

DR PEIPEI WANG (Orcid ID : 0000-0002-7580-9627)

Article type : Methods

Methods

Optimizing the use of gene expression data to predict plant metabolic pathway memberships

Peipei Wang (orcid.org/0000-0002-7580-9627)¹, Bethany M. Moore (orcid.org/0000-0002-2104-7292)², Sahra Uygun (orcid.org/0000-0003-0863-0384)³, Melissa D. Lehti-Shiu (orcid.org/0000-0003-1985-2687)¹, Cornelius S. Barry (orcid.org/0000-0003-4685-0273)⁴, and Shin-Han Shiu (orcid.org/0000-0001-6470-235X)^{1,5}

¹Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA

²Department of Botany, University of Wisconsin-Madison, Madison, WI 53706, USA

³Agendia Inc. Irvine, CA 92618, USA

⁴Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA

⁵Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824, USA

Author for correspondence:

Shin-Han Shiu

Tel: +1-517-353-7196

Email: shius@msu.edu

Received: 1 February 2021

Accepted: 13 March 2021

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/NPH.17355](https://doi.org/10.1111/NPH.17355)

This article is protected by copyright. All rights reserved

Summary

- Plant metabolites from diverse pathways are important for plant survival, human nutrition and medicine. The pathway memberships of most plant enzyme genes are unknown. While co-expression is useful for assigning genes to pathways, expression correlation may exist only under specific spatiotemporal and conditional contexts.
- Utilizing >600 tomato (*Solanum lycopersicum*) expression data combinations, three strategies for predicting memberships in 85 pathways were explored.
- Optimal predictions for different pathways require distinct data combinations indicative of pathway functions. Naive prediction (i.e., identifying pathways with the most similarly expressed genes) is error prone. In 52 pathways, unsupervised learning performed better than supervised approaches, likely due to limited training data availability. Using gene-to-pathway expression similarities led to prediction models that outperformed those based simply on expression levels. Using 36 experimental validated genes, the pathway-best model prediction accuracy is 58.3%, significantly better than that for predicting annotated genes without experimental evidence (37.0%) or random guess (1.2%), demonstrating the importance of data quality.
- Our study highlights the need to extensively explore expression-based features and prediction strategies to maximize the accuracy of metabolic pathway membership assignment. The prediction framework outlined here can be applied to other species and serve as a baseline model for future comparisons.

Keywords: gene expression, machine learning, metabolic pathway, prediction, tomato

Introduction

Metabolites are products of pathways consisting of a linked series of enzymes (Berg *et al.*, 2002). Plants produce diverse metabolites essential for growth and survival (Stitt *et al.*, 2010). Some metabolites are also important for human nutrition and medicine (Martin & Li, 2017; Schlapfer *et al.*, 2017). Although there is an increasing body of knowledge about plant metabolic pathways (Verpoorte, 1998; Pichersky & Gang, 2000; Kim & Buell, 2015), many enzyme genes in known pathways remain to be identified and there are likely unknown pathways (De Luca *et al.*, 2012; Schlapfer *et al.*, 2017). A key challenge in understanding plant metabolic pathways is that genes encoding metabolic enzymes tend to exist as members of large gene families. In addition, experimental assessments of metabolic pathway membership can be laborious. Therefore, it is important to prioritize candidate genes for functional analyses using computational predictions.

Substantial effort has been devoted to predicting pathway membership of plant enzyme genes. One approach is to first identify candidate genes within a genome, and then assign them into pathways according to the reactions catalyzed by the encoded enzymes (Karp *et al.*, 2011; Chae *et al.*, 2014; Schlapfer *et al.*, 2017). This approach was utilized by the Plant Metabolic Network with pathway annotations for 125 plant and green algal species; here the prediction of unknown genes is mainly based on sequence similarity to experimentally evaluated enzyme genes. Another approach utilizes the physical colocalization of enzyme genes to identify biosynthetic gene clusters (Nutzmann *et al.*, 2016; Mao *et al.*, 2020), which tend to contain specialized metabolic pathway genes (Osborn, 2010; Medema *et al.*, 2015). But the colocalization criterion by itself can be error prone (Wisecaver *et al.*, 2017). Chemical-chemical, chemical-protein, and protein-protein interactions (Gao *et al.*, 2012) and information integrated from other methods (e.g., chemical transformations, thermofluor screening, metabolic endpoints) (Calhoun *et al.*, 2018) have also been used for prediction. In addition, localization in the same subcellular compartments was used to identify functionally related genes (Huh *et al.*, 2003), and to reconstruct metabolic networks (Forster *et al.*, 2003). While the above information is useful, not all of them is readily available except in well studied model organisms.

In contrast, transcriptome data have become available for a growing number of plants and have been used to predict metabolic pathway genes based on the assumption that genes from the same pathway are co-expressed (Segal *et al.*, 2003; Kim & Buell, 2015). Considering the ease of generating

transcriptome data, these data are an important resource for computational inference of gene function. There are three general strategies for leveraging expression data for functional inference. First, “naive prediction” (because it is the simplest approach) is to ask, for a gene of unknown function, which genes with known functions have the highest expression similarities to that gene. The utility of this strategy has been shown in a number of single-gene studies (Hirai *et al.*, 2007; Righetti *et al.*, 2015), but its accuracy on a genome-wide scale is unclear. The second strategy is unsupervised machine learning, where genes are first grouped into co-expression clusters and then genes of unknown function are assigned functions based on genes with known functions over-represented within the same cluster (Mutwil *et al.*, 2011; Uygun *et al.*, 2016; Gupta & Pereira, 2019). The third strategy is supervised machine learning where the function of a gene is predicted using models learned from the expression profiles of genes with known functions. While supervised learning has been applied to predict functions (Kaundal *et al.*, 2010; Lloyd *et al.*, 2015; Ni *et al.*, 2016; Moore *et al.*, 2019), it has not been used to predict plant metabolic pathway membership using transcriptome data. Thus, it is unknown which of these three strategies is more effective and whether their accuracy varies depending on the pathway.

In addition, it is unresolved how transcriptome data should be used in pathway membership prediction. One approach is to use as many different expression samples as possible (Aoki *et al.*, 2016; Obayashi *et al.*, 2018), but expression similarities measured using distinct subsets of expression data can be more accurate for inferring functional relationships (Usadel *et al.*, 2009; Uygun *et al.*, 2016). Thus, it is important to optimize the use of specific expression data in pathway prediction. Here, we utilized tomato as a model to optimize pathway prediction because there is considerable knowledge of the diverse metabolic pathways in this species and availability of a large collection of transcriptome data. We investigated the effect of expression dataset, expression value, and gene expression similarity measure on the ability to predict pathway membership using three strategies: naive prediction, unsupervised learning, and supervised learning.

Materials and Methods

Gene functional annotation

Metabolic pathway annotations of genes in *Solanum lycopersicum* were from TomatoCyc V3.0 (Plant Metabolic Network, PMN 12.0, <https://www.plantcyc.org/>) (Schlapfer *et al.*, 2017); the gene annotations were from Solanaceae Genomics Network (SGN)_V3.2 (<https://solgenomics.net/>). A gene in SGN_V3.2 was considered a match to an NCBI_V2.5 entry if the alignment identity score was 100% with no gaps using the BLAST-like alignment tool (Kent, 2002); the matches are in **Table S1**. We assigned EC numbers and/or reactions to 11,036 tomato genes using PMN Ensemble Enzyme Prediction Pipeline V3.0 (<https://gitlab.com/rhee-lab/E2P2/>) with default settings; 2,395 genes with ECs/reactions were in 485 TomatoCyc pathways (**Table S2**). Genes not expressed (Fragments Per Kilobase Million, FPKM=0) in all 372 experiments (see next section) and pathways with <5 annotated genes were excluded, resulting in 2,171 genes in 297 pathways. For pathway membership prediction, 1,050 genes annotated to more than one pathway were removed as well as pathways with < 5 genes after this filtering step, resulting in 972 genes in 85 pathways (**Table S3**).

RNA-seq data and processing

RNA-seq data were obtained from 47 studies (**Table S4,5**). Reads were trimmed using Trimmomatic (Bolger *et al.*, 2014), and mapped to the tomato genome (NCBI_V2.5) using TopHat2 (Kim *et al.*, 2013). Samples with an overall read mapping rate <80% were discarded, and only reads uniquely mapped to the genome were used for the calculation of FPKM using cufflinks (Trapnell *et al.*, 2010). See **Table S6** for software parameters. Read count was calculated using HTSeq v 0.11.2 (Anders *et al.*, 2015), and Transcripts-Per-Million (TPM) was calculated using the calculateTPM function of the R package scater v1.0.4. Fold change (FC) in gene expression level between two samples was calculated using edgeR (Robinson *et al.*, 2010). Median FPKM or TPM value among replicates was used as the expression level estimate for a gene in an experiment.

Samples from the 47 studies were assigned to 41 “datasets”: 36 individual and five combined. The 36 individual datasets were 36 studies with ≥ 3 experiments. The five combined sets were as follows: 1) tissues/stages/circadian—wild-type plants taken from different tissues, at different development stages, or at different times of day; 2) genetic background—comparison studies of wild

type, mutant, or knocked down or overexpression transgenics; 3) hormone treatment—plants treated with hormones and control; 4) stress treatment – plants treated in various stress conditions and control; 5) all experiments (372 total) (**Table S4**).

