Robust predictions of specialized metabolism genes through machine learning

Bethany Moore[1,2], Peipei Wang[1], Pengxiang Fan[3], Bryan Leong[1], Craig A. Schenck[3], John P. Lloyd[1], Melissa Lehti-Shiu[1], Robert Last[1,3], Eran Pichersky[4], and Shin-Han Shiu[1,2*]

[1]Department of Plant Biology, [2]Ecology, Evolutionary Biology, and Behavior program, [3]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA, [4]Department of Molecular, Cellular and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, USA

*Corresponding Authors:
Shin-Han Shiu
Michigan State University
Plant Biology Laboratories
612 Wilson Road, Room 166
East Lansing, MI 48824-1312
517-353-7196

Abstract

Plant specialized metabolism (SM) enzymes produce lineage-specific metabolites with important ecological, evolutionary, and biotechnological implications. Using *Arabidopsis thaliana* as a model, we identified distinguishing characteristics of SM and GM (general metabolism, traditionally referred to as primary) genes through a detailed study of features including duplication patterns, sequence conservation, transcription, protein domain, and gene network properties. Study of benchmark genes revealed that SM genes tend to be tandemly duplicated, co-expressed with their paralogs, narrowly expressed at lower levels, less conserved, and less well connected in gene networks relative to GM genes. Although the values of each of these features significantly differed between SM and GM genes, any single feature was ineffective at predicting SM from GM genes. Using machine learning methods to integrate all features, a well performing prediction model was established with a true positive rate of 0.84 and a false positive rate of 0.23. In addition, 82% of known SM genes not used to create the machine learning model were predicted as SM genes, further demonstrating its accuracy. Application of the prediction model led to the identification of 1,817 *A. thaliana* genes with high confidence of being SM genes, providing a global estimate of SM gene content in a plant genome.

Significance

Specialized metabolites are critical for plant-environment interactions, e.g. attracting pollinators or defending against herbivores, and are important sources of plant-based pharmaceuticals. However, it is unclear what proportion of enzyme genes play roles in specialized metabolism (SM) as opposed to general metabolism (GM) in any plant species. This is because of the diversity of specialized metabolites and the considerable number of incomplete pathways responsible for their production. In addition. SM gene ancestors frequently played roles in GM. Our study evaluates features distinguishing SM and GM genes for building a computational model which accurately predicts SM genes. Our predictions provide candidates for experimental studies, and our modeling approach can be applied to other species that produce medicinally or industrially useful compounds.

Key words

Specialized metabolism, machine learning, predictive biology, data integration

\body

Introduction

Gene duplication and subsequent divergence/loss events led to highly variable gene content between plant species (1, 2). This high rate of differential gain and loss events have led to a repertoire of metabolic enzymes ranging from those involved in generally conserved, primary metabolic processes found in most species, such as carbohydrate metabolism or photosynthesis (referred to as general metabolism, or GM, genes), to those that function in lineage-specific specialized metabolism (SM) (3–6). The proliferation of lineage-specific SM genes in plants resulted in an overall far larger number of specialized than general metabolites. Specialized metabolites are important for niche-specific interactions between plants and environmental agents that can be harmful (e.g. herbivores) or beneficial (e.g. pollinators) (3, 6–8). In addition, specialized metabolites are the basis for thousands of plant-derived chemicals, many of which are used for medicinal and/or nutritional purposes, such as carotenoid derivatives with antioxidant properties or vitamins in tomato (9–11). Thus, identification of the genes led to enzymes that produce SMs (referred to as SM genes) is key to understanding the causes underlying the diversity of plant specialized metabolites as well as for the engineering of plant-derived chemicals and pharmaceuticals.

Despite their importance, most plant metabolites and the enzymes and genes involved in their biosynthesis are yet to be identified (12). Although most SM genes arise by duplication of GM genes (3, 5, 12–14), duplication itself is not sufficient for pinpointing SM genes for four reasons. First, genes encoding GM or SM enzymes can belong to the same family, Second, duplicated GM genes may not necessarily become specialized (1), and minor sequence changes can lead to substantially altered enzyme functions (15, 16). Third, SM genes may arise through lineage-specific loss of the GM function without duplication. Finally, convergent evolution may contribute to unrelated enzymes in different lineages that use the same substrate to make similar products (5). Consequently, despite the presence of a diverse repertoire of enzymes in many plant species, it remains unresolved whether most plant enzyme genes are involved in GM or SM pathways, including the most well annotated plant species, *Arabidopsis thaliana* (3, 5, 13, 14). Therefore, in recent years there has been a renewed focus on identifying SM genes (17, 18).

Despite the challenges, aside from duplication status, multiple properties may be useful in distinguishing SM from GM genes, including restricted phylogenetic distribution, a higher family expansion rate, tandem clustering of paralogs, a propensity for genomic clustering (close physical proximity of genes encoding enzymes in the same pathways), and higher degrees of co-expression (4, 17) compared to GM genes. Recent, pioneering studies have used co-expression with known SM genes (17, 19) or genomic neighborhood and gene-metabolite correlation (20) to predict SM genes. Nonetheless, the association of these properties with SM genes is far from absolute (21). To overcome this limitation, an approach jointly considering sequence similarity, co-expression, inferred enzyme activity, and genomic clustering can be used (18). We developed a machine learning approach to identify SM genes from *A. thaliana* using multiple data types including those that were not considered previously (i.e. duplication type, evolutionary patterns, epigenetic marks, and protein-protein interactions). We first determined the extent to which diverse properties differ between SM and GM genes. Next, we established machine learning models to make global predictions of SM genes as well as predictions of SM genes that belong to glucosinolate pathways and terpene synthase, P450, and methyltransferase families.

Results and Discussion

Benchmark SM and GM genes

Currently there are two sources for plant SM and GM gene annotations: Gene Ontology (GO; (22)) and AraCyc (23). For SM genes, we started with the 357 genes with the GO term secondary metabolic process, and 649 enzyme-encoding genes in 129 AraCyc secondary metabolism pathways (**Dataset S1**). Initial GM genes included 2,009 with the GO term primary metabolic process and 1,557 enzyme-encoding genes in 490 AraCyc non-secondary metabolism pathways (**Dataset S1**). Although 32.4% of GO- and 41.8% of AraCyc-annotated GM genes overlapped, only 24 SM genes (6.7% of GO- and 3.7% of AraCyc-annotated SM genes) overlapped (**Figure 1A**). While this is a significantly higher degree of overlap than expected by chance (**Figure S1A, B)**, it indicates a greater inconsistency in SM annotation criteria between these two datasets compared with GM annotation criteria. As a result, GO- and AraCyc-annotated SM genes have different GO-term (function), AraCyc (pathway), and Pfam (24) protein domain annotations (**Figure 1B**, **Figures S1C,D, Dataset S1,S2**).

For example, GO-annotated SM genes tend to be overrepresented in lignin, coumarin and phenylpropanoid biosynthesis categories. In contrast, AraCyc-annotated SM genes are overrepresented in anthocyanin and flavonoid biosynthetic processes. The only commonly enriched functional category is related to glucosinolate biosynthesis. With regard to pathway annotation, GO-annotated SM genes are overrepresented in, for example, biosynthesis of flavonoids, leucine, suberin monomers and wax. In contrast, AraCyc-annotated SM genes are overrepresented in the terpenoid, camalexin, carotenoid, farnesene, and glucosinolate biosynthesis pathways (**Figure S1C**). The only commonly enriched pathway is flavonoid biosynthesis. Importantly, some SM gene enriched pathways and GO terms, such as leucine, suberin, wax and carotene biosynthesis, are found across all major land plant lineages and genes with these annotations should be considered GM genes. In contrast to SM genes, GO- and AraCyc-annotated GM genes tend be over-represented in the same functional categories and pathways (**Figure 1B**). To include as many annotated SM genes as possible, we defined 392 "benchmark" SM genes as the union of GO and AraCyc SM annotations with Enzyme Commission (EC) numbers (**Dataset S1**). Similarly, 2,226 benchmark GM genes are from the union of GO and AraCyc primary metabolism gene annotations with EC numbers (**Dataset S1**).

