Optimizing the use of gene expression data to predict metabolic pathway memberships with unsupervised and supervised machine learning

Peipei Wang[1], Bethany M. Moore[1,2], Sahra Uygun[3], Melissa D. Lehti-Shiu[1], Cornelius Barry[4], and Shin-Han Shiu[1,2,5] *

[1]Department of Plant Biology, [2]Ecology, Evolutionary Biology, and Behavior Program, [4]Department of Horticulture, [5]Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824
[3]Agendia Inc. Irvine, CA 92618

Abstract

Plants produce diverse metabolites via enzymes metabolic pathways important for plant survival, human nutrition and medicine. However, most plant enzyme genes are of unknown pathway membership. While some genes in the same pathways can be identified based on correlated expression, such correlation may exist only under specific spatiotemporal and conditional contexts. By considering 656 combinations of tomato gene expression datasets calculated with eight co-expression measures, we evaluated the performance of naive prediction (based on expression similarities to pathways), unsupervised and supervised learning methods in predicting memberships in 85 metabolic pathways. We found that optimal predictions for different pathways require different dataset, which tend to be associated with the biological processes related to the pathway functions. In addition, naive prediction has significantly lower performance compared to machine learning methods. Interestingly, the unsupervised learning approach has better performance in 52 pathways than the supervised approach, which may be attributed to the need for more data with supervised learning. Furthermore, machine learning clustering/models using gene-to-pathway expression similarities outperform that with gene expression profiles. Altogether, our study highlights the need to extensively explore expression-based features to maximize the utility of expression data for pinpointing pathway membership. Through this detailed exploration, novel connections between pathways and biological processes can also be identified based on the optimal expression dataset used, improving our mechanistic understanding of the metabolic network.

Introduction

Metabolites are products of metabolic pathways comprised of a linked series of chemical reactions catalyzed by enzymes (Berg et al., 2002). Plants produce diverse metabolites essential for plant growth, development, survival and adaptation (Stitt et al., 2010). Some of these metabolites are also important for human nutrition and medicine (Martin and Li, 2017; Schmidt et al., 2007). Gaining knowledge about metabolic pathways is important for understanding the machine of life, underlying causes of metabolites diversity and for metabolic engineering. Although we have amassed substantial knowledge in plant metabolic pathways (Verpoorte, 1998; Pichersky and Gang, 2000; Kim and Buell, 2015), some genes responsible for biosynthesis of these plant metabolites in known pathways remain to be identified (Schläpfer et al., 2017; De Luca et al., 2012). The challenge is that genes encoding metabolic enzymes tend to be in large gene families, like BAHD Acyltransferases (Yu et al., 2009). In addition, experimental assessments of pathway memberships by biochemical and genetic approaches can be laborious. Therefore, it is important to prioritize candidate genes that belong to various pathways with computational predictions.

Substantial effort has been devoted to predicting metabolic pathway membership of plant enzyme genes. One approach is to first identify candidate enzyme genes from a plant genome, and then assign these enzyme genes into pathways according to the potential reactions the enzymes may participate in (Schläpfer et al., 2017; Chae et al., 2014). This approach has led to the important community resource Plant Metabolic Network database currently with 125 plant and green alga species annotated. Here the prediction of candidate enzyme genes was mainly based on sequence similarity of unknown genes to experimentally evaluated enzyme genes in model species.

Another approach to predict metabolic pathway genes is to associate unknown genes with existing pathway genes by exploring gene expression similarity based on the assumption that genes with similar functions, specifically in the same pathway, have similar expression profiles (Kim and Buell, 2015; Segal et al., 2003). Considering the ease of generating and increasing availability of transcriptome data in model and non-model species, it is an important resource for computational inference of gene functions. There are three general strategies to leverage expression data for functional inference. The first strategy, referred to as “naive prediction” is by asking, for a gene of unknown function, which genes with known functions have sufficiently high expression similarity to the gene in question. Despite its simplicity, the utility of such a strategy has been shown in a number of single gene studies (Huang et al., 2019; Hirai et al., 2007; Righetti et al., 2015; Peng et al., 2015), but it is not clear globally how accurate such strategy is. The second strategy is to globally infer functional association between plant genes through unsupervised learning, where genes are first grouped into co-expression clusters and then genes of unknown functions are assigned functions based on what types of genes with known functions are over-represented within clusters. This strategy

is by far the most widely used (Abu-Jamous and Kelly, 2018; Gupta and Pereira, 2019; Serin et al., 2016; Uygun et al., 2016; Wisecaver et al., 2017; Mutwil et al., 2011). The third strategy is supervised learning where the function of a gene can be predicted with models learned from the expression profiles of genes with known functions. While supervised learning has been applied to predict functions in other contexts (Lloyd et al., 2015; Moore et al., 2019; Kaundal et al., 2010; Ni et al., 2016; Libbrecht and Noble, 2015; Lan et al., 2007), it has not been applied to predicting plant metabolic pathway memberships using gene expression data. Currently, it is not known which strategy may be more effective and for what pathways.

Beyond the specific computational strategies, it is also unresolved how gene expression data should be used in pathway membership prediction. For example, it has been shown that, rather than using as many different conditions as possible (Obayashi et al., 2018; Aoki et al., 2016), expression similarities measured using distinct subsets of expression data can be important for inferring functional relations for specific pathways (Uygun et al., 2016; Usadel et al., 2009; Wisecaver et al., 2017). Thus, in addition to computational strategy, it is important to optimize the use of specific expression data in metabolic pathway prediction. To accomplish this, we used tomato as a model because it is an important crop model species (Quinet et al., 2019; Kimura and Sinha, 2008; Meissner et al., 1997). Most importantly, there is substantial knowledge on metabolic pathways in tomato (Mathieu et al., 2009; Leong et al., 2019; Fan et al., 2017; Matsuba et al., 2015; Brasher et al., 2015; Carrari and Fernie, 2006). Using tomato, we explored the impact of expression datasets, expression values (absolute expression values or fold changes), and gene expression similarity measures in predicting pathway memberships using three strategies: naive prediction, unsupervised learning, and supervised learning.

Results & Discussion

Relationships between gene expression similarity and metabolic pathway memberships

We first examined the extent to which genes within the same metabolic pathways have similar expression profiles. To test this, we collected annotation data of 2171 tomato genes in 297 pathways with ≥ 5 annotated genes each, and transcriptome data from 372 samples (each sample includes multiple biological/technical replicates) (see **Methods**). Using Pearson's correlation coefficient (PCC) of enzyme gene expression levels (Fragments Per Kilobase Million, FPKM) calculated with all samples as the expression similarity measure, we found that 116 out of 297 pathways (39.1%) have significantly higher gene expression similarities between pairs of genes from the same pathways than those between pairs of genes randomly chosen from different pathways (Wilcoxon signed-rank test, $p < 0.05$; **Fig. 1A**). When the maximum expression similarity

of a gene to other genes in the same pathway was examined, 202 of 297 pathways (68.0%) had significantly higher expression similarities than those between pathways (**Fig. S1**). This finding was not simply due to the contribution of paralogs that might have higher expression similarities than non-paralogous pairs (**Fig. S1**). These results suggest that expression data can be useful for distinguishing genes within pathways from genes between pathways. In addition, the maximum expression similarity within pathway was more useful to distinguish genes within and between pathways. However, even using maximum PCC as a criterion, 72.7% of genes do not have maximum PCC values within pathway above threshold PCC (PCC=0.66, the 95 percentile value of between pathway pairs). Thus, we asked next how similarity measures should be generated to maximally identify pathway memberships.

For any gene pair, its expression similarity is influenced by three factors: (1) which expression dataset is used, e.g., all data or a subset ; (2) what expression value is used, i.e., expression levels or contrasts in the form of fold change ; and (3) which similarity measure is used, e.g., PCC vs. Spearman's Rank. For a pair of genes in a pathway, we were not interested in their absolute expression similarity. Rather, we were interested in how their similarity was different from a distribution of between pathway gene similarities that we treated as the background, null distribution. Thus, in all subsequent analyses, we determined the percentile value of each within-pathway similarity based on the between-pathway similarity distribution. By calculating the median of percentile values for all within pathway pairs, the higher median percentile value indicated genes in that pathway had higher similarity compared to between pathway pairs. In the above analysis, all 372 samples were used. To evaluate whether different datasets may be more useful for determining expression similarities of genes in different pathways, 41 datasets including 36 from individual studies, and five combined sets consisted of samples from (1) tissue/developmental stage/circadian clock experiments, (2) different genetic backgrounds (e.g., mutants), (3) different hormone treatments, (4) different stress treatments, and (5) all 372 samples (**Table S3 and S4**, see **Methods**) were examined.

In **Fig. 1B**, using FPKM as the expression value and PCC as the similarity measure, the best datasets for the 13 example pathways differ substantially (results for all the pathways in **Table S7**). Some pathways, such as those involved in photosynthesis (**Fig. 1B**), tended to have high PCC percentiles for most datasets used. While others were much more sensitive to the datasets used, this may reflect that some specific expression experiments were more relevant to pathway functions. For example, the fatty acid α-oxidation I pathway led to products that protects plants against oxidative stress (De León et al., 2002). The chlorogenic acid biosynthesis II pathway protects plants from abiotic stresses (Burchard et al., 2000; Niggeweg et al., 2004). These two pathways have the highest PCC percentiles when the stress dataset was used (**Fig. 1B**). These results demonstrate the need to consider datasets that best reflect the

biological processes different pathways are involved in. Instead of using absolute expression levels (FPKMs), using fold change values (FC, see **Methods**) has led to improved expression similarity percentiles for all example pathways (**Fig. 1C**). In fact, 250 of 297 pathways had higher percentile values than when absolute expression levels were considered (**Fig. 1D**), demonstrating the impact of expression values used.

