SPECIALIZED METABOLISM AND STRESS RESPONSE: STUDIES IN PREDICTING GENE FUNCTION AND REGULATION

By

Bethany Maren Moore

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Biology—Doctor of Philosophy
Ecology, Evolutionary Biology and Behavior—Dual Major

ABSTRACT

SPECIALIZED METABOLISM AND STRESS RESPONSE: STUDIES IN PREDICTING GENE FUNCTION AND REGULATION

By

Bethany Maren Moore

One of the longstanding challenges in biology is to connect the wealth of genome data to the phenotypes it encodes. In plants, phenotypes can encompass a variety of traits, but one of interest is that of specialized metabolism, or the production of compounds unique to only specific lineages of plants rather than all plants. This interest comes from the wide range uses of specialized metabolites from defending a crop plant against insect herbivory to using the plant compound as a base for pharmaceuticals. The challenge faced is that due to their high diversity, among other reasons, the genes underlying the process of making specialized metabolites are not well characterized. In addition, specialized metabolism pathways can be induced by stress, however it is unclear how most SM pathways and stress responsive genes are regulated. Therefore, the research in this dissertation focuses on 1) How to identify the function of genes as being involved specialized metabolism 2) What characteristics are shared among specialized metabolism genes (SM genes) 3) How are SM genes and genes involved in response to stress regulated?

For the first two questions, chapters 2 and 3 use machine learning modeling to predict specialized metabolism (SM) genes in *Arabidopsis thaliana* and *Solanum lycopersicum* (tomato). A shared set of characteristics emerges as being important that includes expression features under biotic stress and in specific tissue types. Additionally, evolutionary and duplication characteristics were important where SM genes tend to be recently and tandemly duplicated, as well as less conserved than genes not in SM pathways. Using these characteristics to build a

machine learning model, 85.6% of SM genes in *A. thaliana* and 76.6% of SM genes in tomato were correctly predicted. Additionally, we show that the superior annotation in *A. thaliana* is able to make cross-species predictions in tomato as well as improve SM gene predictions relative to the model based only on tomato annotation. The improved model predicts 92.4% of SM genes in tomato correctly. Finally, machine learning is used to predict SM genes to a specific pathway.

For the third question, chapter 4 uses machine learning to predict how response to wounding stress is regulated and what regulatory elements are important for an SM pathway that is activated by stress. Important putative cis-regulatory elements were identified for genes differentially expressed under wounding stress and temporal patterns of regulation were discovered. Using machine learning, these putative cis-regulatory elements were found to be important in driving differential expression of genes at different time points after wounding. Additionally, regulatory elements were mapped to the genes in the SM pathway glucosinolate biosynthesis from tryptophan to determine element important for the regulation of this pathway under wounding stress. In this dissertation I examine computational approaches to identify gene function and regulatory mechanisms, highlighting the fact that machine learning can be a powerful tool to make challenging predictions.

This thesis is dedicated to Josh, my partner in life.
Thank you for always believing in me.

ACKNOWLEDGEMENTS

Thank you to the Shiu Lab for your support and helpful discussions. In particular, I'd like to thank present and past members Johnny and Anita Lloyd, Christina Azodi and Dom Del Ponte, Siobhan Cusack, Peipei Wang, Ming-Jung Liu, Nicholas Panchy, Sahra Uygun, Melissa Lehti-Shiu and Shin-Han Shiu for being essential to my progress and success as a Ph.D. student. In addition, I'd like to thank my committee members Rob Last, Gregg Howe, and David Lowry for their guidance and support. Finally, my family Josh and Melbourne were always there with love and encouragement, and I cannot thank them enough.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION

Understanding the link between genes and phenotypes

With the ever-increasing amount of available genome data, a major challenge in the field of genetics is to link genes with the phenotypes they produce (Dowell et al., 2010). Phenotypes are influenced by both genetic and environmental factors, and can encompass a variety of traits (Großkinsky et al., 2015). For example, in plant science a phenotype can be structural, from the whole plant level, such as height or yield to the physiological level of the production of certain metabolites via a metabolic pathway or a cellular response. Traits at the structural level are influenced by those at the physiological level, for example when a set of enzymes make a metabolite that is a pigment, we see this pigment in the color of the fruit or leaf of the plant (Großkinsky et al., 2015). At the molecular level, phenotypes can include transcriptome and proteome data, which are traits that build physiological phenotypes. Despite many new phenotyping technologies, a significant gene-to-phenotype gap remains (Tuberosa et al., 2014).

Connecting genes to their phenotype can be challenging for a number of reasons. One is that plants exhibit plasticity, meaning that in different environments, the same genotype can exhibit different phenotypes (Tardieu et al., 2017). Thus, the expression of certain genes may only be seen under a certain stress or condition and large-scale experiments of different genetic backgrounds may be needed to uncover genes that may be beneficial under specific conditions. Another challenge is that traits are often quantitative, meaning several genes may be involved in producing a phenotypic trait (Tardieu et al., 2017). Additionally, many genomic markers may not actually be found in the causal gene or genes but may rather be linked to them. Thus, a region of the genome may be associated with a trait but not the specific gene(s) (Resende et al., 2014). For these reasons and others, predictive modeling is essential to connect genes to their phenotype.

Modeling via machine learning in biology

Recently, machine learning has emerged as a modeling system in biology. Due to the complex nature of biological systems, and now the overabundance of many data types, machine learning can inform the conversion of large, heterogenous data into biological knowledge by combining them into one model (Larrañaga et al., 2006). Different types of biological data - including genomic, transcriptomic, proteomic, gene networks, metabolomic, and evolutionary - can be used in machine learning models. Machine learning optimizes the performance of a model by learning from past "experience," or known examples that constitute training data. The objective of a machine learning model is to then predict unknown examples based on inferences from the training data (Larrañaga et al., 2006). In a model, instances, or examples of what you want to predict, are categorized with a specific label. Features, or properties of the instances, are then used to classify the instances into different classes (Libbrecht and Noble, 2015). For example, when predicting genes from DNA sequences, the instances would be sequences, some of which are labeled as genes and others are not. Features, or various properties of each instance, are then used to discriminate between sequences that are genes and those that are not. These could be properties such as whether the sequence includes a sequence common to transcription start sites, such as ATG, whether it is in an open chromatin region, or whether that sequence is conserved in other species. The machine learning algorithm then uses these features to build a model that distinguishes gene sequences from non-genic regions. An example of using this type of approach is in predicting essential gene function, where genes that are essential to an organism are predicted relative to genes that are non-essential in *Arabidopsis thaliana* (Lloyd et al., 2015). Based on a set of labeled known lethal genes, the study predicts 1,970 unknown genes to be lethal in *A. thaliana* by combining evolutionary, expression, co-expression, and duplication-

based data. This study also highlights the use of different machine learning algorithms such as Support Vector Machine (SVM) and Random Forest (RF).

Different algorithms can perform differently for a given data set, depending on the overall structure of that data set. I briefly outline the algorithms tried or used in this thesis, however other algorithms for supervised machine learning are available. SVM maps features into a high-dimensional space, where features are used as hyperplanes to separate the data into different labels (Kotsiantis, 2007). RF, in contrast, is based on a series of decision trees, which use features as nodes in the tree to classify instances (Breiman, 2001). Each decision tree is made from random subsets of the training data, and the prediction is made from combining all trees into a 'forest,' where each tree votes for the label of the instance, and the most popular label is assigned to the instance (Breiman, 2001). I chose these algorithms out of a selection of other algorithms in our pipeline (https://github.com/ShiuLab/ML-Pipeline) because they were most compatible with the datasets I used and either RF or SVM consistently performed the best. It is important to understand that the algorithms, while having the ability to take on huge datasets, also assign weights to features, allowing the models to be interpreted. Thus, how important the feature is to your prediction can be determined with some caveats. One is that features are assumed to be independent, thus correlated features may have an impact on interpretability. For example, if features A and B both behave comparably in separating the instances into their respective labels, but feature A is slightly better than B, feature A will be interpreted as a significant contributor by the model. However, feature B, because the instances have already been mostly separated by feature A, drops in significance. Therefore, only one of the features will appear significant even though they both contribute similarly to the model.

Modeling with machine learning can be supervised, as was described in the above sample

where certain sequences are either labeled as genic or non-genic. In contrast, machine learning can also be used in an unsupervised manner, where instances are not labeled at all. Unsupervised machine learning allows for the algorithm to determine how best to group or label the data (Libbrecht and Noble, 2015). This type of machine learning is typically used in co-expression studies to determine gene function and is called clustering. For example, in a study to determine genes that belong to the same metabolic pathway, different clustering algorithms were used to cluster genes into different groups according to their expression patterns (Uygun et al., 2016). This type of study follows guilt-by-association logic where genes expressed the same way under different conditions will likely belong to the same pathway.

Different algorithms can be used in unsupervised learning such as k-means, c-means, or hierarchal clustering, all of which use a type of distance measure to determine if a sample should be placed in a group. For example, partitioning clustering algorithms like k-means or c-means partition the data into clusters based on minimizing the within-group sum of squares. Using these algorithms, Euclidean distance between samples is measured, and samples are added to a group until the sum of squares (based on Euclidean distance) no longer decreases (Larrañaga et al., 2006). In contrast, hierarchal clustering uses either an agglomerative or divisive algorithm, where the agglomerative starts with N groups and merges the 2 most similar clusters based on distance, proceeding to build a tree until all groups are merged. Divisive clustering starts with one cluster and divides the cluster into the most different clusters based on distances, then builds a tree out to N clusters (Larrañaga et al., 2006). Unsupervised machine learning modeling is emphasized when we don't know what we are looking for in the data.

Machine learning has several advantages over traditional statistical modeling. Much like Bayesian statistics, machine learning algorithms give a probability score to a sample being in a

certain group, or the alternative hypothesis. This is unlike the traditional *p*-value which calculates the probability of a sample to fall into the null hypothesis distribution, or outside of a given group. Machine learning also combines heterogenous data, which can be incorporated in a non-linear fashion. For example, both binary and continuous data can be used, with differing random distributions (normal or otherwise) into one model. Machine learning can also take thousands of features and learn which are relevant in predicting a class via a heuristic process (Larrañaga et al., 2006), therefore features not previously known to be important to a given model may turn out to be influential in the prediction. Feature selection can also be used to choose which features are best for distinguishing instances. Then, only these features will be used to build a model, which is often superior to a model with all the features. This can be useful to better understand the significant features so that the biology behind them can be interpreted, but also to simplify the model by removing noisy and redundant features (Libbrecht and Noble, 2015). Finally, machine learning models can be applied to unknowns not present in the model. This dissertation highlights modeling using machine learning in two important areas of biology: the discovery of gene function and identifying elements important for regulating genes.

Specialized metabolism: definitions, significance, and evolution

The study of metabolism, where various metabolites are considered to be phenotypes of a particular plant, is of great interest in the plant science field. Plant metabolism can be divided into two categories: primary or general metabolism, and specialized metabolism. Primary or general metabolism includes metabolites synthesized by all plants, and these metabolites number in the range of 10,000 (Pichersky and Lewinsohn, 2011). In contrast, specialized metabolites are called so because they are unique to a specific species or lineage of plants but are not found in all plants. These types of metabolites are far more diverse, with number estimations at

approximately 200,000 for all plants, with estimates for individual species correlating with the number of genes in that species (Pichersky and Lewinsohn, 2011). Specialized metabolites (SMs) are a result of adaptations of a plant to a particular environment (Hartmann, 2007). They are involved in a diverse array of functions, from defense against a pathogen or insect, to attracting pollinators to flowers or seed dispersers to fruit (Pichersky and Lewinsohn, 2011). For these reasons, many specialized metabolites are studied for agricultural purposes. For example, specialized metabolites found in species of wild tomato, such as acyl sugars, resist arthropod pests commonly found in tomato crops, such as spider mites and whitefly (Alba et al., 2009). Other specialized metabolites, terpenoids, variants of which are found in most plants, protect against fungal or bacterial pathogens, in addition to insects. Many plant species rich in terpenoids are common spices such as mint, basil, oregano, rosemary, and thyme (Freeman, 2008), which not only gives them resistance to certain pests but also the culinary flavors they are known for. Overall, finding genetic resistance to problematic agronomic pests can have beneficial effects, such as reduced use of pesticides in agriculture.

In addition to conferring desirable agronomical traits, many specialized metabolites from plants are also used to derive medicinal compounds. Historically, pharmaceuticals were almost exclusively derived from plants, until the late 19th century when the first drug, aspirin, was synthesized chemically (Schmidt et al., 2008). Despite advancements in synthetic chemistry, compounds from natural products still are used to derive medicines semi-synthetically. Of the medicinal compounds in current use, 25% are botanically derived, including Taxol, an anti-cancer drug, and morphine, an analgesic (Schmidt et al., 2008). It is thought that plants are still able to make more complex molecules and more metabolic diversity than synthetic chemistry, thus making them continually useful in drug discovery (Schmidt et al., 2008). There are many

specialized metabolites currently used as medicines. These include tropane alkaloids, SMs from the Solanaceae family, that can affect many aspects of the central nervous system and have been used to treat Parkinson's disease epilepsy, as vaso-dilators, and as local anesthetics, among many other uses (Grynkiewicz and Gadzikowska, 2008). Another drug made from specialized metabolites is the anti-malarial drug from the plant *Artemisia annua*. Although semi-synthetic versions are feasible, the majority of the anti-malarial product, artemisinin, still comes from the plant (Shen et al., 2016). A better understanding of the biosynthetic pathways that drive the production of these metabolites may help increase production of the medicinally derived compound and reduce overall costs of the pharmaceutical. The significance of specialized metabolism in plants merits further investigation into the genetic basis for these important molecules. From agricultural to medicinal, specialized plant metabolites have a diverse range of applications.

Specialized metabolites are not found in all plants, so it is important to understand how they have evolved. Many specialized metabolites are derived from gene duplications and represent a diverging point from which plant families or plant species diverge from their ancestors. While gene duplication is not the only mechanism by which new gene function can arise, it is one of the most prevalent in plant genomes (Panchy et al., 2016). Genes that are duplicated are most often lost or pseudogenized, but gene duplications can result in neo- or sub-functionalization, where one duplicate may retain its original function while the other develops novel function, or the duplicates retain different parts of the original function of the protein (Panchy et al., 2016). An example of an SM pathway that was a result of gene duplication is the divergence of the glucosinolate biosynthetic pathways in the Brassicaceae family. Glucosinolates represent a well-known specialized metabolite, as they are present in the model plant

Arabidopsis thaliana. The majority of genes in the glucosinolate pathways were derived from a whole genome duplication event at the base of the Brassica family, followed by lineage-specific tandem duplication events. Candidate glucosinolate genes were identified based on their duplication mechanism and timing (Hofberger et al., 2013). This study emphasizes that evolutionary characteristics and duplication events should be considered when attempting to identify the function of a gene.

Predicting gene function: the challenge of predicting specialized metabolism genes

In this section, we focus on identifying gene function, where the function of interest is specialized metabolism. Estimates of the number of genes involved in specialized metabolism are anywhere from 10-20% of genes for a given species, indicating that for the most well-annotated plant species, *Arabidopsis thaliana,* there may be around 1,750-3,500 genes involved in specialized metabolism (Pichersky and Lewinsohn, 2011), however this is a very rough estimate based on the approximate number of specialized metabolites known in *A. thaliana*. Currently in *A. thaliana* there are a little under 400 enzymatic genes annotated as being specialized metabolites and around 500 genes 'dual-annotated' as having both general and specialized metabolic functions by either Gene Ontology or AraCyc, (Moore et al., 2019). Together, this indicates that anywhere from 800 to 2,500 genes, which are estimated to have specialized metabolism function, currently do not have this annotation. Additionally, SM genes are often derived via duplication from GM genes (for examples, see Shoji and Hashimoto, 2011; Ning et al., 2015). Because of this, SM and GM genes often belong to the same gene family, making them difficult to distinguish. For example, the cytochrome p450 family in Arabidopsis contains 81 genes that are classified as SM and 51 genes classified as GM (Moore et al., 2019). Slight changes in sequence can alter the function of an enzyme, making the function specialized

while the ancestral function was generalized, but not all sequence alterations result in specialized function (Schenck et al., 2017). Also, convergent evolution, where two enzymes with different ancestral functions evolve the same function, or the loss of function in one duplicated enzyme but not the other, may make an enzyme specialized (Pichersky and Lewinsohn, 2011). Therefore, it is challenging to detect whether an enzyme is specialized, even if it is in the same gene family and has high similarity to other SM genes. Additionally, as noted above, genes may have both specialized and general functions. This may be a result of genes upstream in a SM pathway being more general, in that they are in other general or specialized metabolic pathways, or they produce general metabolites that are then used in downstream specialized pathways. Therefore, a continuum of general to specialized may apply to the annotation of metabolic genes. In this dissertation, I outline ways in which high confidence predictions of gene function can be made *in-silico* using machine learning.

Plant response to stress and gene regulation

Plant phenotypes, including specialized metabolites, can be cryptic, in that they are not readily displayed under ambient conditions, or in all parts of the plant. For example, some specialized metabolism pathways are induced by stresses such as wounding by an insect. The stress triggers the plant defense hormone, jasmonic acid (JA), among other signals, and JA activates the transcription factor MYC2. The MYC2 then activates specialized metabolism pathways, which increase specialized metabolites that can help defend the plant against the insect (Colinas and Goossens, 2018). Thus, the specialized metabolite may not be readily observed unless the plant is under a specific type of stress. Other specialized metabolites are only found in certain plant tissues. For example, nicotine is synthesized in the roots of Nicotiana tabacum before it is transported to the vacuoles in the leaves (Erb et al., 2009) and artemisinin is found

only in glandular trichomes in the Artemisia annua plant (Shen et al., 2016). Thus, the spatial and temporal components of a phenotype can be critical to understanding how they are regulated. Signals from the stress cause reprogramming of the gene expression network through a variety of mechanisms. For example, transcription factors are induced and bind to specific transcription factor binding sites (TFBS) on the DNA, which then induce stress-responsive genes. These TFBS elements can be cis-acting or trans-acting, where cis-elements are directly adjacent to the gene they regulate, in the promoter for example, and trans-elements are found on remote areas of the DNA away from the gene body, such as an enhancer (Colinas and Goossens, 2018). Other modifications in the structure of chromatin can regulate gene expression. Chromatin structures consist of DNA wrapped around histone marks and packaged tightly into nucleosomes. This condensed version of chromatin is referred to as heterochromatin, while DNA that is not wound as tightly around histone proteins is euchromatin. The fairly open euchromatin has higher amounts of gene expression than the heterochromatin, mainly because it is readily accessible to transcription factors (Asensi-Fabado et al., 2017). Modifying histone proteins with acetylation or methylation can cause them to become more open or more closed, and this may be dependent on stress signals (Asensi-Fabado et al., 2017). Because of the many levels of regulation, and the cryptic response of stress-related genes, it is challenging to determine how stress-responsive genes are regulated. The second part of this dissertation focuses on using machine learning models to identify important regulatory elements in stress-responsive genes.

Dissertation outline and significance

Genes underlying specialized metabolism provide targets for basic research but also for applications in the medicinal and agricultural realms. Given their range of uses and overall importance, identifying genes involved in specialized metabolism pathways remains a critical

goal in plant science. Chapters 1 and 2 represent significant advancements by: 1) using prediction models to identify genes belonging to specialized metabolism pathways and to annotate unknown enzymes as specialized or general metabolism genes in *A. thaliana* and tomato; and 2) quantifying the importance of expression, co-expression, evolutionary, duplication, and protein domain characteristics shared among SM genes and in contrast to GM genes. Chapter 2 goes further to assess the ability of a model built in *A. thaliana* to make cross-species predictions in tomato. Additionally, chapter 2 evaluates the ability of prediction models to classify SM genes into individual SM pathways. Overall, chapters 1 and 2 define distinguishing characteristics of SM genes and make predictions of SM genes in multiple species.

In addition, understanding how genes and SM pathways are turned on under a given stress is a continuing quest in plant research. Chapter 3 outlines models that evaluate the ability of known and putative cis-regulatory elements as well as open chromatin regions to regulate differential gene expression under wounding stress at different times. Additionally, chapter 3 describes regulatory elements important for regulating an SM pathway that is activated under wounding stress. This chapter represents how machine learning can be used to provide insights into important regulatory characteristics for specific clusters of genes. Overall, this dissertation delineates the uses of predictive modeling in biology, both in finding gene function and understanding how genes are regulated by using information *in silico*, to help narrow down experiments *in planta*.

REFERENCES

REFERENCES

Alba JM, Montserrat M, Fernández-Muñoz R (2009) Resistance to the two-spotted spider mite (Tetranychus urticae) by acylsucroses of wild tomato (Solanum pimpinellifolium) trichomes studied in a recombinant inbred line population. Exp Appl Acarol **47**: 35–47

Asensi-Fabado M-A, Amtmann A, Perrella G (2017) Plant responses to abiotic stress: The chromatin context of transcriptional regulation. Biochim Biophys Acta BBA - Gene Regul Mech **1860**: 106–122

Breiman L (2001) Random Forests. Mach Learn **45**: 5–32

Colinas M, Goossens A (2018) Combinatorial Transcriptional Control of Plant Specialized Metabolism. Trends Plant Sci **23**: 324–336

Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, Bernstein DA, Rolfe PA, Heisler LE, Chin B, et al (2010) Genotype to Phenotype: A Complex Problem. Science **328**: 469–469

Erb M, Lenk C, Degenhardt J, Turlings TCJ (2009) The underestimated role of roots in defense against leaf attackers. Trends Plant Sci **14**: 653–659

Freeman (2008) An Overview of Plant Defenses against Pathogens and Herbivores. Plant Health Instr. doi: 10.1094/PHI-I-2008-0226-01

Großkinsky DK, Svensgaard J, Christensen S, Roitsch T (2015) Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. J Exp Bot **66**: 5429–5440

Grynkiewicz G, Gadzikowska M (2008) Tropane alkaloids as medicinally useful natural products and their synthetic derivatives as new drugs. Pharmacol Rep 26

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

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**: 2155–2173

Kotsiantis SB (2007) Supervised Machine Learning: A Review of Classification Techniques. Emerg. Artif. Intell. Appl. Comput. Eng. Real Word AI Syst. Appl. EHealth HCI Inf. Retr. Pervasive Technol. IOS Press, Amsterdam, Netherlands, pp 3–24

Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, et al (2006) Machine learning in bioinformatics. Brief Bioinform **7**: 86–112

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

Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H (2015) Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes. Plant Cell **27**: 2133–2147

Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Lehti-Shiu MD, Last RL, Pichersky E, Shiu S-H (2019) Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci USA **116**: 2344–2353

Ning J, Moghe G, Leong B, Kim J, Ofner I, Wang Z, Adams C, Jones A Daniel, Zamir D, Last R L (2015) A feedback insensitive isopropylmalate synthase affects acylsugar composition in cultivated and wild tomato. Plant Physiol pp.00474.2015

Panchy N, Lehti-Shiu MD, Shiu S-H (2016) Evolution of gene duplication in plants. Plant Physiol pp.00523.2016

Pichersky E, Lewinsohn E (2011) Convergent evolution in plant specialized metabolism. Annu Rev Plant Biol **62**: 549–566

Resende MDV de, Silva FF e, Resende MFR, Azevedo CF (2014) Genome-Wide Selection (GWS). Biotechnol. Plant Breed. Elsevier, pp 105–133

Schenck CA, Holland CK, Schneider MR, Men Y, Lee SG, Jez JM, Maeda HA (2017) Molecular basis of the evolution of alternative tyrosine biosynthetic routes in plants. Nat Chem Biol **13**: 1029–1035

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

Shen Q, Yan T, Fu X, Tang K (2016) Transcriptional regulation of artemisinin biosynthesis in Artemisia annua L. Sci Bull **61**: 18–25

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

Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M (2017) Plant Phenomics, From Sensors to Knowledge. Curr Biol **27**: R770–R783

Tuberosa R, Turner NC, Cakir M (2014) Two decades of InterDrought conferences: are we bridging the genotype-to-phenotype gap? J Exp Bot **65**: 6137–6139

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 ComputBiol **12**:

CHAPTER 2 : ROBUST PREDICTIONS OF SPECIALIZED METABOLISM GENES THROUGH MACHINE LEARNING [1]

[1] The work on this chapter has been published:

Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Lehti-Shiu MD, Last RL, Pichersky E, Shiu S-H (2019) Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci USA **116**: 2344–2353

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 metabolism) genes through a detailed study of features including duplication pattern, sequence conservation, transcription, protein domain content, and gene network properties. Analysis of multiple sets 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 prediction model was established with a true positive rate of 87% and a true negative rate of 71%. In addition, 86% of known SM genes not used to create the machine learning model were predicted. We also demonstrated that the model could be further improved when we distinguished between SM, GM, and junction genes responsible for reactions shared by SM and GM pathways, indicating that topological considerations may further improve the SM prediction model. Application of the prediction model led to the identification of 1,220 *A. thaliana* genes with previously unknown functions, each assigned a confidence measure called an SM score, 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-encoding 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 incompletely characterized pathways responsible for their production. In addition, SM gene ancestors frequently played roles in GM. Here, we evaluate features distinguishing SM and GM genes and build a computational model that 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.

Introduction

Gene duplication and subsequent divergence/loss events led to highly variable gene content between plant species (Hanada et al., 2008; Panchy et al., 2016). These differential gain and loss events have given rise to diverse metabolic enzymes ranging from those involved in generally conserved, primary metabolic processes found in most species (referred to as general metabolism, or GM, genes), to those that function in lineage-specific, specialized metabolism (SM) (Hartmann, 2007; Chen et al., 2011; Pichersky and Lewinsohn, 2011; Chae et al., 2014). The proliferation of SM genes in plants has resulted in an overall far larger number of specialized than general metabolites. These 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) (Ehrlich and Raven, 1964; Chen et al., 2011; Ali and Agrawal, 2014). They are also 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 in tomato (Zhong, 2002; Giuliano et al., 2008; Howat et al., 2014). Thus, identification of the genes encoding enzymes that produce specialized metabolites (referred to as SM genes) is key to understanding the causes underlying the diversity of plant specialized metabolites as well as for engineering 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 (Milo and Last, 2012). Although many SM genes arise by duplication of GM genes (Shoji and Hashimoto, 2011; Ning et al., 2015) or other SM genes (Hofberger et al., 2013), 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 (Panchy et al., 2016), and minor

sequence changes can lead to substantially altered enzyme functions (Moghe and Last, 2015; Schenck et al., 2017). Third, SM genes may arise through lineage-specific loss of the GM function without duplication. Finally, convergent evolution may explain the presence of unrelated enzymes in different lineages that use the same substrate to make similar products (Pichersky and Lewinsohn, 2011). Consequently, it remains unresolved whether most plant enzyme genes are involved in GM or SM pathways, even in the best annotated plant species, *Arabidopsis thaliana* (Initiative, 2000; D'Auria and Gershenzon, 2005; Chen et al., 2011; Pichersky and Lewinsohn, 2011). Therefore, in recent years there has been an enhanced focus on identifying SM genes (Schlapfer et al., 2017; Wisecaver et al., 2017). Multiple properties have been shown to differ between SM and GM genes (Kliebenstein, 2008; Chae et al., 2014; Schlapfer et al., 2017; Wisecaver et al., 2017). For example, whole genome duplications (WGDs) and tandem duplications both contribute to metabolic innovations in glucosinolate biosynthesis genes (Edger et al., 2015). In addition, compared with GM genes, SM genes tend to have a more 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), higher degrees of expression variation, and higher degrees of co-expression. Co-expression with known SM genes (Wei et al., 2006; Wisecaver et al., 2017) or genomic neighborhood and gene-metabolite correlation (Higashi and Saito, 2013) were also used to predict SM pathway genes.

With the influx of more biochemical and -omic data, there is an increasing number of gene properties that can be evaluated for their utility in distinguishing SM/GM genes. Furthermore, the studies published to date have mainly focused on specific SM or GM pathways, raising the question of how SM/GM genes differ globally. This prompted us to examine 10,243

gene properties (referred to as features), including new features and those evaluated in early studies, falling into five categories (gene function, expression/co-expression, gene networks, evolution/conservation, and gene duplication) and evaluate the ability of each feature to distinguish SM genes from GM genes. Earlier studies revealed that the association between features and SM genes is far from absolute (Uygun et al., 2016) and — in most cases — the effect sizes (i.e. the extent to which these specific features can distinguish SM and GM genes) were not reported. To build on these studies, a machine learning approach (Schlapfer et al., 2017), which jointly considers all five categories of heterogenous features, was used to distinguish SM and GM genes. This approach led to machine learning models that were used to predict if an *A. thaliana* enzyme gene is likely an SM gene. Furthermore, we examined the properties of enzyme genes in cases where the annotations and predictions differed. Our findings provide a global estimate of SM gene content in the *Arabidopsis thaliana* genome, and the identified features may pave the way for further improvement of the modeling approach.

Results and Discussion

Benchmark SM and GM genes

Currently there are two major resources for plant SM and GM gene annotations: Gene Ontology (GO; (Botstein et al., 2000)) and AraCyc (Rhee et al., 2006). 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 annotated 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 35 SM genes (15% of GO- and 8.3% of AraCyc-annotated SM genes) overlapped (**Figure 2.1A**). While this is a

significantly higher degree of overlap than expected by chance (**Figure S2.1A, B)**, it indicates a greater inconsistency in SM annotation criteria than in GM annotation criteria between the GO and AraCyc datasets. Furthermore, 152 and 261 genes were annotated as both SM and GM in GO and AraCyc, respectively. This indicates that while SM and GM genes may have distinct properties, several genes can be both and their properties may not be distinct. Here we focus on cases that are not ambiguous, but later we delve into this gene set to see if genes involved in both SM and GM pathways can be uniquely classified.

To further assess the differences in AraCyc and GO annotations, we asked whether SM and GM genes annotated based on these two sources have different functional and pathway annotations and Pfam protein domains. We found that GO- and AraCyc-annotated SM genes have substantially different enriched GO categories (**Figure 2.1B, Dataset S1)**, AraCyc pathways (**Figures S2.1C, Dataset S1)**, and protein domains (**Figure S2.1D, Dataset S2**). In contrast to SM genes, GO- and AraCyc-annotated GM genes tend be over-represented in the same functional categories and pathways (**Figure 2.1B**). Considering the above findings, we defined three benchmark sets (**Dataset S1**). The first (benchmark 1) was defined to include as many annotated SM genes as possible. Here, 393 benchmark 1 SM genes were defined as the union of GO and AraCyc SM annotations that have Enzyme Commission (EC) numbers. Similarly, 2,226 benchmark1 GM genes are from the union of GO and AraCyc primary metabolism gene annotations associated with EC numbers. In the second set (benchmark2), we used only AraCyc annotations, which were likely better annotated because the focus of AraCyc is on metabolic pathways (SM=411, GM=1306, **Figure 2.1A**). In the third set (benchmark 3), we intersection between GO and AraCyc annotations (SM=35, GM=650, **Figure 2.1A**). When we examined which gene feature could distinguish benchmark SM and GM genes (described in the

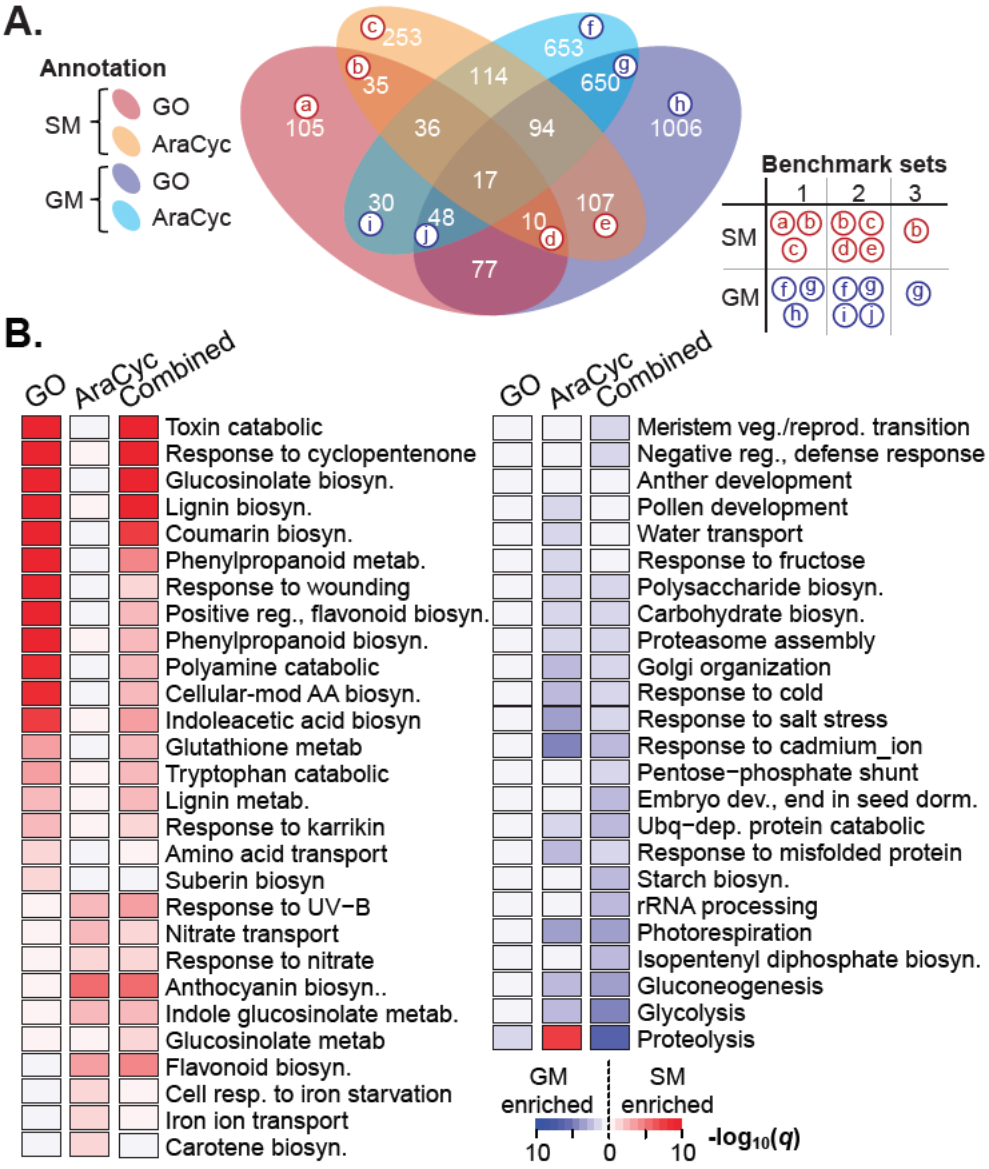


Figure 2.1. Gene Ontology and AraCyc annotation of specialized and primary metabolism genes.

(A) Overlap between Gene Ontology (GO)/AraCyc primary metabolism (PM) and secondary metabolism (SM) gene annotations. The number of genes in an intersection or in a complement set are shown. Three benchmark SM/GM gene sets were defined: benchmark 1 (Union), benchmark 2 (AraCyc), and benchmark 3 (Intersection) (see **Methods**). The table to the right shows the genes (labeled with lowercase letters in the Venn diagram) included in each benchmark set. **(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: represents 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.

following four sections, **Dataset S2**), the *p*-values from testing >10,000 features were highly correlated among the three benchmark definitions ($R^2 \geq 0.55$, **Figure S2.1E-G;** all Pearson Correlation Coefficients (PCCs) ≥0.74, **Dataset S2**). Therefore, we focus on comparing benchmark1 (union-based) and benchmark2 (AraCyc-only) genes, particularly when the conclusions (whether a feature can distinguish between SM and GM genes) were inconsistent.

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

A previous study showed that the expression of genes in some SM pathways tends to be more variable than the expression of genes in "essential pathways" (Kliebenstein, 2008). To further assess differences in SM and GM gene expression, we examined transcriptome datasets encompassing 25 tissue types (development dataset) and 16 abiotic/biotic stress conditions (stress dataset, see **Methods**; for all test *p*-values, see **Dataset S2**). In addition to confirming that benchmark2 SM genes tend to have higher expression variability (*p*=0.003, **Figure 2.2A**), we examined 23 additional expression features. We found that SM genes had significantly narrower breadths of expression (Mann Whitney U tests, for all benchmark sets: *p*<1e-35, **Figure 2.2A**), lower median expression levels (*p*=e-24, **Figure 2.2A**), and lower maximum expression levels (*p*=0.04, **Figure 2.2A**). These findings are consistent with the fact that SM genes have more specialized roles, whereas GM genes are involved in basic cellular functions (Hartmann, 2007; Chen et al., 2011). As expected with the established roles of some specialized metabolites in environmental interactions (e.g. (Steppuhn and Baldwin, 2007; Ali and Agrawal, 2014)), we found that benchmark1 SM genes tend to be up-regulated under a higher number of abiotic and biotic stress conditions compared with GM genes (all *p*<2e-7, **Figure 2.2B**), largely similar to the results based on benchmark2 (*p*=0.24~1e-8). Relatively fewer SM genes were down-regulated in the shoot under stress compared with GM genes (*p*=0.18~3.1e-5, **Figure 2.2B**),

likely reflecting a growth-defense tradeoff (Huot et al., 2014) where GM genes involved in house-keeping functions are down-regulated under stress and SM genes with roles in abiotic and biotic interactions are not. We do not, however, see the same trend in roots. Because CG methylation and histone modification can influence gene expression (Chan et al., 2005; Cedar and Bergman, 2009), we compared the numbers of these sites between SM and GM genes. We found that SM genes tend to have a lower degree of gene body CG-methylation than GM genes (Fisher's exact tests, p<3e-4, **Dataset S2**). On the other hand, the extent of histone modification did not significantly differ between SM and GM genes for seven of the eight histone marks (see **Methods**, **Figure S2.2A**).

Previous studies used expression correlation to evaluate how well genes in distinct SM pathways are correlated (Schlapfer et al., 2017; Wisecaver et al., 2017). To see if similar correlation measures could be used to distinguish SM and GM genes, we used maximum PCCs to evaluate expression correlation between each SM/GM gene and its paralogs (**Figure 2.2C**) as well as to other SM and GM genes (**Figure 2.2D**) in each of four expression datasets (abiotic stress, biotic stress, development, and hormone treatment). We found SM paralogs to have a significantly higher expression correlation than GM paralogs in all four data sets (Mann-Whitney U test, all p<0.05, **Figure 2.2C**). Because SM genes have undergone more recent expansion than GM genes (Hanada et al., 2008; Chae et al., 2014) and the degrees of sequence and expression divergence are positively correlated (Liu et al., 2011; Das et al., 2016), the higher expression similarities between SM paralogs than between GM paralogs may be partly explained by the more recent timing of SM duplication. We next looked at the maximum expression correlation between each SM gene and other SM genes (SM-SM) or GM genes (SM-GM), as well as between each GM gene and other GM genes (GM-GM) or SM genes (GM-SM). The expression

correlations ranked as follows: GM-GM > SM-GM > SM-SM > GM-SM (all benchmark1 $p<0.05$, but all benchmark2 $p>0.05$ for correlation in the development and biotic stress datasets, **Figure 2.2D**). The higher expression correlation for GM-GM compared with SM-SM is likely because GM genes tend to be more broadly expressed and at higher levels than SM genes (**Dataset S2**). The ratio between expression level variance and mean is higher for genes with lower expression levels, such as SM genes, which contributes to the comparatively lower correlation between these genes. Taken together, our findings indicate that expression correlation features can distinguish SM and GM genes.