Differences in gene expression and epigenetic marks between SM and GM genes

To assess differences in expression of SM and GM genes, we examined transcriptome datasets encompassing 25 tissue types (developmental dataset) and 16 abiotic/biotic stress conditions (stress dataset, see **Methods**). We found that SM genes had significantly lower median expression levels (Mann Whitney U, *p*=2e-24, **Figure 2A**), lower maximum expression levels (*p*=0.04, **Figure 2A**), and narrower breadth of expression (*p*=1.2e-35, **Figure 2A**) than GM genes, but no significant difference in expression variability (*p*=0.26, **Figure 2A**). These differences are consistent with the observation that SM genes have more specialized roles, whereas GM genes are involved in basic cellular functions (3, 6). SM genes tend to be up-regulated under a higher number of shoot (*p*=2.1e-7) and root (*p*=9.1e-15) abiotic (**Figure 2B**) and shoot biotic (*p*=1.8e-8), **Figure 2B**) conditions than GM genes, consistent with the roles of some specialized metabolites in environmental interactions (20). Relatively fewer SM genes were down-regulated under stress compared with GM genes (**Figure 2B**), likely reflecting a growth-defense tradeoff (25) where GMs involved in house-keeping functions are down-regulated under stress and SM genes with roles in abiotic and biotic interactions are not. In addition to expression profiles, we compared the numbers of CG methylated and histone modification sites, both of which can influence gene expression (26, 27). We found that SM genes tend to have a lower degree of gene body CG-methylation (**Dataset S2**, see Methods) than GM genes (Fisher's exact test, *p*=2.81e-5). On the other hand, the extent of histone modification does not significantly differ between SM and GM genes for seven of the eight histone marks (see Methods, **Figure S2A**).

To explore general differences in expression patterns between SM and GM genes, we evaluated expression correlation between each SM/GM gene and its paralogs (in the form of maximum Pearson's Correlation Coefficient) in each of the four expression datasets (abiotic stress, biotic stress, development, and hormone treatment). Because SM genes tend to have undergone more recent expansion (2, 4), SM paralogs had a significantly higher expression correlation in all four data sets (Mann-Whitney U test, all *p*<0.05, **Figure 2C**). We next looked at the maximum expression correlation between each SM gene to other SM genes (SM-SM) or to GM genes (SM→GM), as well as between each GM gene to other GM genes (GM-GM) or to SM genes (GM-SM). The expression correlations follow that: GM-GM > SM-GM > SM-SM > GM-SM (all *p*<0.05, **Figure 2D**). The higher expression correlation in GM-GM compared to SM-SM may be due to the more recent duplication of SM genes relative to GM genes. Taken together, our findings indicate that expression correlations can be features distinguishing SM and GM genes.

To further assess if co-expression patterns can also be used to distinguish SM and GM genes, we used six algorithms to define co-expression modules based on expression in three expression datasets (**Figure 2E, Dataset S2**). Among these modules, 99 and 125 contained a significantly larger number of SM genes than randomly expected (α=0.05) and are referred to as SM modules. Similarly, 125 GM modules were significantly enriched in GM genes. Therefore, a subset of annotated GM and SM genes tend to be co-expressed with other GM and SM genes, respectively. However, >50% of SM and GM genes did not belong to SM/GM modules (gray, **Figure 2E**). In addition, 0.3%-14.0% of GM genes were found in SM modules and 0%-32% of SM genes were found in GM modules, depending on the dataset and algorithm (**Figure 2E**). This pattern reflects the fact that some GM genes function upstream of an SM pathway and further highlights the challenge in differentiating SM and GM genes using co-expression patterns alone.

Network properties of SM and GM genes

SM genes tend to have specialized functions and are involved in one or a few pathways, leading us to hypothesize that SM genes would have fewer connections in biological networks than GM genes. To test this prediction, we first assessed the connectivity among SM genes and among GM genes in a protein-protein interaction network (28) and found that SM genes have a significantly smaller number of physical interactions (mean = 1.25) than GM genes (1.84, p=0.03, **Figure S2B**). The smaller number of SM gene interactions is not because SM genes have shorter coding regions (SM>GM, *p*=0.004, **Figure S2C**) or fewer protein domains (SM<GM, *p*=0.352, **Figure S2D**). Thus, the significantly fewer protein-protein interactions

formed by SM proteins is consistent with SM genes having more specific functions than GM genes (6). It is also possible that there are more interaction experiments for GM genes, or GM genes tend to be in larger pathways compared to SM genes. Next, we examined the same relationships using the AraNet functional network (29), which connects genes likely with similar functions through the integration of multiple datasets, including expression and protein-protein interaction. However, there was no significant difference between the number of genes connected to SM genes and to GM genes (p=0.139, **Figure S2E**), suggesting that network connectivity based on gene-gene interactions is not useful for distinguishing between SM and GM genes.

Evolutionary rates of SM and GM genes based on within- and cross-species comparisons

SM genes are frequently involved in plant adaptation to variable environments (30, 31). In contrast, GM genes - which are involved in ancient and stable metabolic functions such as photosynthesis - are expected to be more highly conserved and experience stronger negative selection (32). To assess differences in the evolutionary rates of SM and GM genes, we searched for *A. thaliana* SM and GM paralogs as well as homologs across six plant species spanning more than 300 million years of evolution (see Methods). A significantly higher proportion of SM genes have paralogs than GM genes (*p*=1.2e-10, **Figure S3A**). However, consistently fewer SM genes (14.8-54%) have homologs across species than GM genes (27-76%) (all *p*<2e-4, **Figure S3A**). In addition, only 0.94% of SM genes have homologs in core eukaryotic genomes (33) compared with 14.7% of GM genes (**Figure S3A**). Finally, we determined the timing of GM and SM duplications over the course of land plant evolution (see Methods). The most recent duplicates of 75% of SM genes were found within the Brassicaceae family compared to only 40% of GM genes (**Figure 3A**). Additionally, 25% of SM genes were duplicated within the *Arabidopsis* genus, compared to only 7% of GM genes (**Figure 3A**). Thus, SM genes have higher duplication rates, but do not persist in the long run, leading to observation of fewer homologs across species. In addition, SM genes tend to be more recently duplicated relative to GM genes.

We also found that SM genes and their homologs had significantly higher non-synonymous (*dN*) to synonymous (*dS*) substitution rate ratios (all *p*<1e-06, **Figure 3B**) and higher nucleotide diversity values (**Figure S3B**) compared with GM genes. Together with other measures of selection (**Figure S3C, D**), both within- and cross-species comparisons suggest a lower degree of conservation and weaker negative selection on SM genes relative to GM genes. This is consistent with the roles of SM genes mostly in the production of metabolites important for tolerance to rapidly changing abiotic stress conditions and defense against biotic agents (6).

Duplication mechanisms and genomic clustering of SM and GM genes

Mechanisms in how a gene is duplicated, such as via whole genome duplication (WGD), tandem duplication, or dispersed duplication, may impact subsequent functional divergence and ultimately influence whether a duplicate is under selection and retained (1). Therefore, we first compared the number of GM and SM duplicates generated by WGD in the *A. thaliana* lineage and found no significant difference in the number of SM and GM duplicates (*p*=0.1, **Figure S4A**). However, the expansion of genes involved in the biosynthesis of glucosinolates is due to the α WGD event (34). This may suggest that the glucosinolate pathways are an exception to the general pattern. Compared to WGD, significantly more SM genes tend to be tandem duplicates than GM genes (*p*=1e-48, **Figure S4A**). Considering that genes involved in response to the environment are more likely to be tandem duplicates (2, 35), tandem duplication potentially allowed for rapid evolution of SM gene families that are subject to selection in variable environments.