Thus far, we used PCC as the similarity measure that assesses linear correlation. Because gene expression correlation can be non-linear among other considerations, we explored three additional similarity measures: Spearman's rank, Mutual information (MI), and partial correlation (pcor). The percentile values of within-pathway similarities based on these three new measures in the distribution of between-pathway similarities were calculated like what we have done for PCC . For each of the four measures, we also determined the mutual ranks (MRs, see **Methods**), because the inference of gene function association using similarity measures, like PCC values, can be affected by the method of gene expression database construction and the target gene function (Obayashi and Kinoshita, 2009). For MR, 1 minus expression similarity percentile was calculated because gene pairs with higher percentiles will have smaller numerical rank values. To illustrate the impact, we used all expression data and FPKM and found that some pathways could only have high similarity values when specific measures were used (**Fig. 1E**). For example,19 pathways could only be identified in the first quintile of expression similarity (red) using MI (left panel, **Fig. 1E**). Using MR of MI, an additional four pathways were now in the highest quintile (right panel, **Fig. 1E**). Genes in pathways with high quintiles in all four measures generally tended to have higher similarities in both expression levels and breadths (**Fig. S2**). On the other hand, genes in pathways with higher PCC/Spearman rank but lower MI/pcor quintiles had similar expression patterns in levels but not in breadth (**Fig. S2**). In contrast, genes in pathways with higher MI quintiles compared to other measures are more correlated in expression breadth than expression levels (**Fig. S2**). Taken together, our findings highlight the importance of exploring expression datasets, expression values, and similarity measures in evaluating gene expression similarity for different pathways. Thus, in subsequent analyses, we determined expression correlations using 41 datasets, 2 expression values, and 8 similarity measures. Thus, there were 656 possible data combinations.

Naive prediction of pathway genes

With the impacts of expression datasets, values, and similarity measures demonstrated, we next asked what approaches should be used to predict whether a gene belongs to a metabolic pathway based on the expression similarity. We explored three general approaches: (1) naive prediction, (2) unsupervised learning, and (3) supervised learning. The later two approaches are detailed in subsequent sections. For naive prediction, it was named solely because it was the simplest. We explored two

ways for assigning a gene to a pathway: 1) native median: If gene X has highest median expression similarity with genes in pathway A, then gene X is predicted as in pathway A; and 2) native maximum: If gene X has maximum co-expression with gene Y and gene Y is in pathway B, then gene X is predicted as in pathway B (**Fig. 2A**). Genes annotated in multiple pathways were excluded in the analysis due to the ambiguity in pathway membership assignment using the naive prediction approaches. This resulted in 972 enzyme genes in 85 pathways, each with ≥ 5 genes. To make the naive prediction comparable with prediction using the other two machine learning approaches, five genes in pathways with ≥ 25 genes were randomly held out but not used (holdout subset). The naive prediction methods were applied to the rest of genes, referred to as the prediction subset.

The goal for pathway prediction ultimately is for predicting pathway membership of unknown genes. Thus, we splitted our dataset into "known" (80% for each pathway) and "unknown" (20%) subsets where the maximum and median expression similarities of unknowns were calculated against known genes (**Fig. 2A** and **Methods**). Because we did have annotation information for "unknowns", we could evaluate how well these naive approaches performed when predicting novel genes. This data split was conducted five times similar to the five-fold cross-validation scheme (Fig. 2A), thus, each gene in a pathway would be treated as "unknown" once. The average F1-score (the harmonic mean of precision, proportion of predictions that are true positives, and recall, proportion of true cases that are correctly predicted) on "unknown" genes (referred to as F1unknown) of five data splits was calculated to evaluate the prediction performance (**Fig. 2A**). The F1 ranges from 0 and 1, the later indicates a perfect model. The end results were two $F1_{unk}$ score matrices for the naive medium and maximum criteria that were each 656 (number of data combinations) by 85 (number of pathways) in dimension (**Table S11 and S12**).

We first asked whether one data combination was particularly useful for predicting pathway membership (overall best, red box, **Fig. 2A**). The best data combination had an average $F1_{unknown}$ = 0.04 (all samples, FPKM as the expression value, mutual rank of PCCs as the similarity measure, using the naive maximum method). Although it is better than the median F1 of random guess at 0.01 (dotted line, **Fig. 2B**), the prediction based on the best data combination is far from perfect. Because different data combinations significantly impact whether pathway memberships could be recovered with expression data (**Fig. 1**), we next assessed the extent to which the pathways differ in the data combinations that provided the best prediction. By identifying the data combination that led to the highest $F1_{unknown}$ scores for each pathway (pathway best $F1_{unknown}$, purple boxes, **Fig. 2A**), we found that genes in a pathway can be better predicted when the optimal data combination was used, with the median pathway best $F1_{unknown}$ using the naive median and the naive maximum methods to be 0.08 and 0.11, respectively (**Fig. 2B**). In nearly all cases, using the overall best data combination led to

prediction that had pathway F1 substantially lower than the pathway best F1 in both the naive median (**Fig. 2C**) and the naive maximum (**Fig. 2D**) methods. One notable pattern is that the overall best data combination frequently led to pathways with $F1_{unknown}$=0 (**Fig. 2C,D**), further demonstrate the need to consider optimal combinations of datasets, expression values, and similarity measures.

We also found that the naive maximum method generally performed better than the naive median (**Fig. 2E**), consistent with our finding that the maximum expression similarity of a gene to other genes in pathways provide more information for pathway gene assignment. This is likely because genes in the same pathway can be regulated in many other levels beyond the transcriptional level. Besides, different components of a pathway may branch off to other pathways also, further contributing to regulatory differences within each pathway (Ihmels et al., 2004). Nonetheless, even for the naive maximum method, the median pathway best $F1_{unknown}$ was only 0.11 among pathways. Thus, despite the usefulness of the naive prediction approach for assigning enzyme genes to pathways, there remains substantial prediction errors. In the following sections, we further explored machine learning approaches to see if they could provide better predictions.

Prediction of pathway genes using unsupervised learning methods

Unsupervised learning in the form of clustering is one of the most widely used methods to aggregate genes with similar functions. While unsupervised learning methods have been used to predict metabolic pathway memberships (Wisecaver et al. 2017), earlier studies used one or few datasets without exploring the impacts of datasets and expression values. To explore the impact of clustering algorithms and parameters, we used *k*-means and Affinity Propagation algorithms (see **Methods**). For each algorithm and each of the 82 data combinations (41 datasets and two expression values), two types of input matrices were generated for clustering. The first type was simply the expression value matrix (referred to as Set A, **Fig. 3A**). The second type was generated by first determining the expression similarities of each gene to other non-self genes and then calculating the median and maximum similarities of pathways to the gene in question (Set B, **Fig. 3B**). For all clustering experiments, the same two subsets of genes in naive prediction ("known" and "unknown") were used as clustering data and validation data, respectively.

To predict pathway membership with clustering results, we considered two issues. First, genes from multiple pathways could be grouped in the same cluster. Second, genes from a single pathway could be classified into several clusters. These issues were addressed by, for each cluster C, determining enrichment scores for genes in every pathway and associating C with a pathway P if the enrichment score between C and P was significant and the highest among pathways (see **Methods**). To alleviate the issue of overfitting, a 5-fold cross-validation scheme was applied (**Fig. 3A,B**, see

Methods). This approach also allowed us to compare the clustering-based predictions with those based on supervised learning detailed in the next section. To make predictions, for each gene in the validation set, it was assigned to a cluster C if the distance between the gene in question and cluster C centroid was the shortest.

First, the $F1_{CV}$ (the average F1 value of prediction on validation subsets across the five fold cross-validation) in unsupervised learning was compared to the $F1_{unknown}$ in naive approaches, because the same splitted data were used in unsupervised learning (clustering/validation) and naive approaches ("known"/"unknown") (**Fig. 3A** and **Methods**). As expected, the clustering based unsupervised learning outperformed naive approaches by a large margin (**Fig. 3C**), no matter which algorithm (*k*-means or Affinity propagation) or dataset (setA - gene expression value matrix, or setB - gene-to-pathway expression similarity matrix) was used. Nevertheless, like naive approaches, none of the single clustering had good prediction for most of the pathways, with the maximum average F1 at 0.07, where the FPKM values of genes in the dataset with different genetic backgrounds were used in *k*-means clustering with *k*=500 (**Fig. S3**). We also found that *k*-means outperformed Affinity propagation for most pathways (**Fig. 3C,D**), either using setA (67 pathways) or setB (77 pathways). In addition, k-means based clustering using setB outperformed that using setA for 57 of 85 (67%) pathways (**Fig. 3E**). This could be attributed to two potential reasons. First, there are too few features in setA datasets, as the median number of features was 8 among 82 setA datasets (expression values and contrasts), while all the setB datasets had 170 features (median and maximum gene-pathway expression similarities among 85 pathways). Consistent with this hypothesis, there is significantly positive correlation (PCC = 0.29, *p*-value = 0.01) between the pathway best $F1_{CV}$ and the number of features in the corresponding pathway best clustering when setA was used (Fig. S4A). Second, gene-to-pathway expression similarity likely provided more information for pathway membership prediction than the gene expression profiles for most pathways, which may be because that the gene-pathway expression similarity leveages structures of the metabolic pathway/network which was captured by the unsupervised methods we used.

Prediction of pathway genes using supervised learning methods

Different from the unsupervised clustering where the pathway information of genes were not used for the clustering, supervised machine learning methods build the models by learning from labeled instances, i.e., genes with pathway annotation in our study. The Random Forest (RF) algorithm (ref) was used to build multi-class classification models, where a multitude of decision trees are constructed learning from training data. The same feature datasets (setA and setB), 5-fold cross validation procedure and splitted data used in the unsupervised learning methods were used for RF models (**Fig. 4A,B**).

Interestingly, the supervised learning models had the performance inferior to the clustering-based unsupervised methods (*k*-means), where the median pathway best $F1_{CV}$ values in RF models was 0.23 and 0.3, when setA and setB data were used, respectively, compared to 0.33 and 0.37 in *k*-means clustering (**Fig. 4C**). One potential reason is that there are too few genes in each pathway after data filtering (**Methods**) in our analysis (median pathway size=8), which may be insufficient for supervised machine learning to learn from. Consistent with this, we found that the differences of pathway best $F1_{CV}$ between RF models and *k*-means clustering were weakly, but significantly correlated with the pathway sizes, with PCC=0.29 (*p*-value=0.007) and 0.23 (*p*-value=0.03) when setA and setB were used, respectively (**Fig. S4C,D**). These results suggest that for a multiclass task, when there are too few instances in each class, the clustering-based unsupervised learning approaches may be superior to supervised learning approaches, due to the insufficiency of training data for a supervised learning approach to learn from. In addition, by examining the confusion matrices (or error matrices, showing the mis-prediction in all pathways) in pathway best *k*-means clustering and RF models (**Fig. 5A,B**), we found that genes in most pathways were more likely mis-predicted as in pathways LIPAS-PWY (triacylglycerol degradation) or PWY-1081 (homogalacturonan degradation) in the pathway best RF models. This can be because that LIPAS-PWY and PWY-1081 are the largest pathways (> 50 genes) among others, and have more genes in training subset than other pathways, thus there is a higher chance of predicted genes as in these two pathways. This result indicates that RF models tend to be more vulnerable to unbalanced data.

