in silico Plants Vol. 2, No. 1, pp. 1–19
doi:10.1093/insilicoplants/diaa005
Advance Access publication 30 July 2020.
Original Research

Within- and cross-species predictions of plant specialized metabolism genes using transfer learning

Bethany M. Moore[1,2,9], Peipei Wang[1], Pengxiang Fan[3], Aaron Lee[4], Bryan Leong[1], Yann-Ru Lou[3], Craig A. Schenck[3], Koichi Sugimoto[5,6], Robert Last[1,3], Melissa D. Lehti-Shiu[1], Cornelius S. Barry[7] and Shin-Han Shiu[1,2,8*]

[1]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA
[2]Ecology, Evolutionary Biology, and Behavior Program, Michigan State University, East Lansing, MI 48824, USA
[3]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA
[4]Department of Biology, The College of New Jersey, Ewing, NJ 08628, USA
[5]MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, MI 48824, USA
[6]Science Research Center, Yamaguchi University, Yamaguchi 753-8515, Japan
[7]Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA
[8]Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI 48824
[9]Present address: Department of Botany, University of Wisconsin-Madison, Madison, WI, USA
*Corresponding author's e-mail address: shius@msu.edu

Handling Editor: Amy Marshall-Colon

Citation: Moore BM, Wang P, Fan P, Lee A, Leong B, Lou Y-R, Schenck CA, Sugimoto K, Last R, Lehti-Shiu MD, Barry CS, Shiu S-H. 2020. Within- and cross-species predictions of plant specialized metabolism genes using transfer learning. *In Silico Plants* **2020**: diaa005; doi: 10.1093/insilicoplants/diaa005S

ABSTRACT

Plant specialized metabolites mediate interactions between plants and the environment and have significant agronomical/pharmaceutical value. Most genes involved in specialized metabolism (SM) are unknown because of the large number of metabolites and the challenge in differentiating SM genes from general metabolism (GM) genes. Plant models like *Arabidopsis thaliana* have extensive, experimentally derived annotations, whereas many non-model species do not. Here we employed a machine learning strategy, transfer learning, where knowledge from *A. thaliana* is transferred to predict gene functions in cultivated tomato with fewer experimentally annotated genes. The first tomato SM/GM prediction model using only tomato data performs well (*F*-measure = 0.74, compared with 0.5 for random and 1.0 for perfect predictions), but from manually curating 88 SM/GM genes, we found many mis-predicted entries were likely mis-annotated. When the SM/GM prediction models built with *A. thaliana* data were used to filter out genes where the *A. thaliana*-based model predictions disagreed with tomato annotations, the new tomato model trained with filtered data improved significantly (*F*-measure = 0.92). Our study demonstrates that SM/GM genes can be better predicted by leveraging cross-species information. Additionally, our findings provide an example for transfer learning in genomics where knowledge can be transferred from an information-rich species to an information-poor one.

KEYWORDS: Cross-species gene prediction; specialized metabolism; transfer learning.

1. BACKGROUND

As more genome sequences become available, a major challenge in biology is to connect genotype to phenotype (Dowell *et al.* 2010). At the molecular level, phenotypes can be defined as products derived from genomic sequences, including transcripts, proteins and/or metabolites. Plants produce a diverse array of specialized metabolites, with estimates upwards of 200 000 structurally unique compounds (Ehrlich and Raven 1964; Hartmann 2007), many of which are important in medicine, nutrition and agriculture (Giovannucci 2002; Schmidt *et al.* 2008; Piasecka *et al.* 2015). Plant metabolic activities are broadly classified into two categories. The first is general (or primary) metabolism (GM), which involves the production of metabolites essential for survival, growth and

development in most, if not all, plant species (Hartmann 2007; Chen *et al.* 2011). In contrast, specialized (or secondary) metabolism (SM) leads to the accumulation of lineage-specific metabolites that may confer a fitness advantage in particular environments (Ehrlich and Raven 1964; Hartmann 2007; Pichersky and Lewinsohn 2011; Edger *et al.* 2015). For example, some plant specialized metabolites such as glucosinolates and terpenoids confer resistance against insects and pathogens (Wink 1988; Piasecka *et al.* 2015). Another difference between general and specialized metabolites is that the later tend to accumulate in specific tissues such as in trichomes or fruit (Tohge *et al.* 2013; Nakashima *et al.* 2016). In addition to their ecological and evolutionary importance, specialized metabolites are important for human health; ~25 % of medicinal compounds are derived from plant metabolites (Schmidt *et al.* 2007, 2008). For example, *Solanum nigrum* and *S. lyratum* produce glycosides that have anti-tumour activity in cancer cell lines (Nohara *et al.* 2006). *Atropa belladonna*, nicknamed 'beautiful woman' because in Roman times women used its extract to dilate their pupils (Rajput 2014), is a producer of the tropane alkaloids hyoscyamine and scopolamine, has anticholinergic activity and is used to treat parasympathetic nervous system disorders and asthma (Capasso *et al.* 2000; Grynkiewicz and Gadzikowska 2008). Furthermore, specialized metabolites contribute to desirable agronomic traits such as the aromas and flavours of fruits (Tohge *et al.* 2013) and defence against agricultural pests (Osbourn 1996).

Tomato is a model crop that has emerged as a system for investigating SM pathways. For example, the production of acylsugars, a specialized metabolite, in tomato and its wild relatives is important for repelling herbivores (Lucini *et al.* 2016; Maciel *et al.* 2017; Fan *et al.* 2019). Some specialized metabolites found in the tomato fruit also confer health benefits by, for example, reducing risk of cancers and coronary heart diseases (Giovannucci 2002; Blum *et al.* 2005; Clifford and Brown 2006). Despite recent progress in elucidating tomato SM pathways, our understanding of many of the steps in these pathways is incomplete due to the diversity of specialized metabolites. Many genes that underlie the production of specialized metabolites belong to the same gene families as genes involved in GM (Pichersky and Lewinsohn 2011; De Luca *et al.* 2012; Facchini *et al.* 2012; Milo and Last 2012), which makes them difficult to distinguish. Currently, genetic approaches are used to identify SM genes in tomato, including gene silencing (Itkin *et al.* 2013), genetic mapping (Xu *et al.* 2013) and the use of introgression lines (Schilmiller *et al.* 2010). In addition, genes involved in SM or belonging to a particular pathway can be predicted computationally. For example, protein sequence information can be used to predict enzymatic functions and assign genes to pathways (Karp *et al.* 2011; Chae *et al.* 2014; Schlapfer *et al.* 2017), which can have high error rates (Rost 2002). Gene co-expression networks have also been used to classify genes into specific metabolic pathways (Wisecaver *et al.* 2017). In addition, involvement of genes in a pathway can also be hypothesized using correlation of gene expression with the production of specific metabolites (Tohge *et al.* 2005; Saito *et al.* 2008; Adio *et al.* 2011). Finally, heterogenous gene features including gene duplication status, evolutionary properties, expression levels, placement in co-expression networks and protein domain content have been integrated using supervised machine learning to make SM/GM gene predictions in Arabidopsis (Moore *et al.* 2019).

Supervised learning approaches leverage instances (genes in this study) with known labels (SM or GM) to learn how the properties (i.e. features) of those instances can be best used to distinguish instances with different labels in the form of a predictive model (Fig. 1). There are two factors limiting computational predictions of SM/GM genes. First, although supervised learning methods for SM/GM prediction are effective in Arabidopsis, it remains unclear how these methods may work in species with less complete gene and pathway annotations. Second, as sequence similarity-based approaches have high error rates, it is challenging to transfer annotation information across species (Yu 2004). The goal of this study is to address these limitations using an approach called 'transfer learning' (Torrey and Shavlik 2010), where knowledge of SM/GM annotations from Arabidopsis was transferred to for predicting tomato SM/GM genes.

2. METHODS

2.1 Annotation

Only enzyme genes were included in this study. A gene was considered to be an enzyme gene if it had an EC or RXN number annotation in TomatoCyc or assigned using E2P2 v3.0 (Chae *et al.* 2014). Tomato pathway annotations were downloaded from the Plant Metabolic Network Database, TomatoCyc v. 3.2 (Schlapfer *et al.* 2017). Pathways that were nested under 'Secondary Metabolism Biosynthesis' or 'Secondary Metabolites Degradation' were considered SM pathways and genes within those pathways were considered SM genes. All other pathways were considered to be GM pathways. PMN defines 'secondary metabolism pathways' as 'pathways for the biosynthesis of secondary metabolites, which are organic compounds that are not directly involved in growth, development and reproduction of an organism'. These pathways are defined heuristically with the help of manual curators and are based on the reactions an enzyme catalyses. If a gene was annotated as being in both an SM pathway and a GM pathway, the gene was considered to be dual function (DF). Additionally, the biosynthesis of plant hormones was considered GM even though some hormone pathways fell under the DF category. If a pathway was nested under both 'secondary metabolism biosynthesis' and other general biosynthesis categories, the pathway was determined to be DF. For specific SM pathway annotations, the path ID from TomatoCyc was used.

2.2 Benchmark genes

The benchmark gene set was identified based on expert knowledge and literature mining. Tomato genes were defined as GM, SM or DF based on *in planta* functional analyses of mutant generated through gene silencing or knockout mutations and/or studies of *in vitro* biochemical activity. For the identity of the benchmark genes (i.e. manually curated as SM, GM or DF genes), the evidence used for manual curation and publications supporting the evidence, **see Supporting Information—Table S1**.

2.3 Features used for machine learning

All gene feature values can be found in **Supporting Information—Dataset S1** (available from Zenodo: https://zenodo.org/record/3835883). These 7286 features are divided into several categories, each with different numbers of features: protein domains (4232 features), expression value (280), co-expression (2670), evolution (78) and gene duplication (26). Protein domain Hidden Markov Models from Pfam v.30 (pfam.xfam.org/) was used to identify protein domains in annotated tomato protein sequences with HMMER (https://www.

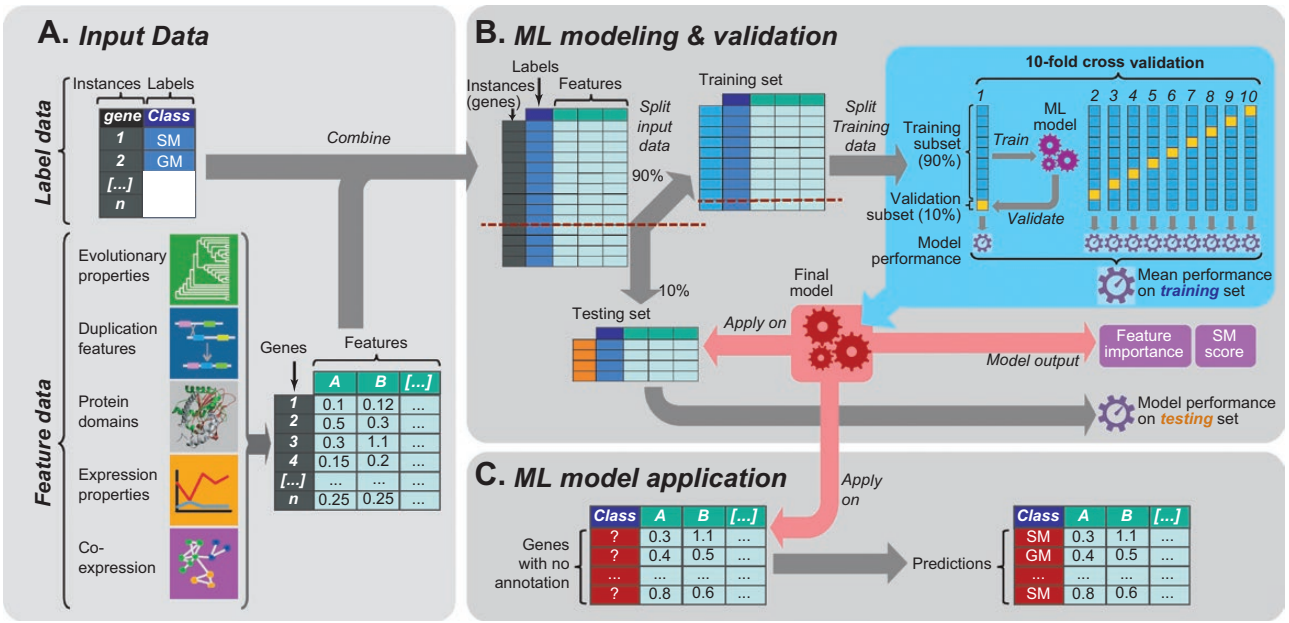

Figure 1. Machine learning workflow used in this study. (A) Schematic showing the input data for machine learning. The first inputs are labelled instances, collectively referred to as the model training set. In this case the instances are genes and the labels are the gene classes (response variable; either specialized or general metabolism, SM or GM). The second input is features, or the predictive variables in the model. In this study, five feature categories, which each contain multiple features, were utilized: evolutionary properties, duplication features, protein domains, expression properties and co-expression data. Each gene (instance) has a value for each feature. (B) The machine learning process. First the data set was split into training (90 %) and testing (10 %) sets. Next, equal numbers of training instances (i.e. 500 GM and 500 SM genes) were randomly selected from the training set to learn prediction models. This step was repeated 100 times, with different subsets of GM/SM genes selected from the training set in each repeat, to assess the robustness of prediction models. For each repeat, a 10-fold cross-validation was performed where the selected instances were further divided into a training subset (90 %) for building the model and a cross-validation subset (10 %; distinct from the testing set withheld from model building) to evaluate the model. After cross-validation, the optimal parameters were chosen to establish the final model for a given training/feature data set. Model performance assessed using the cross-validation sets was represented using the average *F*-measure of all repetitions. In addition to assessing performance based on cross-validation, another *F*-measure was calculated for the final model based on its application to the testing set that was held out from the beginning and never used for training. (C) The final model is applied on unannotated enzymatic genes to make predictions.

ebi.ac.uk/Tools/hmmer/) using the trusted cut-off, then a binary matrix for each gene and domain was created where 1 indicates the protein sequence of a gene has a given domain and 0 indicates it does not.