Because pathway genes tend to be co-expressed and belong to the same co-expression cluster (Schlapfer et al., 2017; Wisecaver et al., 2017), we next assessed if benchmark1 SM and GM genes that belong to distinct pathways were members of distinct co-expression modules (**Figure 2.2E, Dataset S2**). Among these modules, 99 and 125 contained significantly more SM genes than randomly expected ($\alpha=0.05$) and are referred to as SM modules. Similarly, 125 GM modules were significantly enriched in GM genes ($p<0.05$). Therefore, a subset of benchmark 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 2.2E**). 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 2.2E**). This pattern reflects the fact that GM genes which function immediately upstream of an SM pathway may be co-expressed with genes in the SM pathway in question. Examples include 208 "junction" genes interfacing GM and SM pathways based on AraCyc annotations (**Dataset S2**)). These findings further highlight challenges in differentiating SM and GM genes globally using co-expression

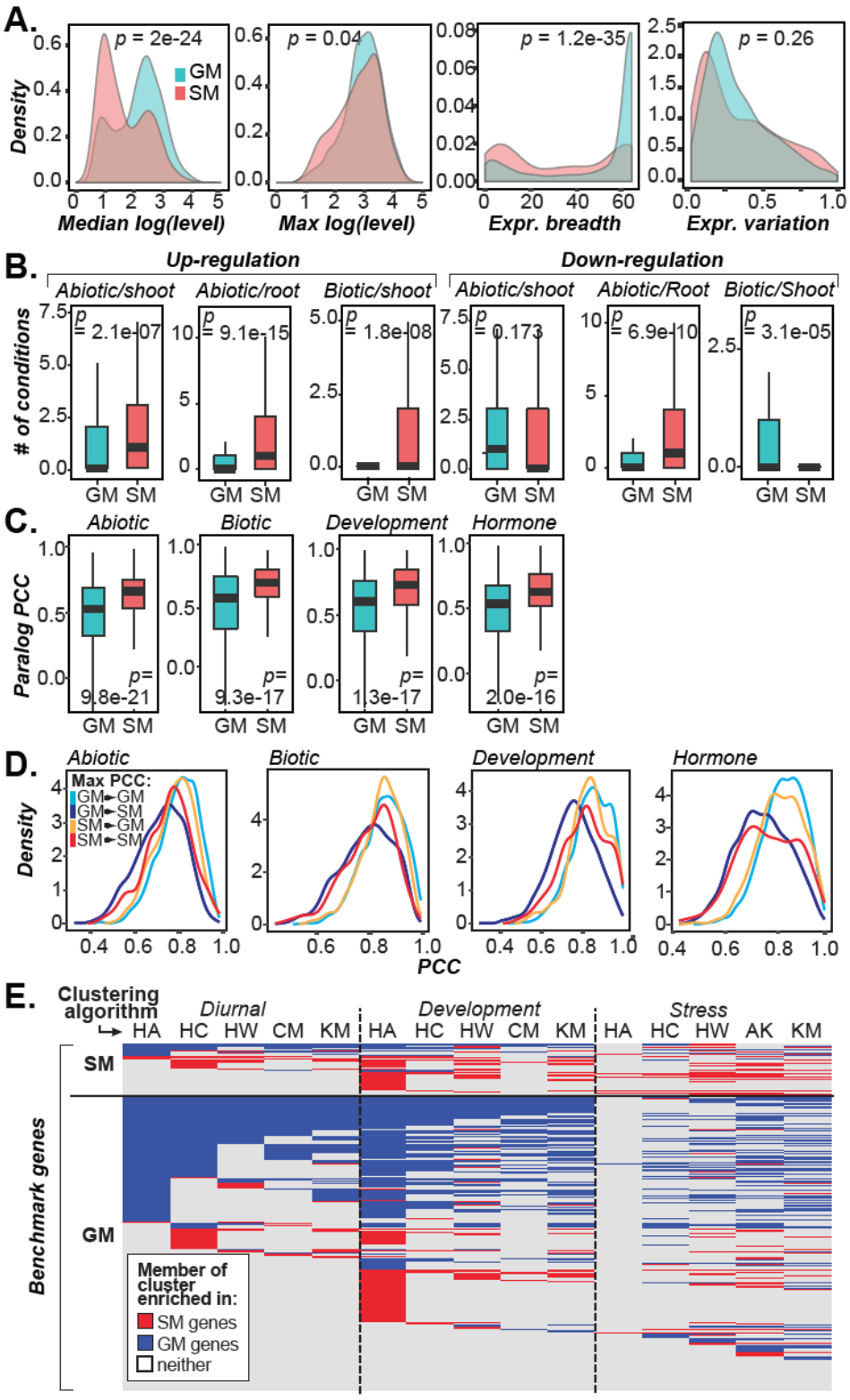


Figure 2.2. Differences in expression and co-expression characteristics of benchmark1 SM and GM genes.

(A) Distributions of SM (red) and GM (blue) gene expression-related values calculated from the

Figure 2.2 (cont'd)

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 stress (root and shoot) and biotic stress (shoot) datasets. **(C)** Distributions of maximum Pearson Correlation Coefficients (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 PCC between GM-GM (light blue), GM-SM (dark blue), SM-GM (orange), and SM-SM (red) gene pairs using the same expression datasets as in **(C)**. **(E)** Clustering of SM and GM genes based on their expression patterns in the diurnal development and stress datasets using 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).

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 (Arabidopsis Interactome Mapping Consortium, 2011) and found that SM genes have a significantly smaller number of physical interactions (mean = 1.25) than GM genes (1.84, benchmark1: p=0.03, benchmark2: p=3.85e-8, **Figure S2.2B**). The smaller number of SM gene interactions is not because SM genes have shorter coding regions (SM>GM, all p=0.004, **Figure S2.2C**) but is possibly due to the presence of fewer protein domains (SM<GM, benchmark1: p=0.35, benchmark2: p=4.3e-6 **Figure S2.2D**). Our finding that significantly fewer protein-protein interactions are known for SM proteins is consistent with SM genes having more specific functions than GM genes (Hartmann, 2007). It is also possible that there have been more interaction experiments for GM genes, or that GM genes tend to function in larger pathways compared with SM genes. Although GM genes tend to have more interactions than SM genes, SM genes with certain domains, such as cytochrome P450, have a higher median number of gene interactions (99.5) when compared with their P450 GM counterparts (15.0). Thus, proteins in some domain families may deviate from the general trend we uncovered.

Next, we examined the same relationships using the AraNet functional network (Lee and Lee, I., 2017), which connects genes with likely similar functions through the integration of multiple datasets, including expression and protein-protein interaction datasets. While the number of protein-protein interactions was significantly higher for GM genes relative to SM

genes (**Figure S2.2B**, all p<0.05), the differences in network connectivity between GM and SM genes in benchmark1 were not significant (*p*=0.139, **Figure S2.2E**) but were significant for benchmark2 genes (*p*=0.027). either were not significant or were marginally significant. AraNet considers multiple gene features including protein interactions, co-expression, shared domains, and homologous genes to construct gene networks, so it is not surprising that this result differs from that for analysis of only protein-protein interactions. These findings suggest that the amount of network connectivity is dependent on the type of network, and this may be useful for distinguishing between SM and GM genes. We should also note that the results from the benchmark1 and 2 sets are inconsistent, highlighting the impact of the benchmark definition on our analyses. In particular, benchmark1 *p*-values were higher than those of benchmark2, despite the fact that benchmark1 was substantially larger and would have lower *p*-values compared to a smaller dataset with the same effect sizes. This suggests that the AraCyc-only-based benchmark2 is likely of higher quality.

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 (Steppuhn and Baldwin, 2007; Ali and Agrawal, 2014; Brachi et al., 2015). In contrast, GM genes, which are involved in ancient and stable metabolic functions such as photosynthesis, are expected to be more highly conserved (Puthiyaveetil et al., 2010) and experience stronger negative selection (De Smet et al., 2013; Lloyd et al., 2015). An earlier study found a high degree of genetic variation in glucosinolate genes across *A. thaliana* accessions (21). Here, by comparing SM to GM genes globally, we found that SM genes tend to have higher nucleotide diversities than GM genes (*p*=3.9e-19, **Figure S2.3B**). In addition, we analyzed 15 evolutionary features based on within species and across species comparisons of SM and GM genes. First, 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 S2.3A**). However, consistently fewer SM genes (14.8-54%) have homologs across species than GM genes (27-76%) (all p<2e-4, **Figure S2.3A**). In addition, as expected for lineage-specific functions, only 0.94% of SM genes have homologs in core eukaryotic genomes (Tatusov et al., 2003) compared with 14.7% of GM genes (**Figure S2.3A**). Finally, we determined the timing of GM and SM duplications over the course of land plant evolution using sequence similarity to determine the most recent duplication point (see **Methods**). We found that 75% of SM genes were products of duplication events after the divergence between the *A. thaliana* and *B. rapa* lineages compared with only 40% of GM genes (**Figure 2.3A**), indicating that SM genes tend to be more recently duplicated relative to GM genes. Additionally, 25% of SM genes were duplicated after the *A. thaliana-A. lyrata* split, compared with only 7% of GM genes (**Figure 2.3A**). Thus, SM genes have higher duplication rates but do not persist in the long run, leading to the observation of fewer homologs across species.

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 2.3B**) compared with GM genes. Together with other measures of selection (**Figure S2.3C, D**), both within- and cross-species comparisons suggest that SM genes are under weaker negative selection relative to GM genes. One reason for this pattern may be that these SM genes initially experienced positive selection (higher rate than GM) followed by negative selection (similar to GM). This would result in SM genes having a higher rate of evolution than GM genes, with the appearance of weaker negative selection. Another possible reason for this pattern is that some of

these SM genes may have experienced strong negative selection (similar to GM) but are now neutrally evolving. This may be because the selective agent (e.g. a particular environmental factor) previously contributing to the selection no longer exists. 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 (Hartmann, 2007).

Duplication mechanisms and genomic clustering of SM and GM genes

Gene duplication mechanism, such as whole genome duplication (WGD), tandem duplication, and dispersed duplication, may impact subsequent functional divergence and ultimately influence whether a duplicate is under selection and retained (Panchy et al., 2016). For example, genes in a few SM pathways, such as aliphatic glucosinolate biosynthesis, tend to be tandemly duplicated and have a higher degree of expression variation (Kliebenstein, 2008). To assess if SM and GM genes differ in their post-WGD retention rate, we compared the number of GM and SM WGD duplicates in the *A. thaliana* lineage. Although two different glucosinolate pathways arose in the α WGD event ~50 million years ago (Hofberger et al., 2013), these two pathways do not lead to a significantly higher number of SM WGD duplicates compared to the number of GM WGD duplicates. This indicates that SM genes from multiple SM pathways (not just those involved in glucosinolate metabolism) are not more likely to be derived from WGDs than GM genes (benchmark1 p=0.1, benchmark2 p=0.85, **Figure S2.4A**). This suggests that the likelihood of long-term retention of SM and GM WGD duplicates does not appear to differ significantly. In contrast, significantly more SM genes tend to be tandem duplicates than GM genes (p<2e-43, **Figure S2.4A**). Genes involved in response to the environment are more likely to be tandem duplicates (Rizzon et al., 2006; Hanada et al., 2008), and tandem duplication potentially allows for rapid evolution of SM gene families that are subject to selection in variable

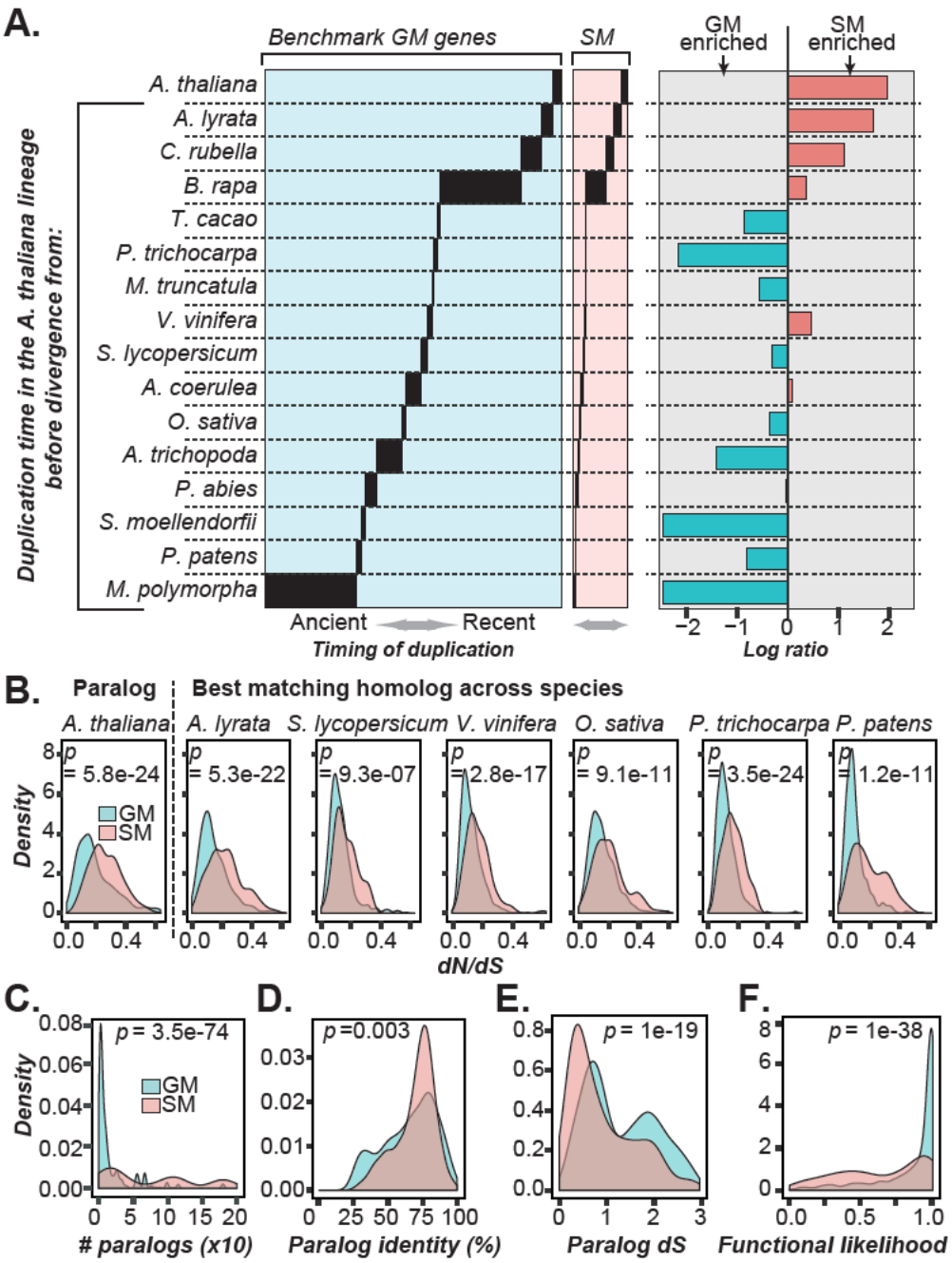


Figure 2.3. Differences in the duplication timing, degree of selective pressure, paralog-related features, and functional likelihood between benchmark1 SM and GM genes.

(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 duplicated prior to the divergence between the *A. thaliana* lineage and the species lineage to the left of the black line. Species order: based on the time of divergence 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 names, see **Methods**. **(B-F)** Density plots showing SM (pink) and GM (blue) gene feature

Figure 2.3 (cont'd)

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 the time of divergence 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, which indicates the likelihood that a gene is under selection.

environments.

The numbers of paralogs and pseudogenes were used as measures of the degree of SM and GM gene gains and losses, respectively. Our analysis revealed that SM genes tend to have more paralogs (p<3e-72, **Figure 2.3C**), higher sequence similarities to their paralogs (benchmark1: p=3e-3, benchmark2: p=0.3 **Figure 2.3D**), and lower synonymous substitution rates (dS) (p<2e-19, **Figure 2.3E**) compared with GM genes. Furthermore, a higher percentage of SM genes duplicated since *A. thaliana* diverged from *A. lyrata* (p<4e-8, **Figure S2.4B**), and SM genes tended not to be found in single copies (p<1e-3, **Figure S2.4C**). These findings all point to more recent expansion of SM gene families. We also compared the functional likelihood, which is a measure of how likely it is that a gene is functional and, thus, under selection (Lloyd et al., 2015), between SM genes, GM genes, and pseudogenes. Interestingly, the functional likelihoods of SM genes are significantly lower than those of GM genes, but higher than those of pseudogenes (ANOVA, Tukey's test, p<2e-16, **Figure 2.3F, Figure S2.4E**). Genes under strong negative selection have high functional likelihoods that are close to one, whereas pseudogenes tend to have values close to zero (Lloyd et al., 2015). In addition, most pseudogenes are eventually removed from the genome (Balakirev and Ayala, 2003) and tend not to be under selection (Moghe et al., 2014). Our finding that SM genes tend to have lower functional likelihood is consistent with the hypothesis 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 (between benchmarks, 9.8-11.1%) compared with GM genes (6.1-6.5%) is not significant overall (p=0.04~0.2, **Figure S2.4D**). Considering that SM genes tend not to have cross-species homologs (**Figure S2.3A**), this finding suggests that pseudogenes are too short lived to be adequate indicators of gene loss.

SM and GM genes that function in the same pathway are sometimes found in genomic clusters (Qi et al., 2004; Sakamoto, 2004; Osbourn, 2010; Schlapfer et al., 2017), and we used two approaches to compare the occurrence of SM and GM genes in close physical proximity. 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 (benchmark1: p=9.5e-121, benchmark2: p=0.02 **Figure S2.4F**) and GM genes tend to be close to GM genes (p<2e-5, **Figure S2.4G**). It is surprising that the p-values for SM clustering differ so greatly between benchmark sets. This may indicate that AraCyc annotation (benchmark 2) is of higher quality. In the second approach, we defined metabolic clusters identified using Plant Cluster Finder (Schlapfer et al., 2017), but the identified clusters were not enriched in either SM or GM genes (**Figure S2.4H**). Taken together, SM genes are more likely to be tandemly duplicated and tend to belong to large gene families. Our findings provide genome-wide confirmation of earlier studies (e.g. 2, 15, 22) that focused on a relatively small number of SM genes or pathways. These characteristics may be useful features in distinguishing SM and GM genes.

Machine learning model for predicting SM and GM genes

In total, we examined 10,243 features (summarized in **Dataset S3**) that differ widely in their ability to distinguish benchmark SM and GM genes. 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. An 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% False Positive Rate (FPR) and a 58% False Negative Rate (FNR). In addition, the majority of the features are not particularly informative

(**Dataset S3**), as the average AuROC for single feature-based models 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 using single features. To remedy this, we next integrated all 10,243 features, regardless of whether they were significantly different between SM and GM genes or not, to build machine-learning models for predicting SM and GM genes. We used machine learning because it allowed us to build an integrated model where multiple features were considered simultaneously. Integrated models offer better predictive power than individual features by lowering FNR and FPR.

Two machine learning algorithms, Support Vector Machine and Random Forest, were used to build predictive models using all three benchmark datasets (**Dataset S3**, **Figure 2.4A, Figure S2.6**, see **Methods**). The best performing SM gene prediction model was based on benchmark2 (AraCyc-only) and Random Forest (AuROC=0.87, FPR=29.4%, FNR=14.8%; **Figure 2.4A**). Randomizing SM/GM labels but maintaining the same feature values associated with the benchmark genes as the initial model resulted in AuROCs=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% of the data were 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 is an SM gene. Based on a threshold SM score defined by minimizing false predictions (see **Methods**), 85.6% of the training SM genes (**Figure 2.4B**) and 73.1% of the training GM genes were correctly predicted (**Figure 2.4B**), which reflects an improvement over the individual feature-based, naïve models.

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

In addition to the SM score, the machine learning result included a list of feature importance values, 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 2.4C**). Based on the AraCyc-only (benchmark2) model, the most informative features for predicting SM genes included specific protein domains as well as multiple gene duplication-related features, such as duplication mechanism (higher degree of tandem duplication), gene family expansion (larger family size), and higher degrees of correlation in expression between an SM gene and other SM genes or its paralogs **(Figure 2.4C)**. In addition, higher evolutionary rates were among the most informative for predicting *A. thaliana* SM genes based on comparison of an SM gene to its *Populus trichocarpa* and *Vitis vinifera* homologs, but not to homologs from more closely related species. This pattern may reflect the fact that at these time points (post divergence between *A. thaliana* and the *P. trichocarpa* or *V. vinifera* lineages) a number of SM genes experienced accelerated, potentially positive, selection that contributed to the diversification of major SM pathways. In contrast, wider expression breadth, measured using the development expression dataset, and higher connectivity in gene networks were among the most important features for predicting GM genes, indicating the more generalizable functions of GM genes and the tendency to interact with a greater number of genes/gene products relative to SM genes. Finally, specific histone marks as well as hierarchical, *k*-means, and approximate *k*-means co-expression clusters based on the stress, diurnal, and development datasets were informative for predicting both SM and GM genes (**Dataset S2**). While earlier studies established that genes belonging to a particular SM pathway

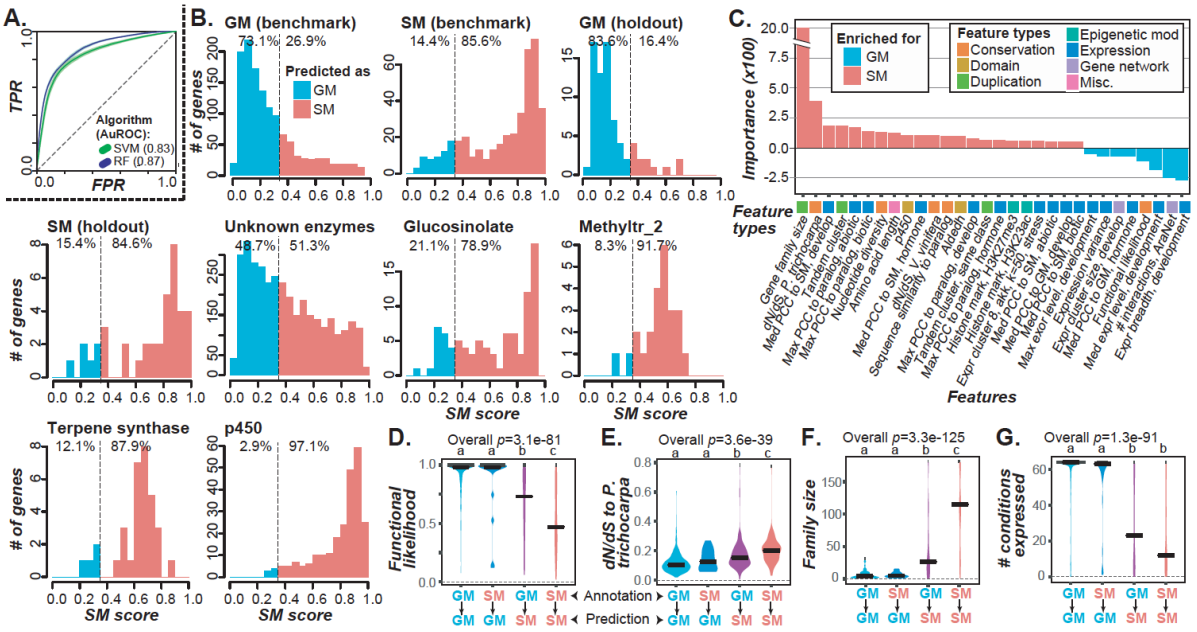


Figure 2.4. SM gene prediction model performance based on benchmark.

(A) AuROC curves of binary SM/GM prediction models built with Support Vector Machine (SVM) and Random Forest (RF) algorithms. TPR: true positive rate. FPR: false positive rate. **(B)** SM score distributions for benchmark GM, benchmark SM, hold-out SM (not included in models), unannotated enzyme, glucosinolate pathway, p450, terpene synthase, and methyltransferase 2 (methyltr_2) 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)** The most important features for SM (red) and GM (blue) gene predictions. **(D-G)** Distributions of the values of representative, predictive features for correctly and incorrectly predicted SM and GM genes. Black horizontal bar: median. Overall *p*-values are from Kruskal-Wallis tests used to evaluate differences among classes. The Dunn post hoc test was used to test differences between classes (**Dataset S3**). **(D)** Functional likelihood. **(E)** *dN/dS* between *A. thaliana* and *P. trichocarpa* homologs. **(F)** Sizes of the gene families the four categories of genes belong to. **(G)** Expression breadth in the development dataset.

tend to be co-expressed (Higashi and Saito, 2013; Wisecaver et al., 2017), our findings demonstrate that there are global differences in expression patterns and properties between SM and GM genes.

With the accuracy of the SM gene prediction models assessed through cross-validation and prominent features identified, we next applied these machine learning models to make predictions for 3,104 known enzymatic genes (with an EC number) not annotated to be SM or GM genes (**Dataset S1)**. Of these genes, 51% (1,592 genes) were predicted to be SM genes. We took three approaches to assess the accuracy of these SM and GM gene predictions. First, we intentionally held out 10% of both known SM and GM genes (**Figure 2.4B, Dataset S1**) from any model training. Upon application of the machine learning model, 84% and 85% of withheld GM and SM genes were correctly predicted, respectively, indicating that the model has an 84% True Positive Rate (or 16% 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 2.4B, Dataset S5**). When applying this new model to glucosinolate genes, 79% of known glucosinolate pathway genes were correctly predicted as SM genes. The FNR was 16% overall, which is much better than the 58% FNR when using the single best feature, gene family size.

Finally, methyltransferase, terpene synthase, and cytochrome P450 families were identified based on their respective protein domains (see **Methods**) and analyzed to test model performance within a specific family (**Figure 2.4B, Dataset S5**). These families were chosen because they tend to be associated with SM. To this end, we built three new models using our benchmark sets, excluding 'hold out' genes from the families we planned to predict. Upon

applying this model to each enzyme family, 97% of P450, 88% of terpene synthase and 92% of methyltransferase genes were predicted as SM genes (**Figure 2.4B**). Thus, these models predicted the majority of hold-out genes with known SM functions, glucosinolate pathway genes, and genes in enzyme families whose members predominantly play roles in SM pathways. In summary, our models allowed assessment of the relative importance of features in distinguishing SM and GM genes, as well as provided predictions for 1,217 SM genes among enzyme genes with no known SM/GM designation. In addition, our findings indicate that our models and this general approach are valuable for predicting unknown enzymes.

Characteristics of Mis-Predicted Genes

Although our SM prediction model performed well, 122 (16.7%) AraCyc annotated GM genes were mis-predicted as SM genes. In addition, 60 (15.3%) AraCyc annotated SM genes were mis-predicted as GM genes. To assess the properties of mis-predicted SM/GM genes, we determined how the values of a subset of the most informative features (**Figure 2.4C, Dataset S3**) differed between four gene classes defined based on the consistency between the gene annotation and the benchmark2 (AraCyc only)-based model prediction. These four classes included: (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 functional likelihood among these four gene classes follow the order: GM→GM > SM→GM > GM→SM > SM→SM (**Figure 2.4D**). The opposite pattern (SM→SM has the highest value) was observed for *dN/dS* values (**Figure 2.4E**), gene family size, (**Figure 2.4F**), the number of conditions expressed (**Figure 2.4G**), and values for other gene

features we examined (**Figure S2.5A-J**). Thus, in the SM→GM mis-predicted class, the annotated SM genes in fact possess multiple properties that are more similar to those of GM genes and vice versa, but no single feature can fully explain why these genes were mis-predicted.

These observations suggest that some of the mis-predicted benchmark genes (**Figure 2.4B**) may in fact be mis-annotated, or alternatively, they may point to a deficiency in our model (addressed in the next section). To assess how many of the mis-predictions are due to mis-annotation, we collated information from 25 genes with predictions (from the benchmark2-based model) matching the AraCyc annotations (GM→GM=4, SM→SM=21), and for 32 genes with predictions that were not consistent with their AraCyc annotations (SM→GM=20, GM→SM=12) (**Dataset S1**, **SI text**). We focused on genes in the P450/terpene synthase families because there is substantial biochemical and functional information available for these genes (Matsuno, et al., 2009; Chen et al., 2011; Renault et al., 2014). For mis-predicted genes, which were manually examined, five (42%) genes in the GM→SM class had supporting SM evidence (**Dataset S1**). In addition, 16 (80%) genes in the SM→GM class have supporting GM evidence (**Dataset S1**). These findings indicate that a subset of these genes (66%) are "mis-predicted" due to mis-annotation, not due to prediction errors.

For the benchmark1 set, which is based on the union between AraCyc and GO annotations, a similar percentage of the mis-predicted genes (5 of 11 GM→SM (45%) genes examined) were likely mis-annotated (**Dataset S1**). This is consistent with our finding that some SM genes enriched in AraCyc pathways and GO terms—such as carotene, leucine, suberin, and wax biosynthesis—are found across all major land plant lineages and should be considered GM genes (**Figure 2.1B**, **Figure S2.1C**). It is also 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 their *in vivo* functions. We should note that genes which were manually examined were mostly from the P450 and terpene synthase families. More enzyme families should be evaluated to obtain a more complete picture of the reasons behind inconsistent annotation and prediction. Together with the finding that nearly all (24/25) benchmark2 genes with consistent annotations and predictions had biochemical evidence supporting their SM or GM classification (**SI text**), these results further demonstrate the feasibility of 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 in addition to their annotated activities.

Impact of dual-annotated genes on model performance

The number of genes mis-predicted with our model which were not mis-annotated according to our manual examination (**Dataset S1**, **SI text**) indicate that our model can be further improved. Our original model focused on distinguishing SM and GM genes as binary classes but genes with both SM and GM functions were excluded. However, there are 261 genes (**Figure 2.1A**) annotated as belonging to both SM and GM pathways in AraCyc (dual-annotated or DA genes, **Figure 2.5A**). We thus explored the possibility that DA genes have properties distinct from SM or GM genes and should be considered a distinct class. We first compared the SM scores between SM, GM, and DA genes based on our AraCyc-only binary model. If DA genes belong to a distinct class that is neither SM nor GM, the SM scores of DA genes should have a unimodal distribution with a median close to 0.5. Contrary to this expectation, the SM score distribution of DA genes is bimodal, where some DA genes resemble SM genes and others resemble GM genes (**Figure 2.5B**). Thus, based on a GM vs SM binary model, DA genes do not appear to belong to a distinct class. These findings raise the question whether the dual annotation

is valid.

To assess whether our inability to distinguish DA genes from SM/GM genes is because the binary model is inadequate, we built a multi-class model assuming SM, GM and DA genes as three distinct classes and plotted the SM scores for each class in a ternary plot (**Figure 2.5C-F**). If the three classes of genes can be perfectly separated, then the highest gene density areas will be toward different corners of the ternary plots. Although the GM/SM/DA model has an F1-score of 0.51 (higher than the F1 of 0.33 for a random model) and an accuracy of 0.53, the inclusion of DA genes as a third class significantly diminished the ability of the model to separate SM (**Figure 2.5C**) and GM (**Figure 2.5D**) genes. Note that SM and GM genes are not well separated in the ternary plots (**Figure 2.5C, D**), but in the binary model, their SM score distributions are highly distinct (**Figure 2.5B**). In addition, the DA gene distribution in the ternary plot overlapped with the distributions of both SM and GM genes (**Figure 2.5E**), consistent with the bimodal SM score distribution observed among DA genes. Thus, the DA genes belong to two sub-classes, with each subclass resembling SM or GM genes, again raising the question whether the dual annotations in AraCyc are valid. Curiously, GM genes separate into two populations in the GM/SM/DA model where one population is located towards the GM corner of the ternary plot (arrow g1) and the second population (arrow g2) overlaps with areas of high SM (arrow s) and DA (arrow d) gene density (**Figure 2.5C**). Therefore, although this three-class model does not separate SM and GM genes well, it raises the question of how the two GM gene populations (g1/g2 peaks) differ and should be further examined.

Consideration of junction genes in predictive model building

Another potential way to improve our model is to consider metabolic network topologies. We hypothesized that SM and GM genes closer to pathway junctions (**Figure 2.5A**, see

Methods) are more likely to be mis-predicted. We identified junction reactions connecting 15 GM (upstream) and 20 SM (downstream) pathways. The 212 genes encoding enzymes responsible for junction reactions were referred to as junction (JC) genes. By further classifying JC genes based on the connectivity of their associated reactions, four topological sub-classes of junction genes were defined: 1→J→1: junction reactions, each connected with one reaction upstream and one reaction downstream, n→J→1: multiple upstream reactions but only one downstream reaction, 1→J→n: one upstream and multiple downstream reactions, and n→J→n: multiple upstream and downstream reactions (**Figure 2.5A**). Although junction genes as a whole also have a bimodal SM score distribution similar to that of DA genes (JC all, **Figure 2.5B**), the score distributions were distinct among the four topological sub-classes, indicating that network topology is a distinguishing characteristic between SM and GM genes. Considering that products of GM pathways serve as substrates for many other pathways, it is expected that GM genes functioning in junction reactions would be connected to multiple downstream pathways. Consistent with this, JC genes in the n→J→n and 1→J→n subclasses where n>1 tend to be more similar to GM genes (**Figure 2.5B**). In contrast, SM enzymes are more likely involved in incorporating substrates from multiple reactions and serve as the committed step for producing specialized metabolites with an expected n→J→1 topology. In addition, a typical SM pathway mostly contains a series of non-branching reactions that lead to specialized metabolite products and is also expected to have a 1→j→1 topology. Consistent with these expectations, JC genes in the n→J→1 and 1→J→1 subclasses are the most similar to SM genes (**Figure 2.5B**).

The GM/SM/JC 3-class model separated SM and GM genes significantly better (F1-score = 0.65, accuracy = 0.65, **Figure 2.5G-J**) than the GM/SM/DA model (**Figure 2.5C-F**), indicating that junction genes have unique characteristics and that some genes intersecting

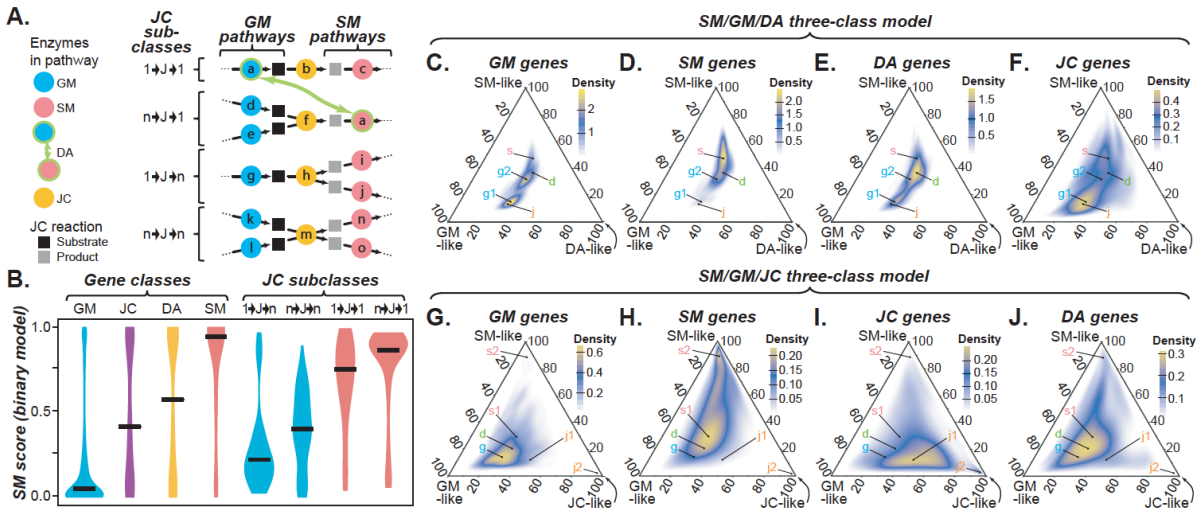


Figure 2.5. Three-class models for classifying SM/GM/DA and SM/GM/JC genes.

(A) Definition of DA (dual annotation) and JC (junction) genes. For JC genes, four sub-classes were defined based on the degree of connectivity, defined as the number of connecting reactions in the metabolic network based on AraCyc annotations. a-o: SM/GM enzymes that are annotated as GM (blue), SM (red), or DA (green outline), or are defined as JC (orange). JC reaction substrates and products are in black and gray, respectively. **(B)** Distributions of SM scores based on the binary model built using benchmark2 data for GM, SM, DA, JC (all), and JC subclass genes. **(C-F)** Ternary plots showing the SM/GM/DA model-based score distributions for GM **(C)**, SM **(D)**, DA **(E)**, and JC **(F)** genes. The g (blue), s (red), d (green), and j (orange) labels indicate the peak gene density areas (brighter yellow) occupied by GM, SM, DA, and JC genes, respectively. **(G-J)** Ternary plots showing the SM/GM/JC model-based score distributions for GM **(G)**, SM **(H)**, JC **(I)**, and DA **(J)** genes. The color scheme follows that in **(C-F)**.

annotated SM and GM pathways can be considered a separate class. In addition, the four topological sub-classes of JC genes are located in different areas in the ternary plots for the SM/GM/DA (**Figure S2.7A**) and GM/SM/JC (**Figure S2.7B**) models. We should emphasize that JC genes were defined based on a network constructed using AraCyc pathway annotations where the criteria for defining pathway boundaries may differ between research groups and/or annotators. Despite this, the GM/SM/JC model predictions demonstrate that JC genes are by and large distinct from GM/SM genes. Although we cannot be certain which JC genes were key enzymes in the committed steps entering SM pathways, the JC genes in the n→J→1 subclass is clearly a class of its own with most genes at the JC-like corner (**Figure S2.7B**). Taken together, these findings demonstrate that further categorization of SM and GM genes based on biologically motivated criteria, such as network topology, could help further distinguish different types of SM or GM genes leading to modest improvement of our models. In addition, the binary classification of SM and GM genes, while meaningful, can be an over-simplification. Finally, the consideration of additional topological characteristics (e.g. pathway depth, terminal reaction) and additional biochemical features (e.g. substrate and product identities) may lead to further improvements in SM and GM predictions.

Conclusions

Machine learning models built using genomic features show considerable promise in predicting the functions of unclassified or unannotated genes (Lloyd et al., 2015; Schlapfer et al., 2017). Prior to establishing such models for predicting SM and GM genes, we first explored how SM and GM genes in *A. thaliana* differ in >10,000 conservation, protein domain, duplication, epigenetic, expression, and gene network-based features. Most of these features have not been examined by other studies contrasting SM and GM genes. We demonstrated that machine

learning models in which these features are integrated to predict SM and GM genes perform well based on cross-validation performed using three benchmark datasets, three predominantly SM gene families, glucosinolate biosynthesis pathway genes, and 39 AraCyc-annotated SM genes that were deliberately withheld from the model building process. Focusing on the AraCyc-only benchmark (benchmark2), although 380 individual features significantly differed between SM and GM genes, the effect sizes are small, and any individual feature does a poor job of distinguishing SM and GM genes compared with the machine learning models. In addition, machine learning models allow the global prediction of SM and GM genes in a plant genome. Based on the SM scores derived from these models, candidate SM genes can be prioritized for further experimental studies.

Although the binary SM/GM gene prediction model performed well, the FPR and FNR were substantial at 28% and 19%, respectively. Through closer examination of experimental evidence for 10 genes annotated as GM genes but predicted as SM genes, we found ~50% had evidence supporting classification as SM genes, indicating that a subset of the mis-predictions is likely due to mis-annotation. Thus, in addition to predicting likely GM/SM functions of un-annotated enzymes, our models can be used to pinpoint potentially mis-annotated GM/SM genes. Mis-predictions can be avoided by further improving the model in two areas: the classes defined, and the features used. Classifying enzyme genes as GM and SM may be an over-simplification. By building two three-class models (GM/SM/JC and GM/SM/DA), we found that SM and GM genes could be further categorized based on the metabolic network topology and, to a lesser extent, based on their dual-annotated roles in both SM and GM pathways. Future studies distinguishing genes at the pathway level can be carried out using similar multi-class modeling methods. Additional features that can distinguish SM and GM genes may also be needed to

further improve model performance. One possibility is to incorporate topological information as features. Another possibility is to examine feature combinations (e.g. combining an expression and a duplication feature linearly or non-linearly) using approaches such as deep learning.

In summary, we have conducted a global analysis of gene features that are useful to distinguish SM and GM genes. We also established well performing machine learning models that provide a global estimate of the SM gene content within a plant genome. The great majority of the predicted SM genes have not been assigned to pathways, highlighting the important next step of combining the GM/SM prediction scheme described here with approaches for pathway discovery and assignment. Considering that the most important features are related to gene duplication, evolutionary rate, and gene expression and that these types of data are readily available for an ever-expanding number of plant species, the machine learning workflow we have developed can be readily applied to any other species for predicting SM genes, or more generally, gene functions. Nonetheless, there is room for further improvement. Our prediction model serves as a baseline model for future studies incorporating additional features and algorithms that are anticipated to further improve the accuracy of predictions.