Consistent with naive prediction and *k*-means clustering, the optimal data combination of expression datasets, expression values and similarity measures needs to be explored for each pathway (**Fig. S3**). To understand whether RF models and *k*-means clustering had similar bias on data combinations, we asked whether pathways have the best prediction in *k*-means clustering and RF models when the same data was used. To simplify the analysis, here only *k*-means clustering and RF models using setB were examined, due to the observation that RF models using setB datasets also outperformed models using setA (**Fig. 4C,D**), like what has been found in *k*-means clustering. We found that only 10 pathways have both the best *k*-means clustering and best RF model when the same setB data was used, e.g., PWY-401 (**Fig. 5D, Fig. S6**) and PWY-6442 (**Fig. 5E, Fig. S7**), while 75 do not, e.g., PWY-6363 (D-myo-inositol (1,4,5)-trisphosphate degradation) (**Fig. 5F, Fig. S8**). This result suggests the different bias on data combination of these two approaches. However, *k*-means clustering and RF models may have similar performance when the same data were used, even though the pathway best prediction differs between these two methods. Consistent with this hypothesis, we found that the performance in *k*-means clustering and RF models were positively correlated, with the PCC=0.4 and 0.34 when the setB data in pathway best *k*-means clustering and in pathway best RF models were used, respectively (**Fig. S9**).

Together with the observation that the relatively good membership prediction (red and orange color in **Fig. 5B,C**) of most pathways was sparsely achieved by a small number of data combinations, these results suggest the common grounds in how *k*-means clustering and RF models have learned from the gene-pathway expression similarity.

With regard to the performance differences of RF models when setA and setB data were used, we found that, like what had been found in *k*-means, the number of features in setA is significantly correlated to the pathway best $F1_{CV}$ (PCC=0.33, *p*-value=0.002, **Fig. S4B**), suggesting that too fewer features in most setA datasets may be the underlying reason for the relative poor performance of RF models built using setA. Next, as hypothesized in *k*-means clustering that the network among pathways may have been considered when setB was used, we examined the most important features in RF models built using setA and the corresponding setB. Two examples were examined here. First, PWY-6442 has best prediction with F1cv=0.6 in a RF model, where a setB data, MR_MI_FC_development (gene-pathway expression similarities were calculated by mutual rank of MI using FC values of genes in samples belonging to tissues/stages/circadian category), was used. In contrast, when the corresponding setA (matrix of FC values of genes among samples) was used, $F1_{CV}$=0 (**Fig. S7**). When the top 20 important features were examined, three genes annotated as in PWY-6442 had very similar profiles in setB, but none of a single pair of genes in PWY-6442 share similar expression profiles in setA. The other example is XX, where XXX

Correlation between pathways and the optimal data combination used

In the previous section, we have shown that *k*-means clustering and RF models share common grounds in the choice of data combinations in pathway membership prediction. Next, we examined the influences of data combinations on pathway membership prediction. To achieve this, we first summarized the pathway best prediction in 4 methods, where setA and setB were used in *k*-means clustering and RF models, and got the pathway optimal prediction among these 4 methods, which ranges from 0.21 to 0.8, with the median pathway optimal $F1_{CV}$=0.4 (**Fig. 5A**). Then we asked among the pathway optimal clustering/models, whether more pathway optimal prediction tends to be achieved when particular data combinations were used.

Regarding the datasets used, we found the “condition-independent” dataset (including all samples) provided optimal prediction for 12 pathways, while the remaining 73 pathways (86%) need pathway specific “condition-dependent” datasets, where ≤5 pathways had optimal prediction in each dataset (**Fig. S5**). In addition, pathway membership tended to be correctly predicted when datasets designed for related biological processes were used. For example, the abscisic acid biosynthesis (ABA, PWY-695), and the superpathway of carotenoid biosynthesis in plants (CAROTENOID-PWY, products from which, carotenoids, serve as precursors for ABA) (Milborrow, 2001), had the optimal F1 when the hormone treated dataset was used (**Table S11,**

S14, S15). The galactolipid biosynthesis I (PWY-401) had the highest F1s when samples with ABA treatment were used (**Table S16**). In *Arabidopsis*, genes involved in PWY-401, MGD1/2/3 and DGD1/2, were found to be regulated in response to phosphate starving, while ABA was shown to play a dominant role in regulating phosphate-responsive candidates (Woo et al., 2012). Interestingly, the substrate of pathway PWY-6363, D-myo-inositol (1,4,5)-trisphosphate, mediates the biological response of a large number of hormones, while the dataset used for PWY-6363 best *k*-means clustering is 2014_ARF (Zouine et al., 2014), where flowers/fruits at three different stages were taken (**Table S4**), which have differential hormone levels (de Jong et al., 2009), and dataset used for PWY-6363 best RF model is 2016_sulfurea, where the sulfurea mutant has less auxin than wildtype (Gouil et al., 2016) (**Table S17**). In contrast, the thioredoxin pathway (THIOREDOX-PWY), which is important for maintaining intracellular redox status and essential for most processes, had the optimal F1 when all the samples were used (**Table S18**). These results not only recognize the importance of reasonable data sampling, but also shed lights on the connections between pathways and biological processes.

For the expression values, we found that 53 and 36 pathways (the sum is 89, because some pathways have optimal prediction when both FPKM and FC values were used) had optimal prediction when FPKM and FC was used, respectively, but with no significant differences in pathway best $F1_{CV}$ (*p*-value of Wilcoxon rank sum test is 0.22, **Fig. S5**). Furthermore, when setB was used, for different similarity measures, features calculated using MR of correlation coefficient achieved optimal prediction for more pathways (45) than the correlation coefficient (32) (**Fig. S5**).

Conclusion

We comprehensively evaluated the utility of gene co-expression in pathway membership prediction, by considering 656 data combinations (expression datasets, expression values and similarity measures), and three prediction algorithms (naive prediction, unsupervised and supervised learning). We demonstrated that optimal data combination should be pursued for each pathway, since none of the 656 data combinations had reasonable prediction for most pathways, and the overall best clustering/model is far from the pathway best one. Optimal pathway membership prediction tended to be achieved when the pathway function associated datasets were used. This indicates that the membership prediction for some pathways can be improved when more relevant datasets are used.

Genes with multiple pathway annotation were removed in our study to prevent the confusion of assignment. Due to the important roles in multiple domains and the immeasurable diversity of metabolites in plants, the identification and understanding of genes involved in plant metabolic pathways has always been a persistent and elusive question. Gene co-expression analysis has been widely used to identify genes with

associated functions since the genome-wide gene expression detection approaches had been developed, e.g., microarrays and RNA-seq techniques. However, the genome-wide utility of gene co-expression analysis in metabolic pathway membership identification remains largely unsolved. Here, by considering 656 data combinations (expression datasets, expression values and similarity measures), and these prediction algorithms (naive prediction, unsupervised and supervised learning), we comprehensively evaluated the utility of gene co-expression in pathway membership prediction.

Transfer learning from model species to other species. About half pathways had optimal prediction in *k*-means when setB was used, suggesting that *k*-means clustering using gene-pathway expression similarity can be potentially the first method worth trying for most pathways. *k*-means clustering using gene-pathway expression similarity calculated using FPKM by mutual rank of similarity measures can be potentially the first method worth trying for most pathways. However, the prediction of candidate enzyme genes in the first place was mainly based on the sequence similarity of unknown genes to experimentally evaluated genes in model species, whereas metabolite composition varies across species, which can be attributed to the emergence of new metabolic pathways after gene duplication and subsequent sub- or neo-functionalization (Pichersky and Gang, 2000), recruitment of genes in new pathways (Pichersky and Gang, 2000)(Shoji and Hashimoto, 2011), and vanish of existing pathways in new environments where the metabolites are not needed anymore (ref). Therefore, the pathway membership predicted in target species are only limited to the known pathways in experimentally tested species, other than pathways or genes specific to the target species.

Methods & Materials

Genes functional annotation

Metabolic pathway annotation of genes in tomato (Solanum lycopersicum) were downloaded from TomatoCyc V3.0 (Plant Metabolic Network, PMN 12.0, https://www.plantcyc.org/) (Schläpfer et al., 2017), where the gene annotation was from Solanum lycopersicum V3.2 (Solanaceae Genomics Network, SGN, https://solgenomics.net/, referred as SGN_V3.2). To make the gene IDs consistent with these in our previous paper (Wang et al., 2018), which was from the S. lycopersicum genome in National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/genome, V2.5, referred as NCBI_V2.5), coding sequences (CDS) in SGN_V3.2 were aligned to the genome sequence in NCBI_V2.5 using the BLAST-like alignment tool (BLAT) (Kent, 2002), with the requirement of identity at 100% and no gap tolerated if located within an aligned block. A NCBI_V2.5 gene ID was associated with the SGN_V3.2 ID if the sequences match (gene ID association was

listed in Supplementary Table S1). SGN_V3.2 genes with no matching NCBI_V2.5 genes were found to be annotated as pseudogenes (e.g., Solyc07g043650) or totally not annotated due to truncated length (e.g., XXX), and these genes normally have no expression in all/most samples. Discrepancy between SGN_V3.2 and NCBI_V2.5 gene annotations can also be due to the mis-annotation of genes in SGN_V3.2 or different open reading frames (ORFs) between two versions. For example, two genes annotated in NCBI_V2.5, XP_010324501.1 (Terpene synthase 41, TPS41) and XP_010324500.1 (TPS18), were annotated as a single gene in SGN_V3.2, Solyc08g005720.3.1 (TPS18), but with different expression profiles (Supplementary Table S1, Supplementary Fig. 2B).

The functional annotation of genes was conducted using PMN Ensemble Enzyme Prediction Pipeline V3.0 (https://gitlab.com/rhee-lab/E2P2/) with default settings. In total, there are 11036 genes with annotated Enzyme Commission (EC) numbers and/or metabolic reactions, out of which 2395 genes were annotated in 485 metabolic pathways in TomatoCyc V3.0 (Supplementary Table S2). For the statistical sense, only pathways with ≥ 5 genes which were expressed in ≥ 1 samples with non-zero RNA-seq reads were included, resulting in 2171 genes in 297 pathways.