2.4 Expression value features

For expression value features, RNA-seq Sequence Read Archive (SRA) files for tomato were downloaded from National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) totalling 47 studies and 926 samples [**see Supporting Information—Table S2**]. These data sets included development (13 studies including fruit, flower, leaf, trichome, anther and meristem tissues), hormone-related (5 studies: cytokinin, auxin, abscisic acid, gibberellic acid and auxin inhibitor treatments), mutant (14 studies which compared various mutants against wild type), stress treatment (16 studies including shade, various pathogens, cold, light and heat treatments) and circadian (1 study with 60 samples). RNA-seq data were processed to determine both fold change and fragments per kilobase of transcript per million mapped reads (FPKM) (https://github.com/ShiuLab/RNAseq_pipeline). The SRA files were converted to fastq format and filtered with Trimmomatic (Bolger *et al.* 2014) for sequence quality with default settings. Bowtie (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) was used to create the genome index from the tomato NCBI *Solanum lycopersicum* genome 2.5, then RNA-seq reads were mapped to the tomato genome using TopHat (Trapnell *et al.* 2009). Samples with <70 % mapped reads were discarded. Cufflinks was then used to obtain FPKM values for mapped reads (Trapnell *et al.* 2010). HTSeq (Anders *et al.* 2015) was used to get raw counts for fold change analysis. Fold change analysis was performed using edgeR version 3.22.5 (McCarthy *et al.* 2012). Using each data set individually or all data sets combined, the median and maximum, and variation values for each gene were calculated. For breadth of differential expression, the number of conditions under which a gene was up- and downregulated was determined using log fold change values for each data set or combination of data sets. A gene was considered upregulated if it had a log fold change > 1 and a multiple-testing corrected *P*-value < 0.05 and downregulated if it had a log fold change < −1 and a corrected *P*-value < 0.05.

2.5 Co-expression features

For co-expression features, expression correlation was calculated using three methods: Pearson's Correlation Coefficient (PCC), Spearman's correlation and Partial Correlation (Corpcor). For each enzymatic gene (annotated and unknown), its expression correlation with each annotated SM/GM/DF gene was calculated (excluding self-correlation) using each method, each expression measure (fold change or FPKM) and each individual expression data set (with a distinct Gene Expression Omnibus GSE number), combination of data sets and all data sets combined [**see Supporting Information—Table S2**]. Then, for an enzymatic gene, E, the median and maximum of the correlation values of gene E for each class (SM, GM or DF) of genes was determined and used as feature values. Next, tomato genes were clustered into co-expression modules using six methods (*k*-means, *c*-means, complete/average/ward hierarchical clustering and weighted correlation network analysis) across each individual expression data set, data set combination and all data sets combined (same as for expression correlation). This was done using both fold change and FPKM values. Using Random Forest (RF) from Python package Scikit-Learn (Pedregosa *et al.* 2011), the top 200 co-expression modules that were the best for distinguishing SM and GM genes for each clustering method were selected to be part of the feature matrix for the models.

2.6 Evolutionary features

Orthologs and duplication nodes were determined using OrthoFinder (Emms and Kelly 2015). For input, protein sequence files from 26 different species were downloaded from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html), Sol Genomics Network (SGN, https://solgenomics.net/), PlantGenIE (http://plantgenie.org/) or NCBI (www.ncbi.nlm.nih.gov/genome): *Physcomitrella patens* 318 v3.3 (Phytozome), *Marchantia polymorpha* 320 v3.1 (Phytozome), *Selaginella moellendorffii* 91 v1.0 (Phytozome), *Picea abies* V1.0 (PlantGenIE), *Amborella trichopoda* 291 v1.0 (Phytozome), *Oryza sativa* 323 v7.0 (Phytozome), *Brassica rapa* 277 V1.3 (Phytozome), *Capsella rubella* 183 V1.0 (Phytozome), *Arabidopsis thaliana* 167 TAIR10 (Phytozome), *Arabidopsis lyrata* v2.1 (Phytozome), *Medicago truncatula* 285 Mt4.0v1 (Phytozome), *Vitis vinifera* 145 *Genoscope* 12x (Phytozome), *Aquilegia coerulea* V3.1 (Phytozome), *Populus trichocarpa* 210 v3.0 (Phytozome), *Theobroma cacao* 233 v1.1 (Phytozome), *Coffea canephora* (SGN), *Ipomoea trifida* V1.0 (NCBI), *Solanum tuberosum* V3.4 (SGN), *Solanum pennellii* SPENNV200 (NCBI), *S. lycopersicum* V2.5 (NCBI), *Capsicum annuum* CM334 v.1.55 (SGN), *Capsicum annuum* var. *glabriusculum* V2.0 (SGN), *Nicotiana tabacum* TN90 AYMY-SS NGS (SGN), *Nicotiana tomentosiformis* V01 (NCBI), *Solanum melongena* r2.5.1 (SGN) and *Petunia axillaris* V1.6.2 (SGN).

To identify putative orthologs, OrthoFinder was first run using default settings, including a BLAST run using protein sequence data for each pair of species with default parameters (*E*-value < 0.001), Markov clustering (inflation parameter = 0.1) to create initial orthogroups and dendroblast to create distance matrices between protein sequences of genes within each initial orthogroup. Initial gene trees were created using OrthoFinder. Three initial orthogroups were found to contain a single copy gene from each of the 26 species. Protein sequences of genes in each of these three orthogroups were aligned with MAFFT (Nakamura *et al.* 2018), and the alignment was used to build a phylogeny with RAXML (-m PROTGAMMAJTT -number of bootstraps 100 -outgroups Mpoly, Ppaten). This putative species tree was used as input into OrthoFinder to reconcile the gene trees for redefining orthogroups. Genes were considered to be homologous if they were in the same orthogroup. *dN/dS* (non-synonymous to the synonymous substitution rate ratio) was calculated with the yn00 program using PAML version 4.4.5 (Xu and Yang 2007). Gene family size was determined by the number of genes in an orthogroup within the species *S. lycopersicum*.

Duplication mechanism was determined using MCScanX-transposed (Wang *et al.* 2013). Four duplication mechanisms were used as features: (i) syntenic duplicates: paralogous genes present in within-species collinear blocks; (ii) dispersed (transposed) duplicates: for a pair of paralogs in species A, only one of their corresponding orthologs in species B is present in the inter-species syntenic block; (iii) tandem duplicate: a gene is adjacent to its paralog; (iv) proximal duplicates: a gene is separated by no more than 10 genes from its paralog. Genomic clustering features were derived from the genome annotation *S. lycopersicum* V2.5. A gene pair X and Y was considered to be in the same genomic cluster if gene X was located within 10 kbps downstream of the 3′-end or upstream of the 5′-end of gene Y, and X and Y were within 10 genes from each other. For gene X, the numbers of genes that qualified as Ys were determined separately for Ys in SM and GM pathways. The time point of the most recent duplication was determined from the most recent speciation node associated with each gene as determined by OrthoFinder (Emms and Kelly 2015). Duplication nodes ranged from most ancient (Node 0) to most recent (Node 24). The most recent duplication points for genes appearing to originate from multiple duplication nodes were defined by the highest-numbered node they belonged to [**see Supporting Information—Fig. S1**]. Pseudogenes in tomato were determined as in Wang *et al.* (2018) where genomic regions with significant similarity to protein-coding genes but with premature stops/frameshifts and/or were truncated were treated as pseudogenes (Wang *et al.* 2018). Detailed methods and parsing scripts for different features can be found in: https://github.com/ShiuLab/SM-gene_prediction_Slycopersicum.

2.7 Statistics

Statistical calculations were performed using R and Python. For discrete features, their relationships with SM/GM designations were determined by the Fisher's exact test. For continuous data, either the Mann–Whitney *U*-test (for comparing two groups) or the Kruskal–Wallis test followed by Dunn Pairwise Comparisons (for >2 groups) was used for tests of significance. Statistical results are in **Supporting Information—Table S3**.

2.8 Machine learning models

Multiple prediction models were made using the Python Sci-kit learn package (Pedregosa *et al.* 2011) with two algorithms, RF and Support Vector Machine (SVM). The pipeline (Fig. 1) used to run the models

can be found here: https://github.com/ShiuLab/ML-Pipeline. For each model, 10 % of the data was withheld from training as an independent, testing set. The remaining 90 % was used for training. Because the data set was unbalanced (2321 GM genes, 537 SM genes), 100 balanced data sets were created from random draws of GM genes to match the number of SM genes. Using the training data, grid searches over the parameter space of RF and SVM were performed. The optimal hyperparameters identified from the search were used to conduct a 10-fold cross-validation run (90 % of the training data set used to build the model, the remaining 10 % used for validation, Fig. 1) for each of the 100 balanced data sets. In total eight models were established using different feature and training data sets as described in Results and Discussion. For a subset of models, feature selection using RF was implemented to reduce the features to 50, 100, 200, 300, 400, 500 and 1000 to determine the optimal number of features. Model performance was evaluated using *F*-measure, the harmonic mean of precision (proportion of predictions that are correct) and recall (proportion of genes correctly predicted). In order to accurately compare across models that had different training sets, we compared using only one algorithm, RF, so that we would know that differences would be due to training sets and not the algorithm.

Each model outputs an SM score for each gene that is defined as the mean of predicted class probabilities of a sample to be in the SM class based on all decision trees in the forest. For each tree, the SM class probability was the fraction of genes predicted as SM. The threshold of the SM score used to determine if a gene was an SM or GM gene was the SM score value when the *F*-measure was maximized. The models also have an importance score for each input feature, which takes into account the weight of the feature by assessing how well the feature (node) splits the data between SM and GM genes in a decision tree in the 'forest' and this is weighted by the proportion of samples reaching that node (impurity score). The decrease in impurity score from each decision tree is averaged across all decision trees in the forest so that the higher the number, the more important the feature (Breiman 2001; Louppe 2014). For our study, the importance score sign was then changed to negative if the feature was correlated with GM genes—but remained positive if the feature was correlated with SM genes.

2.9 Shared features between Arabidopsis and tomato

Supporting Information—Dataset S2 (available from Zenodo: https://zenodo.org/record/3835883) lists the shared features and their values for Arabidopsis and tomato. For binary data, the features that were shared by both species were kept. These included two types of binary features: (i) protein domains: ~4000 Pfam domains common between Arabidopsis and tomato; (ii) evolutionary features: presence of a homolog in one of the 26 species, pseudogene paralog and tandem paralog, and whether the most recent duplication events took place in the lineages leading to the nodes shared by both species (Nodes 0–7). The shared features also included the following continuous features: gene family size, genomic cluster gene count, median/maximum *dN/dS* values between genes and their homologs in each of the 26 species, median/maximum *dN/dS* values between genes and their paralogs and expression-based features. To generate shared expression features, expression data were placed into four categories—abiotic, biotic, hormone and development—in both species. For each category, the Arabidopsis expression breadth, breadth of differential expression and co-expression correlation values using PCC were obtained from an earlier study (Moore *et al.* 2019). The same sets of features were generated for tomato in this study. Continuous values were normalized within each species so that they would be comparable across species. For the normalization script, see https://github.com/ShiuLab/SM-gene_prediction_Slycopersicum.

3. RESULTS

3.1 Identifying SM genes in tomato using machine learning approaches

Prior to applying the transfer learning approach, we first used a supervised learning approach to build a model capable of classifying a tomato gene as either an SM or GM gene to serve as the 'baseline' model for comparing against transfer learning results later on. For model training data, we used TomatoCyc-annotated metabolic enzyme genes (referred to as 'annotated genes', see Methods; for annotation information, **see Supporting Information—Table S1**), where genes in pathways under the category 'secondary metabolism biosynthesis' were considered SM genes (538 genes). Genes in any other pathway not under the SM category were considered to be GM genes (2313 genes). Genes found in both SM and GM pathways (158) were excluded. The remaining annotated genes were divided into a training set (90 %) for model training and a testing set (10 %) for model performance evaluation. For all annotated tomato SM and GM genes (2861), we collected and processed five gene feature categories (Fig. 1A): evolutionary properties, gene duplication mechanism, protein domain content, expression values and co-expression patterns (7286 total features, see Methods; for feature values, **see Supporting Information—Dataset S1**). The values of these features for genes in the training set were then used to train multiple machine learning models for predicting whether a gene was likely an SM or GM gene (see Methods, Fig. 2A).

