, UPLC-HRMS analysis, and mass spectral
data processing were as described (*Methods*).

Figure S3 The mass spectra of several featured metabolites of *indica* and *japonica*
cultivars

846 $[M+H]^+$ and $[M-H]^-$ indicate the protonated and deprotonated precursor ion of
847 metabolites, respectively.

848 **A.** The mass spectra of RSM03724n (tricin O-sulfatohexoside). The m/z 313.03586
849 and 329.06653 are the featured deprotonated ions of triclin, and the neutral losses of
850 m/z 79.95697 and 162.0536 correspond to the sulfate and hexoside groups,
851 respectively. **B.** The mass spectra of RSM04661n (triclin
852 O-acetylramnoside-O-diacetylramnoside). The neutral loss of m/z 418.1954
853 corresponds to the acetylramnoside and diacetylramnoside groups. **C.** The mass
854 spectra of RSM05814p (triclin O-feruloylhexoside-O-hexoside). The m/z 315.04944
855 and 331.08084 are the featured protonated ions of triclin, and the neutral loss of m/z
856 500.1448 corresponds to the feruloylhexoside and hexoside groups. The m/z
857 177.05428 is the featured protonated ion of feruloyl unit. **D.** The mass spectra of
858 RSM03991p (trihydroxy-methoxyflavone C-hexoside). The neutral loss of 120.0449
859 is the characteristic of C-hexosylflavones. **E.** The mass spectra of RSM04767p
860 (di-C,C-pentosyl-apigenin). The neutral loss of 90.05359 is the characteristic of
861 C-pentosylflavones.

862 **Figure S4** The diagram of metabolite association network

863 **A.** The metabolites association network of rice grains according to the metabolic
864 profile of 59 rice cultivars. **B.** The second ranked cluster that mainly consist of
865 terpenoids.

866 **Figure S5** The structure of flavonoids within the subgroup of first-ranked cluster
867 These flavonoids have diverse numbers of hydroxyl and methoxyl groups in their
868 structures.

Figure S6 The structure and mass spectra of characterized flavonolignans

[M+H]⁺ and [M-H]⁻ indicate the protonated and deprotonated precursor ion of flavonolignans, respectively.

A. RSM04382p, aegicin, which has a structure of tricin moiety with a p-coumaroyl alcohol linked by an ether bond. The m/z 315.04950 and 331.08084 are featured protonated ions of tricin, and the neutral loss of m/z 166.0628 corresponds to the p-hydroxyphenylglyceryl unit. **B.** RSM04355p, 5'-Methoxyhydnocarpin-D, which has a structure of methoxyluteolin moiety with a coniferyl alcohol linked by a dioxane bridge. The m/z 315.04874 is featured protonated ion of methoxyluteolin moiety, and the neutral loss of m/z 180.0789 corresponds to the coniferyl alcohol unit. **C.** RSM04702p, salcolin B, which has a structure of tricin moiety with a coniferyl alcohol linked by an ether bond. The neutral loss of m/z 196.0736 corresponds to the guaiacylglyceryl unit. **D.** RSM04691p, palstatin, which has a structure of methoxyluteolin moiety with a sinapyl alcohol linked by a dioxane bridge. The neutral loss of m/z 210.0891 corresponds to the sinapyl alcohol unit. **E.** RSM05479p, which has a structure of salcolin B moiety with a p-coumaroyl alcohol linked by a furan bridge (characterized through its mass spectra). **F.** RSM04546n, which has a structure of salcolin B moiety with a coniferyl alcohol linked by an ether bond (characterized through its mass spectra). **G.** RSM05574p, which has a structure of salcolin B moiety with a coniferyl alcohol linked by a furan bridge (characterized through its mass spectra). **H.** RSM05474p, tricin O-[guaiacyl-(O-p-coumaroyl)-glyceryl] ether, which has an additional coumaroyl unit

at the guaiacylglyceryl group of salcolin B. The neutral loss of m/z 342.1071 corresponds to the O-p-coumaroyl-guaiacylglyceryl group, and the m/z 147.04379 is featured protonated ion of p-coumaroyl unit. **I.** RSM04164p, putative catechyl-type flavonolignans characterized through its mass spectra, which has a structure of dihydroxy-dimethoxyflavone moiety with a caffeyl alcohol linked by a dioxane bridge. The m/z 313.07019 is featured protonated ion of dihydroxy-dimethoxyflavone moiety, and the neutral loss of m/z 166.0626 corresponds to the caffeyl alcohol unit. **J.** RSM04201p, putative catechyl-type flavonolignans characterized through its mass spectra, which has a structure of tetrahydroxy-methoxyflavone moiety with a caffeyl alcohol linked by a dioxane bridge. The m/z 315.04919 is featured protonated ion of tetrahydroxy-methoxyflavone moiety.

Figure S7 The mass spectra of 6 putative tricin derivatives

[$M+H$]⁺ and [$M-H$]⁻ indicate the protonated and deprotonated precursor ion of putative tricin derivatives, respectively (see Table S10 for details).