RNA-seq data and processing

S.lycopersicum RNA-seq data from 47 studies (**Supplementary Table S3 and S4**), including 372 tissue samples (sample with replicates was considered as one single sample), were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/sra/) and DDBJ Center (https://www.ddbj.nig.ac.jp/). There were only 36 of 47 individual datasets with ≥ 3 samples, and each was used to calculate gene expression similarity. All the 47 individual datasets were further classified into four categories: 1) tissues/stages/circadian - samples of different tissues, or at different development stages, or at different time points of days in wild type tomato plants; 2) genetic background - samples from wild type, gene mutant, or transgenic plants where a gene's expression was knocked down or over expressed; 3) hormone treatment - samples from plants treated with hormones (cytokinin, ABA, auxin, etc) and control material; 4) stress treatment - samples from plants with stress treatment (pathogen, light, cold, heat, etc) or control condition (**Fig. 1** and **Supplementary Table S3**). Samples belonging to each of these four categories constitute a single datasets, and all the 372 samples constitute another inclusive dataset. Thus, there were 41 datasets in total: 36 individual datasets, and five integrated datasets.

Mapping the RNA-seq reads to the genome sequences, and calculation of fragments per kilobase of transcript per million mapped reads (FPKM) and fold change (FC), see our previous paper (Wang et al., 2018). Median FPKM value among replicates was used as the expression level for a gene in a sample. Genes with no expression (FPKM=0) in all samples were excluded from our analysis, and pathways

with <5 annotated genes were also excluded, resulting in 2171 genes annotated in 297 pathways.

Gene expression similarity measure

Eight different similarity measures were evaluated in our study. Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (Spearman) between expression values of two genes were calculated using "pearsonr" and "spearmanr" functions in scipy.stats modules (Virtanen et al., 2019) in python environment, respectively. Mutual information (MI) were measured using the "normalized_mutual_info_score" function in python module sklearn.metrics.cluster (Pedregosa et al., 2011). Partial correlation (partCor) were evaluated by "pcor.shrink" function (Schäfer and Strimmer, 2005) in "corpcor" R packages. In addition, mutual rank (MR) were calculated for each of these four methods, with the format

$$MR = \sqrt{rank(gene1 \rightarrow gene2) \times rank(gene2 \rightarrow gene1)}$$

Where the ranks of other non-self genes to a gene ranges from 1 to 2170, with 1 indicating highest expression similarity.

Data split and the measure of prediction performance

To avoid overfitting of parameter setting and make the prediction comparable among three approaches, five genes from each of six pathways, where there are ≥ 25 genes expressed in ≥ 1 samples, were held out and taken as test data in unsupervised and supervised approaches and would never be used in the cluster/model building. The remaining genes within each pathway were further split to 5 folds, of which 4 folds were used as "known", clustering and training genes in naive, unsupervised and supervised approaches, respectively. The remaining 1 fold was taken as "unknown" or validation genes to evaluate the prediction performance of naive and machine learning approaches, respectively. This split was conducted five times to ensure that every gene was taken in a "unknown"/validation subset once.

F-measure (F1) were calculated by comparing the annotated and predicted pathway membership for "unknown"/validation genes in a pathway (referred as F1unknown or F1CV), using the equation: $2*\frac{precision*recall}{precision+recall}$, where precision = $\frac{TP}{TP+FP}$, recall = $\frac{TP}{TP+FN}$, TP: true positive, i.e., number of genes annotated and predicted as in the pathway, FP: false positive, i.e., number of genes mis-predicted as in the pathway, FN: false negative, i.e., number of genes annotated but not predicted as in the pathway. Average F1unknown or F1CV of each pathway across the five data splits mentioned above were used to evaluate the performance of the prediction.

Unsupervised machine learning

The unsupervised clustering was conducted using "KMeans" and "AffinityPropagation" in sklearn.cluster. For *k*-means, the parameter n_clusters (how many clusters to form, k) at 5, 10, 25, 50, 85, 100, 200, 300, 400 and 500 were tested, and for Affinity Propagation, damping factor (the extent to which the current value is maintained relative to incoming values) at 0.5, 0.6, 0.7, 0.8, 0.9 and 0.99 were used. To evaluate which pathway a cluster is enriched for, the enrichment analysis were conducted as: $log_{10}(\frac{a/(a+b)}{(a+c)/(a+b+c+d)})$, where *a* - the number of genes from pathway A grouped in cluster X; *b* - the number of genes from pathway A, but not grouped in cluster X; *c* - the number of genes not from pathway A, but grouped in cluster X; *d* - the number of genes not from pathway A, and also not grouped in cluster X. Cluster X was assigned to pathway A if pathway A has the highest enrichment values to Cluster X among all the other pathways, and the *p*-value of corresponding Fisher's exact test, where *a*, *b*, *c*, and *d* were used in the 2×2 contingency table, <0.05. The closest cluster was predicted for a given gene using the "predict" function in sklearn.cluster, and the gene was predicted as in the corresponding pathway. $\frac{\blacksquare}{precision+recall}\frac{\blacksquare}{TP+FP}\frac{\blacksquare}{TP+FN}$

Supervised machine learning

Multi-class supervised models were established using training genes, with "RandomForestClassifier" in sklearn.ensemble (Pedregosa et al., 2011). Hyperparameters (*max_depth*, *max_features*, and *n_estimators*) were determined by doing the Grid Search in the five cross validation procedures to maximize the F1CV, and were used to build the final model, which was applied to validation and test genes.

Figure legends

Fig. 1 Impact of expression dataset, expression values, and similarity measures on expression similarities between genes in the same pathways. (**A**) Gene expression similarity (in terms of PCC) within and between pathways. Boxplot showing expression similarity between genes within individual pathways. Blue line: median value; light blue box: interquartile range. Grey violin plot shows the PCC distribution of all gene pairs between different pathways (between-pathway distribution), where median value and interquartile range were marked using red and orange dashed lines, respectively. (**B**) Examples showing impact of dataset on gene expression similarity. FPKM was used to calculate the PCCs. X-axis: 41 datasets, y-axis: example pathways. (**C**) Same as (**B**), except that FC was used in PCC calculation. Color scale: percentile of median PCC within a pathway in between-pathway distribution. (**D**) Dotplot shows differences between percentiles of median PCC within pathways in between-pathway distribution, calculated using FPKM (x-axis) and FC (y-axis). (**E**) Gene expression

similarity calculated using different similarity measures. Color scale: percentile or 1-percentile of median expression similarity calculated using different similarity measures or mutual rank of similarities in corresponding between-pathway distribution, respectively. White and black dashed rectangles show pathways with percentiles at range of 0.8~1 and 0.6~0.8 which are specific to a similarity measure, respectively.

Fig. 2 Naive prediction of metabolic pathway genes. (**A**) Methodology. For each of the 82 (41 datasets × 2 expression values) expression matrices, five genes were randomly held out from a pathway which contains ≥ 25 genes. The remaining data were used to determine expression similarities among genes with 8 similarity measures, resulting in 656 expression similarity matrices (m1~656). Genes within each pathway were further splitted to "known" (80%) and "unknown" (20%) subsets, and the data splits were conducted five times. The example "unknown" gene 1 is predicted to be in pathway A in naive median method because it has the highest median expression similarity with "known" genes in pathway A among others, while it is predicted to be in pathway B in naive maximum method since it has the maximum expression similarity with gene 3, which belongs to pathway B. Thickness of arrow and number besides arrow: degree of expression similarity. All the 656 expression similarity matrices were used in both two methods, and F1 score was measured for each of the 85 pathways, resulting in two 656 x 85 F1 score matrices. The prediction with the highest F1 for a pathway was referred to as the pathway best prediction. The average F1 across 85 pathways of each prediction was calculated to measure the overall performance of the prediction, and the one with highest average F1 was referred to as overall best prediction. (**B**) Distribution of pathway best F1 scores in naive median (orange) and naive maximum (pink) methods. Distribution of background pathway F1 scores was shown in grey violin plot, with the median value indicated by dashed line. (**C,D**) Performance of naive median and naive maximum prediction, respectively. X-axis: F1 of each pathway in overall best prediction, Y-axis: pathway best F1 across all 656 predictions. (**E**) Comparison of pathway best F1 in naive median (x-axis) and naive maximum (y-axis) predictions. Three example pathways in (**C-E**) show the differences in pathway predictions in two methods, and were illustrated in (A) as Pathway A, B, and C.

Fig. 3 Prediction of pathway genes using unsupervised clustering methods. (**A**) Workflow for clustering using gene expression profiles. The data split was conducted in the same way as in naive approaches, and the "known" and "unknown" subsets in naive approaches were taken as "clustering" and "validation" subsets, respectively ("Split data" and "5-fold cross validation"). Clusters built using expression values of clustering genes ("*Cluster without validation data*") were assigned to pathways based on an enrichment analysis ("*Assign clusters to pathways*", see **Methods**). Then the clusters were applied to validation genes, which in turn were predicted as in the corresponding

pathways ("*Assign validation genes to clusters*"). The average F1 of validation genes across the 5 clustering/validation splits ($F1_{CV}$) were calculated to evaluate the predictions. The clusters were also applied to test genes, and the average F1 of test genes from 5 clustering ($F1_{test}$) were calculated. (**B**) Similar to (**A**), except that clusters were built using gene-to-pathway co-expression matrices (setB), i.e., the maximum and median expression similarity of a clustering gene to all other non-self clustering genes in each pathway, which was calculated using 8 different similarity measures, resulting in 656 (82×8) data and clustering. For a validation or test gene, the expression similarity was calculated between the gene to the clustering genes in each pathway, and the maximum and median values were used in the matrix. (**C**) Distribution of pathway best $F1_{CV}$ in the clustering, conducted using *k*-means (KM, light green) or Affinity Propagation (AP, green) methods, expression matrix (setA) or gene-pathway co-expression matrix (setB). Orange, pink and grey dashed lines indicate the median pathway best F1CV in Naive median and Naive maximum prediction, and the median background pathway F1, respectively. (**D-F**) Comparison of pathway best $F1_{CV}$ in *k*-means and Affinity Propagation clustering using setA (**D**) or setB (**E**), or in *k*-means clustering when setA and setB were used (**F**). Dots: individual pathways.