We determined model performance by calculating *F*-measure (the hormonic mean of precision and recall, see Methods). For other measure of model performance, **see Supporting Information—Table S2**. The best performing model (Model 1) has *F*-measure = 0.74 [**see Supporting Information—Fig. S2A**]. The Model 1 *F*-measure is significantly better than a random guess (0.5) but far from perfect (1). Using Model 1, 76.6 % of annotated SM genes and 71.0 % of annotated GM genes had predictions consistent with their TomatoCyc annotations (Fig. 2B). To provide an independent validation, the model was then applied to the testing set, which resulted in a similar *F*-measure of 0.73 (Fig. 2C; **see Supporting Information—Table S4**). Because the test set was withheld from model training, this indicated the model could be applied to genes with no annotation and provide reasonable predictions. By applying Model 1, each gene was given a likelihood score, referred to as the SM score (see Methods), which indicates how likely a particular gene is to be an SM gene (Fig. 2B). For SM scores and SM/GM predictions for all tomato enzymatic genes for all models, **see Supporting Information—Table S5**.

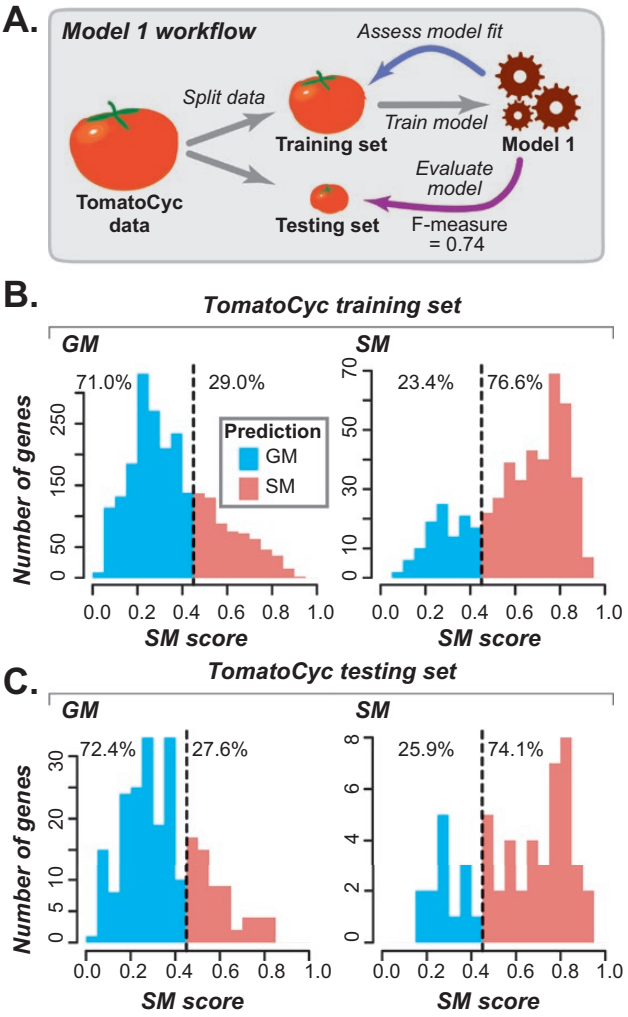

Figure 2. Tomato-based Model 1 and its performance. (A) Schematic illustrating Model 1, in which a tomato data set with 7286 tomato features were used. The model was built using TomatoCyc annotations and applied to tomato genes. (B) Distribution of Model 1 SM gene likelihoods (SM scores) for the TomatoCyc-annotated SM and GM genes in the training set. Prediction threshold, based on the score with the highest *F*-measure, is indicated by the dotted line, and predicted GM (blue) and SM (red) genes are to the left and to the right of the line, respectively. Percentage values indicate the percent total genes predicted as GM or SM. (C) Distribution of Model 1 SM scores for testing SM and GM genes that were withheld from model training.

We identified features with the top 50 importance scores from Model 1 (**see Supporting Information—Fig. S2B**; for feature importance for each model, **see Supporting Information—Table S6**). The higher the importance score, the better the feature is at separating SM and GM genes. By and large, the important features for the tomato Model 1 is similar to those for predicting Arabidopsis SM/GM genes (Moore *et al.* 2019). For example, similar to SM genes in Arabidopsis, tomato SM genes tend to be in larger gene families (median = 8) compared with GM genes (median = 3, Fig. 3A; for test statistics, **see Supporting Information—Table S3**), are more likely to be tandem duplicates (37 %) than GM genes (13 %), have a lower proportion as syntenic duplicates (17 %) compared with GM genes (25 %, Fig. 3B) and have higher synonymous/synonymous substitution rates (*dN/dS*) relative to GM genes in both cross-species (Fig. 3C; **see Supporting Information—Fig. S3A–H**) or within-species (Fig. 3D; **see Supporting Information—Table S6**) comparisons. The lower the *dN/dS* value, the stronger the negative selective pressure a gene has experienced. Thus, SM genes were experiencing less intense negative selection compared to GM genes. We also found that many more homologs of tomato SM genes exist within-species or in closely related species compared to GM genes (Fig. 3E).

Variation in transcriptional levels and patterns between genes may represent differences in their functions and can therefore also be key features distinguishing SM and GM genes. We compiled 47 transcriptome studies (for details on the data sets, **see Supporting Information—Table S2**) spanning a range of environmental conditions, hormone treatments and developmental stages, mostly in wild-type genetic backgrounds. In Model 1, 147 out of the top 200 most informative features were related to expression [**see Supporting Information—Table S6**]. For example, maximum log fold change between developmental stages, circadian time points, mutants vs. wild type, and hormone treatments vs. controls are among the top expression features (ranked 12–30, **see Supporting Information—Fig. S1B**; **Table S6**). Specialized metabolism genes tended to have higher maximum fold change values (Fig. 3F–I; **see Supporting Information—Tables S3 and S2**), but lower expression levels [**see Supporting Information—Fig. S3I and J**] than GM genes. Thus, SM gene expression tends to be more variable across developmental stages, times of day and environment. Consistent with this, expression variation (median absolute deviation, see Methods) is also an important feature [**see Supporting Information—Table S6**]. For example, many specialized metabolites important for fruit flavour and colour are produced during tomato fruit development (Tohge *et al.* 2013). Aside from gene expression, the enrichment of specific protein domains such as the P450 domain among SM genes [**see Supporting Information—Fig. S3K**] is an additional feature that differentiates them from GM genes.

3.2 Characteristics of genes with inconsistent annotations and predictions

Although the tomato SM/GM prediction model *F*-measure (0.74) was significantly better than a random guess (0.5), 29 % of GM genes were mis-predicted as SM and 23 % of SM genes were mis-predicted as GM when using an SM score threshold determined based on the optimal *F*-measure (Fig. 2B). In addition, the tomato model did not perform as well as an earlier model for predicting Arabidopsis SM/GM genes (*F*-measure = 0.79, Moore *et al.* 2019). Note that the tomato model is trained on TomatoCyc annotations, which can be of poorer quality than those of AraCyc (Arabidopsis annotations)—there are only 16 experimentally verified TomatoCyc SM/GM genes compared to 1652 in AraCyc. To understand why we obtained a high rate of mis-predictions, we assessed what features may cause a gene to be mis-predicted.

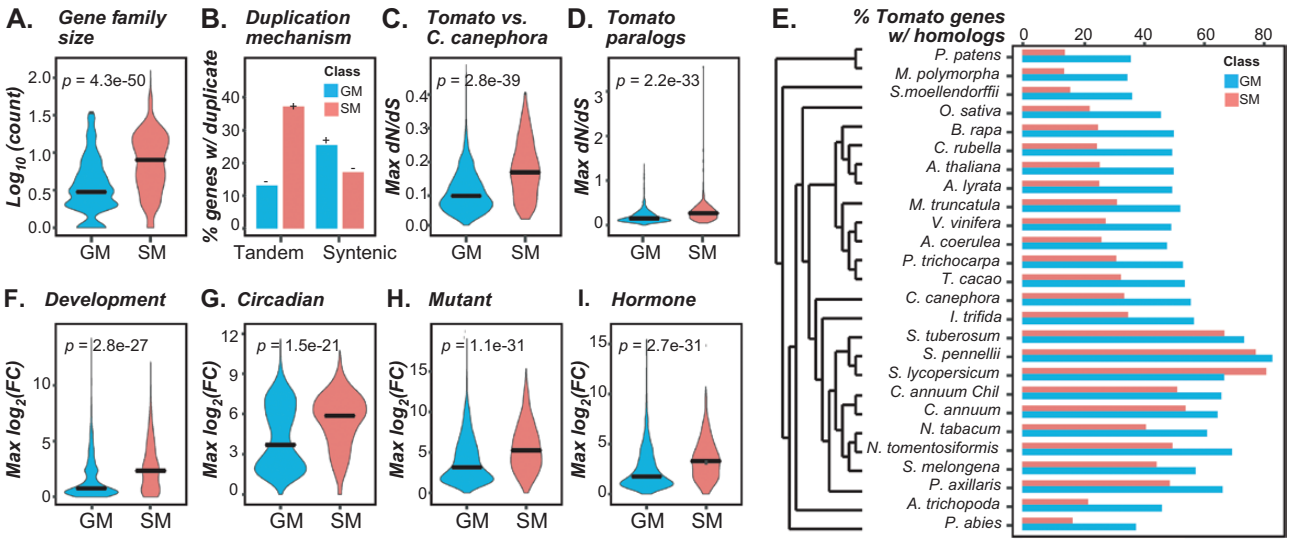

Figure 3. Duplication, evolutionary and expression features important for Model 1 predictions of SM and GM genes. (A) Log_{10} of gene family size (number of paralogs) for the families GM (blue) and SM genes belong to. All *P*-values are from the Mann–Whitney *U*-tests between SM and GM genes. (B) Percent of GM and SM genes with at least one duplicates derived from tandem or syntenic mechanism. (C) Maximum *dN/dS* values from comparisons of tomato SM and GM genes to homologs in *C. canephora*. (D) Maximum *dN/dS* values of tomato SM and GM genes to their paralogs. (E) Phylogenetic tree of 26 species and a bar plot showing the percentage of tomato GM and SM genes that have at least one homologs in each species. (F–I) Distributions of maximum fold change (F–G) across all conditions in expression data sets including (F) the meristem development, (G) the circadian, and (H) between mutant and control in mutant data set and (I) treatment and control in hormone treatment data set.

We found genes annotated as GM but predicted as SM (annotated→predicted: GM→SM) and genes annotated as SM but predicted as GM (SM→GM) defied the general trend in the evolutionary feature, maximum *dN/dS* value. GM→SM and SM→GM genes have higher and lower *dN/dS* values, respectively, compared with those genes with consistent annotations/predictions (GM→GM, SM→SM, Fig. 4A and B; **see Supporting Information—Fig. S4A–H**). For example, one of the GM→SM genes, *XP_010323708* (*Solyc07g054880.3.1*), has a maximum *dN/dS* of 0.25 with its *C. canephora* homolog, which is much higher than that observed for GM→GM genes (median *dN/dS* = 0.10; **see Supporting Information—Dataset S1, Table S3**). This high *dN/dS* value likely contributed to the prediction of this gene as SM. When looking more closely at *XP_010323708*, this gene was reported to encode a methylketone synthase that produces methyl ketones specific to the *Solanum* genus (Yu *et al.* 2010), and should be annotated as an SM gene. Similarly, we found that three tomato *Glycoalkaloid metabolism* (*GAME*) genes were also GM→SM genes with high *dN/dS* values. These genes, *GAME4*, *GAME12* and *GAME17*, are involved in steroidal glycoalkaloids production and should be considered SM genes. These examples demonstrate that a subset of mis-predictions is likely due to mis-annotation. In contrast to GM→SM genes, SM→GM genes have a maximum *dN/dS* score (median = 0.27) from comparisons to tomato paralogs that is significantly below that for SM→SM genes (median = 0.33, Fig. 4B; **see Supporting Information—Table S3**).

Other evolutionary properties, duplication and expression features were skewed for mis-predicted genes. For example, while the general trend of SM genes is to be in larger gene families than GM genes, but GM→SM genes tended to belong to larger gene families (median = 5) than those with consistent GM annotations/predictions (GM→GM, median = 3) and vice versa with SM genes (Fig. 4C). Additionally, we found that GM→SM genes tended to be tandem duplicates, similar to SM→SM genes and in contrast to GM→GM and SM→GM genes (Fig. 4D). Aside from evolutionary properties and duplication features, compared with SM→SM genes, GM→SM genes also had similar maximum expression fold differences (Fig. 4E–H), expression variation values (Fig. 4I and J), median expression levels [**see Supporting Information—Fig. S4I and J**] and protein domain compositions [**see Supporting Information—Fig. S4K**]. In general distributions of feature values for mis-predicted GM→SM genes mirrored those for annotated SM genes and feature distributions for SM→GM genes were similar to the overall distributions for annotated GM genes. These findings indicate that mis-predicted genes tend to possess feature values that are deviated from the norms, where some SM genes in TomatoCyc looked more like GM genes and some GM genes looked more like SM genes. An open question is whether these mis-predicted genes were mis-annotated in the first place or if they were correctly annotated but incorrectly predicted by a faulty model. This prompted us to look more closely at mis-predicted genes to see if their annotations were supported by compelling experimental evidence.

