The utility of single-cell RNA sequencing data in predicting plant metabolic pathway genes

Jingwei Ma[1,2‡], Liting Zou[1‡], Zhenglin Wang[1‡], Xiaoxue Wang[1,2], Xinyi Zuo[1,3], Fei Wang[1,3], Zhiqing Wang[1,4], Zhimei Li[1], Lin Li[2]*, Peipei Wang[1]*

[1] Kunpeng Institute of Modern Agriculture at Foshan, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518124, China

[2] College of Life Science, Shanxi Normal University, Taiyuan 030031, China

[3] College of Horticulture, Northwest A&F University, Yangling, Shaanxi 712100, China

[4] College of Plant Protection, South China Agricultural University, Guangzhou, Guangdong 510642, China

Summary

- It is an ever challenging task to make genome-wide predictions for plant metabolic pathway genes (MPGs) encoding enzymes that catalyze the biosynthesis of plant natural products.
- Here, starting from 1,130 benchmark MPGs that have experimental evidence in *Arabidopsis thaliana*, we investigate the utilities of single-cell RNA sequencing (scRNA-seq) data—a recently arisen omics data that has been used in several other fields—in predicting MPGs using four machine learning (ML) algorithms that support multi-label tasks.
- Compared with traditional bulk RNA-seq data, scRNA-seq data lead to different and tighter co-expression networks among MPGs within metabolic classes, but relatively lower prediction accuracy of MPGs into classes. Splitting the scRNA-seq data into tissue-specific subsets can improve the gene co-expression network tightness and prediction accuracy of MPGs for some classes. Expression features from the same tissue types in bulk RNA-seq and scRNA-seq data have different contributions to the prediction of MPGs into classes. Models built using the ensemble algorithm AutoGluon outperforms those using other three classical ML algorithms.

- Our results demonstrate the usefulness and characteristics of scRNA-seq data in predicting MPGs into metabolic classes, and propose that more effort is needed in the future to improve the model prediction performance.

Keywords: *Arabidopsis thaliana*; single-cell RNA sequencing data; bulk RNA sequencing data; plant metabolic pathway genes; machine learning

Introduction

As sessile organisms, plants protect themselves from herbivores and stresses mainly via producing plant natural products (PNPs) (Al-Khayri *et al.*, 2023). Some PNPs, such as anthocyanin (Appelhagen *et al.*, 2018), paclitaxel (Yang *et al.*, 2020), salicylic acid (Ahmadi *et al.*, 2016) and pyrethroids (Elshafie *et al.*, 2023), have also been widely used in human nutrition, medicine, cosmetics, agriculture and other fields (Tiwari & Rana, 2015; Fang *et al.*, 2019; Jucá *et al.*, 2020; Bhambhani *et al.*, 2021). The content of these natural products in plants is generally quite low, and the extraction of them is usually inefficient and environmentally unfriendly (Zhang *et al.*, 2023). Thus, synthesizing PNPs in engineered microbial cells or plant cells has become a hotspot in synthetic biology, which requires the elucidation of metabolic pathway genes (MPGs) that encode enzymes catalyzing the biosynthesis of these PNPs. However, only about 0.1% of metabolic pathways of the estimated ~1 million metabolites synthesized by plants have been revealed (Schläpfer *et al.*, 2017). Before the genome era, the identification of plant MPGs relied heavily on traditional genetic and/or biochemical approaches, which are time- and labor-consuming. For example, the identification of the genes participating in the biosynthesis pathway of avenacin spanned about 22 years, from the isolation of saponin-deficient (*sad*) mutants (Papadopoulou *et al.*, 1999) to the identification of the 12-gene cluster (Li *et al.*, 2021).

With the development of omics technologies, more and more metabolic pathways have been deciphered by researchers via approaches such as genome-wide association studies (GWAS) (Shang *et al.*, 2014; Dong *et al.*, 2015) and association analysis between transcriptome and metabolome (Magnard *et al.*, 2015; Tohge *et al.*, 2020). Although these approaches have improved the efficiency and accuracy of MPG identification, they generally target one or a few pathways with relatively low throughput, but require large-scale genotyping or RNA-sequencing, in company with metabolomic profiling.

The development of computational tools, such as those for identifying plant metabolic gene clusters (Schläpfer *et al.*, 2017; Kautsar *et al.*, 2017; Töpfer *et al.*, 2017) or annotating enzyme genes (Schläpfer *et al.*, 2017), and machine learning (ML) approaches utilizing one or multiple types of omics data (Moore *et al.*, 2019; Wang *et al.*, 2021), has made it possible to predict plant MPGs on a larger scale. In one of our previous studies, we investigated the utility of transcriptome data in predicting MPGs in tomato using ML approaches, based on the hypothesis that genes in the same metabolic pathways may be transcripted similarly to cooperatively synthesize the target metabolites (Wang *et al.*, 2021). Although the ML approaches outperformed naive methods that soly compared expression similarities among genes in tomato, the global prediction accuracy was still not good enough (the median optimal F1 was 0.4 for 85 pathways) even though exhaustive combinations of the ways how transcriptome data were used and types of ML algorithms were evaluated. We reasoned that the failure in predicting tomato MPGs using transcriptome data may be partially due to the potential shortcomings of the traditional bulk RNA sequencing (bulk RNA-seq) data used in the study: the RNA sequencing is generally processed for a mixed pool of multiple types of cells within the sampled tissues. Considering that the biosynthesis of many PNPs are cell type-specific (Sun *et al.*, 2022), and that single-cell RNA sequencing (scRNA-seq) data have been used to explore the spatial organization of PNPs and the transcripts of contributing MPGs (Sun *et al.*, 2022; Yu *et al.*, 2023), we hypothesize that the prediction of plant MPGs may be potentially improved by using scRNA-seq data compared with bulk RNA-seq data.

Here, we compare the utilities of scRNA-seq vs. bulk RNA-seq data in predicting MPGs in *Arabidopsis thaliana*, for which there are the most comprehensive MPGs with experimental evidence (hereafter referred to as benchmark MPGs) and scRNA-seq data among other plant species. We first compare the degrees of co-expression levels for benchmark MPGs within the same metabolic classes using these two types of RNA-seq data. Next, we establish ML models for predicting MPGs into metabolic classes using these two data, employing four ML algorithms that support multi-class and multi-label classification. Finally, we apply our models to enzyme genes that lack experimental evidence, and compare the predicted metabolic classes of these genes with the annotated classes of them in the Plant Metabolic Network (PMN) database.

87 # Results

88 **Different co-expression modules of benchmark metabolic pathway genes based on two**
89 **types of RNA-seq data**

90 Bulk RNA-seq and scRNA-seq data of tissues collected from Arabidopsis wild-type plants
91 grown under normal conditions or continuous light were collected, since the majority of public
92 scRNA-seq data in Arabidopsis were from these tissues, resulting in a total of 13 scRNA-seq
93 datasets comprising 146 cell types and 5 bulk RNA-seq datasets encompassing 40 tissue types
94 (**Table S1–S4**, see **Methods**). We obtained the metabolic pathway annotations for Arabidopsis
95 genes from the PMN database (Schläpfer *et al.*, 2017), and only retained 1,130 benchmark MPGs
96 belonging to 635 metabolic pathways (**Table S5**, **Methods**). Considering that the majority of
97 these 635 metabolic pathways (75.32%) only contained ≤5 benchmark MPGs (**Fig. 1A**), we
98 grouped them into 13 metabolic classes according to a previous paper (Chae *et al.*, 2014) (**Table**
99 **S5**) to facilitate the following analysis. This classification resulted in an average number of
100 MPGs at 121 across 13 classes (a MPG can belong to multiple classes). Additionally, 34 genes
101 from 37 pathways that do not belong to any of the 13 classes were assigned to the category
102 "Others" (**Fig. 1B**).

103 To compare how well genes within different metabolic classes can be distinguished using
104 bulk RNA-seq and scRNA-seq data, we established co-expression networks for benchmark
105 MPGs using the weighted gene co-expression network analysis (WGCNA) (**Methods**). These
106 1,130 MPGs were clustered into 14 co-expression modules by both data: $M1_{bulk}$–$M14_{bulk}$ (**Fig.**
107 **1C, Table S6**) and $M1_{sc}$–$M14_{sc}$ (**Fig. 1E**, **Table S7**). In general, all the 14 M_{bulk} modules were
108 overrepresented by MPGs from 1~3 metabolic classes (**Fig. 1D**), whereas 5 M_{sc} modules were
109 not overrepresented by MPGs from any classes (**Fig. 1F**). On the other hand, MPGs from the
110 amines and polyamines (AP) and redox classes were not overrepresented in any of the M_{bulk}
111 modules, but overrepresented in $M8_{sc}$ and $M2_{sc}$, respectively. Similarly, MPGs from the
112 detoxification and intermediate metabolites (IM) classes were not overrepresented in any of the
113 M_{sc} modules, but in $M5_{bulk}$ and $M8_{bulk}$/$M13_{bulk}$, respectively.

114 In addition, the gene components differed for M_{bulk} and M_{sc} modules, even those that were
115 overrepresented by MPGs from the same metabolic classes. For example, $M14_{bulk}$ (containing 20
116 hormone genes) and $M9_{sc}$ (15) were overrepresented by MPGs from the hormone class, and

117 MPGs in $M14_{bulk}$ and $M9_{sc}$ were preferentially expressed in flowers/siliques and flowers/embryo-
118 related cell types, respectively (**Fig. 1C,E, Table S3,4,6,7**). However, there was no overlap
119 between hormone genes from $M14_{bulk}$ and $M9_{sc}$ (**Fig. S1A**). MPGs from specialized metabolism
120 (SM) were enriched in $M7_{bulk}$ and $M11_{bulk}$ (containing 50 and 28 SM genes, respectively), and
121 $M4_{sc}$ and $M7_{sc}$ (each containing 16 SM genes, **Fig. 1D,F**). MPGs in $M7_{bulk}$, $M4_{sc}$ or $M7_{sc}$ were
122 preferentially expressed in root-related tissues or cell types, whereas MPGs in $M11_{bulk}$ tended to
123 be expressed in broad types of tissues (**Fig. 1C,E**). There were 10 and 6 SM genes shared
124 between $M7_{bulk}$ and $M4_{sc}$, and between $M7_{bulk}$ and $M7_{sc}$, respectively, but each 1 SM gene shared
125 between $M11_{bulk}$ and $M4_{sc}$, and between $M11_{bulk}$ and $M7_{sc}$ (**Fig. S1B**). Taken together, these
126 results highlight the usefulness of both types of RNA-seq data in distinguishing MPGs from
127 different metabolic classes, while underscoring the differences in co-expression networks among
128 MPGs revealed by these two types of data.

129 **Clustering coefficient of gene expression within 13 metabolic classes based on two types of**
130 **RNA-seq data**

131 To better assess the tightness of the gene co-expression networks for individual metabolic
132 classes, we calculated the clustering coefficients (*C* values) for MPGs within 13 metabolic
133 classes using these two types of RNA-seq data (see **Methods**). The *C* value is normally used to
134 measure the degree to which nodes in a graph tend to cluster together (Onnela *et al.*, 2005), and
135 has been widely used in several real-world networks, particularly in social networks (Opsahl &
136 Panzarasa, 2009); the higher the *C* value, the tighter the network. To illustrate how well the *C*
137 value reflects the tightness of co-expression network within metabolic classes (**Table S8**), we
138 further calculated the *z*-score of the observed *C* value in 10,000 background *C* values where the
139 metabolic class annotations of all the 1,130 MPGs were randomly shuffled (**Table S9**, see
140 **Methods**). A z-score>1.645 (with a corresponding p-value<0.05) indicates that MPGs in a class
141 are significantly more tightly co-expressed than random gene groups (upper inset in **Fig. 2A**),
142 whereas a negative *z*-score indicates not tightly co-expressed (lower inset in **Fig. 2A**). Only the
143 energy metabolism (EM) and cofactors classes had *z*-scores >1.645 both when two types of
144 RNA-seq data were used (**Fig. 2A**); another two (IM and redox) and one (detoxification) class
145 had *z*-scores >1.645 only when scRNA-seq or bulk RNA-seq data were used, respectively. All
146 the remaining eight classes had *z*-scores <1.645 regardless of the types of RNA-seq data used.