To measure the degree of SM and GM gene gain and loss, the numbers of paralogs and pseudogenes were used as indicators of gains and losses, respectively. SM genes tend to have more paralogs (*p*=3.5e-74, **Figure 3C**), higher sequence similarities to their paralogs (*p*=3e-3, **Figure 3D**), and lower synonymous substitution rate (*dS*) (*p*=1e-19, **Figure 3E**) compared to GM genes. Furthermore, a higher percentage of SM genes have been duplicated since *A. thaliana* diverged from *A. lyrata* (*p*=3.2e-8, **Figure S4B**), and SM genes tend not to be found in single copies (*p*=3.5e-6, **Figure S4C**). These findings all point to more recent expansion of SM gene families. We also compared the functional likelihood (a measure of sequence selection) (36) between SM genes, GM genes, and pseudogenes. Interestingly, we found that SM genes have functional likelihoods that are significantly lower than GM genes, but higher than pseudogenes (ANOVA, Tukey's test, *p*=2e-16, **Figure 3F, Figure S4E**). Because genes under strong negative selection have high functional likelihoods that are close to one, whereas pseudogenes tend to have values close to zero (36), this finding indicates that some SM genes are under weaker selection and may be in the process of becoming pseudogenes. The proportion of pseudogene paralogs for SM genes (9.8%) compared with GM genes (6.5%) is not significant (*p*=0.2, **Figure S4D**). Considering that SM genes tend not to have cross-species homologs (**Figure S3A**), this finding suggests that pseudogenes likely did not persist sufficiently long to be an adequate indicator of gene loss.

SM and GM genes that function in the same pathway can sometimes cluster together in the same region of the genome (referred to as genomic clusters) (18, 37–39). We used two approaches to see if SM and GM genes tend to be in genomic clusters. In the first approach, we asked whether SM and GM genes tend to be located near other SM and GM genes, respectively, regardless of whether the neighboring genes are paralogous or not. We found that SM genes cluster near other SM genes (*p*=9.5e-121, **Figure S4F**) and GM genes tend to be close to GM genes (*p*=4.1e-12, **Figure S4G**). In the second approach, we defined metabolic clusters identified using Plant Cluster Finder (18) but the identified clusters were not enriched in either SM or GM genes (**Figure S4H**). Taken together, SM genes are more likely to be tandemly duplicated and tend to belong to large gene families. SM genes also tend to in genomic clusters with other SM genes, indicating that these characteristics may be useful features in distinguishing SM and GM genes. In the following section, we integrate these features to establish an SM-gene prediction model using machine learning approaches.

Machine learning model for predicting SM and GM genes

In total, we examined 10,243 features (summarized in **Dataset S3**) that differ widely in ability to predict benchmark SM and GM genes. We use machine learning because it allows us to leverage multiple features together to build an integrated model, offering better predictive power by lowering the false negative rate (FNR) and the false positive rate (FPR), compared to using each feature individually to build a naïve model. For example, the best performing single feature, gene family size, led to a model with an Area under Receiver Operating Characteristic curve (AuROC) of 0.8. AuROC of 0.5 indicates the performance of random guesses and a value of 1 indicates perfect predictions. However, using this high performing feature alone as the predictor resulted in a 43% FPR and a 58% FNR. In addition, the majority of the features are not particularly informative (**Dataset S3**), as the average AuROC for single features was extremely low (0.5) with an average FPR of 89%.

These findings indicate that SM and GM genes are highly heterogeneous and cannot be distinguished with high accuracy by single features. To remedy this, we integrated all features using two machine learning algorithms, Support Vector Machine and Random Forest (see Methods). The better performing SM gene prediction model with Support Vector Machine had an AuROC=0.82, 16.6% FPR, and 2.7% FNR (**Figure 4A**). Randomizing SM/GM labels but maintaining the same feature values associated with the benchmark genes as the initial model resulted in AuROCs between 0.51-0.57, as expected for random guesses (**Dataset S3**). Note that the performance measures reported above were based on models built with a 10-fold cross-validation scheme where 90% data was used for training the models and 10% for testing them. Based on the prediction outcomes, each gene was given an "SM score" ranging from 0 to 1 indicating the likelihood that the gene was an SM gene. Based on a threshold SM score defined by minimizing false predictions (see Methods), we found that 84.4% of the training SM genes were predicted correctly (**Figure 4B**), while 72.2% of the training GM genes were correctly predicted (**Figure 4B**), a significant improvement from individual feature-based models.

Features important for SM gene prediction and model application on unannotated enzyme genes

In addition to the SM score, the machine learning result included a list of feature "weights" where features with more positive values are more informative for predicting SM genes. In contrast, more negative feature weights are more informative for predicting GM genes (**Dataset S3, Figure S5A**). The most informative features for predicting SM genes included specific protein domains as well as multiple gene duplication-related features, such as duplication mechanism (alpha and tandem), gene family expansion (family size), and timing of duplication **(Figure 4C)**. In addition, duplications immediately prior to the divergence of the *A. thaliana* lineage from the *A. lyrata*, *Capsella rubella*, *Vitis vinifera*, *Solanum lycopersicum*, and *Picea abies* lineages were among the most informative for predicting SM genes. This potentially reflects the major time points of diversification in specialized metabolites and SM pathways. In contrast, duplication prior to the divergence from *Brassica rapa* (the most distant Brassicaceae species used in this analysis), which corresponds to the most recent whole genome duplication in the Brassicaceae lineage, and *Amborella trichopoda* (a basal angiosperm) and *Marchantia polymorpha* (liverwort) lineages are informative for predicting GM genes. Finally, expression correlation with a paralog under abiotic stress as well as hierarchical, *k*-means, and approximate *k*-means co-expressed clusters under stress, diurnal, and development were helpful for predicting both SM and GM genes.

With the accuracy of the SM gene prediction models demonstrated through cross-validation and prominent features identified, we next applied our machine learning models to make predictions for 3,104 enzymatic genes not annotated to be SM or GM genes (see Methods, **Dataset S1)**. Of these genes, 46% (1,610 genes) were predicted to be SM genes. We took three approaches to assess the accuracy of these predictions of SM and GM genes. First, prior to model training, we intentionally held out 17 known SM genes (**Figure 4B, Dataset S1**) from any model training. Upon application of the machine learning model, 14 were correctly predicted as SM genes, indicating that the model has an 82% True Positive Rate (or 18% FNR). Second, we tested how well genes in well-known SM pathways involved in glucosinolate biosynthesis (38, 39) could be predicted. To do this we built a new model using the benchmark SM and GM genes but excluding genes from glucosinolate biosynthetic pathways (see Methods) (**Figure 4B, Dataset S1)**. When applying this new model to glucosinolate genes, 86% of known glucosinolate pathway genes were correctly predicted as SM genes. Thus, the FNR is at 14%, much lower than the 58% FNR using the best gene family size feature.

Finally, three enzyme families (methyltransferase, terpene synthase, and cytochrome P450, see Methods) were analyzed to test model performance within a specific family (**Figure 4B, Dataset S1**). These families were chosen because they tend to be associated with SM. We built separate models for each family, excluding the genes belonging to each respective family during model training. Upon applying this model to each enzyme family, 86% P450, 85% terpene synthase and 67% methyltransferase genes were predicted as SM genes (**Figure 4B**). Taken together, our models allowed assessment of the relative importance of features in distinguishing SM and GM genes, as well as provided predictions for SM genes among enzyme genes with no known SM/GM designation. In addition, these models can predict the majority of hold-out genes with known SM functions, glucosinolate pathway genes, and enzyme families whose members predominantly play roles in SM pathways. These findings indicate that our models and this general approach are valuable for classifying unknown enzymes.