3.3 Manual curation of SM/GM genes to obtain a benchmark set

Based on comparison of feature value distributions, mis-predicted genes tend to possess properties more similar to the class (GM or SM) they were mis-predicted as. This is not a surprising outcome because

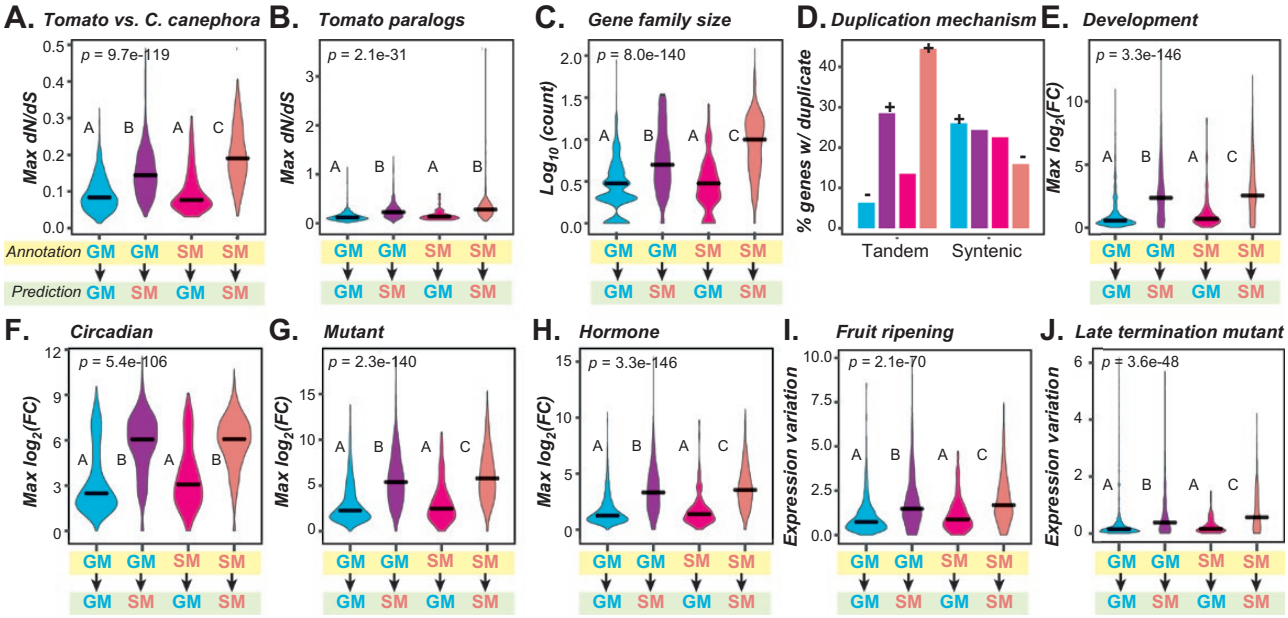

Figure 4. Feature distributions of genes with predictions contrary to their annotated classification. (A) Maximum *dN/dS* values from comparisons of genes in four classes to homologs in *C. canephora* among four Annotation (yellow rectangle) and Prediction (green rectangle) classes. Blue: GM_GM, a GM gene predicted as GM. Purple: GM_SM: a GM gene predicted as SM. Magenta: SM_GM: an SM gene predicted as GM. Red: SM_SM: an SM gene predicted as SM. For this and subsequent figure depicting continuous data, *P*-values are from the Kruskal–Wallis test and *post hoc* comparisons were made using the Dunn's test. Different letters indicate statistically significant differences between groups ($P < 0.05$). (B) Maximum *dN/dS* values of genes in four classes to their paralogs. (C) Log_{10} of gene family sizes. (D) Percentage of genes with at least one duplicates derived from tandem or syntenic mechanism. Colour scheme following that in (A). + and −: significant enrichment of SM and GM genes, respectively at 5 % significance level after Benjamin–Hochberg multiple testing correction. (E–H) Distributions of maximum fold change over the same expression data set as in Fig. 3F–I including (E) the meristem development, (F) the circadian, (G) mutant and (H) hormone treatments. (I and J) Distribution of fold change variation in two data sets: (I) fruit ripening (1 study, 12 samples) and (J) late termination mutant (1 study, 12 samples).

our explicit goal was to learn about generalizable differences between annotated GM and SM genes. The unresolved question is why mis-predictions occur. Three factors may account for mis-predictions: (i) the genes were annotated correctly, and Model 1 was incorrect; (ii) Model 1 made correct predictions, but the annotations were incorrect; and (iii) both annotations and predictions were correct, because these genes have roles in both GM and SM, i.e. they have DFs. To assess these possibilities, we manually curated a set of 88 tomato genes (83 with annotations in TomatoCyc) encoding enzymes classified as SM, GM or DF based on published evidence of *in vitro* enzyme activity and/or *in planta* characterization (see Methods). These 88 genes are collectively referred to as the benchmark set, and the curated evidence supporting their SM/GM/DF designations is shown in **Supporting Information—Table S1**.

Out of 31 TomatoCyc-annotated GM genes analysed, 24, 5 and 2 were manually curated as GM, SM and DF genes, respectively. Among the five annotated GM genes that were manually curated as SM, all five were predicted by Model 1 as SM. Four are the aforementioned genes *Methylketone synthase* (*XP_010323708*), *GAME4*, *GAME12* and *GAME17*. The three *GAME* genes contribute to glycoalkaloid biosynthesis in several Solanaceae species (Itkin *et al.* 2013). The fifth gene correctly predicted by Model 1 is the neofunctionalized gene *Isopropylmalate synthase 3* (*IPMS3*), which acquired a role in an SM pathway after the duplication of an ancestral *IPMS* gene involved in amino acid metabolism (GM pathway). *IPMS3* is a tissue-specific SM gene involved in acylsugar production in glandular-trichome tip cells and is curated as an SM gene based on empirical evidence (Ning *et al.* 2015). Thus, in these cases, Model 1 made the correct predictions, but the annotations were incorrect. Two *Geranylgeranyl diphosphate synthases* (*GGPS, NP_001234087* and *NP_001234302*) are manually curated as DF genes, but annotated by TomatoCyc as GM and predicted by Model 1 as SM. The challenge in classifying these genes might arise from the fact that GGPS enzymes catalyse core reactions in isoprenoid biosynthesis, an ancient and diverse pathway that leads to the synthesis of both GMs and lineage-restricted SMs (Ament *et al.* 2006).

Manual curation of 45 TomatoCyc-annotated SM genes revealed that three were likely GM genes and five were likely DF genes. We chose to look in detail at the three manually curated GM genes that were annotated as SM: two carotenoid biosynthesis genes, *PHYTOENE DESATURASE* and *TANGERINE* (Isaacson *et al.* 2002; Romero *et al.* 2011), and a cytochrome P450, *SlKLUH*, that, when

mutated, disrupts chloroplast homeostasis and has pleiotropic effects on plant growth and development (Chakrabarti *et al.* 2013). As carotenoid biosynthesis is conserved among all photosynthetic organisms (Cunningham and Gantt 1998), and disruptions in basic development processes, such as gametophyte and seed development, are an indicator of essentiality in all plants (Meinke *et al.* 2008), these genes should be considered GM genes. In all three cases, Model 1 predictions agreed with the TomatoCyc SM annotations and, thus both the predictions and annotations were incorrect.

Next, we focused on comparing the manually curated benchmark set to Model 1 predictions. We found that 17 out of 29 (58.6 %) total benchmark GM genes, and 13 of the 24 benchmark GM genes that were annotated as GM by TomatoCyc (54 %), were incorrectly predicted as SM by Model 1 [**see Supporting Information—Fig. S5A**; **Table S5**]. Thus, Model 1 tended to mis-predict benchmark GM genes as SM genes. In contrast, of the 51 total benchmark SM genes, 45 (88.2 %) were correctly predicted by Model 1 [**see Supporting Information—Fig. S4A**; **Table S5**]. Taken together, our Model 1 predictions were mostly consistent with the SM benchmark classifications. However, the model clearly had trouble predicting known GM genes. With regard to TomatoCyc-annotated genes, the opposite was true—24 of 29 (82.8 %) benchmark GM genes were correctly annotated as GM, and 37 of 47 (78.7 %) benchmark SM genes were correctly annotated as SM. Therefore, for SM gene prediction, Model 1 has a lower error rate (11.8 %) compared with the TomatoCyc annotation (21.3 %), indicating that a higher proportion of benchmark SM genes were annotated in TomatoCyc than GM genes. However, for benchmark GM genes, Model 1 has a higher error rate (46 % of benchmark GM genes predicted as SM genes) than the TomatoCyc annotation (14.3 % of benchmark GM genes predicted as SM).

3.4 Using transfer learning to make predictions across species

Based on analysis of the benchmark data, there are two major sources for mis-predictions. The first is that a subset of the TomatoCyc-annotated SM or GM genes were incorrectly annotated, and these mis-annotations were propagated into Model 1. The second is that Model 1 predicts these genes correctly. These two explanations are not mutually exclusive, and the extent to which each contributes to mis-predictions remains to be determined. To determine the most likely reason for the mis-predictions and to improve upon Model 1, we used both the benchmark gene set and the TomatoCyc annotations to build a new model (referred to as Model 2), but this did not improve the prediction accuracy (*F*-measure = 0.74, same as Model 1; **see Supporting Information—Fig. S1A**, **Table S4**). This was likely due to the small proportion of benchmark gene-inspired annotation corrections (30) relative to the large number of TomatoCyc-annotated genes (2858).

We next asked whether information from Arabidopsis, which diverged from the tomato lineage 83–123 million years ago (Ku *et al.* 2000; Tomato Genome Consortium 2012), could be used to improve gene predictions in tomato. Here we use a machine learning approach called transfer learning (Torrey and Shavlik 2010) in which a base model is first built using data from Arabidopsis and then the learned features and/or the base model itself are used to make predictions in tomato using the tomato annotations and features. To accomplish this, a list of 4197 similar features in Arabidopsis and tomato (referred to as shared features, see Methods) were identified. A model was built using previously defined AraCyc GM/SM annotations (Moore *et al.* 2019) and shared features. This model is referred to as Model 3 (Fig. 5A) that performed reasonably well in separating *A. thaliana* GM/SM genes (Fig. 5B). For comparison, we also built a model (Model 4) using TomatoCyc GM/SM annotations and tomato data for the same shared features as in Model 3 and to train the model (Fig. 5C). Model 3 built with Arabidopsis shared feature data had an *F*-measure = 0.81 when it was used to predict Arabidopsis genes as GM/SM [**see Supporting Information—Table S4**]. In comparison, Model 4 built with tomato shared feature data had an *F*-measure = 0.75 when used for predicting tomato annotations [**see Supporting Information—Table S4**]. Additionally, more GM/SM genes in Arabidopsis are predicted correctly by Model 3 (Fig. 5B) than GM/SM genes in tomato by Model 4 (Fig. 5D). The higher *F*-measure and better predictions for Model 3 are consistent with there being more experimentally based gene annotations for Arabidopsis than for tomato that likely contribute to the differences in model performance.

To assess whether the Arabidopsis-based model can be applied to tomato directly, we next applied Arabidopsis-based Model 3 to predict tomato SM and GM genes and obtained an *F*-measure of 0.69 (Fig. 5E; **see Supporting Information—Table S4**). This was substantially lower than the *F*-measure obtained when applying tomato-based Model 4 to tomato genes (*F*1 = 0.75; **see Supporting Information—Table S4**), and fewer TomatoCyc-annotated GM/SM genes were predicted accordingly (Fig. 5F). With a closer look, we found that the poor performance was likely due to substantial mis-annotations, particularly GM genes, in TomatoCyc. Based on SM scores for these models, 21.1 % of TomatoCyc GM genes were predicted as GM genes by tomato Model 4 but predicted as SM genes by Arabidopsis Model 3 (lower right quadrant, Fig. 6A; **see Supporting Information—Table S5**). However, Model 3 predicted 50 % of benchmark tomato GM genes as GM [**see Supporting Information—Fig. S5B**], which—although far from perfect—is substantially better compared with the percentage of benchmark GM genes correctly predicted by tomato Model 4 (25 %; **see Supporting Information—Fig. S5C**). Thus, Arabidopsis data (when used to train Model 3) led to improved tomato GM gene predictions compared with tomato annotation data. Based on our finding that annotated GM genes were more likely to be mis-annotated compared with annotated SM genes [**see Supporting Information—Fig. S5B and C**], this indicates that the decline in model performance was due to mis-annotation of tomato genes.