147 Due to the fact that some PNPs may be synthesized in specific tissues or cell types (Daloso
148 *et al.*, 2023), we hypothesize that co-expression signals of MPGs synthesizing these natural
149 products within tissues/cell types may be diluted by combining RNA-seq data from all the
150 samples. To test this, we split the scRNA-seq data into four subsets based on the tissue types:
151 root, shoot, leaf and flower (the embryo subset was not considered due to the small number of
152 cell types [n=9]); the bulk RNA-seq data were not split further because it contained only 40
153 samples. When the four scRNA-seq subsets were used, more metabolic classes had *z*-
154 scores >1.645 (**Fig. 2B**, **Fig. S3,S4**, **Table S10**). For example, detoxification had *z*-scores >1.645
155 when the shoot, leaf and flower subsets were used; SM had *z*-scores >1.645 when root, shoot and
156 flower subsets were used; the carbohydrates and hormones classes had *z*-scores >1.645 when
157 shoot and flower subsets were used. In contrast, IM had *z*-scores <1.645 when four subsets were
158 used, compared with the *z*-score≈5 when the entire scRNA-seq data were used. However, AP
159 had *z*-scores <1.645 regardless of which types of RNA-seq data or which subsets of scRNA-seq
160 data were used. These results are consistent with our previous findings for bulk RNA-seq data
161 (Wang *et al.*, 2021) that subsets of transcriptome data provided different co-expression patterns
162 for plant MPGs, and highlight the need to explore different RNA-seq data when examining co-
163 expression tightness among MPGs for different metabolic classes.

164 **Prediction performance of machine learning models built using two types of RNA-seq data**

165 Thus far, we showed that the benchmark MPGs within most metabolic classes can be
166 distinguished from others using at least one type of RNA-seq data or one subset of scRNA-seq
167 data. Next, we asked to what degree benchmark MPGs can be correctly predicted into metabolic
168 classes. Considering that some MPGs belong to multiple metabolic classes (**Table S5**), we built
169 ML models using four algorithms that support multi-class and multi-label classification tasks: K-
170 nearest neighbors (KNN, Guo *et al.*, 2003), Random Forest (RF, Breiman, 2001), eXtreme
171 Gradient Boosting (XGBoost, Chen & Guestrin, 2016), and AutoGluon-Tabular (AutoGluon,
172 Erickson *et al.*, 2020). KNN makes predictions of classes for a given gene based on proximity of
173 this gene with genes in the training set; RF builds a number of classification decision trees using
174 a number of sub-samples of the dataset and then aggregates the trees into a single ensemble
175 model; XGBoost is an optimized distributed gradient boosting library that provides parallel tree
176 boosting; AutoGluon ensembles multiple ML algorithms and stacks them in multiple layers.

Twenty percent of benchmark MPGs within each class were first held out as the test set, and the remaining 80% MPGs were used to train the models using a five-fold cross-validation scheme (see **Methods**). F1 scores on the cross-validation ($F1_{CV}$) and test ($F1_{test}$) sets for models with the optimal combination of parameters were both reported, where the F1 score was the harmonic mean of precision and recall of the prediction (see **Methods**).

When measuring F1 scores for MPGs across all the 14 classes (13 metabolic classes plus the "Others" class), we found that models built using bulk RNA-seq data outperformed scRNA-seq-based models in terms of both $F1_{CV}$ and $F1_{test}$ scores, regardless of which algorithm was used (**Fig. 3A**). Although having the highest prediction accuracy, AutoGluon models had potential issues of overfitting, as $F1_{CV}$ scores for AutoGluon models were much higher than the corresponding $F1_{test}$ scores. RF models had the lowest prediction accuracy, with both $F1_{CV}$ and $F1_{test}$ <0.14. Although F1 scores for our models were mostly < 0.3, they were significantly higher than those for 1,000 random guesses where the metabolic classes of all the benchmark MPGs were randomly shuffled (the corresponding *z*-scores >1.645, **Fig. 3B, Fig. S5, Table S11,12**), suggesting the usefulness of these two types of RNA-seq data in predicting MPGs into metabolic classes.

Considering that MPGs within different classes behaved differently in the co-expression analysis as shown in the above sections, we also calculated F1 scores for MPGs within each class. Consistent with the overall F1 scores, bulk RNA-seq data led to significantly higher $F1_{CV}$ and $F1_{test}$ scores for more classes than scRNA-seq data regardless of which algorithms were used (**Fig. 3C,D**, **Fig. S6A,B**), and the F1 scores for individual classes were more similar between KNN, RF and XGBoost models than with AutoGluon models (**Fig. 3E**, **Fig. S6E,F**, only the comparison between $F1_{CV}$ scores were shown). In addition, we found that the prediction performance for a class was positively correlated with the number of MPGs within the class both when bulk RNA-seq and scRNA-seq data were used (**Fig. 3F**, only results of $F1_{CV}$ for XGBoost models were shown as examples). This result suggests that with the increase of knowledge about plant MPGs in the future, the prediction accuracy for these classes is expected to be improved. When examining the *z*-scores of F1 scores, which were supposed to better reflect the prediction performance of models built with different factors than F1 scores *per se*, we found that most of these 13 classes had *z*-scores >1.645 for both $F1_{CV}$ and $F1_{test}$ regardless of which type of RNA-seq data and which algorithms were used (**Fig. 3G,H**, **Fig. S6C,D**, **Fig. S7–14**, **Table S13–20**).

In the above section, we showed that tissue-specific scRNA-seq data led to higher *C* values for some classes (**Fig. 2B**) than the entire scRNA-seq data. We next asked whether the tissue-specific scRNA-seq data can also lead to better prediction for MPGs. $F1_{CV}$ scores of three, nine, four and seven classes can be improved by some of the tissue-specific scRNA-seq data for KNN, RF, XGBoost and AutoGluon models, respectively, while $F1_{test}$ scores of seven, six, three and four classes were improved for KNN, RF, XGBoost and AutoGluon models, respectively (**Fig. S15**). These results further highlight the need to explore different scRNA-seq data when predicting MPGs into different metabolic classes.

Furthermore, we investigated whether combining the two types of RNA-seq data would improve the prediction accuracy for MPGs. For the KNN models, the combined data led to significantly higher $F1_{CV}$ and $F1_{test}$ for four and two classes, respectively, both than when bulk RNA-seq or scRNA-seq data were used (**Fig. S16A**). Particularly, the average $F1_{CV}$ and $F1_{test}$ scores of 10 replicate runs for inorganic nutrients (IN) were improved from 0.034/0.028 (bulk RNA-seq/scRNA-seq) and 0/0 to 0.169 and 0.136, respectively. However, for RF, XGBoost and AutoGluon models, only $F1_{test}$ scores of one (cofactors), two (amino acids [AA] and AP), and two (redox and detoxification) classes were significantly improved, respectively, when combined data were used compared with two individual types of RNA-seq data (**Fig. S16B–D**). These results suggest that, although bulk RNA-seq and scRNA-seq data led to different gene co-expression networks and different prediction performance for MPGs within classes, combining these two data does not necessarily leverage the strengths of both data and lead to better prediction performance than individual data in most cases.

Important features revealed by interpreting the predictive models

To evaluate how the gene expression data from two types of RNA-seq data contribute to the prediction of MPGs into metabolic classes, we examined the feature importance for models built using these two data (see **Methods**, feature importance is not applicable for KNN models). We found that when the same RNA-seq data were used, models built using RF and AutoGluon tended to have more similar feature importance than with XGBoost models (**Fig. 4A, Fig. S17,18, Table S21–26**). In addition, the correlations of feature importance between different algorithms were higher in bulk RNA-seq-based models than in those based on scRNA-seq data (**Fig. 4A**). Together with the above findings that bulk RNA-seq-based models tended to have better

238 prediction of MPGs for more metabolic classes than scRNA-seq-based models, this result
239 suggests that bulk RNA-seq features may be more stable to reflect the co-expression correlations
240 among MPGs than scRNA-seq features.

241 On the other hand, when the same algorithm was used, features from the same tissue types
242 between bulk RNA-seq and scRNA-seq data tended to have different importance ranks. For
243 example, in RF models, bulk RNA-seq features from different types of tissues were generally
244 ranked evenly according to the feature importance (**left panel in Fig. 4B**, **Fig. S17A, Table S21**),
245 while the scRNA-seq features from flowers tended to have higher importance and rank relatively
246 higher than features from other tissues (**right panel in Fig. 4B**, **Fig. S18A, Table S22**). This
247 result is consistent with the findings that RF models built using scRNA-seq data from flowers
248 tended to have the highest or relatively higher prediction accuracy for most classes (**Fig. S15B**),
249 and that the number of MPGs that were specifically expressed in flowers in the scRNA-seq data
250 was the highest among others (**Fig. 1E**). Similarly, for AutoGluon models, scRNA-seq features
251 from flowers and embryos tended to be more important than those from other tissues (**Fig. S18C,**
252 **Table S26**), whereas bulk RNA-seq features from different tissues ranked more evenly (**Fig.**
253 **S17C, Table S25**). AutoGluon models built using scRNA-seq data from flowers and embryos
254 had better prediction accuracies for detoxification and IN than models based on data from other
255 tissues (**Fig. S15D**). In contrast, for XGBoost models, both the bulk RNA-seq and scRNA-seq
256 features from different tissues ranked relatively evenly (**Fig. S17B**, **Fig. S18B, Table S23,24**),
257 and much fewer metabolic classes had improved prediction accuracy when cell type-specific
258 scRNA-seq data were used compared with RF and AutoGluon models (**Fig. S15**). These results
259 suggest that RF and AutoGluon models may outperform XGBoost models in distinguishing the
260 most important predictors.

261 Beside the advantages of having higher prediction accuracy and being potentially better at
262 distinguishing important predictors, the AutoGluon models also allow us to examine the
263 importance of features to the prediction of individual classes (**Table S25,26**). For example, gene
264 expression in the cell culture tissues at an early stage ("Cell_culture_early" in **Fig. 4A**) and gene
265 expression in central cells of female gametophytes ("Central_cell", **Fig. 4A**) were consistently
266 important for models built using three algorithms, and these two features were most important
267 for predicting MPGs into the carbohydrates (**Fig. S19B, Table S25**) and EM (**Fig. S19A, Table**
268 **S26**) classes. To understand why these features were more important for predicting MPGs into

269 the corresponding metabolic classes than others, more analysis is needed to be conducted in the
270 future.

271 **Prediction of metabolic classes for enzyme genes without experimental evidence**

272 To evaluate how our models performed for new genes, we applied our models to 9,441
273 enzyme genes that have not been assigned to any metabolic pathways supported by experimental
274 evidence (hereafter referred to as unknown genes). Only AutoGluon and XGBoost models (the
275 last model from 10 replicate runs) were used due to the relatively better performance compared
276 with models built using the other two algorithms. Using AutoGluon models, 8,361 (88.6%) and
277 8,322 (88.1%) unknown genes were predicted into at least one of the 13 metabolic classes when
278 bulk RNA-seq and scRNA-seq data were used (**Fig. 5A**), whereas 4,623 (49.0%) and 3,425
279 (36.3%) unknown genes were predicted by bulk RNA-seq-based and scRNA-seq-based
280 XGBoost models, respectively(**Fig. S20A**). The number of unknown genes that were predicted
281 into each metabolic class was positively correlated with the number of benchmark MPGs within
282 the class (**Fig. 5A, Fig. S20A**). This result suggests two potential scenarios that are not mutually
283 exclusive: 1) the current pre-knowledge about MPGs are relatively proportionally distributed
284 among metabolic classes; 2) the data structure of predicted instances are influenced by the data
285 structure of the training instances.

286 Next, we examined whether the predicted metabolic classes for unknown genes are
287 consistent between models built using two types of RNA-seq data and with the annotation from
288 the PMN database (**Table S5, S27–30**). Using AutoGluon models, 2,051 unknown genes were
289 predicted into exactly the same metabolic classes using bulk RNA-seq and scRNA-seq data, and
290 119 out of them were annotated in the same classes by PMN (strict criteria, upper panel in **Fig.**
291 **5B**); the numbers were 2,281 and 251 for a loose criteria where a gene had at least one common
292 class between different datasets (lower panel in **Fig. 5B**). In comparison, using XGBoost models,
293 692/33 and 1,003/121 unknown genes shared common class memberships between two
294 predictions/among three datasets, when using strict or loose criteria, respectively (**Fig. S20B**),
295 much fewer than those using AutoGluon models. These results further indicate that the ensemble
296 algorithm AutoGluon outperforms XGBoost (and also the other two algorithms) in predicting
297 MPGs into metabolic classes using transcriptomic data.