Gene expression similarity measure

Eight similarity measures were evaluated. Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (Spearman) between expression values of two genes, were calculated with the `scipy.stats` module (Virtanen *et al.*, 2020). Mutual information (MI), was determined using the `sklearn.metrics.cluster` module (Pedregosa *et al.*, 2011). Partial correlation (partCor) was determined with the `corpcor` R package (Schafer & Strimmer, 2005). Mutual rank (MR) was calculated for each of these four measures (thus generating four more measures) as follows:

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

Where $\text{rank}_{(\text{gene1} \rightarrow \text{gene2})}$ and $\text{rank}_{(\text{gene2} \rightarrow \text{gene1})}$ indicate the rank of expression similarity between *gene1* and *gene2* among all expression similarity values calculated between *gene2* and other genes and between *gene1* and other genes, respectively. Thus, a smaller rank indicates higher expression similarity.

Splitting data for testing and modeling and measuring prediction performance

To assess how well the approaches predict pathway membership, five genes from each of the six pathways with ≥ 25 genes expressed in ≥ 1 sample were held out as test data and never used in any “model” building process. The remaining genes in these six pathways as well as all genes from the remaining 79 pathways were split into five groups. Genes in four of the five groups were referred to as “training” genes for all three approaches: naive prediction, unsupervised, and supervised. Genes in the remaining, fifth, group were used as “validation” genes to evaluate the prediction performance of the three approaches. This split was repeated five times to ensure that every gene was placed in a validation subset once.

F-measure (F1) was calculated by comparing the annotated and predicted pathway membership for validation genes in a pathway (referred as $F1_{CV}$), using the equation: $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$, where $\text{precision} = \frac{TP}{TP + FP}$, $\text{recall} = \frac{TP}{TP + FN}$, *TP*: true positive, i.e., number of genes

annotated and predicted as being a member of the pathway. *FP*: false positive, i.e., number of genes mis-predicted as being in the pathway. *FN*: false negative, i.e., number of genes annotated but not predicted as being in the pathway. The reported $F1_{CV}$ for each pathway in a model built with a specific data combination is the average F1 across all five folds mentioned above. F1 ranges from 0 and 1, with 1 indicating a perfect model.

Unsupervised learning

Unsupervised clustering was conducted using the “KMeans”, “AffinityPropagation”, “Birch” and “MeanShift” functions in `sklearn.cluster` (Pedregosa *et al.*, 2011), which have the “predict” method option. The hyperparameter space for each algorithm is in **Table S6**. To make predictions, we first asked genes in which pathways were enriched statistically ($p < 0.05$, Fisher’s exact test) in a cluster *C*. Then, among this set of enriched pathways, genes in *C* were predicted as in pathway *P* if *P* had the highest enrichment value (*E*) for cluster *C* defined as:

$$E = \log_e \left(\frac{P_C}{P_A} \right),$$

Where P_C is the number of cluster *C* genes in pathway *P* divided by the total number of cluster *C* genes; and P_A is the number of pathway *P* genes divided by the total number of genes analyzed (i.e., randomly expected overlap). Although enrichment values are negatively correlated with *p*-values, the enrichment value was used because it conveyed information about effect size. For validation genes not used for generating clusters, the cluster memberships were predicted using the “predict” function in `sklearn.cluster`. The same prediction procedure was used to predict pathway memberships for test genes.

Supervised learning

RandomForestClassifier (RF), AdaBoostClassifier (AB), LinearSVC (SVC), KNeighborsClassifier (KNN), and MLPClassifier (neural network, NN_1) from `sklearn`, and Sequential API (NN_2) from Keras (<https://keras.io/>) were used to establish multi-class models using training genes. Hyperparameters were determined by performing a grid search using the five-fold cross-validation scheme, with the goal of maximizing $F1_{CV}$ for each of the 82 Set A or 656 Set B data

combinations. A final model for a data combination was built with the best hyperparameter identified and used to predict pathway memberships of validation and test genes. The hyperparameter space is in **Table S6**.

To balance the numbers of genes among pathways, the training data of smaller pathways (i.e., those < 56) were up-sampled to 56 genes per pathway using the SMOTE function from `imblearn.over_sampling` (Blagus & Lusa, 2013), with `sampling_strategy='not majority'`, `random_state=42`, `k_neighbors=3`. The impurity-based feature importance was used to compare the relative degrees of contribution of features to the RF model—the higher the feature importance, the higher the relative contribution to the models—and was determined using the attribute `feature_importances_` of `RandomForestClassifier` with `criterion=gini`.

Results

Relationship between gene expression similarity and metabolic pathway membership

To assess the extent to which genes within the same metabolic pathway have similar expression profiles, we collected annotation data for 2171 tomato genes from 297 pathways, each with ≥ 5 annotated genes (**Table S1,2**), and transcriptome data from 372 experiments (each with multiple biological/technical replicates, **Table S4,5**). Using PCC calculated from enzyme transcript levels (in FPKM) from all 372 experiments as the expression similarity measure, for 124 out of 297 pathways (41.8%) there was significantly higher gene expression similarity between pairs of genes from the same pathway than between pairs randomly chosen from different pathways (Wilcoxon signed-rank test, $p < 0.05$; **Fig. 1a**, **Table S7**). When the maximum expression similarity between a gene and all other genes in the same pathway was examined, genes in 202 of 297 pathways (68.0%) had significantly higher expression similarities within a pathway than between pathways (**Fig. S1a**, **Table S7,8**). This finding was not simply due to paralogs that might have higher expression similarities than non-paralogous pairs (**Fig. S1d**), indicating that the maximum expression similarity within a pathway was more useful for distinguishing genes within and between pathways than the median similarity (**Fig. S1b,c**). However, using maximum PCC as a criterion, 72.7% of genes did not have above-threshold within-pathway PCC values (PCC=0.66, the 95th percentile value of between-pathway gene pairs) (**Fig. S1a**).

We next asked how similarity measures should be generated to best identify pathway membership by considering: (1) expression datasets, e.g., all data or a subset; (2) expression values, i.e., expression levels (in FPKM) or contrasts (FC); and (3) similarity measures, e.g., PCC vs. Spearman's Rank. For a pair of genes in a pathway, we were interested in how their expression similarity compared with a distribution of between-pathway gene similarities, which we treated as the background, null distribution. Thus, in all subsequent analyses, we determined the median percentile value of similarity for all within-pathway pairs in the between-pathway similarity distribution (hereafter referred to as percentile_{BP}). The higher the percentile_{BP} , the higher the similarity.

The effect of expression dataset on expression correlation among pathway genes

To evaluate the effect of dataset on the pathway percentile_{BP} , we used 41 expression datasets: 36 from individual studies and five combined sets (see **Methods**). Using FPKM as the expression value and PCC as the similarity measure, the datasets that produced the highest percentile_{BP} differed substantially between pathways (13 example pathways shown in **Fig. 1b**, results for all pathways in **Table S9**). Some pathways (labeled ① through ⑥ in **Fig. 1b**), such as those relevant to photosynthesis, tended to have high PCC percentile_{BP} values for most datasets. For example, expression of genes in the 3,8-divinyl-chlorophyllide a biosynthesis I pathway (labeled ① in **Fig. 1b**) was well correlated in all datasets except fungal inoculation, cold treatment and treatment with paclobutrazol and gibberellic acid (proGRAS dataset, **Table S4**), consistent with the finding that photosynthesis is disrupted under stress conditions (Nouri *et al.*, 2015). Expression of genes in methylerythritol phosphate (MEP) pathway II (⑤, **Fig. 1b**) are correlated except when the *FRUITFULL1/2* (*FUL1/2*) and *RIPENING INHIBITOR* (*RIN*) mutants, *MutS HOMOLOG1* gene silencing and *Pseudomonas* inoculation datasets were used. It is known that carotenoid biosynthesis pathway genes downstream from MEP are regulated by *FUL1/2* and *RIN* (Fujisawa *et al.*, 2014). This indicates that, when *FUL1/2* and *RIN* are mutated, MEP II genes may not be properly regulated and thus not co-expressed.

In contrast, genes in pathways 7 through 13 were either correlated in a highly dataset-specific manner or are not correlated. For example, expression of genes in the fatty acid α -oxidation I pathway (⑧, **Fig. 1b**) and chlorogenic acid biosynthesis II pathway (⑨, **Fig. 1b**) was only correlated when

the stress combined dataset was used, consistent with the role of these pathways in protecting plants against environmental perturbations (De Leon *et al.*, 2002; Niggeweg *et al.*, 2004). These results demonstrate the need to consider datasets that best reflect the biological processes in which different pathways participate.

Using FC values instead of FPKM to determine PCC led to improved percentile_{BP} values for all example pathways (**Fig. 1c**, **Table S10**); 250 of 297 pathways had higher percentile_{BP} when using FC (**Fig. 1d**), demonstrating the importance of expression value format. We should note that the crocetin ester biosynthesis pathway in crocus (Carmona *et al.*, 2006) and the 3 β -hydroxysesquiterpene lactone biosynthesis pathway in feverfew (Majdi *et al.*, 2011) (⑪ and ⑬ in **Fig. 1b,c**) are not expected to be present in tomato. The nonrandom degree of co-expression between genes in these pathways observed for FC values (**Fig. 1c**) may reflect the existence of similar as yet uncharacterized pathways in tomato.

Thus far, we used PCC as a similarity measure to assess linear correlation. Because gene expression correlation can be non-linear or context dependent, we explored three additional similarity measures—Spearman's rank, MI, and partCor—as well as the MR for each of the four measures. Genes in some pathways only displayed high similarity when specific measures were used (**Fig. 1e**, **Table S11**)—e.g., the similarity scores of genes from 19 pathways were in the highest quintile only when MI was used (left panel, **Fig. 1e**). Using MR of MI, an additional four pathways had similarity scores in the highest quintile (right panel, **Fig. 1e**). The same patterns are illustrated by the three pathways shown in **Fig. S2**. Our findings highlight the importance of exploring expression datasets, expression values, and similarity measures when evaluating gene expression similarity for different pathways. Thus, in subsequent analyses, we determined expression correlations using 41 datasets, 2 types of expression values, and 8 similarity measures, yielding 656 possible data combinations.