Fig. 4 Prediction of pathway genes using supervised machine learning methods. (**A**) Workflow for RandomForest machine learning using setA as features. Labels of the pathway a gene belongs to were merged into the expression matrix. The same genes were held out as test genes as in clustering methods, and never used in the model building. The same 5-fold cross-validation procedure was used, and 80% of genes were used to train the model (training genes), and the 20% of genes were used to validate the model (validation genes) ("Split data" and "5-fold cross validationData split"). Grid search procedure was conducted to get the best combination of parameters (hyperparameters) when the model had the maximum $F1_{CV}$ (see Methods). Final model was built using the same cross-validation procedure with the hyperparameters, and was applied to validation and test genes ("Apply"). (**B**) Same as (**A**), except that gene-pathway co-expression matrices (setB) were used as features. (**C**) Distribution of pathway best F1PR and $F1_{CV}$ in Random Forest (RF) models using setA (light blue) or setB (blue). Light green and green dashed line: median pathway best F1PR or $F1_{CV}$ in *k*-means using setA and setB, respectively; orange and pink dashed line: median pathway best F1PR in Naive median and Naive maximum prediction, respectively; grey dashed line: median background pathway F1. (**D**) Comparison of pathway best $F1_{CV}$ in RF models when setA and setB were used. (**E**) Comparison of pathway best $F1_{CV}$ in RF models and *k*-means clustering when setB was used. Dots: individual pathways.

Fig. 5 (**A**) Confusion matrix for prediction of pathway in question in pathway best *k*-means clustering. (**B**) Confusion matrix for prediction of pathway in question in pathway

best RF models. Color scale: proportion of genes in a pathway (y-axis) was predicted as in one of the 85 pathways (x-axis) in pathway best clustering.

Fig. 5 Summary of pathway membership prediction using unsupervised clustering and supervised machine learning models. (**A**) Pathway best $F1_{CV}$ in *k*-means clustering using setA (orange) and setB (purple), and RF models using setA (green) and setB (blue) using setB, and the background pathway F1 (grey). Six pathways with test genes were marked using colorful fonds. (**B**) Heatmap showing normalized pathway $F1_{CV}$ in 6560 (656 setB matrix × 10 *k*s) *k*-means clustering. (**C**) Heatmap showing normalized pathway $F1_{CV}$ in 656 RF models. Pathway optimal $F1_{CV}$ was set as 1, while the minimal $F1_{CV}$ was 0. (**D-F**) Three example pathways, indicated using arrows in (**A**) and purple rectangles in (**B,C**), where the pathway best $F1_{CV}$ in *k*-means was larger (PWY-401, **D**), smaller (PWY-6442, **E**) than that in RF, or equaled to that in RF (PWY-6363, **F**), when setB was used. Dots: 656 data used for *k*-means clustering or RF models.

Supplementary figure legends

Fig. S1 Differences of gene expression similarity within pathways and between pathways. (**A**) Maximum PCC of a gene to all other non-self genes within the same pathway, where the expression levels (FPKM) of genes in all 372 samples were used to calculate the PCC. Blue line: median value; light blue box: interquartile range. Violin plot shows distribution of maximum PCC of a gene to genes from a different pathway. Median value and interquartile range in between pathway distribution were marked using red and orange dashed lines, respectively. (**B**) Percentile of a median value in a pathway in corresponding between-pathway distribution, where the value is median (x-axis) or maximum (y-axis) PCC of a gene to other non-self genes within the same pathway. (**C**) Distribution of PCC values between gene pairs within pathway and between pathways. X-axis: PCC values at different percentiles within pathways. All *p*-values of Wilcoxon rank sum test < 5.3e-10. es: effect size, which is the median difference. (**D**) Distribution of PCC values between homologous (blue) and non-homologous (light blue) gene pairs within the same pathway. X-axis: different percentiles within pathways. Pink dashed line: median PCC values of gene pairs between different pathways in (**C**) in corresponding percentile.

Fig. S2 Impact of similarity measure on expression similarity. (**A**) Three examples showing pathways with different percentiles of expression similarity in between-pathway distribution calculated with different similarity measures. Color scale: five range of expression similarity percentile in between-pathway distribution. (**B**) Expression profiles

(normalized FPKM) in all 372 samples for three pathways shown in (**A**). The expression levels were scaled to 0-1, minimum level: 0 (blue), maximum: 1 (red).

Fig. S3 Pathway membership prediction in unsupervised clustering and supervised models. (**A**) Distribution of pathway best F1_clustering or F1_training. The corresponding pathway best F1 in naive prediction and background were indicated using orange (Naive_median), red (Naive_maximum), and grey (background) dashed lines. (**B**) F1_clustering/F1_training and F1_cv in clustering/models which had the highest F1_clustering/F1_training for each pathway. Dots: individual pathways. (**C**) Distribution of average F1_cv of clustering and models. (**D**) F1_clustering/F1_training, F1_cv, and F1_test of six pathways, which had ≥ 25 genes and five genes of which were held out as test data, in clustering/models with pathway best F1_cv. (**E**) Pathway F1_test of six pathways in *k*-means clustering and RF models when setB was used. Dots: clustering/models when 656 setB data were used. Horizontal and vertical dashed line: background pathway F1_test.

Fig. S4. Comparison between pathway F1_cv in overall best clustering/model which has highest average F1_cv, and pathway best F1_cv. (**A**) *k*-means and (**B**) Affinity Propagation clustering when setA was used, *k*-means (**C**) and Affinity Propagation (**D**) clustering when setB was used, RF models when setA (**E**) and setB (**F**) was used. Dots: individual pathways.

Fig. S5. Underlying factors in pathway optimal clustering/models. (**A-D**) Distribution of pathway optimal F1_cv (y-axis, highest pathway F1_cv in all four methods in **Fig. 5A**) in *k*-means clustering/RF models when different expression values (**A**), datasets (**B**), feature matrices (**C**), similarity measures (**D**), machine learning methods (**E**), and number of clusters in *k*-means clustering (**F**) were used (x-axis). Blue dots: individual pathways.

Fig. S6. Expression profiles of genes annotated or predicted as in PWY-401. The optimal membership prediction (F1_cv = 0.61) for PWY-401 was achieved using *k*-means clustering when setB data MR_spearman_FPKM_2013_ABA was used, where FPKM of genes in samples with ABA and control treatment were used to calculate the mutual rank of Spearman's rank correlation coefficient between a gene to other non-self genes in each pathway. The pathway best F1_cv in RF models was also achieved using the same setB data, but with F1_cv at 0.53 (**Fig. 5F**). Only values of genes in top 20 important features (based on the importance in the RF model when the same setB data was used) were shown in the right heatmap, and values were normalized across genes (y-axis), i.e., for each feature, gene with the highest value will have red color (1), while gene with lowest value will have blue color (0). Expression profiles (FPKM) of genes in

the corresponding setA were also shown in the left heatmap, and were normalized across different samples (x-axis).

Fig. S7. Expression profiles of genes annotated or predicted as in PWY-6442. The optimal F1_cv (0.6) for PWY-6442 was achieved using RF machine learning, using setB data MR_MI_FC_development, where FC of genes in samples belonging to the category "Tissues/stages/circadian" were used to calculate the mutual rank of MI. The pathway best F1_cv in *k*-means was also achieved using the same setB data, with an F1_cv at 0.3 (**Fig. 5G**). Only values of genes in top 20 important features in setB and the corresponding setA were shown in the heatmaps, and the values were normalized across genes.

Fig. S8. Expression profiles of genes annotated or predicted as in PWY-6363. PWY-6363 has the optimal membership prediction (F1_cv=0.28) in both *k*-means clustering and RF models, but using setB data MR_corpcor_FC_2016_sulfurea and corpcor_FC_2014_ARF, respectively. Top 20 important features for genes in there two setB data were shown in the heatmaps, normalized across genes.

Fig. S9. Comparison of pathway membership prediction between *k*-means clustering and RF models when setB was used. (**A**) Pathway F1_cv in RF models (y-axis) which had the highest pathway F1_cv, and in *k*-means clustering (x-axis) when the same setB data used for the pathway best RF models was used. (**B**) Pathway F1_cv in *k*-means clustering (x-axis) which had the highest pathway F1_cv, and in RF models (y-axis) when the same setB data used for the pathway best *k*-means clustering was used. Dots: individual pathways.

Funding

This work was partly supported by the National Science Foundation (IOS-1546617, DEB-1655386) and U.S. Department of Energy Great Lakes Bioenergy Research Center (BER DE-SC0018409) to S.-H.S.

Computational resources

All the scripts used in this study are available on Github at:
https://github.com/peipeiwang6/Pathway_gene_prediction_in_tomato

References

Abu-Jamous, B. and Kelly, S. (2018). Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. Genome Biology **19**.

Aoki, Y., Okamura, Y., Ohta, H., Kinoshita, K., and Obayashi, T. (2016). ALCOdb: Gene Coexpression Database for Microalgae. Plant Cell Physiol. **57**: e3.

Berg, J.M., Tymoczko, J.L., and Stryer, L. (2002). Biochemistry, Fifth Edition: International Version (W. H. Freeman).

Brasher, M.I., Surmacz, L., Leong, B., Pitcher, J., Swiezewska, E., Pichersky, E., and Akhtar, T.A. (2015). A two-component enzyme complex is required for dolichol biosynthesis in tomato. The Plant Journal **82**: 903–914.

Burchard, P., Bilger, W., and Weissenbock, G. (2000). Contribution of hydroxycinnamates and flavonoids to epidermal shielding of UV-A and UV-B radiation in developing rye primary leaves as assessed by ultraviolet-induced chlorophyll fluorescence measurements. Plant, Cell and Environment **23**: 1373–1380.

Carrari, F. and Fernie, A.R. (2006). Metabolic regulation underlying tomato fruit development. J. Exp. Bot. **57**: 1883–1897.

Chae, L., Kim, T., Nilo-Poyanco, R., and Rhee, S.Y. (2014). Genomic signatures of specialized metabolism in plants. Science **344**: 510–513.

De León, I.P., Sanz, A., Hamberg, M., and Castresana, C. (2002). Involvement of the Arabidopsis alpha-DOX1 fatty acid dioxygenase in protection against oxidative stress and cell death. Plant J. **29**: 61–62.

De Luca, V., Salim, V., Atsumi, S.M., and Yu, F. (2012). Mining the biodiversity of plants: a revolution in the making. Science **336**: 1658–1661.

Fan, P., Miller, A.M., Liu, X., Jones, A.D., and Last, R.L. (2017). Evolution of a flipped pathway creates metabolic innovation in tomato trichomes through BAHD enzyme promiscuity. Nat. Commun. **8**: 2080.

Gouil, Q., Novák, O., and Baulcombe, D.C. (2016). SLTAB2 is the paramutated SULFUREA locus in tomato. J. Exp. Bot. **67**: 2655–2664.