298 Considering that a benchmark MPG can be annotated with more than one class, and the
299 multiple class annotations for a gene were not randomly distributed (e.g., MPGs in detoxification
300 tended to be also annotated as in redox, compared with other classes, **Fig. 5C**), we further
301 investigated whether the mismatches of classes between PMN annotations and model predictions
302 have a similar bias among classes (**Fig. 5D,E, Fig. S20C,D**). For some cases the mismatches
303 resembled the multi-label annotations. For example, unknown genes that were annotated as
304 within SM by PMN tended to be predicted as within the class hormones by the bulk RNA-seq-
305 based AutoGluon model (**Fig. 5D**) and both XGBoost models built using bulk RNA-seq and
306 scRNA-seq data (**Fig. S20C,D**). However, the global mismatch patterns did not necessarily
307 repeat the multi-label annotation patterns. For example, benchmark MPGs in EM tended to also
308 belong to carbohydrates and AA (**Fig. 5C**), whereas unknown genes in EM from PMN
309 annotations tended to be predicted into cofactors and hormones by bulk RNA-seq-based
310 AutoGluon model (**Fig. 5D**), fatty acids and lipids by scRNA-seq-based AutoGluon model (**Fig.
311 5E**), and SM by both XGBoost models (**Fig. S20C,D**). This discrepancy can be mainly due to the
312 unbalanced numbers of benchmark MPGs from different classes used to train the models, and the
313 far-from-perfect prediction accuracy of our models built using transcriptomic data. The potential
314 low quality of PMN annotations for these unknown genes can be another rationale that should
315 not be excluded at the current stage.

316 Discussion

317 Models built using two types of RNA-seq data had significantly higher prediction accuracy
318 for most metabolic classes than random guesses, suggesting the usefulness of both RNA-seq data
319 in predicting MPGs into metabolic classes. Generally, bulk RNA-seq-based models led to higher
320 performance in predicting MPGs into classes than scRNA-seq-based models, opposed to our
321 assumption that scRNA-seq data may provide more detailed expression profiles for genes than
322 bulk RNA-seq data. There can be three potential reasons for this result. Firstly, the sequencing
323 and analytical techniques for single-cell transcriptome at present still have multiple limitations,
324 such as the low capture efficiency for transcripts within cells, especially for genes with low
325 expression, resulting in inherently sparse and noisy nature of the scRNA-seq data (Kharchenko,
326 2021); the analysis of scRNA-seq data is complex and challenging, and requires prior knowledge
327 about marker genes (Ke *et al.*, 2022). With the development and improvement of sequencing and

328 analytical techniques, scRNA-seq data is expected to show more promise in multiple fields,
329 including predicting plant MPGs.

330 Secondly, the biosynthesis of a PNP may occur sequentially in a series of specific cells. For
331 example, the transcription of 20 genes responsible for the biosynthesis of monoterpenoid indole
332 alkaloids occur in three types of cells (namely, internal phloem-associated parenchyma cells,
333 epidermal cells, and laticifer cells) in the *Catharanthus roseus* leaves, and the intermediates
334 produced during the biosynthesis process are transported between different cell types (Sun *et al.*,
335 2022). If this phenomenon is common for PNPs, bulk RNA-seq data that capture gene expression
336 profiles for the entire tissues rather than individual cell types are supposed to be superior to
337 scRNA-seq data in predicting MPGs. Bioinformatic approaches that take into account the
338 temporal or spatial frameshift in expression when measuring the expression correlation between
339 genes could potentially make better use of the scRNA-seq data in predicting MPGs.

340 Thirdly, our predictive models used the expression profiles of genes in bulk RNA-seq and
341 scRNA-seq data, rather than the expression correlations among genes (e.g., *C* values of genes).
342 In our clustering coefficient analysis, scRNA-seq data led to higher *C* values (*z*-scores > 1.645)
343 for more classes than bulk RNA-seq data (**Fig. 2A**), and for even more classes when tissue-
344 specific scRNA-seq data were used (**Fig. 2B**). Models built using *C* values of genes may
345 improve the prediction performance when scRNA-seq data were used. However, a major
346 challenge for using *C* values of genes as predictors in the models is that the *C* value of a gene is
347 calculated by considering all the neighbors that are directly connected with the gene. It is a tricky
348 problem to calculate the *C* values for genes in the validation and test subsets with neighbors from
349 the training subset while controlling the risks of data leakage during the calculation.

350 The finding that prediction performance for a given class was influenced by the number of
351 MPGs within the class regardless of the types of RNA-seq data used (**Fig. 3F**) suggests an urgent
352 need to expand our knowledge about plant MPGs to improve the prediction accuracy, which in
353 turn needs assistance from other approaches (such as ML approaches in our study) beyond
354 conventional forward genetics. One potential solution for this seemingly "endless loop" is to
355 experimentally evaluate part of the preliminary predictions (error-prone) and feed the validated
356 knowledge back to the ML models to boost the prediction performance. Another potential
357 solution would be the development of ML algorithms which can handle with giant amount of
358 features (e.g., expression and epigenetic profiles of genes across a large number of tissues, cell

359 types, stages, environments, and so on) while learning from a limited numbers of instances (i.e.,
360 genes). The AutoGluon algorithm has shown great performance in prediction tasks such as
361 OpenML AutoML and Kaggle Benchmark (Erickson *et al.*, 2020). Although outperformed
362 models built using other three classical ML algorithms in predicting MPGs into metabolic
363 classes in our study, AutoGluon models had the risk of overfitting. This indicates that for some
364 prediction tasks in biology with limited numbers of instances (e.g., plant MPGs), the algorithms
365 for ensemble learning (e.g., AutoGluon) that show great promise in other fields (e.g., image
366 recognition) should be used with caution.

367 In our study, due to the limitation of publicly released scRNA-seq data, we only focused on
368 scRNA-seq and bulk RNA-seq data from wild-type tissues that were grown under normal
369 conditions. A number of PNPs, especially specialized metabolites, are only synthesized in
370 specific tissues or cell types, or under specific conditions, such as biotic and abiotic stresses. For
371 example, acyl sugars are specifically synthesized in the glandular trichomes of certain plants
372 (Schilmiller *et al.*, 2008); the rupture of a cassava leaf mesophyll cell caused by herbivore attack
373 or animal chewing leads to the production of cyanide which has the bitter taste for mammals
374 (McMahon *et al.*, 1995). Our findings showed that tissue-specific scRNA-seq data led to better
375 prediction for some classes than the entire scRNA-seq data we collected. More scRNA-seq data
376 for samples collected under different conditions are needed to improve the prediction accuracy
377 for MPGs responsible for the biosynthesis of the PNPs that are specific to the corresponding
378 conditions (Wang *et al.*, 2021).

379 In summary, the prediction of MPGs is a complex and challenging task. More efforts should
380 be put in the future to expand our knowledge about the contributing genes underlying PNPs with
381 higher efficiency and to accelerate the development of synthetic biology.

382 Materials and Methods

383 RNA-seq data collection and preprocessing

384 A total of 13 scRNA-seq datasets of *A. thaliana* (**Table S1**) were downloaded from the
385 PlantscRNAdb (http://ibi.zju.edu.cn/plantscrnadb/, accessed on Mar 20th, 2023) and Gene
386 Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/, accessed on March 20th, 2023)
387 databases, or from the publication (Song *et al.*, 2020). These scRNA-seq datasets contained
388 samples taken from root, shoot, leaf, flower and embryo tissues of wild-type plants grown under

389 normal conditions or continuous lights. Five bulk RNA-seq datasets that contained samples from
390 the same types of tissues of wild-type plants grown under normal conditions or continuous lights
391 (**Table S2**) were downloaded from the Expression Atlas database
392 (https://www.ebi.ac.uk/gxa/sc/home) and GEO (accessed on Mar 20th, 2023) databases, or from
393 the publication (Mergner *et al.*, 2020).

394 For the preprocessing of scRNA-seq data, genes expressed in <3 cells and cells with <200
395 genes expressed were first filtered out from the following analysis. Missing values in the
396 resulting expression matrix were filled with zeros. The Seurat (v4.3.0) package (Hao *et al.*, 2021)
397 in R (v4.2.0) was used to preprocess the scRNA-seq data. Specifically, the "RunPCA" function
398 was used for dimensionality reduction; "NormalizeData" was used to normalize gene expression
399 levels among different cells, where read count of each gene was divided by the total counts
400 within individual cells, then multiplied by a scale factor of 10,000 and log1p-transformed;
401 "FindClusters" was used to cluster cell types; "UMAP" was used to visualize the clusters.
402 Marker genes that were specifically expressed in certain cell types were obtained from the
403 PlantscRNAdb and were used to annotate the corresponding cell types. A total of 146 cell types
404 were annotated and gene expression levels were averaged among cells with the same type (**Table**
405 **S3**).

406 For the preprocessing of bulk RNA-seq data, the gene expression levels (transcripts per
407 million, TPM) were first averaged across biological replicates, and then were normalized using
408 the "normalize.quantiles" function in R among different samples, where the expression values
409 within individual samples were adjusted to ensure a consistent distribution and to eliminate
410 technical variations between samples. The normalized expression values were further log1p-
411 transformed to enable comparison with those in scRNA-seq data. (**Table S4**).

412 **Metabolic pathway gene annotation collection**

413 We first downloaded the metabolic pathway gene annotation from the PMN database
414 (https://plantcyc.org/, accessed on Sep 9th, 2022, PlantCyc v15.1.0, Schläpfer *et al.*, 2017), which
415 contained 3,508 MPGs annotated in 635 metabolic pathways. To obtain benchmark MPGs, we
416 wrote a python script to crawl the PMN webpages of individual MPGs (accessed on Dec 19th,
417 2022) and kept MPGs with experimental evidence, such as those inferred from direct assay,
418 mutant phenotype, etc. This resulted in a total of 1,131 benchmark MPGs, which belong to 636

419 metabolic pathways (**Table S5**). The gene *AT5G23960* and its corresponding pathway β-
420 caryophyllene biosynthesis (PWY-6275) was removed from the following analysis since it
421 lacked expression information in the scRNA-seq data. Additional potential enzyme genes of *A.*
422 *thaliana* were obtained from The Arabidopsis Information Resource (TAIR) databases
423 (https://www.arabidopsis.org/), or annotated using the Ensemble Enzyme Prediction Pipeline
424 (E2P2, v3.1) software. In summary, a total of 10,572 potential enzyme-encoding genes were
425 annotated, including 1,131 benchmark MPGs (**Table S5**).

426 **Weighted gene co-expression network analyses**

427 Gene co-expression network analysis was conducted using the WGCNA (v1.72.1, Zhang &
428 Horvath, 2005) package in R. During the network analysis, the matrix of Pearson correlation
429 coefficient (PCC) values for gene pairs was transformed into an adjacency matrix to achieve a
430 scale-free network structure, with the parameter 'networkType' being set to 'signed', where the
431 PCC values were processed using the equation:

432 $$\boldsymbol{a_{i,j}} = \left(\frac{\mathbf{1+cor(i,j)}}{\mathbf{2}}\right)^{\boldsymbol{\beta}} \quad (1)$$

433 where, $\boldsymbol{a}_{i,j}$ represents the weight of the edge between two genes *i* and *j*; *cor*(*i, j*) represents the
434 PCC between the two genes; β is the best soft threshold power estimated by the software
435 WGCNA using the function "pickSoftThreshold", which was estimated at 12 for both bulk
436 RNA-seq and scRNA-seq data. The "cutreeDynamic" function was employed to divide the
437 network into 14 co-expression modules for both RNA-seq data, using parameters as
438 minModuleSize=30, mergeCutHeight=0.1 for both data, while deepSplit=3 and 4 for scRNA-seq
439 and bulk RNA-seq data, respectively.