Naive prediction of pathway genes

We next asked what approaches should be used to predict pathway memberships based on expression similarity by exploring three general approaches: naive prediction, unsupervised learning, and supervised learning. For naive prediction, we explored two methods: 1) naive median: If gene X has the highest median expression similarity with genes in pathway A, then gene X is predicted to be

in pathway A; and 2) naive maximum: If gene X has the maximum expression similarity with gene Y and gene Y is in pathway B, then gene X is predicted to be in pathway B (**Fig. 2a**). A gene was not predicted to be in any pathway if it had ≥ 2 pathway assignments using either approach.

The ultimate goal for pathway prediction is to predict pathway membership of unknown genes. Thus, we split annotated genes in a pathway into two subsets: (1) a “known” subset consisted of “training” genes used for building prediction “models”, which here were simply the naive median and naive maximum rules for naive predictions and (2) an “unknown” subset consisted of “validation” genes—for which annotation information was actually available—used to validate the models in a five-fold cross-validation scheme (**Fig. 2, Methods**). Both naive approaches were applied to all 656 data combinations, resulting in 2×656 naive prediction models. To evaluate model predictions, the average F1 was calculated for each pathway based on prediction of validation genes across all five CV folds (referred to as $F1_{CV}$ with 1 indicating a perfect model, **Fig. 2a**). The end results were two $F1_{CV}$ matrices, one for the naive median and one for the naive maximum, that were each 656 (number of data combinations) by 85 (number of pathways) in dimension (**Table S12,13**).

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 $F1_{CV}=0.04$ (all experiments, FPKM as the expression value, MR of PCCs as the similarity measure, using the naive maximum method). Although better than a random guess ($F1=0.01$, dotted line, **Fig. 2b**, and **Table S14**), the prediction is suboptimal. Because different data combinations affect pathway membership recovery (**Fig. 1**), we next identified the data combination that led to the highest $F1_{CV}$ scores for each pathway (pathway best $F1_{CV}$, purple boxes, **Fig. 2a**). Genes in a pathway were better predicted when the pathway-specific optimal data combination was used instead of the overall best combination; the median pathway best $F1_{CV}$ values using the naive median and the naive maximum methods were 0.08 and 0.11, respectively (**Fig. 2b**). In nearly all cases, the pathway $F1_{CV}$ values obtained using the overall best data combination were substantially lower (in many cases 0) than the pathway best $F1_{CV}$ values (**Fig. 2c,d**). These findings demonstrate the need to identify the optimal combinations of datasets, expression values, and similarity measures when making predictions for different pathways.

We also found that the naive maximum method generally performed better than the naive median (**Fig. 2e**), consistent with our finding that maximum expression similarity yields higher

percentile_{BP} values (**Fig. S1**). This is likely because genes in the same pathway can be regulated at multiple levels beyond transcription and/or function in different branches of the pathway, resulting in relatively lower median expression similarity within pathways. Nonetheless, even for the naive maximum method, the median value of pathway best $F1_{CV}$ among the 85 pathways was only 0.11. Thus, despite the usefulness of the naive prediction approach for assigning enzyme genes to pathways, there remains substantial prediction errors (**Fig. S3**).

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, but the effects of using different datasets and types of expression values are mostly unexplored. To investigate the effect of clustering algorithms and parameters on prediction performance, we focused on four algorithms with “predict” functions: *k*-means, Affinity Propagation, Birch, and MeanShift. Clusters built using training genes can be used to predict the cluster membership (and pathway membership) of “unknown” genes. 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 (FPKM or FC, referred to as Set A, **Fig. 3a**). The second type was generated by first determining the expression similarities of each gene to other genes and then calculating the median and maximum similarities (8 measures) of genes in pathways to the gene in question (Set B, **Fig. 3b**). There were 656 Set B data combinations ($41 \times 2 \times 8$), with each combination consisting of 170 variables (2 similarity values \times 85 pathways). Note that the same training/validation data used for naive predictions were also used for unsupervised learning.

To predict pathway membership for a gene, we first assigned each cluster *C* to a pathway *P* based on enrichment values (see **Methods**). Next, a gene in the validation set was assigned to pathway *P* if the distance between the gene in question and the cluster *C* centroid was less than that between the gene and other cluster centroids. Because these clusters were used in making predictions, they are referred to as “clustering models”. The above process was repeated for each of the 82 Set A and 656 Set B data combinations to identify the dataset that led to clusters yielding the best predictions (i.e., highest $F1_{CV}$) for each pathway. Independently of algorithm or dataset, clustering-

based, unsupervised learning greatly outperformed naive approaches (**Fig. 3c**). This may be because the naive approaches only considered one expression correlation value (maximum or median), while the unsupervised approach utilized multiple expression correlation values between a gene and many other genes in multiple pathways.

Nevertheless, like the naive approaches, no single clustering model yielded good predictions for most pathways; the best overall $F1_{CV}$ averaged across pathways was only 0.09 (**Fig. S4**). Although MeanShift had the highest overall $F1_{CV}$, k -means performed best in terms of pathway best $F1_{CV}$, while Affinity Propagation was least effective (**Fig. 3c,d and Fig. S4**). In addition, k -means models using Set B outperformed those using Set A for 57 of 85 (67%) pathways (**Fig. 3e**). This may be because there were too few features in some Set A datasets for clustering; the median number of features (expression values or contrasts) was 8 for Set A, whereas each Set B dataset had 170 features (median and maximum gene-to-pathway expression similarities). Consistent with this, there was a significant positive correlation (Spearman's $\rho=0.47$, $p=6.4e-6$) between the pathway best $F1_{CV}$ and the number of features in the corresponding Set A dataset (**Fig. S5**). Another possibility is that gene-to-pathway expression similarity provided more information than the gene expression profiles for most pathways, allowing the structures of the metabolic pathway/network to be captured by the unsupervised methods.

Prediction of pathway genes using supervised learning methods

Different from unsupervised methods, where pathway information is not used for clustering, supervised machine learning methods build predictive models by learning from pathway annotations. Pathway prediction was framed as a multi-class learning problem, i.e., predicting which of the 85 pathways (classes) a gene belongs to using the Set A and Set B datasets (**Fig. 4a,b**). We first started with RF, which typically performs well in bioinformatic applications (Boulesteix *et al.*, 2012; Qi, 2012). Because supervised learning methods directly associate the pathway labels with the underlying data, our expectation was that RF models would outperform clustering-based predictions. Contrary to our expectation, the RF models had an overall lower performance (median pathway best $F1_{CV}=0.23$ for Set A and 0.3 for Set B data) compared with k -means models (0.33 for Set A and 0.37 for Set B; **Fig. 4c, Table S14**).

One potential reason for the difference in performance is that there were too few genes in most pathways (median pathway size=8 after filtering out small pathways with size<5) for RF to effectively generalize the features shared by genes in a pathway. Consistent with this, the differences between pathway best $F1_{CV}$ values of RF models and k -means models were weakly, but significantly correlated with pathway size, with $\rho=0.41$ (Set A, p -value=1.1e-4) and 0.28 (Set B, p -value=0.01; **Fig. S5**). These results suggest that when there are too few genes in each pathway, k -means may be superior to RF. Nonetheless, the weak correlations between pathway size and $F1_{CV}$ differences suggests that other factors may explain the poorer performance of RF models.

Upon examination of the confusion matrices (i.e., matrices showing the proportion of genes in a pathway predicted as being in each of the 85 pathways) for k -means and RF models (**Fig. 5a,b, Table S15,16**), we found that 260 genes from 64 pathways and 84 genes from 42 pathways were mis-predicted to be in two pathways with the most annotations, which jointly contributed to 12.6% of the training instances: triacylglycerol degradation (LIPAS-PWY, 75 genes) and homogalacturonan degradation (PWY-1081, 54), respectively (**Table S3**). These mis-predictions were a major reason for the poorer performance of RF models. To alleviate the impact of sampling bias on model performance, we balanced the training data by randomly up-sampling the minority pathways to the size of the largest pathway in the training subset so that all 85 pathways were the same size ($n=56$) in the training subset. The instances in the validation subset were kept unchanged for performance comparisons against models based on unbalanced data. Only six and 15 genes from three and nine pathways were mis-predicted as being in LIPAS-PWY and PWY-1081, respectively, and the prediction errors were relatively evenly distributed across pathways (**Fig. 5c, Table S17**). The resulting median pathway best $F1_{CV}$ for balanced RF Set B models was 0.33, higher than that of unbalanced RF Set B models (0.30), but still lower than that of k -means Set B models (0.37) (**Fig. 4c, and Fig. 5d-f**). Only balanced RF models are further discussed.

In addition to the differences in pathway best $F1_{CV}$, the best data combination also differs for k -means and RF Set B models—only 16 of the 85 pathways had the same best data combinations (**Table S18**). Nonetheless, the prediction performances for the same data combinations were significantly correlated between k -means and RF models (**Fig. 5g,h, and Fig. S6**), suggesting that there are some commonalities in how k -means and RF models learn from the gene-to-pathway

expression similarity. The RF models allowed us to ask what expression features contributed the most to predicting memberships in different pathways. By examining the feature importance in RF Set B models, we found that for a gene in pathway P, its expression similarity to other genes in pathway P (gene-to-target-pathway) tend to be more important for predicting membership in pathway P than its similarity to genes in other, non-P pathways (gene-to-other-pathways) (**Fig. S7**). However, expression similarities between genes in one pathway and genes in the other pathways are also required for the predictions. This is supported by the finding that in 81 out of 85 pathways, the most important features were not the gene-to-target-pathway features. This result explains why naive prediction models based solely on gene-to-target-pathway expression correlation performed poorly (**Fig. 2b**).