Methods

Specialized and general metabolism gene annotation and enrichment analysis

Gene sets were identified based on GO ((Botstein et al., 2000); http://www.geneontology.org/ontology/go.obo), and/or AraCyc ((Rhee et al., 2006); http://www.plantcyc.org/) annotations, but not MapMan (Thimm et al., 2004). We did not analyze MapMan annotations because all GO and AraCyc SM genes, which include a large number of well-known SM examples, were annotated as GM in MapMan, raising questions about the utility of MapMan SM/GM designations. GO annotations for *A. thaliana* were

downloaded from The Arabidopsis Information Resource (TAIR) (Berardini et al., 2015) 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 more specialized primary and secondary metabolism child GO terms were also classified as GM and SM 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. For AraCyc genes, the v.15 pathway annotations were retrieved from the Plant Metabolic Network database (http://www.plantcyc.org) (Rhee et al., 2006). Potential SM genes were those associated with "secondary metabolites biosynthesis" pathways. Potential GM genes were those found in non-secondary metabolite biosynthesis pathways. In addition, genes without experimental evidence in AraCyc (EV-EXP) were not included in the benchmark. Some genes were annotated in both SM and non-SM pathways and were defined as dual-annotated (DA) genes, not as SM or GM. Potential SM and GM genes from GO or AraCyc were required to have an enzyme commission (EC) number annotation from AraCyc or from Pfam v.30 (http://pfam.xfam.org/) (Finn et al., 2016). Five benchmark gene sets were defined. In addition, glucosinolate pathway genes were also defined to test model performance. The criteria for defining benchmarks and glucosinolate pathway genes are detailed in **SI Methods**. Terpene synthase, P450, and methyltransferase genes were identified from *A. thaliana* annotated protein sequences using the following domain matches from Pfam: terpene_synth, p450 and methyltr_2. Details of gene set enrichment analysis is available in **SI Methods**.

Expression dataset processing and co-expression and gene network analysis

Expression datasets were downloaded from TAIR. Target datasets included plant

development (Schmid et al., 2005), biotic stress (Wilson et al., 2012), abiotic stress (Kilian et al., 2007; Wilson et al., 2012), hormone treatment (Goda et al., 2008) and diurnal expression (Mockler et al., 2007). Genes that were considered significantly expressed relative to the 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, (Lloyd et al., 2015)). The median and maximum expression levels and expression variation and breadth across the developmental expression dataset were calculated as previously described (Lloyd et al., 2015). 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 (Gautier et al., 2004; Ritchie et al., 2015). 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 (Chen, 2006). 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 *K*-means, approximate kernel *K*-means, c-means, and hierarchical clustering algorithms as described in our earlier study (Uygun et al., 2016) resulting in 5,303 binary features. For *K*-means-related analyses, the within cluster 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 ((Lee and Lee, I., 2017); http://www.functionalnet.org/aranet/), number of protein interactions (Arabidopsis Interactome Mapping Consortium, 2011), domain number, and amino acid length were calculated in our earlier study (Lloyd et al., 2015). 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 studies (Moghe et al., 2013; Lehti-Shiu et al., 2015; Lloyd et al., 2015). Details on determining the timing of duplication of an *A. thaliana* gene is available in **SI Methods.** 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 β/γ whole genome duplication status, and tandem duplication status **(Dataset S2**), were obtained from (Lloyd et al., 2015). CG methylation and $\log_2$ fold change of histone marks relative to background were taken from (Lloyd et al., 2015) (detailed in **SI methods**). Three approaches were used to evaluate the degree of metabolic gene clustering (see **SI Methods**).

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 using the Python package sci-kit

learn (Pedregosa et al.). To build binary machine learning models, we used three benchmark sets (benchmark1, 2, and 3). For each benchmark set, SM and GM genes were first divided into a modeling set (90%) and a hold-out set for independent validation (10%). 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. Details for the performance measure are available in **SI Methods**. A confidence score between 0 and 1 was produced by the model and was used as the SM prediction score. For the procedure to define threshold SM score classifying a gene as SM or not, the performance measures used, and the random background model, see **SI Methods**. Dual-annotation (DA) genes are genes annotated as both GM and SM pathway genes in AraCyc. Junction (JC) genes were defined based on the pathway annotation data (pathway.dat) from the PlantCyc *A. thaliana* v.12 dataset. Two three-class models were built. The first SM/GM/DA model used SM, GM, and DA genes (benchmark4) as the three classes. The second SM/GM/JC model used SM, GM, and JC genes (benchmark5). Additional information for defining the JC gene type is available in **SI** Methods.

Acknowledgements

We thank Christina B. Azodi, Joshua J. Moore, Nicholas L. Panchy, and Sahra Uygun for helpful discussion and support. We also thank the editor and anonymous reviewers for critical comments that led to new findings, particularly those on dual function and junction genes. This work was partly supported by grants from the National Science Foundation (NSF) IOS-1546617 to R.L., E.P., and S.-H.S., NSF DEB-1655386 to S.-H.S., and the DOE Great Lakes Bioenergy Research Center (DOE Office of Science BER DE-SC0018409) to R.L. and S.-H.S.

APPENDIX

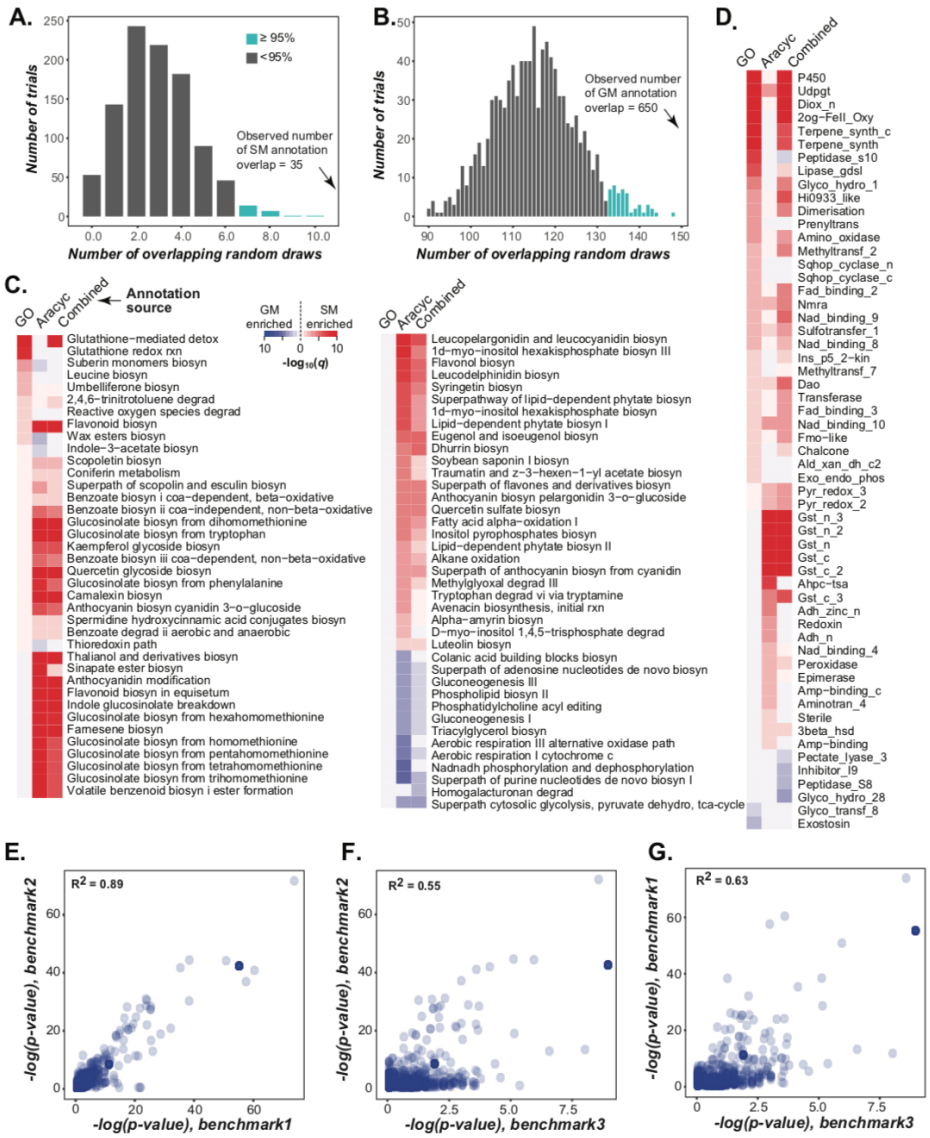


Figure S 2.1. Overlap in SM/GM gene annotations from GO and AraCyc and enrichment of SM/GM genes in pathways and protein domains.

(A) Distribution of the numbers of overlapping entries in trials where genes were randomly drawn based on the number of SM genes annotated by GO or AraCyc. The random draws were repeated 1,000 times. The blue region shows the 95th percentile of the random draw overlap distribution. **(B)** Same as **(A)** except the random sample sizes were based on the number of GO and AraCyc-annotated GM genes. **(C)** AraCyc pathway enrichment of SM genes relative to GM genes. SM and GM genes were annotated by GO, AraCyc, or both (Combined). The color in each cell represents the transformed q-value of the Fisher's exact test for a pathway or domain enriched in SM genes, with darker red and blue indicating overrepresentation in SM and GM genes, respectively. **(D)** Same as **(C),** except that Pfam domain enrichment is shown. **(E-G)** Scatterplot comparing between two benchmark sets the -log(*p*-values) from tests of feature differences between SM and GM genes. The log(*p*-values) for all features are shown, and darker

Figure S 2.1 cont'd. blue indicates regions with higher concentrations of data points. **(E)** Benchmark 1 vs. benchmark 2 **(F)** Benchmark 2 vs. benchmark 3 **(G)** Benchmark 1 vs, benchmark 3.

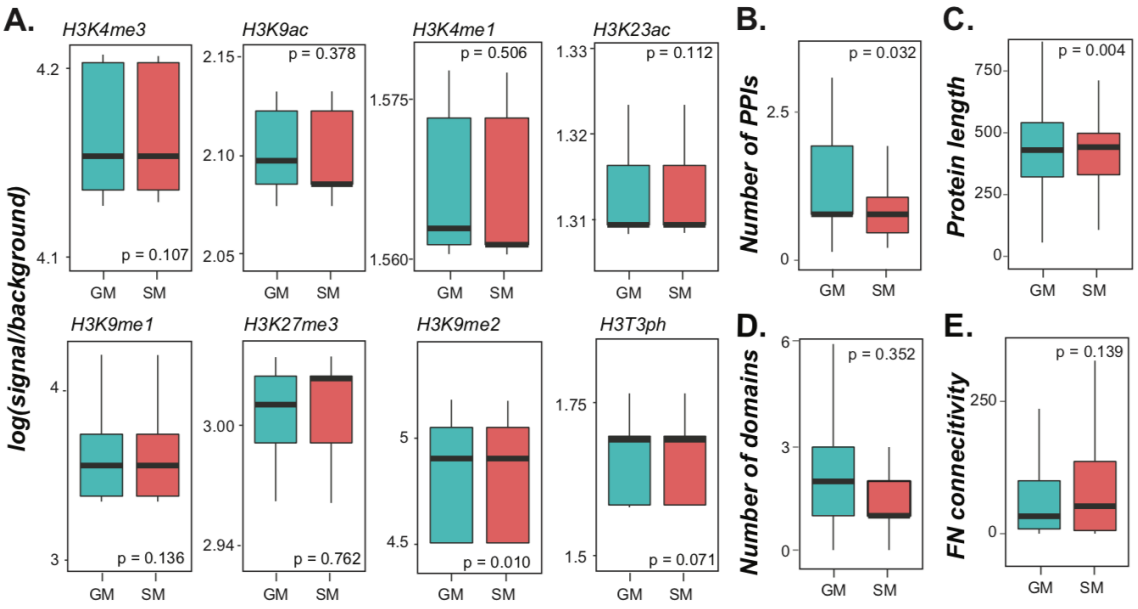


Figure S 2.2. Differences in histone marks, protein-related features, and functional network-related features between SM and GM genes.

(A) Histone modifications. For each GM (blue) and SM (red) gene, the log (base 2) ratio between immunoprecipitation and background signals was calculated for each histone mark. **(B)** The numbers of protein-protein interactions (PPI) for GM and SM proteins. **(C)** GM and SM protein lengths. **(D)** The number of protein domains in GM and SM proteins. **(E)** The number of AraNet functional network (FN) interactions for GM and SM genes. All *p*-values shown are based on Mann-Whitney tests.

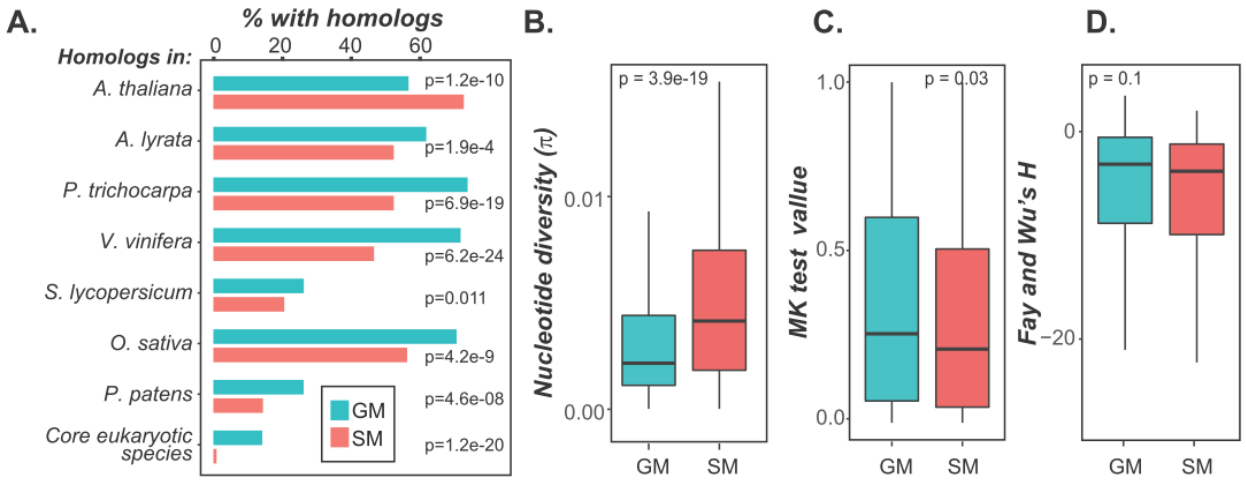


Figure S 2.3. Differences in conservation-related features between SM and GM genes.

(A) The percentage of *A. thaliana* GM (blue) and SM (red) genes with homologs in seven plant species and in core eukaryotic species. **(B)** Difference in nucleotide diversity (π) between GM and SM genes among 80 *A. thaliana* accessions. **(C)**. Difference in the MacDonald-Kreitman (MK) test statistic (see Methods), a measure of selection that compares evolutionary rates both within species and across species, between GM and SM genes. **(D)** Difference in the Fay and Wu's H statistic between GM and SM genes. H is a measure of selection where a positive value indicates a deficit of SNPs, potentially indicating a selective sweep, and a negative value indicates an excess of high-frequency derived SNPs. All *p*-values shown are based on Mann-Whitney tests.

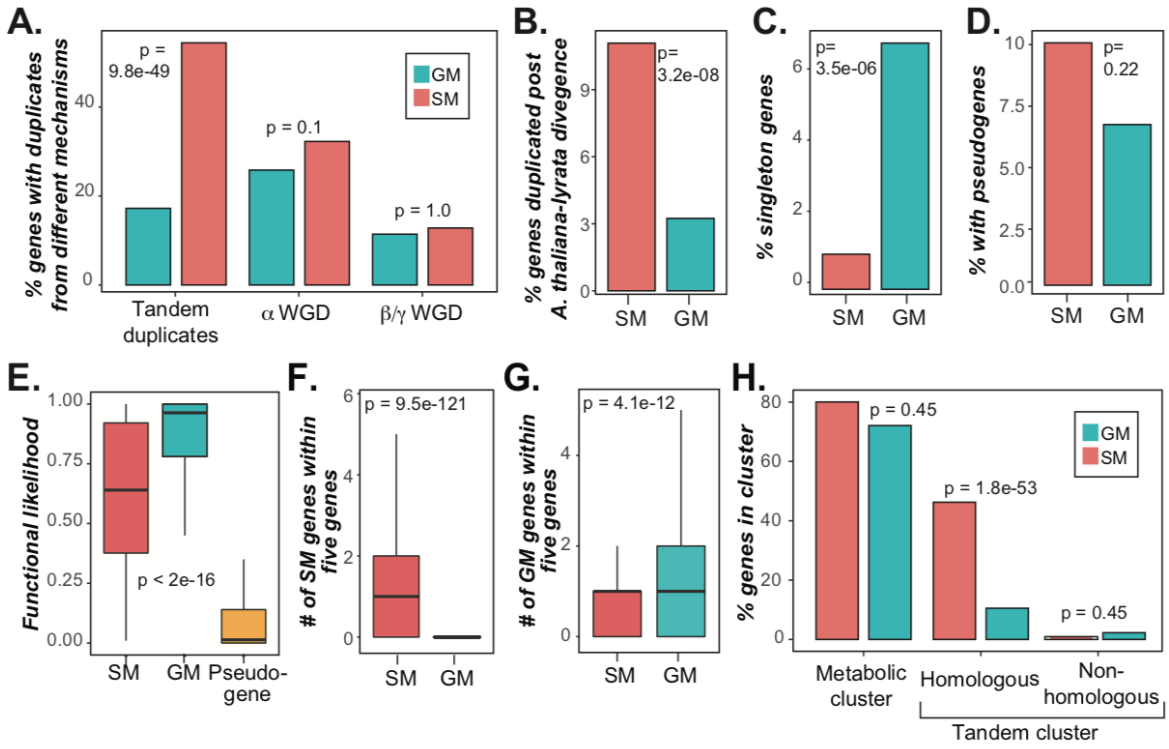


Figure S 2.4. Differences in gene duplication and genome co-localization features between SM and GM genes.

(A-D) Percentage of GM (blue) and SM (red) genes: **(A)** with ≥1 tandem duplicates, α whole genome duplicates (WGD), or β/γ WGD, **(B)** with ≥1 duplicates derived from duplication events after the *A. thaliana-A. lyrata* split, **(C)** that are singletons defined as *A. thaliana* genes with no duplicate within species but with homologous genes in *O. sativa* and *P. patens*, and **(D)** with related pseudogenes**.** All *p*-values are from Fisher's exact tests. **(E)** Functional likelihood distributions of SM and GM genes and pseudogenes. *P*-values were determined by ANOVA and post-hoc Tukey's test. **(F)** Number of SM genes that are located ≤5 genes away from another SM (red) or GM (blue) gene. The *p*-value is from the Mann-Whitney U test. **(G)** Same as **(F)** except that the number for GM genes is shown. **(H)** Percentage of genes in a metabolic cluster, a homologous tandem cluster that includes SM/GM genes, or a non-homologous tandem cluster that includes SM/GM genes. All *p*-values are from Fisher's exact tests.

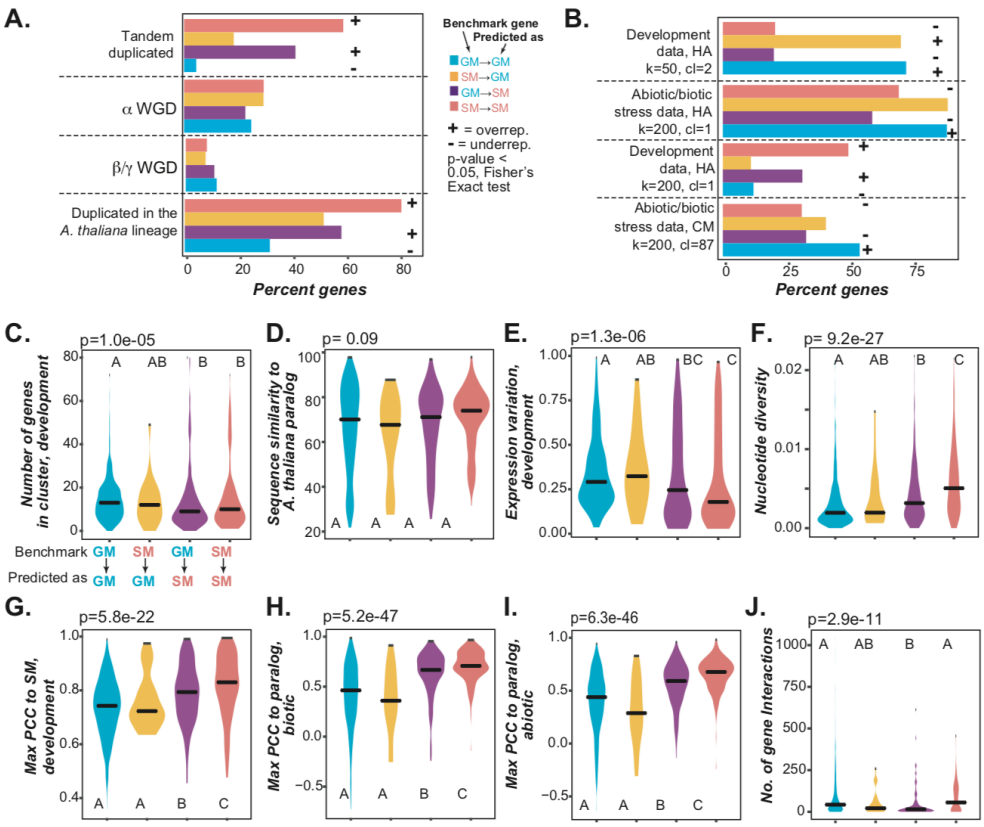


Figure S 2.5. Feature importance and properties of predictions consistent or inconsistent with annotations.

(A) Percentage of SM and GM genes (benchmark 2) duplicated via tandem or WGD mechanisms or duplicated in the *A. thaliana* lineage. Four categories are defined depending on annotation (left)→prediction (right) consistency: GM→GM (blue), SM→GM (orange), GM→SM (purple), and SM→SM (red). (+) indicates significant overrepresentation of a category, while (-) indicates significant underrepresentation of a category using the Fisher's Exact test. **(B)** Same as **(A)** except that the analysis is based on co-expression cluster-based features. HA: hierarchical clustering, average linkage. CM: c-means. k: number of clusters generated. cl: specific cluster name a gene resides in. **(C-J)** Distributions of the values of representative, predictive features for correctly and incorrectly predicted SM and GM genes. Black horizontal bar: median. Overall *p*-values are from Kruskal-Wallis tests used to evaluate differences among classes. The Dunn post hoc test was used to test differences between classes. Colors: same as **(A)**. For genes in each of the four categories, the features shown include: **(C)** number of genes found in a cluster C (developmental dataset, k=2000) where an SM or GM gene is present, **(D)** % sequence identity to the best matching *A. thaliana* paralog, **(E)** logarithm of the expression variation calculated using the development dataset, **(F)** nucleotide diversity among *A. thaliana* accessions, **(G)** maximum PCC value among all pairwise PCCs (development expression data) between each gene and an SM gene, **(H)** maximum PCC value among all pairwise PCCs (biotic expression data) between each gene and each of its paralogs, **(I)** maximum PCC value among all pairwise PCCs (abiotic expression data) between each gene and each of its paralogs, and **(J)** number of gene-gene interactions based on AraNet. (**Dataset S3**).

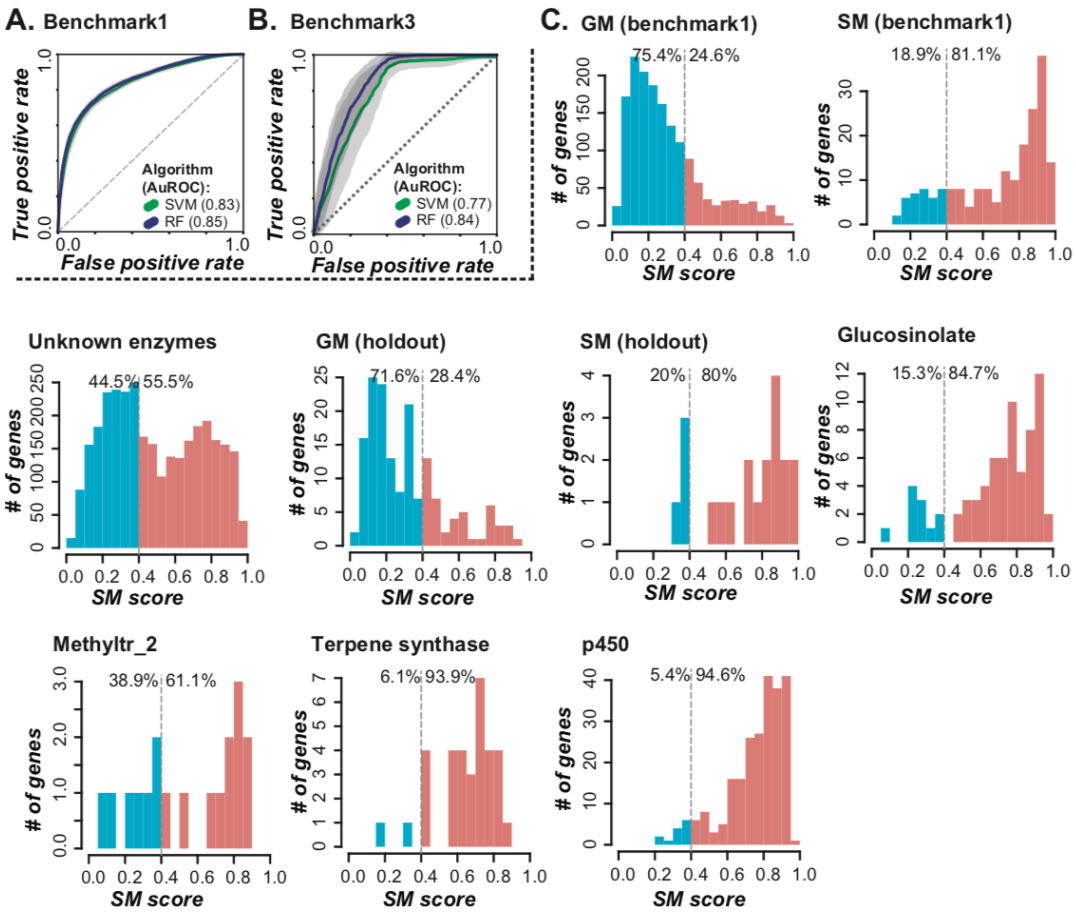


Figure S 2.6. Results for models based on benchmark1 and benchmark3.

(A, B) ROC curves and AuROCs of models using Support Vector Machine (SVM) Random Forest (RF) algorithms. **(A)** Models based on benchmark1 (GO-AraCyc union). **(B)** Models based on benchmark3 (GO-AraCyc intersection). **(C)** Distributions of SM scores based on the benchmark1 model for benchmark, unknown, and holdout genes, as well as for genes in glucosinolate biosynthesis pathways and selected families with predominantly SM genes. Unknown: genes with no SM/GM annotation in either GO or AraCyc. Holdout: GM and SM genes deliberately set aside for validation purposes that were not part of the training/testing data used for building the model. methyltr_2: methyltransferase 2. Dotted line: SM score threshold (see Methods). Red and blue shading indicate genes predicted to be SM and GM genes, respectively.

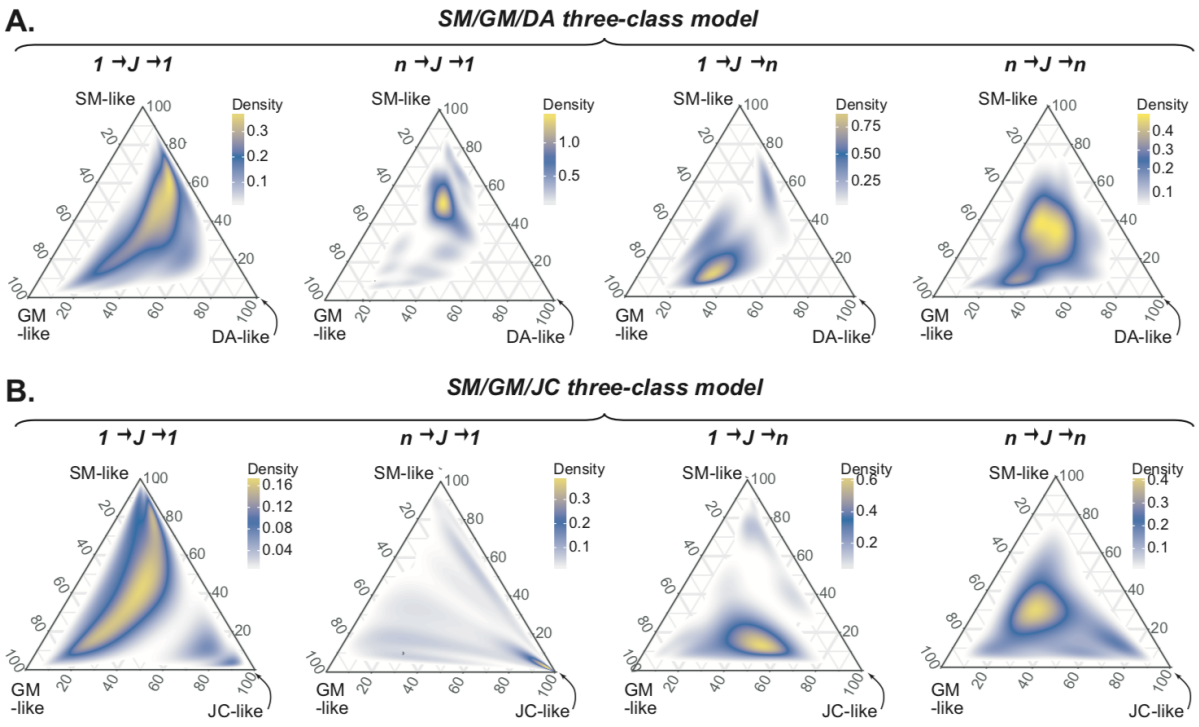


Figure S 2.7. Distributions of scores for junction subclasses in the three-class models.

(A) Ternary plots showing SM/GM/Dual-annotation (DA) model-based score distributions of genes in four junction subclasses. The four junction subclasses are as defined in **Figure 5A**. Areas with high gene density are in brighter yellow. J: Junction (JC) genes. n>1. **(B)** Ternary plots showing SM/GM/JC model-based score distributions of genes in four junction subclasses.

SI Text

Manual annotation of enzyme genes with inconsistent predictions and annotations

Our machine-learning model incorporated various characteristics such as gene duplication status and gene expression patterns not typically used in gene annotation, and accurately predicted SM genes the majority of the time. Thus, we assessed if the inconsistency in annotation and prediction is due to errors of the model or mis-annotation. We took a detailed look at SM genes predicted as GM (SM→GM) or GM genes predicted as SM (GM→SM) in the following two families: (1) terpene synthase, (2) cytochrome P450, and several biosynthetic pathways including carotenoid biosynthesis, inositol metabolism, and phenylacetaldehyde biosynthesis. For these selected families, mis-predicted genes and manual annotations were performed based experimental evidence from the literature including biochemical activity and/or genetics. For approximately 400 genes, there was a discrepancy between our model and the AraCyc annotation. A subset of these genes (32) were selected for manual annotation. For 20 SM genes predicted as GM (SM→GM) 16 were manually annotated as GM (AraCyc mis-annotation) and 4 were manually annotated as SM (model mis-prediction). For 12 GM genes predicted as SM (GM→SM) 5 were manually annotated as SM (AraCyc mis-annotation) and 7 were manually annotated as GM (model mis-prediction). For genes with inconsistent predictions and annotations we analyzed features in the model that may explain why these genes were mis-predicted or mis-annotated.

Terpene synthase genes

We investigated terpene synthases because they are a well-studied group of enzymes mostly involved in the production of specialized metabolites (Chen et al., 2011). Of the 31 terpene synthases the AraCyc-only (benchmark2) model was applied to, 29 encode enzymes

involved in the synthesis of monoterpenes, sesquiterpenes, diterpenes, and triterpenes, which are not universal to all angiosperms (Chen et al., 2011) and, thus, should be considered SM genes. Among these 29 terpene synthase genes, 28 were predicted as SM based on the machine learning model. In addition, 24 of these 29 genes were annotated as SM (SM→SM category) and four annotated as GM (SM→GM category). Based on literature information, these four AraCyc-annotated GM genes (*AT2G23230, AT3G29190, AT4G20200, AT5G48110*) that are predicted as SM are most likely mis-annotated (Chen et al., 2011). Here we do not have example where the model makes incorrect prediction.

Cytochrome P450 genes

For cytochrome P450 (referred to as P450) mis-predictions, we have examples of both incorrect model predictions and potential mis-annotations. We examined three P450 GM genes predicted as SM. Two of these genes, *AT1G01280* and *AT1G69500* (median SM score 0.75 and 0.68, respectively), are involved in sporopollenin biosynthesis (Morant et al., 2007; Dobritsa et al., 2009), a metabolite deposited on the outer layer of pollen grains that protects against desiccation, and are conserved in land plants (Liu and Fan, 2013). These two P450 enzyme genes show strong tissue-specific expression in young flowers, contributing to their mis-prediction as SM genes (**Dataset S1, S2**). When general metabolites are distributed in a tissue-specific pattern, and their biosynthetic genes have a tissue-specific expression pattern, obtaining an accurate prediction is challenging.

We further found experimental evidence suggesting that the third P450 GM gene predicted as SM may actually be an SM gene. The P450 encoded by *AT5G04660* operates at the interface of GM and SM pathways and is an enzyme catalyzing epoxidation of free fatty acids in plants (median SM score 0.91, **Dataset S1**). This enzyme was the first cytochrome P450 reported

to epoxidize unsaturated C18 fatty acids (Sauveplane et al., 2009), and is annotated as GM based on its activity on primary metabolites. However, a previous report showed that certain fatty acid epoxides have antifungal properties in plants (Kato et al., 1993). Our model predicts *AT5G04660* as an SM gene, likely due to its increased gene expression in response to environmental stresses.

Carotenoid biosynthesis genes

Carotenoids are a group of structurally diverse C_{40} hydrocarbon compounds broadly distributed in plants, which serve as important accessory pigments in the photosynthetic antenna complex and as precursors for the plant hormone abscisic acid (Bartley and Scolnik, 1995; Hirschberg, 2001; Nambara and Marion-Poll, 2005). Thus, enzymes involved in carotenoid biosynthesis should be considered GM genes. At least three genes (*AT5G17230, AtPYS1,* SM score=0.25; *AT4G14210, AtPDS3,* SM score=0.28; *AT3G04870; AtZDS,* SM score=0.24, **Dataset S2**) involved in the earlier steps of carotenoid biosynthesis are mis-annotated by AraCyc as SM genes, whereas our model correctly predicts them as GM. For example, *AtPDS3,* which is also the target of photobleaching herbicides (ChamovitzSO and Sandmannll), is a phytoene desaturase that introduces two double bonds into 15-*cis*-phytoene (Bartley et al., 1999). Mutants defective in AtPDS3 show an albino phenotype and arrested growth, likely due to impaired chlorophyll, carotenoid, and gibberellin biosynthesis (Qin et al., 2007).

Inositol metabolic genes

Inositol phosphate metabolism is conserved in eukaryotes, including plants, and is involved in cell signaling and homeostasis mechanisms, such as phosphate sensing (Tsui and York, 2010). At least one gene involved in phosphate sensing and homeostasis (*AT5G42810*; *AtIPK1,* SM score=0.19, **Dataset S2**) is predicted as a GM gene but mis-annotated by AraCyc as an SM gene. AtIPK1 possesses *in vitro* activity on inositol polyphosphate intermediates, and

atipk1 mutants show severe growth defects and aberrant phosphate homeostasis (Stevenson-Paulik et al., 2005; Kuo et al., 2014). In addition to *AtIPK1*, eight other genes associated with inositol phosphate or inositol metabolism were annotated as SM genes by AraCyc, but GM by our model. These genes all possess roles consistent with involvement in GM.

Phenylacetaldehyde biosynthesis gene

Phenylacetaldehyde is a volatile specialized metabolite involved in plant defense and is induced upon herbivory (Gutensohn et al., 2011). *AT2G20340* (*AtAAS*, median SM score 0.17, **Dataset S2**), which encodes an aromatic aldehyde synthase, converts phenylalanine to phenylacetaldehyde (Gutensohn et al., 2011; Torrens-Spence et al., 2013). Furthermore, *AtAAS* RNAi knockdown lines had increased phenylalanine and decreased phenylacetaldehyde levels (Gutensohn et al., 2011). These data suggest that AtAAS is responsible for making phenylacetaldehyde and acts in plant defense. However, *AtAAS* was mis-predicted by our model as GM, likely because *AtAAS* is expressed broadly (58 expression data points), whereas the average SM gene has a narrower expression pattern (27.5 expression data points, **Dataset S2**), has high connectivity in AraNet gene networks (91 gene-gene interactions compared with the median SM gene number of 27, **Dataset S2**), and is a member of a small gene family (only 2 family members, whereas the median SM gene family size is 62.5, **Dataset S2)**.

SI Methods

Definition of benchmark and glucosinolate pathway genes

The benchmark1 SM and GM gene dataset was established by merging potential SM and GM genes identified based on the union of GO and AraCyc annotations but removing genes with ambiguous SM and GM classifications (e.g. SM in GO but PM in AraCyc). The second benchmark gene set (benchmark2) consisted of genes based solely on AraCyc (with an EC number) annotation

excluding DA genes. The third benchmark gene set (benchmark3) consisted of the intersection between AraCyc and GO annotations excluding DA genes as well as genes with conflicting annotation from the two databases. In addition to the binary classes, we defined a benchmark dataset for classifying AraCyc SM, AraCyc GM, and AraCyc DA genes (benchmark4). For benchmark5, a new junction (JC) class was defined using AraCyc pathway annotations by identifying connected reactions between SM and GM pathways. The full list of GM, SM, DA, and JC genes is available in **Dataset S1**. Glucosinolate pathway genes were defined as those annotated by AraCyc as being involved in one of multiple glucosinolate pathways (**Dataset S1**) or annotated to the GO terms: glucosinolate metabolic process (GO:0019760), glucosinolate biosynthetic process (GO:0019761), indole glucosinolate metabolic process (GO:0042343), glucosinolate transport (GO:1901349), or regulation of glucosinolate biosynthetic process (GO:0010439). This resulted in 72 genes annotated to glucosinolate pathways and processes (**Dataset S1**).

Gene set enrichment analysis

Enrichment of SM genes relative to GM genes (for each of the three binary benchmark sets) in AraCyc pathways or GO categories was assessed with Fisher's exact tests, and test *p*-values were corrected for multiple testing (Benjamin and Hochberg, 1995). 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 Hidden Markov Models (v.30) was used to identify protein domains in proteins encoded by SM and GM genes with HMMER (Finn et al., 2015). 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).

Timing of duplication, histone mark data analysis and metabolic gene clustering

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 ($S_{X,Z}$): *Arabidopsis lyrata, Capsella rubella, Brassica rapa, Theobroma cacao, Populus trichocarpa, Medicago truncatula, Vitis vinifera, Solanum lycopersicum, Aquilegia coerulea, Oryza sativa, Amborella trichopoda, Picea abies, Selaginella moellendorffii, Physcomitrella patens,* and *Marchantia polymorpha*. Among cases where $S_{X,Z} > S_{X,Y}$, the species with gene Z most distantly related to *A. thaliana* was identified. Thus, gene X duplication likely occurred immediately prior to the divergence between *A. thaliana* and the species harboring gene Z **(Dataset S2).**

The average of the $\log_2$ 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 for each gene *X* (GM or SM) defined as the number of GM or SM genes within five or ten 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 as SM or GM genes where the neighboring SM/GM gene was separated by <10 non-SM/GM genes and <100 kb. The clusters were then determined to be homologous or non-homologous based on the presence of a significant BLAST match (E-value < 1e-05) in a given cluster (two features, **Dataset S2**). In the third approach, metabolic gene clusters were identified using the Plant Cluster Finder tool (Schlapfer et al., 2017) 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**.