Gupta, C. and Pereira, A. (2019). Recent advances in gene function prediction using context-specific coexpression networks in plants. F1000Res. **8**.

Hirai, M.Y. et al. (2007). Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc. Natl. Acad. Sci. U. S. A. **104**: 6478–6483.

Huang, A.C., Jiang, T., Liu, Y.-X., Bai, Y.-C., Reed, J., Qu, B., Goossens, A., Nützmann, H.-W., Bai, Y., and Osbourn, A. (2019). A specialized metabolic

network selectively modulates Arabidopsis root microbiota. Science **364**: eaau6389.

Ihmels, J., Levy, R., and Barkai, N. (2004). Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nat. Biotechnol. **22**: 86–92.

de Jong, M., Mariani, C., and Vriezen, W.H. (2009). The role of auxin and gibberellin in tomato fruit set. J. Exp. Bot. **60**: 1523–1532.

Kaundal, R., Saini, R., and Zhao, P.X. (2010). Combining machine learning and homology-based approaches to accurately predict subcellular localization in Arabidopsis. Plant Physiol. **154**: 36–54.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res. **12**: 656–664.

Kim, J. and Buell, C.R. (2015). A Revolution in Plant Metabolism: Genome-Enabled Pathway Discovery. Plant Physiol. **169**: 1532–1539.

Kimura, S. and Sinha, N. (2008). Tomato (Solanum lycopersicum): A Model Fruit-Bearing Crop. Cold Spring Harbor Protocols **2008**: db.emo105–pdb.emo105.

Lan, H., Carson, R., Provart, N.J., and Bonner, A.J. (2007). Combining classifiers to predict gene function in Arabidopsis thaliana using large-scale gene expression measurements. BMC Bioinformatics **8**: 358.

Leong, B.J., Lybrand, D.B., Lou, Y.-R., Fan, P., Schilmiller, A.L., and Last, R.L. (2019). Evolution of metabolic novelty: A trichome-expressed invertase creates specialized metabolic diversity in wild tomato. Sci Adv **5**: eaaw3754.

Libbrecht, M.W. and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat. Rev. Genet. **16**: 321–332.

Lloyd, J.P., Seddon, A.E., Moghe, G.D., Simenc, M.C., and Shiu, S.-H. (2015). Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes. Plant Cell **27**: 2133–2147.

Martin, C. and Li, J. (2017). Medicine is not health care, food is health care: plant metabolic engineering, diet and human health. New Phytologist **216**: 699–719.

Mathieu, S., Cin, V.D., Fei, Z., Li, H., Bliss, P., Taylor, M.G., Klee, H.J., and Tieman, D.M. (2009). Flavour compounds in tomato fruits: identification of loci and potential pathways affecting volatile composition. J. Exp. Bot. **60**: 325–337.

Matsuba, Y., Zi, J., Jones, A.D., Peters, R.J., and Pichersky, E. (2015). Biosynthesis of the diterpenoid lycosantalonol via nerylneryl diphosphate in Solanum lycopersicum. PLoS One **10**: e0119302.

Meissner, R., Jacobson, Y., Melamed, S., Levyatuv, S., Shalev, G., Ashri, A.,

Elkind, Y., and Levy, A. (1997). A new model system for tomato genetics. The Plant Journal **12**: 1465–1472.

Milborrow, B.V. (2001). The pathway of biosynthesis of abscisic acid in vascular plants: a review of the present state of knowledge of ABA biosynthesis. J. Exp. Bot. **52**: 1145–1164.

Moore, B.M., Wang, P., Fan, P., Leong, B., Schenck, C.A., Lloyd, J.P., Lehti-Shiu, M.D., Last, R.L., Pichersky, E., and Shiu, S.-H. (2019). Robust predictions of specialized metabolism genes through machine learning. Proc. Natl. Acad. Sci. U. S. A. **116**: 2344–2353.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F.M., Wilkins, O., Campbell, M.M., Fernie, A.R., Usadel, B., Nikoloski, Z., and Persson, S. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell **23**: 895–910.

Niggeweg, R., Michael, A.J., and Martin, C. (2004). Engineering plants with increased levels of the antioxidant chlorogenic acid. Nat. Biotechnol. **22**: 746–754.

Ni, Y., Aghamirzaie, D., Elmarakeby, H., Collakova, E., Li, S., Grene, R., and Heath, L.S. (2016). A Machine Learning Approach to Predict Gene Regulatory Networks in Seed Development in Arabidopsis. Front. Plant Sci. **7**: 1936.

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index. Plant and Cell Physiology **59**: e3–e3.

Obayashi, T. and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res. **16**: 249–260.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. **12**: 2825–2830.

Peng, C., Uygun, S., Shiu, S.-H., Robert, and Last, L. (2015). The Impact of the Branched-Chain Ketoacid Dehydrogenase Complex on Amino Acid Homeostasis in Arabidopsis. Plant Physiology: 00461.2015.

Pichersky, E. and Gang, D.R. (2000). Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. Trends Plant Sci. **5**: 439–445.

Quinet, M., Angosto, T., Yuste-Lisbona, F.J., Blanchard-Gros, R., Bigot, S., Martinez, J.-P., and Lutts, S. (2019). Tomato Fruit Development and Metabolism. Front. Plant Sci. **10**: 1554.

Righetti, K., Vu, J.L., Pelletier, S., Vu, B.L., Glaab, E., Lalanne, D., Pasha, A., Patel, R.V., Provart, N.J., Verdier, J., Leprince, O., and Buitink, J. (2015). Inference of Longevity-Related Genes from a Robust Coexpression Network of Seed Maturation

Identifies Regulators Linking Seed Storability to Biotic Defense-Related Pathways. Plant Cell **27**: 2692–2708.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. **4**: Article32.

Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A.K., Nilo-Poyanco, R., Bernard, T., Kahn, D., and Rhee, S.Y. (2017). Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. Plant Physiol. **173**: 2041–2059.

Schmidt, B.M., Ribnicky, D.M., Lipsky, P.E., and Raskin, I. (2007). Revisiting the ancient concept of botanical therapeutics. Nat. Chem. Biol. **3**: 360–366.

Segal, E., Wang, H., and Koller, D. (2003). Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics **19 Suppl 1**: i264–71.

Serin, E.A.R., Nijveen, H., Hilhorst, H.W.M., and Ligterink, W. (2016). Learning from Co-expression Networks: Possibilities and Challenges. Front. Plant Sci. **7**: 444.

Shoji, T. and Hashimoto, T. (2011). Recruitment of a duplicated primary metabolism gene into the nicotine biosynthesis regulon in tobacco. Plant J. **67**: 949–959.

Stitt, M., Sulpice, R., and Keurentjes, J. (2010). Metabolic networks: how to identify key components in the regulation of metabolism and growth. Plant Physiol. **152**: 428–444.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F.M., Bassel, G.W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., and Provart, N.J. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. **32**: 1633–1651.

Uygun, S., Peng, C., Lehti-Shiu, M.D., Last, R.L., and Shiu, S.-H. (2016). Utility and Limitations of Using Gene Expression Data to Identify Functional Associations. PLoS Comput. Biol. **12**: e1005244.

Verpoorte, R. (1998). Exploration of nature's chemodiversity: the role of secondary metabolites as leads in drug development. Drug Discovery Today **3**: 232–238.

Virtanen, P. et al. (2019). SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. arXiv:1907.10121.

Wisecaver, J.H., Borowsky, A.T., Tzin, V., Jander, G., Kliebenstein, D.J., and Rokas, A. (2017). A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell **29**: 944–959.

Woo, J., MacPherson, C.R., Liu, J., Wang, H., Kiba, T., Hannah, M.A., Wang, X.-J.,

Bajic, V.B., and Chua, N.-H. (2012). The response and recovery of the Arabidopsis thaliana transcriptome to phosphate starvation. BMC Plant Biol. **12**: 62.

Yu, X.-H., Gou, J.-Y., and Liu, C.-J. (2009). BAHD superfamily of acyl-CoA dependent acyltransferases in Populus and Arabidopsis: bioinformatics and gene expression. Plant Mol. Biol. **70**: 421–442.

Zouine, M., Fu, Y., Chateigner-Boutin, A.-L., Mila, I., Frasse, P., Wang, H., Audran, C., Roustan, J.-P., and Bouzayen, M. (2014). Characterization of the tomato ARF gene family uncovers a multi-levels post-transcriptional regulation including alternative splicing. PLoS One **9**: e84203.

Figure 1.