We also examined five other supervised learning algorithms that support multi-class classification: Support Vector Classification (SVC), *k*-nearest neighbors (KNN), Adaptive Boosting (AB) and two neural network approaches (NN_1, NN_2, see **Methods**). KNN, and AB models had similar performances as RF models. However, SVC models and the two neural network algorithms performed worse (**Fig. 4c**), which in the latter case may be due to the relatively larger sample sizes required for training neural network models (Liu *et al.*, 2017). We next examined the performance of supervised learning algorithms in predicting membership in individual pathways and found that the predictions were the best in 41, 33, 24, 20, 3, and 1 pathways (some pathways have the best F1 for >1 algorithms, so total is larger than the number of pathways) when using KNN, RF, AB, SVC, NN_1, and NN_2, respectively. This finding indicates that the choice of supervised learning algorithms can impact predictions of specific pathways.

Additional factors influencing pathway membership predictions

Thus far, metabolic pathway membership prediction has been treated as a multi-class classification task (i.e., one prediction—whether a gene belongs to one of 85 pathways) rather than multiple binary classification (i.e., 85 predictions—genes in a pathway vs. those in the other pathways). This is due to the relatively small sizes of most pathways (only 35 pathways have ≥ 10 genes). To assess the impact of classification scheme (multi vs. binary), we used the LinearSVC approach, which supports both “crammer_singer” (one multi-class classification) and “ovr” (multiple one-vs-rest binary classifications) strategies. Better performances were obtained using

“crammer_singer” for 46 (56%) Set A and 276 (42%) Set B data, indicating that the classification scheme may also impact predictions.

In addition, we also explored the effect of expression level estimation measures (FPKM vs. TPM) and normalization of feature data on predictions. We found that FPKM and TPM values were highly correlated (median PCC= 0.95, **Fig. S8a**). Although the performances of *k*-means and balanced RF models using FPKM and TPM values were statistically indistinguishable ($p=0.58$ and 0.49 , respectively, Wilcoxon signed-rank test, **Fig. S8b-e**), improved performances for 38 and 36 pathways were obtained using FPKM values in *k*-means and RF models, respectively, while increased performances for 41 and 42 pathways were obtained using TPM for *k*-means and RF models, respectively (**Fig. S8b-e**). Because some distance metrics used in unsupervised learning (e.g., Euclidean distance used in *k*-means) may be sensitive to different value scales, feature standardization can affect model performance. The original and standardized FPKM-based *k*-means models had statistically indistinguishable performances ($p=0.67$, Wilcoxon signed-rank test, **Fig. S8f**). However, pathway best F1s were improved for some pathways (**Fig. S8g**), indicating that expression value scales affect pathway membership prediction, although normalized values are not always better.

Thus far, model prediction performance has been based on cross-validation. Given that most pathway memberships in tomato are annotated computationally, we examined the prediction of 36 benchmark genes (among 942 annotated genes in the training set) that have been experimentally validated (**Table S19**) using 85 pathway best models. Here, the pathway best model is the model with highest pathway $F1_{CV}$ among *k*-means/RF and data Set A/B-based models. Only *k*-means and RF models are discussed here, as other unsupervised and supervised algorithms had similar or poorer performances. Among these 36 benchmark genes, 21 (58.3%) were correctly predicted, which is substantially better than the random guess accuracy of 1.2% (1/85). We should also emphasize that the benchmark accuracy is significantly better than 37.0%, the accuracy for predicting the remaining 906 annotated genes that do not have experimental data support ($p=0.01$, Fisher’s exact test) indicating that annotation quality affects predictions. In addition, by further assessing the cause of mis-predictions and excluding eight pathways with an unusually high number of genes annotated for a single reaction, significantly improved performance was obtained (**Fig. S9, Table S20**),

demonstrating the impact of annotation quality on predictions. Thus, our current models can be regarded as “baseline” predictions that can be further improved as pathway annotation improves.

Optimizing pathway predictions by identifying optimal data and method combinations

To provide optimal predictions for different pathways, we summarized the pathway best predictions when different data combinations were used. Because the naive prediction models performed poorly (**Fig. 3c and Fig. 4c**), they are not discussed here. The pathway optimal F1 values ranged from 0.19 to 0.73 (median=0.4 (**Fig. 6**)). An F1=0.4 may seem low at the first glance, but it is much higher compared with the background median F1=0.01 (achieved by randomly guessing, **Table S14**). To recover 10 genes for a pathway with 20 members with F1=0.4, 30 predictions would need to be made where 10 (33%) predictions are true positives and 20 (67%) are false positives. But for F1=0.01, 2000 predictions would need to be made with 0.5% being true positives and 99.5% being false positives. Thus, while there is room for improvement, the optimal predictions are substantially better than background and facilitate hypothesis development for experimental testing.

We next asked whether the optimal prediction tends to be achieved when particular data combinations are used. The “condition-independent” dataset (including all experiments) provided optimal predictions for seven pathways (**Fig. S10, Table S21**), which are involved in general cellular processes, including the thioredoxin pathway, which is important for maintaining cellular redox status. In contrast, the remaining 78 pathways (91.8%) were better predicted using “condition-dependent” datasets (**Fig. S10**), echoing our findings on the impact of dataset on expression correlation (**Fig. 1b,c**). Pathway membership tended to be better predicted when datasets for related biological processes were used. For example, the galactolipid biosynthesis I pathway (PWY-401), had the highest F1s when samples with ABA treatment were used (**Table S21**), consistent with the finding that *Arabidopsis* PWY-401 genes respond to phosphate starvation in an ABA-dependent manner (Woo *et al.*, 2012). Similarly, the optimal F1 for the trichome monoterpenes biosynthesis pathway (PWY-6447) was obtained when a pathogen treatment dataset was used (**Table S21**), supporting a role of monoterpenes in the defense against pathogens (Lackus *et al.*, 2018). These results not only illustrate the importance of the data set used for prediction, but also highlight the possibility of using

the prediction framework described here to identify connections between pathways and biological processes.

Discussion

Gene co-expression analysis is widely used to associate genes with specific functions. We assessed the utility of transcriptome data for predicting metabolic pathway membership by considering 82 expression values, 656 gene-to-pathway expression similarity data combinations, and three prediction strategies (naive prediction, unsupervised and supervised learning). We demonstrated that the optimal data combination and prediction strategy should be identified for each pathway. Among the 85 pathways examined, 90 different data combinations (≥ 1 data combination may lead to an optimal prediction for a pathway) led to optimal membership predictions. Our examples demonstrate that optimal pathway membership predictions tend to be achieved when pathway function-associated datasets are used, although in many cases, it is not obvious why a data combination is optimal for a pathway.

We show that machine learning approaches out-perform naive methods. The unsupervised learning approach tended to outperform supervised learning when the pathway size was small, likely because there were insufficient data for supervised learning to be effective. However, 39 of 85 pathway best predictions were made by supervised learning approach. The substantial number of pathway best predictions made by supervised approach indicates the importance of exploring both data combinations and prediction approaches. It will be helpful to dissect the prediction models using model interpretation methods (Azodi *et al.*, 2020) to further identify which data features are particularly important for predicting pathway membership; this will facilitate strategies for identifying optimal data combinations for modelling.

While the prediction performance was better than random guessing, the error rates of some pathways were relatively high no matter which data or algorithms were examined (**Fig. 6**). There are three potential reasons for this. First, a prediction model relies on the quality of the input. However, the current pathway annotation in tomato mainly relies on sequence similarity to genes in other model species. The composition of metabolites varies across species due to repeated metabolic innovation via gene duplication and subsequent sub- or neo-functionalization (Pichersky & Gang, 2000),

recruitment of genes to new pathways (Shoji & Hashimoto, 2011), and loss of pathway genes (Cutter & Jovelin, 2015; Baggs *et al.*, 2020). Thus, membership of genes in lineage-specific pathways may not be readily inferred using information from model organisms based on sequence similarity. In addition, we found that model accuracy for the experimentally validated gene set was higher than for those without such validation. Thus, as annotation improves, the accuracy of predictions is expected to increase.

Second, we removed genes with multiple pathway annotations to simplify pathway assignment, further reducing the sizes of training data sets. To overcome this issue, multilabel learning approaches (Herrera *et al.*, 2016) can be used in which multiple pathways can be assigned to a gene. Third, additional features may be needed to improve the pathway predictions. For example, pathways can have interwoven reactions, and pathway genes can have negatively correlated expression profiles (Zeng & Li, 2010). Thus, the nature of reactions that enzymes catalyze could be used to hypothesize interactions and be incorporated as features. In addition, only transcriptome-based features were used in this study. Considering that enzymes in the same pathway may be located in tandem clusters (Field *et al.*, 2011) and interact genetically and/or physically (Gao *et al.*, 2012; Weissenborn & Walther, 2017), clustering and interaction data could be informative.

Another consideration is that our models can only predict genes in these 85 pathways. In the future, an additional, “other” class could be included in supervised learning models so a gene belonging to another pathway would not be forced into one of the 85 pathways. To assign genes into the “other” class, one approach is by holding out one or more known pathways and designate them as “other”, which is conceptually the same as the multi-class framework used here. For unsupervised learning models, the “other” class is baked into the approach because some clusters may not be associated with any pathway, and genes in these clusters are not predicted to belong to any pathway.

Together, this study provides quantitative measures of the usefulness of expression data in predicting metabolic pathway memberships, and lays the foundation for further method comparison studies that seek to improve the use of expression data for similar purposes. Although this prediction exercise focused on annotated enzyme genes, the *k*-means and RF models can be applied to unknown genes and provide pathway membership predictions with estimated likelihood scores. Most importantly, our study underscores the feasibility and limitations of solely using transcriptome data

for predicting metabolic pathway membership. The exploration of methods and data subsets in this study provides a baseline for future modeling efforts and highlights the need for further exploration, particularly of the causes of mis-predictions, for improving future models.