Performance measure, threshold SM score, and random/background model

Performance of the RF and SVM models was determined based on both AuROC, or the area under the plot of the true positive (TP) rate against the false positive (FP) rate, calculated in R using the ROCR package, and F-measure, the harmonic mean of precision (TP/TP+FP) and recall (TP/TP+FN), where FN= false negative. The threshold SM score for calling an SM gene was defined as the SM score with the highest F-measure in the RF or 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 E.C. number annotations from AraCyc or Pfam v.30. 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.

Definition of DA and JC genes for multi-class classification

Dual-annotation (DA) genes are genes annotated as both GM and SM pathway genes in AraCyc. This classification was performed for testing if DA genes belong to a class of its own, distinct from GM and SM genes. Junction (JC) genes were defined based on the pathway annotation data (pathway.dat) from the PlantCyc *A. thaliana* v.12 dataset. Two types of JC genes were defined. For each reaction *R* in a GM pathway, if *R* was also found in an SM pathway, *R* was defined as a type 1 JC reaction, and the gene(s) encoding enzyme(s) for *R* was(were) referred to as type 1 JC genes. Type 2 JC genes were identified based on the overlap between the final products of GM pathways and the beginning substrate of SM pathways (**Figure 5A**). For a metabolic intermediate or product *M* in a GM pathway, if *M* was used as a substrate in an SM pathway, then the GM reaction(s) R_G responsible for producing *M* and the SM reaction(s) R_S using *M* as a substrate were defined as type 2 JC reactions. The genes encoding enzymes for R_G and R_S were referred to as type 2 JC genes. Two three-class models were built. The first SM/GM/DA model used SM, GM, and DA genes (benchmark4) as the three classes. The second SM/GM/JC model used SM, GM, and JC genes (benchmark5). For the three-class models the same Python package sci-kit learn and the same algorithms (RF and SVM) as the binary classification models were used; the only difference was that three class labels were used instead of two.

Supplemental Datasets

Dataset S1: Gene annotation and prediction scores

Dataset S2: Feature values

Dataset S3: Model scores and feature weights

REFERENCES

REFERENCES

Ali JG, Agrawal AA (2014) Asymmetry of plant-mediated interactions between specialist aphids and caterpillars on two milkweeds. Funct Ecol **28**: 1404–1412

Arabidopsis Interactome Mapping Consortium (2011) Evidence for Network Evolution in an Arabidopsis Interactome Map. Science **333**: 601–606

Balakirev ES, Ayala FJ (2003) Pseudogenes: Are They "Junk" or Functional DNA? Annu Rev Genet **37**: 123–151

Bartley GE, Scolnik PA (1995) Plant carotenoids: pigments for photoprotection, visual attraction, and human health. Plant Cell **7**: 1027

Bartley GE, Scolnik PA, Beyer P (1999) Two *Arabidopsis thaliana* carotene desaturases, phytoene desaturase and ζ-carotene desaturase, expressed in *Escherichia coli*, catalyze a poly- *cis* pathway to yield pro-lycopene. Eur J Biochem **259**: 396–403

Benjamin Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Statistical Society 57:

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E (2015) The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome: Tair: Making and Mining the "Gold Standard" Plant Genome. genesis **53**: 474–485

Botstein D, Cherry JM, Ashburner M, Ball CA, Blake JA, Butler H, Davis AP, Dolinski K, Dwight SS, Eppig JT (2000) Gene Ontology: tool for the unification of biology. Nat Genet **25**: 25–9

Brachi B, Meyer CG, Villoutreix R, Platt A, Morton TC, Roux F, Bergelson J (2015) Co-selected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. Proc Natl Acad Sci USA **112**: 4032–4037

Cedar H, Bergman Y (2009) Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet **10**: 295–304

Chae L, Kim T, Nilo-Poyanco R, Rhee SY (2014) Genomic Signatures of Specialized Metabolism in Plants. Science **344**: 510–513

Chan SW-L, Henderson IR, Jacobsen SE (2005) Erratum: Gardening the genome: DNA methylation in *Arabidopsis thaliana*. Nat Rev Genet **6**: 351–360

ChamovitzSO D, Sandmannll G (1993) Molecular and biochemical characterization of herbicide-resistant mutants of cyanobacteria reveals that phytoene desaturation is a rate-limiting step in carotenoid biosynthesis. *J Biol Chem* 268(23):17348–17353.

Chen F (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res **34**: D363–D368

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**: 212–229

Das M, Haberer G, Panda A, Laha SD, Ghosh TC, Schäffner AR (2016) Expression pattern similarities support the prediction of orthologs retaining common functions after gene duplication events. Plant Physiol **171** (4): 2343-2357

D'Auria JC, Gershenzon J (2005) The secondary metabolism of Arabidopsis thaliana: growing like a weed. Curr Opin Plant Biol **8**: 308–316

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci USA **110**: 2898–2903

Dobritsa AA, Shrestha J, Morant M, Pinot F, Matsuno M, Swanson R, Moller BL, Preuss D (2009) CYP704B1 Is a Long-Chain Fatty Acid -Hydroxylase Essential for Sporopollenin Synthesis in Pollen of Arabidopsis. Plant Physiol **151**: 5

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, et al (2015) The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci USA **112**: 8362–8366

Ehrlich PR, Raven PH (1964) Butterflies and Plants: A Study in Coevolution. Evolution **18**: 586

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR (2015) HMMER web server: 2015 update. Nucleic Acids Res **43**: W30–W38

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res **44**: D279–D285

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) Affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics **20**: 307–315

Giuliano G, Tavazza R, Diretto G, Beyer P, Taylor MA (2008) Metabolic engineering of carotenoid biosynthesis in plants. Trends Biotechnol **26**: 139–145

Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, Ogawa M, Yamauchi Y, Preston J, Aoki K, et al (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J **55**: 526–542

Gutensohn M, Klempien A, Kaminaga Y, Nagegowda DA, Negre-Zakharov F, Huh J-H, Luo H, Weizbauer R, Mengiste T, Tholl D, et al (2011) Role of aromatic aldehyde synthase in wounding/herbivory response and flower scent production in different Arabidopsis ecotypes: Phenylacetaldehyde biosynthesis in *A. thaliana*. Plant J **66**: 591–602

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**: 993–1003

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

Higashi Y, Saito K (2013) Network analysis for gene discovery in plant-specialized metabolism: Gene discovery in plant specialized metabolism. Plant Cell Environ **36**: 1597–1606

Hirschberg J (2001) Carotenoid biosynthesis in flowering plants. Curr Opin Plant Biol **4**: 210–218

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**: 2155–2173

Howat S, Park B, Oh IS, Jin Y-W, Lee E-K, Loake GJ (2014) Paclitaxel: biosynthesis, production and future prospects. New Biotechnol **31**: 242–245

Huot B, Yao J, Montgomery BL, He SY (2014) Growth–Defense Tradeoffs in Plants: A Balancing Act to Optimize Fitness. Mol Plant **7**: 1267–1287

Initiative AG (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. nature **408**: 796

Kato T, Yamaguchi Y, Namai T, Hirukawa T (1993) Oxygenated Fatty Acids with Anti-rice Blast Fungus Activity in Rice Plants. Biosci Biotechnol Biochem **57**: 283–287

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (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**: 347–363

Kliebenstein DJ (2008) A Role for Gene Duplication and Natural Variation of Gene Expression in the Evolution of Metabolism. PLoS ONE **3**: e1838

Kuo H-F, Chang T-Y, Chiang S-F, Wang W-D, Charng Y, Chiou T-J (2014) Arabidopsis inositol pentakisphosphate 2-kinase, AtIPK1, is required for growth and modulates phosphate homeostasis at the transcriptional level. Plant J **80**: 503–515

Lee T, Lee, I. (2017) A Network Biology Server for Arabidopsis thaliana and Other Non-Model Plant Species. Plant Gene Regul. Netw. Methods Mol. Biol. 1629:

Lehti-Shiu MD, Uygun S, Moghe GD, Panchy N, Fang L, Hufnagel DE, Jasicki HL, Feig M, Shiu S-H (2015) Molecular Evidence for Functional Divergence and Decay of a Transcription Factor Derived from Whole-Genome Duplication in *Arabidopsis thaliana*. Plant Physiol **168**: 1717–1734

Liu S-L, Baute GJ, Adams KL (2011) Organ and Cell Type–Specific Complementary Expression Patterns and Regulatory Neofunctionalization between Duplicated Genes in *Arabidopsis thaliana*. Genome Biol Evol **3**: 1419–1436

Liu L, Fan X (2013) Tapetum: regulation and role in sporopollenin biosynthesis in Arabidopsis. Plant Mol Biol **83**: 165–175

Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H (2015) Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes. Plant Cell **27**: 2133–2147

Matsuno, M, Compagnon V, Schoch GA, Schmitt M, Debayle D, Bassard J-E, Pollet B, Hehn A, Heintz D, Ullmann P, et al (2009) Evolution of a Novel Phenolic Pathway for Pollen Development. Science **325**: 1688–1692

Milo R, Last RL (2012) Achieving Diversity in the Face of Constraints: Lessons from Metabolism. Science **336**: 1663–1667

Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, McEntee C, Kay SA, Chory J (2007) The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. Cold Spring Harb. Symp. Quant. Biol. Cold Spring Harbor Laboratory Press, pp 353–363

Moghe G, Last RL (2015) Something old, something new: Conserved enzymes and the evolution of novelty in plant specialized metabolism. Plant Physiol pp.00994.2015

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S-H (2014) Consequences of Whole-Genome Triplication as Revealed by Comparative Genomic Analyses of the Wild Radish *Raphanus raphanistrum* and Three Other Brassicaceae Species. Plant Cell **26**: 1925–1937

Moghe GD, Lehti-Shiu MD, Seddon AE, Yin S, Chen Y, Juntawong P, Brandizzi F, Bailey-Serres J, Shiu S-H (2013) Characteristics and Significance of Intergenic Polyadenylated RNA Transcription in Arabidopsis. Plant Physiol **161**: 210–224

Morant M, Jorgensen K, Schaller H, Pinot F, Moller BL, Werck-Reichhart D, Bak S (2007) CYP703 Is an Ancient Cytochrome P450 in Land Plants Catalyzing in-Chain Hydroxylation of Lauric Acid to Provide Building Blocks for Sporopollenin Synthesis in Pollen. The Plant Cell **19**: 1473–1487

Nambara E, Marion-Poll A (2005) Abscisic Acid Biosynthesis and Catabolism. Annu Rev Plant Biol **56**: 165–185

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 Physiol **169**: 1821

Osbourn A (2010) Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends Genet **26**: 449–457

Panchy N, Lehti-Shiu MD, Shiu S-H (2016) Evolution of gene duplication in plants. Plant Physiol **171**(4): 2294

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res **12:** 2825–2830

Pichersky E, Lewinsohn E (2011) Convergent evolution in plant specialized metabolism. Annu Rev Plant Biol **62**: 549–566

Puthiyaveetil S, Ibrahim IM, Jeličić B, Tomašić A, Fulgosi H, Allen JF (2010) Transcriptional Control of Photosynthesis Genes: The Evolutionarily Conserved Regulatory Mechanism in Plastid Genome Function. Genome Biol Evol **2**: 888–896

Qi X, Bakht S, Legget M, Maxwell C, Melton R, Osbourn A (2004) A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants. Proc Natl Acad Sci USA **101**: 8233–8238

Qin G, Gu H, Ma L, Peng Y, Deng XW, Chen Z, Qu LJ (2007) Disruption of phytoene desaturase gene results in albino and dwarf phenotypes in Arabidopsis by impairing chlorophyll, carotenoid, and gibberellin biosynthesis. Cell Res **17**: 471–482

Renault H, Bassard J-E, Hamberger B, Werck-Reichhart D (2014) Cytochrome P450-mediated metabolic engineering: current progress and future challenges. Curr Opin Plant Biol **19**: 27–34

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 57: Springer

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers

differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res **43**: e47–e47

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**: e115

Sakamoto T (2004) An Overview of Gibberellin Metabolism Enzyme Genes and Their Related Mutants in Rice. Plant Physiol **134**: 1642–1653

Sauveplane V, Kandel S, Kastner P-E, Ehlting J, Compagnon V, Werck-Reichhart D, Pinot F (2009) Arabidopsis thaliana CYP77A4 is the first cytochrome P450 able to catalyze the epoxidation of free fatty acids in plants: CYP77A4, an epoxy fatty acid-forming enzyme. FEBS J **276**: 719–735

Schenck CA, Holland CK, Schneider MR, Men Y, Lee SG, Jez JM, Maeda HA (2017) Molecular basis of the evolution of alternative tyrosine biosynthetic routes in plants. Nat Chem Biol **13**: 1029–1035

Schlapfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T (2017) Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants. Plant Physiol pp–01942

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet **37**: 501–506

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

Steppuhn A, Baldwin IT (2007) Resistance management in a native plant: nicotine prevents herbivores from compensating for plant protease inhibitors. Ecol Lett **10**: 499–511

Stevenson-Paulik J, Bastidas RJ, Chiou S-T, Frye RA, York JD (2005) Generation of phytate-free seeds in Arabidopsis through disruption of inositol polyphosphate kinases. Proc Natl Acad Sci **102**: 12612–12617

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4**: 41

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M (2004) mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J **37**: 914–939

Torrens-Spence MP, Liu P, Ding H, Harich K, Gillaspy G, Li J (2013) Biochemical Evaluation of the Decarboxylation and Decarboxylation-Deamination Activities of Plant Aromatic Amino Acid Decarboxylases. J Biol Chem **288**: 2376–2387

Tsui MM, York JD (2010) Roles of inositol phosphates and inositol pyrophosphates in development, cell signaling and nuclear processes. Adv Enzyme Regul **50**: 324–337

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**: e1005244

Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A (2006) Transcriptional Coordination of the Metabolic Network in Arabidopsis. Plant Physiol **142**: 762–774

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**: 237

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

Zhong J-J (2002) Plant cell culture for production of paclitaxel and other taxanes. J Biosci Bioeng **94**: 591–599

CHAPTER 3 : WITHIN AND CROSS SPECIES PREDICTIONS OF PLANT SPECIALIZED METABOLISM GENES USING TRANSFER LEARNING

Abstract

Plant specialized metabolites mediate interactions between plants and abiotic/biotic environmental factors and have significant agronomical and pharmaceutical value. However, most genes involved in specialized metabolism (SM) are unknown because of the large number of specialized metabolites and the challenge in differentiating SM genes from general metabolism (GM) genes. To meet this challenge, we employed transfer learning, a type of machine learning strategy in which information from one species with substantially more experimentally derived function data is used to build a model to predict gene functions in another species. We focused on *Solanum lycopersicum* (tomato), a model crop for investigating SM pathways, and *Arabidopsis thaliana* (Arabidopsis), the best annotated plant species. Using machine learning methods to integrate five categories of gene features, predictive models distinguishing tomato SM and GM genes were built using different annotations and feature sets: (1) tomato annotations and gene features, (2) Arabidopsis annotations and gene features shared between tomato and Arabidopsis, and (3) tomato annotations filtered based on the Arabidopsis model predictions and tomato gene features. Although SM/GM genes can be predicted with reasonable accuracy based on tomato data alone (F-measure=0.74, compared with 0.5 for random guesses and 1.0 for perfect predictions), using information from Arabidopsis to filter likely misannotated genes significantly improves the predictions (F-measure= 0.92). This improvement is mainly due to significantly better GM predictions, most likely because these two species have multiple distinct SM pathways with properties that are not all shared, and thus cannot be readily transferred across species. This study demonstrates that SM/GM genes can be better predicted by leveraging functional annotation information across species. It also highlights the utility of transfer learning methods in biological applications.

Introduction

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; Schmidt et al., 2008). For example, members of the Solanaceae family, *Solanum nigrum* and *S. lyratum*, produce glycosides that have anti-tumor activity in cancer cell lines (Nohara et al., 2006). *Atropa belladonna*, also in the Solanaceae family, produces the tropane alkaloids hyoscyamine and scopolamine. This plant is named ‘beautiful woman’ because

in Roman times women used the extract to dilate their pupils (Rajput, 2014). The plant also has anticholinergic activity and are 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 flavors of fruits (Tohge et al., 2013) and defense against agricultural pests (Osbourn, 1996).

Tomato is a model crop species 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; Andersen and Markham, 2006). Despite recent progress in elucidating tomato SM pathways, our understanding of many of the steps in these pathways are incomplete due to the diversity of specialized metabolites within the tomato lineage. 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). However, inferring gene functions using sequence information alone can lead to high error rates (Rost, 2002). In addition to sequence similarity, gene co-expression networks have been used to

classify genes into specific metabolic pathways (Wisecaver et al., 2017). Similarly, 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 examples or instances (e.g., genes) with known labels (SM or GM) to learn how the properties (features) of those instances can be best used to distinguish instances with different labels in the form of a predictive model (**Figure 3.1**). 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 by improving computational approaches for distinguishing genes with SM and GM functions. To determine if the supervised learning approach to identify SM/GM genes developed for Arabidopsis can be used in another species (e.g., tomato), we first identified gene features (e.g., how a gene is expressed, what protein domains it contains) that were the most important for distinguishing SM genes from GM genes in tomato. Next, we assessed the ability to leverage annotation information from Arabidopsis to make predictions in tomato using an approach called "transfer learning" (Soria Olivas, 2010), where knowledge of SM/GM annotations from Arabidopsis was applied to models for tomato.

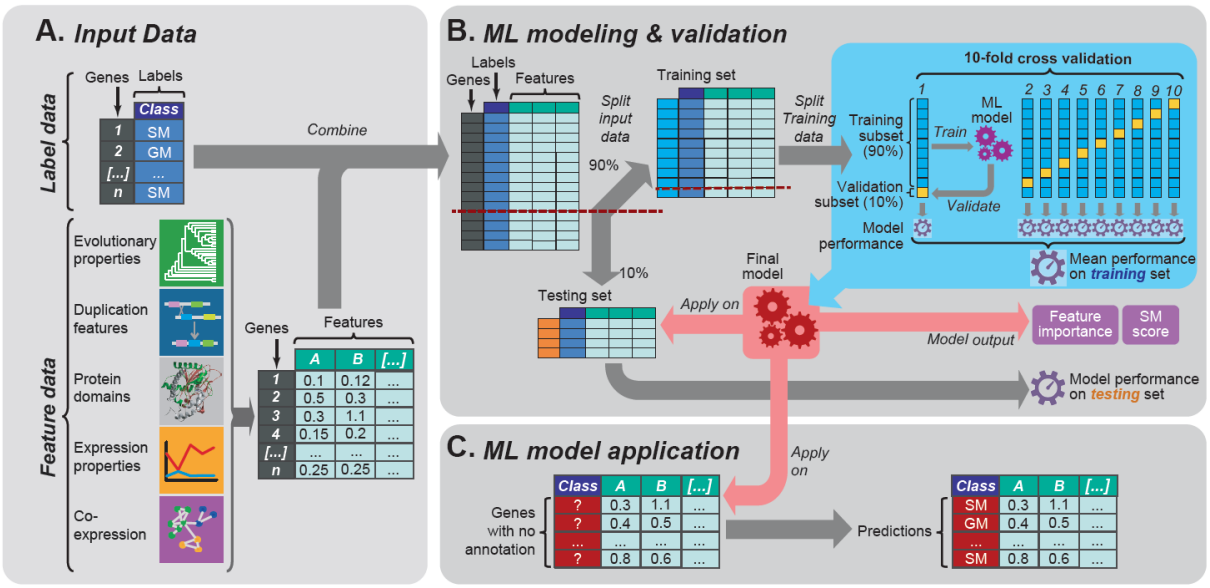


Figure 3.1. Machine learning diagram

(A) Schematic showing the input data for machine learning. The first inputs are labeled 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 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 very beginning and never used for training. (**C)** The final model is applied on unannotated enzymatic genes to make predictions.

Results and Discussion

Identifying specialized metabolism genes in tomato using machine learning approaches

To predict SM and GM genes in tomato and to understand what gene features are most important for driving the distinction between these genes, a supervised learning approach was used to build a model capable of classifying a gene as either an SM or GM gene. We focused solely on genes predicted to encode metabolic enzymes rather than regulatory genes such as transcription factors. The first step in building a machine learning model was to select the genes on which to train the model (**Figure 3.1A**). We based the training data on TomatoCyc annotated genes (referred to as "annotated genes", see **Methods**, for annotation information see **Dataset S4**), where genes in pathways under the category “secondary metabolism biosynthesis” were considered SM genes (538 genes). Genes in any other pathway not under this category were considered to be GM genes (2,313 genes). Genes found in both SM and GM pathways (158 genes) were excluded from feature analysis and model building. The remaining annotated genes were divided into two sets; 90% of genes were used as the training set, which was used for training the model. The remaining 10% of annotated genes were withheld from the model and used as an independent testing set to evaluate the performance of the model. For all annotated SM and GM genes (2,861), we collected and processed five general categories of tomato gene features (**Figure 3.1A**): evolutionary properties, gene duplication mechanism, protein domain content, expression values, and co-expression patterns (7,286 total features, see **Methods**, for feature values see **Dataset S5**). The values of these features for genes in the training set were then used to train models for predicting whether a gene is likely an SM or GM gene (see **Methods**, **Figure 3.2A**).

Multiple models were built using two machine learning algorithms, as well as different numbers of features (see **Methods**) to determine the best performing model for predicting SM and GM genes. We determined model performance by calculating precision (proportion of predictions that are correct) and recall (proportion of instances correctly predicted). The best performing model had a precision of 0.70 at a recall of 0.78. To jointly consider precision and recall, the harmonic mean of precision and recall (F-measure) was determined. The F-measure of the best performing model was 0.74 (highlighted in pink and labeled Model 1 in **Figure S3.1A**) compared with the first 9 models in **Figure S3.1A** which use the same training set but different algorithms or numbers of features (for other measure of model performance see **Dataset S6**). This model score is significantly better than a random guess (F-measure = 0.5) but is not perfect (F-measure = 1). Using this model, referred to as Model 1, 76.6% of annotated SM genes and 71.0% of annotated GM genes had predictions consistent with their TomatoCyc annotations (**Figure 3.2B)**. To provide an independent validation, the model was then applied to the test set, which resulted in a similar F-measure of 0.73 (**Figure 3.2C, Dataset S6**). Because the test set was withheld completely from the model, this indicated the model could be applied to genes with no annotation and provide reasonable predictions. In addition to model performance, 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 (**Figure 3.2B**). For SM scores and SM/GM predictions for all tomato enzymatic genes for all models, see **Dataset S7**.

Important features for predicting tomato SM genes

To better understand what gene features are important for predicting SM and GM genes, we identified features with the top 50 importance scores from Model 1 (**Figure S3.1B**, for feature importance for each model, see **Dataset S8**). The importance score for a feature is a

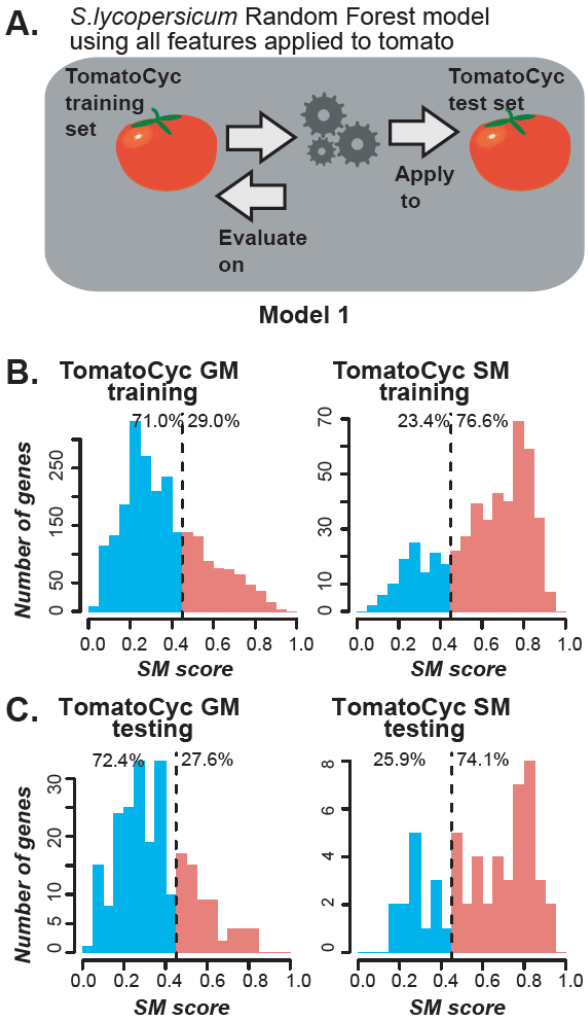


Figure 3.2. Model 1 machine-learning results.

(A) Schematic illustrating the first model, in which a tomato data set with 7,286 tomato features was used. The model was built using TomatoCyc annotations and applied to tomato genes. For B-C: SM likelihood score is represented 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. (B) Distribution of Model 1 gene likelihood scores for the TomatoCyc-annotated SM and GM genes used in training the model. (C) Distribution of Model 1 gene likelihood scores for test SM and GM genes (which were withheld from the model completely).

measurement of how much information is gained by including it in the model (see **Methods**); the higher the importance score, the better the feature is at separating SM and GM genes. Nine out of the top 10 important features are in the evolutionary property and duplication categories (**Figure S3.1B**). Gene family size, i.e., the number of paralogs of a gene, was the most important feature for the tomato SM/GM prediction Model 1. This is consistent with an earlier study in Arabidopsis (Moore et al., 2019); similar to SM genes in Arabidopsis, tomato SM genes tend to be in larger gene families (median = 8) compared with GM genes (median = 3, **Figure 3.3A,** for test statistics between all SM and GM gene features, see **Dataset S9**). Thus, SM genes tend to have a higher rate of duplication and/or duplicate retention than GM genes. SM genes are also more likely to be tandem duplicates (37%) than GM genes (13%). In addition, a lower proportion of SM genes have syntenic duplicates (17%), which are likely derived from whole genome duplication, compared with GM genes (25%, **Figure 3.3B**). This is consistent with the previous finding that genes that respond to environmental stimuli tend to be retained after duplication, particularly if they occur in tandem (Hanada et al., 2008; Kliebenstein, 2008).

It was determined previously that Arabidopsis SM genes tend to experience more relaxed selection pressure relative to GM genes (Moore et al., 2019). Consistent with this, 6 out of the top 10 most important features for tomato Model 1 are maximum or median non-synonymous/synonymous substitution rates (*dN/dS*) from comparisons of tomato genes to homologs in six other land plant species (**Figure 3.3C, Figure S3.2A-H**). The lower the *dN/dS* value, the stronger the negative selective pressure a gene has experienced. Similar to Arabidopsis, we found that SM genes in tomato tend to have a higher median or maximum *dN/dS* rate relative to between-species homologs compared with GM genes (**Figure S3.2A-H**). In addition, within-species maximum *dN/dS* values between tomato paralogs were also important

(ranked 4th, **Figure 3.3D, Dataset S8).** This is likely because GM genes are conserved among plant species and are therefore under stronger negative selection while many SM genes are derived from homologous GM genes but have experienced less stringent negative selective pressure. One possible reason for the elevated *dN/dS* is that SM genes may be under positive selection for producing specialized metabolites. Another possibility is that some SM genes are no longer under strong purifying selection because of environmental changes and are becoming pseudogenes. These explanations are supported by the observation that many more homologs of SM genes exist within species or in closely related species than in distantly related species (**Figure 3.3E**). It has also been shown that more recent duplicates tend to have higher *dN/dS* values (Lynch, 2000). Considering that SM genes tend to belong to large gene families with a high duplication rate, recent duplication events are also likely a contributor to the higher *dN/dS* values of SM genes compared with GM genes.

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. To assess how expression data may be used to distinguish SM and GM genes in tomato, we compiled 47 transcriptome studies (for details on the datasets, see **Dataset S10**) 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 (**Dataset S8**). Among the top expression features (ranked between 12-30) were maximum log fold change between developmental stages, circadian time points, mutants vs. wild type, or hormone treatments vs. controls (**Figure S3.1B**, **Dataset S8**), where SM genes tended to have higher maximum fold change values than GM genes (**Figure 3.3F-I, Dataset S9, S6**), in contrast to absolute expression values where GM genes had higher expression levels than SM

genes (**Figure S3.2I-J**). Thus, when considering gene transcription, SM gene expression tends to differ between developmental stages, varying times of day, and in response to different environments (stress or hormone treatment) to a more extreme extent than that of GM genes. Consistent with this, expression variation (median absolute deviation, see **Methods**) is also an important feature (**Dataset S8**). Examples include expression variation among fruit ripening samples (ranked 46 out of 200) and between the mutant *late termination* (Tal et al., 2017) and wild-type plants (ranked 44 out of 200, **Figure 3.3J-K**). Higher expression variation indicates that SM genes are expressed at higher levels in certain development stages and/or environments. For example, many specialized metabolites important for fruit flavor and color 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 (**Figure S3.2K**) is an additional feature that differentiates them from GM genes.

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 (**Figure 3.2B**). 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 1,652 in AraCyc. To understand why we obtained a high rate of mispredictions, we assessed what features may cause a gene to be mis-predicted. For

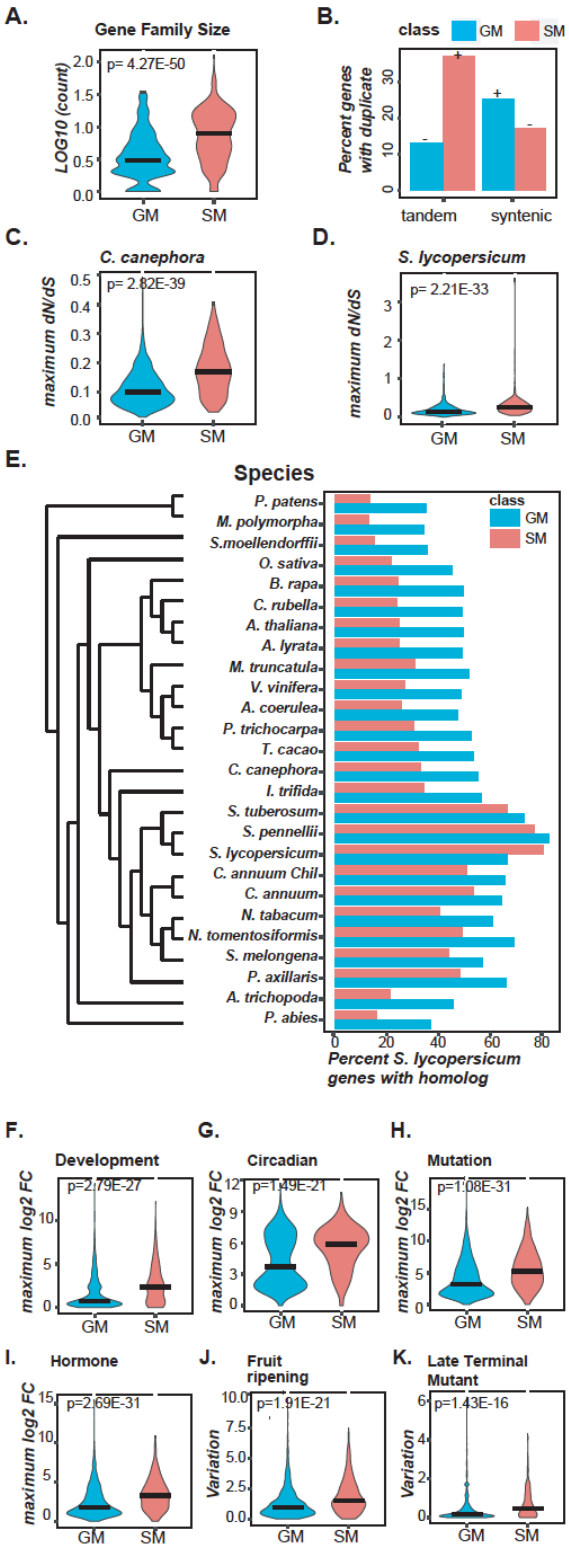


Figure 3.3. Duplication, evolutionary, and expression features important to the SM vs. GM model 1

(A-F) GM genes are denoted in blue, SM genes are denoted in red. (A) Log 10 of the number

Figure 3.3 (cont'd)
gene family members (paralogs) for each class of genes (SM and GM). (B) Percent of genes with a known duplicate (tandem or syntenic) for each class (SM and GM). (C) Maximum *dN/dS* values from comparisons of SM and GM genes to homologs in *C. canephora* and (D) *S. lycopersicum*. (E) Phylogenetic tree of 26 species showing speciation nodes, and a bar plot showing the percentage of tomato genes in each class (SM and GM) that have a homolog in an orthologous group in a given species. (F-I) Distribution of maximum fold change between all samples in a given dataset for genes in each class (GM and SM) over a (F) meristem development dataset (1 study, 18 samples), (G) circadian dataset (1 study, 86 samples), (H) mutant dataset (14 studies, 239 samples, see **Dataset S10** for list of mutants) and (I) hormone treatment dataset (5 studies, 89 comparisons, see **Dataset S10** for hormone treatments). (J-K) Distribution of variation in fold change in expression over a (J) fruit ripening dataset (1 study, 12 samples) and (K) a dataset from the *late termination mutant*, which shows delayed flowering and precocious doming of the shoot apical meristem (1 study, 12 samples, see **Dataset S10,** LTM mutant) for each gene class (GM and SM). *P*-values are from the Mann-Whitney U test between SM and GM genes.

example, SM genes in general tend to be in larger gene families than GM genes, and genes annotated as GM but predicted as SM (annotated→predicted: GM→SM) tended to belong to larger gene families (median = 5) than those having consistent GM annotations/predictions (GM→GM, median = 3, **Figure 3.4A**). Similarly, annotated SM genes predicted as GM (SM→GM) belonged to smaller families (median = 3) compared with correctly annotated/predicted SM genes (SM→SM, median = 10, **Figure 3.4A**). 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 (**Figure 3.4B**). These findings indicate that mis-predicted genes tend to possess feature values that are deviated from the norms.

Another example where GM→SM and SM→GM genes defied the general trend was in maximum *dN/dS* value, having higher and lower *dN/dS* values, respectively, compared with those genes with consistent annotations/predictions (**Figure 3.4C, D, Figure S3.3A-H**). For example, one of the GM→SM genes, *XP_010323708* (*Solyc07g054880.3.1*), has a maximum *dN/dS* of 0.25 relative to its *Coffea canephora* homolog, which is much higher than that observed for GM→GM genes (*dN/dS* of 0.10) (**Dataset S5, Dataset S9**). This high *dN/dS* value likely contributed to the prediction of this gene as SM. When looking more closely at *XP_010323708*, we found that this gene was previously reported to encode a methylketone synthase that produces specialized methyl ketones specific to the *Solanum* genus (Yu et al., 2010), and should be annotated as an SM gene. Other GM genes with high *dN/dS* values from comparisons to their tomato paralogs were also predicted as SM genes. For example, three *Glycoalkaloid metabolism* (*GAME*) genes involved in steroidal glycoalkaloids production – *GAME4*, *GAME12*, and *GAME17* – stand out as SM genes in our model while TomatoCyc incorrectly annotated them as GM genes. *GAME4* and *GAME12* both have high maximum *dN/dS* values relative to tomato

paralogs (0.30 and 0.26, respectively), a feature that many other SM genes share (SM median = 0.27, GM median= 0.15). *GAME17* belongs to a large protein family (30), another feature common to SM genes (SM median = 8, GM median = 3) and the most important feature for Model 1. 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, **Figure 3.4C, Dataset S9**). Aside from evolutionary properties and duplication features, compared with SM→SM genes, GM→SM genes also had similar maximum expression fold differences (**Figure 3.4E-H**), expression variation values (**Figure 3.4I, J**), median expression levels (**Figure S3.3I, J**), and protein domain compositions (**Figure S3.3K**).

In summary, we found that the distributions of feature values for mis-predicted GM→SM genes mirrored those for annotated SM genes. Likewise, the feature values distributions for SM→GM genes were similar to the overall distributions for annotated GM genes. These observations indicated that some SM genes in TomatoCyc looked more like GM genes and some GM genes looked more like SM genes which contributed to the discrepancies between annotation and prediction. An open question is whether these mis-predicted genes were misannotated 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.

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 our explicit goal was to learn about generalizable differences between annotated GM and SM genes. The unresolved question is why mis-predictions occur.

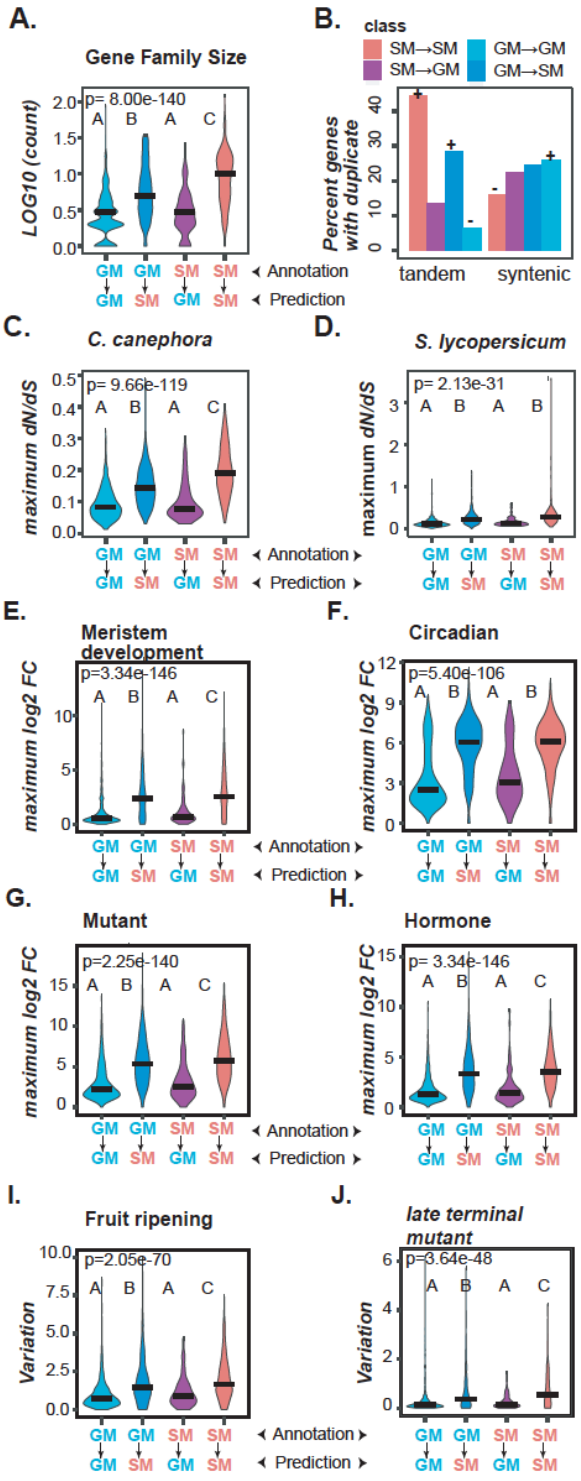


Figure 3.4. Feature distributions of genes which are predicted contrary to their annotated classification

All plots show four classes of predictions from Model 1 (GM→GM indicates a GM gene

Figure 3.4 (cont'd)
predicted as GM, GM→SM: a GM gene predicted as SM, SM→GM: an SM gene predicted as GM, SM→SM: an SM gene predicted as SM). (A) Log 10 of the number of gene family members (paralogs). (B) Percentage of genes with a known duplicate (tandem or syntenic). (C) Maximum *dN/dS* values from comparisons to homologs in *C. canephora* and (D) *S. lycopersicum*. (E-H) Distributions of maximum fold change over (E) the meristem development dataset (1 study, 18 samples), (F) the circadian dataset (1 study, 86 samples), (G) the mutation dataset (14 studies, 239 samples, see **Dataset S10** for list of mutants) and (H) the hormone dataset (5 studies, 89 comparisons, see **Dataset S10** for hormone treatments). (I-J) Distribution of variation in fold change in expression over (I) the fruit ripening dataset (1 study, 12 samples) and (J) the *late terminal mutant* dataset (1 study, 12 samples). For 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 ($P < 0.05$). For binary data (B) overrepresentation (+) and underrepresentation (-) were determined using the Fisher's Exact test, where (+) is significant enrichment of SM genes and (-) is significant enrichment of GM genes. A *p*-value less than 0.05 after Benjamin-Hochberg multiple testing correction was considered significant.