3.5 Reasons why Arabidopsis-based Model 3 had suboptimal performance on tomato genes

To further assess the possibility of mis-annotation, we asked how well Models 3 and 4 predict benchmark SM genes. We found that benchmark tomato SM genes were less well predicted using Arabidopsis Model 3 (84 % correctly predicted; **see Supporting Information—Fig. S5B**), a substantial drop from the near perfect predictions (97 %) using tomato Model 4 [**see Supporting Information—Fig. S5C**]. This indicated that Arabidopsis data may provide more useful information about true GM genes in other species than about SM genes,

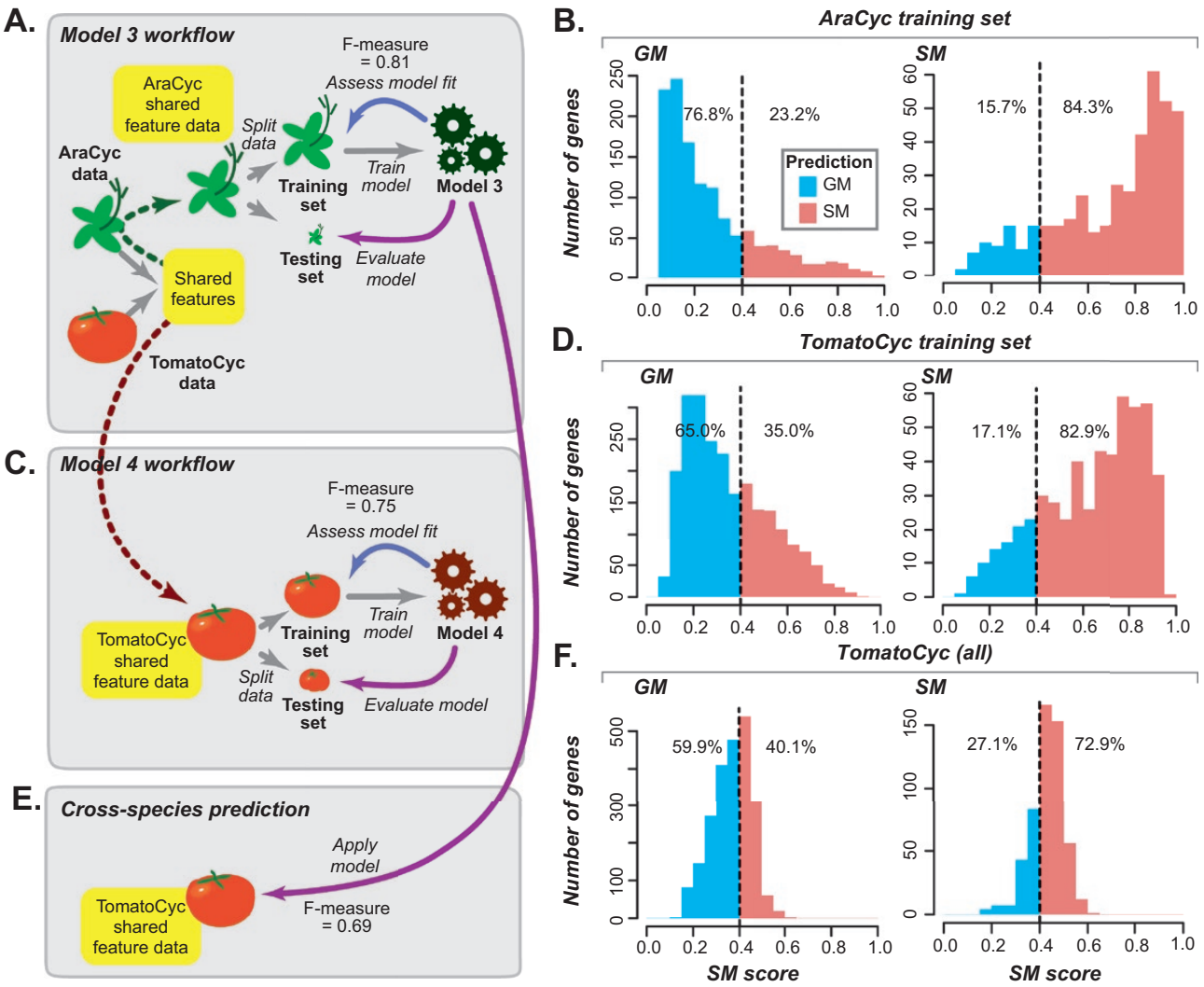

Figure 5. Arabidopsis Model 3 and tomato Model 4 predictions. (A) Model 3 built with the Arabidopsis training data using only shared feature set between Arabidopsis and tomato. (B) Distribution of Model 3 SM scores of Arabidopsis SM and GM training set genes. Dotted line: same as in Fig. 2. Blue and red: SM and GM genes, respectively. (B) Model 4 built with the tomato training data using only shared feature set between Arabidopsis and tomato. (D) Distribution of Model 4 SM scores of tomato training set GM and SM genes. (E) Application of Model 3 on tomato genes using the shared feature set between Arabidopsis and tomato. (F) Distribution of Model 3 SM scores on TomatoCyc-annotated GM and SM genes.

likely because GM genes are conserved among plant species, and many have been studied using Arabidopsis as a model. Thus, it is more straightforward to transfer knowledge about Arabidopsis GM genes to tomato. Specialized metabolism genes, in contrast, are by definition lineage-specific and not all SM gene properties will be shared across species, which explains the drop in prediction accuracy in Model 3 compared with Model 4. Nonetheless, the SM likelihood scores are largely consistent between Models 3 and 4 (Fig. 6A and B; **see Supporting Information—Fig. S6A and B**; **Table S5**), indicating there remain substantial similarities among SM genes across species.

When we looked into the models in more detail, we found that the major reason why Arabidopsis Model 3 predicted genes differently from tomato Model 4 is because they have different important features (Fig. 6C). Aside from the three most consistently important ones, which are gene family size, expression correlation between SM genes during development and expression correlation between GM genes in the hormone data set (Fig. 6C), many features such as maximum *dN/dS* relative to *C. canephora* homologs are highly important in tomato Model 4 but much less important in Arabidopsis Model 3. Upon examination of feature value distributions, we found that, in general, the feature values of the tomato Model 4-based predictions more closely aligned with those of the annotated genes in the tomato training set than with Arabidopsis Model 3-based predictions (Fig. 6D–F). For example, annotated tomato SM genes predicted as

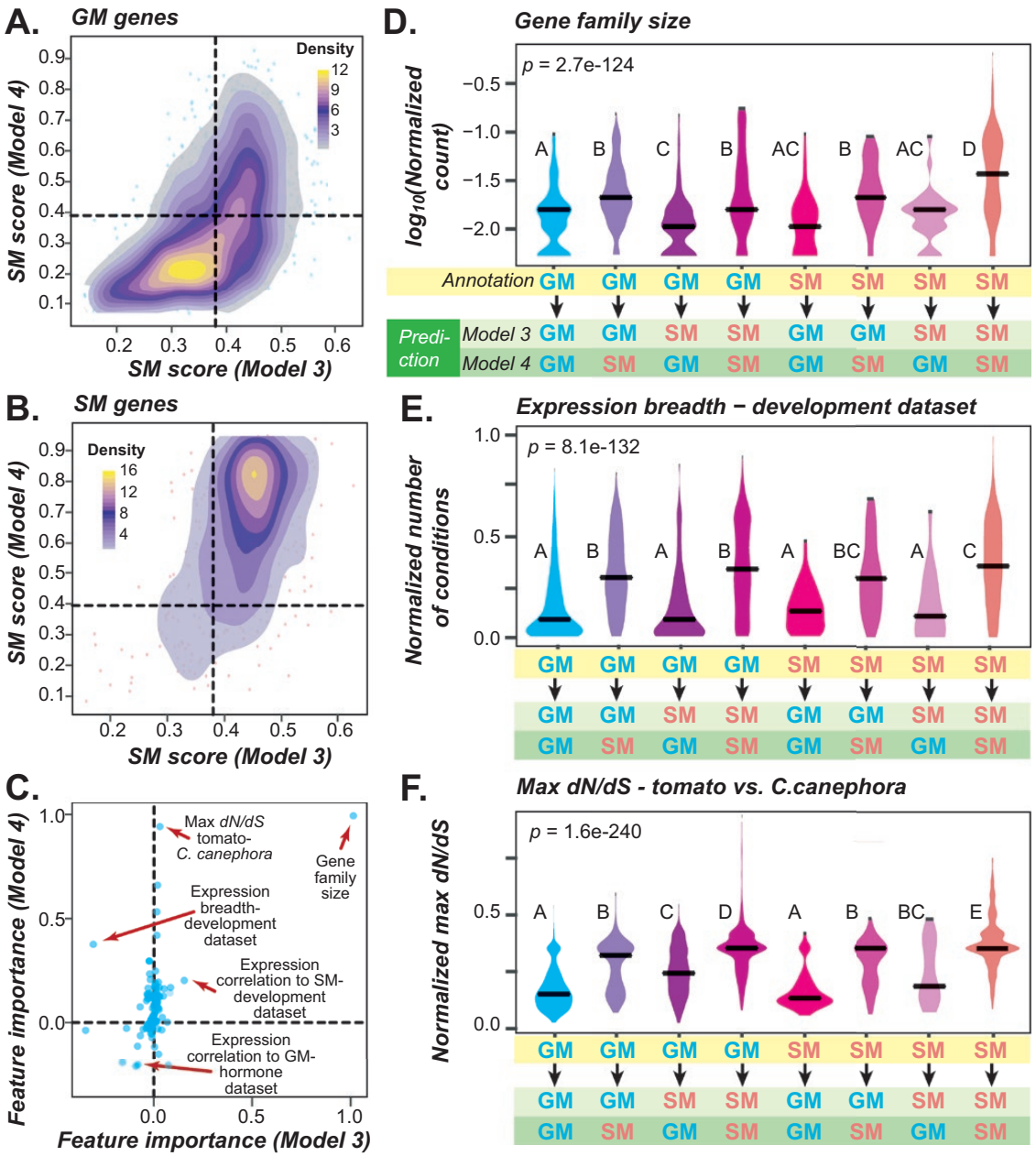

Figure 6. Tomato Model 4 and Arabidopsis Model 3 comparison. (A) Correlations between TomatoCyc GM genes SM scores based on Model 3 and Model 4. Colour: data point density ranges from high (yellow) to medium (purple), to low (fading purple). (B) Correlation of TomatoCyc SM genes SM scores based on Model 3 and Model 4. (C) Correlations in feature importance values based on Model 3 and Model 4. Arrows point to example consistent and inconsistent features. (D–F) Feature value distributions for annotated SM and GM genes that are predicted as SM or GM genes by Model 3 and Model 4. *P*-values are from Kruskal–Wallis tests and *post hoc* comparisons were made using the Dunn's test. Different letters indicate statistically significant differences between groups ($P < 0.05$). (D) Log_2 of normalized gene family size. (E) Normalized expression breadth based on the meristem development data. (F) Normalized maximum *dN/dS* between tomato genes and their homologs in *C. canephora*.

GM genes by Arabidopsis Model 3 but as SM genes by tomato Model 4 (referred to as $SM{\rightarrow}GM_3/SM_4$ genes, the plot in pink, Fig. 6D) tend to be in large gene families like $SM{\rightarrow}SM_3/SM_4$ genes (the orange plot, Fig. 6D). In contrast, $SM{\rightarrow}SM_3/GM_4$ genes (the brown plot, Fig. 6D) tend to be in small gene families. This indicates that tomato Model 4 is more strongly influenced by gene family sizes when differentiating SM and GM genes than Arabidopsis Model 3. This general pattern is also true for expression-based and *dN/dS* features (Fig. 6E and F; **see**

Supporting Information—Fig. S6C–F). For example, GM→GM_3/SM_4 genes are likely predicted as SM genes by tomato Model 4 (the second plot, Fig. 6F) because they have high *dN/dS* values similar to those of the SM genes used to train the model (the eighth plot, Fig. 6F). However, GM→SM_3/GM_4 genes (the third plot, Fig. 6F) tend to have lower *dN/dS* values similar to those of the GM genes used to train the model (the first plot, Fig. 6F). In the above example, the Arabidopsis Model 3 yields predictions contrasting with those from tomato Model 4. Most notably, the Arabidopsis Model 3-based predictions have feature values that mostly defy the general trends of the GM and SM genes in the tomato training data. This indicates that there are differences between the training data for Arabidopsis Model 3 and tomato Model 4 that bias each model.

3.6 Improving the tomato-based model by removing potentially mis-annotated genes identified based on the Arabidopsis model predictions

We hypothesized that if the Arabidopsis Model 3-based predictions are correct, then the genes with contrasting predictions and annotations are mis-annotated and their removal from the training data would lead to significantly improved predictions. This is because training the model from incorrect examples (i.e. mis-annotated entries) will lead to suboptimal models making erroneous predictions. On the other hand, if the Arabidopsis Model 3-based predictions are completely uninformative, the removal of genes from the training set would not improve the prediction. Thus, to further test the above hypotheses, we removed TomatoCyc-annotated GM and SM genes that had contradictory predictions from Arabidopsis-based Model 3 (i.e. GM→SM_3 and SM→GM_3) from the training set. Using this filtered training data set, a new tomato data-based model, Model 5, was generated using the same shared feature set between Arabidopsis and tomato for Models 3 and 4 (Fig. 7A, see Methods).

When we applied this filter to build tomato Model 5, there was a dramatic improvement in tomato GM/SM gene predictions (*F*-measure=0.92; **see Supporting Information—Fig. S2A**; **Table S4**) compared with predictions based on Model 3 (*F*-measure = 0.69; **see Supporting Information—Fig. S1A**; **Table S4**) and Model 4 (*F*-measure = 0.75; **see Supporting Information—Fig. S2A**; **Table S4**). In particular, we were able to predict 90.9 % of all annotated GM genes and 92.4 % of all annotated SM genes in the filtered training data as GM and SM genes, respectively (Fig. 7B; **see Supporting Information—Table S4**). Thus, Model 5, trained on a data set where GM→SM_3 and SM→GM_3 genes have been removed, is significantly improved compared with previous models. To validate Model 5 with an independent data set, we applied it to a testing set of 159 SM and GM genes withheld from Model 5 during training. We found that 84 % and 88 % of the test set GM and SM genes, respectively, were predicted consistently with their annotations [**see Supporting Information—Fig. S7A**].