440 **Clustering coefficient calculation**

441 To measure the tightness of gene co-expression within metabolic classes, we calculated the
442 clustering coefficient (*C* value) for each class. The gene expression values in two RNA-seq data
443 first plus one and were log-transformed with base 10, and then processed through the WGCNA
444 pipeline to obtain the gene co-expression networks. In the networks, the nodes were benchmark
445 MPGs, edges between nodes were co-expression relationships between MPGs with the thickness
446 of edges (i.e., the strength of the connection) being the transformed PCC values. The networks

were next processed using the NetworkX (v3.1, Hagberg *et al.*, 2008) package in Python to calculate the local *C* values for individual genes, with the equation (Onnela *et al.*, 2005):

$$c_u = \frac{1}{deg(u)(deg(u)-1)} \sum_{vw} \quad (\widehat{w}_{uv}\widehat{w}_{uw}\widehat{w}_{vw})^{1/3} \qquad (2)$$

where, for a node (i.e., a gene) *u* with two adjacent nodes *v* and *w* as an example, c_u represents the local *C* value of the node *u*; *deg(u)* is the degree of the node *u*, representing the number of neighboring nodes connected to the node *u*; w_{uv}, w_{uw}, and w_{vw} represent the weights of the edges between nodes *u*, *v*, and *w*; the weights are normalized by being divided by the maximum weight in the network, e.g., $\widehat{w}_{uv} = w_{uv}/max(w)$; $(\widehat{w}_{uv}\widehat{w}_{uw}\widehat{w}_{vw})^{1/3}$ indicates the geometric mean of the normalized weights between nodes *u*, *v*, and *w*; $\sum_{vw}$ represents the summation over all pairs of neighboring nodes (i.e., *v*,*w*) that are directly connected to node *u*. The local *C* value ranges from 0 to 1. The average local *C* value of all genes within a metabolic class (the global *C* value, hereafter referred to as *C* value for short) was used to represent the compactness of the class in terms of gene expression (**Table S8**).

Since the *C* values can be influenced by the number of nodes within a network (Kaiser, 2008), to understand what *C* values would indicate a tight network and to make the *C* values among different metabolic classes comparable, we also estimated the *C* values for 10,000 random simulations (**Table S9,10**) where the class memberships of all the benchmark MPGs were randomly shuffled to simulate the out-of-order relationships among these MPGs. Then a *z*-score of the observed *C* values in the distribution of 10,000 background *C* values were calculated using the equation:

$$z = \frac{x-\mu}{\sigma} \qquad (3)$$

where, *z* represents the *z*-score value, *x* represents the observed *C* value, μ represents the mean of 10,000 background *C* values, and σ represents the standard deviation of 10,000 background *C* values.

Machine learning model building

Four algorithms were used to build ML models: KNeighborsClassifier (K-Nearest Neighbors, KNN), RandomForestClassifier (RF), and XGBClassifier (eXtreme Gradient Boosting, XGBoost) from scikit-learn (v1.2.2, implemented in Python v3.9.16, Pedregosa *et al.*, 2011), and AutoGluon-Tabular (AutoGluon, Erickson *et al.*, 2020) from the AutoGluon library

(v0.8.2, Python). Out of the benchmark MPGs within each class, 20% were used as the test set to evaluate the performance of the models. The remaining 80% of benchmark MPGs within each class were further split into five folds; MPGs in four folds were used to train the model and MPGs in the remaining fifth fold were used to validate the model performance; this training-validation step was conducted five times to make sure that MPGs in all five folds were used as in the validation subset once. Caution should be taken to avoid that genes that belong to multiple classes are held out as in the test set for one class but remain as in the training set for another class, and the same for the training/validation splits.

Hyperparameter tuning was conducted for KNN, RF and XGBoost models (this step is automatically embedded within AutoGluon) to select the optimal combination of parameters using GridSearch. Specifically, the hyperparameter spaces for KNN models were {3, 4, 5, 6, 7} for the 'n_neighbors' (*k* values) and {'uniform', 'distance'} for the 'weights'; for RF models: {3, 5, 10} for the 'max_depth', {0.1, 0.25, 0.5, 0.75, 'sqrt', 'log2', None} for the 'max_features' and {10, 50, 100, 500} for the 'n_estimators'; for XGBoost modes: {3, 5, 10} for 'max_depth', {0.01, 0.1, 0.5, 1} for 'learning_rate', {10, 50, 100, 500} for 'n_estimators' and {0.1, 5} for 'gamma'. The best combination of parameters for each model was determined based on the F1 scores on the cross-validation set ($F1_{CV}$), where F1 was calculated with the equations:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

$$precision = \frac{TP}{TP + FP} \quad (5)$$

$$recall = \frac{TP}{TP + FN} \quad (6)$$

where TP (true positive) represents the number of benchmark MPGs that are correctly predicted into a class; FP (false positive) represents the number of benchmark MPGs that are mis-predicted into a class; FN (false negative) represents the number of benchmark MPGs that are not predicted into the correct class.

Finally, the feature importance was obtained for RF ("gini"), XGBoost ("weight"), and AutoGluon ("permutation-based") models (**Table S21–26**), and the feature importance for all the classes were averaged for AutoGluon models to make them comparable to those for RF and XGBoost models.

Acknowledgments

We thank Xinhao Fan, Qiaowei Li, Kangli Li, Ruipu Chen, Zhuobiao Wang, for their helpful suggestions in preprocessing the scRNA-seq data. This work was supported by the Scientific Research Foundation for Principle Investigator, Kunpeng Institute of Modern Agriculture at Foshan (KIMAQD2022003 to PW) and National Natural Science Foundation of China (32370241 to PW).

Author contributions

PW conceived and designed this study. ZW collected the metabolic pathway gene annotations; LZ and JM collected and preprocessed the bulk RNA-seq and scRNA-seq data, respectively. ZW wrote the machine learning pipelines. JM conducted all the other analyses with help from LZ, XW and ZW. JM and PW wrote the manuscript with inputs from all the other authors. All authors read and approved the final manuscript.

Data availability

All the data used in this study are provided in the supplementary materials, and all the scripts used are available on Github at:
https://github.com/peipeiwang6/Manuscript/tree/main/2024_scRNA_in_pathway_prediction.

Figure legends

Fig. 1. Different expression patterns for genes within metabolic classes revealed using bulk RNA-seq and single-cell RNA-seq (scRNA-seq) data. (**A**) A pie plot showing the percentages of 635 metabolic pathways that have a certain number of benchmark metabolic pathway genes (MPGs): blue, 1 MPG; purple, 2; cyan, 3; orange, 4; green, 5; red, ≥ 6 MPGs. (**B**) A bar chart showing the numbers of MPGs belonging to 13 metabolic classes and the "others" category that contains all the other pathways not belonging to any classes. (**C,E**) Heatmaps showing the expression patterns of 1,130 MPGs, which were grouped into 14 co-expression modules ($M1_{bulk}$–$M14_{bulk}$ for **C**, $M1_{sc}$–$M14_{sc}$ for **E**) inferred by weighted gene co-expression network analysis using the bulk RNA-seq (**C**) and scRNA-seq (**E**) data. Color scale in the heatmaps: the

530 standardized expression values (transcripts per kilobase million) using *z*-score normalization; red:
531 high expression, blue: low expression; all values > 3 were assigned a value of 3. Tissue types
532 that bulk RNA-seq samples or scRNA-seq cell types belong to are indicated above the heatmaps
533 with colors: dark green, root; green, shoot; olive, leaf; orange, flower; pink, embryo; light pink,
534 others. (**D,F**) Heatmaps showing enrichment of genes from 13 different metabolic classes
535 (columns) within 14 co-expression modules (rows) when bulk RNA-seq (**D**) and scRNA-seq (**F**)
536 data were used. The ordering of metabolic classes is the same as in (**B**). Color scale: odds ratio
537 from Fisher's exact test; all values > 4 were assigned a value of 4. Asterisk(s) indicate(s)
538 significance levels of the Fisher's exact test: *, *p*-value<0.05; **, *p*-value<0.01; ***, *p*-
539 value<0.001; only the over-representative cases (odds ratio > 1) are indicated with the
540 significance levels. SM: specialized metabolism; Carb: carbohydrates; FAL: fatty acids and
541 lipids; AA: amino acids; Cof: cofactors; Horm: hormones; EM: energy metabolism; NT:
542 nucleotides; IN: inorganic nutrients; DTX: detoxification; AP: amines and polyamines; IM:
543 intermediate metabolism.
544
545 **Fig. 2. Clustering coefficients (*C* values) for genes within 13 metabolic classes calculated**
546 **using two types of RNA-seq data**. (**A**) The *z*-scores of *C* values for genes within 13 metabolic
547 classes compared with background values, when bulk RNA-seq (blue) and scRNA-seq (orange)
548 data were used. The dashed horizontal line indicates a *z*-score of 1.645, corresponding to a *p*-
549 value of 0.05. Light gray background: metabolic classes with significantly higher *C* values than
550 background values when both types of RNA-seq data were used; light orange and light blue
551 backgrounds: classes with significantly higher *C* values only when scRNA-seq or bulk RNA-seq
552 data were used, respectively. The two inset density plots show the distribution of 10,000
553 background *C* values for genes in the energy metabolism (EM, upper) and specialized
554 metabolism (SM, lower) classes as examples when scRNA-seq data were used. Red solid vertical
555 line: the observed *C* value for genes within EM or SM class; black dashed vertical line: the
556 average *C* value of 10,000 background values. (**B**) *z*-scores of *C* values for genes within 13
557 metabolic classes when scRNA-seq data from four different types of tissues (root, shoot, leaf,
558 and flower) were used.
559

560 **Fig. 3. Prediction accuracy of genes in metabolic classes for models built using two types of**
561 **RNA-seq data and different algorithms**. (**A**) Overall $F1_{CV}$ (left panel) and $F1_{test}$ (right) scores
562 for all 13 metabolic classes when bulk RNA-seq (blue) and scRNA-seq (orange) data were used
563 (n=10 replicate runs). KNN, RF, XGBoost and AutoGluon: models built using K-Nearest
564 Neighbors, Random Forest, eXtreme Gradient Boosting and AutoGluon-Tabular, respectively.
565 (**B**) *z*-scores of $F1_{CV}$ (left panel) and $F1_{test}$ (right) for models built using two types of RNA-seq
566 data and different algorithms. A *z*-score > 1.645 (dashed line, *p*-value=0.05) indicates that the
567 model's performance is significantly better than random guess. (**C,D**) $F1_{CV}$ (left panel) and $F1_{test}$
568 (right) scores for individual metabolic classes for XGBoost (**C**) and AutoGluon (**D**) models. **E:**
569 Correlation of $F1_{CV}$ between XGBoost (x-axis) and KNN models (y-axis) built using bulk RNA-
570 seq (left panel) and scRNA-seq (right) data. **(F**) Correlation between the number of MPGs in a
571 class and the $F1_{CV}$ for the class in XGBoost models, when bulk RNA-seq (left panel) and
572 scRNA-seq (right) data were used. Error bars indicate the standard deviation of results from 10
573 replicate runs. PCC: Pearson correlation coefficient. (**G,H**) *z*-score values of $F1_{CV}$ (left panel)
574 and $F1_{test}$ (right) scores for 13 metabolic classes compared with values of random guess for 1,000
575 times for XGBoost (**G**) and 100 times for AutoGluon (**H**) models. Asterisk(s) indicate(s)
576 significant levels for the two-sided Wilcoxon rank-sum test: *, *p*-value < 0.05; **, *p*-value < 0.01;
577 ***: *p*-value < 0.001. EM: energy metabolism; Cof: cofactors; IM: intermediate metabolism;
578 DTX: detoxification; AA: amino acids; NT: nucleotides; AP: amines and polyamines; Carb:
579 carbohydrates; FAL: fatty acids and lipids; IN: inorganic nutrients; SM: specialized metabolism;
580 Horm: hormones.
581
582 **Fig. 4. Feature importance for RF, XGBoost and AutoGluon models**. (**A**) Correlation of
583 importance for bulk RNA-seq (upper panels) and scRNA-seq (lower) features between RF,
584 XGBoost and AutoGluon models. PCC: Pearson correlation coefficient. (**B**) Importance of bulk
585 RNA-seq (left panel) and scRNA-seq (right) features in RF models. For the details of feature
586 names, see **Fig. S17,18**. Different colors indicate features from different tissue types: dark green,
587 root; green, shoot; olive, leaf; orange, flower; pink, embryo; light pink, others.
588
589 **Fig. 5. Prediction of metabolic classes for unknown genes by AutoGluon models**. (**A**) Count
590 of unknown genes that were predicted into 13 metabolic classes by AutoGluon models built

using bulk RNA-seq (upper panel) and scRNA-seq (lower) data. (**B**) Venn diagrams showing the overlap of unknown genes that were predicted into metabolic classes by AutoGluon models built using bulk RNA-seq (blue) and scRNA-seq (orange) data, and those that were annotated as within metabolic classes in the PMN database (purple). In the upper and lower panels, overlapping regions show the number of genes with exactly the same classes or at least one common class between datasets, respectively. (**C**) A heatmap showing the numbers of benchmark MPGs that were annotated within a certain metabolic class (diagonal cells) and within two different classes (non-diagonal cells). For example, from 242 benchmark MPGs that were annotated as in SM, 31 MPGs were also annotated as in the hormone class. (**D**,**E**) Heatmaps showing the numbers of unknown genes that were predicted as in a class (x-axis) by XGBoost models built using bulk RNA-seq (**D**) and scRNA-seq (**E**) data, and were also annotated as within the same or other classes (y-axis) by PMN database. Values in the column "Not-assigned" indicate the numbers of unknown genes that were predicted into neither any of the 13 metabolic classes nor the "Others" category. Cell colors in (**C,D,E**) indicate the numbers of genes. EM: energy metabolism; Cof: cofactors; IM: intermediate metabolism; DTX: detoxification; AA: amino acids; NT: nucleotides; AP: amines and polyamines; Carb: carbohydrates; FAL: fatty acids and lipids; IN: inorganic nutrients; SM: specialized metabolism; Horm: hormones.