Figure legends

Fig. 1. Impact of expression dataset, expression values, and similarity measures on expression similarities between genes in the same pathway in *Solanum lycopersicum*. **(a)** Gene expression similarity (in terms of Pearson correlation coefficient, PCC) within and between pathways. Boxplot showing expression similarity between genes within individual pathways. Blue line: median value; light blue box: interquartile range. Gray violin plot shows the PCC distribution of all gene pairs between different pathways (between-pathway distribution), where median value and interquartile range are marked with red and orange dashed lines, respectively. Circled numbers indicate example pathways in **(b)**. **(b)** Examples showing the effect of dataset on gene expression similarity. Fragments Per Kilobase Million (FPKM) was used to calculate the PCCs. X-axis: 41 datasets, y-axis: example pathways. Color scale: percentile of the median PCC within a pathway in the between-pathway distribution (percentile_{BP}). The percentile_{BP} values were scaled to 0–1 here and hereafter, where 1 indicates the 100th percentile. Pathway names in green are those relevant to photosynthesis. **(c)** Same as **(b)**, except that fold change (FC) was used in PCC calculation. **(d)** Scatter plot showing the differences between percentile_{BP} of median PCC calculated using FPKM (x-axis) and FC (y-axis). **(e)** Gene expression similarity calculated using different similarity measures. Color scale: percentile_{BP} of median expression similarity calculated using different similarity measures (left) and $1 - \text{percentile}_{BP}$ of the mutual rank of similarity values (right). For mutual rank, $1 - \text{percentile}_{BP}$ was used because lower ranks (lower percentiles) indicate higher degrees of expression correlation. Yellow and cyan rectangles indicate pathways with high (red) and median high (light red) expression similarities, respectively, only when a specific similarity measure was used.

Fig. 2. Naive prediction of metabolic pathway genes in *Solanum lycopersicum*. **(a)** Methodology. For each of the 82 (41 expression datasets \times 2 expression values) expression matrices, five genes (blue)

were randomly held out from each pathway containing ≥ 25 genes. The remaining data (gray) were used to determine expression similarities among genes using 8 similarity measures, resulting in 656 expression similarity matrices (m1–m656). Genes within each pathway were further split into “training” (80%, gray) and “validation” (20%, magenta) subsets, and the data splitting was conducted five times. The example validation gene 1 is predicted to be in pathway A using the naive median method because it has the highest median expression similarity with training genes in pathway A; however, it is predicted to be in pathway B using the naive maximum method because it has the maximum expression similarity with gene 3, which belongs to pathway B. The thickness of the arrow and number beside the arrow indicate the degree of expression similarity. All 656 expression similarity matrices were used for both methods, and the $F1_{CV}$ score was calculated for each of the 85 pathways, resulting in two 656×85 F1 score matrices. The prediction with the highest $F1_{CV}$ for a pathway (purple) was referred to as the pathway best prediction. The average $F1_{CV}$ across 85 pathways for each naive model (made using one of the 656 matrices) was calculated to measure the overall prediction performance, and the model with the highest average F1 (red) was referred to as the overall best model (dashed box). **(b)** Distribution of pathway best $F1_{CV}$ obtained using the naive median (orange) and naive maximum (pink) methods. Distribution of background pathway F1 scores is shown as a gray violin plot, with the median value indicated by a dashed line. ***: p -value < 0.001 , Wilcoxon signed-rank test. **(c,d)** Performances of naive median **(c)** and naive maximum **(d)** models. X-axis: $F1_{CV}$ values for the overall best model—values in the dashed box in the table on the right in **(a)**. Y-axis: pathway best $F1_{CV}$ across all 656 models—values in the purple cells in the table on the right in **(a)**. **(e)** Comparison of pathway best $F1_{CV}$ for naive median (x-axis) and naive maximum (y-axis) predictions. Three example pathways in **(c-e)** show the differences in pathway predictions made by the two approaches and are shown in the table on the right in **(a)** as Pathway A (CHLOROPHYLL-SYN, 3,8-divinyl-chlorophyllide a biosynthesis I), B (PWY-841, superpathway of purine nucleotides de novo biosynthesis I), and C (PWY-6773, 1,3- β -D-glucan biosynthesis).

Fig. 3. Prediction of pathway genes using unsupervised clustering methods in *Solanum lycopersicum*. **(a)** Workflow for clustering using gene expression profiles in the form of 82 matrices (41 expression datasets \times 2 expression values, Set A). Data splitting was conducted in the same way as for naive

prediction approaches (*Split data and five-fold cross-validation steps*). The pathway annotations (colored circles, i.e., Pathways A, B, and C) were removed from the expression matrices before clustering. Clusters (blue closed lines, C1–C4) built using expression values of “training” genes (*Clustering without validation data*) were assigned to pathways based on enrichment analysis (*Assign clusters to pathways*, see **Methods**). Then the validation genes were assigned to clusters (blue arrow, see **Methods**), and then were assigned to the pathways the clusters were assigned to (*Assign validation genes to clusters*). Red arrow indicates mis-prediction of a gene. The average F1 scores of validation genes across the five training/validation splits ($F1_{CV}$) were calculated to evaluate the predictions. The test genes were assigned to pathways using the same method used for the validation genes, and the average F1 of test genes from five clustering models (F1 based on holdout data, $F1_{test}$) were calculated. **(b)** Similar to **(a)**, except that clusters were built using gene-to-pathway co-expression matrices (Set B), i.e., the maximum and median expression similarity of a training gene to all other training genes in each pathway, which was calculated using eight different similarity measures, resulting in 656 (82×8) matrices and clustering models. For a validation or test gene, the expression similarity was calculated between the gene and the training genes in each pathway, and the maximum and median values were used in the matrix. **(c)** Distribution of pathway best $F1_{CV}$ obtained from clustering models, performed using the *k*-means (KM), Affinity Propagation (AP), Birch, or MeanShift (MS) method, and the expression matrix (Set A, light green) or gene-pathway co-expression matrix (Set B, green). Orange, pink and gray dashed lines indicate the median pathway best $F1_{CV}$ from naive median and naive maximum prediction models, and the median background pathway F1 (as in **Fig. 2b**), respectively. **: p -value < 0.01, ***: p -value < 0.001, Wilcoxon signed-rank test. **(d)** Comparison of pathway best $F1_{CV}$ for *k*-means and MeanShift clustering models, using Set A (left panel) or Set B (right panel). Dots: individual pathways. Two examples showing performance differences between two clustering methods: CHLOROPHYLL-SYN (3,8-divinyl-chlorophyllide a biosynthesis I) and PWY-6823 (molybdenum cofactor biosynthesis). p -values are from Wilcoxon signed-rank test. **(e)** Comparison of pathway best $F1_{CV}$ for *k*-means models using Set A or Set B data.

Fig. 4. Prediction of pathway genes using supervised machine learning methods in *Solanum lycopersicum*. **(a)** Workflow for supervised machine learning using Set A as features. Gene pathway membership information (i.e., labels) was merged with the expression matrix. Data splitting was conducted in the same way as for naive and unsupervised approaches (*Split data* and *five-fold cross-validation*) in **Fig. 2, 3**. A grid search was conducted to get the best combination of parameters (hyperparameters) that yielded the maximum $F1_{CV}$ (see **Methods**). The final model was built using the same cross-validation scheme with the hyperparameters and was applied to validation and test genes (*Apply*). **(b)** Same as **(a)**, except that gene-to-pathway expression similarity matrices (Set B) were used as features. **(c)** Distribution of pathway best $F1_{CV}$ from unbalanced RandomForest (RF), balanced RF, Support Vector Classification (SVC), k -nearest neighbors (KNN), Adaptive Boosting (AB), and two neural network (NN_1, NN_2, see **Methods**) models using Set A (light blue) or Set B (blue). Light green and green dashed line: median pathway best $F1_{CV}$ from k -means models using Set A and Set B, respectively (as in **Fig. 3c**); orange and pink dashed line: median pathway best $F1_{CV}$ from naive median and naive maximum prediction models, respectively (as in **Fig. 2b**); gray dashed line: median background pathway $F1$. **(d)** Comparison of pathway best $F1_{CV}$ from balanced RF models when Set A and Set B were used. Dots: individual pathways. p -value is from Wilcoxon signed-rank test.

Fig. 5. Performance difference between k -means clustering and Random Forest models in *Solanum lycopersicum* when Set B was used. **(a-c)** Confusion matrix, which shows the proportion of genes that are predicted in each pathway for pathway best k -means models **(a)**, unbalanced RF models **(b)** and balanced RF models **(c)**. Color scale: proportion of genes in a pathway (y-axis) predicted as being in one of the 85 pathways (x-axis) by the pathway best model. **(d-f)** Comparison of pathway best $F1_{CV}$ between k -means, unbalanced RF and balanced RF models when Set B was used. p -value is from Wilcoxon signed-rank test. **(g-h)** Pathway $F1_{CV}$ from k -means (x-axis) and balanced RF models (y-axis) when the pathway best k -means Set B data **(g)** or the pathway best balanced RF Set B data **(h)** was used. Dots: individual pathways.

Fig. 6. Summary of pathway membership prediction using unsupervised and supervised models in *Solanum lycopersicum*. Pathway best $F1_{CV}$ in k -means models using Set A (orange) and Set B (pink), in balanced RF models using Set A (blue) and Set B (purple), and the background pathway $F1$ (black). The largest six pathways with test genes are marked using different colored fonts. $F1_{test}$ values of these six pathways in the pathway best models are shown in Supporting Information **Fig. S4b** and **Table S22**.