Three factors may account for mis-predictions: (1) the genes were annotated correctly, and Model 1 was incorrect, (2) Model 1 made correct predictions, but the annotations were incorrect, and (3) both annotations and predictions were correct, because these genes have roles in both GM and SM, i.e., they have dual functions (DF). 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 are shown in **Dataset S4**.

Out of 31 TomatoCyc-annotated GM genes analyzed, 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 catalyze 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 3 were likely GM genes and 5 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, is a strong 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 (**Figure S3.4A**; **Dataset S7**). 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 (**Figure S3.4A**; **Dataset S7**). 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).

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 predict 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, **Figure S3.1A**, **Dataset S6**). This was likely due to the small proportion of benchmark gene-inspired annotation corrections (30) relative to the large number of TomatoCyc-annotated genes (2,858).

We next asked whether information from Arabidopsis, which diverged from the tomato lineage 83-123 million years ago (Ku et al., 2000; Sato et al., 2012), could be used to improve gene predictions in tomato. We chose to use a machine learning approach called transfer learning (Soria Olivas, 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 4,197 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 (**Figure 3.5A**). 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 (**Figure 3.5B**). 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 (**Dataset S6**). In comparison, Model 4 built with tomato shared feature data had an F-measure = 0.75 when used for predicting tomato annotations (**Dataset S6**). Additionally, more GM/SM genes in Arabidopsis are predicted correctly by Model 3 (**Figure 3.5C**) than GM/SM genes in tomato by Model 4 (**Figure 3.5D**).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.

We next applied Arabidopsis-based Model 3 to predict tomato SM and GM genes and obtained an F-measure of 0.69 (**Figure 3.5E, Dataset S6**). This was substantially lower than the F-measure obtained when applying tomato-based Model 4 to tomato genes (0.75, **Dataset S6**), and fewer TomatoCyc annotated GM/SM genes were predicted correctly (**Figure 3.5F**). 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, **Figure 3.6A**, **Dataset S7**). However, Model 3 predicted 50% of benchmark tomato GM genes as GM (**Figure S3.4B**), which – although far from perfect – is substantially better compared with the percentage of benchmark GM genes correctly predicted by tomato Model 4 (25%, **Figure S3.4C**). 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

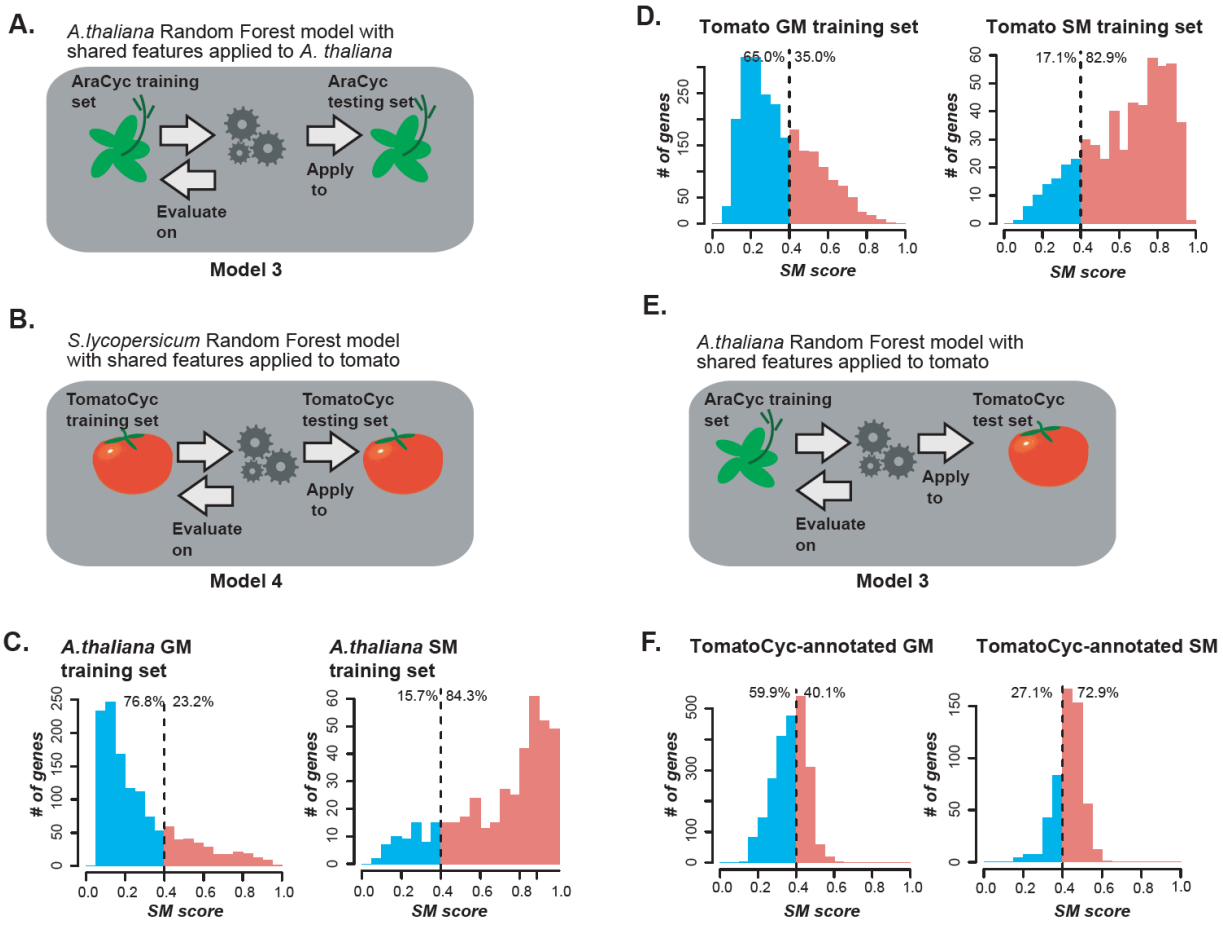


Figure 3.5. Schematics and prediction of the Arabidopsis model 3 and tomato model 4 with shared features.

(A) Schematic diagram showing the Arabidopsis model (Model 3) built using the shared feature set between Arabidopsis and tomato. Model 3 was trained using Arabidopsis annotations and was then applied to Arabidopsis genes. (B) Schematic diagram showing the tomato model built using the shared feature set between Arabidopsis and tomato. (Model 4). Model 4 was trained using tomato annotations and was then applied to tomato genes. (C) Distribution of SM likelihood scores for Arabidopsis SM and GM training set genes from Arabidopsis Model 3. (D) Distribution of SM likelihood scores from tomato Model 4. Scores for tomato training set GM and SM genes are shown. (E) Schematic diagram of Arabidopsis Model 3 built using the shared feature set between Arabidopsis and tomato. Model 3 was trained using Arabidopsis annotations and then applied to tomato genes. (F) Distribution of SM likelihood scores from Arabidopsis Model 3. Scores for annotated tomato GM and SM genes are shown. For figures C, D, and 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.

genes were more likely to be misannotated compared with annotated SM genes (**Figure S3.4B, C**), this indicates that the decline in model performance was due to mis-annotation of tomato genes.

Next, we asked how well Model 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, **Figure S3.4B**), a substantial drop from the near perfect predictions (97%) using tomato Model 4 (**Figure S3.4C)**. This indicated that Arabidopsis data may provide more useful information about true GM genes in other species than about 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. SM 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 (**Figure 3.6A, B**; **Figure S3.5A, B**; **Dataset S7**), 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 (**Figure 3.6C**). 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 dataset (**Figure 3.6C**), 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 (**Figure 3.6D-F**). For example, annotated tomato SM genes predicted as GM genes by Arabidopsis Model 3 but as SM genes by tomato Model 4 (referred to as SM→GM_3/SM_4 genes, the plot in pink, **Figure 3.6D**) tend to be in large gene families like SM→SM_3/SM_4 genes (the orange plot, **Figure 3.6D**). In contrast SM→SM_3/GM_4 genes (the brown plot, **Figure 3.6D**), 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 (**Figure 3.6E, F**; **Figure S3.5C-F**). For example, GM→GM_3/SM_4 genes are likely predicted as SM genes by tomato Model 4 (the second plot, **Figure 3.6F**) because they have high *dN/dS* values similar to those of the SM genes used to train the model (the eighth plot, **Figure 3.6F**). However, GM→SM_3/GM_4 genes (the third plot, **Figure 3.6F**) tend to have lower *dN/dS* values similar to those of the GM genes used to train the model (the first plot, **Figure 3.6F**). 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.

Improved tomato-based model by removal of potentially mis-annotated genes 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

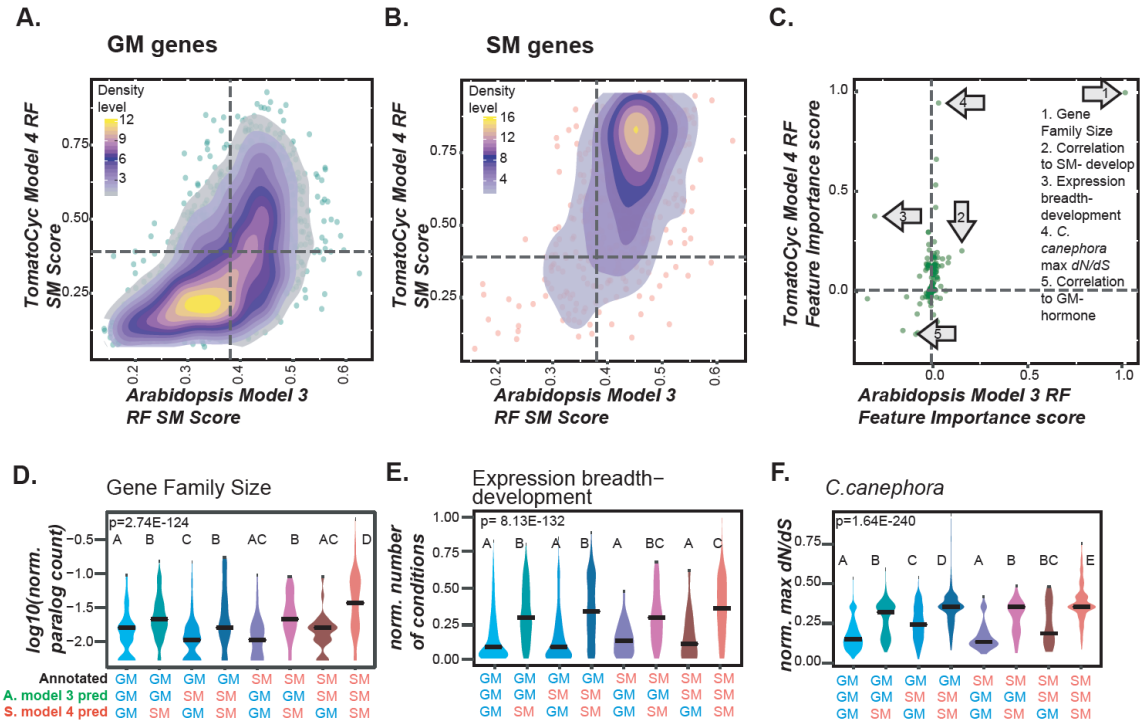


Figure 3.6. Tomato Model 4 and Arabidopsis Model 3 comparison

(A-B) Comparison of the SM score distributions from tomato Model 4 (y-axis) and Arabidopsis Model 3 (x-axis). For both models Random Forest (RF) and a shared feature set were used. Density of data points ranges from high (yellow) to medium (blue-purple), to low (white). (A) SM scores for TomatoCyc-annotated GM genes. (B) SM scores for TomatoCyc-annotated SM genes. (C) Comparison of importance score distributions for features of tomato Model 4 (y-axis) and Arabidopsis Model 3 (x-axis). Arrows point to important features: (1) Gene Family Size; (2) PCC (Pearson's correlation coefficient) between SM genes, development data; (3) Breadth of expression, development data; (4) the normalized maximum *dN/dS* between Arabidopsis or tomato genes and their *C. canephora* homologs; (5) PCC between GM genes, hormone data. (D-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, how they were predicted using Arabidopsis Model 3, and how they were predicted using 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 ($P < 0.05$). (D) Gene family size; (E) Expression breadth under development; (F) normalized maximum *dN/dS* between Arabidopsis or tomato genes and their homologs in *C. canephora*.

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. $GM \rightarrow SM_3$ and $SM \rightarrow GM_3$) 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 Model 3 and 4 (**Figure 3.7A,** 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, **Figure S3.1A**, **Dataset S6**) compared with predictions based on Model 3 (F-measure= 0.69, **Figure S3.1A, Dataset S6**) and Model 4 (F-measure = 0.75, **Figure S3.1A, Dataset S6**). 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 (**Figure 3.7B, Dataset S6**). Thus, Model 5, trained on a data set where $GM \rightarrow SM_3$ and $SM \rightarrow GM_3$ genes have been removed, is significantly improved compared with previous models. To validate Model 5 with an independent dataset, 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 (**Figure S3.6B**).

To test whether model improvement was due to the filtering out of a subset of misannotated 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; **Figure S3.1A**, **Dataset S6, see Methods**), showing no model improvement. Thus, the improvement in model performance of tomato Model 5 could not be attributed to random gene removal and was likely achieved because the filtered tomato training data did not contain mis-annotated genes that would confuse the model.

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 75% of benchmark GM genes were correctly predicted by Model 5 (**Figure S3.6A, Dataset S7**), compared with 25% for tomato Model 4 and 50% for Arabidopsis Model 3 (**Figure S3.4F, G**). In contrast, there was no improvement in benchmark SM predictions when comparing Model 4 (94% correct, **Figure S3.4F, Dataset S7**) to Model 5 (92% correct, **Figure S3.6A, Dataset S7**). 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 mostly corrected for GM genes misannotated as SM genes in TomatoCyc. Consistent with this conclusion, 83.1% 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 (**Figure S3.6C**). This indicates that introducing GM genes that were likely misannotated as SM genes into the training set led to a sub-optimal model. After their removal, the new model was able to better identify GM genes misannotated as SM. In contrast, among annotated GM genes removed from the training set because they were predicted as SM genes by Model 3, only 6.1% were predicted by Model 5 as SM genes (**Figure S3.6C**). Furthermore, GM genes identified as SM genes by Model 3, were mostly still predicted as GM genes, indicating

that the removal of these genes was relatively inconsequential, and the main issue was that a substantial number of GM genes were mis-annotated as SM genes.

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, 4, and 5; **Figure S3.6D**). 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, **Dataset S6**, **Figure S3.6E-G**). Both Models 7 and 8 were significantly improved compared with Model 1 (F-measure = 0.74), particularly when predicting GM genes (similar to Model 5). 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 (Model 5 and 7), especially for predicting GM genes.

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 5,627 unannotated enzyme genes and found that 2,865 are likely involved in SM pathways (**Figure S3.6H**). 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 **Dataset S7** for scores for each gene); thus, those unannotated enzymes that are highly likely to be an SM gene can be prioritized for further investigation.

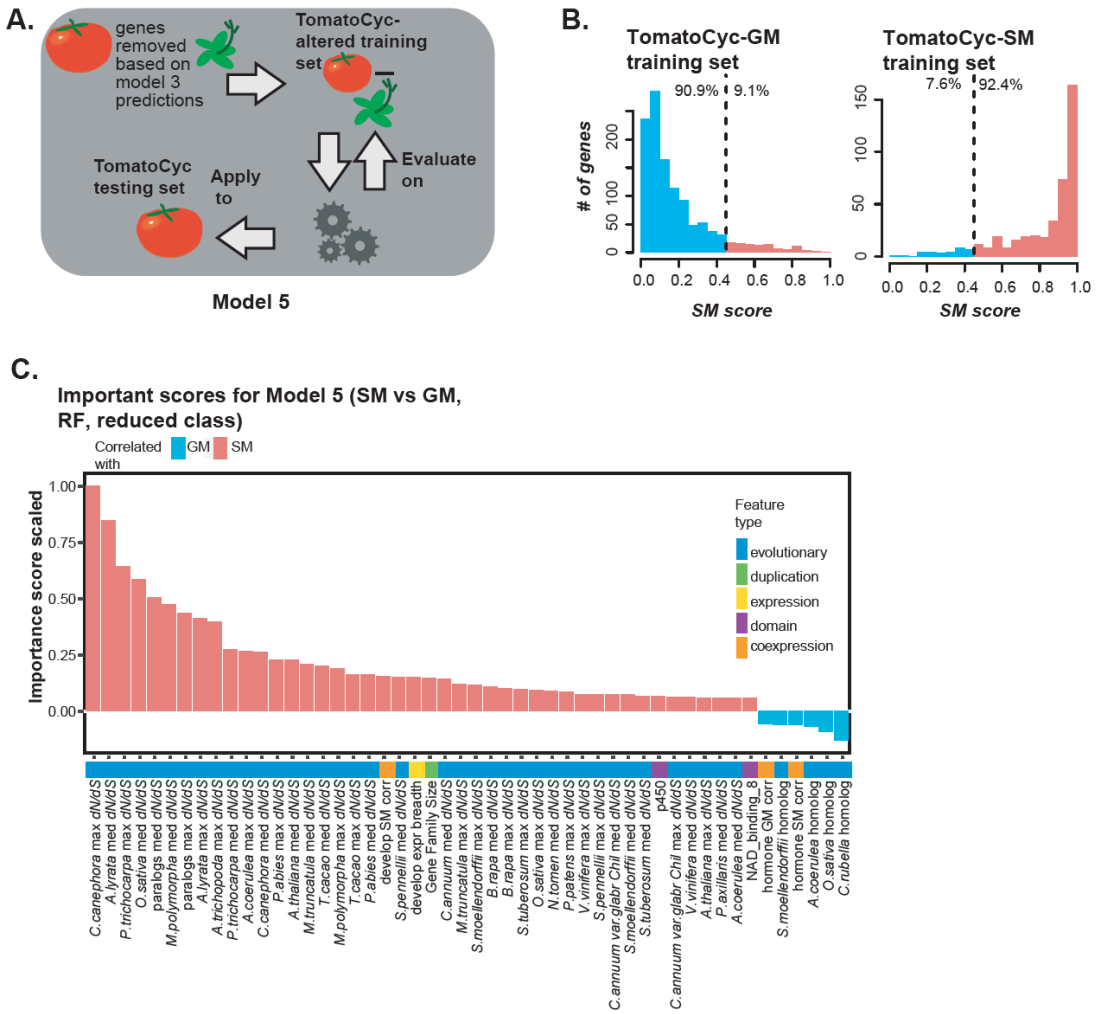


Figure 3.7. Important features for tomato Model 5

(A) Schematic diagram showing the tomato model trained on filtered annotations (Model 5) applied to tomato. The shared data set was used to build a binary model using tomato SM and GM annotations after removing those annotations that were mis-predicted by Arabidopsis Model 3. The model was then applied to tomato genes. (B) Distribution of SM likelihood scores from Model 5 using Random Forest (RF). Scores are for tomato training set GM and SM genes. 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. (C) Importance scores for Model 5. Importance scores were normalized, with 1 or -1 being the highest importance score, and 0 being the lowest. Red and blue bars indicate whether a feature is correlated with SM genes and GM genes, respectively. Normalized importance scores are shown on the x-axis and features are shown on the y-axis. Feature type is shown as a bar on the y-axis where the color indicates the feature type: evolutionary (blue), duplication (green), expression (yellow), functional domain (purple), and co-expression (orange).

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 mispredictions were likely due to gene expression-related features. While *LeACO1* exhibited a maximum log2 fold change of 7.0 based on the fruit ripening dataset (**Dataset S5**), 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 log2 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 up-regulated 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 (**Dataset S8**). 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) (**Dataset S8**). 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 *Arabidopsis lyrata*, and maximum *dN/dS* between tomato genes and homologs in *Populus trichocarpa*. In Model 5 these features were ranked 1, 2, and 3, respectively; in Model 4 they were ranked 2, 3, and 8, respectively; **Dataset S8**); in Model 7 they were ranked 1, 2, and 7, respectively; and in Model 1 they were ranked 2, 9, and 16, respectively **Dataset S8**. *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 P-450 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 sub-functionalization 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 misprediction 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.

Predicting specialized metabolism pathways

While I was able to predict SM genes globally using a binary classification of SM or GM genes, the next logical step is to identify what pathways these SM genes belong to. This would facilitate targeting candidate genes responsible for the biosynthesis of different classes of specialized metabolites for functional analyses. To assess the feasibility, we first chose seven SM pathways that had sufficient gene numbers (see **Methods**) involved in the biosynthesis of phenylpropanoid derivatives, ranging from lignin and lignan derivatives to flavonoids and volatiles (**Dataset S4**). To characterize genes from multiple pathways at the same time, we built a multi-class model that distinguishes 8 classes (i.e. seven pathways + "mix") of genes from each other (**Figure 3.8A,** see **Methods**). Genes present in multiple pathways were put together into one class called “mix” in the multi-class model. The multi-class model has an average F-measure of 0.68, much better than a random model with F-measure = 0.14). Some pathways were predicted better than others – the two most extreme examples were PWY-6199, quercetin sulfate biosynthesis, with a near perfect F-measure = 0.99 and PWY-4203, volatile benzenoid biosynthesis I (ester formation), with the lowest F-measure = 0.44 (**Figure 3.8B, Dataset S6**). Note that the class containing genes belonging to multiple pathways (referred to as "mix", **Figure 3.8B**) had a F-measure = 0.24, which is close to random. This is because for the most part the multi-pathway genes were mostly placed in other pathways rather than lumped into the “multi” class (multi, **Figure 3.8B**, **Dataset S6**). Overall, we found genes unique to a particular pathway can be predicted with supervised learning but with variable accuracy.

Why are some pathways predicted better than others and how do features of genes from these seven pathways differ? To answer this question, we compared pairwise genes within a pathway to genes between pathways (see **Methods**) and found significant differences using the

Fisher's exact test for binary data and Mann-Whitney U test for continuous data, and then examined the top 10 features with largest effect size (i.e. the lowest p-values) for each of the seven pathways. We found that 47 out of 53 features (**Figure 3.8C**) with the largest effect size for the seven pathways were expression or co-expression features, indicating this to be the major distinction among SM pathways. The largest distinguishing factor was down regulation under various hormone treatments (**Figure 3.8C**). Pathways 361, 5466, and 6673 had higher numbers of genes within each pathway down-regulated than between pathways, while for pathways 5751,7139,4203, and 6199 it was the opposite. This indicates that the genes in pathways 361, 5466, and 6673 were more responsive under hormone treatment than genes in pathways 5751, 7139, 4203, and 6199, and this influenced how these genes were predicted.

Additionally, different combinations of co-expression modules were useful in distinguishing between pathways. For example, in the well-predicted pathway 6199, modules under development, particularly fruit development, hormone, and biotic stress of *Pseudomonas syringae* were enriched. PWY-6199 produces quercetin sulfates, which are sulfonated flavonoids that are involved in pigmentation as well as auxin transport, and has anti-microbial activity (Teles et al., 2018). Thus, because of the function of the pathway, expression under specific conditions and in specific plant tissues that produce a lot of pigmentation (i.e. fruits), can help distinguish the genes in this pathway from other pathways. In contrast, PWY-361, lignin and phenylpropanoid biosynthesis is enriched in many clusters relating to development and mutation, particularly meristem development (**Figure 3.8C**). Because lignin is involved in making cell walls and contributing to rigidity of the plant, clustering in development modules for meristem may be predictive of genes being in the lignin pathway. Finally, one pathway, 4203, was not predicted well (F-measure = 0.44), and this may be because it closely mirrors pathways 6199,

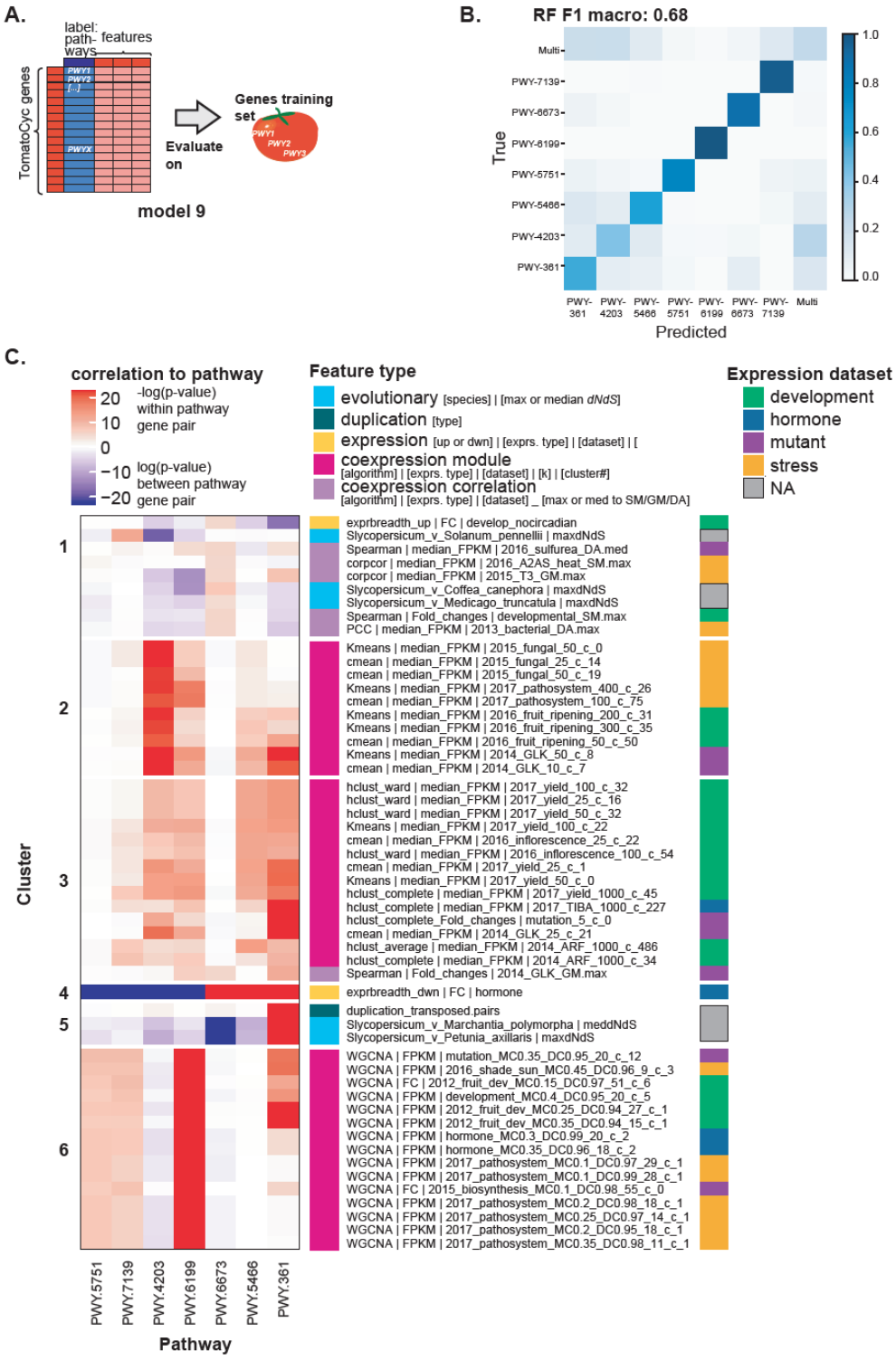


Figure 3.8. Pathway model

(A) Schematic diagram showing the pathway model 9 applied to tomato. The full tomato data set is used to build a multi-class model using tomato pathway annotations from seven SM pathways.

Figure 3.8 (cont'd)
The model is then applied to tomato enzymatic genes. (B) Percent genes correctly predicted for each pathway of the Random Forest (RF) SM pathway model 9. Actual class is on the y-axis and predicted class is on the x-axis. Scale is shown from dark blue (100%) to white (0%). Genes present in each of the seven pathways are used to build the model (labeled as PWY-XX), and genes which are present in one of these seven SM pathways but also in present in any other pathway are labeled as "Multi". (C) Feature enrichment of SM pathways comparing gene pairs within a pathway to gene pairs between pathways. Features are shown on the y-axis and pathways are shown on the x-axis. Feature enrichment is indicated by color ranging from red where it is most highly overrepresented in within-pathway gene pairs to blue where it is most overrepresented in between-pathway gene pairs. P-value is determined by the Fisher's exact test with false discovery rate correction for binary features and the Wilcox Rank sum test for continuous features. Feature type is shown as a bar on the y-axis where colors are associated with the feature type for each feature and are as flows: light blue: evolutionary features; dark blue green: duplication features; yellow: expression features; dark pink: co-expression module; light purple: co-expression correlation. The expression data set used for each pertinent feature is also shown as a bar on the y-axis where colors are associated with the following expression data sets: green: development; dark blue: hormone; purple: mutant; dark yellow: stress.
a greater distance across hormone down regulated conditions within their respective pathway

5466, and 361 in being enriched in the same co-expressed modules, but lacks unique modules that only genes in this pathway are enriched in.

Overall we found expression and co-expression features to be more important for distinguishing genes at the pathway level (**Figure 3.8C**), while evolutionary features are more important for distinguishing SM genes from GM genes (**Figure 3.7A**). Features such as evolutionary properties, duplication type, and timing of duplication are helpful in distinguishing genes to a particular pathway but are not as important as expression features which have larger of effect sizes. Using only features with the largest effect sizes, however, may not be able to discern genes as well to a specific pathway as incorporating features of both small and large effects. Thus, using machine learning methods to model pathway specific genes can better capture the variation between pathways and make better predictions of genes to the finer pathway level scale.

Conclusions

SM and GM genes are difficult to distinguish due to the vast number of specialized metabolites that are limited to specific species and the fact that SM genes are often derived from GM genes. Additionally, most gene annotations are derived from the model plant *A. thaliana*, while many specialized metabolites of interest are found in medicinal plants or crops. Thus, if data from a better annotated species such as Arabidopsis can 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. We used machine learning to establish models that could classify genes with SM and GM functions in tomato, but these models had relatively poor performance compared to models built in A. thaliana. Together with findings based on manually curated, benchmark genes, we discovered that the differences in features 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 a cross-species knowledge transfer by using the machine learning approach called transfer learning (Soria Olivas, 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 tomato-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 demonstrated that this improvement 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.

It is important to note that a limitation to the transfer learning approach we used 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 for GM genes but it did not have an appreciable impact on the prediction of SM genes, likely because SM pathways are by definition specialized – thus, 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.

We also find machine learning methods can be used to distinguish genes to a specific pathway, and that different types of co-expression, domain, and duplication features are important in distinguishing a gene to a particular pathway. A challenge for using supervised machine learning methods when distinguishing pathways is the lack of positive examples. We found that most SM pathways either did not have enough genes annotated to make predictions or had genes present in multiple pathways. This shows a need for using unsupervised machine learning methods with a heterogenous set of features to help predict pathways which currently only have one or two genes.

In future studies, transfer learning can be used to predict GM 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 will be informative and potentially can further improve cross-species predictions via transfer learning. Using transfer learning we may also be able to better

annotate less well known species. Another consideration is that we treated our research problem as a binary (SM or GM) classification problem. Over the course of evolution, 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 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.

Methods

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 specialized metabolism (SM) pathways and genes within those pathways were considered SM genes. All other pathways were considered to be general metabolism (GM) pathways. 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.

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 **Dataset S4.**

Features used for machine learning

All gene feature values can be found in **Dataset S5**. These 7,286 features are divided into several categories, each with different numbers of features: protein domains (4,232 features), expression value (280), co-expression (2,670), 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.ebi.ac.uk/Tools/hmmer/https://www.ebi.ac.uk/Tools/hmmer/) using the trusted cutoff, 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.

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/) totaling 47 studies and 926 samples (**Dataset S10**). 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).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 *S. 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. 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 down-regulated was determined using log fold change values for each data set or combination of data sets. A gene was considered up-regulated if it had a log fold change > 1 and a multiple-testing corrected p-value < 0.05 and down-regulated if it had a log fold change < -1 and a corrected p-value < 0.05.

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 dataset (with a distinct Gene Expression Omnibus GSE number), combination of datasets, and all datasets combined (see **Dataset S10**). 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 dataset, dataset combination, and all datasets combined (same as for expression correlation). This was done using both fold change and FPKM values. Using Random Forest 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.

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), *Solanum 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 (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: 1) syntenic duplicates: paralogous genes

present in within-species collinear blocks; 2) 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; 3) tandem duplicate: a gene is adjacent to its paralog; 4) proximal duplicates: a gene is separated by no more than 10 genes from its paralog. Genomic clustering features were derived from the genome annotation *Solanum 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 (**Figure S3.7**). 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. Detailed methods and parsing scripts for different features can be found in: https://github.com/ShiuLab/SM-gene_prediction_Slycopersicum.

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) were used for tests of significance. Statistical results are in **Dataset S9**.

Machine learning models

Multiple prediction models were made using the Python Sci-kit learn package (Pedregosa et al., 2011) with two algorithms, Random Forest (RF) and Support Vector Machine (SVM). The pipeline (**Figure 3.1**) used to run the models can be found here: https://github.com/ShiuLab/ML-Pipelinehttps://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 dataset was unbalanced (2,321 GM genes, 537 SM genes), 100 balanced datasets 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 dataset used to build the model, the remaining 10% used for validation, **Figure 3.1**) for each of the 100 balanced datasets. In total eight models were established using different feature and training datasets as described in **Results & 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 and recall. 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).

Shared features between Arabidopsis and tomato

Dataset S11 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: (1) protein domains: ~4,000 Pfam domains common between Arabidopsis and tomato; (2) 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.

Pathway characteristics

Within and between pathway gene pairs were determined by taking all within pathway pairs and taking a random equal number of between pathway pairs. For each pair, the overlap for

binary features and the distance of continuous features was calculated. Significance was then calculated for each within pathway and between pathway gene sets using the Fisher's exact test for binary data and the Mann Whitney U test for continuous data. The log of the p-value (or q-value for Fisher's exact test) was taken for features enriched for the between-pathway class and the negative log of the p-value or q-value for the Fisher's exact test was taken for features enriched for the within-pathway class.

APPENDIX

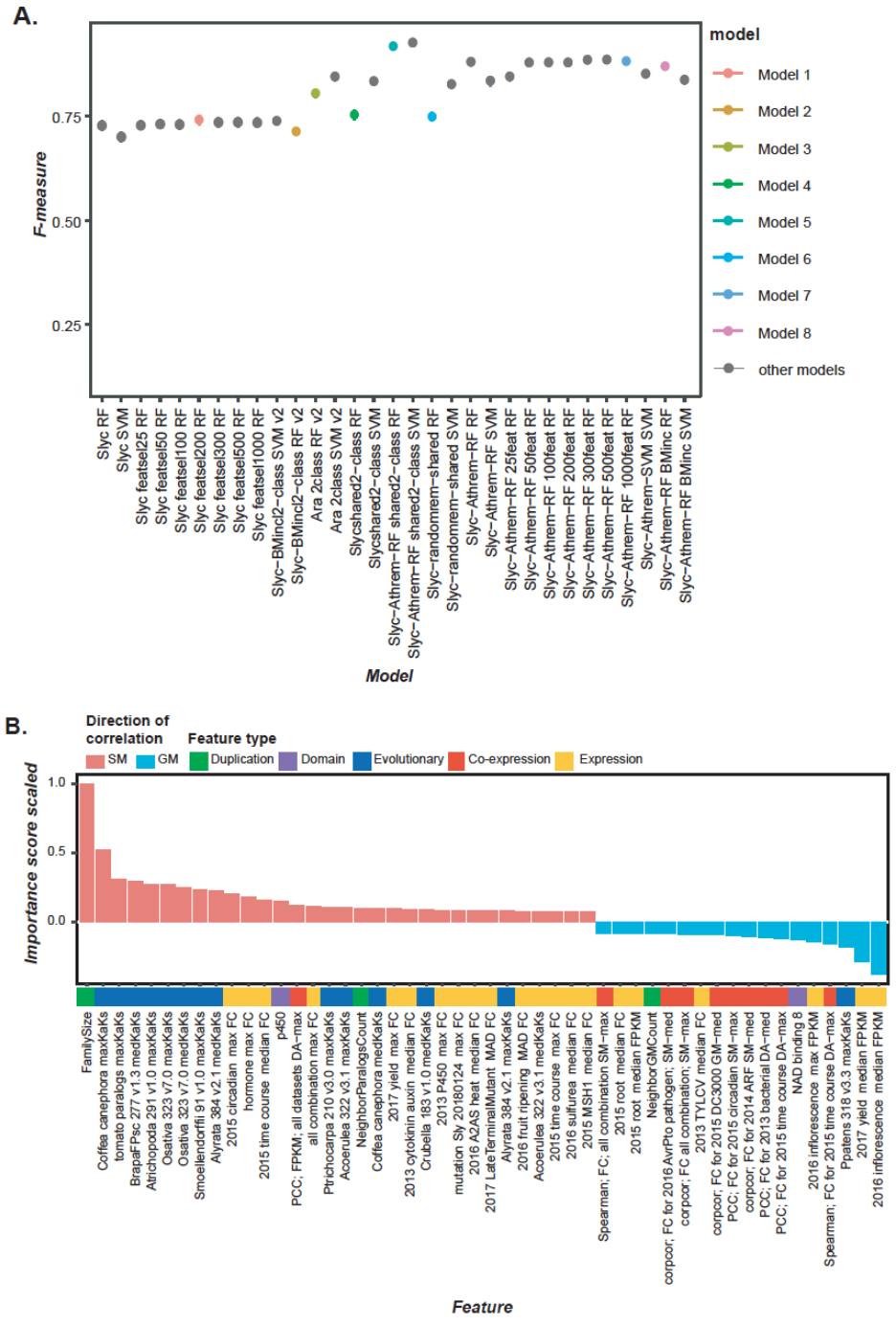


Figure S 3.1. Comparison of all model scores and Model 1 feature importance

(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 color. Gray indicates variations of Models 1-8 that are not described in the text. (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 SM genes while the blue bars represent features enriched for GM genes. Features are listed along the x-axis, with the color denoting the feature category.

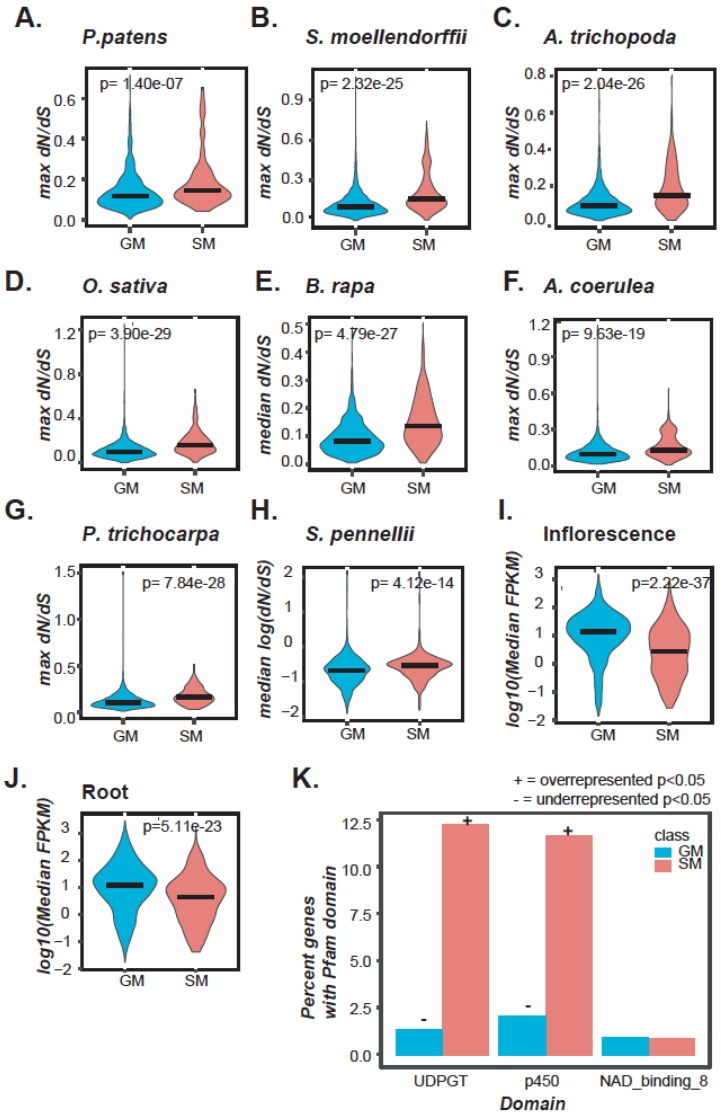


Figure S 3.2. Important features for Model 1.