Supplementary figure legends

Fig. S1. Shared benchmark MPGs within co-expression modules inferred by bulk RNA-seq and scRNA-seq data. (**A**) A venn diagram showing the numbers of benchmark MPGs shared by $M3_{bulk}$, $M4_{bulk}$, $M14_{bulk}$, $M7_{sc}$, and $M9_{sc}$, for which MPGs from the hormone class were overrepresented. (**B**) A venn diagram showing the numbers of benchmark MPGs shared by $M7_{bulk}$, $M11_{bulk}$, $M4_{sc}$, and $M7_{sc}$, for which MPGs from the specialized metabolism class were overrepresented.

Fig. S2. Clustering coefficients (*C* values) for MPGs within each metabolic class calculated using bulk RNA-seq (A) and scRNA-seq (B) data. The density plot shows the distribution of background *C* values for MPGs within each metabolic class for 10,000 random simulations where the metabolic class annotations for MPGs were randomly shuffled. The red solid line

denotes the observed *C* value for the co-expression network of genes within each class; the black dashed line indicates the average *C* value of 10,000 random simulations. The *z* represents the *z*-score of the observed *C* value in the background distribution; *p* represents the corresponding *p*-value of the *z*-score.

Fig. S3. *C* values for MPGs within each metabolic class calculated using scRNA-seq data of roots (A) and shoots (B). The density plot shows the distribution of background *C* values. The density plot shows the distribution of background *C* values for MPGs within each metabolic class for 10,000 random simulations where the metabolic class annotations for MPGs were randomly shuffled. The red solid line denotes the observed *C* value for the co-expression network of genes within each class; the black dashed line indicates the average *C* value of 10,000 random simulations. The *z* represents the *z*-score of the observed *C* value in the background distribution; *p* represents the corresponding *p*-value of the *z*-score.

Fig. S4. *C* values for MPGs within each metabolic class calculated using scRNA-seq data of leaves (A) and flowers (B). The density plot shows the distribution of background *C* values for MPGs within each metabolic class for 10,000 random simulations where the metabolic class annotations for MPGs were randomly shuffled. The red solid line denotes the observed *C* value for the co-expression network of genes within each class; the black dashed line indicates the average *C* value of 10,000 random simulations. The *z* represents the *z*-score of the observed *C* value in the background distribution; *p* represents the corresponding *p*-value of the *z*-score.

Fig. S5. Overall F1 scores for KNN, RF, XGBoost, and AutoGluon models when bulk RNA-seq (A) and scRNA-seq (B) data were used. The density plot shows the distribution of overall F1 scores for models where the labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red solid line: the average $F1_{CV}$ or $F1_{test}$ score for models built using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000 random guesses for KNN, RF, and XGBoost models and 100 random guesses for AutoGluon models. *z*: *z*-score of the average F1 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-score.

652 **Fig. S6. Prediction accuracy of MPGs within individual metabolic classes for KNN and RF**
653 **models**. (**A,B**) $F1_{CV}$ (left panel) and $F1_{test}$ (right) scores for individual metabolic classes for KNN
654 (**A**) and RF (**B**) models when bulk RNA-seq (blue) and scRNA-seq (orange) data were used.
655 Error bars indicate the standard deviation of results from 10 replicate runs. Asterisk(s) indicate(s)
656 significant levels for the two-sided Wilcoxon rank-sum test: *, *p*-value < 0.05; **, *p*-value < 0.01;
657 ***: *p*-value < 0.001. (**C**,**D**) *z*-scores of $F1_{CV}$ (left panel) and $F1_{test}$ (right) scores for 13 metabolic
658 classes compared with values of random guess for 1,000 times for KNN (**C**) and RF (**D**) models.
659 A *z*-score above the dashed line (*z*-score=1.645, *p*-value=0.05) indicates that the model's
660 performance is significantly better than random guess. (**E,F**) The correlation of $F1_{CV}$ between
661 models built using XGBoost (x-axis) and RF (y-axis in **E**), and between models built using
662 XGBoost and AutoGluon (y-axis in **F**), when bulk RNA-seq (left panel) and scRNA-seq (right)
663 data were used. PCC: Pearson correlation coefficient. EM: energy metabolism; Cof: cofactors;
664 IM: intermediate metabolism; DTX: detoxification; AA: amino acids; NT: nucleotides; AP:
665 amines and polyamines; Carb: carbohydrates; FAL: fatty acids and lipids; IN: inorganic nutrients;
666 SM: specialized metabolism; Horm: hormones.
667
668 **Fig. S7. $F1_{CV}$ (A) and $F1_{test}$ (B) scores of individual classes for KNN models built using bulk**
669 **RNA-seq data**. The density plot shows the distribution of F1 scores for models where the labels
670 of benchmark MPGs in the training set were randomly shuffled (random guesses). Red solid line:
671 the average F1 score for models built using the true labels of MPGs; pink rectangle area: the
672 range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000 random
673 guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the background distribution; *p*:
674 the corresponding *p*-value of the *z*-score.
675
676 **Fig. S8. $F1_{CV}$ (A) and $F1_{test}$ (B) scores of individual classes for KNN models built using**
677 **scRNA-seq data**. The density plot shows the distribution of F1 scores for models where the
678 labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red
679 solid line: the average F1 score for models built using the true labels of MPGs; pink rectangle
680 area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000
681 random guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the background
682 distribution; *p*: the corresponding *p*-value of the *z*-score.

Fig. S9. $F1_{CV}$ (A) and $F1_{test}$ (B) scores of individual classes for RF models built using bulk RNA-seq data. The density plot shows the distribution of F1 scores for models where the labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red solid line: the average F1 score of models built using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000 random guesses. *z*: z-score of the average F1 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-score.

Fig. S10. $F1_{CV}$ (A) and $F1_{test}$ (B) scores of individual classes for RF models built using scRNA-seq data. The density plot shows the distribution of F1 scores for models where the labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red solid line: the average F1 score of models built using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000 random guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-score.

Fig. S11. $F1_{CV}$ (A) and $F1_{test}$ (B) scores of individual classes for XGBoost models built using bulk RNA-seq data. The density plot shows the distribution of F1 scores for models where the labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red solid line: the average F1 score of models built using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000 random guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the background distribution; *p*: the corresponding *p*-value of the *z*-score.

Fig. S12. $F1_{CV}$ (A) and $F1_{test}$ (B) scores of individual classes for XGBoost models built using scRNA-seq data. The density plot shows the distribution of F1 scores for models where the labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red solid line: the average F1 score of models built using the true labels of MPGs; pink rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 1,000

713 random guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the background
714 distribution; *p*: the corresponding *p*-value of the *z*-score.
715
716 **Fig. S13. $F1_{CV}$ (A) and $F1_{test}$ (B) scores of individual classes for AutoGluon models built**
717 **using bulk RNA-seq data.** The density plot shows the distribution of F1 scores for models
718 where the labels of benchmark MPGs in the training set were randomly shuffled (random
719 guesses). Red solid line: the average F1 score of models built using the true labels of MPGs; pink
720 rectangle area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1
721 score of 100 random guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the
722 background distribution; *p*: the corresponding *p*-value of the *z*-score.
723
724 **Fig. S14. $F1_{CV}$ (A) and $F1_{test}$ (B) scores of individual classes for AutoGluon models built**
725 **using scRNA-seq data.** The density plot shows the distribution of F1 scores for models where
726 the labels of benchmark MPGs in the training set were randomly shuffled (random guesses). Red
727 solid line: the average F1 score of models built using the true labels of MPGs; pink rectangle
728 area: the range of F1 scores of 10 replicate runs; black dashed line: the average F1 score of 100
729 random guesses. *z*: *z*-score of the average F1 score of 10 replicate runs in the background
730 distribution; *p*: the corresponding *p*-value of the *z*-score.
731
732 **Fig. S15. Performance of models built with four algorithms using the entire and subsets of**
733 **scRNA-seq data from five different tissues**. KNN (**A**), RF (**B**), XGBoost (**C**), and AutoGluon
734 (**D**) models were built using the K-Nearest Neighbors, Random Forest, eXtreme Gradient
735 Boosting, and AutoGluon-Tabular algorithms, respectively. Left panel: $F1_{CV}$; right panel: $F1_{test}$.
736 Models built using different scRNA-seq datasets are indicated using different colors: dark blue,
737 the entire scRNA-seq data; dark green, scRNA-seq data from root; green, those from shoot; olive,
738 leaf; orange, flower; pink, embryo. Error bars indicate the standard deviation of results from 10
739 replicate runs. Asterisk(s) indicate(s) significant levels for the two-sided Wilcoxon rank-sum test,
740 showing that the models' performance using tissue subset data is significantly higher than when
741 using the entire scRNA-seq data: *, *p*-value < 0.05; **, *p*-value < 0.01; ***: *p*-value < 0.001. EM:
742 energy metabolism; Cof: cofactors; IM: intermediate metabolism; DTX: detoxification; AA:

743 amino acids; NT: nucleotides; AP: amines and polyamines; Carb: carbohydrates; FAL: fatty
744 acids and lipids; IN: inorganic nutrients; SM: specialized metabolism; Horm: hormones.
745
746 **Fig. S16. Performance of models built with four algorithms using bulk RNA-seq (blue),**
747 **scRNA-seq data (orange) and the combined data (gray).** KNN (**A**), RF (**B**), XGBoost (**C**),
748 and AutoGluon (**D**) models were built using K-Nearest Neighbors, Random Forest, eXtreme
749 Gradient Boosting, and AutoGluon-Tabular algorithms, respectively. Left panel: $F1_{CV}$; right
750 panel: $F1_{test}$. Error bars indicate the standard deviation of results from 10 replicate runs.
751 Asterisk(s) indicate(s) significant levels for the two-sided Wilcoxon rank-sum test: *, p-value <
752 0.05; **, p-value < 0.01; ***: p-value < 0.001. The asterisks are only marked for the classes for
753 which the combined data-based models significantly outperformed both models built using bulk
754 RNA-seq and scRNA-seq data. EM: energy metabolism; Cof: cofactors; IM: intermediate
755 metabolism; DTX: detoxification; AA: amino acids; NT: nucleotides; AP: amines and
756 polyamines; Carb: carbohydrates; FAL: fatty acids and lipids; IN: inorganic nutrients; SM:
757 specialized metabolism; Horm: hormones.
758
759 **Fig. S17. Global feature importance for RF (A), XGBoost (B) and AutoGluon (C) models**
760 **built using bulk RNA-seq data**. Bar chart shows the average importance of bulk RNA-seq
761 features for 10 replicate runs; bar colors indicate different tissue types to which the features
762 belong: dark green, root; green, shoot; olive, leaf; orange, flower; pink, embryo; light pink,
763 others.
764
765 **Fig. S18. Global feature importance for RF (A), XGBoost (B) and AutoGluon (C) models**
766 **built using scRNA-seq data.** The bar chart shows the average importance of scRNA-seq
767 features for 10 replicate runs; bar colors indicate different tissue types to which the features
768 belong: dark green, root; green, shoot; olive, leaf; orange, flower; pink, embryo; light pink,
769 others.
770
771 **Fig. S19. Feature importance for individual metabolic classes for AutoGluon models when**
772 **scRNA-seq (A) and bulk RNA-seq (B) data were used**. Colors within the heatmaps indicate
773 the average importance of features for 10 replicate runs; the background colors of the texts

indicate different tissue types to which the features belong: dark green, root; green, shoot; olive, leaf; orange, flower; pink, embryo; light pink, others.