Acknowledgments

We thank John P. Lloyd, Christina B. Azodi, Serena G. Lotreck, Elizabeth M. Gibbons, Koichi Sugimoto, Yann-Ru Lou, Bryan Leong, Brian St. Aubin, Pengxiang Fan, Paul Fiesel, Craig Schenck, Robert Last, and Eran Pichersky for helpful discussions. We also thank the reviewers on insights and suggestions that led to further improvement of the study. This work was partly supported by the National Science Foundation IOS-1546617 to CSB. and SHS and NSF DEB-1655386 and U.S. Department of Energy Great Lakes Bioenergy Research Center BER DE-SC0018409 to SHS. CSB is supported in part by Michigan AgBioResearch and through USDA National Institute of Food and Agriculture Hatch project number MICL02552.

Author contributions

PW and SHS conceived and designed the study. PW, BMM, SU and CSB performed the analysis. PW, MDL, CSB and SHS wrote the paper. All authors read and approved the final manuscript.

Data availability

Gene expression data are available on Zenodo at: <https://zenodo.org/record/4585635#.YEJ2zshvrrM>.

All the scripts used in this study are available on Github at:

https://github.com/ShiuLab/Pathway_gene_prediction_in_tomato.

References

- Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**:166-169.
- Aoki Y, Okamura Y, Ohta H, Kinoshita K, Obayashi T. 2016. ALCOdb: gene coexpression database for microalgae. *Plant and Cell Physiology* **57**:e3.
- Azodi CB, Tang J, Shiu SH. 2020. Opening the black box: interpretable machine learning for geneticists. *Trends in Genetics* **36**:442-455.
- Baggs EL, Monroe JG, Thanki AS, O' Grady R, Schudoma C, Haerty W, Krasileva KV. 2020. Convergent loss of an EDS1/PAD4 signaling pathway in several plant lineages reveals coevolved components of plant immunity and drought response. *Plant Cell* **32**:2158-2177.
- Berg JM, Tymoczko JL, Stryer L. 2002 *Biochemistry, Fifth Edition: International Version*. New York, USA: W. H. Freeman.
- Blagus R, Lusa L. 2013. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* **14**:106.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114-2120.
- Boulesteix A-L, Janitza S, Kruppa J, König IR. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Mining and Knowledge Discovery* **2**:493-507.
- Carmona M, Zalacain A, Sanchez AM, Novella JL, Alonso GL. 2006. Crocetin esters, picrocrocin and its related compounds present in *Crocus sativus* stigmas and *Gardenia jasminoides* fruits. Tentative identification of seven new compounds by LC-ESI-MS. *Journal of Agricultural and Food Chemistry* **54**:973-979.
- Chae L, Kim T, Nilo-Poyanco R, Rhee SY. 2014. Genomic signatures of specialized metabolism in plants. *Science* **344**:510-513.
- Cutter AD, Jovelín R. 2015. When natural selection gives gene function the cold shoulder. *Bioessays* **37**:1169-1173.

- De Leon IP, Sanz A, Hamberg M, Castresana C. 2002.** Involvement of the *Arabidopsis* alpha-DOX1 fatty acid dioxygenase in protection against oxidative stress and cell death. *Plant Journal* **29**:61-62.
- De Luca V, Salim V, Atsumi SM, Yu F. 2012.** Mining the biodiversity of plants: a revolution in the making. *Science* **336**:1658-1661.
- Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE. 2011.** Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proceedings of the National Academy of Sciences of the United States of America* **108**:16116-16121.
- Forster J, Famili I, Fu P, Palsson BO, Nielsen J. 2003.** Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research* **13**:244-253.
- Fujisawa M, Shima Y, Nakagawa H, Kitagawa M, Kimbara J, Nakano T, Kasumi T, Ito Y. 2014.** Transcriptional regulation of fruit ripening by tomato FRUITFULL homologs and associated MADS box proteins. *Plant Cell* **26**:89-101.
- Gao YF, Chen L, Cai YD, Feng KY, Huang T, Jiang Y. 2012.** Predicting metabolic pathways of small molecules and enzymes based on interaction information of chemicals and proteins. *PLoS One* **7**:e45944.
- Gupta C, Pereira A. 2019.** Recent advances in gene function prediction using context-specific coexpression networks in plants. *F1000Research* **8**:F1000 Faculty Rev-1153.
- Herrera F, Charfe F, Rivera AJ, Jesus MJd. 2016** *Multilabel classification. Problem analysis, metrics and techniques*. Switzerland: Springer International Publishing.
- Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K *et al.* 2007.** Omics-based identification of *Arabidopsis* MYB transcription factors regulating aliphatic glucosinolate biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* **104**:6478-6483.
- Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003.** Global analysis of protein localization in budding yeast. *Nature* **425**:686-691.
- Karp PD, Latendresse M, Caspi R. 2011.** The pathway tools pathway prediction algorithm. *Standards in Genomic Sciences* **5**:424-429.

- Kaundal R, Saini R, Zhao PX. 2010.** Combining machine learning and homology-based approaches to accurately predict subcellular localization in *Arabidopsis*. *Plant Physiology* **154**:36-54.
- Kent WJ. 2002.** BLAT--the BLAST-like alignment tool. *Genome Research* **12**:656-664.
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013.** TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**:R36.
- Kim J, Buell CR. 2015.** A revolution in plant metabolism: genome-enabled pathway discovery. *Plant Physiology* **169**:1532-1539.
- Lackus ND, Lackner S, Gershenzon J, Unsicker SB, Kollner TG. 2018.** The occurrence and formation of monoterpenes in herbivore-damaged poplar roots. *Scientific Reports* **8**:17936.
- Liu B, Wei Y, Zhang Y, Yang Q. 2017.** Deep Neural Networks for High Dimension, Low Sample Size Data. *Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne, Australia. International Joint Conferences on Artificial Intelligence.
- Lloyd JP, Seddon AE, Moghe GD, Simenc MC, 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.
- Majdi M, Liu Q, Karimzadeh G, Malboobi MA, Beekwilder J, Cankar K, Vos Rd, Todorović S, Simonović A, Bouwmeester H. 2011.** Biosynthesis and localization of parthenolide in glandular trichomes of feverfew (*Tanacetum parthenium* L. Schulz Bip.). *Phytochemistry* **72**:14-15.
- Mao L, Kawaide H, Higuchi T, Chen M, Miyamoto K, Hirata Y, Kimura H, Miyazaki S, Teruya M, Fujiwara K et al. 2020.** Genomic evidence for convergent evolution of gene clusters for momilactone biosynthesis in land plants. *Proceedings of the National Academy of Sciences of the United States of America* **117**:12472-12480.
- Martin C, Li J. 2017.** Medicine is not health care, food is health care: plant metabolic engineering, diet and human health. *New Phytologist* **216**:699-719.
- Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, Bruijn Id, Chooi YH, Claesen J, Coates RC et al. 2015.** Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology* **11**:625-631.

- 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 of the United States of America* **116**:2344-2353.
- Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S. 2011.** PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* **23**:895-910.
- Ni Y, Aghamirzaie D, Elmarakeby H, Collakova E, Li S, Grene R, Heath LS. 2016.** A machine learning approach to predict gene regulatory networks in seed development in *Arabidopsis*. *Frontiers in Plant Science* **7**:1936.
- Niggeweg R, Michael AJ, Martin C. 2004.** Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nature Biotechnology* **22**:746-754.
- Nouri M-Z, Moumeni A, Komatsu S. 2015.** Abiotic stresses: insight into gene regulation and protein expression in photosynthetic pathways of plants. *International Journal of Molecular Sciences* **16**:20392–20416.
- Nutzmann HW, Huang A, Osbourn A. 2016.** Plant metabolic clusters - from genetics to genomics. *New Phytologist* **211**:771-789.
- Obayashi T, Aoki Y, Tadaka S, Kagaya Y, 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**:440.
- Osbourn A. 2010.** Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends in Genetics* **26**:449-457.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V *et al.* 2011.** Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**:2825-2830.
- Pichersky E, Gang DR. 2000.** Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends in Plant Science* **5**:439-445.
- Qi Y. 2012** *Random Forest for Bioinformatics*. In: Zhang C., Ma Y. (eds) *Ensemble Machine Learning*. Boston, MA: Springer.

- Righetti K, Vu JL, Pelletier S, Vu BL, Glaab E, Lalanne D, Pasha A, Patel RV, Provart NJ, Verdier J *et al.* 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.
- Robinson MD, McCarthy DJ, Smyth GK. 2010.** edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139-140.
- Schafer J, Strimmer K. 2005.** A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**:Article32.
- Schlapfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T *et al.* 2017.** Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology* **173**:2041-2059.
- Segal E, Wang H, Koller D. 2003.** Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19 Suppl 1**:i264-271.
- Shoji T, Hashimoto T. 2011.** Recruitment of a duplicated primary metabolism gene into the nicotine biosynthesis regulon in tobacco. *Plant Journal* **67**:949-959.
- Stitt M, Sulpice R, Keurentjes J. 2010.** Metabolic networks: how to identify key components in the regulation of metabolism and growth. *Plant Physiology* **152**:428-444.
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010.** Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**:511-515.
- Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ. 2009.** Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell & Environment* **32**:1633-1651.
- Uygun S, Peng C, Lehti-Shiu MD, Last RL, Shiu SH. 2016.** Utility and limitations of using gene expression data to identify functional associations. *PLoS Computational Biology* **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, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J *et al.* 2020.** SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**:261-272.
- Weissenborn S, Walther D. 2017.** Metabolic pathway assignment of plant genes based on phylogenetic profiling - a feasibility study. *Frontiers in Plant Science* **8**:1831.
- Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, 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 CR, Liu J, Wang H, Kiba T, Hannah MA, Wang XJ, Bajic VB, Chua NH. 2012.** The response and recovery of the *Arabidopsis thaliana* transcriptome to phosphate starvation. *BMC Plant Biol.* **12**:62.
- Zeng T, Li J. 2010.** Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways. *Nucleic Acids Research* **38**:e1.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1. Differences in gene expression similarity within pathways and between pathways

