Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants[1][OPEN]

Pascal Schläpfer[2], Peifen Zhang[2], Chuan Wang[2,3], Taehyong Kim[4], Michael Banf, Lee Chae[3], Kate Dreher[5], Arvind K. Chavali, Ricardo Nilo-Poyanco[6], Thomas Bernard, Daniel Kahn, and Seung Y. Rhee*

Carnegie Institution for Science, Plant Biology Department, Stanford, California 94305 (P.S., P.Z., C.W., T.K., M.B., L.C., K.D., A.K.C., R.N.-P., S.Y.R.); and Laboratoire Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, Centre National de la Recherche Scientifique, Institut National de la Recherche Agronomique, Unité Mixte de Recherche 5558, 69622 Villeurbanne, France (T.B., D.K.)

ORCID IDs: 0000-0002-0828-8681 (P.S.); 0000-0001-5143-1714 (C.W.); 0000-0001-8183-2674 (L.C.); 0000-0003-4652-4398 (K.D.); 0000-0001-8093-0854 (D.K.); 0000-0002-7572-4762 (S.Y.R.).

Plant metabolism underpins many traits of ecological and agronomic importance. Plants produce numerous compounds to cope with their environments but the biosynthetic pathways for most of these compounds have not yet been elucidated. To engineer and improve metabolic traits, we need comprehensive and accurate knowledge of the organization and regulation of plant metabolism at the genome scale. Here, we present a computational pipeline to identify metabolic enzymes, pathways, and gene clusters from a sequenced genome. Using this pipeline, we generated metabolic pathway databases for 22 species and identified metabolic gene clusters from 18 species. This unified resource can be used to conduct a wide array of comparative studies of plant metabolism. Using the resource, we discovered a widespread occurrence of metabolic gene clusters in plants: 11,969 clusters from 18 species. The prevalence of metabolic gene clusters offers an intriguing possibility of an untapped source for uncovering new metabolite biosynthesis pathways. For example, more than 1,700 clusters contain enzymes that could generate a specialized metabolite scaffold (signature enzymes) and enzymes that modify the scaffold (tailoring enzymes). In four species with sufficient gene expression data, we identified 43 highly coexpressed clusters that contain signature and tailoring enzymes, of which eight were characterized previously to be functional pathways. Finally, we identified patterns of genome organization that implicate local gene duplication and, to a lesser extent, single gene transposition as having played roles in the evolution of plant metabolic gene clusters.

Plants are prodigious producers of complex small molecules called natural products, secondary metabolites, or specialized metabolites (Wink, 2010; Pichersky and Lewinsohn, 2011). These compounds are often restricted to specific taxon lineages, tissues, or are produced only under certain environmental conditions; hence, the term specialized metabolites. Among many other things, specialized metabolites are important in shaping the ecological interactions and adaptation of plants (Ehrlich and Raven, 1964; Bednarek and Osbourn, 2009). For humans, plant-derived specialized metabolites represent an important source of medicine. Among anticancer drugs, over half are plant derived or inspired by plant natural products (Schmidt et al., 2007). Although more than 25,000 plant species are cataloged as medicinal (Farnsworth, 1988), our understanding of the biochemical basis for these properties is extremely limited. Moreover, of the about 1 million metabolites estimated to be synthesized by plants (Afendi et al., 2012), we know the biosynthetic pathways of only about 0.1% (Caspi et al., 2012).

How plants generate and maintain such enormous chemical diversity is not known. However, several theories have been proposed, including coevolution (Ehrlich and Raven, 1964), sequential evolution (Jermy, 1984), random screening (Jones and Firn, 1991), duplication and subfunctionalization of transcriptional regulators (Grotewold, 2005), and catalytic promiscuity and the radiation of specialized metabolic enzymes (Weng et al., 2012). In bacteria and fungi, which also produce a large array of specialized metabolites, many biosynthetic pathways are organized into physically colocalized metabolic gene clusters (Cimermancic et al., 2014; Wisecaver and Rokas, 2015). Plant metabolic pathways are generally not known to occur in clusters, although there are now about 20 reported cases of specialized metabolic pathways that occur as gene clusters (Boycheva et al., 2014; Nützmann and Osbourn, 2014; Nützmann et al., 2016). Bioinformatic approaches have recently identified hundreds of metabolic gene clusters in Arabidopsis (*Arabidopsis thaliana*), rice (*Oryza sativa*), and sorghum (*Sorghum bicolor*; Chae et al., 2014) and collocated gene pairs between terpene synthases and oxidoreductases in several plant species (Boutanaev et al., 2015). The phenomenon of clustering may point to and assist in the discovery of unknown metabolic pathways and novel enzymes. Nevertheless, the prevalence and genesis of metabolic gene clustering in plants remain open questions. There are many other largely unanswered questions about plant metabolism. For example, what biological roles do these metabolites play in how plants adapt to their

environmental niches, and how do plants communicate with their beneficial partners and in the ongoing warfare against viruses, pathogens, and parasites? How did plants evolve to gain and maintain the metabolic repertoire? Finally, how can we use this knowledge to produce more and better food, industrial materials, and medicine?

In order to comprehensively study and tackle the fascinating questions of plant metabolism, we need a unified set of plant-specific, genome-wide metabolic reconstructions that integrate genome information with metabolic genes, enzymes, reactions, pathways, and compounds. Existing plant pathway databases have been built using different methods, sometimes not transparent to users, making it difficult to compare metabolism among plant species (Urbanczyk-Wochniak and Sumner, 2007; Bombarely et al., 2011; Dharmawardhana et al., 2013; Van Moerkercke et al., 2013; Jung et al., 2014). Creating a metabolic reconstruction consists of identifying enzyme sequences in a target plant genome, associating enzymes with reactions and pathways, and validating pathways using expert curation (Zhang et al., 2010). Each phase has considerable challenges. For example, accurately identifying enzymes remains a primary obstacle (Dale et al., 2010). Undetected enzymes result in incomplete pathways with unsupported reactions (Osterman and Overbeek, 2003). Falsely predicted pathways also are problematic, and, in the absence of clearly benchmarked statistics, manual curation is the only means of addressing the quality of reconstructed pathways. To establish a unified resource of plant metabolism, we developed a pipeline to build high-quality metabolic pathway databases for any sequenced plant genome. The pipeline consists of a machine learning-based enzyme annotation algorithm (Ensemble Enzyme Prediction Pipeline [E2P2]) that boosts prediction performance by leveraging a set of customized molecular function prediction programs, a gold-standard data set of protein sequences with expanded coverage of metabolic reaction types (Reference Protein Sequence Database [RPSD]), a pathway prediction algorithm (PathoLogic, Pathway Tools; Karp et al., 2011), a semiautomated validation software (Semi-Automated Validation and Integration [SAVI]) that greatly reduces manual validation time, and an algorithm to detect metabolic gene clusters (PlantClusterFinder). Using the pipeline, we predicted metabolic enzymes and pathways genome wide for 21 plant and one algal species and metabolic gene clusters from 17 plant and one algal species.

In this study, we present, to our knowledge for the first time, a detailed description of the algorithms, tools, and data sets we created and provide detailed analyses of metabolic gene clusters that exemplify how these resources can be used in discovering novel metabolic pathways. We have substantially improved both the quality and quantity of enzyme predictions described in our previous work (Chae et al., 2014) by enhancing E2P2 and expanding the reference protein sequences used by the algorithm. We also enhanced our gene cluster prediction method by allowing intervening nonenzymatic genes and considering sequence gaps that may interrupt a predicted gene cluster. We extended enzyme annotations to include 12 new plant genomes and extended metabolic gene cluster predictions from four to 18 species. We discovered a widespread occurrence of metabolic gene clusters in plants, with more than 1,700 (15%) containing enzymes that could potentially generate and modify scaffolds of specialized metabolites. We identified patterns of genome organization that implicate local gene duplication and single gene transposition as having played roles in the evolution of plant metabolic gene clusters.

RESULTS

An Integrative Metabolic Network Reconstruction Pipeline

We created a comprehensive computational pipeline that produces genome-scale metabolic reconstructions as well as metabolic gene cluster predictions consistently, efficiently, and accurately based on the protein

[1] This work was supported by the National Science Foundation (grant nos. IOS-1026003 and DBI-0640769), the Department of Energy (grant no. DE-SC0008769), the National Institutes of Health (grant no. 1U01GM110699-01A1), Becas Chile-CONICYT (postdoctoral fellowship to R.N.-P.), the Swiss National Foundation (postdoctoral fellowship to P.S.), and the Alexander Humboldt Foundation (postdoctoral fellowship to M.B.).

[2] These authors contributed equally to the article.

[3] Present address: Hampton Creek, San Francisco, CA 94103.

[4] Present address: Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104.

[5] Present address: International Maize and Wheat Improvement Center, El Batan 56130, Mexico.

[6] Present address: Center for Genome Regulation, Santiago 8370415, Chile.

* Address correspondence to srhee@carnegiescience.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Seung Y. Rhee (srhee@carnegiescience.edu).

S.Y.R. conceived the project; C.W. and L.C. developed the enzyme prediction pipeline with help from T.K., T.B., and D.K.; K.D., T.K., P.Z., and R.N.-P. developed the SAVI pipeline; P.Z., K.D., C.W., P.S., and R.N.-P. generated the pathway databases and validated pathways; R.N.-P., P.Z., and P.S. classified reactions and pathways into metabolic domains; P.S. and T.K. developed the gene cluster prediction pipeline with help from M.B., C.W., R.N.-P., and S.Y.R.; P.S. and S.Y.R. collected genome gap information; P.Z. and S.Y.R. compiled signature and tailoring enzymes; P.Z., P.S., C.W., M.B., A.K.C., and S.Y.R. analyzed gene clusters; P.S., C.W., and S.Y.R. performed gene duplication and transposition analysis; M.B. developed the coexpression-based gene cluster-ranking algorithm with help from P.S., P.Z., and A.C.; M.B., P.Z., and A.K.C. analyzed gene expression data; P.Z., C.W., P.S., and S.Y.R. wrote the article with contributions from all authors; S.Y.R. oversaw the project.