Characteristics of Mis-Predicted Genes

Although our SM prediction model performed well, we found that 431 GM-annotated genes (27.8% of benchmark) were mis-predicted to be SM, while 31 SM genes (15.5% of benchmark) were mis-predicted as GM genes. To assess the causes of mis-predictions, we used a subset of the most informative features (**Figure 4C, Dataset S3**) to determine how their values differed between four gene classes based on the consistency between annotations and predictions. These four classes include: (1) annotated GM predicted as GM (GM [annotation] →GM [prediction]), (2) annotated SM predicted as SM (SM→SM), (3) annotated GM predicted as SM (GM→SM), and (4) annotated SM predicted as GM (SM→GM). Genes in the mis-predicted classes (3 and 4) tend to have feature values between those of genes in correctly predicted classes (1 and 2). For example, the median values of the feature functional likelihood among these four gene classes follow the order, from high to low: GM→GM, SM→GM, GM→SM, SM→SM (**Figure 4D**). The same pattern is also true for median expression levels (**Figure 4E**), the number of expressed conditions (**Figure 4F**), *dN/dS* values (**Figure 4G**), and values for other gene features we have examined (**Figure S5A-J**). Thus, in the SM→GM mis-predicted class, the annotated SM genes in fact possess properties that are more similar to those of GM genes and vice versa (**Figure 4D-G, Figure S5A-J**).

These observations suggest the hypothesis that some of the mis-predicted benchmark genes (**Figure 4B**) may in fact be mis-annotated. To assess this idea, we collated literature information on 17 genes with consistent GO/Aracyc annotations and our model predictions (GM→GM=4, SM→SM=13), as well as 13 genes that were inconsistent (SM→GM=3, GM→SM=8, **Dataset S1**). We focused on genes in the P450/ terpene synthase families because they were among the best characterized (**SI text**). For mis-predicted genes, one of three genes in the SM→GM class had supporting GM evidence (cinnamate-4-hydroxylase) and four of eight genes in the GM→SM class had supporting SM evidence (**SI text**). These findings indicate that a subset of these genes (45% based on the genes examined) are mis-annotated, rather than mis-predicted, particularly among GM annotated gene. It is possible that some of the erroneous annotations are based on *in vitro* biochemical activity and/or sequence similarities alone, criteria that may not accurately represent the *in vivo* functions. We also examined genes with conflicting annotation information between GO and AraCyc (11) or with no GM/SM annotation (4) (**Dataset S1**) and found our predictions and biochemical data are consistent in 87% cases (**SI text**). Together with the finding that all (17/17) genes with consistent annotations and predictions had biochemical evidence supporting their SM/GM classification (**SI text**), these results further support the utility of our machine learning model and demonstrate the feasibility in using the model prediction outcome to prioritize future experiments to determine the *in planta* role of SM or GM genes, including those that may be mis-annotated or have functions beyond their annotated activities.

Conclusions

Machine learning models using genomic features show considerable promise in predicting the functions of unclassified or unannotated genes (18, 36). We have identified significant differences between SM and GM genes in *A. thaliana* related to expression, evolution, and duplication, (see **Dataset S2** for full feature list) and have shown these features can be used to distinguish SM genes from GM genes using machine learning approaches. In addition, while individual characteristics significantly differ between SM and GM genes, the effect sizes are small and any individual characteristic does a poor job of predicting SM genes relative to a model that includes a combination of characteristics. Therefore, we developed SM vs. GM models that predict SM genes among enzymatic genes as well as within a specific pathway or family of genes. It is important to note scores can be used to prioritize genes of unknown function for analysis based on SM likelihood.

Although the SM gene prediction model performs well, there are two areas that can be improved. First, high quality benchmark data is essential for building a good machine learning model. Our model predictions, particularly those disagree with existing annotations should be examined experimentally to assess the reasons behind the discrepancy. The information can then be incorporated for iteratively refine the prediction model. Second, additional features that can distinguish SM and GM genes may be needed to further improve model performance. In this study, a collection of >10,000 features were examined either because there was prior knowledge of their importance or because they were potentially useful. The challenge is how to identify these missing features. Another consideration is that our model is for distinguishing SM and GM genes. There are 151 GO and 259 AraCyc genes that are annotated to play roles in both specialized and general metabolism (**Figure 1A**). In addition, there are >90 SM pathways in AraCyc. Future studies distinguishing SM, GM, and genes in both SM and GM pathways can be carried out through multi-class modeling methods. It will also be desirable to devise models that can predict SM genes at the pathway level.

In summary, our study provided a detailed analysis of features where a subset of them represent signatures of SM and GM genes that have not been reported previously. Using machine learning models integrating heterogenous features, a global prediction of enzyme-encoding SM genes in *A. thaliana* is generated which can be used for selection of candidate genes for functional characterization. The machine learning framework we have adopted can be applied broadly to identify SM genes in other plant species. Most importantly, the prediction outcome is the SM score, a quantitative measure reflecting the likelihood an enzyme gene likely plays roles in specialized metabolism, indicative of the degree of biochemical specialization of an enzyme.

Methods

Specialized and general metabolism gene annotation and enrichment analysis

Benchmark SM and GM genes were identified based on GO ((22); http://www.geneontology.org/ontology/go.obo) and AraCyc (23); http://www.plantcyc.org/) annotations. GO annotations for *A. thaliana* were downloaded from The Arabidopsis Information Resource (TAIR;(40)) and genes annotated with the secondary metabolism term (GO:0019748) and primary metabolism term (GO:0044238) were selected as potential SM genes and GM genes, respectively. Genes that were associated with a more specialized child GO term of SM and PM terms (but not the SM/PM terms themselves) were also classified as SM and GM genes, respectively. Only genes annotated with either SM or PM terms, but not both, were included in the analysis and only those with experimental evidence codes IDA, IEP, IGI, IPI and/or IMP were included in the model. AraCyc v.15 pathway annotations were retrieved from the Plant Metabolic Network database (23); http://www.plantcyc.org). Potential SM genes were those associated with “secondary metabolites biosynthesis” pathways. Potential GM genes were those found in non-secondary metabolite biosynthesis pathways. Some genes were annotated in both SM and non-SM pathways and were excluded from further analysis. In addition, genes without experimental evidence as annotated in AraCyc (EV-EXP) were excluded from the model. To be included in further analysis, potential SM and GM genes from GO or AraCyc were required to have an enzyme commission number annotation from AraCyc. Sets of benchmark SM and GM genes were then established by merging potential SM and GM genes identified based on GO and AraCyc annotations but removing genes with ambiguous SM and GM classifications (e.g. SM in GO but PM in AraCyc). The full list of GM and SM genes are available in **Dataset S1**.

Glucosinolate pathway genes were defined as AraCyc annotated genes in multiple glucosinolate pathways (**Dataset S1**), as well as genes with the GO terms: glucosinolate metabolic process (GO:0019760), glucosinolate biosynthetic process (GO:0019761), indole glucosinolate metabolic process (GO:0042343), glucosinolate transport (GO:1901349), regulation of glucosinolate biosynthetic process (GO:0010439). This results in 72 genes annotated to glucosinolate pathways and processes available in **Dataset S1.** Terpene synthase, P450, and methyltransferase genes were identified from *A. thaliana* annotated protein sequences by using the following domain matches from Pfam: terpene_synth, p450, methyltr_7.

Enrichment of SM genes relative to GM genes in AraCyc pathways or GO categories was assessed with Fisher’s exact tests, and test *p*-values were corrected for multiple testing (41). The enrichment results are available in **Dataset S2**. GO slim terms and AraCyc pathways that mapped to a particular gene were used as binary features in prediction models (resulting in 636 features, or 1 feature for each GO slim term and pathway). Pfam v.30 (24); http://pfam.xfam.org/) was used to identify protein domains in proteins encoded by SM and GM genes with HMMER (42); http://hmmer.org/). A domain match was considered significant if the score was above the trusted cutoff parameter. Enriched domains were then used as model features (totaling 4,217 features).