Fig. S2. Impact of similarity measure on expression similarity

Fig. S3. Confusion matrix for naive prediction models

Fig. S4. Pathway membership prediction using unsupervised and supervised models

Fig. S5. Potential underlying reasons for the relatively poorer performances of Set A and RF models than Set B and *k*-means models, respectively

Fig. S6. Pathway $F1_{CV}$ values from *k*-means models and balanced RF models

Fig. S7. Important features for predictions in RF Set B models

Fig. S8. Effects of expression values on pathway membership prediction

Fig. S9. Potential reasons for low pathway F1

Fig. S10. Pathway optimal expression datasets

Table S1. Correspondence between gene annotations in SGN_V3.2 and NCBI_V2.5

Table S2. Pathway gene annotation

Table S3. Number of genes annotated in a pathway after genes with multiple pathway annotations were filtered out

Table S4. Expression data information

Table S5. Expression experiment information

Table S6. Parameters for software in RNA-seq data processing, and hyperparameter space for unsupervised and supervised algorithms

Table S7. Median and maximum PCC values for correlation between a gene to all other genes in each pathway -- calculated using FPKM from all experiments

Table S8. Percentile of pathway PCC in between-pathway distribution -- calculated using FPKM from all experiments

Table S9. Percentile of pathway median PCC values in between-pathway distribution calculated using FPKM

Table S10. Percentile of pathway median PCC values in between-pathway distribution calculated using FC

Table S11. Percentile of pathway median correlation values in between-pathway distribution calculated using FPKM with different similarity measures

Table S12. Pathway $F1_{CV}$ values for the naive median method

Table S13. Pathway $F1_{CV}$ values for the naive maximum method

Table S14. Pathway best $F1_{CV}$ from different approaches

Table S15. Confusion matrix in pathway best k-means Set B clustering

Table S16. Confusion matrix in pathway best unbalanced RF Set B models

Table S17. Confusion matrix in pathway best balanced RF Set B models

Table S18. Pathway best clustering/models

Table S19. Prediction of genes in pathway best models

Table S20. Improvement of prediction by excluding pathways with the highest number of genes in a single reaction

Table S21. Factors impacting clustering/model performance

Table S22. Pathway $F1_{CV}$ and $F1_{test}$ data in pathway best model with highest pathway $F1_{CV}$