[OPEN] Articles can be viewed without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.16.01942

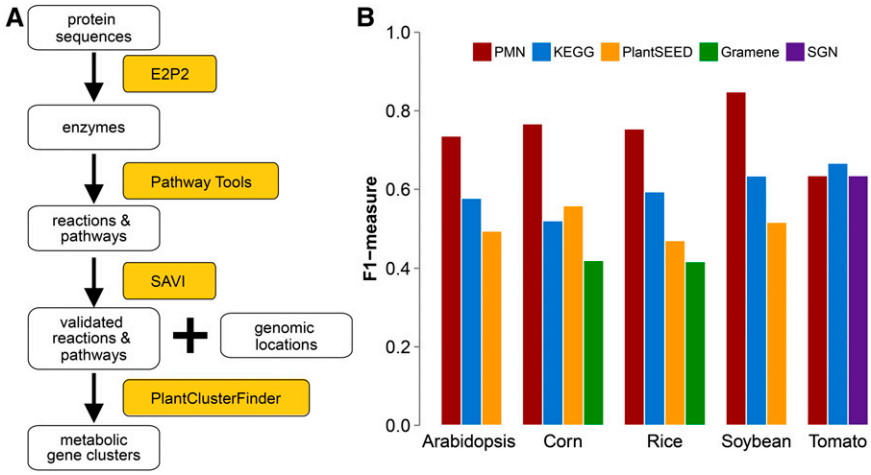

Figure 1. Prediction of metabolic complements from plant genomes. A, Overview of the pipeline to predict metabolic enzymes, pathways, and gene clusters. Protein sequences from sequenced genomes are processed by E2P2 to identify putative enzymes. Pathway Tools (Karp et al., 2011) assigns reactions to enzymes and predicts pathways. SAVI refines the reaction and pathway predictions. PlantClusterFinder uses genomic location information with enzyme and reaction data to identify metabolic gene clusters. B, Quality comparison of our databases (Plant Metabolic Network [PMN]; red) and other databases, KEGG (blue; Kanehisa et al., 2014), PlantSEED (yellow; Seaver et al., 2014), Gramene (green; Monaco et al., 2014), and the Solanaceae Genomics Network (SGN; purple; Bombarely et al., 2011), using independent, experimentally verified enzyme-reaction associations for the reactions that exist in all the compared databases. F1 measure is the harmonic mean of precision and recall.

sequences of a genome of interest (Fig. 1A). The pipeline integrates an enzyme annotation process called E2P2 (Chae et al., 2014), pathway prediction using Pathway Tools software (Karp et al., 2011), and a pathway prediction validation process called SAVI to predict enzymes, reactions, and pathways of an organism of interest. We substantially improved E2P2 by doubling the enzymes in the gold-standard protein sequences (RPSD v3.1) from 25,562 enzyme and 91,267 nonenzyme sequences (Chae et al., 2014) to 50,184 enzyme and 91,855 nonenzyme sequences (Supplemental Fig. S1A). To extend the coverage of metabolic reactions, we enhanced E2P2 (E2P2 v3.0) to represent catalytic functions as Enzyme Function (EF) classes encompassing MetaCyc reaction identifiers (7,393 identifiers) in addition to Enzyme Commission (EC) numbers (4,509 identifiers). Overall, E2P2 v3.0 contains 11,902 EF classes, a more than 5-fold increase of EF classes used in our previous study (Supplemental Fig. S1A). To assess the performance of E2P2, we used the enzyme and nonenzyme sequences of RPSD and conducted a 5-fold cross-validation test (Mosteller and Tukey, 1968). The enzyme annotation by E2P2 performed well, with 78.2% precision and 69.3% recall, for an overall F1 measure of 73.5% (see "Materials and Methods"; Supplemental Fig. S1B), optimizing the tradeoff between precision and recall of individual methods (BLAST [Altschul et al., 1990] and PRIAM [Claudel-Renard et al., 2003]) and resulting in the highest F1 measure.

Our pipeline associates the predicted enzymes with reference reactions and pathways in MetaCyc (Caspi et al., 2012) using PathoLogic (Karp et al., 2011). PathoLogic attempts to increase prediction sensitivity at the cost of false positives (Karp et al., 2011). Manually validating predictions based on the literature can fix these errors but is time consuming and cannot scale with the rate of genome sequencing. To expedite the manual validation of pathways, we developed a pipeline called SAVI that incorporates knowledge gained from manual curation. SAVI relies on a library of manually curated pathways that are applied to all inferred pathways (Supplemental Fig. S2; see "Materials and Methods"), improving the information quality in three ways: (1) decreasing false-negative pathway prediction by identifying unpredicted pathways thought to exist in a given taxonomic lineage; (2) lowering the number of false positives by invoking decision rules when assessing known lineage-specific pathways; and (3) removing redundant pathways such as nonplant pathway variants and pathways already represented as part of a larger pathway. On average, SAVI removed ~20% of the predicted pathways for each species identified as false positives and added ~5% of the pathways in the final databases identified as false negatives (Supplemental Fig. S2).

A Unified Resource of Plant Metabolic Pathway Databases: The Plant Metabolic Network

We applied this computational pipeline to predict metabolic pathways for 21 plant and one algal species that span a broad phylogenetic range and include major crops as well as model organisms. These species are *Chlamydomonas reinhardtii*, *Physcomitrella patens*,

Selaginella moellendorffii, *Spirodela polyrhiza* (common duckweed), *Brachypodium distachyon*, *Triticum urartu*, *Aegilops tauschii*, *Hordeum vulgare* (barley), rice, *Panicum virgatum* (switchgrass), *Setaria italica*, sorghum, *Zea mays* (maize), Arabidopsis, *Brassica rapa* (Chinese cabbage), *Carica papaya*, *Vitis vinifera* (grape), *Glycine max* (soybean), *Manihot esculenta* (cassava), *Populus trichocarpa* (poplar), *Solanum lycopersicum* (tomato), and *Solanum tuberosum* (potato). We used the Core Eukaryotic Genes Mapping Approach (CEGMA; Parra et al., 2007) to evaluate the completeness of a genome annotation by determining the fraction of highly conserved core eukaryotic genes annotated in a given genome. All genomes in our study had complete CEGMA scores of at least 75% (Supplemental Table S1). Overall, we generated higher quality plant metabolic pathway databases compared with resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al., 2014), PlantSEED (Seaver et al., 2014), and Gramene (Monaco et al., 2014; Fig. 1B) and a broader coverage of enzymes, reactions, and pathways (Supplemental Fig. S3C; Supplemental Table S2).

In total, we predicted 152,009 enzymes from the 22 species (PMN release version 10), with switchgrass containing the most (15,295) and *C. reinhardtii* the fewest (3,235) enzymes. Arabidopsis and *C. reinhardtii* contained the highest (620) and lowest (374) number of pathways. In addition to the enzymes and pathways, the databases contain 2,597 to 3,635 reactions and 1,755 to 2,802 compounds for a given genome (Supplemental Fig. S3A). All the pathways, reactions, enzymes, and genes can be searched, browsed, and downloaded online (www.plantcyc.org).

All components of this pipeline are generic except for the SAVI input files, which are currently customized for plants and green algae. The pipeline can be extended to bacteria, fungi, and animals by customizing the SAVI pathway library files for different lineages (Supplemental Fig. S2). The pipeline reduces the time to generate a species pathway database from 3 to 4 weeks to 1 to 2 d. Since all databases were reconstructed using the same pipeline, it enables the comparison of metabolic networks across species. E2P2 v3.0 and SAVI v3.02 are freely available online (https://dpb.carnegiescience.edu/labs/rhee-lab/software).

Prediction of Metabolic Gene Clusters

Previously, we detected a greater than expected presence of clustered metabolic genes in four flowering plant species (Chae et al., 2014). To determine if this phenomenon is general for flowering plants and whether it is also found in lower plants, we extended the analysis to 18 species, including a green alga, two lower land plants, and 15 higher plants, using a new prediction software called PlantClusterFinder (Supplemental Fig. S4; Supplemental Tables S1 and S3). Of the 22 species for which we created the metabolic pathway databases, we excluded switchgrass, two wheat (*Triticum aestivum*) progenitor species, and barley because of the quality of the assembly (for details, see "Materials and Methods").

In this study, we define a metabolic gene cluster as a minimal contiguous stretch of the genome that includes (1) at least three enzyme-coding genes involved in small molecule metabolism (referred to as metabolic genes herein) that catalyze at least two reactions; (2) more than just a single group of tandemly duplicated genes; and (3) an enrichment of metabolic genes (within the top 5% of the distribution of theoretical clusters of the same size in the genome; Supplemental Fig. S4A; see "Materials and Methods"). Surprisingly, we found that about half of the metabolic genes were clustered in all species examined, with an average of 665 metabolic gene clusters per species and 11,969 clusters in 18 species (Fig. 2A; Supplemental Fig. S5; Supplemental Table S4). This represents ~20 clusters predicted per 1,000 genes on average (Supplemental Fig. S6A).

To assess the performance of our cluster-finding algorithm, we examined the published cases of plant metabolic pathways that are clustered (Supplemental Table S5). The predicted clusters contained genome-wide averages of three to five metabolic genes (Fig. 2A; Supplemental Fig. S7B) and five to nine non-metabolic genes (median = 7; Supplemental Fig. S7C) and had average physical sizes ranging from 28 to 109 kb (median = 52.5 kb; Fig. 2C), well within the range of the sizes of experimentally verified clustered pathways in plants (four to 18 genes and 33 to 284 kb; Supplemental Table S5). Thirteen published clustered pathways are from species whose genomes have been analyzed in our study. We identified all but one clustered pathway; we failed to predict the potato chaconine/solanine cluster (Itkin et al., 2013) because there is a large chromosomal assembly gap in this region. Of the recovered known clustered pathways, one predicted cluster was identical to the published cluster, one was missing known genes at both ends of the predicted cluster, and 10 predicted clusters included additional metabolic genes that have not yet been characterized (Supplemental Table S5). Three of the 31 newly predicted genes with available expression data coexpress with previously characterized genes in three clustered pathways, thalianol (Arabidopsis), arabidiol (Arabidopsis), and 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (maize) biosynthesis pathways, suggesting that the uncharacterized enzymes could be involved in these pathways (Supplemental Table S5).

Inferring High-Confidence Gene Clusters Using Coexpression

A functional gene cluster consists of metabolic genes that are not only collocated but also involved in the same pathway (Boycheva et al., 2014; Nützmann and Osbourn, 2014). Coexpression is a strong indication of genes functioning in the same biological process (Wei et al., 2006). To further assess the quality of our gene