To test whether model improvement was due to the filtering out of a subset of mis-annotated genes from the tomato training data and not just to the removal of genes in general, we built 10 additional models (collectively referred to as Model 6) using the same number of tomato SM and GM training genes as used for training Model 5, except that the genes were removed randomly. We found the median *F*-measure to be the same as that from Model 4 (where no SM or GM genes were removed; **see Supporting Information—Fig. S2A**, **Table S4**, see Methods), showing no model improvement. Thus, the improvement in model performance of tomato Model 5 could not be attributed to random gene removal.

This improvement was likely achieved for two reasons. First, consistent with the manual annotation result reported earlier, this improvement is likely because the filtered tomato training data did not contain as many mis-annotated genes that would confuse the model. The second possibility is that filtered genes were hard-to-predict or edge cases—i.e. they simply have properties unlike the average SM or GM genes. If the filtered genes were edge cases, we would expect that their SM score would be close to the threshold SM score. To assess this, we plotted the filtered SM and GM gene SM scores between Models 3 and 5 [**see Supporting Information—Fig. S8**]. We found that in the case of filtered GM genes [**see Supporting Information—Fig. S8A**], the majority are most similar to the remaining, unfiltered SM genes [**see Supporting Information—Fig. S8B**], indicating they are mostly like average SM genes and, thus, are not simply hard-to-predict GM genes. In contrast, the SM scores of filtered SM genes [**see Supporting Information—Fig. S8C**] were much closer to the threshold compared to the remaining GM genes, indicating they may in fact be atypical, hard-to-predict SM genes [**see Supporting Information—Fig. S8D**]. However, consider separation of the filtered SM genes in the SM score space based on both Models 3 and 5, they are much more similar to GM genes than they are to SM. Thus, while there are edge cases, the improvement is likely mainly due to removal of mis-annotated genes.

After showing that Model 5 performed significantly better on training data, we next asked how Model 5 faired in predicting benchmark GM genes. We found that 55 % of benchmark GM genes were correctly predicted by Model 5 [**see Supporting Information—Fig. S7C**; **Table S5**], compared with 25 % for tomato Model 4 and 50 % for Arabidopsis Model 3 [**see Supporting Information—Fig. S5F and G**]. In contrast, there was no improvement in benchmark SM predictions when comparing Model 4 (94 % correct; **see Supporting Information—Fig. S5F**; **Table S5**) to Model 5 (92 % correct; **see Supporting Information—Fig. S7C**; **Table S5**). These findings indicate that the improvement in Model 5 is likely due to its ability to determine true GM genes while maintaining true SM gene prediction performance. In addition, our results suggest that the filtering step corrected for genes mis-annotated in TomatoCyc. Consistent with this conclusion, 73.8 % of the annotated SM genes that were removed from the Model 5 training data because Model 3 called them as GM, were predicted as GM genes by Model 5 [**see Supporting Information—Fig. S7B**]. Similarly, among annotated GM genes removed from the training set because they were predicted as SM genes by Model 3, 72.3 % were predicted by Model 5 as SM genes [**see Supporting Information—Fig. S7B**]. This indicates that Model 3 was able to identify both SM and GM genes that were likely mis-annotated and their introduction into the training set of Model 4 led to a suboptimal model. After their removal, the new model was able to better identify mis-annotations, resulting in improved predictions. Importantly, randomly removed genes from Model 6 did not improve predictions, indicating Model 3 is identifying probable mis-annotations.

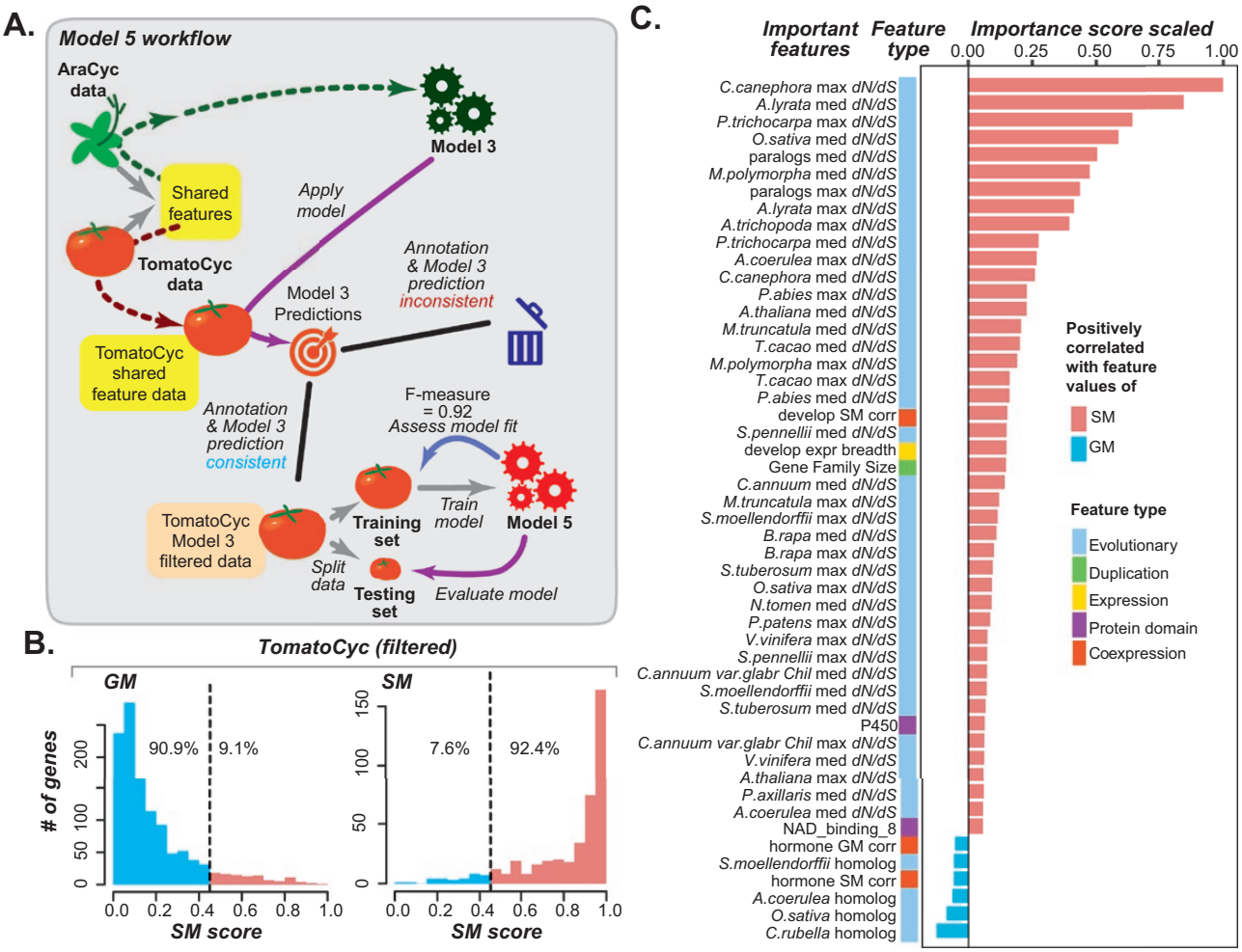

Figure 7. Model 5 performance and important features. (A) Diagram showing the procedures leading to Model 5 for predicting tomato GM/SM genes. (B) Distribution of Model 5 SM scores of tomato training set GM and SM genes. Dotted line: same as in Fig. 2. Blue and red: SM and GM genes, respectively. (C) Feature importance values for Model 5. Blue: importance scores normalized to between −1 and 0 for top features positively correlated with GM gene feature values (more negative is more important). Red: importance scores normalized between 0 and 1 for top features positively correlation with SM gene feature values (more positive is more important).

Additional models (Models 7 and 8) were trained using the same filtered gene set used in training Model 5 but with the full tomato feature data set (instead of just the shared features used in Models 3–5; **see Supporting Information—Fig. S9A**). The training set for Model 8 also included the benchmark gene annotations. Models 7 and 8 had similar performances (*F*-measure = 0.88 and 0.86, respectively; **see Supporting Information—Table S4**). Both Models 7 and 8 were significantly improved compared with Model 1 (*F*-measure = 0.74) and are similar to Model 5 in predictions of the training, testing, benchmark genes and removed gene set [**see Supporting Information—Fig. S9B–E**]. Overall, using Arabidopsis Model 3 to remove potentially mis-annotated tomato genes, i.e. genes that were not good training examples, led to substantially improved models (Models 5 and 7).

While TomatoCyc provides annotations for many genes in SM pathways, the global SM gene content in tomato is unknown. To provide a genome-wide estimate of SM gene content in the tomato genome, we used Model 7 to classify 5627 unannotated enzyme genes and found that 2865 are likely involved in SM pathways [**see Supporting Information—Fig. S9F**]. This indicates that substantially more SM genes are yet to be identified because only 696 genes are currently annotated in TomatoCyc. As noted earlier, each enzyme gene has an SM score from the model application, which can be interpreted as the probability that a gene is an SM gene (**see Supporting Information—Table S5** for scores for each gene); thus, those unannotated enzymes that are highly likely to be an SM gene can be prioritized for further investigation.

3.7 Relationships between improved performance and feature rankings

Models 5 and 7 substantially improved gene predictions in tomato compared with all other models because mis-annotated genes, mostly genes annotated as SM but predicted as GM by Arabidopsis Model 3, were removed from the training data. To better understand the reasons for the improvement in GM gene predictions, we looked into three examples where Models 5 and 7 predicted manually curated GM benchmark genes as GM genes, but where tomato-based Models 1 and 4 predicted the genes as SM genes: *1-aminocyclopropane-1-carboxylate oxidase 1* (*LeACO1*, NP_001234024), *abscisic acid 8′-hydroxylase* (*CYP707A1*, NP_001234517) and the cytochrome P450 *SlKLUH* (XP_004236064). In these cases, the mis-predictions were likely due to gene expression-related features. While *LeACO1* exhibited a maximum $\log_2$ fold change of 7.0 based on the fruit ripening data set [**see Supporting Information—Dataset S1**], which is consistent with the higher values observed for SM genes (median = 1.9) than for GM genes (1.2, P = 1.3e-15). Similarly, the variance of $\log_2$ fold change in expression during fruit ripening for *SlKLUH* is 2.5, which is consistent with significantly higher median variance for SM genes (1.5) compared with GM genes (1.0, P = 1.9e-21). *CYP707A1* is upregulated under many developmental conditions (13), which is not typical for tomato GM genes (SM median = 16, GM median = 9, P = 9.3e-26). Additionally, the expression of *LeACO1, CYP707A1* and *SlKLUH* correlates highly with that of other SM genes (PCC = 0.87, 0.63 and 0.83, respectively). The similarity of these expression feature values as those of SM genes likely contributed to their mis-prediction by Models 1 and 4.

Importantly, Models 5 and 7 likely predict these three genes correctly as GM genes because of the reduced reliance of these models on features associated with gene expression. Models 1 and 7 both use the full feature set, but filtered training data were used to train Model 7. In Model 1, expression variance in fruit ripening was ranked 46 among important features, while in Model 7 it was ranked 120 [**see Supporting Information—Table S6**]. Similarly, when comparing Models 4 and 5, which both use the shared feature set but differ in whether filtered training data were used, the features expression breadth under development and expression correlation between SM genes were ranked higher for Model 4 (6 and 16, respectively) than for Model 5 (22 and 20, respectively) [**see Supporting Information—Table S6**]. Model improvement is also due to higher ranking of evolutionary features, such as maximum *dN/dS* between tomato genes and *C. canephora* homologs, median *dN/dS* between tomato genes and homologs in *A. lyrata*, and maximum *dN/dS* between tomato genes and homologs in *P. trichocarpa*. In Model 5 these features were ranked 1, 2 and 3, respectively; in Model 4 they were ranked 2, 3 and 8, respectively; **see Supporting Information—Table S6**; in Model 7 they were ranked 1, 2 and 7, respectively; and in Model 1 they were ranked 2, 9 and 16, respectively; **see Supporting Information—Table S6**. *LeACO1* and *CYP707A1* both have maximum *dN/dS* values from comparisons to *C. canephora* homologs (0.07) more similar to those of GM genes (median = 0.10) than to SM genes (0.17). Similarly, *SlKLUH* has a maximum *dN/dS* value from comparisons to *A. lyrata* of 0.11, which is closer to the GM median (0.09) than to the SM median (0.15). Because in Models 5 and 7 these *dN/dS* features were weighted more heavily and certain expression features were weighted less heavily, the *dN/dS* feature values contributed to their correct classification as GM genes.

In addition to the features discussed thus far, we also found that gene family size was no longer the most important feature in Models 5 and 7, ranked 24 and 27, respectively, as it was Models 1, 3 and 4. Considering that some of the largest enzyme families—such as cytochrome P450 and terpene synthases—contain both SM and GM genes, this reduced importance likely contributed to improved predictions. Despite the improvement, Models 5 and 7 are by no means perfect and erroneous predictions still occur. For example, *PSY1* is a fruit ripening-related gene manually curated as an SM benchmark gene, but it was predicted as a GM gene by both Models 4 and 5. *PSY1* represents an unusual case of duplication-associated subfunctionalization and is specifically expressed in chromoplast-containing tissues such as ripening fruits and petals (Fray and Grierson 1993). *PSY1* has comparatively low *dN/dS* values (similar to GM genes), especially between tomato and *C. canephora* (maximum *dN/dS* = 0.06). Because this *dN/dS* feature was the most important feature for Model 5, this ultimately contributed to the mis-prediction of *PSY1* as a GM gene.