Figure 1

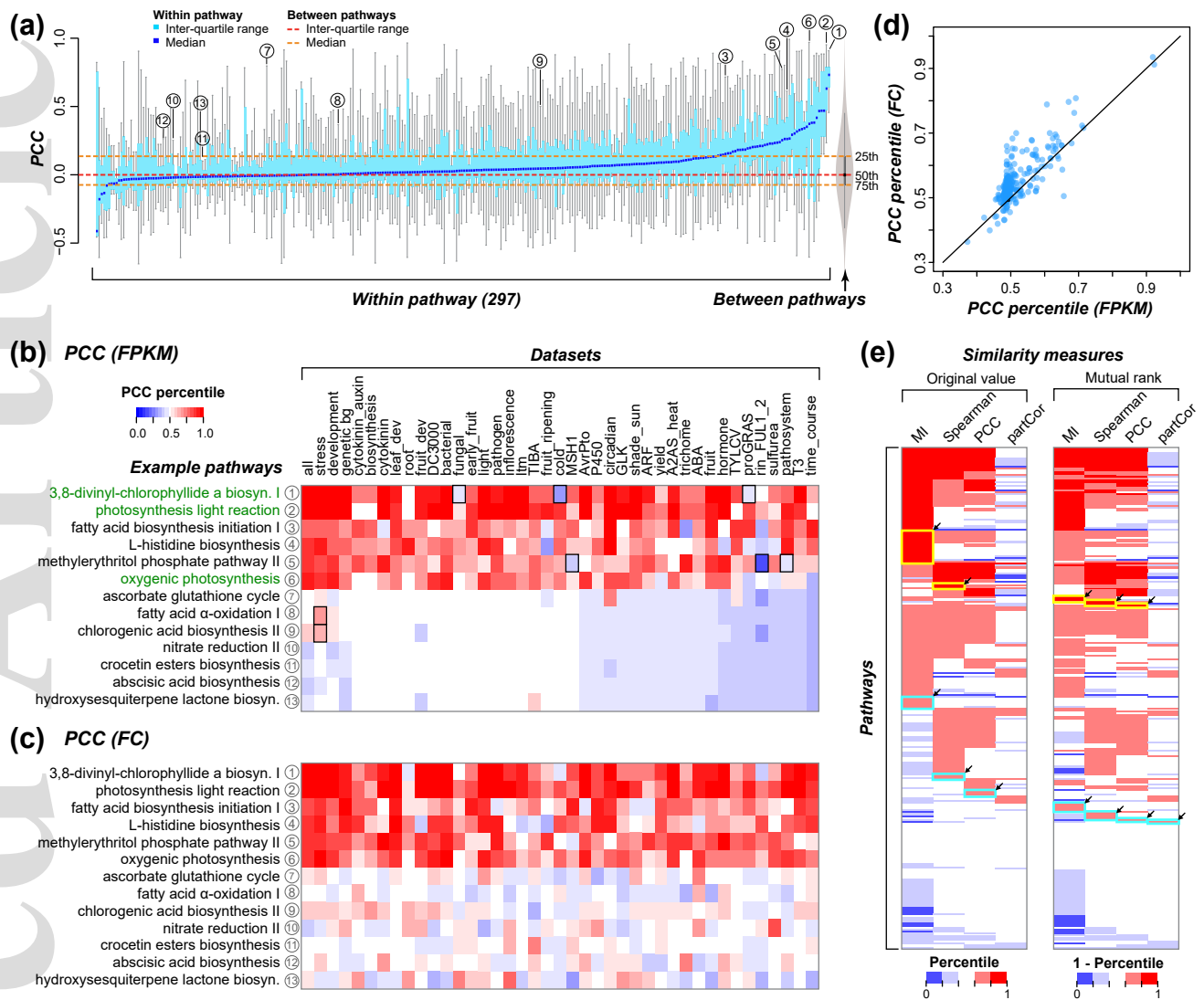


Figure 2

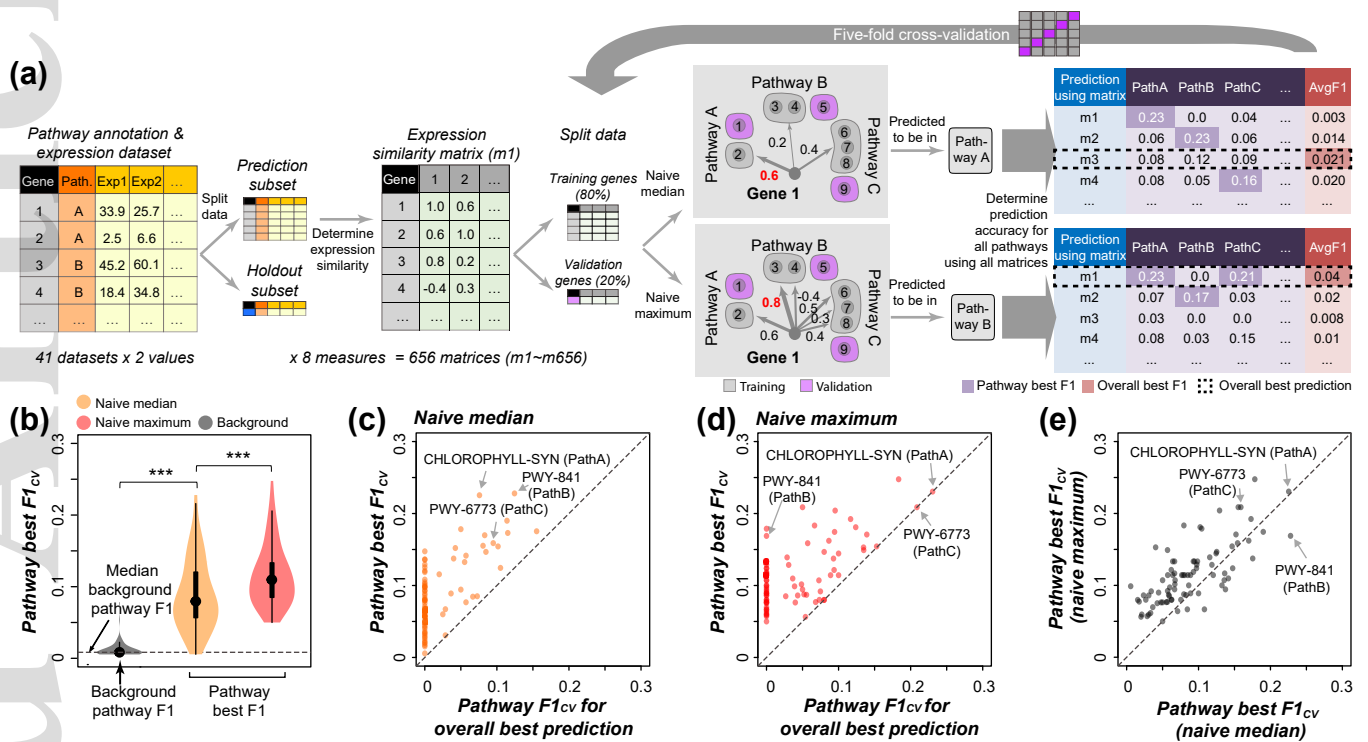


Figure 3

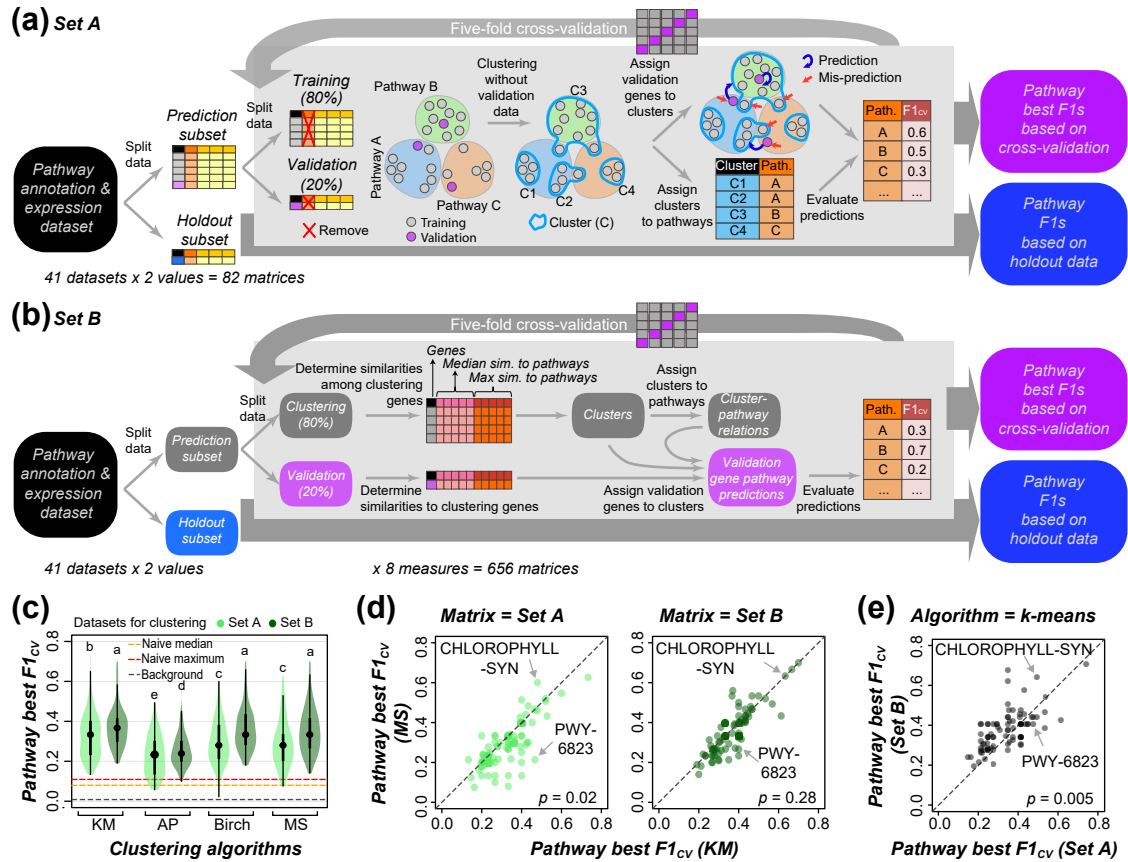


Figure 4

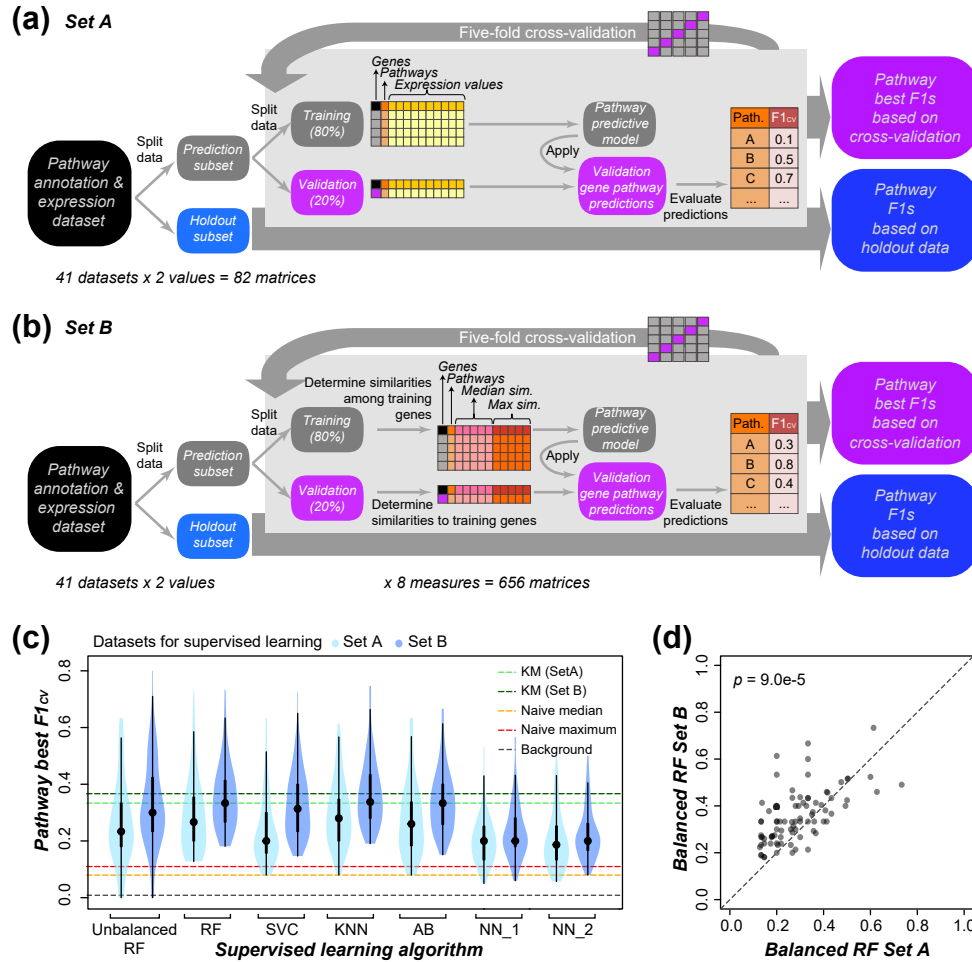


Figure 5

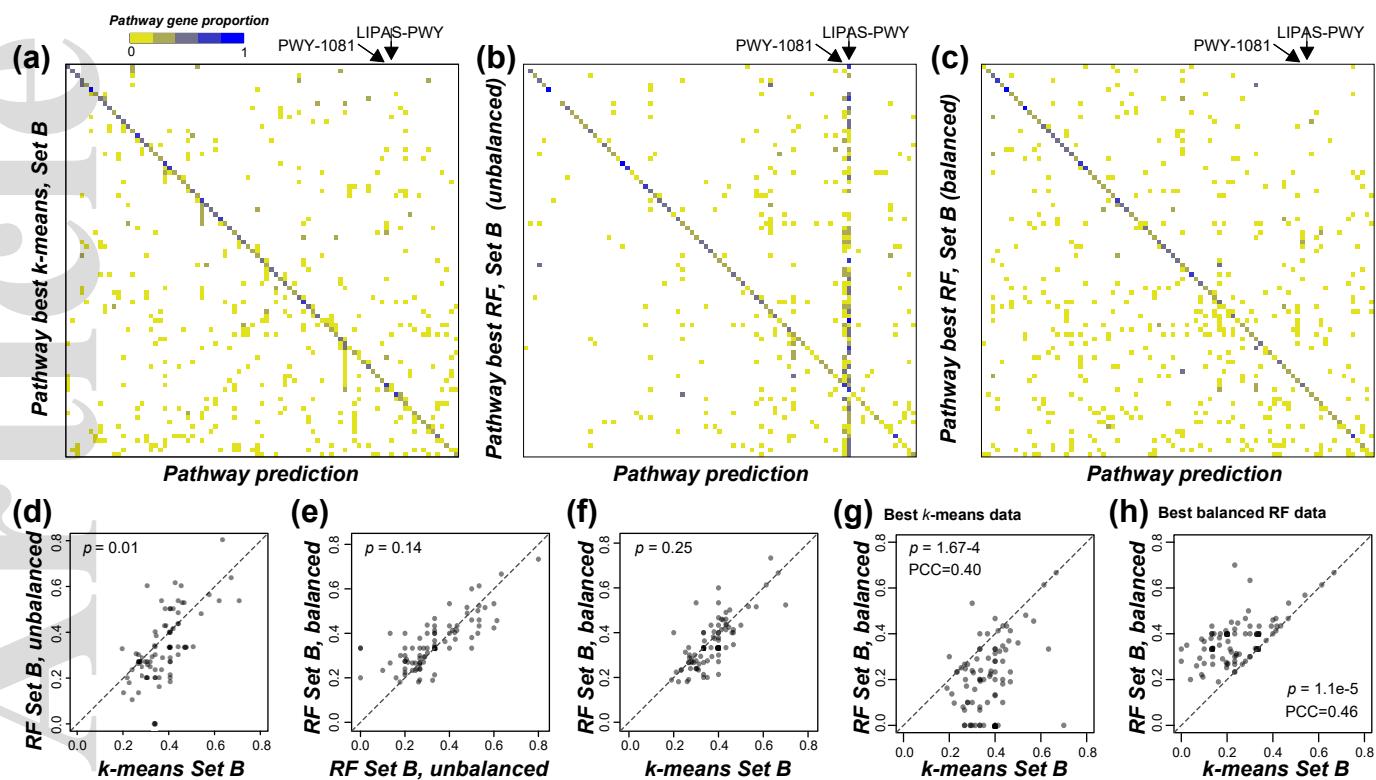


Figure 6