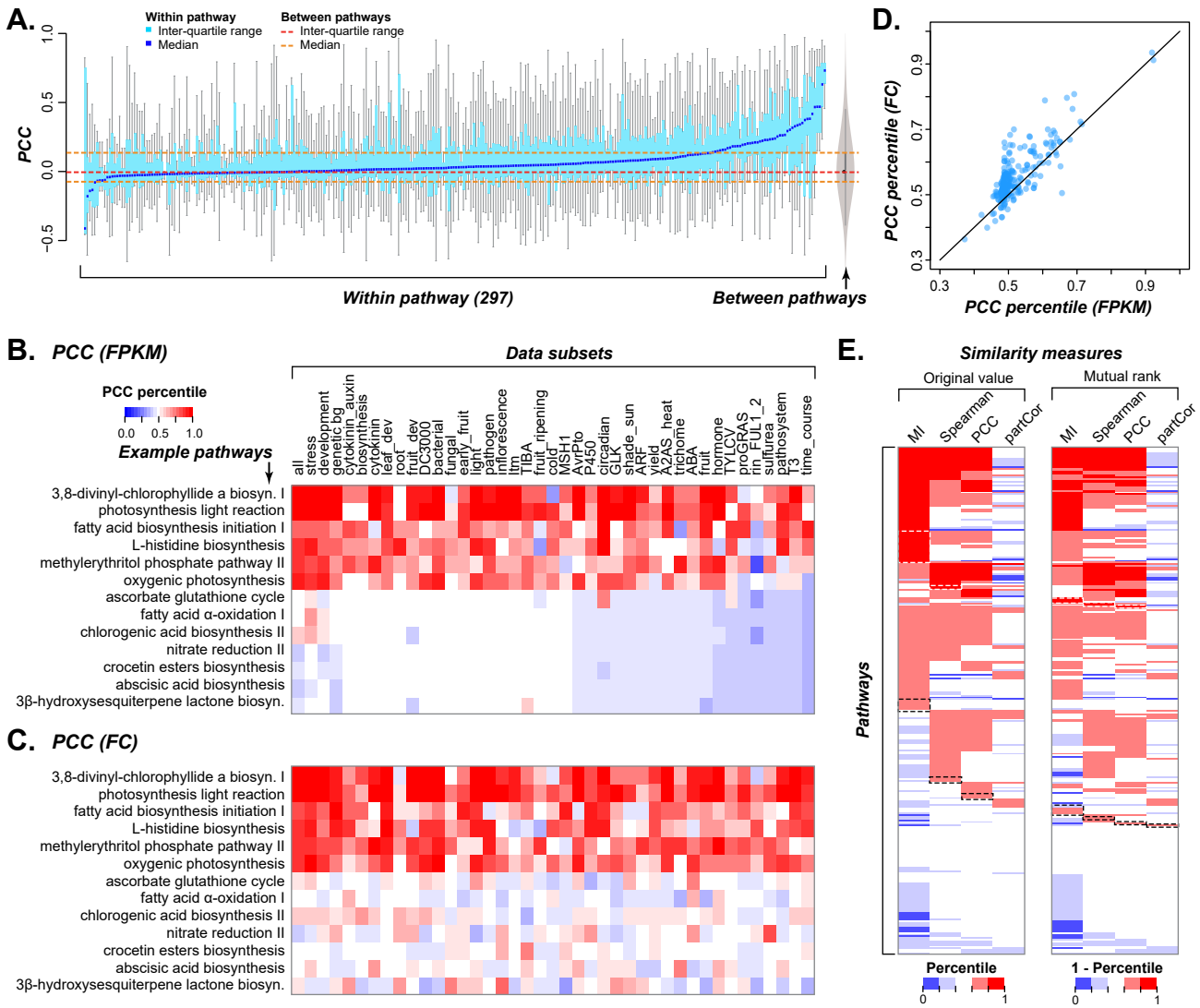

A.
Within pathway
Inter-quartile range
Median
Between pathways
Inter-quartile range
Median
PCC
Within pathway (297)
Between pathways
D.
PCC percentile (FC)
PCC percentile (FPKM)
B. PCC (FPKM)
Data subsets
PCC percentile
Example pathways
all
stress
development
genetic_bg
cytokinin_auxin
biosynthesis
cytokinin
leaf_dev
root
fruit_dev
DC3000
bacterial
fungal
early_fruit
light
pathogen
inflorescence
ltm
TIBA
fruit_ripening
cold
MSH1
AvrPto
P450
circadian
GLK
shade_sun
ARF
yield
A2AS_heat
trichome
ABA
fruit
hormone
TYLCV
proGRAS
rin_FUL1_2
sulfurea
pathosystem
T3
time_course
3,8-divinyl-chlorophyllide a biosyn. I
photosynthesis light reaction
fatty acid biosynthesis initiation I
L-histidine biosynthesis
methylerythritol phosphate pathway II
oxygenic photosynthesis
ascorbate glutathione cycle
fatty acid α-oxidation I
chlorogenic acid biosynthesis II
nitrate reduction II
crocetin esters biosynthesis
abscisic acid biosynthesis
3β-hydroxysesquiterpene lactone biosyn.
C. PCC (FC)
E.
Similarity measures
Original value
Mutual rank
MI
Spearman
PCC
partCor
Pathways
Percentile
1 - Percentile

Figure 2

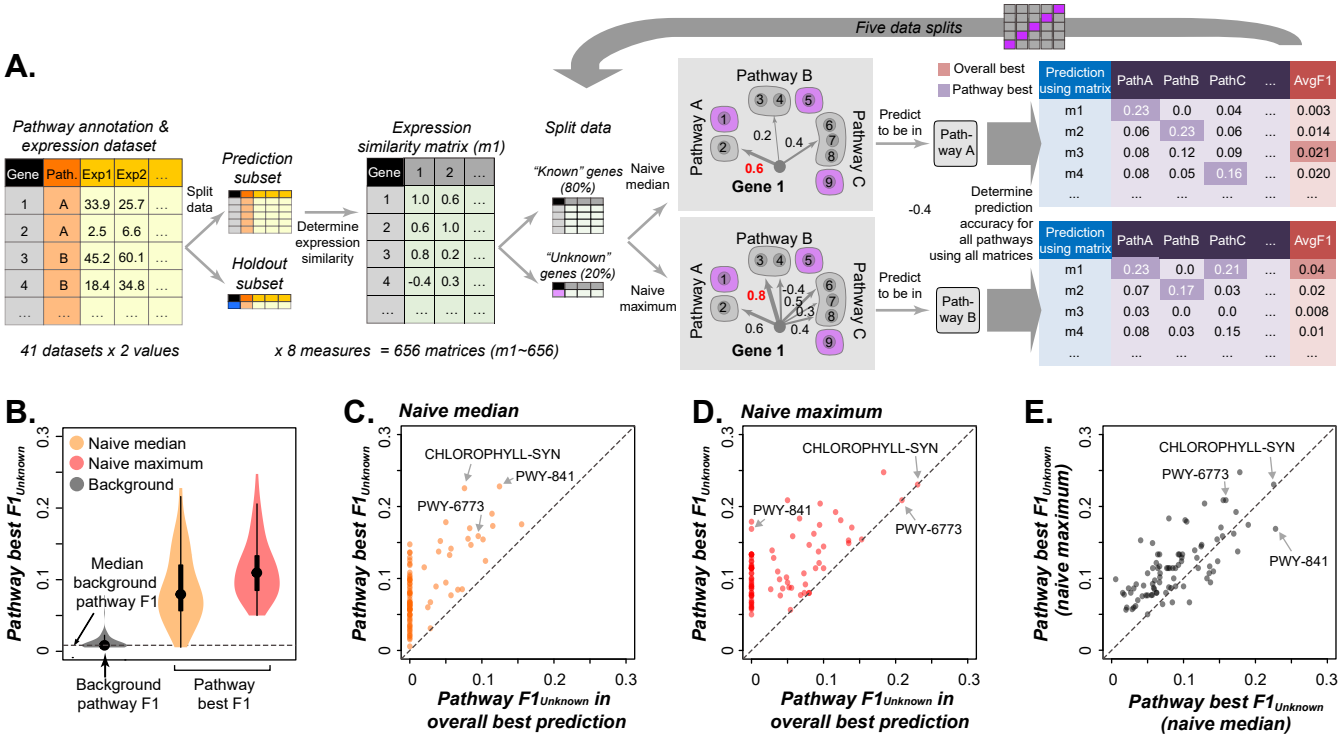

A.
Five data splits
Pathway annotation & expression dataset
Gene | Path. | Exp1 | Exp2 | ...
1 | A | 33.9 | 25.7 | ...
2 | A | 2.5 | 6.6 | ...
3 | B | 45.2 | 60.1 | ...
4 | B | 18.4 | 34.8 | ...
41 datasets x 2 values
Split data
Prediction subset
Holdout subset
Determine expression similarity
Expression similarity matrix (m1)
Gene | 1 | 2 | ...
1 | 1.0 | 0.6 | ...
2 | 0.6 | 1.0 | ...
3 | 0.8 | 0.2 | ...
4 | -0.4 | 0.3 | ...
x 8 measures = 656 matrices (m1~656)
Split data
"Known" genes (80%)
"Unknown" genes (20%)
Naive median
Naive maximum
Pathway A
Pathway B
Pathway C
Gene 1
Predict to be in
Path-way A
Path-way B
Overall best
Pathway best
Determine prediction accuracy for all pathways using all matrices
Prediction using matrix | PathA | PathB | PathC | ... | AvgF1
m1 | 0.23 | 0.0 | 0.04 | ... | 0.003
m2 | 0.06 | 0.23 | 0.06 | ... | 0.014
m3 | 0.08 | 0.12 | 0.09 | ... | 0.021
m4 | 0.08 | 0.05 | 0.16 | ... | 0.020
Prediction using matrix | PathA | PathB | PathC | ... | AvgF1
m1 | 0.23 | 0.0 | 0.21 | ... | 0.04
m2 | 0.07 | 0.17 | 0.03 | ... | 0.02
m3 | 0.03 | 0.0 | 0.0 | ... | 0.008
m4 | 0.08 | 0.03 | 0.15 | ... | 0.01
B.
Naive median
Naive maximum
Background
Median background pathway F1
Background pathway F1
Pathway best F1
Pathway best F1Unknown
C. Naive median
CHLOROPHYLL-SYN
PWY-841
PWY-6773
Pathway F1Unknown in overall best prediction
D. Naive maximum
CHLOROPHYLL-SYN
PWY-841
PWY-6773
Pathway F1Unknown in overall best prediction
E.
CHLOROPHYLL-SYN
PWY-6773
PWY-841
Pathway best F1Unknown (naive maximum)
Pathway best F1Unknown (naive median)