Other examples are two GM terpene synthases involved in the biosynthesis of gibberellin, a plant hormone (Yamaguchi 2008): *copalyl diphosphate synthase* (*CPS*, NP_001234008) and *kaurene synthase* (*KS*, XP_004243964). Both *CPS* and *KS* are mis-predicted as SM genes in all models, presumably because of their high *dN/dS* values from comparisons to homologs in several species (*CPS* median *dN/dS* = 0.20, *KS* median *dN/dS* = 0.26). These two enzymes were derived from an ancestral dual functional enzyme containing both copalyl diphosphate synthase and kaurene synthase activities (Chen *et al.* 2011). Angiosperm terpene synthases seem to have lost one activity or the other, but the ancient timing of the CPS/KS duplication (after divergence between bryophytes and the other land plant lineages) makes the high rate of evolution unusual. It is unknown what effect the loss of activity has on the evolution of the terpene synthase sequence. For all three genes, *PSY1*, *CPS* and *KS*, the atypical evolutionary rates, either unusually low or high, led to mis-prediction. Overall, our machine learning approach led to a highly accurate SM/GM model with an *F*-measure of 0.91 (where a value of 1 indicates a perfect model). However, while our approach ensures the identification of typical SM/GM genes, SM/GM genes with atypical properties that defy the general trend still are likely mis-predicted.

4. DISCUSSION

Many SM genes are unknown due to the vast number of specialized metabolites are limited to specific species and SM and GM genes are difficult to distinguish because SM genes are often derived from GM genes. Additionally, many specialized metabolites of interest are found in medicinal plants or crops that are not well annotated. If data from a better annotated species such as Arabidopsis could be used, directly or indirectly, to make cross-species predictions in another species, such as tomato, this could greatly improve annotations in non-model species. Here we used machine learning to establish models for classifying genes with SM and GM functions in tomato, but consistent with the

lower quality of the tomato annotation, these models established using tomato features had relatively poor performance compared with models built in Arabidopsis. We also found that a substantial number of important features and predictions differed between the models based on Arabidopsis (Model 3) and tomato (Model 4). We discovered that the differences in feature importance and model performance were likely the result of mis-annotation of some tomato genes, which contributed negatively to the performance of machine learning models. Therefore, we attempted to perform cross-species knowledge transfer by using a machine learning approach called transfer learning (Torrey and Shavlik 2010), where knowledge learned from a previously trained model (e.g. our Arabidopsis Model 3) is used (in this case, to remove predictions inconsistent with annotations) to train another model (e.g. tomato Model 5).

By filtering out TomatoCyc-annotated genes that had predictions opposite from those of the Arabidopsis-based Model 3 from the training data, we significantly improved the accuracy of tomato SM/GM gene predictions. We should emphasize that, for building the tomato Model 5, the testing genes were filtered similarly to the training set. This was intentional for testing whether removing genes with the Arabidopsis Model 3 would lead to improvement or not. If the Arabidopsis model was helpful in filtering out mislabelled tomato genes, we would expect the remaining tomato SM/GM genes to be better separated and lead to an improved tomato model. This separation would have been artificial if we used the Arabidopsis model to filter out genes and then applied the Arabidopsis model to classify tomato genes again. The improvement seen would be completely circular. Instead of the above, we established a new tomato model based on filtered SM/GM tomato annotations. Another important reason why the performance is not artificially inflated by removing genes based on Model 3 is because, if the Arabidopsis model was not effective in filtering out mislabelled data, we expected that the resulting tomato model would not improve. Compared to Model 6 that had genes randomly removed and did not see an improvement in the overall score, the model based on Arabidopsis filtered data fared much better. This improvement could also be due to the removal of hard-to-predict edge cases. We showed that filtered GM genes and, to a lesser extent, filtered SM genes have similar SM scores as remaining SM and GM genes, respectively. Thus, we demonstrated that this improvement is not simply due to the removal of hard-to-predict cases and would not have been possible without informed removal of potentially mis-annotated data. This approach can be applied more generally to any problem in a species that is relatively information-poor by transferring knowledge from an information-rich one. Using transfer learning we may also be able to better annotate less well-studied species.

It is important to note that a limitation of the transfer learning approach is that it is only useful for transferring knowledge, mechanisms or phenomena that are similar across species. In our study, the transfer learning approach worked well by identifying GM genes conserved across species, and likely by default impacted the prediction of SM genes. This is likely because SM pathways are by definition specialized—what you learn in one species does not necessarily apply to another. A specific example of where transfer learning can suffer is in predicting genes with atypical properties. The machine learning approach excels at spotting patterns in data, and the performance of machine learning models improves as more high-quality instances (e.g. experimentally validated SM/GM genes) and more informative features (e.g. *dN/dS*) are incorporated. However, it is a challenge to generate high-quality instances, and expert knowledge dictates what kinds of features are incorporated. In addition, the representation of genes that are considered 'atypical' in the model can be limited by our ability to scour the literature for novel features to represent these genes. The inability to apply a general model like this one to atypical genes could include genes which encode enzymes but do not have direct roles in plant metabolism, such as regulatory genes like transcription factors or kinases, or ambiguous SM genes that may have roles in GM processes. For these types of genes, new training sets must be obtained in order to build a model that could correctly predict them.

In future studies, transfer learning can be used to predict GM genes and, to a lesser extent, SM genes in species that lack annotations and/or experimental evidence such as non-model, medicinal plant species. An open question in this area that needs to be addressed is whether more closely related species, even though they may not be as well annotated, are better candidates for transfer learning than better annotated but more distantly related species. In addition, as discussed above, our models can potentially be further improved by incorporating additional features, particularly those that are shared between species, using transfer learning. For example, data that are incorporated as features for across species models should come from experiments performed in more similar ways in terms of treatments applied and tissues investigated. Furthermore, we found that SM gene annotations can vary across species, so reliance on information from a particular species may skew the model predictions and the features that are most important for the model. Thus, in future studies comparisons between models using data from single and multiple species can potentially inform further efforts to improve cross-species predictions via transfer learning. Another consideration is that we treated our research problem as a binary (SM or GM) classification problem. Over the course of evolution, SM pathways may branch off from GM pathways or some SM pathways may ultimately become GM pathways because of increasingly wider taxonomic distribution. Thus, the extent to which a gene is considered to be SM is likely continuous, where genes at the end of an SM pathway may be more 'SM-like' than genes at the beginning of the pathway, which may be linked to GM pathways. The question is how to define the degree of involvement of a gene in SM pathways and determine whether continuous SM scores, where GM and SM genes have low and high scores, respectively, are good proxies for involvement in these pathways. This can be accomplished by mapping SM scores to pathways to see if they are predictive of where a gene lies in a pathway.

SUPPORTING INFORMATION

The following additional information is available in the online version of this article—

Figure S1. Speciation nodes. Phylogenetic tree of 26 species showing speciation nodes (N0–N24). Most recent gene duplication node in text refers to the speciation node where gene was last duplicated.

Figure S2. Comparison of all model scores and feature importance values for Model 1. (A) Comparison of model scores. *F*-measure is shown

on the *y*-axis and model is shown on the *x*-axis. Model type is denoted by colour. Gray indicates Models 1–8 variants (i.e. different ML algorithms and/or numbers of features used) that are not described in the text. RF: Random Forest. SVM: Support Vector Machine. featsel25-1000: features selected, sets of 25–1000. For model names, **see Supporting Information—Table S4**. (B) Bar plot of the top 50 most important features for Model 1. The importance score is on the *y*-axis and all scores are normalized to the score of the most important feature, which was set as 1. Red bars represent features that are enriched for specialized metabolism (SM) genes while the blue bars represent features enriched for general metabolism (GM) genes. Features are listed along the *x*-axis, with the colour denoting the feature category.

Figure S3. Important features for Model 1. (A–K) Distributions or bar plots of feature values for TomatoCyc-annotated specialized metabolism (SM) and general metabolism (GM) genes. (A–J) Significance determined by the Mann–Whitney *U*-test. (A–H) Distributions of the maximum or median *dN/dS* value for a given gene relative to their homolog in *P. patens*, *S. moellendorffii*, *A. trichopoda*, *O. sativa*, *B. rapa*, *A. coerulea*, *P. trichocarpa* and *S. pennellii*. (I and J) Distributions of log 10 of median FPKM values for the Inflorescence data set and Root data set. (K) Percent of genes with a given Pfam domain. Overrepresentation (+) and underrepresentation (−) was determined using those genes with a *P*-value less than 0.05 from a Fisher's exact test between SM and GM genes with Benjamin–Hochberg multiple testing correction.

Figure S4. Features important for specialized metabolism (SM) vs. general metabolism (GM) predictions. For all distributions of each predicted class, GM→GM represents GM genes predicted by Model 1 as GM, GM→SM represents GM genes predicted by Model 1 as SM, SM→GM represents SM genes predicted by Model 1 as GM and SM→SM represents SM genes predicted by Model 1 as SM. Significant differences between continuous variables were determined by the Kruskal–Wallis test (A–J) and *post hoc* comparisons were made using Dunn's test. Different letters indicate statistically significant differences between groups ($P < 0.05$). For binary data (*K*), overrepresentation (+) and underrepresentation (−) were determined by the Fisher's exact test where (+) is significant overrepresentation of a predicted class and (−) is significant underrepresentation. A *P*-value < 0.05 after Benjamin–Hochberg multiple testing correction was considered significant. (A–H) Distributions of the maximum or median *dN/dS* value for a given gene from comparisons to its homolog in *P. patens*, *S. moellendorffii*, *A. trichopoda*, *O. sativa*, *B. rapa*, *A. coerulea*, *P. trichocarpa* and *S. pennellii*. (I and J) Distributions of $\log_{10}$ (median FPKM) values for the Inflorescence (I) and Root (J) data sets. (K) Percentage of genes with a given Pfam domain.

Figure S5. Specialized metabolism (SM) likelihood scores for manually annotated genes. (A–C) Bar plots showing the percentage of manually annotated benchmark genes predicted as SM or general metabolism (GM). The original annotation from TomatoCyc is shown first, followed by the benchmark annotation and then the prediction. (A) Predictions for Model 1, (B) predictions for Model 3, and (C) predictions for Model 4.

Figure S6. *Solanum lycopersicum* and *A. thaliana* model comparison and model performance. (A and B) Comparison of the specialized metabolism (SM) score distributions for tomato Model 4 (*y*-axis) and Arabidopsis Model 3 (*x*-axis). Support Vector Machine (SVM) and a shared feature set were used for both models. Density of data points ranges from high (yellow) to medium (blue-purple) to low (white). (A) SM scores for general metabolism (GM) genes; (B) SM scores for SM genes; (C–F) feature distributions for annotated SM and GM genes that are predicted as SM or GM genes by Arabidopsis Model 3 and tomato Model 4. The *x*-axis lists the annotations for each group of genes predicted using Arabidopsis Model 3 and tomato Model 4. *P*-values are from the Kruskal–Wallis test and *post hoc* comparisons were made using the Dunn's test. Different letters indicate statistically significant differences between groups ($P < 0.05$). (C) Maximum Pearson's Correlation Coefficient (PCC) between a given gene and all other SM genes under stress conditions; (D) maximum PCC between a given gene and all other SM genes during development; (E) maximum PCC between a given gene and all other GM genes under hormone treatment; (F) normalized median *dN/dS* values between tomato or Arabidopsis genes and their homologs in *O. sativa*.

Figure S7. Benchmark and test set predictions from Model 5 (*A. thaliana* mis-predictions removed). Plots A and B show distributions of specialized metabolism (SM) likelihood scores. (A) Model 5 test set SM and general metabolism (GM) genes, which were held out from the model building process completely. (B) TomatoCyc SM and GM genes with annotations opposite to Arabidopsis Model 3 predictions removed from the filtered training set. For plots (A and B): SM likelihood score is shown on the *x*-axis, number of genes is shown on the *y*-axis. Prediction threshold, based on the score with the highest *F*-measure, is indicated by the dotted line, and predicted SM genes are shown to the right of the line in red while predicted GM genes are shown to the left of the line in blue. (C) Bar plots showing the percentage of manually annotated benchmark genes predicted as SM or GM by Model 5. For the first bar plot, the original annotation from TomatoCyc is shown first, followed by the benchmark annotation and then the prediction. The second bar plot shows overall benchmark predictions (not divided by TomatoCyc annotations).

Figure S8. Arabidopsis Model 3 and Tomato Model 5 comparison. Plots A–D show gene scores from Arabidopsis Model 3 on the *x*-axis and Tomato Model 5 on the *y*-axis. Colour: data point density ranges from high (yellow) to medium (purple), to low (fading purple). (A) Filtered general metabolism (GM) genes—removed from Model 5 training; (B) remaining specialized metabolism (SM) genes—kept in Model 5 training; (C) filtered SM genes—removed from Model 5 training; (D) remaining GM genes—kept in Model 5 training.