Fig. S20. Prediction of metabolic classes for unknown genes by XGBoost models. (**A**) Count of unknown genes that were predicted into 13 metabolic classes by XGBoost models built using bulk RNA-seq (upper panel) and scRNA-seq data (lower). (**B**) Venn diagrams showing the overlap of unknown genes that were predicted into metabolic classes by XGBoost models built using bulk RNA-seq (blue) and scRNA-seq data (orange), and those that were annotated as within metabolic classes in the PMN database (purple). In the upper and lower panels, overlapping regions show the number of genes with exactly the same classes or at least one common class between datasets, respectively. (**C**,**D**) Heatmaps showing the numbers of unknown genes that were predicted as in a class (x-axis) by XGBoost models built using bulk RNA-seq (**C**) and scRNA-seq data (**D**), and were also annotated as within the same or other classes (y-axis) by PMN database. Values in the column "Not-assigned" indicate the numbers of unknown genes that were predicted into neither any of the 13 metabolic classes nor the "Others" category. Cell colors in (**C,D**) indicate the numbers of genes. EM: energy metabolism; Cof: cofactors; IM: intermediate metabolism; DTX: detoxification; AA: amino acids; NT: nucleotides; AP: amines and polyamines; Carb: carbohydrates; FAL: fatty acids and lipids; IN: inorganic nutrients; SM: specialized metabolism; Horm: hormones.

Supplementary tables

Table S1. Information of 13 single-cell RNA sequencing (scRNA-seq) datasets.

Table S2. Information of five bulk RNA-seq datasets.

Table S3. Gene expression matrix of scRNA-seq data.

Table S4. Gene expression matrix of bulk RNA-seq data.

Table S5. 10,572 enzyme-encoding genes obtained from the PMN, TAIR databases, or annotated with EC numbers using the E2P2 software.

Table S6. 14 gene co-expression modules determined by WGCNA using bulk RNA-seq data.

Table S7. 14 gene co-expression modules determined by WGCNA using scRNA-seq data.

Table S8. Clustering coefficients of metabolic pathway genes within each class based on two types of RNA-seq data and four tissue-specific scRNA-seq data.

Table S9. Clustering coefficients for 10,000 random simulations calculated using two types of RNA-seq data.

Table S10. Clustering coefficients for 10,000 random simulations calculated using four tissue-specific scRNA-seq data.

Table S11. Overall $F1_{CV}$ and $F1_{test}$ scores for 1,000 random simulations of KNN, RF, and XGBoost models and for 100 random simulations of AutoGluon model when bulk RNA-seq data were used.

Table S12. Overall $F1_{CV}$ and $F1_{test}$ scores for 1,000 random simulations of KNN, RF, and XGBoost models and for 100 random simulations of AutoGluon model when scRNA-seq data were used.

Table S13. $F1_{CV}$ and $F1_{test}$ scores for 1,000 random simulations of KNN models when bulk RNA-seq data were used.

Table S14. $F1_{CV}$ and $F1_{test}$ scores for 1,000 random simulations of KNN models when scRNA-seq data were used.

Table S15. $F1_{CV}$ and $F1_{test}$ scores for 1,000 random simulations of RF models when bulk RNA-seq data were used.

Table S16. $F1_{CV}$ and $F1_{test}$ scores for 1,000 random simulations of RF models when scRNA-seq data were used.

Table S17. $F1_{CV}$ and $F1_{test}$ scores for 1,000 random simulations of XGBoost models when bulk RNA-seq data were used.

Table S18. $F1_{CV}$ and $F1_{test}$ scores for 1,000 random simulations of XGBoost models when scRNA-seq data were used.

Table S19. $F1_{CV}$ and $F1_{test}$ scores for 100 random simulations of AutoGluon models when bulk RNA-seq data were used.

Table S20. $F1_{CV}$ and $F1_{test}$ scores for 100 random simulations of AutoGluon models when scRNA-seq data were used.

Table S21. Global feature importance of 40 bulk RNA-seq features for RF models.

Table S22. Global feature importance of 146 scRNA-seq features for RF models.

Table S23. Global feature importance of 40 bulk RNA-seq features for XGBoost models.

Table S24. Global feature importance of 146 scRNA-seq features for XGBoost models.

834 **Table S25.** Feature importance for individual metabolic classes of 40 bulk RNA-seq features for
835 AutoGluon models.
836 **Table S26.** Feature importance for individual metabolic classes of 146 scRNA-seq features for
837 AutoGluon models.
838 **Table S27.** Prediction of 9,441 unknown genes into individual metabolic classes by the
839 AutoGluon model when bulk RNA-seq data were used.
840 **Table S28.** Prediction of 9,441 unknown genes into individual metabolic classes by the
841 AutoGluon model when scRNA-seq data were used.
842 **Table S29.** Prediction of 9,441 unknown genes into individual metabolic classes by the
843 XGBoost model when bulk RNA-seq data were used.
844 **Table S30.** Prediction of 9,441 unknown genes into individual metabolic classes by the
845 XGBoost model when scRNA-seq data were used.

846 References

847 **Ahmadi MK, Fang L, Moscatello N, Pfeifer BA**. **2016**. *E. coli* metabolic engineering for gram scale
848 production of a plant-based anti-inflammatory agent. *Metabolic Engineering* **38**: 382–388.
849 **Al-Khayri JM, Rashmi R, Toppo V, Chole PB, Banadka A, Sudheer WN, Nagella P, Shehata WF,**
850 **Al-Mssallem MQ, Alessa FM, Almaghasla MI, Rezk AA-S**. **2023**. Plant secondary metabolites:
851 the weapons for biotic stress management. *Metabolites* **13**: 716.
852 **Appelhagen I, Wulff-Vester AK, Wendell M, Hvoslef-Eide A-K, Russell J, Oertel A, Martens S,**
853 **Mock H-P, Martin C, Matros A**. **2018**. Colour bio-factories: towards scale-up production of
854 anthocyanins in plant cell cultures. *Metabolic Engineering* **48**: 218–232.
855 **Bhambhani S, Kondhare KR, Giri AP**. **2021**. Diversity in chemical structures and biological properties
856 of plant alkaloids. *Molecules* **26**: 3374.
857 **Breiman L**. **2001**. Random Forests. *Machine Learning* **45**: 5–32.
858 **Chae L, Kim T, Nilo-Poyanco R, Rhee SY**. **2014**. Genomic signatures of specialized metabolism in
859 plants. *Science* **344**: 510–513.
860 **Chen T, Guestrin C**. **2016**. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM
861 SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco
862 California USA: ACM, 785–794.
863 **Daloso DDM, Morais EG, Oliveira E Silva KF, Williams TCR**. **2023**. Cell-type-specific metabolism in
864 plants. *The Plant Journal* **114**: 1093–1114.

865 **Dong X, Gao Y, Chen W, Wang W, Gong L, Liu X, Luo J**. **2015**. Spatiotemporal distribution of
866 phenolamides and the genetics of natural variation of hydroxycinnamoyl spermidine in rice.
867 *Molecular Plant* **8**: 111–121.
868 **Elshafie HS, Camele I, Mohamed AA**. **2023**. A comprehensive review on the biological, agricultural
869 and pharmaceutical properties of secondary metabolites based-plant origin. *International Journal*
870 *of Molecular Sciences* **24**: 3266.
871 **Erickson N, Mueller J, Shirkov A, Zhang H, Larroy P, Li M, Smola A**. **2020**. AutoGluon-tabular:
872 robust and accurate AutoML for structured data. https://doi.org/10.48550/arXiv.2003.06505.
873 **Fang C, Fernie AR, Luo J**. **2019**. Exploring the diversity of plant metabolism. *Trends in Plant Science*
874 **24**: 83–98.
875 **Guo G, Wang H, Bell D, Bi Y, Greer K**. **2003**. KNN model-based approach in classification. In: On The
876 Move to Meaningful Internet Systems: CoopIS DOA and ODBASE. Berlin, Germany: Springer,
877 986–996.
878 **Hagberg A, Swart PJ, Schult DA**. **2008**. Exploring network structure, dynamics, and function using
879 networkx. In: Proceedings of the 7th Python in Science Conference. Los Alamos National
880 Laboratory (LANL), Los Alamos, NM (United States), 11–15.
881 **Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C,**
882 **Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T,**
883 **Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija**
884 **R**. **2021**. Integrated analysis of multimodal single-cell data. *Cell* **184**: 3573–3587.
885 **Jucá MM, Cysne Filho FMS, De Almeida JC, Mesquita DDS, Barriga JRDM, Dias KCF, Barbosa**
886 **TM, Vasconcelos LC, Leal LKAM, Ribeiro JE, Vasconcelos SMM**. **2020**. Flavonoids:
887 biological activities and therapeutic potential. *Natural Product Research* **34**: 692–705.
888 **Kaiser M**. **2008**. Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures
889 for small-world networks. *New Journal of Physics* **10**: 083042.
890 **Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH**. **2017**. plantiSMASH: automated
891 identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic*
892 *Acids Research* **45**: W55–W63.
893 **Ke M, Elshenawy B, Sheldon H, Arora A, Buffa FM**. **2022**. Single cell RNA-sequencing: a powerful
894 yet still challenging technology to study cellular heterogeneity. *BioEssays* **44**: 2200084.
895 **Kharchenko PV**. **2021**. The triumphs and limitations of computational methods for scRNA-seq. *Nature*
896 *Methods* **18**: 723–732.
897 **Li Y, Leveau A, Zhao Q, Feng Q, Lu H, Miao J, Xue Z, Martin AC, Wegel E, Wang J, Orme A, Rey**
898 **M-D, Karafiátová M, Vrána J, Steuernagel B, Joynson R, Owen C, Reed J, Louveau T,**

Stephenson MJ, Zhang L, Huang X, Huang T, Fan D, Zhou C, Tian Q, Li W, Lu Y, Chen J, Zhao Y, Lu Y, Zhu C, Liu Z, Polturak G, Casson R, Hill L, Moore G, Melton R, Hall N, Wulff BBH, Doležel J, Langdon T, Han B, Osbourn A. **2021**. Subtelomeric assembly of a multi-gene pathway for antimicrobial defense compounds in cereals. *Nature Communications* **12**: 2563.

Magnard J-L, Roccia A, Caissard J-C, Vergne P, Sun P, Hecquet R, Dubois A, Hibrand-Saint Oyant L, Jullien F, Nicolè F, Raymond O, Huguet S, Baltenweck R, Meyer S, Claudel P, Jeauffre J, Rohmer M, Foucher F, Hugueney P, Bendahmane M, Baudino S. **2015**. Biosynthesis of monoterpene scent compounds in roses. *Science* **349**: 81–83.

McMahon JM, White WL, Sayre RT. **1995**. Cyanogenesis in cassava (*Manihot esculenta* crantz). *Journal of Experimental Botany* **46**: 731–741.

Mergner J, Frejno M, List M, Papacek M, Chen X, Chaudhary A, Samaras P, Richter S, Shikata H, Messerer M, Lang D, Altmann S, Cyprys P, Zolg DP, Mathieson T, Bantscheff M, Hazarika RR, Schmidt T, Dawid C, Dunkel A, Hofmann T, Sprunck S, Falter-Braun P, Johannes F, Mayer KFX, Jürgens G, Wilhelm M, Baumbach J, Grill E, Schneitz K, Schwechheimer C, Kuster B. **2020**. Mass-spectrometry-based draft of the arabidopsis proteome. *Nature* **579**: 409–414.

Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Lehti-Shiu MD, Last RL, Pichersky E, Shiu S-H. **2019**. Robust predictions of specialized metabolism genes through machine learning. *Proceedings of the National Academy of Sciences, USA* **116**: 2344–2353.

Onnela J-P, Saramäki J, Kertész J, Kaski K. **2005**. Intensity and coherence of motifs in weighted complex networks. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics* **71**: 065103.

Opsahl T, Panzarasa P. **2009**. Clustering in weighted networks. *Social Networks* **31**: 155–163.

Papadopoulou K, Melton RE, Leggett M, Daniels MJ, Osbourn AE. **1999**. Compromised disease resistance in saponin-deficient plants. *Proceedings of the National Academy of Sciences, USA* **96**: 12923–12928.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D. **2011**. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12**: 2825–2830.