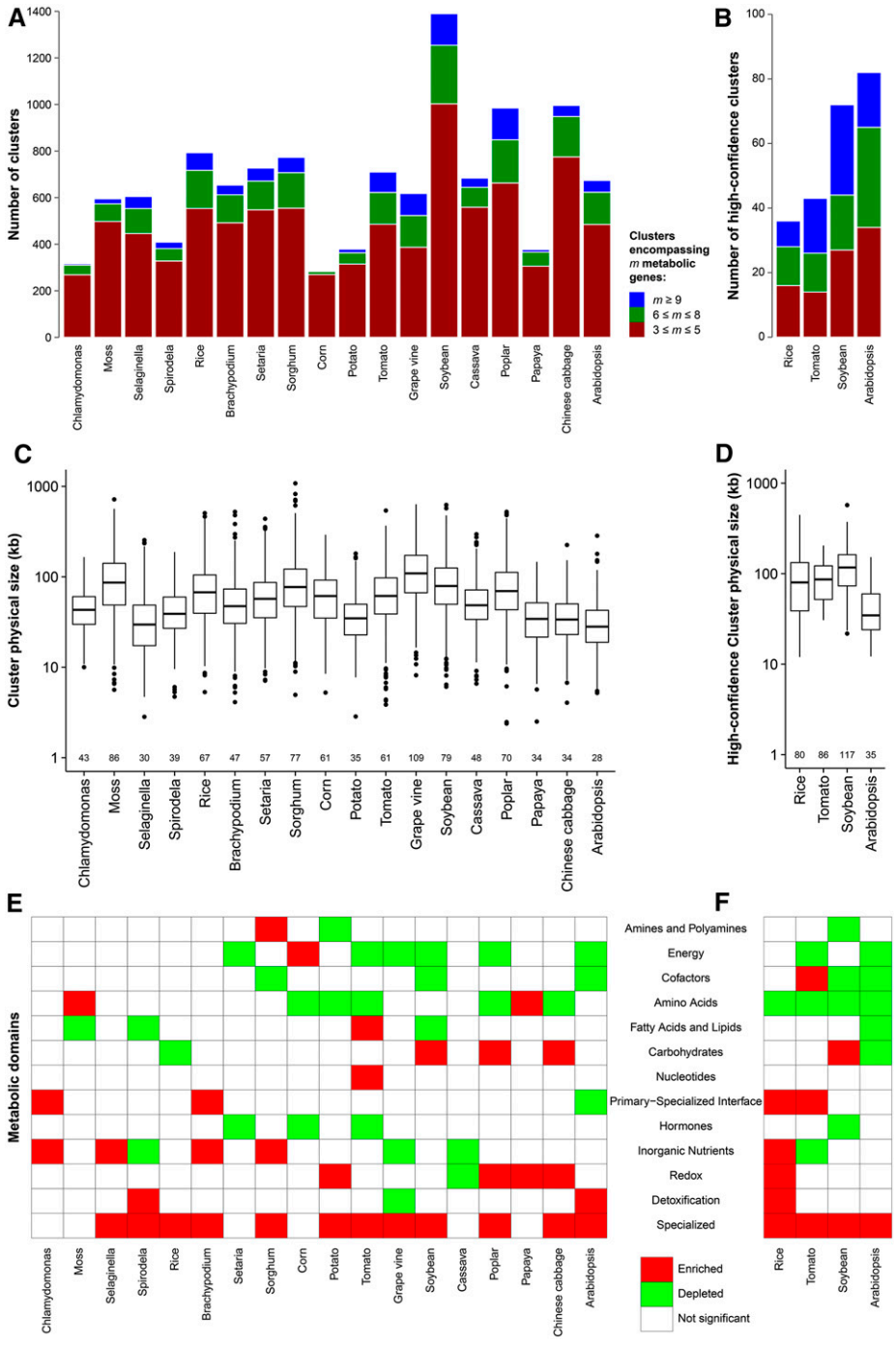

Figure 2. Prevalence of metabolic gene clusters in plants. A and B, Number of all predicted metabolic gene clusters of different sizes (number of clustered metabolic genes) across 17 plant and one algal species (A) or high-confidence metabolic gene clusters based on coexpression (B). C and D, Distribution of the physical size of all predicted metabolic gene clusters (C) or high-confidence metabolic gene clusters (D). Numbers at the bottom indicate median physical size. Outliers (dots) represent physical sizes beyond 1.5-interquartile ranges. E and F, Enrichment of metabolic domains in clustered metabolic genes (E) or high-confidence clustered metabolic genes (F). Significantly enriched (red) or depleted (green) metabolic domains are shown as $\log_2$ ratios ($P < 0.05$, hypergeometric test) for each species.

cluster predictions, we examined evidence of coexpression among metabolic genes within the same cluster with coexpression data in ATTED-II (Aoki et al., 2016; Supplemental Table S6). To ensure sufficient data coverage for the analysis, four species were chosen: Arabidopsis, soybean, rice, and tomato. We considered a gene pair to be coexpressed if its Pearson's correlation coefficient was above 99% of all the gene pairs in the transcriptome. Of the predicted clusters in the four species (3,620 clusters), 14% to 40% contained at least one pair of coexpressed metabolic genes (Supplemental Fig. S8). To identify high-confidence clusters (highly coexpressed), we ranked these clusters by rewarding coexpressed metabolic gene pairs per cluster while penalizing those pairs without coexpression support and normalizing the rank of each cluster with the total number of metabolic gene pairs in the cluster (see "Materials and Methods"). From the distribution of the calculated ranks, we defined high-confidence clusters as those whose binomial likelihood of cluster-level

coexpression by chance was less than 1%. A total of 233 clusters from the four species were considered as high-confidence clusters (Fig. 2B; Supplemental Fig. S8; Supplemental Table S7). All eight known clustered pathways from these four species were recovered within this set of high-confidence clusters. The high-confidence clusters had a higher proportion of medium and large clusters (more than five metabolic genes in a cluster; Fig. 2B) compared with all predicted clusters (Fig. 2A). However, the physical sizes of the high-confidence clusters (Fig. 2D) were similar to those of all predicted clusters (Fig. 2C), indicating a higher density of metabolic genes in the high-confidence clusters.

Enrichment of Specialized Metabolism in Metabolic Gene Clusters

We next asked which metabolic functions the predicted gene clusters might carry out. We adopted the previously compiled classification system (Chae et al., 2014) to classify all metabolic reactions into 13 major metabolic domains and propagated their classification to the encoding metabolic genes (see "Materials and Methods"; Supplemental Table S8). On average, 19.6% of metabolic genes were associated with specialized metabolism per species, with the least proportion in *C. reinhardtii* (8.6%) and the most in potato (25.2%; Supplemental Table S8). To evaluate whether annotations to any metabolic domains were overrepresented in clustered metabolic genes relative to all metabolic genes, we used hypergeometric tests (Rivals et al., 2007) to perform enrichment analysis (see "Materials and Methods"). Enriched and depleted metabolic domains varied across the species (Fig. 2E). However, the metabolic domain of specialized metabolism was slightly but significantly enriched in clustered genes in the majority of species (Fig. 2E; average fold change, 1.14; $P < 3.29E-36$, hypergeometric test). The enrichment was more prominent in the high-confidence clusters (Fig. 2F; average fold change, 1.64; $P < 2.48E-37$, hypergeometric test).

Local (also called tandem [Freeling, 2009]) duplication was suggested previously to be a main driving force for enzyme expansion in plant specialized metabolism (Chae et al., 2014). This prompted us to examine whether local duplication (LD) of metabolic genes also could be a reason for the enrichment of specialized metabolic domains in gene clusters. Locally duplicated metabolic genes were significantly enriched in clusters in all 18 species (average fold change, 1.29; $P < 5.55E-10$; Supplemental Fig. S9D), whereas locally duplicated nonmetabolic genes were significantly depleted in the clusters in all species except papaya (average fold change, 0.58; $P < 0.01$, hypergeometric test; Supplemental Table S9; Supplemental Fig. S9D), indicating that the enrichment of the LD was specific to the metabolic genes in the clusters. When the locally duplicated genes were excluded from the analysis, the specialized metabolic domain was no longer enriched in clusters (Supplemental Fig. S9C), indicating that the locally duplicated genes were responsible for the enrichment of specialized metabolic domains in clusters.

Metabolic Gene Clusters with Hallmarks of Specialized Metabolic Pathways

A common feature (or hallmark) of specialized metabolic pathways is the presence of key enzymes (signature enzymes) that generate a specialized metabolite scaffold (or skeleton) and enzymes that modify the scaffold (tailoring enzymes) with various chemical groups to produce end products (Boycheva et al., 2014; Nützmann and Osbourn, 2014; Nützmann et al., 2016). Typical signature enzymes include terpene cyclases and polyketide synthases, whereas typical tailoring enzymes include oxidoreductases, methyltransferases, acyltransferases, and glycosyltransferases (Osbourn, 2010). In addition to the well-characterized signature enzymes, cytochrome P450s also can generate a scaffold for some pathways (Nützmann et al., 2016).

To identify clusters that contain both signature and tailoring enzymes, we compiled a list of signature enzymes for four categories of specialized metabolism in plants, terpenoids, phenylpropanoids (including flavonoids and stilbenes), alkaloids, and polyketides (excluding flavonoids and stilbenes), as well as a list of the most commonly found tailoring enzymes in plant specialized metabolism (including enzymes introducing hydroxylation, glycosylation, methylation, and acylation), and subsequently cataloged all predicted clusters based on the presence of signature and tailoring enzymes (Supplemental Table S5; see "Materials and Methods"). We identified 664 clusters containing both signature and tailoring enzymes. Terpenoids, phenylpropanoids, and alkaloids were prominent, whereas polyketides were a minor class among the 664 clusters, consistent with what has been observed in plant specialized metabolism (Gunatilaka, 2008; Wink, 2010). An additional 1,063 clusters were identified as an atypical class of the specialized metabolic category, not carrying signature enzymes but containing cytochrome P450s or 2-oxoglutarate-dependent dioxygenases (2ODDs); both types of enzymes have been found at branching points of plant specialized metabolic pathways and, thus, both could be potential signature enzymes. In total, the 1,727 hallmark-containing clusters represent 15% of all predicted clusters (Fig. 3, A and C).

Among the 543 hallmark-containing clusters from Arabidopsis, soybean, rice, and tomato, 43 (8%) also were identified as high-confidence clusters based on coexpression (Supplemental Table S7). Of these 43 clusters, 18 have evidence of coexpression between a signature and a tailoring enzyme or a cytochrome P450/2ODD and a tailoring enzyme within the same cluster. All eight previously known clustered pathways were predicted as high-confidence clusters that contain signature and tailoring enzymes. All but one of these had coexpression between signature and tailoring