Figure S9. Benchmark and test set predictions from Model 7 (*A. thaliana* mis-predictions removed). (A) Schematic diagram showing the application of tomato Model 7 to tomato. The full tomato feature data set was used to build a binary model using TomatoCyc specialized metabolism (SM) and general metabolism (GM) annotations after removing genes mis-predicted by Arabidopsis Model 3. The model was then applied to tomato genes. (B) TomatoCyc filtered training set SM and GM genes from tomato Model 7. (C) Model 7 test set: SM and GM genes, which were held out completely from the tomato Model 7 building process. (D) Bar plot showing the percentage of manually annotated benchmark genes predicted as SM or GM by Model 7. The original annotation from TomatoCyc is shown first, followed by the

benchmark annotation and then the prediction. (E) SM and GM genes removed from Model 7 training set. (F) Unannotated tomato enzymes. For plots (B, C and E, F): SM likelihood score is shown on the *x*-axis, number of genes is on the *y*-axis. Prediction threshold, based on the score with the highest *F*-measure, is indicated by the dotted line, and predicted SM genes are shown to the right of the line in red while predicted GM genes are shown to the left of the line in blue.

Table S1. Tomato gene annotation information. Annotation information based on TomatoCyc and manual annotation.

Table S2. Feature Statistics. Statistics for original and shared features.

Table S3. Transcriptome studies. Information about all expression data sets used in the models.

Table S4. Model scores. Scores and information for all models.

Table S5. Specialized metabolism (SM) gene scores. Specialized metabolism prediction scores for all genes for each of the models.

Table S6. Feature Importance. Feature importance scores for all models discussed in the text.

Dataset S1. Original features. Data set includes all of the features used for Models 1, 2, 7 and 8.

Dataset S2. Shared features. Data set includes all of the shared features between Arabidopsis and tomato used for Models 3, 4 and 5.

DATA AVAILABILITY

Data and materials availability are as noted in the Methods section.

ACKNOWLEDGEMENTS

We thank the Shiu lab, Last lab and Barry lab members for their help during the preparation of this manuscript.

SOURCES OF FUNDING

This work was supported by NSF grant IOS-1546617 to R.L., C.S.B. and S.-H.S.; National Institute of General Medical Sciences of the National Institutes of Health graduate training grant T32-GM110523 to B.L.; a postdoctoral fellowship from the National Science Foundation (NSF) IOS-1811055 to C.A.S.; U.S. Department of Energy Great Lakes Bioenergy Research Center (BER DE-SC0018409) grant to R.L. and S.-H.S.; Michigan AgBioResearch and U.S. Department of Agriculture National Institute of Food and Agriculture Hatch project number MICL02552 to C.S.B; and NSF grant DEB-1655386 to S.-H.S.

CONTRIBUTIONS BY THE AUTHORS

B.M.M. and S.-H.S. conceived the study; all authors contributed to the design of the study; B.M.M., P.W. and A.L. implemented the computational analysis; P.F., B.L., Y.-R.L., C.A.S., K.S. and C.S.B. contributed to manual annotation; B.M.M., P.W., A.L. and M.D.L. contributed to interpretation of computational analysis results, all authors contributed to interpreting annotation results; all authors contributed to the manuscript draft.

CONFLICT OF INTEREST

None declared.

LITERATURE CITED

Adio AM, Casteel CL, De Vos M, Kim JH, Joshi V, Li B, Juéry C, Daron J, Kliebenstein DJ, Jander G. 2011. Biosynthesis and defensive function of Nδ-acetylornithine, a jasmonate-induced *Arabidopsis* metabolite. *The Plant Cell* **23**:3303–3318.

Ament K, Van Schie CC, Bouwmeester HJ, Haring MA, Schuurink RC. 2006. Induction of a leaf specific geranylgeranyl pyrophosphate synthase and emission of (E,E)-4,8,12-trimethyltrideca-1,3,7,11-tetraene in tomato are dependent on both jasmonic acid and salicylic acid signaling pathways. *Planta* **224**:1197–1208.

Anders S, Pyl PT, Huber W. 2015. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**:166–169.

Blum A, Monir M, Wirsansky I, Ben-Arzi S. 2005. The beneficial effects of tomatoes. *European Journal of Internal Medicine* **16**:402–404.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120.

Breiman L. 2001. Random Forests. *Machine Learning* **45**:5–32.

Capasso R, Izzo AA, Pinto L, Bifulco T, Vitobello C, Mascolo N. 2000. Phytotherapy and quality of herbal medicines. *Fitoterapia* **71**:S58–S65.

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

Chakrabarti M, Zhang N, Sauvage C, Muños S, Blanca J, Cañizares J, Diez MJ, Schneider R, Mazourek M, McClead J, Causse M, van der Knaap E. 2013. A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proceedings of the National Academy of Sciences of the United States of America* **110**:17125–17130.

Chen F, Tholl D, Bohlmann J, Pichersky E. 2011. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *The Plant Journal* **66**:212–229.

Clifford M, Brown J. 2006. *Flavonoids: chemistry, biochemistry, and applications*. Boca Raton, FL: CRC, Taylor & Francis.

Cunningham FX, Gantt E. 1998. Genes and enzymes of carotenoid biosynthesis in plants. *Annual Review of Plant Physiology and Plant Molecular Biology* **49**:557–583.

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

Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, Bernstein DA, Rolfe PA, Heisler LE, Chin B, Nislow C, Giaever G, Phillips PC, Fink GR, Gifford DK, Boone C. 2010. Genotype to phenotype: a complex problem. *Science* **328**:469.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, Hofberger JA, Smithson A, Hall JC, Blanchette M, Bureau TE, Wright SI, dePamphilis CW, Schranz ME, Barker MS, Conant GC, Wahlberg N, Vogel H, Pires JC, Wheat CW. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences of the United States of America* **112**:8362–8366.

Ehrlich PR, Raven PH. 1964. Butterflies and plants: a study in coevolution. *Evolution* **18**:586.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**:157.

Facchini PJ, Bohlmann J, Covello PS, De Luca V, Mahadevan R, Page JE, Ro DK, Sensen CW, Storms R, Martin VJ. 2012. Synthetic biosystems for the production of high-value plant metabolites. *Trends in Biotechnology* **30**:127–131.

Fan P, Leong BJ, Last RL. 2019. Tip of the trichome: evolution of acylsugar metabolic diversity in Solanaceae. *Current Opinion in Plant Biology* **49**:8–16.

Fray RG, Grierson D. 1993. Identification and genetic analysis of normal and mutant phytoene synthase genes of tomato by sequencing, complementation and co-suppression. *Plant Molecular Biology* **22**:589–602.

Giovannucci E, Rimm EB, Liu Y, Stampfer MJ, Willett WC. 2002. A prospective study of tomato products, lycopene, and prostate cancer risk. *Journal of the National Cancer Institute* **94**:391–398.

Grynkiewicz G, Gadzikowska M. 2008. Tropane alkaloids as medicinally useful natural products and their synthetic derivatives as new drugs. *Pharmacological Reports* **60**:439–463.

Hartmann T. 2007. From waste products to ecochemicals: fifty years research of plant secondary metabolism. *Phytochemistry* **68**:2831–2846.

Isaacson T, Ronen G, Zamir D, Hirschberg J. 2002. Cloning of tangerine from tomato reveals a carotenoid isomerase essential for the production of beta-carotene and xanthophylls in plants. *The Plant Cell* **14**:333–342.

Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B, Cardenas PD, Bocobza SE, Unger T, Malitsky S, Finkers R, Tikunov Y, Bovy A, Chikate Y, Singh P, Rogachev I, Beekwilder J, Giri AP, Aharoni A. 2013. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341**:175–179.

Karp PD, Latendresse M, Caspi R. 2011. The pathway tools pathway prediction algorithm. *Standards in Genomic Sciences* **5**:424–429.

Ku HM, Vision T, Liu J, Tanksley SD. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proceedings of the National Academy of Sciences of the United States of America* **97**:9121–9126.

Louppe G. 2014. Understanding Random Forests: from theory to practice. *ArXiv*, ArXiv:14077502.

Lucini T, Resende JTV, Oliveira JRF, Scabeni CJ, Zeist AR, Resende NCV. 2016. Repellent effects of various cherry tomato accessions on the two-spotted spider mite *Tetranychus urticae* Koch (Acari: Tetranychidae). *Genetics and Molecular Research*. **16**. doi:10.4238/gmr.15017736.

Maciel GM, Almeida RS, da Rocha JPR, Andaló V, Marquez GR, Santos NC, Finzi RR. 2017. Mini tomato genotypes resistant to the silverleaf whitefly and to two-spotted spider mites. *Genetics and Molecular Research*. **16**. doi:10.4238/gmr16019539.

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**:4288–4297.

Meinke D, Muralla R, Sweeney C, Dickerman A. 2008. Identifying essential genes in *Arabidopsis thaliana*. *Trends in Plant Science* **13**:483–491.

Milo R, Last RL. 2012. Achieving diversity in the face of constraints: lessons from metabolism. *Science* **336**:1663–1667.

Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Lehti-Shiu MD, Last RL, Pichersky E, Shiu SH. 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.

Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**:2490–2492.

Nakashima T, Wada H, Morita S, Erra-Balsells R, Hiraoka K, Nonami H. 2016. Single-cell metabolite profiling of stalk and glandular cells of intact trichomes with internal electrode capillary pressure probe electrospray ionization mass spectrometry. *Analytical Chemistry* **88**:3049–3057.

Ning J, Moghe GD, Leong B, Kim J, Ofner I, Wang Z, Adams C, Jones AD, Zamir D, Last RL. 2015. A feedback-insensitive isopropylmalate synthase affects acylsugar composition in cultivated and wild tomato. *Plant Physiology* **169**:1821–1835.

Nohara T, Ikeda T, Fujiwara Y, Matsushita S, Noguchi E, Yoshimitsu H, Ono M. 2006. Physiological functions of solanaceous and tomato steroidal glycosides. *Journal of Natural Medicines* **61**:1–13.

Osbourn AE. 1996. Preformed antimicrobial compounds and plant defense against fungal attack. *The Plant Cell* **8**:1821–1831.

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

Piasecka A, Jedrzejczak-Rey N, Bednarek P. 2015. Secondary metabolites in plant innate immunity: conserved function of divergent chemicals. *The New Phytologist* **206**:948–964.

Pichersky E, Lewinsohn E. 2011. Convergent evolution in plant specialized metabolism. *Annual Review of Plant Biology* **62**:549–566.

Rajput H. 2014. Effects of *Atropa belladonna* as an anti-cholinergic. *Natural Products Chemistry and Research* **1**. doi:10.4172/2329-6836.1000104.

Romero I, Tikunov Y, Bovy A. 2011. Virus-induced gene silencing in detached tomatoes and biochemical effects of phytoene desaturase gene silencing. *Journal of Plant Physiology* **168**:1129–1135.

Rost B. 2002. Enzyme function less conserved than anticipated. *Journal of Molecular Biology* **318**:595–608.

Saito K, Hirai MY, Yonekura-Sakakibara K. 2008. Decoding genes with coexpression networks and metabolomics - 'majority report by precogs'. *Trends in Plant Science* **13**:36–43.

Schilmiller A, Shi F, Kim J, Charbonneau AL, Holmes D, Daniel Jones A, Last RL. 2010. Mass spectrometry screening reveals widespread diversity in trichome specialized metabolites of tomato chromosomal substitution lines. *The Plant Journal* **62**:391–403.

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

Schmidt BM, Ribnicky DM, Lipsky PE, Raskin I. 2007. Revisiting the ancient concept of botanical therapeutics. *Nature Chemical Biology* **3**:360–366.

Schmidt B, Ribnicky DM, Poulev A, Logendra S, Cefalu WT, Raskin I. 2008. A natural history of botanical therapeutics. *Metabolism: Clinical and Experimental* **57**:S3–S9.

Tohge T, Alseekh S, Fernie AR. 2013. On the regulation and function of secondary metabolism during fruit development and ripening. *Journal of Experimental Botany* **65**:4599–4611.

Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J, Awazuhara M, Inoue E, Takahashi H, Goodenowe DB, Kitayama M, Noji M, Yamazaki M, Saito K. 2005. Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. *The Plant Journal* **42**:218–235.

Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**:635–641.

Torrey L, Shavlik J. 2010. *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. Hershey, PA: Information Science Reference.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**:1105–1111.

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.

Wang Y, Li J, Paterson AH. 2013. MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. *Bioinformatics* **29**:1458–1460.

Wang P, Moore BM, Panchy NL, Meng F, Lehti-Shiu MD, Shiu SH. 2018. Factors influencing gene family size variation among related species in a plant family, Solanaceae. *Genome Biology and Evolution* **10**:2596–2613.

Wink M. 1988. Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. *Theoretical and Applied Genetics* **75**:225–233.

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. *The Plant Cell* **29**:944–959.

Xu J, Ranc N, Muños S, Rolland S, Bouchet JP, Desplat N, Le Paslier MC, Liang Y, Brunel D, Causse M. 2013. Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. *Theoretical and Applied Genetics* **126**:567–581.

Xu B, Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**:1586–1591.

Yamaguchi S. 2008. Gibberellin metabolism and its regulation. *Annual Review of Plant Biology* **59**:225–251.

Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. 2004. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Research* **14**:1107–1118.

Yu G, Nguyen TT, Guo Y, Schauvinhold I, Auldridge ME, Bhuiyan N, Ben-Israel I, Iijima Y, Fridman E, Noel JP, Pichersky E. 2010. Enzymatic functions of wild tomato methylketone synthases 1 and 2. *Plant Physiology* **154**:67–77.