Schilmiller AL, Last RL, Pichersky E. **2008**. Harnessing plant trichome biochemistry for the production of useful compounds. *The Plant Journal* **54**: 702–711.

Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T, Kahn D, Rhee SY. **2017**. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology* **173**: 2041–2059.

933 **Shang Y, Ma Y, Zhou Y, Zhang H, Duan L, Chen H, Zeng J, Zhou Q, Wang S, Gu W, Liu M, Ren J,**
934 **Gu X, Zhang S, Wang Y, Yasukawa K, Bouwmeester HJ, Qi X, Zhang Z, Lucas WJ, Huang**
935 **S**. **2014**. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**:
936 1084–1088.
937 **Song Q, Ando A, Jiang N, Ikeda Y, Chen ZJ**. **2020**. Single-cell RNA-seq analysis reveals ploidy-
938 dependent and cell-specific transcriptome changes in *Arabidopsis* female gametophytes. *Genome*
939 *Biology* **21**: 178.
940 **Sun S, Shen X, Li Y, Li Y, Wang S, Li R, Zhang H, Shen G, Guo B, Wei J, Xu J, St-Pierre B, Chen**
941 **S, Sun C**. **2022**. Single-cell RNA sequencing provides a high-resolution roadmap for
942 understanding the multicellular compartmentation of specialized metabolism. *Nature Plants* **9**:
943 179–190.
944 **Tiwari R, Rana CS**. **2015**. Plant secondary metabolites: a review. *International Journal of Engineering*
945 *Research and General Science* **3**: 661–670.
946 **Tohge T, Scossa F, Wendenburg R, Frasse P, Balbo I, Watanabe M, Alseekh S, Jadhav SS, Delfin**
947 **JC, Lohse M, Giavalisco P, Usadel B, Zhang Y, Luo J, Bouzayen M, Fernie AR**. **2020**.
948 Exploiting natural variation in tomato to define pathway structure and metabolic regulation of
949 fruit polyphenolics in the lycopersicum complex. *Molecular Plant* **13**: 1027–1046.
950 **Töpfer N, Fuchs L-M, Aharoni A**. **2017**. The PhytoClust tool for metabolic gene clusters discovery in
951 plant genomes. *Nucleic Acids Research* **45**: 7049–7063.
952 **Wang P, Moore BM, Uygun S, Lehti-Shiu MD, Barry CS, Shiu S**. **2021**. Optimising the use of gene
953 expression data to predict plant metabolic pathway memberships. *New Phytologist* **231**: 475–489.
954 **Yang Y-H, Mao J-W, Tan X-L**. **2020**. Research progress on the source, production, and anti-cancer
955 mechanisms of paclitaxel. *Chinese Journal of Natural Medicines* **18**: 890–897.
956 **Yu C, Hou K, Zhang H, Liang X, Chen C, Wang Z, Wu Q, Chen G, He J, Bai E, Li X, Du T, Wang**
957 **Y, Wang M, Feng S, Wang H, Shen C**. **2023**. Integrated mass spectrometry imaging and
958 single-cell transcriptome atlas strategies provide novel insights into taxoid biosynthesis and
959 transport in *Taxus mairei* stems. *The Plant Journal* **115**: 1243–1260.
960 **Zhang B, Horvath S**. **2005**. A general framework for weighted gene co-expression network analysis.
961 *Statistical Applications in Genetics and Molecular Biology* **4**: 1–45.
962 **Zhang S, Ye T, Liu Y, Hou G, Wang Q, Zhao F, Li F, Meng Q**. **2023**. Research advances in clinical
963 applications, anticancer mechanism, total chemical synthesis, semi-synthesis and biosynthesis of
964 paclitaxel. *Molecules* **28**: 7517.

Fig. 1

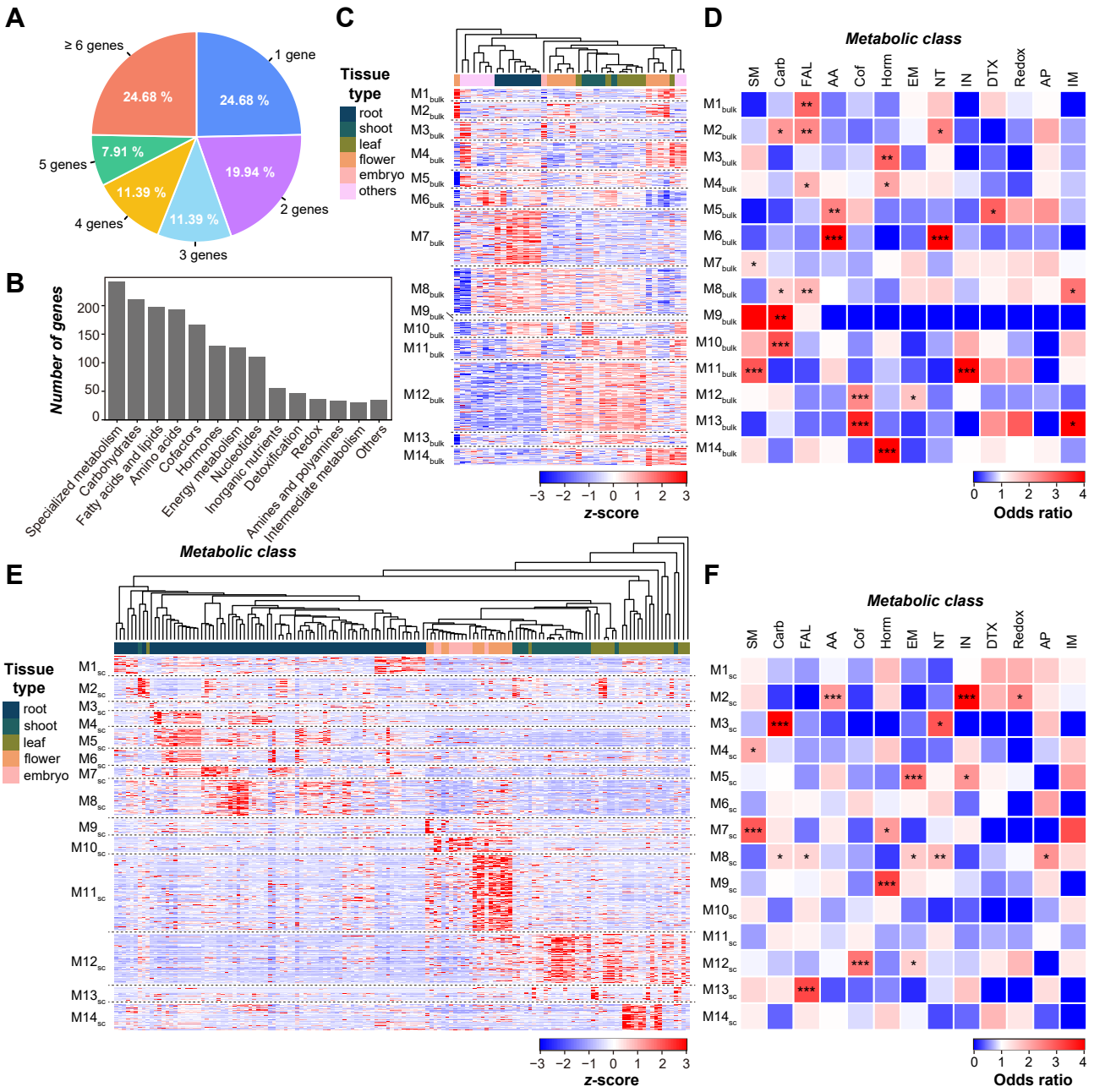

A
≥ 6 genes
1 gene
24.68 %
24.68 %
7.91 %
5 genes
19.94 %
11.39 %
2 genes
4 genes
11.39 %
3 genes
B
Number of genes
200
150
100
50
0
Specialized metabolism
Carbohydrates
Fatty acids and lipids
Amino acids
Cofactors
Hormones
Energy metabolism
Nucleotides
Inorganic nutrients
Detoxification
Redox
Amines and polyamines
Intermediate metabolism
Others
C
Tissue type
root
shoot
leaf
flower
embryo
others
M1bulk
M2bulk
M3bulk
M4bulk
M5bulk
M6bulk
M7bulk
M8bulk
M9bulk
M10bulk
M11bulk
M12bulk
M13bulk
M14bulk
−3 −2 −1 0 1 2 3
z-score
D
Metabolic class
SM Carb FAL AA Cof Horm EM NT IN DTX Redox AP IM
0 1 2 3 4
Odds ratio
E
Metabolic class
Tissue type
root
shoot
leaf
flower
embryo
M1sc
M2sc
M3sc
M4sc
M5sc
M6sc
M7sc
M8sc
M9sc
M10sc
M11sc
M12sc
M13sc
M14sc
−3 −2 −1 0 1 2 3
z-score
F
Metabolic class
SM Carb FAL AA Cof Horm EM NT IN DTX Redox AP IM
0 1 2 3 4
Odds ratio

Fig. 2

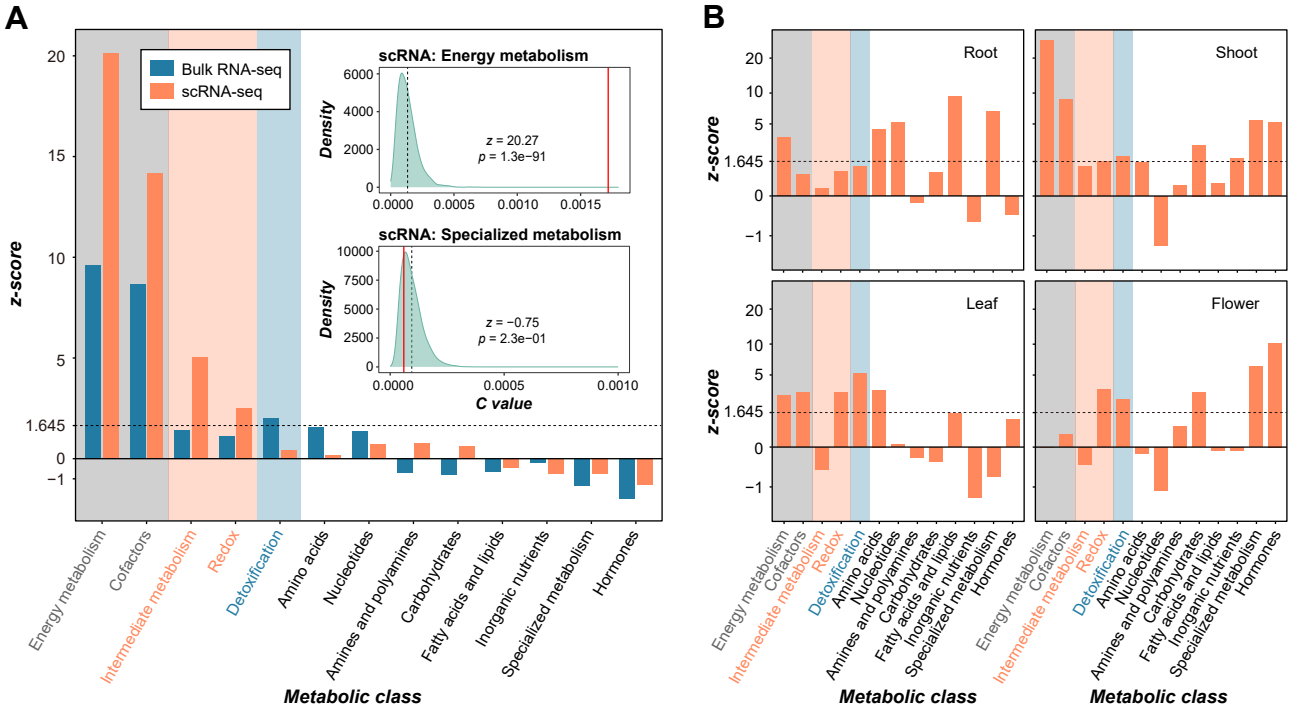

A
Bulk RNA-seq
scRNA-seq
z-score
scRNA: Energy metabolism
Density
z = 20.27
p = 1.3e−91
scRNA: Specialized metabolism
z = −0.75
p = 2.3e−01
C value
Energy metabolism
Cofactors
Intermediate metabolism
Redox
Detoxification
Amino acids
Nucleotides
Amines and polyamines
Carbohydrates
Fatty acids and lipids
Inorganic nutrients
Specialized metabolism
Hormones
Metabolic class
B
Root
Shoot
Leaf
Flower
z-score
Metabolic class