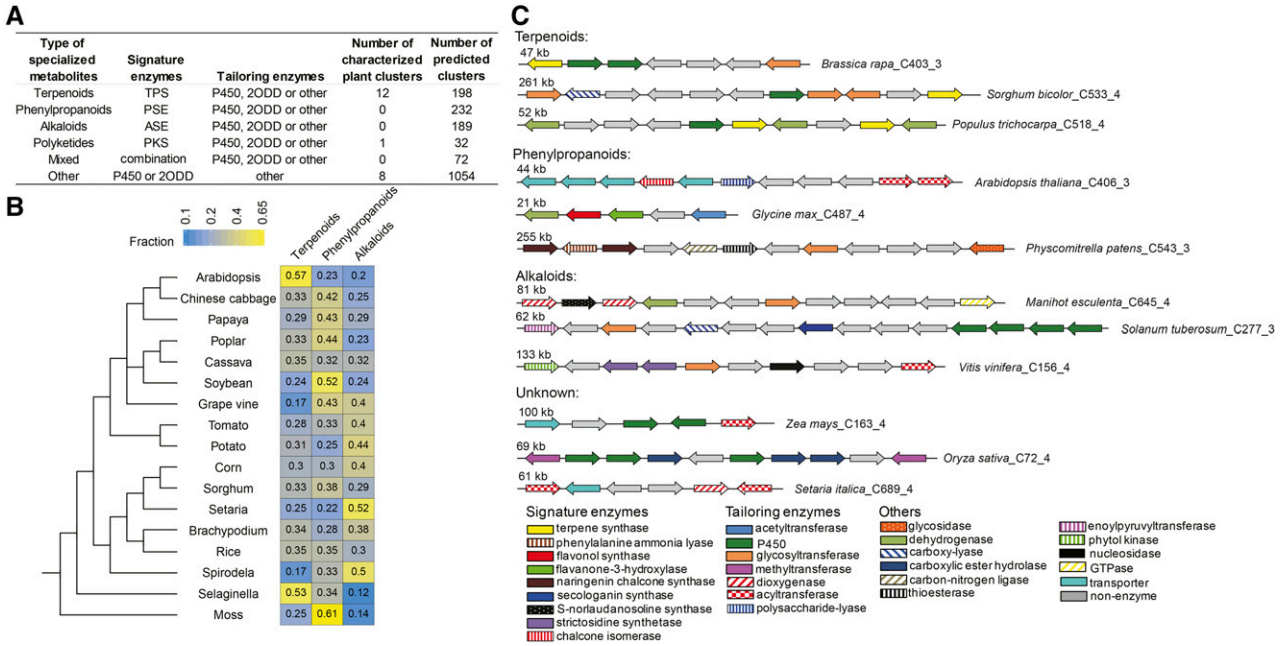

Figure 3. Patterns of metabolic gene clusters in specialized metabolism. A, Categories of specialized metabolic gene clusters classified by signature enzymes (TPS, terpene synthases; PSE, phenylpropanoid signature enzymes; ASE, alkaloid signature enzymes; PKS, polyketide synthases; combination, combination of any two types of signature enzymes; P450, cytochrome P450 enzyme. The category Other represents clusters that might harbor atypical signature enzymes such as cytochrome P450 or 2ODD. Glycosyltransferases, methyltransferases, or acyltransferases are termed other tailoring enzymes. B, Relative proportion of specialized metabolic categories found in metabolic gene clusters for each species. C, Examples of metabolic gene clusters in each specialized metabolic domain.

enzymes. Tomato cluster C495_4 (Matsuba et al., 2013) had coexpressed genes, but signature and tailoring enzymes were not coexpressed with each other (Supplemental Table S7). One example of the clusters with coexpression between signature and tailoring enzymes is a tomato cluster, C584_4, which spans 141 kb on chromosome 9. The cluster includes the signature enzyme naringenin chalcone synthase, three methyltransferases, one cytochrome P450, and 13 other genes (Fig. 4A). The three methyltransferases were highly coexpressed with the naringenin chalcone synthase, and all four enzymes also were coexpressed with an ankyrin repeat protein (nonenzyme; Fig. 4B). One of the methyltransferases, SOLYC09G091550, can methylate salicylic acid in vivo (Tieman et al., 2010). The presence of a naringenin chalcone synthase suggests that the cluster is involved in producing phenylpropanoids. Tomato indeed produces hydroxylated naringenin chalcone and methyl ethers of hydroxylated naringenin chalcone (Mintz-Oron et al., 2008). However, the genes that encode enzymes catalyzing these reactions have not been identified. Therefore, this tomato cluster provides a compelling model for a pathway that makes these compounds (Fig. 4C).

Partial Clustering of Metabolic Pathways

To determine how many of the metabolic pathways in our databases contained clustered metabolic genes, we looked for metabolic pathways that contain at least two reactions that were encoded by at least two genes in a cluster. On average, 14.9% of the pathways belong to specialized metabolism per genome (Supplemental Table S8). Specialized metabolic pathways were more than 2-fold likely to contain clustered genes than nonspecialized metabolic pathways across the genomes (fold change, 2.41; $P = 0.0012$, Kolmogorov-Smirnov test; Fig. 5A). The proportion of specialized metabolic pathways containing clustered genes varied among species (Fig. 5B). For example, *C. reinhardtii* did not have any specialized metabolic pathways that contained clustered genes, whereas in rice, ~24.1% of specialized metabolic pathways contained clustered genes (Fig. 5B).

Evidence of Genetic Mechanisms That Could Have Contributed to the Formation of Plant Metabolic Gene Clusters

Considering the widespread occurrence of metabolic gene clusters in plants, we wondered how these clusters might have formed. Multiple models were proposed for the origin of prokaryotic clusters (operons), including gene duplication, gene recruitment, and horizontal gene transfer (Fondi et al., 2009). In fungi, both gene duplication and horizontal transfer are more pronounced in clustered genes than in their nonclustered

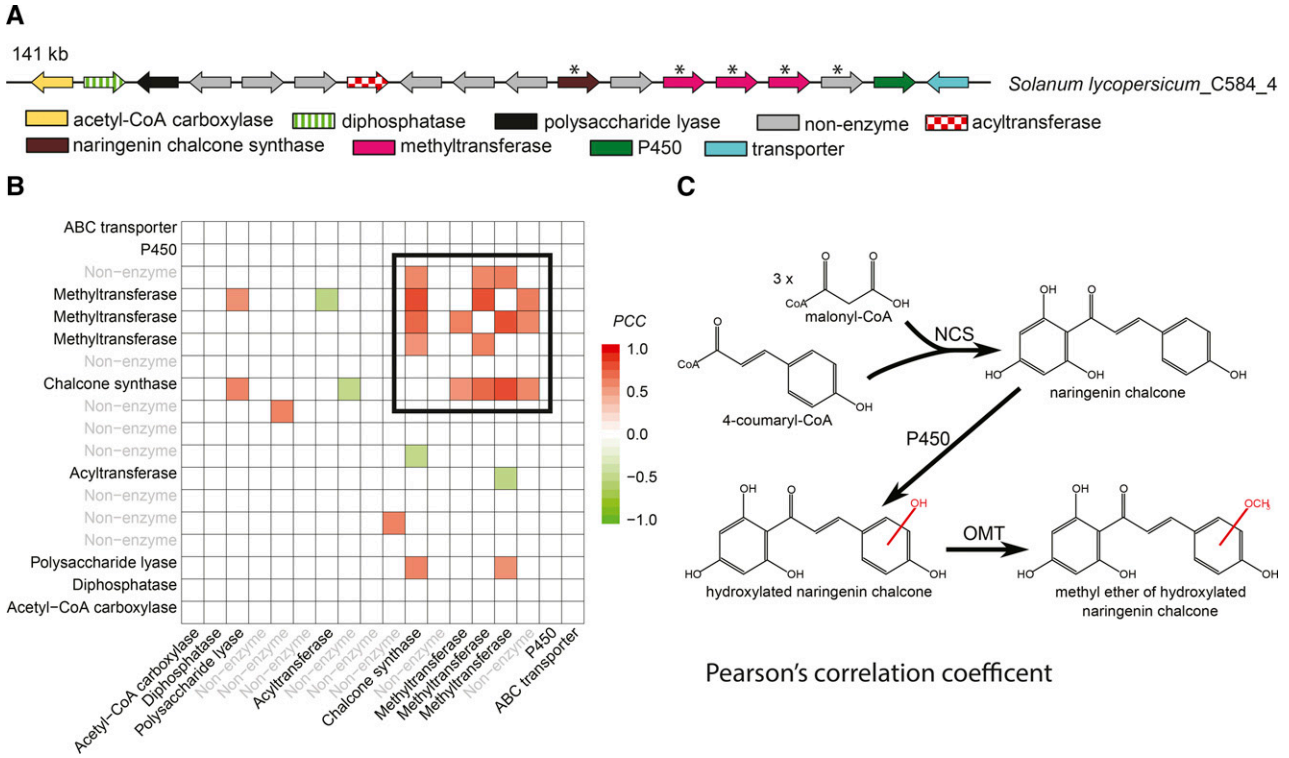

Figure 4. Hypothetical pathway model for the predicted tomato C584_4 cluster. A, Gene composition of the tomato cluster C584_4 with highly coexpressed genes marked with asterisks. B, Heat map of coexpression among clustered genes, with highly coexpressed genes framed in black. C, Hypothetical pathway catalyzed by metabolic genes of the cluster. NCS, Naringenin chalcone synthase; OMT, *O*-methyltransferase.

counterparts (Wisecaver et al., 2014). Furthermore, fungal clusters are often located near telomeres and transposable elements (Wisecaver and Rokas, 2015). In plants, it is well established that plant-specialized metabolic enzymes evolved from plant primary metabolic enzymes (Weng et al., 2012; Moghe and Last, 2015). Therefore, plant clusters might have evolved from mechanisms other than horizontal gene transfer. Supporting this idea is the significant enrichment of LD in clustered metabolic genes compared with all metabolic genes (average fold change, 1.29; $P < 5.55E-10$; Supplemental Fig. S9D; Supplemental Table S9). Thus, functional divergence via LD could have played a role in plant metabolic gene cluster formation.

We also examined whether single gene transpositions contributed to metabolic gene cluster formation. Over 4,500 Arabidopsis genes (~21% of the genome) were reported to have transposed after the divergence of Arabidopsis from poplar about 100 million years ago (Woodhouse et al., 2011). We found that single gene transpositions were not enriched in clustered metabolic genes compared with all metabolic genes (fold change, 1.04; $P = 0.08$, hypergeometric test). However, clustered LD metabolic genes were slightly but significantly enriched with single gene transpositions compared with all LD metabolic genes (fold change, 1.07; $P = 0.01$, hypergeometric test; Supplemental Table S9). Thus, single gene transposition also may have contributed, either directly or via local gene duplication, to the formation of clusters.

Finally, we found a significant enrichment of transposable elements associated with locally duplicated genes in Arabidopsis (fold change, 1.19; $P = 8.95E-35$, hypergeometric test; Supplemental Table S9). However, the locally duplicated metabolic genes in clusters were equally likely to associate with transposons as the locally duplicated nonclustered metabolic genes (fold change, 1; $P = 0.28$, hypergeometric test; Supplemental Table S9). Interestingly, we did not observe any bias of cluster distribution toward subtelomeric regions (Supplemental Fig. S9E). These observations indicate that plant metabolic genes might have formed using mechanisms that differ from their counterparts in bacterial and fungal genomes.