Figure 3

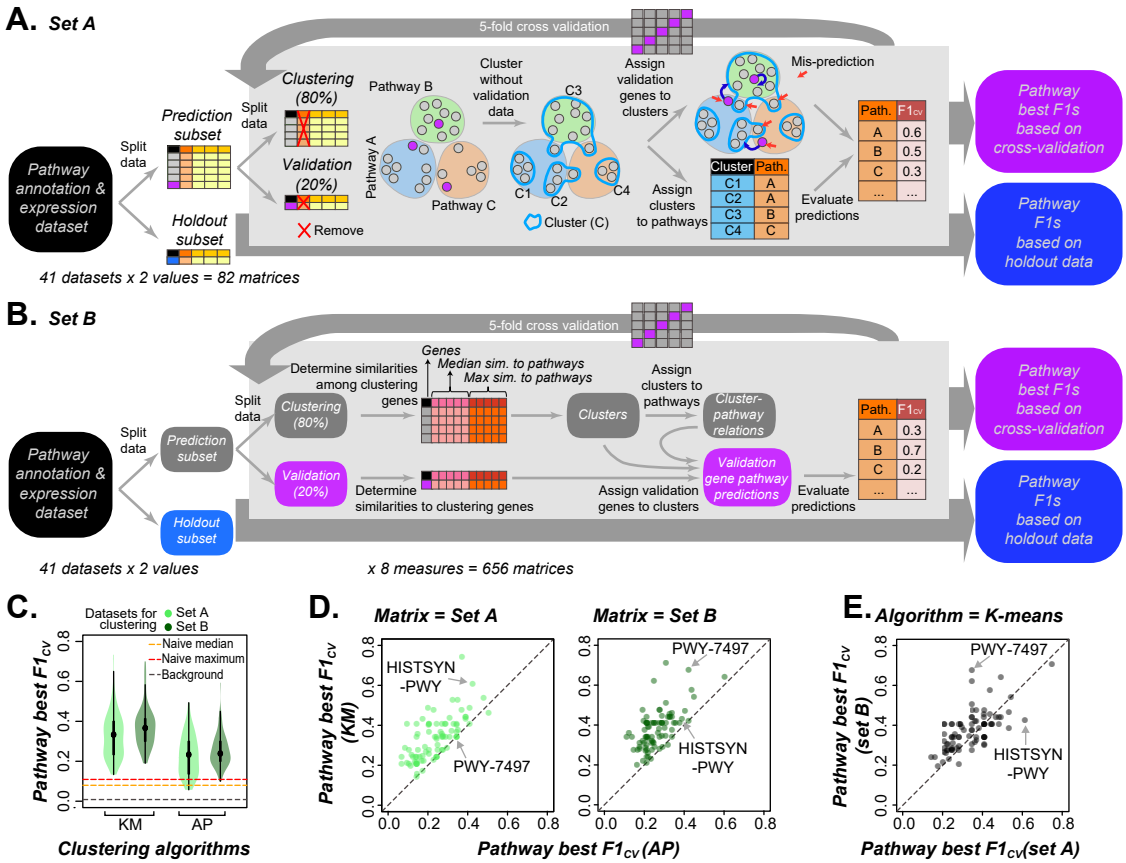
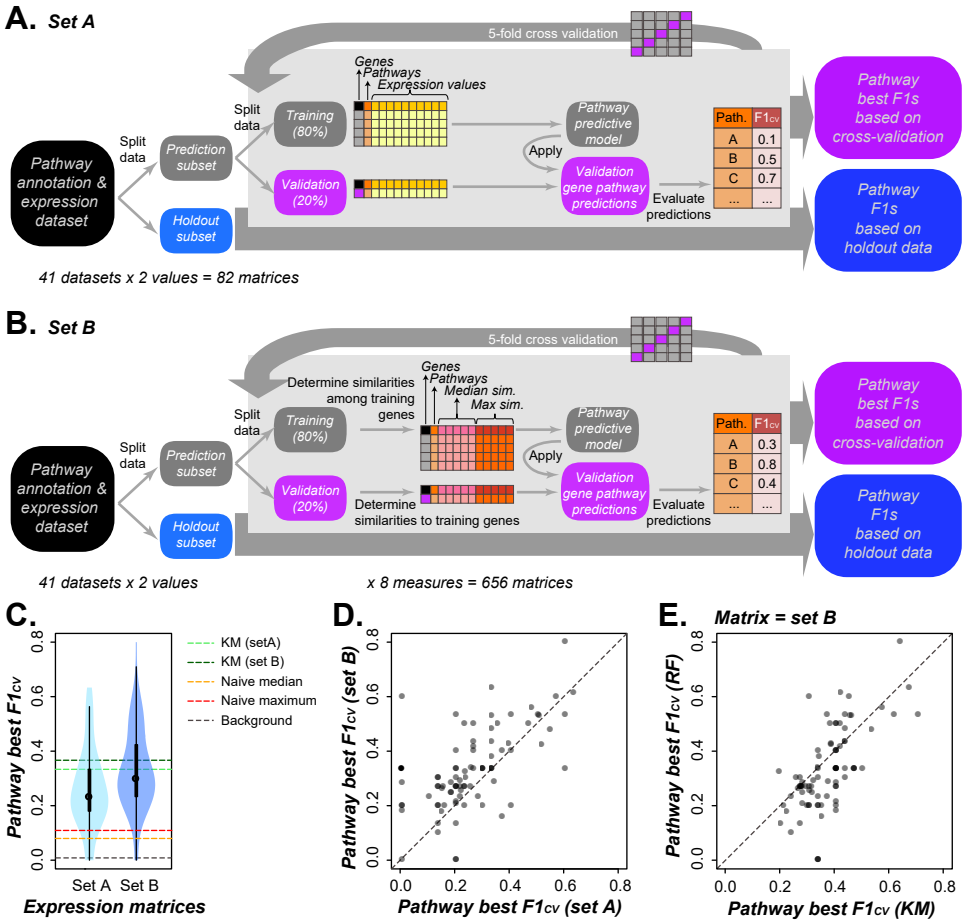
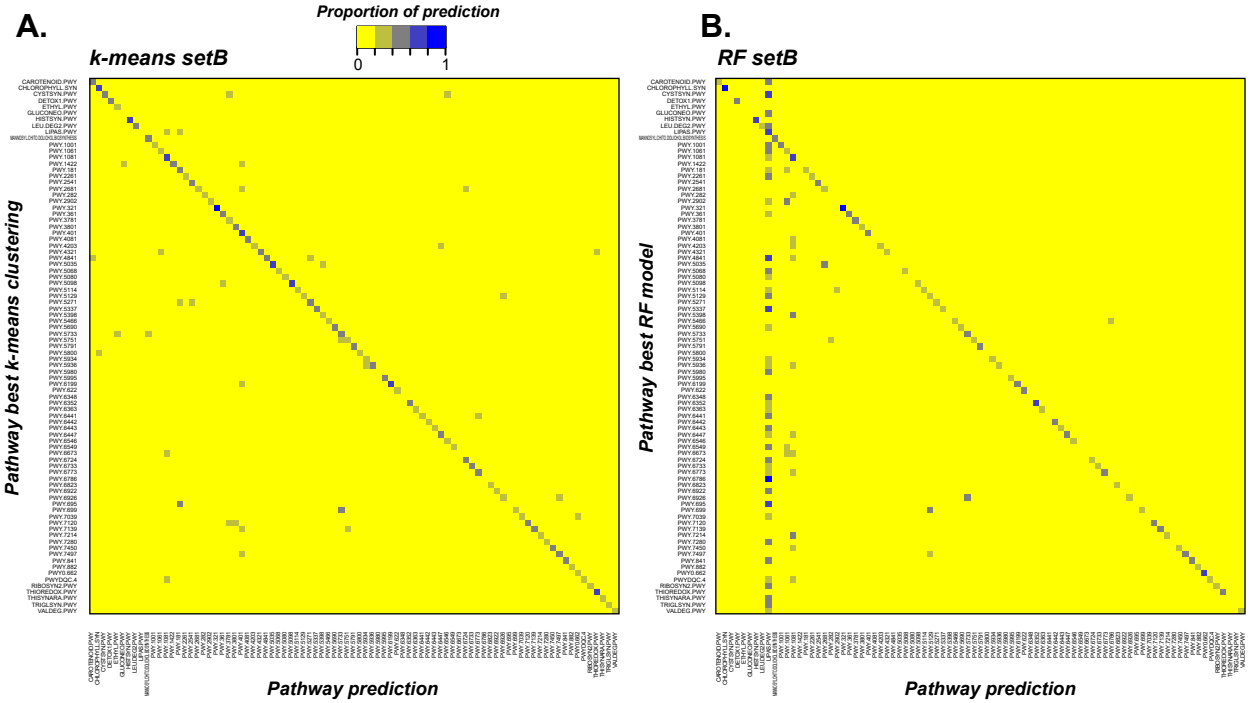

A. Set A
5-fold cross validation
Pathway annotation & expression dataset
Split data
Prediction subset
Holdout subset
Split data
Clustering (80%)
Validation (20%)
Remove
Pathway A
Pathway B
Pathway C
Cluster without validation data
C1
C2
C3
C4
Cluster (C)
Assign validation genes to clusters
Assign clusters to pathways
Mis-prediction
Cluster Path.
C1 A
C2 A
C3 B
C4 C
Path. F1cv
A 0.6
B 0.5
C 0.3
... ...
Evaluate predictions
Pathway best F1s based on cross-validation
Pathway F1s based on holdout data
41 datasets x 2 values = 82 matrices
B. Set B
5-fold cross validation
Pathway annotation & expression dataset
Split data
Prediction subset
Holdout subset
Split data
Clustering (80%)
Validation (20%)
Determine similarities among clustering genes
Genes
Median sim. to pathways
Max sim. to pathways
Clusters
Assign clusters to pathways
Cluster-pathway relations
Determine similarities to clustering genes
Assign validation genes to clusters
Validation gene pathway predictions
Evaluate predictions
Path. F1cv
A 0.3
B 0.7
C 0.2
... ...
Pathway best F1s based on cross-validation
Pathway F1s based on holdout data
41 datasets x 2 values
x 8 measures = 656 matrices
C.
Datasets for clustering
Set A
Set B
Naive median
Naive maximum
Background
Pathway best F1cv
KM
AP
Clustering algorithms
D.
Matrix = Set A
Matrix = Set B
Pathway best F1cv (KM)
HISTSYN -PWY
PWY-7497
Pathway best F1cv (AP)
E. Algorithm = K-means
Pathway best F1cv (set B)
PWY-7497
HISTSYN -PWY
Pathway best F1cv(set A)

Figure 4


A. Set A
5-fold cross validation
Genes
Pathways
Expression values
Split data
Training (80%)
Validation (20%)
Pathway predictive model
Apply
Validation gene pathway predictions
Evaluate predictions
Path.
F1cv
A 0.1
B 0.5
C 0.7
... ...
Pathway annotation & expression dataset
Split data
Prediction subset
Holdout subset
Pathway best F1s based on cross-validation
Pathway F1s based on holdout data
41 datasets x 2 values = 82 matrices
B. Set B
5-fold cross validation
Determine similarities among training genes
Genes
Pathways
Median sim.
Max sim.
Split data
Training (80%)
Validation (20%)
Determine similarities to training genes
Pathway predictive model
Apply
Validation gene pathway predictions
Evaluate predictions
Path.
F1cv
A 0.3
B 0.8
C 0.4
... ...
Pathway annotation & expression dataset
Split data
Prediction subset
Holdout subset
Pathway best F1s based on cross-validation
Pathway F1s based on holdout data
41 datasets x 2 values
x 8 measures = 656 matrices

C.
Pathway best F1cv
0.0
0.2
0.4
0.6
0.8
KM (setA)
KM (set B)
Naive median
Naive maximum
Background
Set A
Set B
Expression matrices
D.
Pathway best F1cv (set B)
Pathway best F1cv (set A)
E.
Matrix = set B
Pathway best F1cv (RF)
Pathway best F1cv (KM)

Figure 5


A.
k-means setB
Proportion of prediction
0
1
Pathway best k-means clustering
Pathway prediction
B.
RF setB
Pathway best RF model
Pathway prediction