Figure S8 The performance evaluation of the annotation approaches with test set of Fiehn HILIC Library

The test set used for performance evaluation was collected from the MassBank of North America (see Table S11 for details). **A.** The percentages of top 3 correct annotation of the first annotation approach (INCOS). The evaluations were performed with Metlin and MassBank as reference, and the cutoff of similarity score was 0.75. **B.** The percentages of top 3 correct annotation of the second annotation approach. The evaluations were performed with the *in silico* mass spectra generated from KEGG and

913 SDBRC database as reference, and the cutoff of similarity score was 0.3.

914 **Figure S9** The matching results of MS/MS spectra between 17 metabolite features
915 with standard compounds

916 The MS/MS spectra for standard compounds were acquired by the same metabolic
917 analysis methods with metabolite features in MS2T library. The detailed information
918 for identification of 17 metabolite features were listed in Table S12.

919 **Figure S10** The detailed steps of structural motif search combined with neutral loss
920 scanning

921 1) The raw structure data of flavones and flavonols were collected from the
922 LIPID MAPS Structure Database. The OpenBabel software was used to convert
923 the raw structure data into the machine-readable structural information, including
924 formula, exact mass, Simplified Molecular-Input Line-Entry System (SMILES),
925 and the IUPAC International Chemical Identifier (InChI). The non-redundant
926 structure data for 3145 flavones and flavonols were obtained through merging the
927 compounds with identical structure.

928 2) The theoretical MS/MS spectra for flavones and flavonols were predicted by
929 CFM-ID software with structure data as input.

930 3) The fragments with top 25% of intensity in each mass spectra were retained
931 for further analysis. The fragments with lower structure similarity (calculated by
932 OpenBabel software) with the diphenylpropane backbone (C6-C3-C6) of
933 flavonoids, were filtered out as non-aglycone derived fragments.

934 4) The remaining fragments with identical mass were merged and sorted by
935 their frequency. The merged fragments were considered as the characteristic
936 fragments of flavones and flavonols. The fragments with comparatively high
937 frequency represent the structural motifs frequently found in flavones and
938 flavonols, such as m/z at 287.0550145 (featured ion of kaempferol derivatives),
939 m/z at 303.0499291 (featured ion of quercetin derivatives), m/z at 271.0600999

940 (featured ion of apigenin derivatives), and m/z at 301.0706646 (featured ion of
941 chrysoeriol derivatives), *etc.*

942 5) The mass difference between precursor ion and characteristic fragments in
943 each MS/MS spectra were calculated as neutral losses. The neutral losses with
944 similar mass (tolerance: 10ppm) were merged and sorted by their frequency. The
945 merged neutral losses were considered as the featured neutral losses of flavones
946 and flavonols. The neutral losses with comparatively high frequency represent the
947 modifications frequently occurred in flavones and flavonols, such as the neutral
948 losses of hexoside (m/z 162.0530308), pentoside (m/z 132.0423309), rhamnoside
949 (m/z 146.0576808), hexuronide (m/z 176.0322455), sulfate (m/z 79.9568149),
950 and coumaroylhexoside (m/z 308.0892455) groups, *etc.*

951 6) The characteristic fragments and featured neutral losses were used as the
952 queries to search unknown MS/MS spectra for structural characterization. We
953 firstly searched for the presence of the characteristic fragments (mass tolerance:
954 10ppm) in unknown MS/MS spectra. And if there were characteristic fragments,
955 the mass difference between precursor ion and each characteristic fragment was
956 calculated. If the mass difference was similar (mass tolerance: 10ppm) with
957 featured neutral losses, the unknown MS/MS spectra was putatively characterized
958 as flavones or flavonols.

959 7) The putatively characterized structures were further inspected with other
960 available phytochemical information to improve the confidence of structural
961 characterization.

962 **Table S1** Structural Database of Biologically Relevant Compounds (SDBRC)

963 **Table S2** The information of test sets

964 **Table S3** The information of rice varieties

965 **Table S4** Metabolite Reporting Checklist

966 **Table S5** The MS2T library of rice grains

967 **Table S6** The relative abundance data of 3409 metabolites for 59 rice cultivars

968	Table S7	The information of featured metabolites of indica and japonica cultivars
969	Table S8	The Gaussian graphical model (GGM)-based network
970	Table S9	The 64 clusters within GGM network
971	Table S10	The information of metabolites within the clusters of GGM network
972	Table S11	The detailed information of the test dataset of Fiehn HILIC Library
973	Table S12	The detailed information of the identification with standard compounds
974		for 17 metabolite features
975	Table S13	The mass spectra of metabolites mentioned in paper
976	Appendix S1	The detailed steps for the acquisition and processing of relative
977		abundance data of metabolites

MS/MS spectra of rice grains

MS/MS spectra annotation with experimental mass spectra

MS/MS spectra annotation with *in silico* mass spectra

Structural characterization of MS/MS spectra

Metabolite association network

Structural motif search combined with neutral loss scanning

Figure 1

Figure 2

A

B

C

Figure 3

A

B

C

D

Figure 4

A

B

Figure 5