DISCUSSION

In this study, we present a comprehensive computational pipeline to create a unified resource of metabolism information for plants. This resource can be used in a variety of ways to compare and analyze plant metabolic complements. We used the resource to discover 11,969 metabolic gene clusters in 18 species. Of

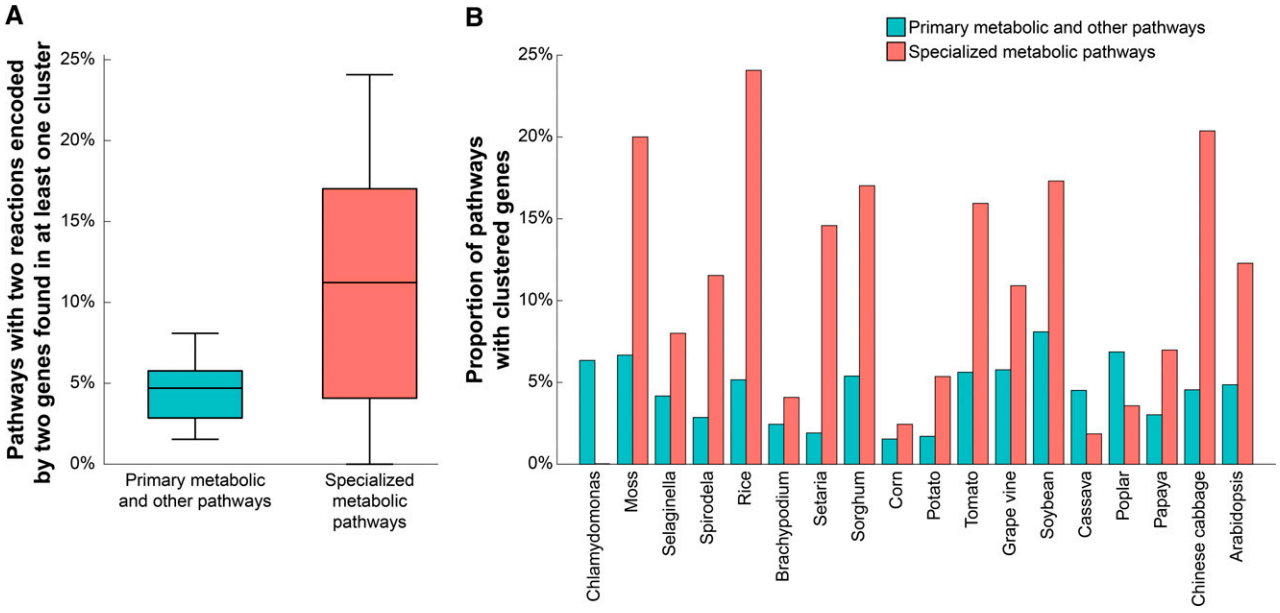

Figure 5. Metabolic pathways partially encoded by clustered genes. The percentage of metabolic pathways with at least two reactions encoded by different genes in a metabolic gene cluster is shown as a distribution over all organisms (A) and for each organism independently (B). Specialized metabolic pathways are shown in red, and all other pathways are shown in green.

these, 233 clusters in four species were determined to be of high confidence by coexpression analysis. This represents a dramatic increase over the ~20 metabolic gene clusters that are known in plants (Boycheva et al., 2014; Nützmann and Osbourn, 2014; Nützmann et al., 2016), revealing a potential source of new metabolic pathways in plant genomes.

Quality Assessment of Metabolic Enzyme Prediction

The updated E2P2 pipeline can predict 11,902 different enzymatic functions, covering 5-fold more EF classes than in our previous study (Chae et al., 2014). E2P2's performance is superior to other methods (Supplemental Fig. S1B), and when combined with PathoLogic and SAVI (Supplemental Fig. S2), the generated pathway databases are more accurate than other published databases (Fig. 1B; Supplemental Table S2).

Despite the performance of E2P2, sequence similarity-based function prediction may not distinguish specific enzyme functions for closely related members of a family. To determine the extent of uncertainty of enzyme function prediction of E2P2, we searched for EF classes whose protein sequence similarity shared among proteins within an EF class is similar to or lower than the sequence homology shared with proteins in other EF classes. Overall, ~29.2% of EF classes formed 10,091 pairs that may be potentially misidentified by each other (Supplemental Table S10). The majority (76.3%) of these pairs shared the first three EC numbers. Similarly, when we examined experimentally characterized Arabidopsis enzymes that were not used in training E2P2 and compared them with their predictions by E2P2, we found 22% false-positive predictions. Again, the majority of the false positives (63.8%) were misannotations caused by high sequence similarities between two EF classes (Supplemental Table S10). As expected, E2P2 achieved a higher performance at the three-part EC level, with a 95.8% precision, a 91.5% recall, and a 93.6% F1 measure (see "Materials and Methods"; Supplemental Fig. S10), than the EF class level (four-part EC level or distinct MetaCyc reaction identifier level). Therefore, future improvements on enzyme function annotation are likely to come from methods that can distinguish functions between closely related enzyme sequences. Approaches to boost accuracy at the EF class level include integrating phylogenetics, conserved residues, and other non-sequence homology features.

Prediction of Plant Metabolic Gene Clusters

To predict metabolic gene clusters de novo, we added a component to the metabolic network prediction pipeline called PlantClusterFinder. Currently, PlantClusterFinder uses four types of information: the relative physical location of genes, the presence or absence of metabolic gene prediction, LD information, and sequencing gap information. The initial prediction was further validated and ranked using coexpression. Future improvements could integrate other types of data, such as the physical chromosomal span, the prevalence of certain biochemical reactions within a cluster, protein-protein interactions, epigenetic modification marks, and evolutionary patterns.

Besides PlantClusterFinder, there are several existing tools that predict metabolic gene clusters from microbes, such as ClusterFinder (Cimermancic et al., 2014), SMURF (Khaldi et al., 2010), and antiSMASH (Weber et al., 2015). However, they are trained on experimentally characterized gene clusters from bacteria and fungi, which predominantly make polyketides, nonribosomal peptides, and sugar derivatives (Hoffmeister and Keller, 2007; Cimermancic et al., 2014). On the other hand, plants predominantly make terpenoids, phenylpropanoids, and alkaloids (Gunatilaka, 2008; Wink, 2010). Therefore, we developed software to predict metabolic gene clusters de novo.

To identify the clusters that are more likely to function as a pathway, we used coexpression information. A total of 233 clusters from four species with sufficient expression data were considered as high-confidence clusters (Fig. 2B; Supplemental Fig. S8; Supplemental Table S7) and included all eight known clustered pathways from these four species. The coexpression data sets provided by ATTED-II (Aoki et al., 2016; Supplemental Table S6) included a diverse set of treatments, developmental stages, and tissue types. This diversity generates high variances in expression values per gene, which leads to a higher statistical significance of Pearson's correlation values between coexpressed gene pairs. However, the incorporation of diverse samples also may cause condition-specific correlations to be masked. This may explain the large difference in the total number of predicted gene clusters and the number of high-confidence clusters.

Comparison of Metabolic Gene Cluster Prediction within and across Kingdoms

Using PlantClusterFinder to find metabolic gene clusters in plants de novo, we discovered ~600 clusters per species on average, and all but one known cluster were recovered (Fig. 2A; Supplemental Fig. S5; Supplemental Table S4). Five species showed substantial deviation from the average number of gene clusters per species: *C. reinhardtii*, maize, potato, and papaya had fewer predicted clusters, whereas soybean had more predicted clusters than average (Fig. 2A; Supplemental Fig. S7A). The genomes with fewer predicted clusters had more sequencing gaps. For example, the maize genome contained many sequencing gaps that likely prevented the prediction of additional metabolic gene clusters. Both potato and papaya also had more sequencing gaps, with an average of one out of 12 intergenic regions containing sequencing gaps. All other species had an average of one out of 42 intergenic regions containing sequencing gaps. Additionally, the potato and papaya genomes had smaller proportions of metabolic genes (12.2% and 17.3%, respectively) compared with the average of all other organisms (20.1%). The low number of clusters in *C. reinhardtii* and the high number in soybean can be explained by the size of their genomes, as the frequency of metabolic gene clusters generally increases linearly with the number of genes in the genome (Supplemental Fig. S6). A linear relationship between metabolic gene clusters and protein-encoding genes in the genome also is found in bacteria (Supplemental Fig. S6, A and B), but the relationship seems more complex for fungi (Supplemental Fig. S6, C and D).

We found ~20 clusters per 1,000 genes on average (Supplemental Fig. S6A), which is much higher than the frequency found in bacterial and fungal genomes (Khaldi et al., 2010; Cimermancic et al., 2014; Wisecaver et al., 2014). However, when only the coexpressed, high-confidence metabolic gene clusters are considered, the frequency of plant clusters is ~1.6 per 1,000 genes, which is similar to the frequency found in bacteria (~3.3 per 1,000 genes) and fungi (~1.3–2.6 per 1,000 genes; Supplemental Fig. S6, B–D; Khaldi et al., 2010; Cimermancic et al., 2014; Wisecaver et al., 2014).

Formation and Evolution of Plant Metabolic Gene Clusters

It is unclear when and how metabolic gene clusters formed in plants. To assess the degree of cluster conservation, we examined nine experimentally characterized metabolic gene clusters that were reported to be conserved in more than one species (Qi et al., 2004; Field et al., 2011; Takos et al., 2011; Dutartre et al., 2012; Itkin et al., 2013; Miyamoto et al., 2016). To characterize the degree of conservation of these clusters, we expanded the analysis to include additional species. Most clusters have evolved recently, and conservation is limited to closely related species (largely at the genus level; Supplemental Table S11). For example, Arabidopsis's thalianol cluster was detected in *Arabidopsis lyrata* (Field et al., 2011). However, we found it to be absent in *B. rapa* (Supplemental Table S11), which suggests that the cluster formed after the split of the *Brassica* and *Arabidopsis* genera 14.5 to 20.4 million years ago (Yang et al., 1999; Supplemental Table S11). Of the experimentally characterized clusters we examined, only the maize 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one cluster is conserved at the family level. It is partially conserved in the grass family, which evolved about 50 million years ago (Paterson et al., 2004). Orthologs of the core enzymes Bx1 to Bx5 were found in wheat and rye (*Secale cereale*), albeit in two separate chromosome regions (Dutartre et al., 2012). Overall, our analysis combined with previously reported information for these characterized plant gene clusters indicate that they evolved in a lineage-specific manner.

Consistent with the lineage specificity of the known metabolic gene clusters, we found that the metabolic composition of predicted metabolic gene clusters in the 18 species was species specific and phylogeny independent. Overrepresented metabolic domains (Fig. 2E) and the proportion of specialized metabolic subdomains (Fig. 3B) in the clusters varied across species and did not recapitulate phylogeny. For example,

within the Brassicales, terpenoids predominate in Arabidopsis clusters, whereas phenylpropanoids predominate in papaya, consistent with the chemical compositions of specialized metabolites found in Arabidopsis (D'Auria and Gershenzon, 2005) and papaya (Gogna et al., 2015). In the moss *P. patens*, phenylpropanoid clusters represent the predominant type (Fig. 3B). This coincides with the evolutionary advantage for early land plants to devote their specialized metabolism to synthesize and meet the demand for UV protection when they migrated to terrestrial habitats (Weng, 2014). In *S. moellendorffii*, however, half of the clusters belong to the terpenoid class. Compared with moss, *S. moellendorffii* has a much more complex profile of terpenoids (Li et al., 2012; Zi et al., 2014), which might explain the expansion of potentially terpenoid-producing clusters in this species. The predominant presence of alkaloid clusters in *S. polyrhiza* was a surprise, given that we know little about alkaloids in this species.