Expression dataset processing and co-expression and gene network analysis

Expression datasets were downloaded from TAIR ((40); http://www.arabidopsis.org/). Target datasets included plant development (43), biotic stress (44), abiotic stress (44, 45), hormone treatment (46) and diurnal expression (47). Genes that were considered significantly expressed relative to background in the development expression dataset were those with a $\log_2$ microarray hybridization intensity value of ≥4 (the cutoff value is based on our earlier study, (36)). The median and maximum expression levels and expression variation and breadth across the developmental expression dataset were calculated as previously described (36). Differentially expressed genes under biotic stress, abiotic stress, and hormone treatments were defined as those that had an absolute $\log_2$ fold change ≥1 and adjusted $p<0.05$ following analysis using the affy and limma packages in R (48, 49). For each gene, the number of conditions in which the gene in question was significantly differentially regulated was also calculated. This resulted in 16 expression values that were used as model features (**Dataset S2**).

For each expression dataset (development, abiotic, biotic, and hormone), Pearson Correlation Coefficients (PCC) were calculated between each gene and genes in the same paralogous cluster as defined by ORTHOMCL v1.4 (50). For the gene in question, the maximum PCC <1 for genes in the paralog cluster was used as the PCC value. In addition to examining expression correlation, co-expressed genes in the biotic stress, abiotic stress, diurnal, and developmental datasets were classified into co-expression clusters using standard *K*-means, approximate kernel *K*-means, c-means, and hierarchical clustering algorithms as described in our earlier study (21) resulting in 5,303 binary features. For *K*-means-related analyses, the within sum of squares was plotted against the number of clusters, and *K* was chosen based on the number of clusters at the elbow or bend of the plot. Gene clusters that were significantly enriched in SM or GM genes were identified using Fisher's exact tests (adjusted-$p<0.05$). The number of AraNet gene network interactions ((29); http://www.functionalnet.org/aranet/), number of protein interactions (28), domain number, and amino acid length were calculated in our earlier study (36). There were 23 model features related to PCC values, significant cluster membership, and gene network data (**Dataset S2**).

Conservation, duplication, methylation, histone modification, and genome location related features

Nonsynonymous (*dN*)/synonymous (*dS*) substitution rates between plant homologs, core eukaryotic gene status, nucleotide diversity data, Fay and Wu's H and MacDonald-Kreitman test statistics were the same as used in our earlier study (36, 51, 52). The timing of duplication of an *A. thaliana* gene X was defined based on a comparison of the BLAST scores between X and its closest paralog Y ($S_{X,Y}$) and between X and its closest homolog Z in each of 15 other plant species ($S_{X,Z}$): *A. lyrata, C. rubella, B. rapa, Theobroma cacao, Populus trichocarpa, Medicago truncatula, V. vinifera, S. lycopersicum, Aquilegia coerulea, Oryza sativa, A. trichopoda, P. abies, Selaginella moellendorffii, Physcomitrella patens,* and *M. polymorpha*. Among cases where $S_{X,Z} > S_{X,Y}$, the species with gene Z most distantly related to *A. thaliana* was identified. Thus, the timing of X duplication must be immediately prior to the divergence between *A.*

thaliana and the species harboring gene Z **(Dataset S2).** Pseudogenes were defined using a published pipeline (53). The lethal gene scores, which represent the relative likelihood that a mutation in a gene is lethal, and additional gene duplication-related features, including gene family size, rates of synonymous substitutions, α and βγ whole genome duplication status, and tandem duplicate status **(Dataset S2**), were obtained from (36).

CG methylation and log fold change of histone marks relative to background were taken from (36). The average of the log fold change of each histone mark was calculated for all histones that overlapped with a gene. There were 37 feature values related to conservation, duplication, methylation, and histone modification (**Dataset S2**).

Three approaches were used to evaluate the degree of metabolic gene clustering. The first approach involved a co-localization measure defined as, for each gene X (GM or SM), the number of GM or SM genes within five or 10 genes from X (four features, **Dataset S2**). For the second approach, we first defined a metabolic gene cluster as a group of >1 genes annotated to as SM or GM genes where the neighboring SM/GM gene pair were within 10 non-SM/GM genes and 100 kb of each other. The clusters were then determined to be homologous or non-homologous by the presence of a significant BLAST match in a given cluster (two features, **Dataset S2**). In the third approach, metabolic gene clusters were identified using the Plant Cluster Finder tool (18) with the following parameters: has >2 metabolic genes, two reaction identifiers, all genes in the cluster are on the same chromosome, clusters of only tandem duplicates are not allowed, and number of metabolic genes > number of non-metabolic genes (one feature, **Dataset S2**). The genomic clustering values (resulting in seven features) are shown in **Dataset S2**.

Machine learning classification of SM and GM genes

The prediction models were built based on 10,243 features using the Random Forest (RF) and Support Vector Machine (SVM) algorithms implemented in the Weka program (54). The benchmark SM genes were first divided into a modeling set (92%) and a hold-out set for independent validation (8%). Since there were significantly more GM genes than SM genes, 100 balanced data sets were constructed by randomly selecting GM genes equal to the number of SM genes in each balanced set. Additionally, ten-fold cross validation was performed for 100 random draws of a balanced data set for each machine learning run, and grid searches were performed to obtain the best performing parameters for each model. Performance of the RF and SVM models was determined by both AuROC, or the area under the plot of the true positive (TP) rate against the false positive (FP) rate, calculated in R by the ROCR package, and F-measure, the harmonic mean of precision (TP/TP+FP) and recall (TP/TP+FN), where FN= false negative. A confidence score between 0 and 1 was produced by the model and was used as the SM prediction score.

To call a gene as SM or not, the threshold SM score was defined as the SM score with the highest F-measure in the RF or the SVM model. AuROC and F-measures were also calculated for each feature to determine their individual predictive value. In addition to cross-validation, the hold-out data were used to further assess model performance. Finally, the predictive value of each feature was calculated individually with a custom python script using the known SM and GM genes and the individual feature values to calculate the FP, FN, TP, and TN rates and subsequently the F-measure and AuROC values (**Dataset S3**. For the random model, we first randomized the SM/GM labels of benchmark genes but with feature values associated with each gene unchanged. This randomized feature table was then used to establish machine learning models and the model performance was evaluated with F-measure and AuROC values. Models were applied to enzyme genes not classified by GO or AraCyc as SM or GM, but with known enzyme commission (E.C.) number annotations from AraCyc. Additional models were built to exclude genes in glucosinolate pathways or specific enzyme families (terpene synthases, cytochrome P450s, methyltransferases). The model built with genes excluding the designated pathway or family was then applied to classify genes in the pathway or family in question.

Acknowledgements

We thank Christina B. Azodi, Joshua J. Moore, Nicholas L. Panchy, and Sahra Uygun for helpful discussion and support. This work was partly supported by grants from the National Science Foundation IOS-1546617 to R.L., E.P., and S.-H.S. and DEB-1655386 to S.H.S.

References

1. Panchy N, Lehti-Shiu MD, Shiu S-H (2016) Evolution of gene duplication in plants. *Plant Physiol* 171(4): 2294–2316.
2. Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H (2008) Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. *Plant Physiol* 148(2):993–1003.
3. 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: Terpene synthase family. *Plant J* 66(1):212–229.
4. Chae L, Kim T, Nilo-Poyanco R, Rhee SY (2014) Genomic Signatures of Specialized Metabolism in Plants. *Science* 344(6183):510–513.
5. Pichersky E, Lewinsohn E (2011) Convergent evolution in plant specialized metabolism. *Annu Rev Plant Biol* 62:549–566.
6. Hartmann T (2007) From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry* 68(22–24):2831–2846.
7. Ehrlich PR, Raven PH (1964) Butterflies and Plants: A Study in Coevolution. *Evolution* 18(4):586.
8. Herde M, Howe GA (2014) Host plant-specific remodeling of midgut physiology in the generalist insect herbivore Trichoplusia ni. *Insect Biochem Mol Biol* 50:58–67.
9. Howat S, et al. (2014) Paclitaxel: biosynthesis, production and future prospects. *New Biotechnol* 31(3):242–245.
10. Zhong J-J (2002) Plant cell culture for production of paclitaxel and other taxanes. *J Biosci Bioeng* 94(6):591–599.
11. Giuliano G, Tavazza R, Diretto G, Beyer P, Taylor MA (2008) Metabolic engineering of carotenoid biosynthesis in plants. *Trends Biotechnol* 26(3):139–145.
12. Milo R, Last RL (2012) Achieving Diversity in the Face of Constraints: Lessons from Metabolism. *Science* 336(6089):1663–1667.
13. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814):796-815.