(A-K) Distributions or bar plots of feature values for TomatoCyc-annotated SM and 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, 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.

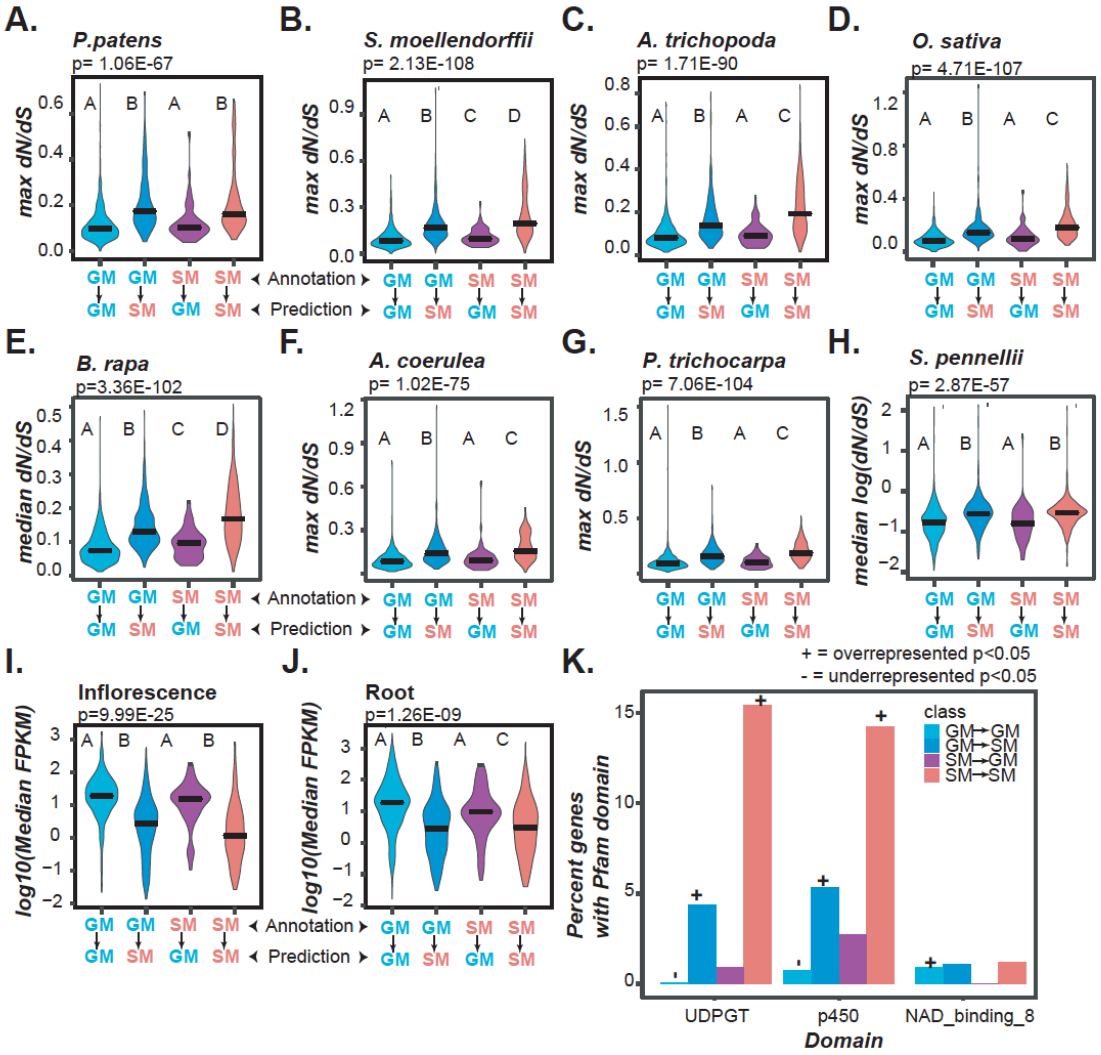


Figure S 3.3. How features shape predictions

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, J) Distributions of log10 (median FPKM) values for the Inflorescence (I) and Root (J) data sets. (K) Percentage of genes with a given Pfam domain.

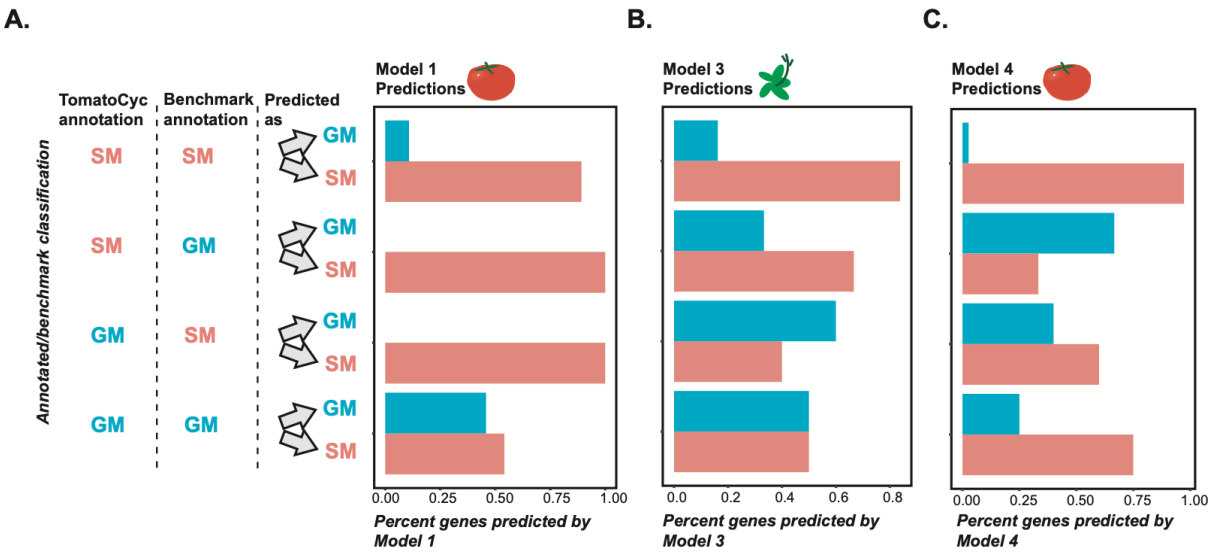


Figure S 3.4. Manually annotated gene predictions

(A) Bar plot showing the percentage of manually annotated benchmark genes predicted as SM or GM by Model 1. The original annotation from TomatoCyc is shown first, followed by the benchmark annotation and then the prediction. (B) Same as (A), except that the predictions were made using the Arabidopsis Model 3 with shared features. (C) Same as (A), except that the predictions were made using the tomato Model 4 with shared features.

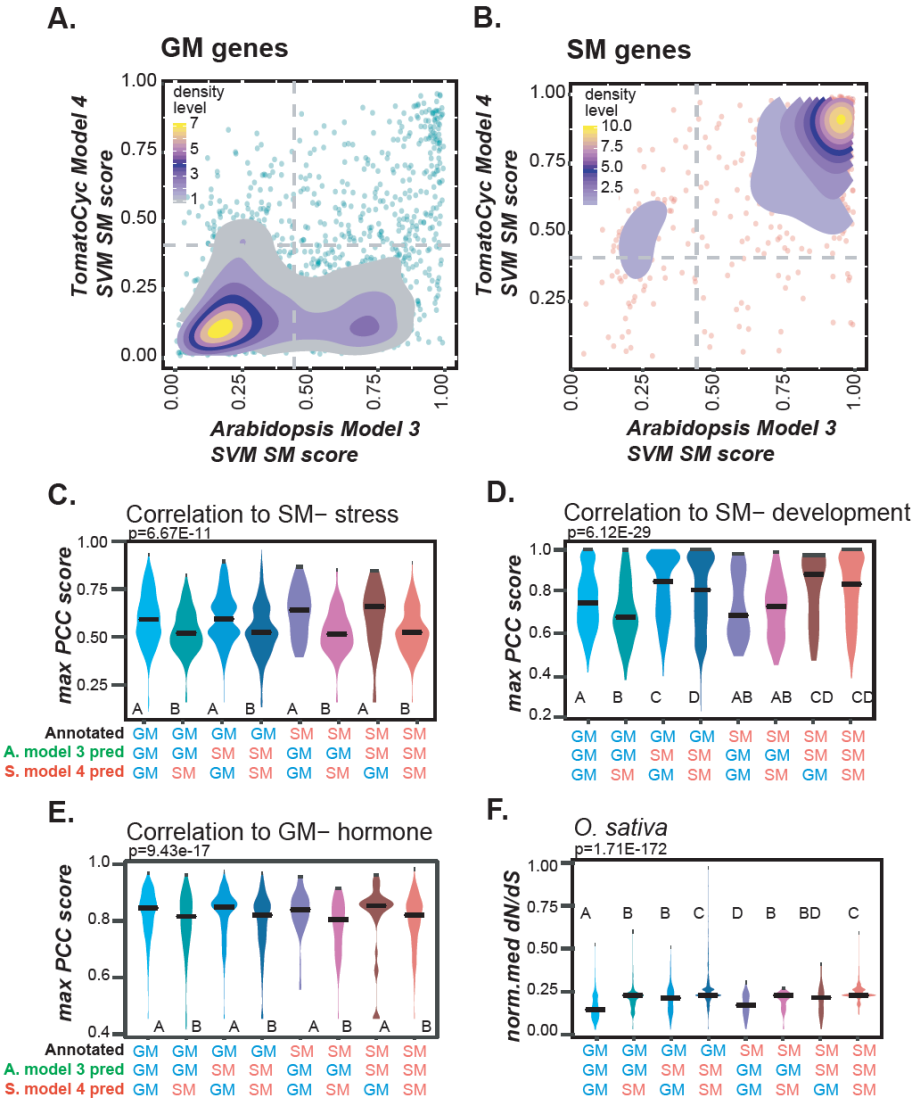


Figure S 3.5. *S. lycopersicum* and *A. thaliana* model comparison and model performance

(A-B) Comparison of the 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 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*.

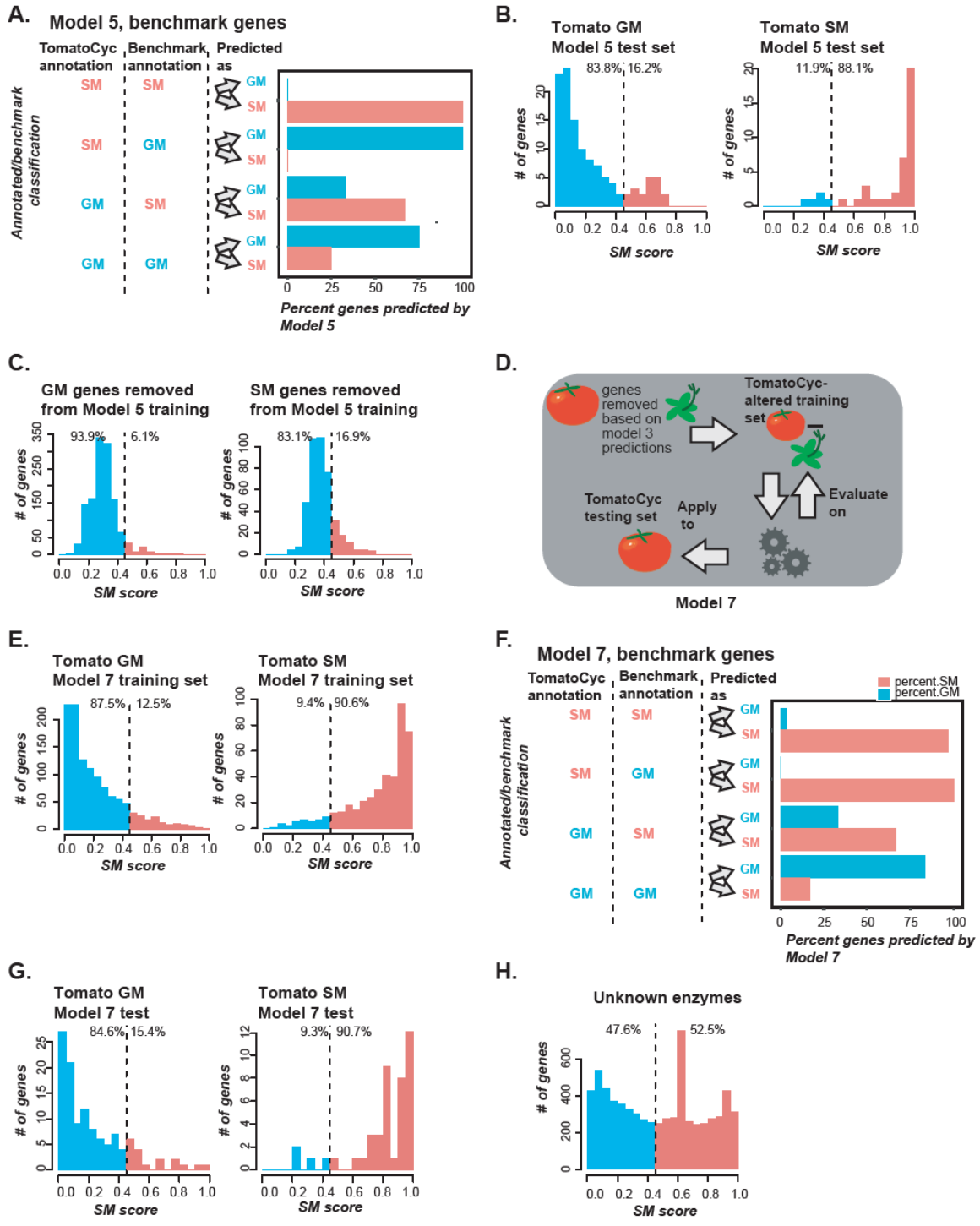


Figure S 3.6. Finalized models with *A. thaliana* mis-predictions removed, benchmark and test predictions

(A) Bar plot showing the percentage of manually annotated benchmark genes predicted as SM or GM by Model 5. The original annotation from TomatoCyc is shown first, followed by the benchmark annotation and then the prediction. Distributions of SM likelihood scores are shown in plots B, C, G, and H. (B) Model 5 test set SM and GM genes, which were held out from the

Supplemental Figure 3.6 (cont'd)
model building process completely. (C) TomatoCyc SM and GM genes with annotations opposite to Arabidopsis Model 3 predictions removed from the filtered training set. (D) Schematic diagram showing the application of tomato Model 7 to tomato. The full tomato feature dataset was used to build a binary model using TomatoCyc SM and GM annotations after removing genes mis-predicted by Arabidopsis Model 3. The model was then applied to tomato genes. (E) TomatoCyc filtered training set SM and GM genes from tomato Model 7. (F) 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. (G) Model 7 test set: SM and GM genes, which were held out completely from the tomato Model 7 building process and (H) unannotated tomato enzymes. For plots (B, D, F, and G): 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.

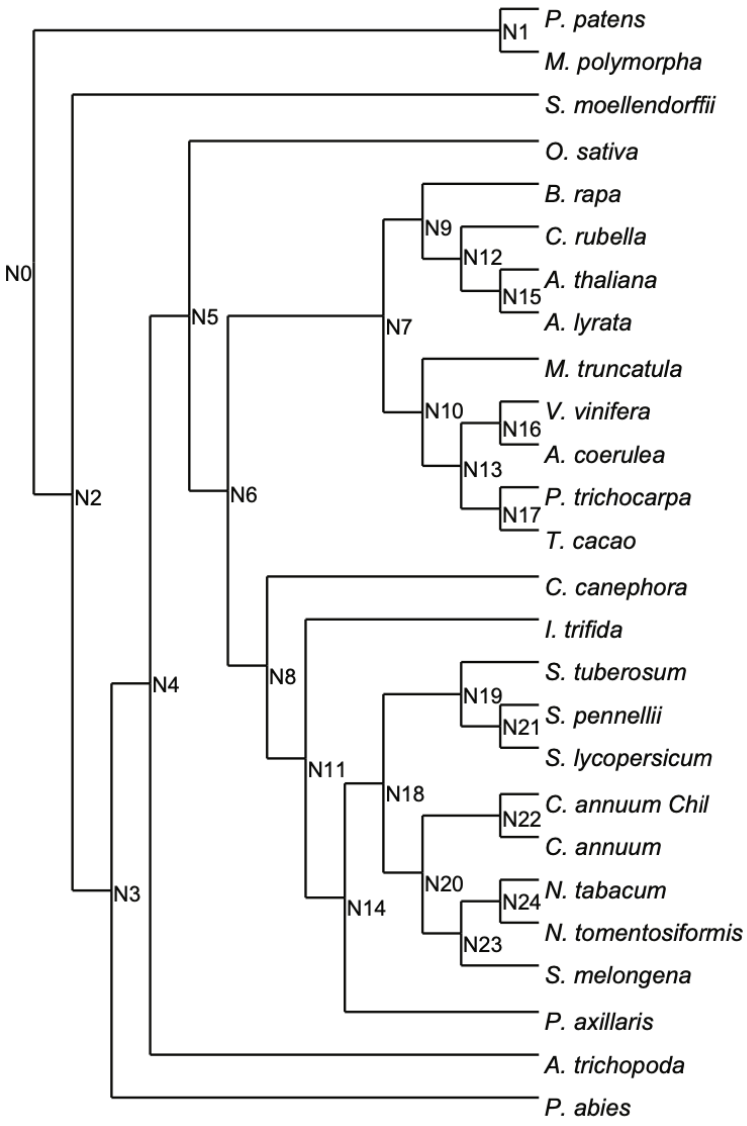


Figure S 3.7. Speciation nodes

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.

Supplemental Data

Dataset S4: Tomato gene annotation information

Annotation information based on TomatoCyc and manual annotation.

Dataset S5: Original features

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

Dataset S6: Model scores

Scores and information for all models.

Dataset S7: SM gene scores

SM prediction scores for all genes for each of the models.

Dataset S8: Feature Importance

Feature importance scores for all models discussed in the text.

Dataset S9: Feature Statistics

Statistics for original and shared features.

Dataset S10: Transcriptome studies

Information about all expression datasets used in the models.

Dataset S11: Shared features

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

Acknowledgements

We would like to thank Christina Azodi and Siobhan Cusack for helpful discussions. This work was supported by a postdoctoral fellowship from the National Science Foundation (NSF) IOS-1811055 to C.A.S; NSF grant IOS-1546617 to R.L., C.S.B, and S.-H.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.

REFERENCES

REFERENCES

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 N^{δ}-Acetylornithine, a Jasmonate-Induced *Arabidopsis* Metabolite. 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

Andersen ØM, Markham KR, eds (2006) Flavonoids: chemistry, biochemistry, and applications. CRC, Taylor & Francis, Boca Raton, FL

Blum A, Monir M, Wirsansky I, Ben-Arzi S (2005) The beneficial effects of tomatoes. Eur J Intern Med **16**: 402–404

Breiman L (2001) Random Forests. Mach Learn **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, Munos S, Blanca J, Canizares J, Diez MJ, Schneider R, Mazourek M, McClead J, et al (2013) A cytochrome P450 regulates a domestication trait in cultivated tomato. Proc Natl Acad Sci USA **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: Terpene synthase family. Plant J **66**: 212–229

Cunningham FX, Gantt E (1998) Genes and enzymes of carotenoid biosynthesis in plants. Annu Rev Plant Physiol Plant Mol Biol **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, et al (2010) Genotype to Phenotype: A Complex Problem. Science **328**: 469–469

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, et al (2015) The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci USA **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 Biol. **16**: 157

Facchini PJ, Bohlmann J, Covello PS, De Luca V, Mahadevan R, Page JE, Ro D-K, Sensen CW, Storms R, Martin VJJ (2012) Synthetic biosystems for the production of high-value plant metabolites. Trends Biotechnol **30**: 127–131

Fan P, Leong BJ, Last RL (2019) Tip of the trichome: evolution of acylsugar metabolic diversity in Solanaceae. Curr Opin Plant Biol **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 Mol Biol **22**: 589–602

Giovannucci E (2002) A Prospective Study of Tomato Products, Lycopene, and Prostate Cancer Risk. CancerSpectrum Knowl Environ **94**: 391–398

Grynkiewicz G, Gadzikowska M (2008) Tropane alkaloids as medicinally useful natural products and their synthetic derivatives as new drugs. Pharmacol Rep **60(4)**: 439-63

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**: 993–1003

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 β-Carotene and Xanthophylls in Plants. 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, et al (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. Stand Genomic Sci **5**: 424–429

Kliebenstein DJ (2008) A Role for Gene Duplication and Natural Variation of Gene Expression in the Evolution of Metabolism. PLoS ONE **3**: e1838

Ku H-M, 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. Proc Natl Acad Sci USA **97**: 9121–9126

Louppe G (2014) Understanding Random Forests: From Theory to Practice. ArXiv14077502 Stat

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). Genet Mol Res. **15(1)**

Lynch M (2000) The Evolutionary Fate and Consequences of Duplicate Genes. Science **290**: 1151–1155

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. Genet Mol Res. **16(1)**

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

Meinke D, Muralla R, Sweeney C, Dickerman A (2008) Identifying essential genes in Arabidopsis thaliana. Trends Plant Sci **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 S-H (2019) Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci USA **116**: 2344–2353

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. Anal Chem **88**: 3049–3057

Ning J, Moghe G, Leong B, Kim J, Ofner I, Wang Z, Adams C, Jones A Daniel, Zamir D, Last R L (2015) A feedback insensitive isopropylmalate synthase affects acylsugar composition in cultivated and wild tomato. Plant Physiol **69(3)**: 1821-35

Nohara T, Ikeda T, Fujiwara Y, Matsushita S, Noguchi E, Yoshimitsu H, Ono M (2006) Physiological functions of solanaceous and tomato steroidal glycosides. J Nat Med **61**: 1–

13

Osbourn AE (1996) Preformed Antimicrobial Compounds and Plant Defense against Fungal Attack. 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, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res **12**: 2825–2830

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

Pichersky E, Lewinsohn E (2011) Convergent evolution in plant specialized metabolism. Annu Rev Plant Biol **62**: 549–566

Rajput H (2014) Effects of Atropa belladonna as an Anti-Cholinergic. Nat Prod Chem Res. **1(1)**

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

Rost B (2002) Enzyme Function Less Conserved than Anticipated. J Mol Biol **318**: 595–608

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

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, et al (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**: 635–641

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: Solanum trichome chemistry. Plant J **62**: 391–403

Schlapfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T (2017) Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants. Plant Physiol **173(4)**: 2041-2059

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

Schmidt BM, Ribnicky DM, Lipsky PE, Raskin I (2007) Revisiting the ancient concept of botanical therapeutics. Nat Chem Biol **3**: 360–366

Soria Olivas E, ed (2010) Handbook of research on machine learning applications and trends:

algorithms, methods, and techniques. Information Science Reference, Hershey, PA

Tal L, Friedlander G, Gilboa NS, Unger T, Gilad S, Eshed Y (2017) Coordination of Meristem Doming and the Floral Transition by Late Termination, a Kelch Repeat Protein. Plant Cell **29**: 681–696

Tohge T, Alseekh S, Fernie AR (2013) On the regulation and function of secondary metabolism during fruit development and ripening. J Exp Bot **65**: 4599–4611

Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J, Awazuhara M, Inoue E, Takahashi H, Goodenowe DB, Kitayama M, et al (2005) Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor: Metabolomics and transcriptomics. Plant J **42**: 218–235

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

Wang P, Moore BM, Panchy NL, Meng F, Lehti-Shiu MD, Shiu S-H (2018) Factors Influencing Gene Family Size Variation Among Related Species in a Plant Family, Solanaceae. Genome Biol Evol **10**: 2596–2613

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

Wink M (1988) Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. Theor Appl Genet **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. Plant Cell **29**: 944–959

Xu J, Ranc N, Muños S, Rolland S, Bouchet J-P, Desplat N, Le Paslier M-C, Liang Y, Brunel D, Causse M (2013) Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. Theor Appl Genet **126**: 567–581

Yamaguchi S (2008) Gibberellin Metabolism and its Regulation. Annu Rev Plant Biol **59**: 225–251

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol **24**: 1586–1591

Yu H (2004) Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs. Genome Res **14**: 1107–1118

CHAPTER 4 : MODELING GENE REGULATION IN RESPONSE TO WOUNDING: TEMPORAL VARIATIONS, HORMONAL VARIATIONS, AND SPECIALIZED METABOLISM PATHWAYS INDUCED BY WOUNDING

Abstract

Plants respond to wounding stress by changing gene expression patterns and inducing jasmonic acid (JA), as well as other plant hormones, to cope with stress. This includes activating some specialized metabolism pathways, including the glucosinolate pathways, in the case of *Arabidopsis thaliana*. We model how these responses are regulated by using machine learning to incorporate putative cis-regulatory elements (pCREs), known transcription factor binding sites from literature, *in-vitro* DNA affinity purification sequencing (DAP-seq) sites, and DNase I hypersensitive sites to predict gene expression for genes clustered by their wound response. We found temporal patterns where regulatory sites and regions of open chromatin differed between clusters of genes up-regulated at early and late wounding time points as well as clusters where JA response was induced relative to clusters where JA response was not induced. Overall, we found identifying pCREs improved model predictions and discovered 4,255 pCREs related to wound response at different time points and 2,569 pCREs related to differences between JA-induced and non-JA induced wound response. In addition, pCREs found to be important at different wounding time points were mapped to the promoters of genes in a glucosinolate biosynthesis pathway to determine the regulation of this pathway under wounding stress.

Introduction

Plants cope with many environmental stresses by reprogramming their pattern of gene expression to trigger chemical and physiological responses (Bostock et al., 2014). These stress responses are essential to plant survival in their respective niches and are optimized for a plant's particular environment (Bostock et al., 2014). Gene expression reprogramming is a complex process that involves multiple levels of regulation. At the DNA sequence level, short stretches of DNA (regulatory elements) are recognized and bound by transcription factors that can activate or repress gene expression (Zou et al., 2011). Beyond the level of DNA sequence, chromatin structure can impact whether a regulatory element is accessible to a transcription factor. Chromatin structure can be modified based on signals stress response signals (Asensi-Fabado et al., 2017). Finally, reprogramming can also occur by modifying (Glisovic et al., 2008) or turning over (Hutvagner and Simard, 2008) messenger RNA.

Stress responses change over time, adding an additional level of temporal complexity to transcriptional response to stress. For example, after an initial response, genes that are turned on may act to turn on or off other genes, resulting in a cascading effects. This type of gene expression reprogramming mechanism is beneficial when different responses are needed at different times. For example, response to wounding stress in plants changes over time as the plant first needs to recognize damaging patterns, then respond by sending various hormone signals, and ultimately repair the wound (Ikeuchi et al., 2017). This means that stress responsive genes may be regulated differently depending on when they are expressed.

The production of various hormone signals allows plants to coordinate their response to different stresses because the interactions of certain hormones can regulate a specific response from the plant by changing the expression of certain genes. For example, response to wounding

stress involves several hormones, with the most ubiquitous signal being jasmonic acid (JA) (Howe and Jander, 2008). After wounding, JA levels increase and bind to JAZ repressor proteins, which allows *Myc2* transcription factors to become active (Chung et al., 2008). *Myc2* transcription factors then activate wounding responses, such as JA biosynthesis, to amplify the JA signal and activate other defensive processes (Chung et al., 2008). Additional hormones interact with JA to moderate wounding response. For example, while JA induces the expression of certain wounding response genes, ethylene simultaneously represses the expression of these genes at the damaged site in order to make sure the correct spatial response pattern is produced (Rojo et al., 1999). Ethylene also works in a synergistic fashion with JA to fine-tune wounding response by inducing the expression of proteinase inhibitor genes (O'Donnell et al., 1996) and by activating ERF1, another transcription factor that triggers defense responses (Lorenzo et al., 2003). Abscisic acid (ABA), which responds to many abiotic stresses, is also induced by wounding (León et al., 2001). While ethylene, ABA, and JA rapidly respond to wounding, other hormones such as auxin and cytokinin, start to accumulate around 12 hours after wounding and are involved in signaling for the expression of genes that ultimately work to repair the wound (Ikeuchi et al., 2017). While a great deal is known about hormone signaling in response to wounding, it is unclear what other regulatory mechanisms are involved in response to wounding and how these mechanisms interact with hormone signals. In particular, regulatory mechanisms for wounding responses not directly regulated by JA are less well understood.

Wounding can also induce the production of specialized metabolites that can deter further stress. For example, after wounding stress, *Arabidopsis thaliana* activates glucosinolate pathways. These glucosinolates and the bioproducts generated from their degradation affect the plant's interactions with biotic stresses, such as microbes and herbivores (Yan and Chen, 2007).

Additionally, mutants with decreased glucosinolate levels show greater susceptibility to the necrotrophic fungus *Fusarium oxysporum* (Tierens et al., 2001). Glucosinolate production is shown to be regulated by JA, salicylic acid (SA), and ethylene (ACC). These hormones work together to modulate glucosinolate levels in response to stress, by activating *Myb* and *Dof* transcription factors (Yan and Chen, 2007). Additionally, glucosinolates can be divided into different types, such as indole or aliphatic glucosinolates, and these types may be induced by various stresses and regulated in different ways (Yan and Chen, 2007). While specific transcription factors have been shown to turn on glucosinolate biosynthesis (Frerigmann and Gigolashvili, 2014), the regulatory elements or chromatin structure of how and when these transcription factors bind has not been resolved.

Here we assessed the extent of divergence in gene expression among various time points following wounding by correlating wounding data with other types of stress or hormone treatment. By using a time course data set, where transcriptional response is recorded over a 24-hour period (Kilian et al., 2007), we captured differences in differential gene expression and the regulatory elements required to regulate this transcriptional response. In addition, by clustering wound-responsive genes into groups based on whether or not they also respond to JA, we were able to single out differences between JA and non-JA regulatory mechanisms in regard to wounding. Finally, by using a time course study, we were able to identify important regulatory elements for the specialized metabolism pathway glucosinolate biosynthesis from tryptophan, which is induced by wounding. The goals of this study were to uncover the cis-regulatory code involved in regulating temporal responses to wounding stress, to see how wounding stress independent of the wound-induced hormone JA is regulated, and finally to understand how certain specialized metabolism pathways are regulated.

Results and Discussion

Transcriptional response to wounding varies functionally across time points

To understand how transcriptional response to wounding varies across time points, we used expression data downloaded from TAIR, in which a range of abiotic stress treatments (seven in total, including wounding) were applied to 18 day old *A. thaliana* seedlings (Kilian et al., 2007). Samples were harvested at multiple time points after treatments ranging from 15 minutes to 24 hours after treatment. Control samples were performed in parallel to exclude circadian effects (see **Methods**, Kilian et al., 2007). We identified genes that were up- or down-regulated at time points ranging from 15 minutes to 24 hours after wounding (diagonal values; **Figure 4.1A**) and how frequently the same genes were differentially expressed in these different time points (lower triangle; **Figure 4.1A)**. We found a cascading effect, where the majority of genes up-regulated at 15 and 30 minutes after wounding are still up-regulated at one hour (63% and 70% respectively), but by three hours <25% of those genes were still up-regulated (**Figure 4.1A, Dataset S13**). Consequently, the genes up- or down-regulated at later time points tended to be different than those differentially expressed earlier, with the genes responsive at 12 and 24 hours after wounding having the least amount of overlap with genes from previous wounding time points (**Figure 4.1A, Dataset S13**). Thus, different time points after wounding have overlapping but distinct sets of genes which are up- or down-regulated, suggesting temporal variation in how wound response is regulated.

To determine how response to wounding differs from response to other environmental conditions, we measured how similar the pattern of differential gene expression was between (also downloaded from TAIR, see **Methods**). The Pearson's correlation coefficient (PCC) was used to compare the $\log_2$ fold change values across genes between wounding and other

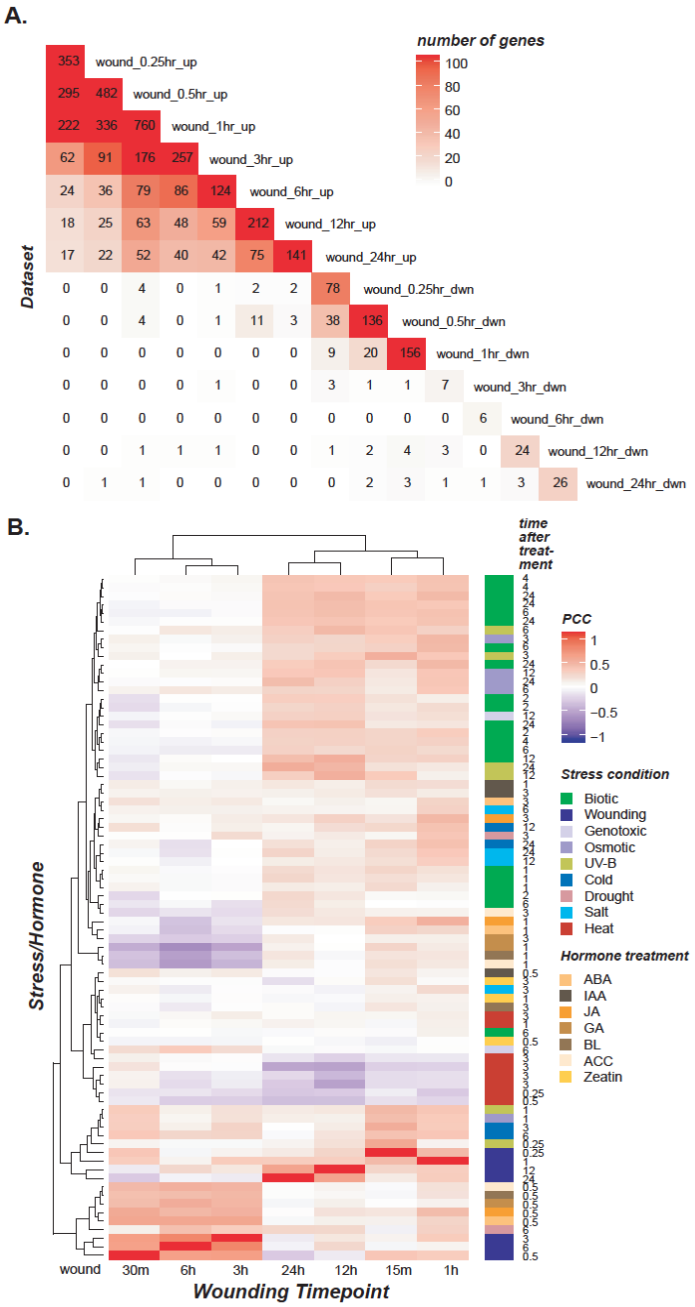


Figure 4.1. Gene expression correlation across stress and hormone data sets and the overlap of wound and JA differentially expressed genes.

A. Heatmap showing the number of genes overlapping in each wounding time point cluster. The order of rows and columns are the same, based on time point. Number of genes range from 0 (white) to 760 (red) and actual value is printed in the heatmap. B. Heatmap of Pearson's correlation coefficient (PCC) between data sets based on the log2 fold change between treatment and control. PCC values in heatmap range from 1 (red) to -1 (blue). The order of the rows and

Figure 4.1 (cont'd)
columns are the same, which is based on hierarchal clustering. The stress and hormone treatments are labeled by color and stress time point is labeled on y-axis.
Different wounding time points and other abiotic stress, biotic stress, and hormone treatments

stress/hormone treatments and then hierarchal clustering was used to find conditions where differential gene expression was most similar (**Figure 4.1B**). We found that gene expression patterns 30 minutes, 3 hours, and 6 hours after wounding clustered together, 24 hours and 12 hours after wounding cluster together, and then 15 minutes and 1 hour after wounding cluster together. To better understand this pattern, we first looked at how wounding time points correlated with abiotic stress response.

We found a cascading effect, where the majority of genes up-regulated at 15 and 30 minutes after wounding are still up-regulated at one hour (63% and 70% respectively), but by three hours <25% of those genes were still up-regulated (**Figure 4.1A, Dataset S13**). Consequently, the genes up- or down-regulated at later time points tended to be different than those differentially expressed earlier, with the genes responsive at 12 and 24 hours after wounding having the least amount of overlap with genes from previous wounding time points (**Figure 4.1A, Dataset S13**). Thus, different time points after wounding have overlapping but distinct sets of genes which are up- or down-regulated, suggesting temporal variation in how wound response is regulated.

To determine how response to wounding differs from response to other environmental conditions, we measured how similar the pattern of differential gene expression was between different wounding time points and other abiotic stress, biotic stress, and hormone treatments (also downloaded from TAIR, see **Methods**). The Pearson's correlation coefficient (PCC) was used to compare the $\log_2$ fold change values across genes between wounding and other stress/hormone treatments and then hierarchal clustering was used to find conditions where differential gene expression was most similar (**Figure 4.1B**). We found that gene expression patterns 30 minutes, 3 hours, and 6 hours after wounding clustered together, 24 hours and 12

hours after wounding cluster together, and then 15 minutes and 1 hour after wounding cluster together. To better understand this pattern, we first looked at how wounding time points correlated with abiotic stress response.

Patterns of DGE 15, 30 minutes or 1 hour after wounding, correlated more strongly with those of other early abiotic stresses compared to late wounding time points (12 or 24 hours after wounding). For example, DGE response 15 minutes after wounding had a PCC of 0.51 to the DGE response 15 minutes after UV-B light treatment and a PCC of 0.48 to the response to cold treatment after 3 hours. In contrast, the correlation of the response patterns between 15 minutes after wounding and 12 or 24 hours after wounding was 0.23 and 0.11, respectively. Gene expression patterns at 30 minutes and 1 hour after wounding were also similar to those under certain abiotic stresses, such as cold, UV-B, osmotic, and genotoxic stress. Additionally, early DGE 15 minutes and 1 hour after wounding were more similar to each other (PCC= 0.39) and to 30 minutes after wounding (PCC= 0.33 and 0.30 respectively) than to later time points (for PCC results, see **Dataset S12**). Curiously, DGE 30 minutes after wounding was highly correlated with DGE 3 and 6 hours after wounding (PCC = 0.59 and 0.57, respectively), while responses 3 and 6 hours after wounding are also highly correlated with each other (PCC= 0.55, **Dataset S12**). Thus, there is some association between response at mid-range time points and early time points. Similarly, expression patterns of later time points after wounding (12 and 24 hours after wounding) correlated most strongly with each other (PCC=0.55) compared with other earlier wounding time points (PCC ranging from -0.25 to 0.28). The DGE response from the 12 and 24 hour time points also had a high correlation to the DGE response at 12 or 24 hours after UV-B or osmotic treatment, showing late wound response is more similar to other late abiotic stress responses than to early wound response. Thus, transcriptomic responses were more similar

among comparable time points between treatments than largely differing time points within wounding. This indicates that temporal patterns can impact gene expression more than the type of abiotic stress.

Certain types of biotic stresses, such as insect chewing, can create wounds in plants. Therefore, although wounding is an abiotic stress, we wanted to see how different wounding timepoints may correlate with a range of biotic stresses. When observing wounding response patterns in relation to biotic stress, 15 minutes, 1 hour, 12 hours, and 24 hours after wounding have the highest correlations to biotic stress DGE response out of all wounding time points (**Figure 4.1B**). The biotic stresses included pathogens *Pseudomonas syringae* and *Phytophthora infestans*, as well as pathogen-derived elicitors Flagellin (bacterial), necrosis-inducing *Phytophthora* protein (oomycete), and Hairpin Z (bacterial). While the 15 minute, 1, 12, and 24 hour time point responses all correlate with the different types of biotic stress listed above, the 12 and 24 hour time point DGE responses have a higher correlation to (PCC range from 0.35 to 0.47) the DGE responses to *P. infestans* than any other wounding time point (**Figure 4.1B, Dataset S12**). *P. infestans* is a necrotrophic oomycete that creates extensive tissue damage in the plant, which may be similar to wounding damage, and the later time points of both stresses may be most similar because of similarities in immune response and recovery. It is interesting that wound DGE response at 3 and 6 hours after wounding do not correlate with biotic stress response (PCC range -0.09-0.07, **Dataset S12**). One hypothesis is that initial response to wounding triggers some of the same pathways involved in response to other biotic stresses and the late response to wounding triggers pathways involved in recovery from other biotic stresses, but the middle time points are involved in separate functions from biotic stress. This could explain similarities in DGE response seen between early and late time points.