Several mechanisms have been proposed to lead to the formation of gene clusters in bacteria and fungi (Fondi et al., 2009; Wisecaver et al., 2014; Wisecaver and Rokas, 2015). LD of metabolic genes was enriched in clusters and could have played a role in metabolic gene cluster formation in plants (Supplemental Fig. S9D; Supplemental Table S9). LD has been shown to be lineage specific by several groups (Hanada et al., 2008; Jacquemin et al., 2014). Thus, we examined whether LDs in experimentally characterized clusters are lineage specific and found this to be the case: five clusters that have been examined for conservation contained locally duplicated genes (Supplemental Table S11). In four clusters, the LDs were specific to the species that contain the clusters and absent in related species that do not have the cluster (Supplemental Table S11). One interesting exception is the phytocassane cluster in rice (Miyamoto et al., 2016). There are two LD groups in the rice cluster, and one of the LD groups preceded the emergence of the cluster in this lineage (Supplemental Table S11).

Examples of Primary Metabolic Gene Clusters

While all known metabolic gene clusters encode for specialized metabolic pathways, ~50% of all clusters and ~24% of high-confidence metabolic gene clusters do not contain any genes annotated to specialized metabolism. For example, cluster C438_4 in soybean contains 12 metabolic genes, none of which is associated with the specialized metabolic domain. Three genes encode enzymes that are involved in the early steps of Glc degradation (GLYMA.07G013800, GLYMA.07G014300, and GLYMA.07G015100; Supplemental Fig. S11, A–C). The gene encoding Glc-6-P isomerase (GLYMA.02G212600) was not found in the cluster but located on chromosome 2 and was coexpressed with the Glc-6-P dehydrogenase (GLYMA.07G013800) in the cluster (data not shown). The next step of the Glc degradation pathway is carried out by Fru-1,6-bisP aldolase, which is encoded by several genes in soybean that are not part of this cluster. However, a subset of these genes (GLYMA.04G008300, GLYMA.14G010900, GLYMA.11G111100, GLYMA.12G037400, GLYMA.11G111400, and GLYMA.02G303000) was coexpressed with both Glc-6-P dehydrogenase (GLYMA.07G013800) and 6-phosphate fructokinase (GLYMA.07G014300) in the cluster (data not shown). Partial clustering of the glycolytic pathway was not detected in the other plants in this study.

Another example of a gene cluster not associated with specialized metabolism is cluster C195_3 in Arabidopsis (Supplemental Fig. S12A). This cluster contains five metabolic genes that were predicted to catalyze three out of four reactions of the mitochondrial electron transfer chain (AT2G07689 and AT2G07785, NADPH dehydrogenase; AT2G07695 and AT2G07687, cytochrome *c* oxidase; and AT2G07698, ATP synthase; Supplemental Fig. S12B). This cluster is part of a recent integration of the mitochondrial genome into the Arabidopsis nuclear genome (Lin et al., 1999; Arabidopsis Genome Initiative, 2000; Stupar et al., 2001) and is absent in other closely related species like *A. lyrata* (Hu et al., 2011; Goodstein et al., 2012).

These examples show that primary metabolic pathways can be at least partially clustered. In addition to known pathways, primary metabolic gene clusters can potentially encode alternative metabolic paths that have not yet been described. The same selective pressures that drive or retain the clustering of enzymes annotated to the specialized metabolic domain might act to form primary metabolic gene clusters. For example, when a pathogen attacks a host, both specialized and primary metabolic pathways can be activated, the former to defend against the aggressor and the latter to reduce the concentration of potential resources that otherwise could be exploited by the pathogen. It is also possible that some metabolic gene clusters might encode poorly characterized primary metabolic pathways, such as pathways that control metabolite repair and/or damage control that are yet to be discovered and, therefore, not present in our database (Hanson et al., 2016).

An alternative explanation for the clusters that do not contain any specialized metabolic enzymes is that they are involved in specialized metabolism but the enzyme sequences are more similar to primary metabolic enzymes than specialized metabolic enzymes. Some specialized metabolic enzymes have evolved their function by divergence from primary metabolic enzymes (Weng et al., 2012). Since E2P2 has a precision of 78.2%, specialized metabolic enzymes are potentially misidentified by E2P2 as (the ancestral) primary metabolic enzymes if the specialized metabolic function has not yet been discovered.

CONCLUSION

By developing a high-throughput, high-quality computational pipeline to annotate metabolic genes

and gene clusters, we found a widespread occurrence of metabolic gene clusters in plants that are enriched for specialized metabolism, driven by local gene duplications, with more than 1,700 gene clusters containing the hallmark of known clustered specialized metabolic pathways. In four species with sufficient gene expression data, we defined a stringent cutoff to identify 233 coexpressed clusters. Evidence of coexpression among clustered enzymes highlights high-confidence clusters that can be prioritized for experimental testing. Signature enzymes can suggest the type of metabolites a cluster might produce and inform targeted experimental design. Additionally, any characterized enzyme in a cluster can provide leads for functional dissection of the entire cluster. In Arabidopsis, where most enzymes have been experimentally characterized, 65% of clusters contain at least one experimentally determined enzyme (Supplemental Table S4). The combination of bioinformatics analysis with synthetic biology platforms could provide a new direction for systematically cataloging the chemical diversity of plants and revealing Nature's pharmacopeia in the green world.

MATERIALS AND METHODS

Plant Data

In Supplemental Table S1, we provide the data source and associated information for each species used in this study. The information includes the National Center for Biotechnology Information (NCBI) taxonomy identifiers, common and scientific names, PubMed reference identifiers of articles describing the initial genome sequence release, assessments of genome quality and scaffold size, sources of genomic location information, and names, versions, and sources of protein sequence files. The CEGMA (Parra et al., 2007) score evaluates the fraction of highly conserved core eukaryotic genes annotated in a given genome and allows us to assess the completeness of a genome annotation. Complete reflects the proportion of proteins annotated in a given species with an alignment length of more than 70% of the protein length of a set of 248 core eukaryotic genes. Partial scores deliver a less stringent evaluation of the annotation quality; a score is calculated even if a protein is not complete but still exceeds a precomputed minimum alignment score (Parra et al., 2007). We included genomes with complete CEGMA scores of at least 75% (Supplemental Table S1). Splice variants of genes in each species were removed, and only the longest transcript/protein for each gene was analyzed. While we annotated enzymes, reactions, and pathways for barley (*Hordeum vulgare*), switchgrass (*Panicum virgatum*), and two wheat (*Triticum aestivum*) progenitor genomes (*Triticum urartu* and *Aegilops tauschii*), most of the sequences were in short contigs, and these genomes were not subjected to metabolic gene cluster identification and subsequent analyses.

Acquisition of Genomic Location Information Used to Identify Plant Metabolic Gene Clusters

Data were downloaded from the online BIOMART tool of Phytozome (Goodstein et al., 2012) or Ensembl Plants (Kersey et al., 2014; Supplemental Table S1). For each species, the genome option was chosen from the Dataset dropdown menu. Features from the Attributes option was selected. Under Features, the following fields were selected: (1) Gene stable ID (or Gene name if there's no Gene stable ID); (2) Gene start; (3) Gene end; (4) Chromosome/scaffold name; and (5) Strand. The results were exported as .txt format. Mitochondrial or plastid-encoded genes were excluded from the analysis, since information regarding organelle genomes is not available for all species.

Enzyme Annotation with E2P2 v3.0

E2P2 is designed to identify enzymes based on protein sequence data and classify them according to predicted catalytic functions (Chae et al., 2014). E2P2 v1.0 (Chae et al., 2014) used only four-part EC (IUBMB, 1992) numbers as catalytic functions. In E2P2 v3.0, catalytic functions are defined as EF classes, which are based on either four-part EC numbers or MetaCyc reaction identifiers (Caspi et al., 2014). MetaCyc reaction identifiers often represent reactions at finer resolutions than EC numbers. Including MetaCyc reaction identifiers allowed us to expand the predictable reaction classes to 11,902 (9,095 classes were composed of distinct protein sets) from ~2,300 (Chae et al., 2014).

To train E2P2, we compiled a custom data set of annotated protein sequences from any organism, which we refer to as the RPSD. E2P2 v1.0 used RPSD v1.0 (Chae et al., 2014). E2P2 v3.0 uses an expanded version called RPSD v3.1 that contains 50,184 enzyme and 91,855 nonenzyme sequences (Supplemental Fig. S1A). RPSD v3.1 was compiled from manually curated or experimentally supported data in SwissProt (UniProt Consortium, 2011; November 2014 release), BRENDA (Schomburg et al., 2013; November 2014 release), MetaCyc (Caspi et al., 2014; November 2014 release), and PlantCyc (Zhang et al., 2010; November 2014 release). We obtained an enzymatic protein sequence database by filtering for those sequences annotated with (1) a four-part EC number (IUBMB, 1992), (2) a MetaCyc reaction identifier in MetaCyc (Caspi et al., 2014) or PlantCyc (Zhang et al., 2010), or (3) a leaf-node Gene Ontology (Blake et al., 2013) term under catalytic activity (GO:0003824). Those that fulfilled at least one of these criteria were considered as enzymes. The nonenzyme sequence database was extracted by retaining those that did not have any EC number (IUBMB, 1992; full or partial), catalytic Gene Ontology term (Blake et al., 2013), MetaCyc reaction identifier (Caspi et al., 2014), or enzyme-related keywords in SwissProt (UniProt Consortium, 2011). RPSD v3.1 contanis more than twice the number of enzymes compared to RPSD v1.0 (Chae et al., 2014).

E2P2 uses a two-tiered classification process. The first tier consists of a set of individual classifiers that evaluates query sequences for homology to enzymes using pairwise and profile-sequence approaches. Pairwise sequence comparisons are performed using BLAST (Altschul et al., 1990; e value threshold $\leq$ 1e-2) against RPSD v3.1, and only the top hit is considered. E2P2 v3.0 also implements PRIAM (Claudel-Renard et al., 2003) to perform profile-sequence searches, using custom profile libraries trained on enzyme sequence data in RPSD v3.1. CatFam (Yu et al., 2009) was tested but not used in E2P2 v3.0, since it introduced many false-positive predictions in the ensemble mode (second tier).

The second classification tier of E2P2 consists of an ensemble classifier that integrates predictions from the individual classifiers to produce a final prediction of whether a sequence encodes an enzyme from a given EF class. We tested a number of ensemble integration schemes over all test partitions to determine the highest performing ensemble classifier to use as the final classifier (Supplemental Fig. S1C). For E2P2 v3.0, prediction integration is based on a maximum-weight voting scheme, where the weights represent performance measurements (F1 measure; see below) of each individual classifier on each EF class, as learned during a large-scale training and testing regimen (see below). For each EF class, the maximum-weight voting scheme chooses the prediction from the individual classifier that has the highest performance weight. To handle multifunctional enzymes, the scheme incorporates a threshold to keep EF predictions that fall above a certain performance weight. We empirically tested thresholds stepped by 10%. The threshold of the top 50% performed best and was used to implement E2P2 v3.0 (Supplemental Fig. S1C).