Fig. 3

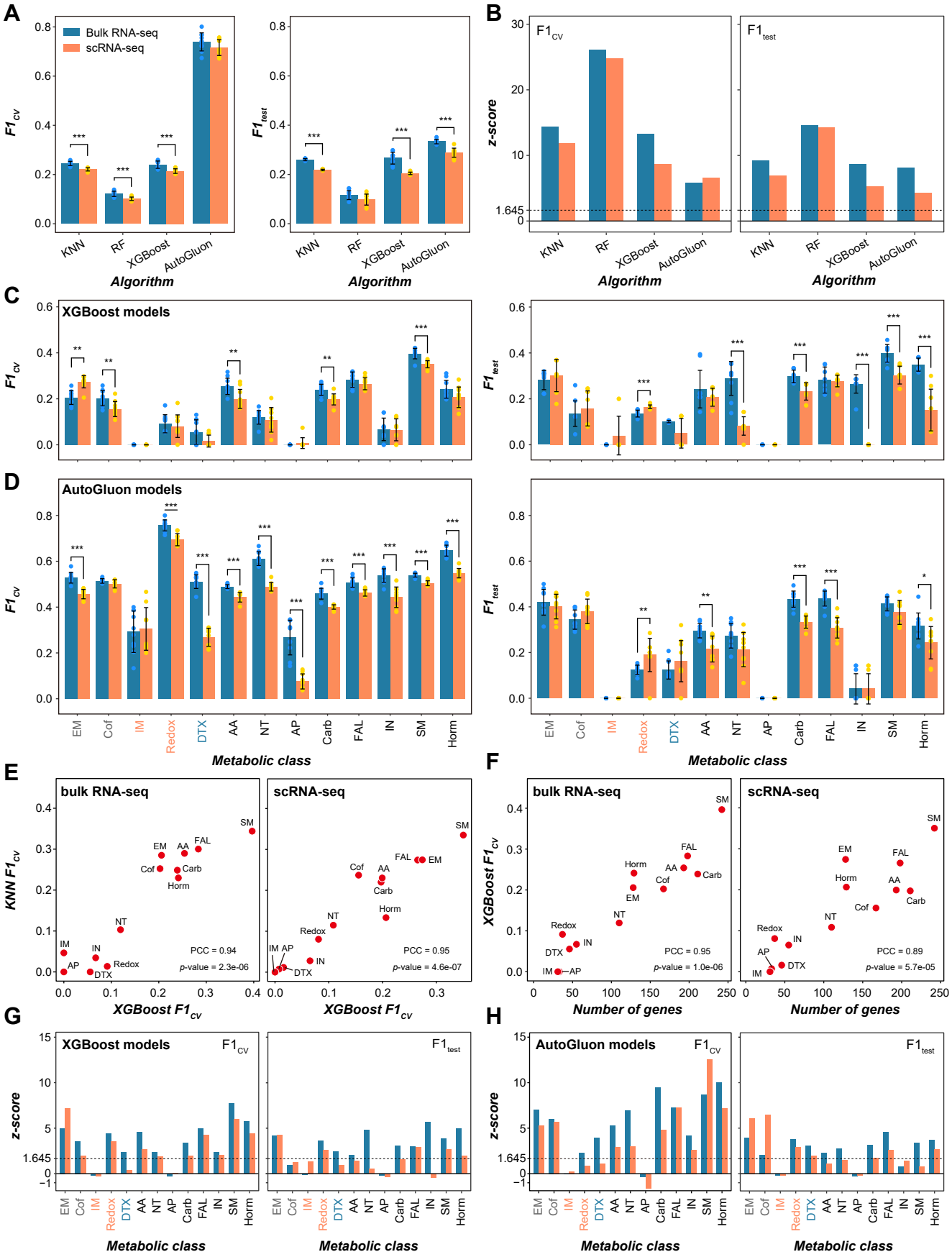

A
Bulk RNA-seq
scRNA-seq
F1cv
F1test
KNN
RF
XGBoost
AutoGluon
Algorithm
B
F1cv
F1test
z-score
1.645
C
XGBoost models
D
AutoGluon models
EM
Cof
IM
Redox
DTX
AA
NT
AP
Carb
FAL
IN
SM
Horm
Metabolic class
E
bulk RNA-seq
scRNA-seq
KNN F1cv
XGBoost F1cv
PCC = 0.94
p-value = 2.3e-06
PCC = 0.95
p-value = 4.6e-07
F
XGBoost F1cv
Number of genes
PCC = 0.95
p-value = 1.0e-06
PCC = 0.89
p-value = 5.7e-05
G
H

Fig. 4

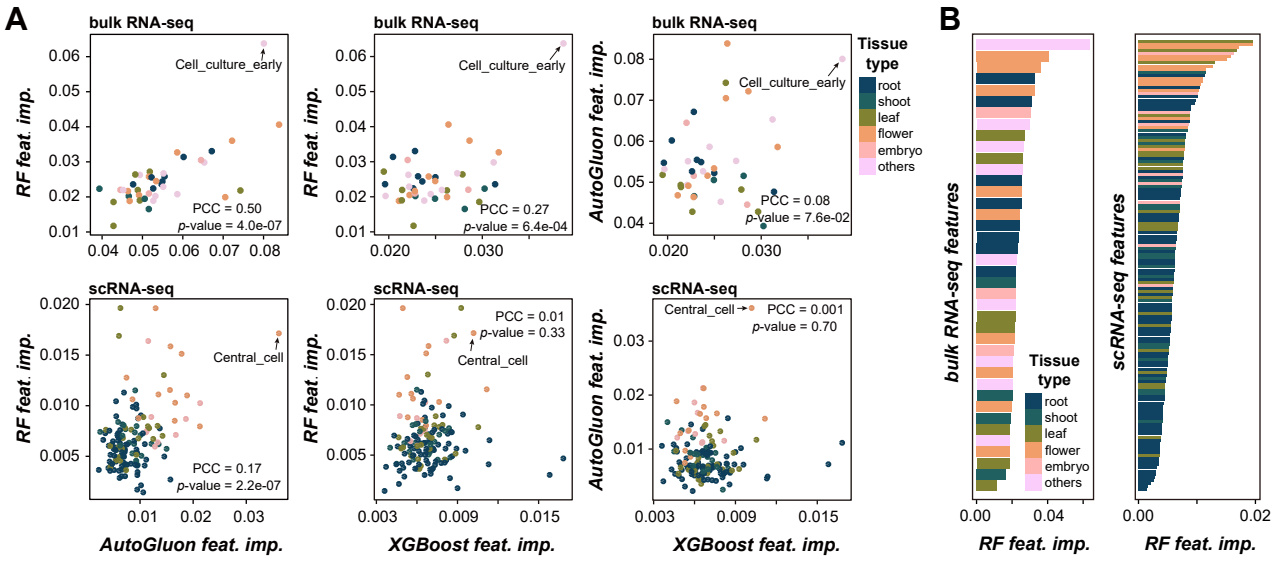
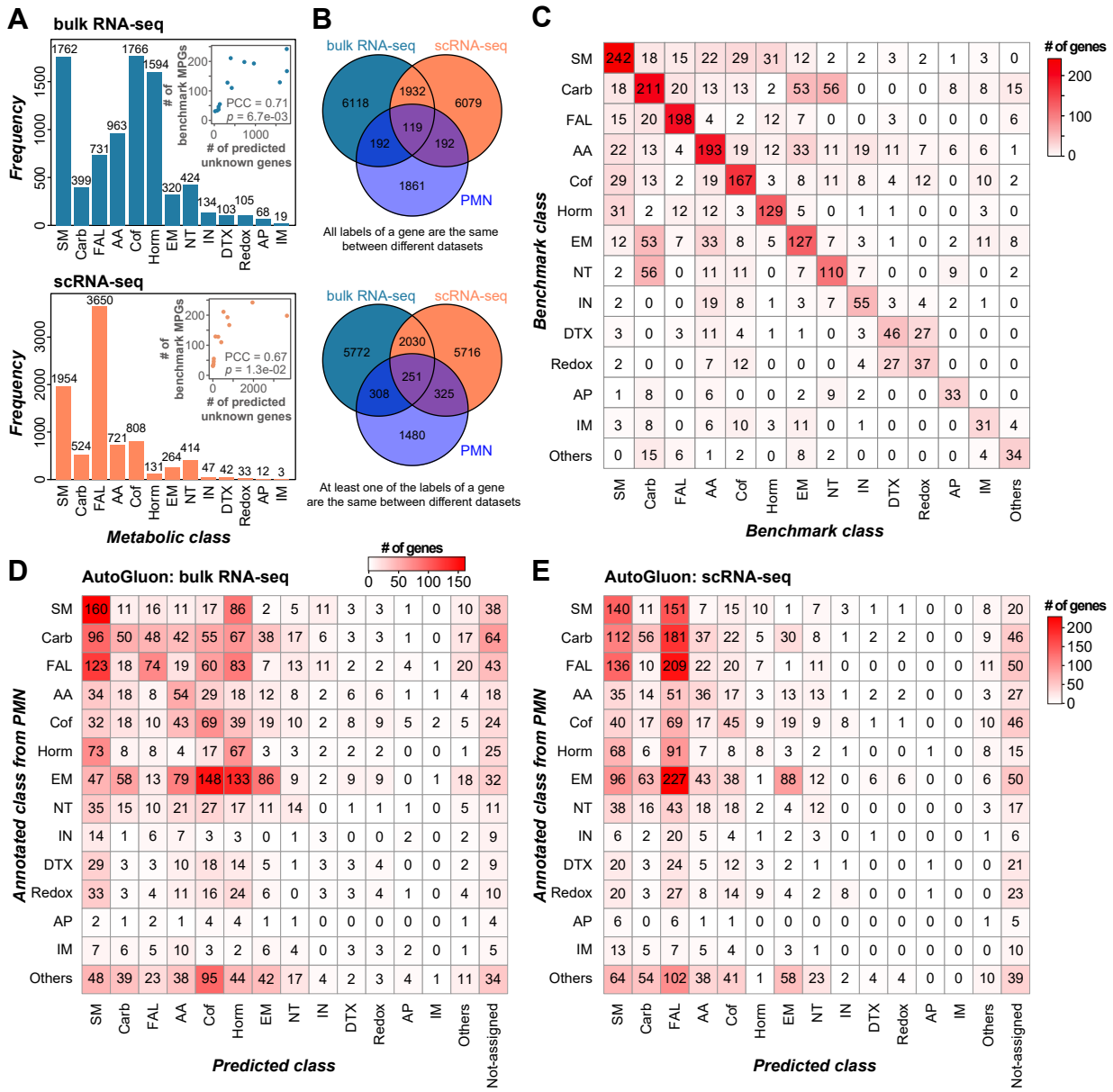

A
bulk RNA-seq
Cell_culture_early
PCC = 0.50
p-value = 4.0e-07
RF feat. imp.
PCC = 0.27
p-value = 6.4e-04
AutoGluon feat. imp.
PCC = 0.08
p-value = 7.6e-02
Tissue type
root
shoot
leaf
flower
embryo
others
scRNA-seq
Central_cell
PCC = 0.17
p-value = 2.2e-07
PCC = 0.01
p-value = 0.33
PCC = 0.001
p-value = 0.70
XGBoost feat. imp.
B
bulk RNA-seq features
scRNA-seq features
RF feat. imp.

Fig. 5


A
bulk RNA-seq
Frequency
1762
399
731
963
1766
1594
320
424
134
103
105
68
19
SM Carb FAL AA Cof Horm EM NT IN DTX Redox AP IM
# of benchmark MPGs
PCC = 0.71
p = 6.7e-03
# of predicted unknown genes
scRNA-seq
1954
524
3650
721
808
131
264
414
47
42
33
12
3
PCC = 0.67
p = 1.3e-02
Metabolic class
B
bulk RNA-seq
scRNA-seq
6118
1932
6079
119
192
192
1861
PMN
All labels of a gene are the same between different datasets
5772
2030
5716
251
308
325
1480
At least one of the labels of a gene are the same between different datasets
C
Benchmark class
# of genes
D
AutoGluon: bulk RNA-seq
Annotated class from PMN
Predicted class
Not-assigned
E
AutoGluon: scRNA-seq