14. D'Auria JC, Gershenzon J (2005) The secondary metabolism of Arabidopsis thaliana: growing like a weed. *Curr Opin Plant Biol* 8(3):308–316.
15. Schenck CA, et al. (2017) Molecular basis of the evolution of alternative tyrosine biosynthetic routes in plants. *Nat Chem Biol* 13(9):1029–1035.
16. Moghe G, Last RL (2015) Something old, something new: Conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiol* 169(3):1512-1523.
17. Wisecaver JH, et al. (2017) A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* 29(5):944–959.
18. Schlapfer P, et al. (2017) Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants. *Plant Physiol* 173(4):2041-2059.
19. Wei H, et al. (2006) Transcriptional Coordination of the Metabolic Network in Arabidopsis. *Plant Physiol* 142(2):762–774.
20. Higashi Y, Saito K (2013) Network analysis for gene discovery in plant-specialized metabolism: Gene discovery in plant specialized metabolism. *Plant Cell Environ* 36(9):1597–1606.
21. Uygun S, Peng C, Lehti-Shiu MD, Last RL, Shiu S-H (2016) Utility and Limitations of Using Gene Expression Data to Identify Functional Associations. *PLoS Comput Biol* 12(12):e1005244.
22. Botstein D, et al. (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* 25(1):25–9.
23. Rhee SY, Zhang P, Foerster H, Tissier C (2006) AraCyc: Overview of an Arabidopsis Metabolism Database and its Applications for Plant Research. *Biotechnology in Agriculture and Forestry* (Springer, Berlin, Heidelberg).
24. Finn RD, et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44(D1):D279–D285.
25. Huot B, Yao J, Montgomery BL, He SY (2014) Growth–Defense Tradeoffs in Plants: A Balancing Act to Optimize Fitness. *Mol Plant* 7(8):1267–1287.
26. Cedar H, Bergman Y (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet* 10(5):295–304.
27. Chan SW-L, Henderson IR, Jacobsen SE (2005) Erratum: Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nat Rev Genet* 6(5):351–360.
28. Arabidopsis Interactome Mapping Consortium (2011) Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science* 333(6042):601–606.
29. Lee T, Lee, I. (2017) A Network Biology Server for Arabidopsis thaliana and Other Non-Model Plant Species. *Plant Gene Regulatory Networks. Methods in Molecular Biology* (Humana Press, New York, NY).
30. Edger PP, et al. (2015) The butterfly plant arms-race escalated by gene and genome duplications. *Proc Natl Acad Sci USA* 112(27):8362–8366.
31. Steppuhn A, Baldwin IT (2007) Resistance management in a native plant: nicotine prevents herbivores from compensating for plant protease inhibitors. *Ecol Lett* 10(6):499–511.
32. Rojas CM, Senthil-Kumar M, Tzin V, Mysore KS (2014) Regulation of primary plant metabolism during plant-pathogen interactions and its contribution to plant defense. *Front Plant Sci* 5. doi:10.3389/fpls.2014.00017.

33. Tatusov RL, et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4(1):41.
34. Hofberger JA, Lyons E, Edger PP, Chris Pires J, Eric Schranz M (2013) Whole Genome and Tandem Duplicate Retention Facilitated Glucosinolate Pathway Diversification in the Mustard Family. *Genome Biol Evol* 5(11):2155–2173.
35. Rizzon C, Ponger L, Gaut BS (2006) Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in Arabidopsis and Rice. *PLoS Comput Biol* 2(9):e115.
36. Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H (2015) Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes. *Plant Cell* 27(8):2133–2147.
37. Osbourn A (2010) Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends Genet* 26(10):449–457.
38. Qi X, et al. (2004) A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants. *Proc Natl Acad Sci USA* 101(21):8233–8238.
39. Sakamoto T (2004) An Overview of Gibberellin Metabolism Enzyme Genes and Their Related Mutants in Rice. *Plant Physiol* 134(4):1642–1653.
40. Berardini TZ, et al. (2015) The Arabidopsis Information Resource: Making and Mining the “Gold Standard” Annotated Reference Plant Genome. *Genesis* 53(8):474–485.
41. Benjamin Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Statistical Society Ser B* 57(1): 289-300.
42. Finn RD, et al. (2015) HMMER web server: 2015 update. *Nucleic Acids Res* 43(W1):W30–W38.
43. Schmid M, et al. (2005) A gene expression map of Arabidopsis thaliana development. *Nat Genet* 37(5):501–506.
44. Wilson TJ, Lai L, Ban Y, Steven XG (2012) Identification of metagenes and their interactions through large-scale analysis of Arabidopsis gene expression data. *BMC Genomics* 13(1):237.
45. Kilian J, et al. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses: AtGenExpress global abiotic stress data set. *Plant J* 50(2):347–363.
46. Goda H, et al. (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. *Plant J* 55(3):526–542.
47. Mockler TC, et al. (2007) The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. *Cold Spring Harbor Symposia on Quantitative Biology* (Cold Spring Harbor Laboratory Press), pp 353–363.
48. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20(3):307–315.
49. Ritchie ME, et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7):e47–e47.
50. Chen F (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 34(90001):D363–D368.

51. Lehti-Shiu MD, et al. (2015) Molecular Evidence for Functional Divergence and Decay of a Transcription Factor Derived from Whole-Genome Duplication in *Arabidopsis thaliana*. *Plant Physiol* 168(4):1717–1734.
52. Moghe GD, et al. (2013) Characteristics and Significance of Intergenic Polyadenylated RNA Transcription in Arabidopsis. *Plant Physiol* 161(1):210–224.
53. Zou C, et al. (2009) Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice. *Plant Physiol* 151(1):3–15.
54. Smith TC, Frank E (2016) Introducing Machine Learning Concepts with WEKA. *Statistical Genomics*, eds Mathé E, Davis S (Springer New York, New York, NY), pp 353–378.