The training routine consists of a 5-fold cross-validation approach (Mosteller and Tukey, 1968). The enzyme and nonenzyme components of RPSD v3.1 were divided into five partitions. Four of the partitions were used to create the BLAST (Altschul et al., 1990) and PRIAM (Claudel-Renard et al., 2003) reference databases. Sequences from the fifth partition were used as test queries against those databases by the respective individual classifiers. Using precision, recall, and F1 measure as metrics, the performance of each classifier on each EF class was measured and recorded for the fifth partition. For every EF class, a set of true positive (TP), false positive (FP), and false negative (FN) sequences were identified as such: true positives were sequences that were correctly predicted to the same EF class as they were originally assigned in the reference database RPSD; false positives were sequences that were predicted to an EF class, although they were not assigned to the EF class in the reference database; false negatives were sequences that were not predicted to an EF class but were originally assigned to the EF class in the reference database. We repeated the process until every partition has been used as a test partition. Precision, recall, and F1 measure were calculated as follows: precision, TP/(TP + FP), which measures how many of the positively retrieved results are likely correct predictions; recall, TP/(TP + FN), which measures the coverage of positive results that were identified; F1 measure, 2TP/(2TP + FP + FN), which is the harmonic mean of both precision and recall and thus integrates both coverage and the likelihood that a positive result is a correct prediction.

The final average F1 measure for each classifier was recorded for each EF class and used as the weight by the ensemble classifiers.

E2P2 v3.0, which includes RPSD v3.1, is available as a package (https://dpb.carnegiescience.edu/labs/rhee-lab/software).

Assessing E2P2 Predictions at the Three-Part EC Level

EF classes can be associated with EC numbers or MetaCyc reaction identifiers. When an EF class was associated with an EC number, we used the first three parts of this EC number. In cases where an EF class was associated with a MetaCyc reaction identifier, we retrieved the EC number of the reaction from MetaCyc and subsequently used the first three parts of this EC number. We then extracted all protein sequences under the same three-part EC and removed duplicates; for example, an enzyme was predicted to both EC 1.1.1.1 and EC 1.1.1.2, and this enzyme was counted as one entry under EC 1.1.1. We then repeated the quality tests as described above (for the EF classes) for each three-part EC number.

Pathway Inference

The Pathway Tools' PathoLogic software (Karp et al., 2011; version 18.5), in conjunction with the reference pathway database MetaCyc (Caspi et al., 2014; version 18.5), was used in pathway inference and pathway database construction. PlantCyc (Zhang et al., 2010; version 9.5) was utilized as an external custom reference database to create pathway databases. The databases generated in this study are available from the PMN project's Web site (www.plantcyc.org). The Pathway Tools default option for taxonomic range-based pruning was used. Enzyme annotations from E2P2 in the form of four-part EC numbers were converted to the corresponding MetaCyc reaction identifiers. In cases where multiple MetaCyc reactions have the same EC number, only the MetaCyc reaction labeled official for the EC number was used. To focus on small molecule metabolism, enzymes that metabolize macromolecules, such as Ser protein kinase, were filtered out. Enzyme annotations from E2P2 were then provided as input to Pathway Tools (Karp et al., 2011). Predicted reactions and pathways that were in PlantCyc but not MetaCyc were incorporated into the pathway databases by an in-house script.

Pathway Validation with the SAVI Pipeline

The SAVI pipeline version 3.02 categorizes the predicted pathways to be retained, deleted, or manually reviewed (Supplemental Fig. S2). SAVI removes false-positive pathways and adds false-negative pathways from PathoLogic's predictions. Traditionally, this step was performed by manual curation, which was time consuming, often requiring several weeks per genome, to validate the predicted pathways from a genome. SAVI accelerates this process by codifying a set of rules derived from years of manual curation. For each species, the SAVI program uses six pathway library files, two taxonomy files, an E2P2 enzyme annotation output file, and four PathoLogic output files to enable semi-automated changes to the predicted pathway databases.

Pathway Library Files

All pathway library files were curated based on published scientific literature and are available online (ftp://ftp.plantcyc.org/Pathways/SAVI_validation_lists/SAVI_lists_pmn10_June_2015/).

(1) Ubiquitous Plant Pathways (UPP; Zhang et al., 2010). Plants are generally photosynthetic autotrophs. All plants have to synthesize many compounds, especially those in primary metabolism. All land plants also synthesize hormones to adapt to their land-living environments. We expect the pathways on the UPP list to exist in all Embryophyta (land plants), such as individual amino acid biosynthetic pathways and abscisic acid biosynthesis. When these pathways are predicted for any Embryophyta species, they are automatically approved. Furthermore, any UPP that have not been predicted for a particular Embryophyta species are added to the database with an Inferred by Curator evidence code. UPP version 6.0 was used for the 21 Embryophyta species.

(2) Common Viridiplantae Pathways (CVP). Pathways on this list are expected to be found in all Viridiplantae. This is a subset of the UPP list that can be used for non-Embryophyta species, including algae. This list is used instead of the UPP for Viridiplantae species outside of Embryophyta. Like the UPP file, it is used to automatically approve any CVP that are predicted and to add any CVP that were not predicted. CVP version 4.0 was used for *Chlamydomonas reinhardtii*.

(3) Non-PMN Pathways (NPP; Zhang et al., 2010). Pathways on this list are excluded from PMN databases because they are nonplant variants of common primary metabolic pathways (e.g. bacterial glycolysis and glycogen biosynthesis), redundant short pathways that are wholly contained within larger pathways, or non-small-molecule metabolic pathways (e.g. those related to protein modification). NPP version 6.0 was used to remove these pathways when they were predicted for any plant species.

(4) Accept-If-Predicted Pathways (AIPP; version 3.0). Pathways on this list are expected to exist in Viridiplantae species but are not considered ubiquitous (e.g. variant pathways for hormone degradation). Therefore, all AIPP predicted by PathoLogic for any species were automatically accepted, but no unpredicted pathways were added to databases from the AIPP.

(5) Conditionally Accepted Plant Pathways (CAPP; version 3.0). This was used to determine which pathways predicted for a given species should be kept based on reaction and/or expected taxonomic range criteria. More details on how a pathway is added to this list are provided below in the section called "*SAVI Procedure*."

(6) Manually Checked Pathways (version 1.0). This small set of pathways was reserved for manual curation in each species because they were considered metabolically important and difficult to predict accurately. All four pathways in this list pertain to different variants of C3 and C4 photosynthesis.

Taxonomy Files

Taxonomic names from NCBI (names.dmp) were downloaded on March 25, 2013. The taxonomic tree structure from NCBI (nodes.dmp) was downloaded on March 25, 2013.

E2P2 Output File

The E2P2 output file (in .pf format) contained enzyme annotations from E2P2 v3.0.

PathoLogic Output Files

PathoLogic output files were as follows: (1) PathoLogic output of pathways (pathways.dat); (2) PathoLogic output of reactions (reactions.dat); (3) PathoLogic output of species (species.dat); and (4) PathoLogic output of proteins (proteins.dat). For species that had previously predicted databases, we used an additional file from PathoLogic, which includes all pathways from the previous version (renamed to pathways_pgdb.dat).

SAVI Procedure

Predicted base pathways and superpathways (Caspi et al., 2013) that are on the UPP/CVP or AIPP lists are automatically accepted, and pathways on the NPP list are automatically rejected. The remaining base pathways are passed through the CAPP filtering process. Pathways can be accepted based on taxonomic and/or reaction criteria. If the species of the database falls within the expected taxonomic range of the pathway in the CAPP file, then the pathway is accepted. Under the reaction criterion, a pathway is accepted if one or more of its key reactions specified in the CAPP file is annotated with a gene. We selected key reactions for each pathway based on several criteria, most prominently: (1) it is associated with only a single pathway in the reference database; (2) it differentiates a variant pathway from other related variants; or (3) it appears as the final step in a biosynthetic pathway or as the first step in a degradation pathway. Any predicted base pathways that are not found in the pathway library files are passed to a list for manual review along with the pathways on the Manually Checked Pathways list. The pathways to be reviewed are manually curated with information from the literature and added to the appropriate SAVI library file. After processing all the base pathways, SAVI rechecks any remaining predicted superpathways (Caspi et al., 2013) and accepts them if all of their constituent base pathways have been accepted. All accepted pathways are associated with an appropriate evidence code to indicate the reason for their inclusion in the accepted pathway file. All rejected pathways are reported in an output file with information about the reason for their rejection. For previously generated databases, SAVI has an optional preliminary step to automatically accept experimentally supported pathways from the previous version (pathways_pgdb.dat).

SAVI v3.02 is written in Java and available as a package (https://dpb.carnegiescience.edu/labs/rhee-lab/software).

Database Construction

Following the initial pathway database creation using PathoLogic and validation using SAVI, custom Perl scripts were used to add (from MetaCyc v18.5 and PlantCyc v9.5) or remove pathways from each database and to update the evidence codes and references for all the retained pathways supported solely by computational prediction or curator inference. Experimentally supported pathways for a given species in MetaCyc v18.5, PlantCyc v9.5, or previous version of the species-specific database, which were not predicted by Pathway Tools, were manually imported.

To preserve previously curated data from the literature, pathway databases of Arabidopsis (*Arabidopsis thaliana*; AraCyc), maize (*Zea mays*; CornCyc), *C. reinhardtii* (ChlamyCyc), rice (*Oryza sativa*; OryzaCyc), and soybean (*Glycine max*; SoyCyc) were updated using the Pathway Tools function Incremental update. Previous enzyme annotations without experimental support were deleted if they were no longer supported by E2P2 v3.0. We ran the Rescore pathways function in Pathway Tools following the Incremental update process. The resulting pathways were validated by SAVI, and the databases were edited as described above.

All the metabolic pathway databases created and analyzed in this study are part of the PMN release version 10.0 and available for downloading online (www.plantcyc.org).

Comparison of Database Quality

To compare the quality of our databases with other publicly available resources such as KEGG (Kanehisa et al., 2014), PlantSEED (Seaver et al., 2014), and Gramene (Monaco et al., 2014), we used 10% of RPSD v3.1 as an independent test set. Within the 10% test set, we further selected the enzymes present in all databases for each species we compared (Arabidopsis, maize, rice, soybean, and tomato [*Solanum lycopersicum*]). These species were chosen for comparison because they were (1) covered by PMN as well as the databases compared with PMN and (2) had at least 15 enzyme reaction pairs in the independent test set. To make a fair comparison, reactions from KEGG (Kanehisa et al., 2014), PlantSEED (Seaver et al., 2014), and Gramene (Monaco et al., 2014) were mapped to MetaCyc reaction identifiers using reaction identifier mapping files from MetaCyc (Caspi et al., 2014), Rhea (Morgat et al., 2015), and the supplemental data file from the PlantSEED publication (Seaver et al., 2014). To compare the quality of databases, we calculated F1 measures of enzyme-reaction associations of the test set in each database for each species. For every MetaCyc reaction identifier, a set of true-positive, false-positive, and false-negative enzyme sequences were identified from each compared database: true positives are sequences that were correctly associated with the same MetaCyc reaction identifier that was assigned in RPSD; false positives are sequences that were predicted to a different MetaCyc reaction identifier in RPSD; false negatives are sequences that were not predicted to a MetaCyc reaction identifier but were assigned to a MetaCyc reaction identifier in RPSD. The cause of false positives in AraCyc was further analyzed by checking the sequence similarity between sequences that were assigned to false-positive EF classes and sequences that were assigned to the correct EF classes in RPSD (Supplemental Table S10).