Hormonal responses are also triggered by wounding, including JA, ABA, and ACC (ethylene), and may correlate with wound response. While JA response is known to be induced by wounding (Chung et al., 2008), ABA is induced during multiple abiotic stress responses (Nakashima et al., 2009), and ACC can also have both positive and negative interactions with JA to promote a synergistic response to wounding (Lorenzo et al., 2003). Also, certain hormones, including IAA, ABA, and JA have a positive relationship where the hormones up- or down-regulate the same genes (Goda et al., 2008). While DGE 15 minutes after wounding was not similar to DGE after hormone treatment, by 30 minutes after wounding, DGE was similar to DGE 30 minutes after treatment with five different hormones (ABA, ACC, brassinosteroid (BL), gibberellic acid (GA), and JA, PCCs ranging from 0.37-0.52; **Figure 4.1B**, **Dataset S12**), indicating a strong temporal component to how genes are expressed, and the initial response to wounding may involve multiple hormones. The DGE responses at time points at 3 and 6 hours after wounding were more similar to the DGE response to plant hormones at 30 minutes including ABA, ACC, BL, GA, and JA (PCC range from 0.54-0.39), than most other wounding time points with the exception of 30 minutes after wounding (PCC range from -0.04-0.21; **Figure 4.1B, Dataset S12**). Similarities between several hormonal responses and response to wounding at mid-range time points indicate that many stress-responsive hormones may still be involved in wound response even after 6 hours. Finally, 12 and 24 hours after wounding, transcriptomic responses show little correlation with DGE responses to the hormones correlated with earlier time points, with the highest correlation to DGE response to JA treatment after 3 hours (PCC = 0.26 and 0.14, respectively, **Dataset S12**). This indicates that the later responses to wounding may not signal stress-responsive hormones or that they do not correlate with the early transcriptomic responses to many hormone treatments. Overall, the high association of DGE

patterns in early and mid-range time points after wounding to early hormone treatment DGE patterns indicates an interaction between wound response and hormone response, in which wound response likely signals various stress-related hormones, but this interaction lessens over more time after wounding.

Modeling temporal wound response using machine learning

The temporal differences in transcriptional response to wounding described above suggest that the regulation of wounding response changes over time, with regulatory control being more similar within early time points (0.25, 0.5, 1 hours), middle time points (3 and 6 hours) and late time points (12 and 24 hours) compared to between these time points. In order to compare what regulatory mechanisms were important across different time points, we first needed to model the regulatory code of transcriptional response to wounding for time point. We used a machine learning approach to generate models of the regulatory code that could classify a gene as being differentially regulated or non-differentially regulated at a specific time point.

First, we tested how well-known sequence based regulatory information was able to model wounding response. We collected 52 known *cis*-regulatory elements (CREs) associated with JA, wounding, or insect response identified previously using experimental or computational approaches (see **Methods**; Dataset S14). To incorporate the known CRE information into a model, we mapped each putative regulatory sequence to the promoters (defined as 1 kb upstream of the transcription start site, see **Methods**) of each gene in a cluster (each cluster consisting of genes up- or down- regulated after wounding at a given time point), as well as to genes in a "null" cluster, consisted of genes that are not significantly up-regulated or down-regulated under any stress or hormone treatment. Two algorithms, Random Forest (RF) and Support Vector Machine (SVM) were used to build models for each wounding response timepoint. To measure

model performance, F-measure was used which jointly considers precision, or the number of genes that were differentially expressed and were predicted as differentially expressed over the number of all of genes the model predicted as differentially expressed, and recall, or the number of genes which were differentially expressed and were predicted as differentially expressed over the number of genes which were differentially expressed (predicted or not). The F-measures for models built for each wounding time point cluster ranged from 0.67 to 0.71, scores that show our models performed better than random guessing (F-measure = 0.5) but were not perfect predictors (F-measure = 1) (**Figure 4.2, Dataset S15**). Note that three wounding response timepoints were not included in this analysis: down-regulated 3 and 6 hours after wounding because too few genes had these responses to train a machine learning model (<10) and down-regulated 12 hours after wounding category because no known regulatory elements were present in the promoters of the genes in this group.

Second, we incorporated additional levels of regulatory information into our models. We included *in vitro* DNA binding data of 510 TFs in *A. thaliana* generated with DNA affinity purification sequencing (DAP-seq) (O'Malley et al., 2016) and information about Dnase I Hypersensitive Sites (DHS) in *A. thaliana* at different developmental stages including seedling (leaf samples) and two-week old plants (flower buds) (Zhang et al., 2012). Each DAP-seq and DHS feature was considered present if its peak coordinates overlapped with the promoter region of a gene. Machine learning models trained using both known sequence and DAP-seq and DHS features performed slightly better overall than known sequence-based models alone, with the F-measure ranging from 0.66 to 0.74 (**Figure 4.2, Dataset S15**). Models for genes up-regulated in early wounding response (0.25, 0.5, and 1 hours) benefited the most from the addition of these two data sets, with a +0.03, +0.03, and +0.02 improvement in F-measure, respectively. This may

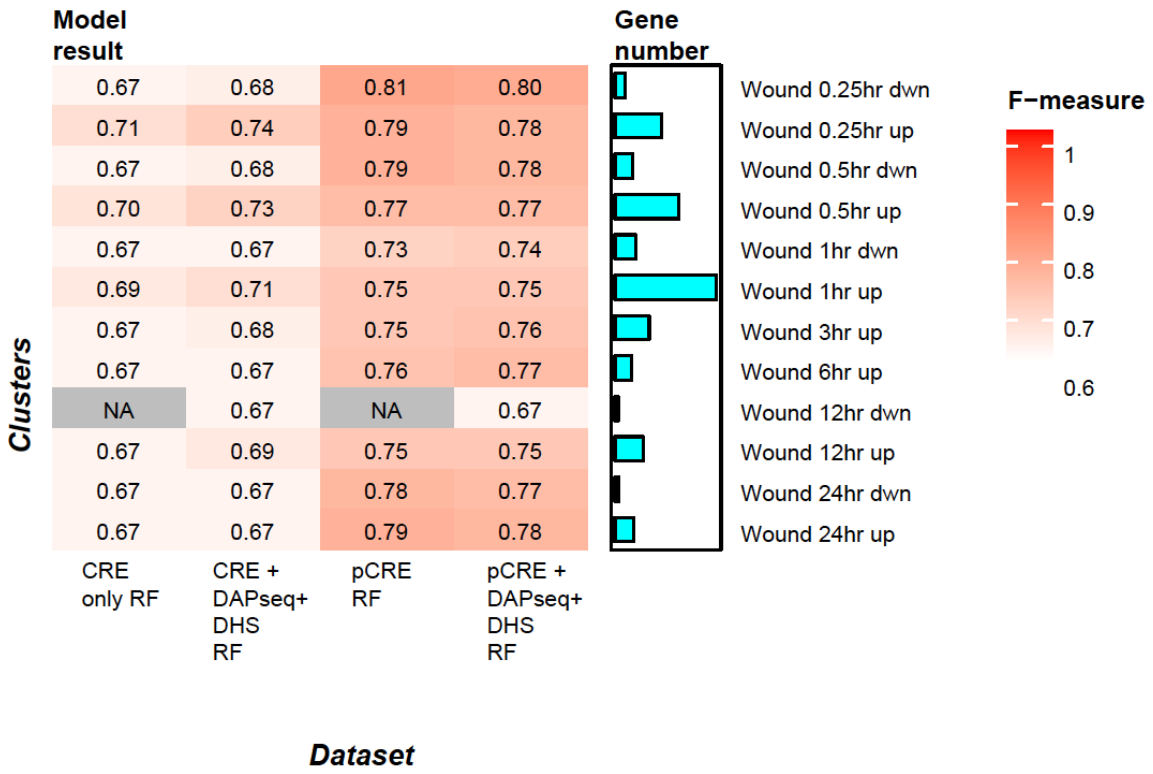


Figure 4.2. Heatmap of the F-measure for wounding time point models.

Each row is a different cluster which was used to build a separate model. Each column represents the datasets used as features in the model and the algorithm used (RF= Random Forest). Known only refers to CREs found in the literature (**Dataset S14**). DAPseq and DHS refer to the DAP-seq and Dnase I hypersensitivity sites. FET enriched 6mer refers to the pCREs which were enriched for a specific cluster. The F-measure range is from 0.5 (white) to 1 (red), and gradient as well as actual F-measure is shown in each cell. The bar chart next to the heat map corresponds to each row/cluster and represents the number of genes in that cluster.

be because most known CREs are known from early wound response. Thus, more known information in the form of the DAP-seq data may improve the performance of early time point clusters more than later time points. Overall, while known sequence-based information and DAP-seq and DHS information is predictive of differential gene expression in response to wounding across time points, the models still have substantial room for improvement.

Determining important known motifs for temporal wound response

To understand what known elements are important for driving expression at different times after wounding, we measured the importance of each feature (known CRE, DAP-seq, or DHS, see **Methods**) in each model in order to rank features in terms of how important they were for our ability to predict differential gene expression at different time points (**Dataset S16**). For early wound response (genes up-regulated 0.25, 0.5, and 1 hour after wounding), the most important known wound CREs identified by our models were CGCGTT (first ranked), a known regulatory elements for Rapid Wound Response (RWR) (Walley et al., 2007) and CACGTG (second ranked) that is bound by some *Myc* TFs in the basic Helix-Loop-Helix (bHLH) family in response to wounding and JA treatment (Fernández-Calvo et al., 2011). Genes with the RWR elements are known to respond quickly to wounding and have a variety of functions in the downstream response, including chromatin remodeling, signal transduction, and mRNA processing (Walley et al., 2007). These functions are consistent with stress-induced transcriptional changes, where chromatin conformation is changed to modulate binding of stress-related TFs, mRNAs are modified post-transcriptionally, and signaling pathways up and down-stream to transcription are involved in response to wounding stress. Other TFs that respond to wounding stress, *Myc* 2, 3, and 4 TFs, respond to both JA and wounding, and induce other JA responsive genes, ultimately triggering defense response to herbivory (Fernández-Calvo et al.,

2011). In addition to genes up-regulated 0.25 ~ 1 hour post wounding, CACGTG, the wound response element that *Myc* TFs bind, was still important (ranked 1 or 2) among genes up-regulated 3, 6 and 12 hours after wounding, but not the RWR element. By 24 hours after wounding, the CACGTG element was no longer important.

DAP-seq binding sites were less important in predicting wound response than the known sequence-based sites, but still rank among the top 10 most important features for models predicting early wound response (**Dataset S16**). For example, the CAMTA TF family binding site, AAGCGCGTG, was ranked 3rd most important for genes up-regulated 0.25 or 0.5 hours after wounding but dropped to 11th at 1 hour after wounding, and even lower in later time points. At 1 hour, the AP2EREBP TF family binding site, GGCGGCGGCGG, started to become more important, ranking 10th in the model and increasing to 4th at 3 hours after wounding. In contrast, all DAP-seq sites became less important, or not important at all, for predicting genes up-regulated 6, 12, and 24 hours after wounding. These findings highlight temporal differences in wounding regulation, but also that in-vitro TF binding sites do not capture the entirety of how wounding response is regulated, especially at later time points.

In addition to known *cis*-regulatory elements and DAP-seq sites, open chromatin sites (DHS) were important for predicting expression at all time points after wounding (top ranked DHS sites for each cluster ranged from rank 1~4). However, DHS features tended to become more important at later time points (**Dataset S16**). For example, while different types of features were important at earlier time points, at 24 hours after wounding, the top 12 most important features were all DHS-related. This is rather intriguing and suggests epigenetic modifications are more important for later response to wounding. We propose three potential explanations for this finding. The first is that at the late time point transcription factor binding is no longer the major

determinant of regulation since the chromatin state for wounding stress has been established. While chromatin state does change under JA or wound stress, it is not clear to what extent or for how long (Berr et al., 2012). Second, at this later time point, the functional diversity of genes has increased to a point that their transcriptional regulatory mechanisms have become more heterogeneous and thus no single CRE or DAP-seq feature has high importance. The third possibility is that the known CRE or DAP-seq features important for later time points are not present in our dataset. This could be because the later time points are not as well studied or because DAP-seq data is only available for ~38% of known TFs (Weirauch et al., 2014; O'Malley et al., 2016), which would suggest that there are novel regulatory sequences that have not yet been identified.

Finding important temporal putative cis-regulatory elements for wound response using machine learning

Although known CREs, *in vitro* TF binding data, and DHS are useful for building wound response prediction models for the clusters in **Figure 4.2**, the model performance is far from perfect which raises the question whether additional CREs remain to be discovered that can better explain the gene expression patterns seen. To discover novel putative CREs (pCREs), a *k*-mer finding approach was used (modified from Liu et al., 2018), where all possible 6-30-mer sequences were tested for enrichment (p<0.01, see **Methods**) in the putative promoters of genes for each cluster (see **Methods**). Based on this criteria, between 42-1,081 pCREs were identified as enriched in genes from each wound response cluster, with the exception of the down-regulated wounding after 12 hours cluster, which had no enriched pCREs (**Dataset S17**). For each wound response cluster, the pCREs were used to build a wound response prediction model. We found that models built with pCREs alone (e.g. RF algorithm, F-measure range = 0.73-0.81) perform

better than models built with known CREs, DAP-seq and DHS for all clusters (e.g. RF algorithm, median F-measure range = 0.66 to 0.74, **Figure 4.2, Dataset S15**). Interestingly, models that combined pCREs with known CREs, DAP-seq, and DHS data did not perform better than pCRE-based models alone (e.g. RF-algorithm, median F-measure = 0.67-0.80). This indicates that these pCREs, some are variants of known CREs and other novel, may contribute substantially to the regulation of wound response at different time points.

To understand why the models improve with the addition of pCREs, and what impact pCREs have across wounding time points relative to known information and open chromatin sites, we looked at the relative importance (normalized importance score, see **Methods**) of pCREs, DAP-seq sites, and chromatin accessibility sites across the post-wounding time course (**Figure 4.3).** Overall, DHS sites tended to be most important, followed by pCREs, and then finally by DAP-seq sites. When looking closer at individual up-regulated clusters, we found for early time point clusters (0.25, 0.5, and 1 hour after wounding, **Figures 4.3A-C**), a small percentage of pCREs have higher or as high of an importance score as the most important DHS sites. This indicates that a few unique regulatory elements, or variations of known elements which are not captured by the known TF binding sites are important for distinguishing differential expression at early time points. Other than this small subset of pCREs, DH-sites are in general more important than the majority of pCREs or DAP-seq sites. For middle range to later time points (3, 6, 12, 24 hours after wounding, **Figures 4.3D-F**), DH sites have the highest importance, but certain pCREs are also important, ranking just below the most important DH sites. In fact, there is a general shift where the majority of pCREs are of higher importance at these time points than at earlier time points (**Figure 4.3G**). This indicates that even though the most important features at late time points are the DH sites, there are more putative TF binding sites that distinguish mid-range to

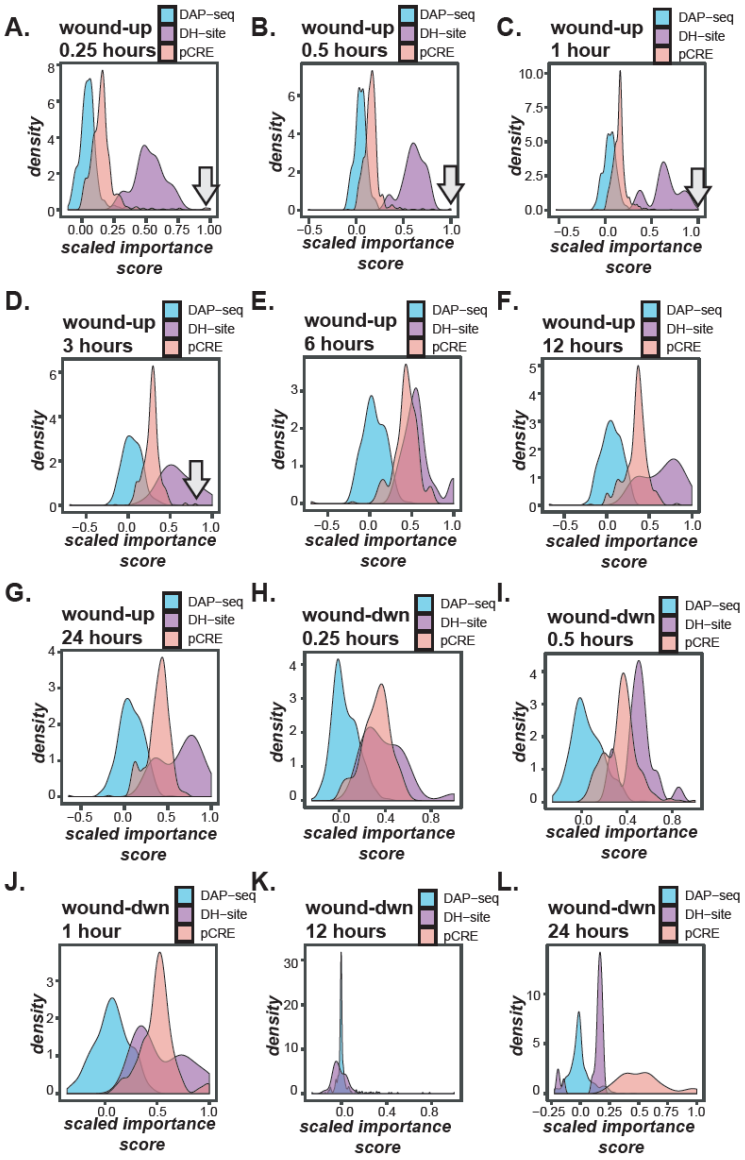


Figure 4.3. Scaled importance value for each up-regulation wounding time point model (rows) for all features used in the final model.

For A-L, density plots show importance value on the x-axis and the density of the feature on the y-axis. DAP-seq sites are in blue, DH sites are in purple, and pCREs are in pink. The importance value is scaled from -1 to 1 for each model, where positive value correlates with differential expression in a given cluster and -1 correlates with the null cluster. The higher absolute value correlates with higher importance of a given feature for a given model. Arrows point to a small peak of pCREs in figures A-D. A. wound up-regulated at 0.25 hours model, B. wound up-regulated at 0.5 hours model, C. wound up-regulated at 1 hour model, D. wound up-regulated at 3 hours model, E. wound up-regulated at 6 hours model, F. wound up-regulated at 12 hours model, G. wound up-regulated at 24 hours model, H. wound down-regulated at 0.25 hours model, I. wound down-regulated at 0.5 hours model, J. wound down-regulated at 1 hour model, K. wound down-regulated at 12 hours model, and L. wound down-regulated at 24 hours model.

late wound induced gene expression compared to earlier time points.

Finally, down-regulation importance patterns are less consistent (**Figures 4.3H-L**) While genes down-regulated at 0.25, 0.5, and 1 hour are fairly similar where either DH sites or pCREs are the most important for regulation, down-regulated genes at 12 hours have no important pCREs, and down-regulated genes at 24 hours have almost exclusively pCREs as being important for regulation. This indicates that 24 hours after wounding is under the most unique regulation compared to all other time points, where up-regulated genes are mostly regulated by open chromatin, and down-regulated by pCREs. Also, it is important to note that open chromatin sites appear to be important features for models at all time points. While some sites are more important to earlier time points than they are to later ones, many sites have an equal relative importance across wounding time point, indicating open chromatin sites cannot distinguish the different expression patterns at different times after wounding. Together, the most distinguishing regulation of wounding time points may be the pCREs.

Correlation to transcription factor families and cis-regulatory differences across time

Next we wanted to determine which pCRE were similar to a known TF binding motif and which were likely to be novel regulatory elements. To do this, we first calculated the sequence similarity between each pCRE and each known binding motif in order to find the TF family who's binding motifs best matched the pCRE. **Figure 4.4** shows the importance rank across all time points for the top 10 most important pCREs for each wounding model. Similar to how more of the same genes were differentially expressed at nearby time point (i.e. the cascade effect), we found more important pCREs were shared with close time points, however some pCREs were uniquely important at a single time point. In fact, none of the top 10 most important pCREs were shared across all time points.

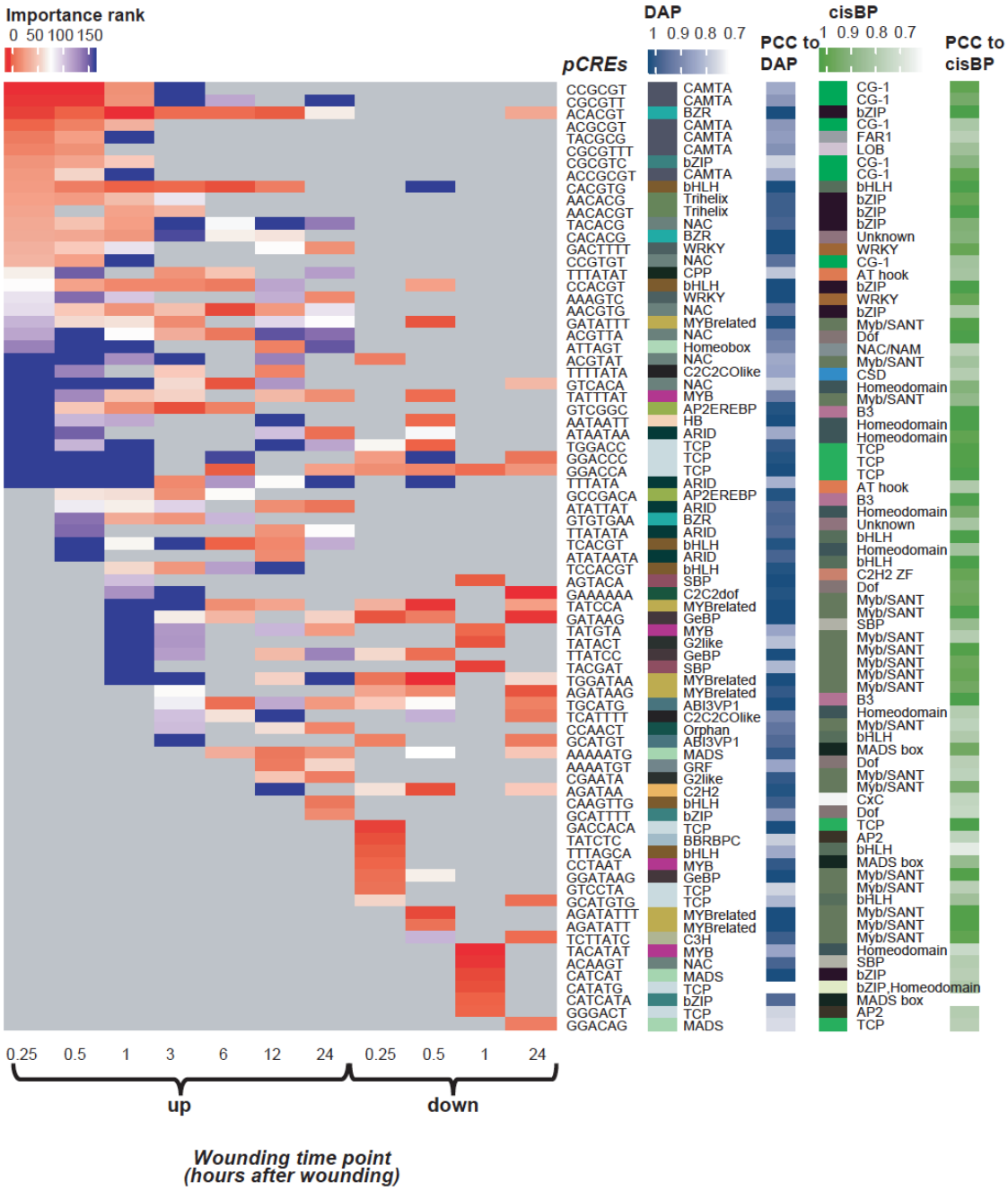
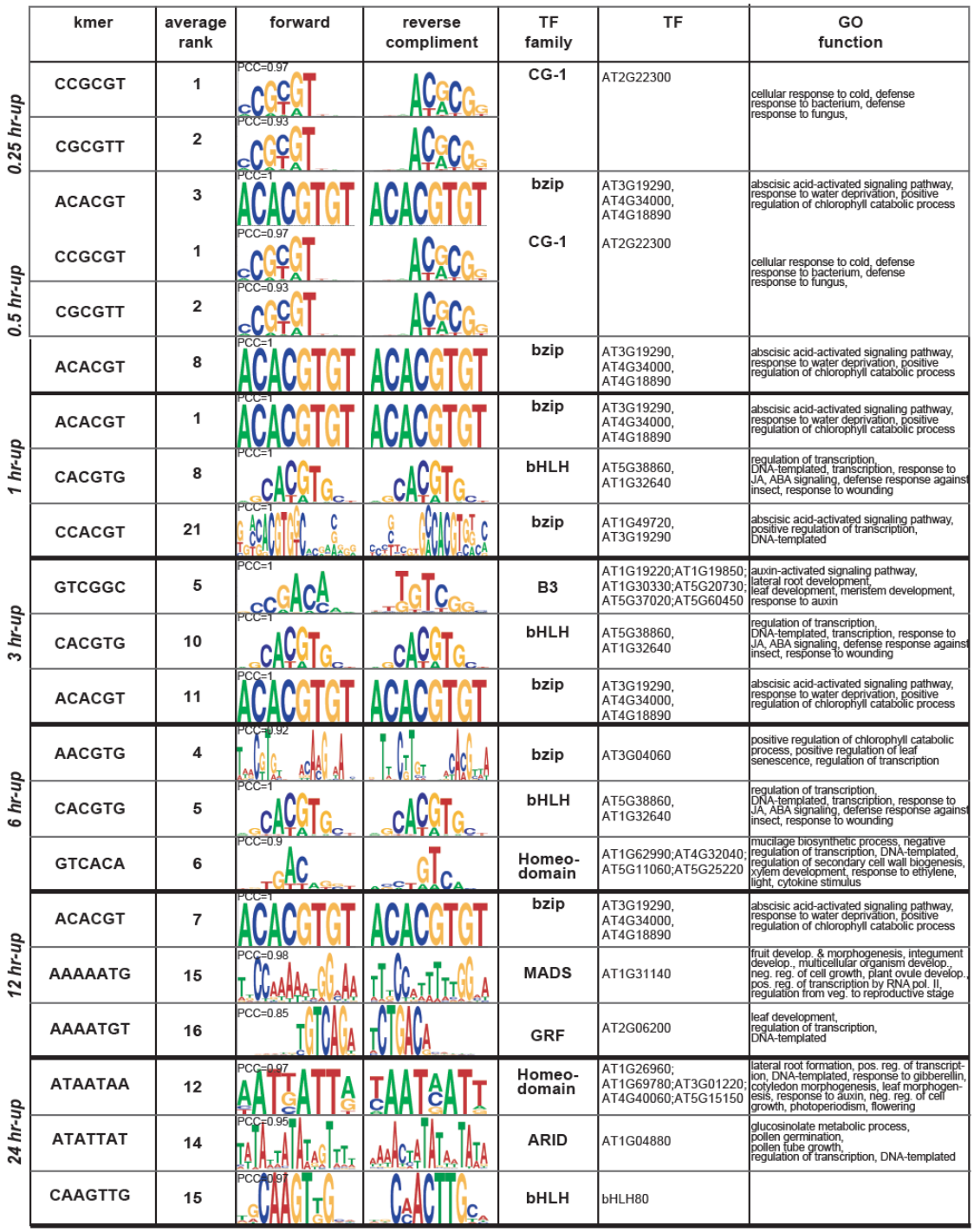


Figure 4.4. Average importance rank for the top 10 pCREs for each wounding time point model and their TF family.

Wound time point models are the columns while pCREs are the rows. Average importance rank is the average rank across five duplicate models ran for the same time point. Highest rank (1) is red and ranks 150 or lower are blue. TF family association is based on the maximum PCC to known TF binding sites. PCC is shown for both DAP-seq and TFBM sites.

For early time points after wounding (i.e. 0.25, 0.5, and 1 hour), many of the top important pCREs were shared and resemble TF binding sites in the *CG-1*, *bZIP*, *FAR1*, *LOB*, and *bHLH* TF families (right two panels; **Figure 4.4**). Different binding sites which bind multiple TF families is consistent with the notion that a variety of signals are induced by wounding. For example, while JA is induced by wounding, other hormones or signals involved, such as ACC, hydrogen peroxide, and ABA, can amplify the JA response (Howe, 2004). Focusing on the top 3 most important pCREs for each time point (excluding DAP-seq or DH sites, **Figure 4.5**), we found that 0.25 and 0.5 hours after wounding, CCGCGT, which is most similar to the binding motif of a *CG-1* TF family TF, was the most important pCRE for up-regulated gene models, it then dropped to the 29th most important pCRE at 1 hour after wounding, and by 6 hours after wounding is not enriched in the cluster (**Figure 4.5**; for the importance rank of pCRE and PCC to known TF binding motifs from each TF family, see **Dataset S16**). This binding site is associated with TFs which respond to both abiotic and biotic stress. On the other hand, one hour after wounding, the pCRE CACGTG, which was not as important previously (0.25 hour rank = 30, 0.5 hour rank = 17), was the 8th most important pCRE. This pCRE was most similar to the known binding motif for *Myc2*, a *bHLH* TF that responds to both JA and wounding (Dombrecht et al., 2007) (**Figure 4.5, Dataset S16**). This element remained important at both 3 and 6 hours after wounding (ranked 10 and 5, respectively), indicating a change in response at 1, 3, and 6 hour time points and that JA hormones have been activated. Other important early pCREs remained important across the wider range of time points. One example is ACACGT, a pCRE most similar to the known binding motif for *bZIP* family TFs, which are activated by ABA (Yamamoto et al., 2011) and regulate responses to water deprivation (**Figure 4.5**). This pCRE was enriched in the promoters of genes from all time points and was important (rank < 11) for



	kmer	average rank	forward	reverse compliment	TF family	TF	GO function
0.25 hr-up	CCGCGT	1	PCC=0.97		CG-1	AT2G22300	cellular response to cold, defense response to bacterium, defense response to fungus,
	CGCGTT	2	PCC=0.93				
0.5 hr-up	ACACGT	3	PCC=1		bzip	AT3G19290, AT4G34000, AT4G18890	abscisic acid-activated signaling pathway, response to water deprivation, positive regulation of chlorophyll catabolic process
	CCGCGT	1	PCC=0.97		CG-1	AT2G22300	cellular response to cold, defense response to bacterium, defense response to fungus,
	CGCGTT	2	PCC=0.93				
1 hr-up	ACACGT	8	PCC=1		bzip	AT3G19290, AT4G34000, AT4G18890	abscisic acid-activated signaling pathway, response to water deprivation, positive regulation of chlorophyll catabolic process
	ACACGT	1	PCC=1		bzip	AT3G19290, AT4G34000, AT4G18890	abscisic acid-activated signaling pathway, response to water deprivation, positive regulation of chlorophyll catabolic process
	CACGTG	8	PCC=1		bHLH	AT5G38860, AT1G32640	regulation of transcription, DNA-templated, transcription, response to JA, ABA signaling, defense response against insect, response to wounding
	CCACGT	21	PCC=1		bzip	AT1G49720, AT3G19290	abscisic acid-activated signaling pathway, positive regulation of transcription, DNA-templated
3 hr-up	GTCGGC	5	PCC=1		B3	AT1G19220;AT1G19850; AT1G30330;AT5G20730; AT5G37020;AT5G60450	auxin-activated signaling pathway, lateral root development, leaf development, meristem development, response to auxin
	CACGTG	10	PCC=1		bHLH	AT5G38860, AT1G32640	regulation of transcription, DNA-templated, transcription, response to JA, ABA signaling, defense response against insect, response to wounding
	ACACGT	11	PCC=1		bzip	AT3G19290, AT4G34000, AT4G18890	abscisic acid-activated signaling pathway, response to water deprivation, positive regulation of chlorophyll catabolic process
6 hr-up	AACGTG	4	PCC=0.92		bzip	AT3G04060	positive regulation of chlorophyll catabolic process, positive regulation of leaf senescence, regulation of transcription
	CACGTG	5	PCC=1		bHLH	AT5G38860, AT1G32640	regulation of transcription, DNA-templated, transcription, response to JA, ABA signaling, defense response against insect, response to wounding
	GTCACA	6	PCC=0.9		Homeo-domain	AT1G62990;AT4G32040; AT5G11060;AT5G25220	mucilage biosynthetic process, negative regulation of transcription, DNA-templated, regulation of secondary cell wall biogenesis, xylem development, response to ethylene, light, cytokine stimulus
12 hr-up	ACACGT	7	PCC=1		bzip	AT3G19290, AT4G34000, AT4G18890	abscisic acid-activated signaling pathway, response to water deprivation, positive regulation of chlorophyll catabolic process
	AAAAATG	15	PCC=0.98		MADS	AT1G31140	fruit develop. & morphogenesis, integument develop., multicellular organism develop., neg. reg. of cell growth, plant ovule develop., pos. reg. of transcription by RNA pol. II, regulation from veg. to reproductive stage
	AAAATGT	16	PCC=0.85		GRF	AT2G06200	leaf development, regulation of transcription, DNA-templated
24 hr-up	ATAATAA	12	PCC=0.97		Homeo-domain	AT1G26960; AT1G69780;AT3G01220; AT4G40060;AT5G15150	lateral root formation, pos. reg. of transcript-ion, DNA-templated, response to gibberellin, cotyledon morphogenesis, leaf morphogen-esis, response to auxin, neg. reg. of cell growth, photoperiodism, flowering
	ATATTAT	14	PCC=0.95		ARID	AT1G04880	glucosinolate metabolic process, pollen germination, pollen tube growth, regulation of transcription, DNA-templated
	CAAGTTG	15	PCC=0.97		bHLH	bHLH80	

Figure 4.5. Motif logos for the top 3 pCREs for each up-regulated wounding time point.

Chart is divided by time point (0.25 to 24 hours after wounding). The first column is the top 3 ranked pCREs for that time point. The second column is the average rank for that pCRE in the given model. The third and fourth columns are the best matched TF binding motif logos, forward and reverse compliment with PCC value. Columns 5-7 are the TF which binds a given logo (column 6), the TF family the TF belongs to (column 5) and GO functions of the TF (column 7).

models of wounding response at 0.25, 0.5, 1, 3, 6, and 12 hours (**Figure 4.4, Dataset S16**).

Two pCREs (GTCGGC and GTCACA) were uniquely important for models built for mid-range time points (i.e. 3 and 6 hours after wounding), as the 5th and 6th most important pCREs for the genes up-regulated at 3 hours and 5th and 18th most important at 6 hours (**Figure 4.5**). These elements were most similar to binding motifs of B3 and Homeodomain family TFs, respectively. Given these TF families are involved in development, response to auxin, and secondary wall biogenesis, this indicates that by 3 to 6 hours after wounding, the damage is likely being repaired.

At the latest time points (i.e. 12 and 24 hours after wounding), we found that some important pCREs were the same as the pCREs important for earlier time points, while others were unique to the later response. As discussed previously, ACACGT, which was ranked 7th at 12 hours after wounding and was also important for earlier time points (0.25, 0.5, 1, 3, and 6 hours after wounding). While ATATTAT, which was most similar to binding motifs of TFs in the ARID family, was ranked 14th at 24 hours after wounding (**Figure 4.5, Dataset S16)** This TF family is involved in regulating glucosinolate metabolism. Other important pCREs at the latest time points, ATAATAA and AAAATGT, were elements that bind TF families which regulate development (**Figure 4.5, Dataset S16**).

In summary, we found that pCREs important for our models of response at early time points (0.25 to 0.5 after wounding) tend to be associated with many stress and hormone responses, while pCREs 1 to 6 hours after wounding tend to be associated with TFs involved in JA signaling and ABA signaling. Finally, from 3-24 hours after wounding the pCREs tend to be associated with TFs involved in growth and very late responses (12-24 hours after wounding) are associated with TFs related to metabolic defense. Overall, we generated models of the cis-

regulatory code in response to wounding that demonstrate how different sets of pCREs, which are likely bound by a variety of TFs, are important at different response times after wounding and could work to regulate a dynamic response to wounding over time.

Modeling the regulatory code of JA-induced and non-JA-induced response to wounding

Having demonstrated how wounding response regulation changes over time, we next wanted to study the regulatory differences between JA-induced and non-JA-induced wounding responsive genes. Non-JA-induced wounding responses include those induced by rnase and nuclease activities that are triggered by wounding but not the application of JA (LeBrasseur et al., 2002). Thus to understand how non-JA induced wound responses are regulated, we used the hormone treatment data described above (Goda et al., 2008) to identify genes that were differentially expressed in response to wounding but not in response to JA at each timepoint. For this analysis, we only included timepoints (30 minutes, 1 hour, and 3 hours) for which we had data for both JA treatment and wounding. Across these three timepoints only 16%, 26%, and 28% of genes up-regulated after wounding were also up-regulated after JA treatment, respectively (**Figure 4.6A**). The large number of non-JA induced wounding responsive genes is consistent with other studies of wounding which relay a JA-independent mechanism for wounding response (León et al., 1998; LeBrasseur et al., 2002). However, the regulatory code that governs these responses is less well understood.

To determine the regulatory differences between JA-induced and non-JA induced wounding response, we first needed to generate models of these different regulatory codes. Using the same approach as described above we generated machine learning models of wounding response based on known CREs, DAP-seq sites, DH sites, and pCREs for up- and down-regulation responses at different time points. However, here we divided our genes in each

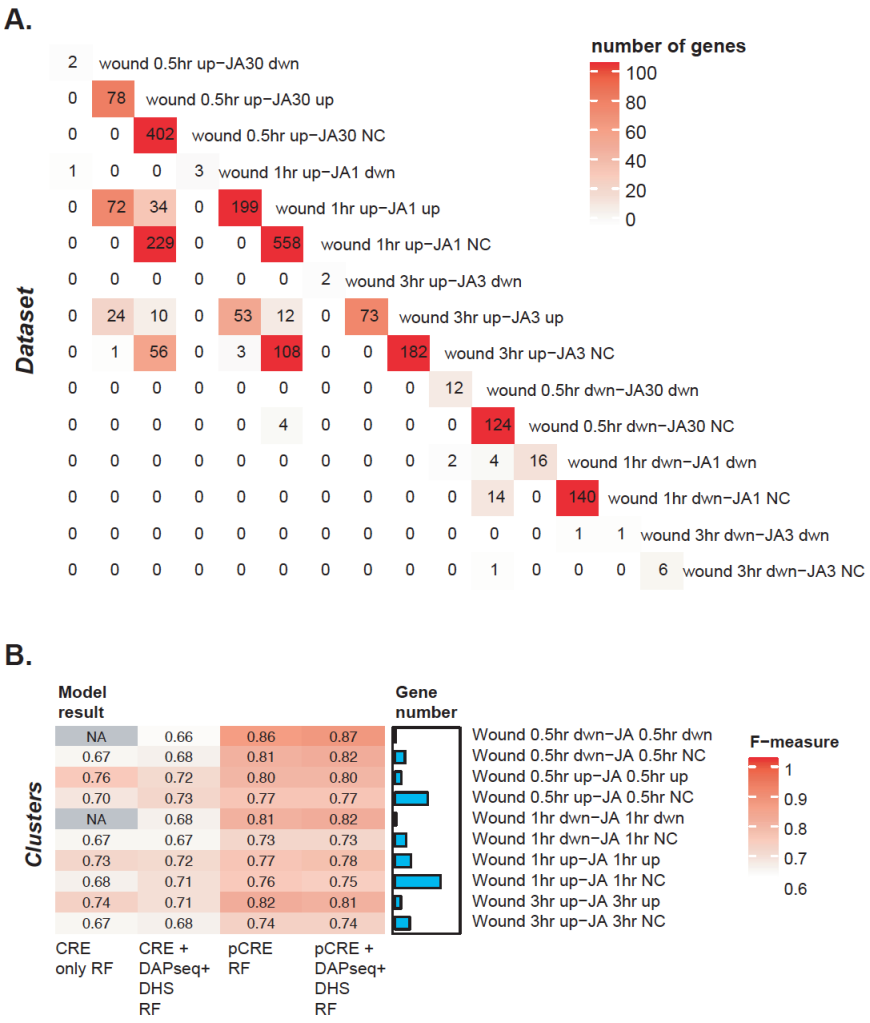


Figure 4.6. Gene overlap and model performance of each wound JA-induced and wound non JA-induced cluster.