Figures and Figure Legends

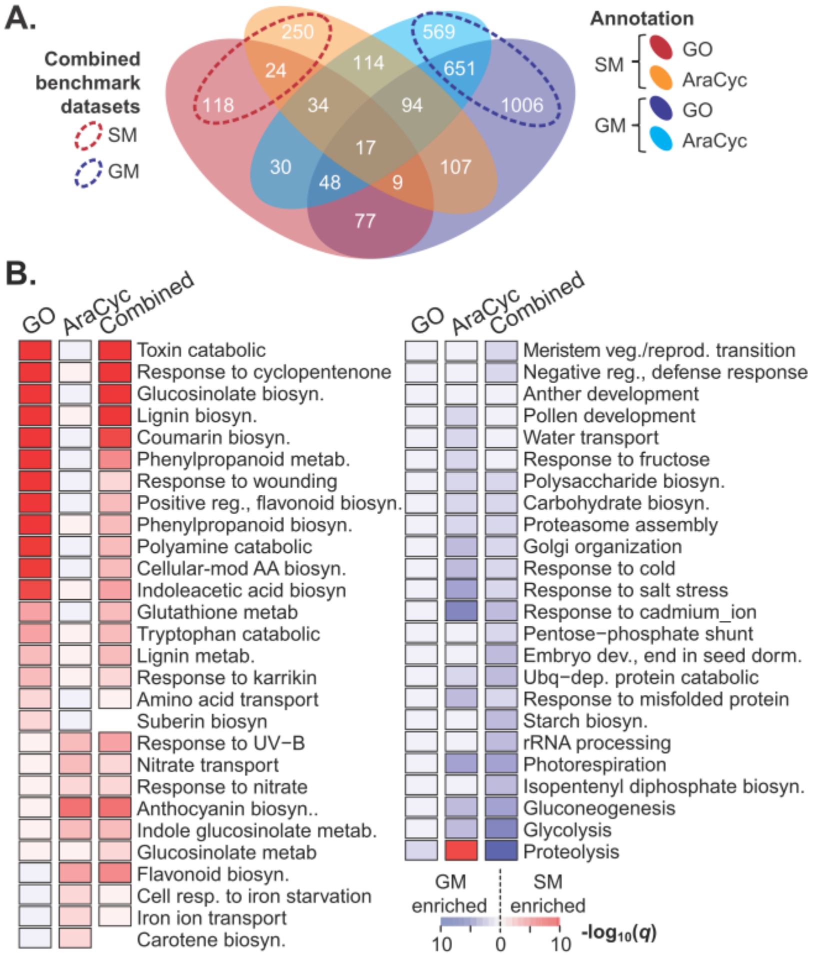

Figure 1. Gene Ontology and AraCyc annotation of specialized and primary metabolism genes. (A) Overlap between Gene Ontology (GO)/AraCyc primary metabolism (PM) and secondary metabolism (SM) gene annotations. The number of genes in an intersected or a complement set are shown, and dotted lines indicate sets of SM (red) and GM (purple) genes used for machine learning models. **(B)** GO term enrichment in SM genes (left panel) and in GM genes (right panel). The three columns show statistics for GM/SM genes that are GO-annotated, AraCyc-annotated, or belong to a combined set (union between GO and AraCyc). Rows: GO terms. Color: representing the *q*-value (multiple testing corrected *p*-value) of the Fisher's exact test for a GO term enriched in either GM (blue) or SM (red) genes (**Dataset S2**). White: no significant enrichment.

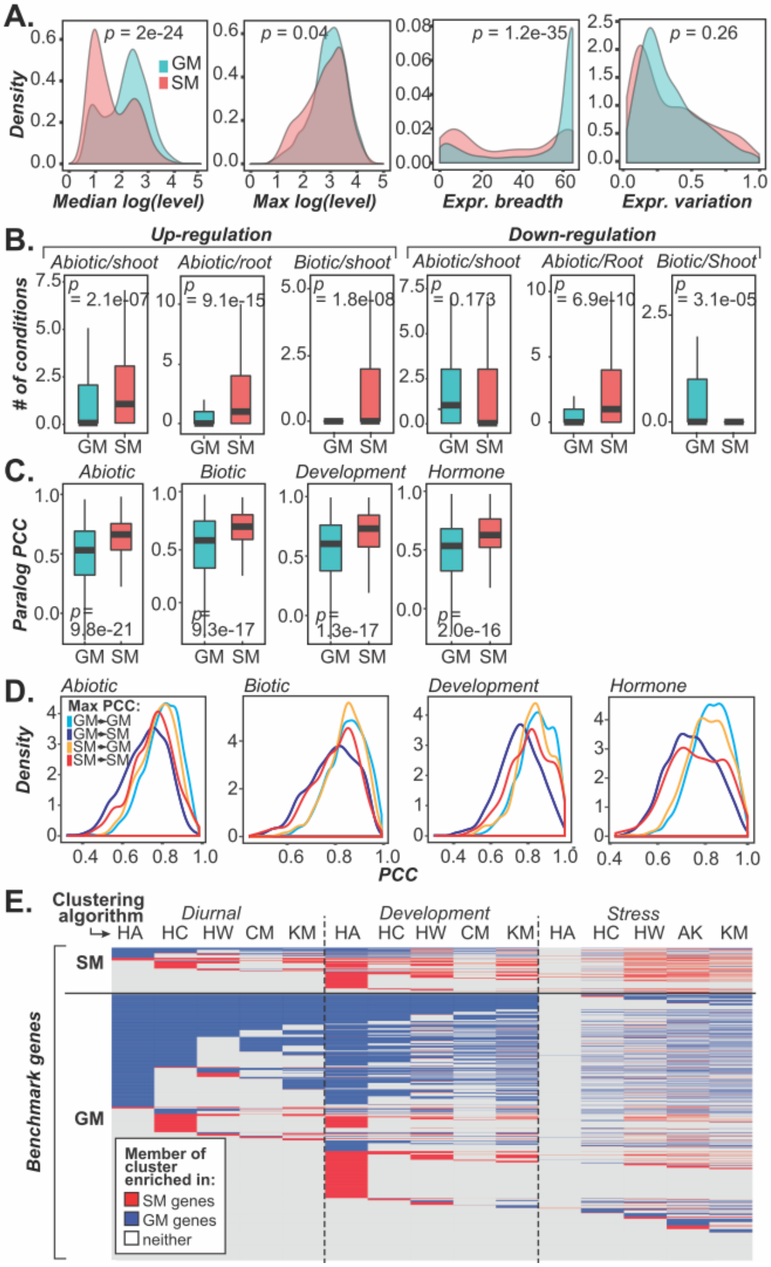

Figure 2. Differences in expression and co-expression characteristics of SM and GM genes. (A) Distributions of SM (red) and GM (blue) gene expression-related values calculated from the development dataset. Level: microarray intensity. Expression breadth: the number of tissues/developmental stages in which a gene is expressed. Expression variation: median absolute deviation/median. **(B)** Distributions of the number of conditions in which a gene is up- or down-regulated in the abiotic/biotic stress dataset from root or shoot. **(C)** Distributions of maximum PCC values between SM or GM genes and their paralogs in four expression datasets. All test statistics from **(A-C)** were generated using Mann-Whitney U tests. **(D)** Distributions of maximum Pearson's Correlation Coefficients (PCC) between GM-GM (light blue), GM-SM (dark blue), SM-GM (light purple), and SM-SM (pink) gene pairs using the same expression datasets as in **(C)**. **(E)** Clustering of SM and GM genes based on their expression patterns in the three datasets with six algorithms: HA (hierarchical, average linkage), HC (hierarchical, complete linkage), HW (hierarchical, Ward's method), CM (*c*-means), KM (*k*-means), and AK (approximate *k*-means). Row: a benchmark SM/GM gene. Blue and red shading: the gene belongs a cluster with an over-represented number of GM genes and SM genes, respectively, compared with the background (*p*<0.05, Fisher's exact test).

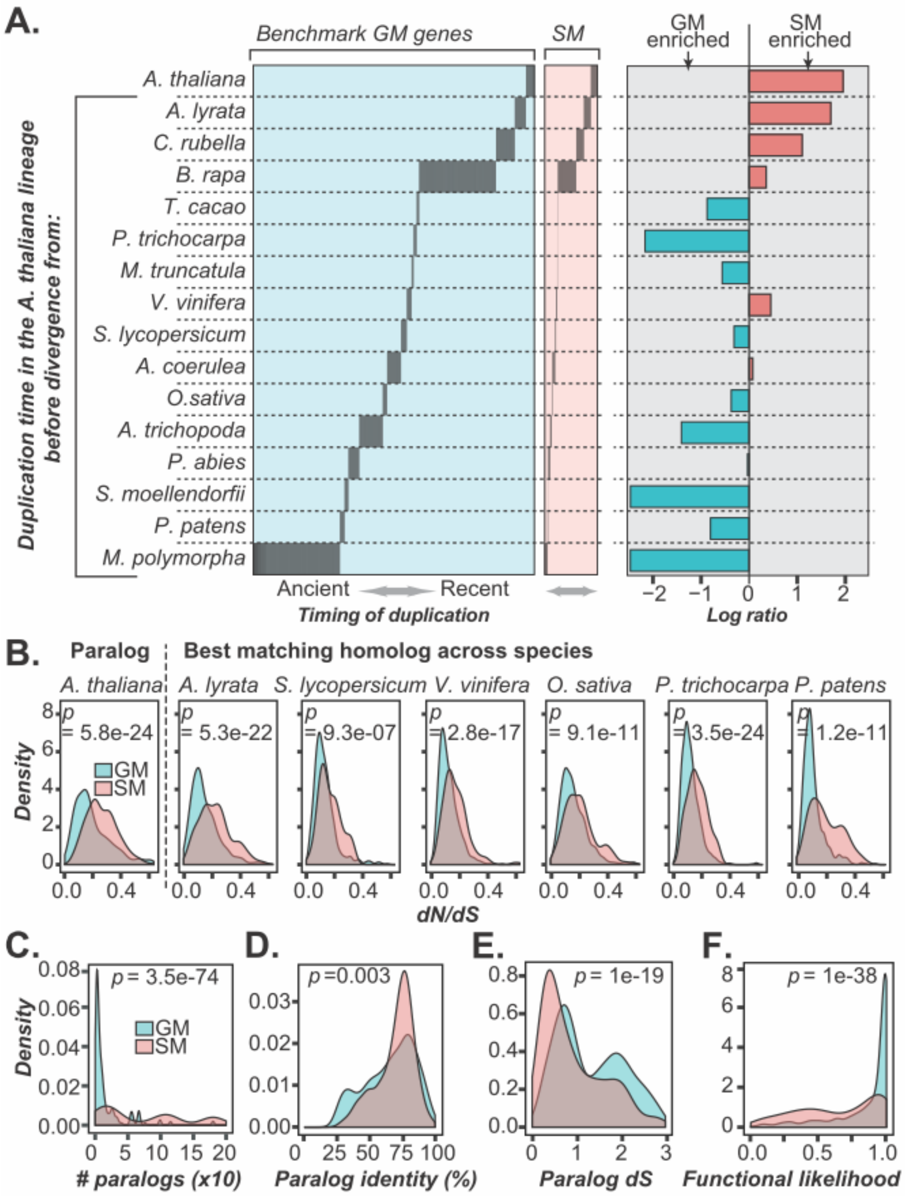

Figure 3. Differences in the timing, degree of selective pressure, and gene family expansion of duplication between SM and GM genes. (A) The distribution of duplication time points (y-axis) for each GM/SM gene (x-axis). Left/middle panel: a black line indicates that the GM (left panel) or SM (middle panel) gene in question likely have duplicated prior to the divergence between the *A. thaliana* lineage and the species lineage left to the black line. Species order: based on their divergence time from *A. thaliana.* Right panel: each bar represents the log2 ratio (x-axis) between the proportions of SM and GM genes duplicated at each duplication time point (y-axis). For full species name, see Methods. **(B-F)** Density plots showing SM (pink) and GM (blue) gene feature distributions. Test statistics were generated using Mann-Whitney U tests. **(B)** Median nonsynonymous substitution rate/synonymous substitution rate (*dN*/*dS*) values between *A. thaliana* SM/GM genes and their *A. thaliana* paralogs or best matching homologs in six other species, arranged based on their divergence time from *A. thaliana*. **(C)** The number of *A. thaliana* paralogs of SM or GM genes. **(D)** The maximum percent identity of an SM or GM gene to its paralogs. **(E)** The *dS* distribution between each SM or GM gene and its paralog. **(F)** The functional likelihood ranging from 0 to 1 that indicates the likelihood that a gene is under selection.

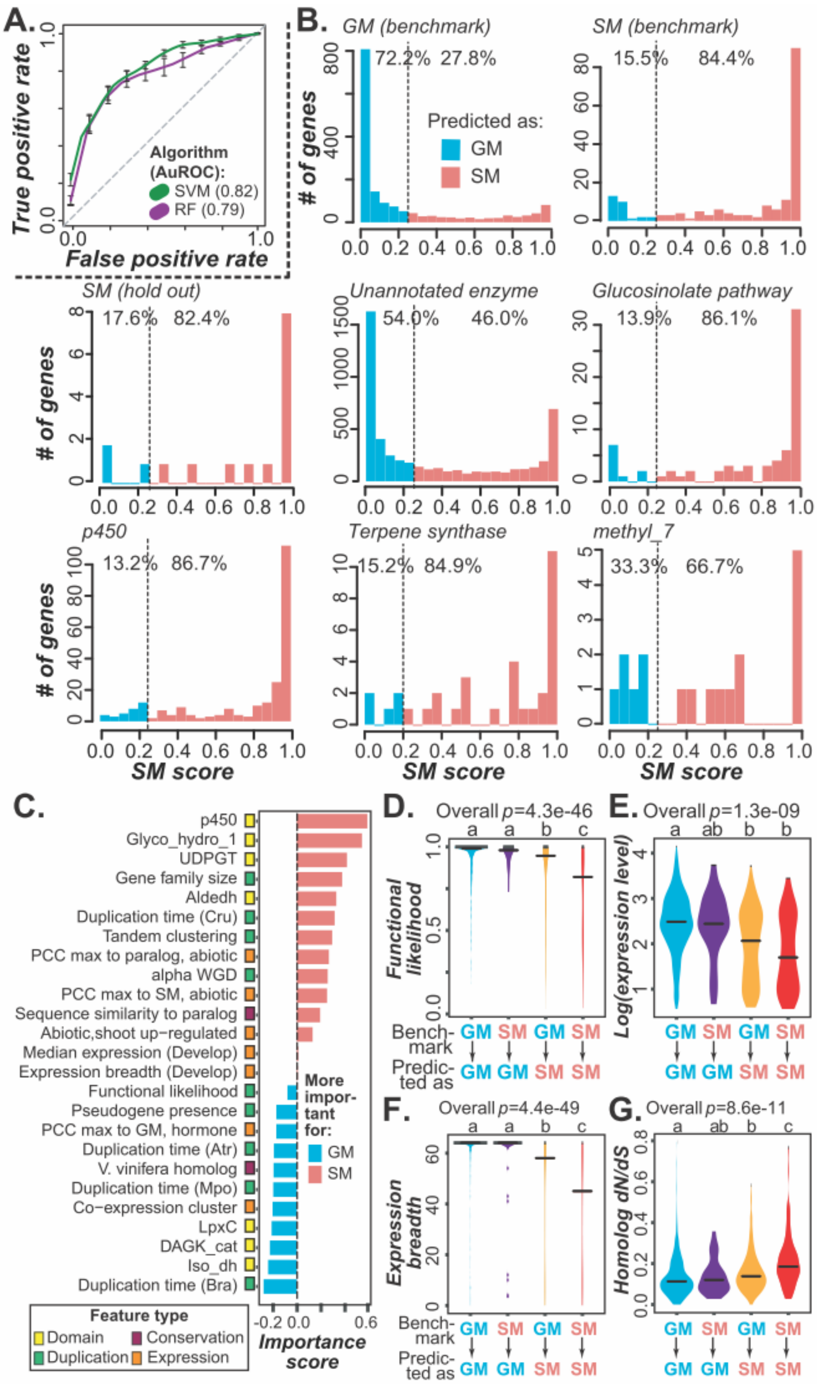

Figure 4. SM gene prediction model performance. (A) AuROC curves of models built with Support Vector Machine (SVM) and Random Forest (RF). **(B)** SM score distributions among benchmark GM, benchmark SM, hold-out SM, unannotated enzyme, glucosinolate pathway, p450, terpene synthase, and methyltransferase 7 (methyltr_7) domain-containing genes. Dotted line: SM score threshold (see Methods). Red and blue shading indicate genes predicted to be SM and GM genes, respectively. **(C)** Top 10 most important features for SM (red) and GM (blue) gene predictions. **(D-G)** Distributions of values of representative, predictive features of correctly and incorrectly predicted SM and GM genes. Overall *p*-values were from ANOVA. Tukey's test was then applied to pairwise classes (**Dataset S3**). **(D)** Functional likelihoods. **(E)** Logged median expression levels in the development dataset. **(F)**, Expression breadth in the development dataset. **(G)** *dN/dS* between *A. thaliana* and *A. lyrata* homologs.