A. Heatmap showing the number of genes overlapping in each wounding-JA cluster. The order of rows and columns are the same, based on time point. Number of genes range from 0 (white) to 558 (red) and actual value is printed in the heatmap. B. Each row is a different cluster (JA-induced or JA non-induced) which was used to build a separate model. Each column represents the datasets used as features in the model and the algorithm used (RF= Random Forest). Known only refers to CREs found in the literature (**Dataset S14**). DAPseq and DHS refer to the DAP-seq and Dnase I hypersensitivity sites. FET enriched 6mer refers to the pCREs which were enriched for a specific cluster. The F-measure range is from 0.5 (white) to 1 (red), and gradient as well as actual F-measure is shown in each cell. The bar chart next to the heat map corresponds to each row/cluster and represents the number of genes in that cluster.

cluster further by if they were differentially regulated under wounding and JA treatment (JA-induced) or under wounding but not JA treatment (non-JA induced). Note that models were not generated for genes down-regulated 3 hours after wounding because not enough genes were available for training. Similar to our earlier results, we found that pCRE based models (F-measures: 0.73 ~ 0.87) performed better than both known CREs based models (0.67 ~ 0.74) and known CREs and DAP-seq and DHS based models (0.66 ~ 0.73; **Figure 4.6B, Dataset S15**). This, again, indicated that pCREs were better able to model the regulation of JA-induced and non-JA-induced wounding response across time points, than using only known TF sites.

Nonresponsive JA CREs: Known and putative

To better understand differences between how JA-induced and non-JA-induced wounding responses are regulated, we next compared the importance of known CREs, DAP-seq sites, DH sites, and pCREs across models. We identified differences in which known CREs were important at 30 minutes and 1 hour after wounding between JA-induced and non-JA-induced responses. For example, CGCGTT, the *RWR* element, was the most important element for the non-JA-induced model, while for JA-induced models, the most important element was the *Myc* element, CACGTG (**Dataset S17**). Interestingly, the *Myc* element also ranks as the third most important feature in the non-JA-induced models. This could be because other TFs not involved in JA response (e.g. *Myc*-LIKE and *BIM3* TFs) can bind to this element (O'Malley et al., 2016) or because the *Myc* element may be necessary to facilitate TF binding to a different regulatory element important for non-JA-induced response. Finally, we found that chromatin accessible sites have a higher overall importance for non-JA-induced than for JA-induced wounding response. Out of the top 10 most important features, 4 to 8 were DH sites for non-JA-induced up-regulated models. In contrast, for JA-induced up-regulated models, none of the top 10 most

important features were DH sites (**Dataset S17**). This indicates that open chromatin sites are important for distinguishing genes that are non-JA-induced from those that are JA-induced.

Next we compared the importance of pCREs between JA-induced and non-JA-induced models and, with the exception of the G-box motif (CACGTG) and the bZIP binding site (ACGTGT), found there to be little overlap between the two (**Figure S4.2**). For example, AACGTG and CACGTTT were ranked from 1st to 7th across timepoints in JA-induced models but were not enriched or were ranked much lower (69th to 157th) for non-JA-induced models (**Figure S4.2, Dataset S16**). These pCREs were most similar to binding motifs of TFs in the NAC and CAMTA families, respectively. In contrast, CCGCGT and GCCGAC, were the most important pCREs 0.5 and 3 hours after wounding in the non-JA-induced models but were not enriched or were ranked much lower (232th importance) for JA-induced models **(Dataset S16**). These pCREs were most similar to the binding motifs of TFs in the CG-1 and B3 TF families, respectively. Interestingly, these TF families are known involve TFs that have a general response to stress as well as those which are involved in auxin signaling and development, indicating two functions of wound stress response which do not involve JA. Together, this highlights how JA-induced and non-JA-induced differential gene expression is likely regulated by different sets of regulatory elements that are recognized by different families of TFs.

Modeling SM pathway regulation using wound stress data

Another way to study response to wounding is by focusing on the response of whole metabolic pathways instead of the response of individual genes. Here, we measured the degree to which genes annotated as belonging to a particular specialized metabolism pathway were enriched in the genes up-regulated across the time series (**Dataset S17**). At earlier to mid-range time points (from 0.25 to 3 hours after wounding) JA biosynthesis was the most enriched

pathway (*p*-values range from 0.0015 to 3.5e-07; **Dataset S17**). However, by 6 hours after wounding, JA biosynthesis genes were less enriched (*p*-values = 0.0018) and by 12 hours it is not enriched at all. This demonstrates how the JA biosynthesis pathway is only 176ctive early in wounding response. Genes from the glucosinolate biosynthesis from tryptophan (Gluc-Trp) pathway, on the other hand, were enriched 0.5 hours after wounding (*p*-value = 0.008), were most enriched 12 hours after wounding (*p*-value = 0.0008) and were not enriched by 24 hours after wounding (*p*-value = 0.1). Finally, two genes from the anthocyanin biosynthesis pathway were up-regulated 0.5 hours after wounding and the same two genes remain up-regulated through 24 hours after wounding. These examples demonstrate that some wounding responsive pathways are dynamic over time, while other wounding responsive pathways are steady.

To determine how dynamic changes in a metabolic pathway are regulated, we used the glucosinolate biosynthesis from tryptophan (Gluc-Trp) pathway as an example. No Gluc-Trp pathway genes were up-regulated at the earliest time point, however by 0.5 hours after wounding three genes were significantly up-regulated and by the one hour time point three additional genes were significantly up-regulated (see stars; **Figure 4.7A**). Looking beyond the first hour, we saw a cascading effect, where by 3 hours after wounding, the genes turned on at one hour still were still up-regulated, but the three genes that were first up-regulated at 0.5 hours were turned off. Continuing this trend, by six hours after wounding, only one gene that was up-regulated at one and three hours after wounding was still significantly up-regulated. (**Figure 4.7A**). This type of pattern could be due to genes upstream in the pathway being involved in up-regulating genes downstream in the pathway or could be due to having different TFs, not in the Gluc-Trp pathway, regulating up and downstream pathway genes.

To understand how the cascading response is regulated, , we mapped the pCREs found from each of the models for up-regulated genes at a given time point back to the promoters of the Gluc-Trp pathway genes (see Methods). **Figure 4.7B** shows the overlap of pCREs discovered from the wounding time point models that map to Gluc-Trp genes and their importance level at a given time point. We can see that starting at 0.5 hours after wounding, there is little overlap of important pCREs across time points with the exception of pCREs present at 6 and 12 hours after wounding. This indicates that for the Gluc-Trp pathway, genes turned on at different times have different regulatory elements which are specific to those genes. **Figure 4.7C** shows the highest ranked important pCREs for Gluc-Trp pathway genes at a given time point, and whether this pCRE is present or absent in Gluc-Trp pathway genes at other time points. For example, ACACGT, which is perfectly similar to the binding motif of a TF from the *bZIP* family (PCC=1), when specifically finding pCREs in Gluc-Trp genes, this pCRE is the most important element at 0.5 hours after wounding (**Figure 4.7C**). Additionally, this element is not found in Gluc-Trp pathway genes up-regulated at other time points. Other pCREs are important for regulating expression of pathway genes at later time points. For example, AACGTG, which is most similar to the binding motif of a *bZIP* family TF, is enriched in the promoters of Gluc-Trp pathway genes up-regulated 1, 3, 6, 12, and 24 hours after wounding, but has the highest importance at 6 hours after wounding. Overall genes belonging to the Gluc-Trp pathway have varied cis-regulatory elements depending on when they are up-regulated after wounding, and timing of response can be an important consideration when finding CREs related to certain pathways.

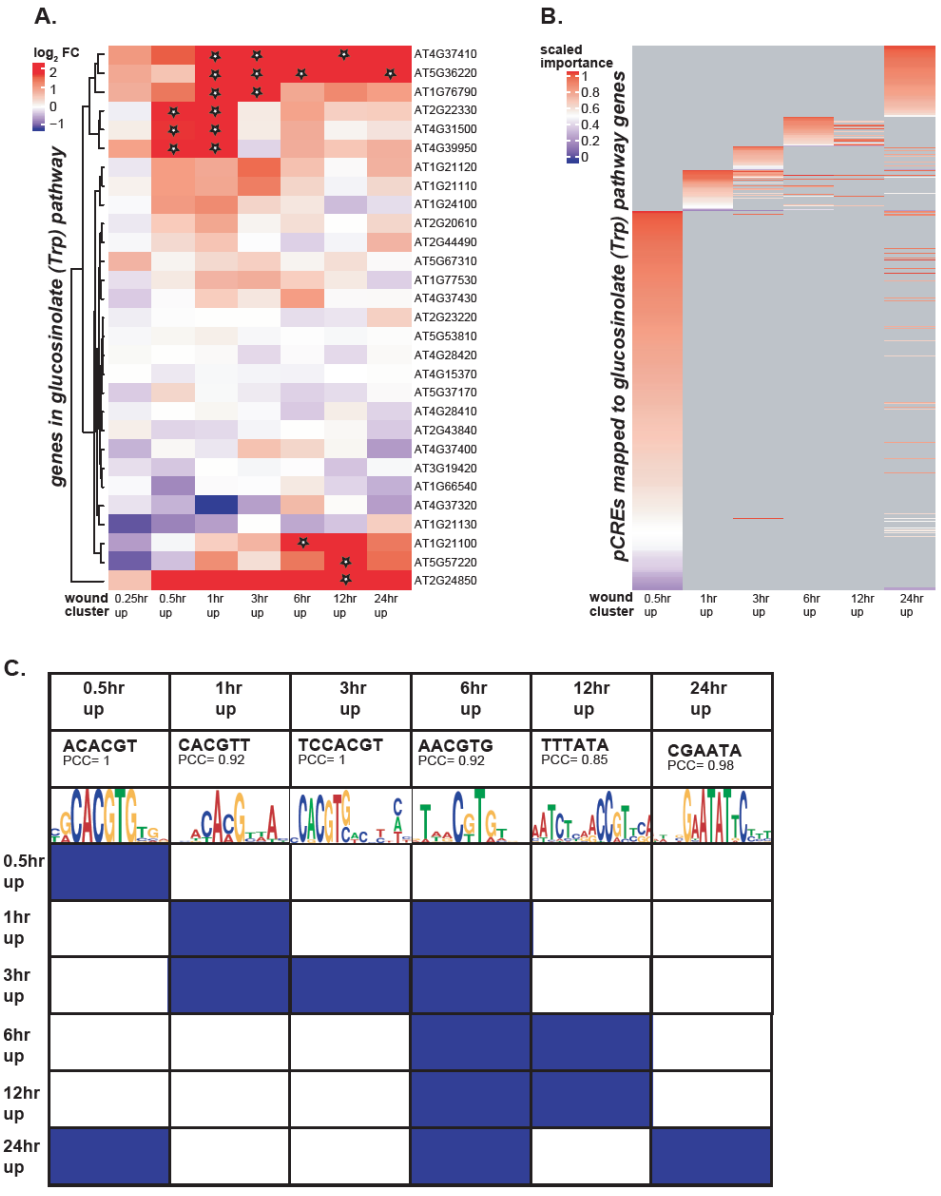


Figure 4.7. Co-expression and regulation of Glucosinolate from Tryptophan pathway genes.

A. Heatmap showing the log_2 fold change values of all genes in the Gluc-Trp pathway across the 7 wounding time points. Genes are clustered using hierarchal clustering. Genes are on the y-axis, wounding time points are on the x-axis, and log_2 fold change is represented as the color gradient from a value of 2 or greater (red) to a value of -1 or less (blue). Stars indicate gene is significantly up-regulated at a given time point. B. Scaled importance score of pCREs mapped to Gluc-Trp genes which are up-regulated at a given wounding time point. Importance is scaled from 0 to 1, where 1 is most important and 0 is least important. Each row is a pCRE and each column is the wounding time point. C. The most important pCRE for Gluc-Trp pathway genes at a given time point. First row is the pCRE and correlation to a known TF binding site. The second

Figure 4.7 (cont'd)
row is the motif logo for the known TF binding site. The rest of the rows show whether that particular pCRE overlaps with Gluc-Trp genes at other time points.

Conclusion

The aim of this study was to better understand the temporal differences in transcriptional response to wounding stress in *A. thaliana.* We accomplished this by integrating multiple levels of regulatory information (e.g. sequence based and epigenetic features) into machine learning models of the regulatory code that could be used to predict if a gene was up- or down-regulated at a specific timepoint after wounding. We demonstrated that wounding response is regulated by a diverse set regulatory elements that are likely bound by TFs from a wide range of TF families. We identify 4,255 pCREs derived from wounding co-expression clusters up-regulated at different timepoints, with 3,493 (82%) having significant sequence similarity (PCC > 0.8) to known TF binding sites. These pCREs were more predictive of differential expression at each wounding time point than models based on known TF binding sites (derived from the literature and the DAP-seq database) and information about open chromatin sites. From our machine learning models, we were also able to quantify the relative importance of each pCRE included in the model for each time point. While some pCREs were important across multiple timepoints, we generally found that pCREs were either important for early or late time points after wounding. By modeling JA-induced and non-JA-induced transcriptional responses separately, we were able to identify 2,569 pCREs important for predicting genes up-regulated in response to wounding but not in response to JA treatment. Of these, 2,371 (92%) had significant sequence similarity (PCC > 0.8) to known TF binding sites. Finally, by focusing on genes in the Gluc-Trp pathway, we were able to identify pCREs important for predicting genes in this wound responsive specialized metabolite pathway.

While our models perform notably better than random expectation, there is room for improvement. One possible reason we could not predict differential expression perfectly is that

we limited our study to focus on CRE sites in the promoter region (+1kb upstream of the transcription start site). However, CREs can be located in other regions, including in the downstream untranslated regions of the gene, in introns, or coding regions (Rose et al., 2008), which could be evaluated in future studies. Another limitation is that genes up- or down-regulated at a particular time point might not all be regulated the same way. This is especially likely for large time point gene groups, like the cluster of up-regulated genes one hour after wounding, which contains 760 genes. If we could further break down this group, perhaps based on the gene's response to other stresses, we may be able to model more specific responses at one hour, which could improve the overall performance. Finally, data regarding DAP-seq and DH sites did not come from wounded plants, and therefore are not capturing any changes that may occur to chromatin state or TF-binding sites after wounding.

Many of the important pCREs found in this study have not been shown to be associated with wounding. This is especially true for pCRE found at later timepoints that have been less well studied. However, new technologies, such as CRISPR-cas9, make it possible to generate precise edits to the DNA allowing for the role of these pCREs in temporal wounding response to be tested experimentally. To that end, our study provides a set of important putative targets that could be used to prioritize experiments to can confirm novel pCREs associated with different types of wounding response. Finally, more can be done to find regulatory elements associated with different pathways. Because the Gluc-Trp pathway was associated with wounding, we were able to find elements which may help regulate that pathway. However, other pathways may respond to different types of stress or may be active during certain stages of development or in particular tissues. Therefore, future studies should focus on determining regulatory elements for particular pathways by using an associated expression data set.

Methods

Expression datasets and analysis

Microarray data from three different AtGenExpress studies were downloaded from TAIR and CEL files were processed using Affy program in R. The studies included biotic stress (Wilson et al., 2012), abiotic stress (Kilian et al., 2007; Wilson et al., 2012), and hormone treatment (Goda et al., 2008), where wounding is part of the abiotic stress dataset. These studies grew plants under similar conditions, were treated 18 days after germination, and were all part of the AtGenExpress project. Each study had 8 different treatments of either different stresses or hormones, for a total of 24 data sets. Samples from each data set were collected after treatment at a range of time points, including 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, 4 hours, 6 hours, 12 hours, and 24 hours after treatment. Note that not all time points were used for each treatment. For each data set, controls were collected at the same time in order to control for circadian effects.

Differential expression was calculated using affy and limma packages in R (Gautier et al., 2004; Ritchie et al., 2015), and significantly differentially expressed genes were those that had an absolute $\log_2$ fold change ≥ 1 and adjusted p-value < 0.05. Up-regulated genes were those genes which were differentially expressed but with a $\log_2$ fold change ≥ 1, while down-regulated genes were those genes which were differentially expressed with a $\log_2$ fold change ≤ -1. For each expression dataset, Pearson's Correlation Coefficient (PCC) was calculated between each treatment.

Gene clusters

Wounding time point clusters were determined by differential expression at each time point of wounding stress (0.25, 0.5, 1, 3, 6, 12, and 24 hours after wounding). For example, genes which were up-regulated at the time point of 1 hour after wounding were placed in cluster

1 while genes down-regulated at 1 hour after wounding were placed in cluster 2. This created a total of 14 wounding clusters. For wounding and JA clusters, genes were placed in a cluster based on whether they were differentially expressed in one or both treatments at the same time point. For example, a gene X up-regulated in both 1 hour after wounding and 1 hour after JA treatment would be placed in cluster 1, while gene Y up-regulated in 1 hour after wounding but not changed under 1 hour after JA treatment would be placed in cluster 2. Finally, for a gene Z up-regulated in 1 hour after wounding but down-regulated in 1 hour after JA treatment would be placed in cluster 3. Therefore, at each time point which is in both wounding and JA treatment datasets (0.5, 1, and 3 hours) each up- or down-regulated after wounding cluster was divided into 3 separate clusters, for a total of 18 clusters. Three of these potential clusters actually contained no genes and were subsequently omitted (up-regulated after wounding but down-regulated after JA treatment at 0.5 hours, 1 hour, and 3 hours). A non-differentially expressed cluster was determined by genes which were not differentially expressed across all stresses and timepoints as well as all hormone treatments. For all gene clusters and overlap of clusters, see **Dataset S13**.

Known cis-regulatory elements literature search

Known regulatory elements were curated from a literature search. They included elements shown to be responsive to JA, wounding, or insect stress. The studies for this search can be found in **Dataset S14**. Both experimental and computational data was included.

Putative Cis-regulatory finding

Promoter regions of each gene (identified as 1-kb upstream of the transcription start site) were downloaded from TAIR for *A. thaliana*. Using homemade python scripts (https://github.com/ShiuLab/MotifDiscovery) were used to identify all combinations of 6-mers present in gene promoters. The Fisher's Exact Test (FET) was then used to determine

overrepresented putative cis-regulatory elements (pCREs) in the promoter region (defined as 1000 bp upstream of gene start site) by comparing a given wounding up- or down-regulated cluster to the non-differentially expressed cluster. A range of p-value cutoffs (adjusted $P < 0.01$, $P<0.01$, adjusted $P < 0.05$, and $P < 0.05$) was used, however for later machine-learning models, the best results were with the non-adjusted $P < 0.01$. Using the Motif Discovery pipeline, kmers (oligomer sequences of length k) were searched for in the promoters of genes of interest. Starting with all possible 6-mers, sequences which were found to be significantly overrepresented in the clusters based on the p-values listed above, were kept. Another round of kmer finding then occurred where the significant 6mer was extended on either side, producing two 7mers, and these 7mers were again tested to see if they were significantly overrepresented in the given cluster, and if their p-value was lower than the parent 6mer. If this was true, the 7mer was kept and the 6mer discarded. If not, the 7mer was discarded and the 6mer was kept. This procedure of "growing" kmers continued until the longest kmer with a p-value lower than its predecessor was obtained. These pCREs were then used as features to predict expression in machine-learning models. TAMO/1.0 (Gordon et al., 2005)was also used to create tamo files for each motif, which was used later to correlate to known transcription factor binding sites.

Arabidopsis cistrome and epicistrome

Two datasets providing *in-vitro* transcription factor (TF) binding sites were used to correlate to pCREs. First, *A. thaliana* motifs (position weight matrices) determined from protein binding arrays (called TF binding motifs or TFBMs) (Weirauch et al., 2014) were downloaded from http://cisbp.ccbr.utoronto.ca. DNA affinity purification sequencing (DAP-seq) peaks (O'Malley et al., 2016) were downloaded from http://neomorph.salk.edu/PlantCistromeDB. The peaks were then mapped to *A. thaliana* genome using python scripts. If the peak overlapped with

the promoter of a gene of interest, the peak was considered present as a feature for that gene. To provide insight into chromatin structure, Dnase I hypersensitivity (DH) sites (Zhang et al., 2012) were obtained from the National Center for Biotechnology Information database under the ID number GSE34318 as bed files. Bed files were parsed using python scripts to obtain gff files, which were then mapped to the *A. thaliana* genome. If the peak overlapped with the promoter of a gene of interest, the peak was considered present as a feature for that gene.

Machine learning models

Prediction models were built for each wounding time point cluster as well as for wounding-JA cluster where enriched pCREs from the promoter analysis were used as features to predict expression patterns in each expression class (up- or down-regulated genes in each cluster). Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting, (GB) were the machine learning algorithms implemented for each cluster using Python package sci-kit learn (Pedregosa et al., 2011). Python scripts 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. Because the dataset was unbalanced (i.e. 6,855 null genes, 760 up-regulated genes under wounding at 1 hour), 100 balanced datasets were created from random draws of the null gene cluster to match with the number of genes in the differentially expressed cluster. 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 dataset used to build the model, the remaining 10% used for validation) for each of the 100 balanced datasets. Model performance was evaluated using F-measure, the harmonic mean of precision and recall, where precision is defined as the number of true positives divided by the sum of true and false positives, and recall

is defined as the number of true positives divided by the sum of true positives and false negatives. Thus, in a binary model, a perfect prediction has an F-measure of 1 and the random expectation is 0.5. The models also have an importance score for each input feature, which is determined by the decrease in impurity of a node in a decision tree, and then averaged across the trees in the forest. Thus, the higher the number, the more important the feature (Breiman, 2001; Louppe, 2014). Importance value was scaled by normalizing based on the minimum and maximum values (where the minimum importance value is subtracted from a given value I, then divided by the difference between the maximum and minimum importance value). Importance rank was taken by ranking taking the importance value and ranking from highest to lowest. Percentile rank is taken by taking the rank of the feature and dividing it by the total number of features.

For each cluster, models with only known (derived from literature) CREs were built (model set 1), then models with known CREs plus DAP-seq and DH site information were built (model set 2). Finally, models with DAP-seq, DH site and enriched pCRE information were built (model set 3). Additionally, for model set 3, five separate models for each wounding time point cluster were run to determine the average importance score for each feature. This was then used to rank the features (pCREs) from most important to least important based on the average of the importance rank for each feature from the five models. Before ranking, reverse compliment pCREs were removed, so that essentially the same pCRE was not ranked twice. To assess random expectation, gene clusters chosen randomly from the expression data sets were enriched for pCREs. These were then used to build machine learning models using the methods above. Random gene clusters were made for genes at n= 30, 50, 100, 150, 200, and 250 at 20 repetitions each. Model results are reported in **Dataset S15**.

Sequence similarity of pCREs to known TF binding sites

To compare pCREs to potential know TF binding sites, pairwise PCC (Pearson's correlation coefficient) distance between pCREs and TF binding sites (both DAP-seq and TFBMs) was generated using the TAMO program (Gordon et al., 2005). After calculating the PCC distance to all possible TF binding sites, the lowest distance (highest PCC) was determined for each pCRE as its best match. The best match was then used for visualization of the binding site logo.

Pathway enrichment and pCRE mapping

Pathway annotations were downloaded from the Plant Metabolic Network Database (https://www.plantcyc.org/). Enrichment tests were performed by using python scripts (https://github.com/ShiuLab/GO-term-enrichment) and the python fisher 0.1.9 package which implements the Fisher Exact test. In order to map pCREs back to the genes in the glucosinolate from tryptophan (Gluc-Trp) pathway, gff files were created which contained the coordinates of pCREs in the promotors of all *A. thaliana* genes. Genes which were part of the Gluc-Trp pathway which were expressed at a wounding time point were matched up with pCREs which mapped to them. Finally, the importance for pCREs which map to Gluc-Trp genes was determined for each wounding time point from the previous wounding models.

APPENDIX

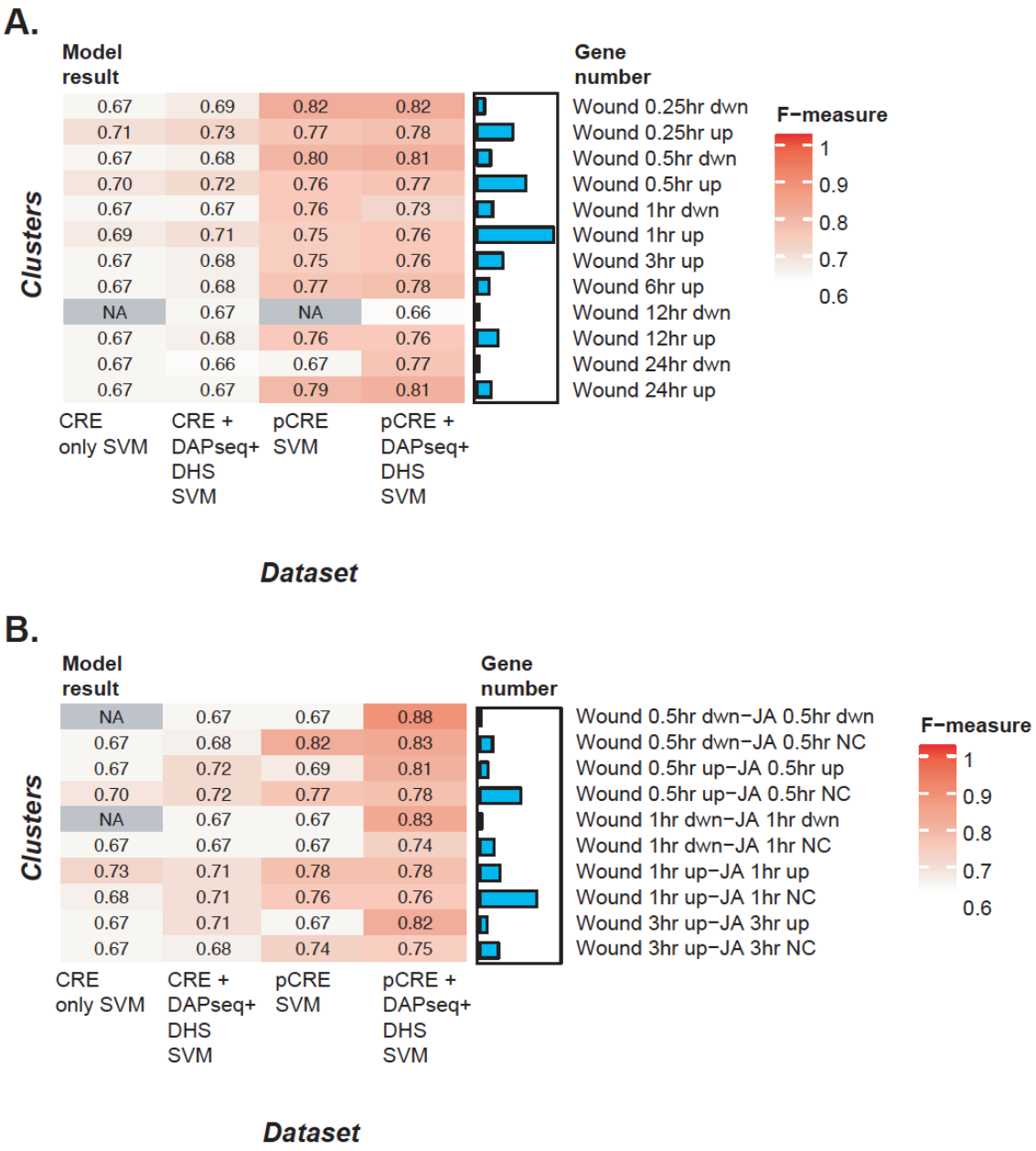


Figure S 4.1. Heatmap of the F-measure for all wounding SVM models.

Each row is a different cluster which was used to build a separate model. Each column represents the datasets used as features in the model and the algorithm used (SVM= Support Vector Machine). Known only refers to CREs found in the literature (**Dataset S14**). DAPseq and DHS refer to the DAP-seq and Dnase I hypersensitivity sites. FET enriched 6mer refers to the pCREs which were enriched for a specific cluster. The F-measure range is from 0.5 (white) to 1 (red), and gradient as well as actual F-measure is shown in each cell. The bar chart next to the heat map corresponds to each row/cluster and represents the number of genes in that cluster. A. Wounding time point models. B. wounding JA-induced and non JA-induced models.

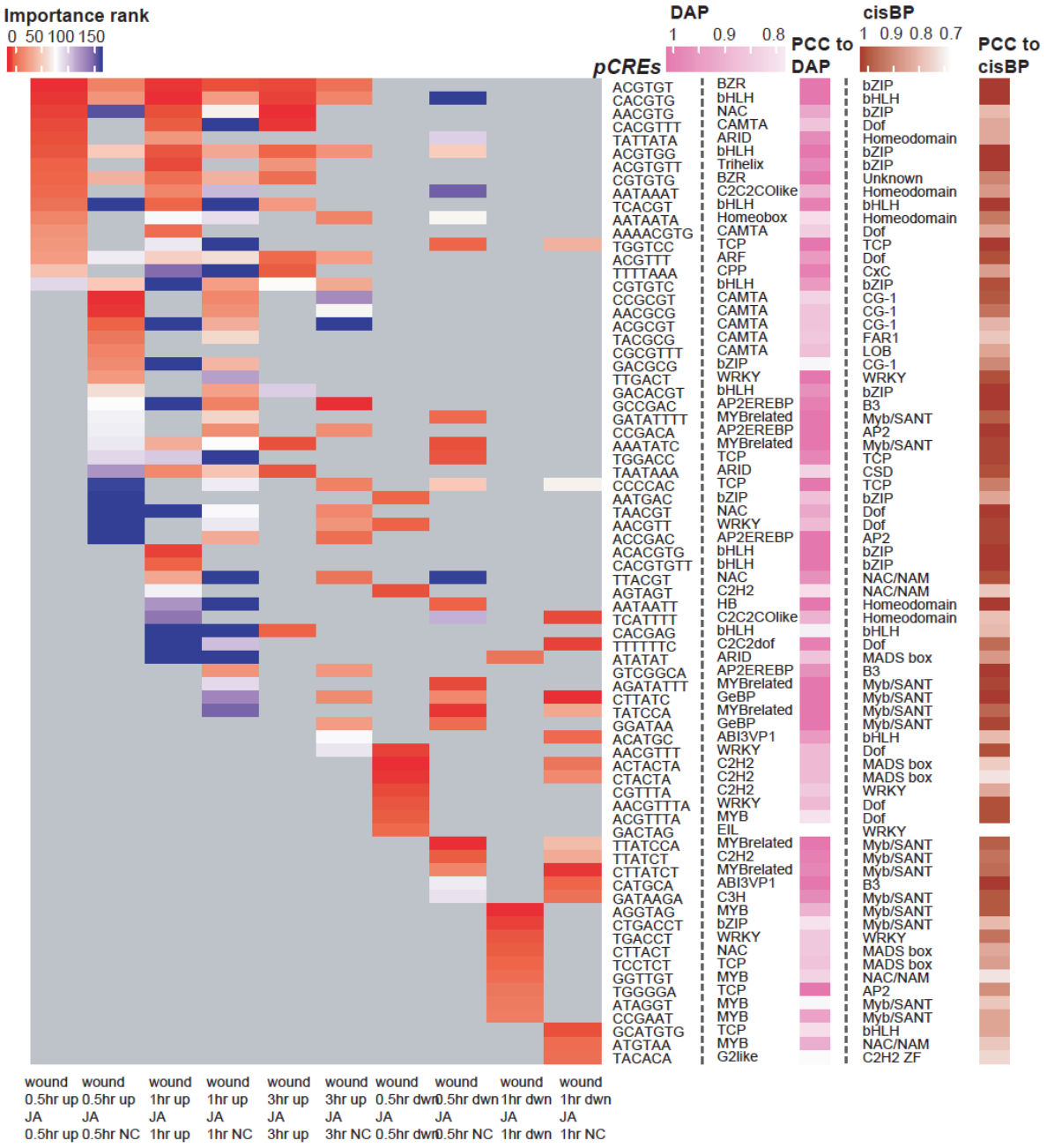


Figure S 4.2. Average importance rank for the top 10 pCREs for each wounding JA-induced and non JA-induced model and their TF family.

Wound time point models are the columns while pCREs are the rows. Average importance rank is the average rank across five duplicate models ran for the same time point. Highest rank (1) is red and ranks 150 or lower are blue. TF family association is based on the maximum PCC to known TF binding sites. PCC is shown for both DAP-seq and TFBM sites.

Supplemental Data

Dataset S12: Between sample PCC results

Dataset S13: Sample cluster overlap and genes in each cluster

Dataset S14: Known cis-regulatory elements derived from literature

Dataset S15: All machine learning model results

Dataset S16: Feature importance for models using only known elements or sites

Dataset S17: Summary table for the importance rank of each pCRE for each cluster and their correlation to DAP-seq or TFBM sites

Dataset S18: Overall feature importance score for wounding JA-induced and non JA-induced clusters

Dataset S19: All pCREs enriched for each wounding time point cluster and their p-values

Dataset S20: Pathway enrichment for each wounding time point cluster and their p-values

REFERENCES

REFERENCES

Asensi-Fabado M-A, Amtmann A, Perrella G (2017) Plant responses to abiotic stress: The chromatin context of transcriptional regulation. Biochim Biophys Acta BBA - Gene Regul Mech **1860**: 106–122

Berr A, Ménard R, Heitz T, Shen W-H (2012) Chromatin modification and remodelling: a regulatory landscape for the control of Arabidopsis defence responses upon pathogen attack: Chromatin regulation of plant defence. Cell Microbiol **14**: 829–839

Bostock RM, Pye MF, Roubtsova TV (2014) Predisposition in Plant Disease: Exploiting the Nexus in Abiotic and Biotic Stress Perception and Response. Annu Rev Phytopathol **52**: 517–549

Breiman L (2001) Random Forests. Mach Learn **45**: 5–32

Chung HS, Koo AJK, Gao X, Jayanty S, Thines B, Jones AD, Howe GA (2008) Regulation and Function of Arabidopsis *JASMONATE ZIM* -Domain Genes in Response to Wounding and Herbivory. Plant Physiol **146**: 952–964

Dombrecht B, Xue GP, Sprague SJ, Kirkegaard JA, Ross JJ, Reid JB, Fitt GP, Sewelam N, Schenk PM, Manners JM, et al (2007) MYC2 Differentially Modulates Diverse Jasmonate-Dependent Functions in *Arabidopsis*. Plant Cell **19**: 2225–2245

Fernández-Calvo P, Chini A, Fernández-Barbero G, Chico J-M, Gimenez-Ibanez S, Geerinck J, Eeckhout D, Schweizer F, Godoy M, Franco-Zorrilla JM, et al (2011) The *Arabidopsis* bHLH Transcription Factors MYC3 and MYC4 Are Targets of JAZ Repressors and Act Additively with MYC2 in the Activation of Jasmonate Responses. Plant Cell **23**: 701–715

Frerigmann H, Gigolashvili T (2014) Update on the role of R2R3-MYBs in the regulation of glucosinolates upon sulfur deficiency. Front Plant Sci. doi: 10.3389/fpls.2014.00626

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics **20**: 307–315

Glisovic T, Bachorik JL, Yong J, Dreyfuss G (2008) RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett **582**: 1977–1986

Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, Ogawa M, Yamauchi Y, Preston J, Aoki K, et al (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J **55**: 526–542

Gordon DB, Nekludova L, McCallum S, Fraenkel E (2005) TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics **21**: 3164–3165

Howe GA (2004) Jasmonates as Signals in the Wound Response. J Plant Growth Regul **23**: 223–237

Howe GA, Jander G (2008) Plant Immunity to Insect Herbivores. Annu Rev Plant Biol **59**: 41–66

Hutvagner G, Simard MJ (2008) Argonaute proteins: key players in RNA silencing. Nat Rev Mol Cell Biol **9**: 22–32

Ikeuchi M, Iwase A, Rymen B, Lambolez A, Kojima M, Takebayashi Y, Heyman J, Watanabe S, Seo M, De Veylder L, et al (2017) Wounding Triggers Callus Formation via Dynamic Hormonal and Transcriptional Changes. Plant Physiol **175**: 1158–1174

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (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**: 347–363

LeBrasseur ND, MacIntosh GC, Perez-Amador MA, Saitoh M, Green PJ (2002) Local and systemic wound-induction of RNase and nuclease activities in Arabidopsis: RNS1 as a marker for a JA-independent systemic signaling pathway. Plant J **29**: 393–403

León J, Rojo E, Sánchez- Serrano JJ (2001) Wound signalling in plants. J Exp Bot **52**: 1–9

León J, Rojo E, Titarenko E, Sánchez-Serrano JJ (1998) Jasmonic acid-dependent and -independent wound signal transduction pathways are differentially regulated by Ca 2+ /calmodulin in Arabidopsis thaliana. Mol Gen Genet MGG **258**: 412–419

Liu M-J, Sugimoto K, Uygun S, Panchy N, Campbell MS, Yandell M, Howe GA, Shiu S-H (2018) Regulatory Divergence in Wound-Responsive Gene Expression between Domesticated and Wild Tomato. Plant Cell **30**: 1445–1460

Lorenzo O, Piqueras R, Sánchez-Serrano JJ, Solano R (2003) ETHYLENE RESPONSE FACTOR1 Integrates Signals from Ethylene and Jasmonate Pathways in Plant Defense. Plant Cell **15**: 165–178

Louppe G (2014) Understanding Random Forests: From Theory to Practice. ArXiv14077502 Stat

Nakashima K, Ito Y, Yamaguchi-Shinozaki K (2009) Transcriptional Regulatory Networks in Response to Abiotic Stresses in Arabidopsis and Grasses: Figure 1. Plant Physiol **149**: 88–95

O'Donnell PJ, Calvert C, Atzorn R, Wasternack C, Leyser HMO, Bowles DJ (1996) Ethylene as a Signal Mediating the Wound Response of Tomato Plants. Science **274**: 1914–1917

O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **165**: 1280–1292

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res **12**: 2825–2830

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res **43**: e47–e47

Rojo E, León J, Sánchez-Serrano JJ (1999) Cross-talk between wound signalling pathways determines local versus systemic gene expression in Arabidopsis thaliana: Alternative wound signalling pathways in Arabidopsis. Plant J **20**: 135–142

Rose AB, Elfersi T, Parra G, Korf I (2008) Promoter-Proximal Introns in *Arabidopsis thaliana* Are Enriched in Dispersed Signals that Elevate Gene Expression. Plant Cell **20**: 543–551

Tierens KFM-J, Thomma BPHJ, Brouwer M, Schmidt J, Kistner K, Porzel A, Mauch-Mani B, Cammue BPA, Broekaert WF (2001) Study of the Role of Antimicrobial Glucosinolate-Derived Isothiocyanates in Resistance of Arabidopsis to Microbial Pathogens. Plant Physiol **125**: 1688–1699

Walley JW, Coughlan S, Hudson ME, Covington MF, Kaspi R, Banu G, Harmer SL, Dehesh K (2007) Mechanical Stress Induces Biotic and Abiotic Stress Responses via a Novel cis-Element. PLoS Genet **3**: 13

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al (2014) Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell **158**: 1431–1443

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**: 237

Yamamoto YY, Yoshioka Y, Hyakumachi M, Maruyama K, Yamaguchi-Shinozaki K, Tokizawa M, Koyama H (2011) Prediction of transcriptional regulatory elements for plant hormone responses based on microarray data. BMC Plant Biol **11**: 39

Yan X, Chen S (2007) Regulation of plant glucosinolate metabolism. Planta **226**: 1343–1352

Zhang W, Zhang T, Wu Y, Jiang J (2012) Genome-Wide Identification of Regulatory DNA Elements and Protein-Binding Footprints Using Signatures of Open Chromatin in *Arabidopsis*. Plant Cell **24**: 2719–2731

Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, Shiu S-H (2011) Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proc Natl Acad Sci **108**: 14992–14997