Classification of Reactions and Pathways into Metabolic Domains

New reactions and pathways in the reference databases MetaCyc (v18.5) and PlantCyc (v9.5) used in this study were classified as described previously (Chae et al., 2014). In total, 13 parent classes were used as our main metabolic domain categories: (1) amides and polyamides; (2) amino acids; (3) carbohydrates; (4) cofactors; (5) detoxification; (6) energy; (7) fatty acids and lipids; (8) hormones; (9) inorganic nutrients; (10) nucleotides; (11) reduction and oxidation (redox); (12) primary-specialized interface; and (13) specialized metabolism. Pathways that channel metabolites from primary to specialized metabolism were labeled as primary-specialized interface metabolism pathways. Pathways associated with processes that do not fit into these 13 main classes were grouped under Other. After the pathways were annotated, reactions that were associated with pathways were classified based on the pathway classification. The reactions that are not associated with pathways were classified manually based on the associated compounds of the reactions. For these reactions, the classification was based on information from the literature, if available. If no data were available, we used the chemical nature of substrates and/or products that the reaction catalyzes to infer the metabolic domain. Reactions whose substrates or products were not specific or did not have any literature information were labeled as unclassified. The metabolic domain classification of the individual reactions was then propagated to the individual metabolic enzymes according to the gene-reaction information stored in the individual PMN databases. The metabolic domain classification of all the genes used in this study are available online (ftp://ftp.plantcyc.org/Pathways/Data_dumps/PMN10_June2015/Gene_classification_to_metabolic_domain).

Identification of Metabolic Gene Clusters

We developed a pipeline called PlantClusterFinder version 1.0 to identify metabolic gene clusters using an iterative approach (Supplemental Fig. S4A). We excluded switchgrass and two wheat progenitor genomes because they did not meet the genome assembly criterion for our analysis (more than 50% of their genomes were assembled in scaffolds with at least 50 genes per scaffold). We also excluded barley because the shotgun assembly of its genome had too many gaps (every third intergenic region on average).

For each species, we labeled genes encoding enzymes by their annotation to metabolic reactions from each species' pathway database. These metabolic genes are associated with MetaCyc reaction identifiers. We iteratively searched for groups of labeled metabolic genes that are contiguously located on the same chromosome using sliding windows. We used the following criteria to identify metabolic gene clusters: (1) at least three metabolic genes must be present in a cluster; (2) at least two distinct MetaCyc reaction identifiers must be represented in the cluster; (3) all genes in the cluster must be located contiguously on the same chromosome; and (4) clusters that consisted only of locally duplicated metabolic genes were not allowed (see below for the identification of locally duplicated metabolic genes). In the first iteration, we did not allow the presence of any intervening nonmetabolic genes; thus, only metabolic enzymes were identified in the cluster. In the second iteration, we allowed one intervening nonmetabolic gene between metabolic genes within a cluster. We iteratively identified clustered genes by increasing the number of intervening nonmetabolic genes by one in each round. Because we were interested in metabolic gene clusters, we stopped this process before the total number of metabolic enzymes found in clusters dropped below the total number of nonmetabolic genes encoded in clusters (Supplemental Fig. S4B).

The gene clusters were further filtered for the enrichment of metabolic genes. For every size of a cluster (composed of n genes), we drew all theoretical clusters from the genome by applying a window size of n to the gene position file and computed the distribution of the theoretical clusters containing 0 to n metabolic genes. A predicted cluster (of size n) is kept if the number of its metabolic genes is above 95% of those theoretical clusters of size n.

PlantClusterFinder v1.0 is written in Java and MATLAB and is available online as a package (https://dpb.carnegiescience.edu/labs/rhee-lab/software).

Handling Sequencing Gaps in Metabolic Gene Cluster Prediction

To identify gene clusters, PlantClusterFinder exploits the sequential order of genes on a scaffold or chromosome. Therefore, sequencing gaps within a genome must be considered to avoid predictions that span unaccounted sequencing gaps. For each species, we downloaded its gene position information from Phytozome or the Ensembl Plants database (Supplemental Table S1). We then produced a sequencing gap position file based on the genome assemblies (Goodstein et al., 2012; Harper et al., 2016; Kersey et al., 2016), where a sequencing gap is defined as a stretch of DNA encoded as the letter N. These two files were used to find all intergenic regions interrupted by a sequencing gap.

From extensive communications with scientists who assembled the genomes, we found that there are two types of sequencing gaps. First, sequencing gaps of unknown length but with an estimated length on the order of several megabases (Supplemental Table S3) were used as natural boundaries for predicting individual gene clusters. Second, there were gaps of known length, which may contain genes. For these, we estimated the number of potential genes (hypothetical genes) and evenly inserted them within the intergenic region containing a sequencing gap (called the affected sequence). Each hypothetical gene was labeled as a nonmetabolic gene.

We employed a Monte Carlo sampling procedure to estimate the most likely number of hypothetical genes per affected sequence. First, we generated theoretical distributions of genomic distances for a sequence encoding 1, 2, 3, …, M genes (including flanking intergenic regions) by randomly sampling the physical size of M genes and M+1 intergenic regions 10,000 times. We then

evaluated whether each affected sequence exceeded the 99th percentile of any of the M distributions sampled. If so, the affected sequence was chosen to encode M hypothetical genes. If an affected sequence exceeded more than one distribution, the maximal M was chosen. M was set to be maximally 20 to ensure the evaluation of all likely hypothetical gene sizes and to reduce the computational time.

Comparison of Metabolic Gene Cluster Frequency among Bacterial, Fungal, and Plant Genomes

To determine the relative frequency of metabolic gene clusters per genome, we used the number of protein-encoding genes as a proxy for genome size because plant genomes vary tremendously in physical size due to repetitive sequences (Mehrotra and Goyal, 2014). We obtained the total number of protein-encoding genes from the databases where we obtained the genomes (Goodstein et al., 2012; Kersey et al., 2014) and plotted the number of predicted clusters as a function of the number of protein-encoding genes (Supplemental Fig. S6A). To obtain the frequency of clusters in bacterial genomes, we used the data from Supplemental Table S11 from Cimermancic et al. (2014). To convert the physical bacterial genome size to the number of protein-encoding genes, we used the genome sizes and the number of genes of bacteria from the Joint Genome Institute-Integrated Microbial Genomes Web site (https://img.jgi.doe.gov/cgi-bin/edu/main.cgi; Supplemental Fig. S6B). To obtain the frequency of clusters in fungal genomes, we used data from Khaldi et al. (2010) and Wisecaver et al. (2014) (Supplemental Fig. S6, C and D).

Whole-Genome Duplication Information

Whole-genome duplication (WGD) gene blocks for 15 species (*C. reinhardtii*, *Physcomitrella patens*, *Selaginella moellendorffii*, rice [*Oryza sativa*], *Brachypodium distachyon*, sorghum [*Sorghum bicolor*], maize, potato [*Solanum tuberosum*], tomato, grape [*Vitis vinifera*], soybean, poplar [*Populus trichocarpa*], papaya [*Carica papaya*], Chinese cabbage [*Brassica rapa* ssp. *pekinensis*], and Arabidopsis) were downloaded from the Plant Genome Duplication Database on June 12, 2015 (Lee et al., 2013). WGD information for the other species analyzed in our study was not available in the Plant Genome Duplication Database. Chinese cabbage's WGD gene block information used a different version of gene identifiers, which could not be mapped to the version we used. Therefore, Chinese cabbage also was left out for WGD-related analyses.

Identification of LD Genes

PlantClusterFinder includes a subpipeline to identify LD metabolic genes. Protein sequences from a genome are first subjected to an all-against-all BLAST followed by clustering using the Markov cluster algorithm (Enright et al., 2002) with the inflation value (cluster granularity parameter) set to 2. PlantClusterFinder takes the Markov cluster algorithm result and the gene position information (Supplemental Table S1) as input to identify LD metabolic genes in a genome using the following criteria for LD genes that were used previously in the literature (Hanada et al., 2008): (1) separated by no more than a 10-gene interval; and (2) within 100 kb from its nearest duplicate.

Enrichment Analysis

To evaluate the enrichment of specific metabolic domains in gene clusters, we compared the proportion of genes belonging to each metabolic domain in clustered metabolic genes (Supplemental Fig. S7D) with the proportion of all metabolic genes in each metabolic domain (Supplemental Fig. S7E). For example, if q is the number of clustered metabolic genes in domain A, m is the number of all clustered metabolic genes, k is the number of all metabolic genes in domain A, and n is the number of all metabolic genes in the genome, the fold change for domain A will be $f = (q/m)/(k/n)$. Domain A is enriched in clustered genes if $f > 1$ and depleted if $f < 1$. We performed a similar analysis of metabolic domain enrichment by removing LD genes from gene clusters (Supplemental Fig. S9C).

Similarly, to evaluate the enrichment of metabolic domains in LD or WGD metabolic genes, we compared the proportion of LD or WGD metabolic genes belonging to each metabolic domain with the proportion of all metabolic genes in each metabolic domain (Supplemental Fig. S9, A and B). To test whether LD or WGD metabolic genes were enriched in gene clusters, we compared the proportion of LD or WGD metabolic genes in clusters with the proportion of all metabolic genes in the genome (Supplemental Fig. S9D).

Statistical significance for all enrichment analyses was calculated by conducting hypergeometric tests in R (phyper).

Coexpression Analysis

We performed coexpression analyses for four species (Arabidopsis, soybean, rice, and tomato) based on coexpression data sets provided by ATTED-II (Aoki et al., 2016; Supplemental Table S6). To identify gene clusters with coexpressed enzyme pairs, we defined a genome-wide coexpression threshold, based on Pearson's correlation coefficient (Usadel et al., 2009), beyond which a gene pair would be considered coexpressed. We used the correlation coefficient of the upper 99th percentile of the coexpression distribution for each species' coexpression matrix.

The genome size and the number of experimental samples influence the selection of a coexpression threshold (Usadel et al., 2009). Therefore, we applied Student's t tests with multiple hypothesis test correction (Usadel et al., 2009) to ensure that all 99th percentile-based Pearson's correlation coefficient thresholds exceeded minimal coexpression values for meaningful statistical significance. Information about the number of genes and experimental samples was acquired from ATTED-II (Aoki et al., 2016; Supplemental Table S6).

For soybean, rice, and tomato, ATTED-II gene identifiers were mapped to Phytozome version 10 gene identifiers. To do this, we retrieved protein sequences from GenBank using RefSeq protein identifiers corresponding to each ENTREZ identifier (via GenBank's internal mapping file ftp://ftp.ncbi.nih.gov/gene/DATA/gene2refseq.gz). Then, protein sequences were BLASTed against the Phytozome protein sequences, all splicing variants included, using BLAST+ 2.2.28+ (Camacho et al., 2009). All perfect matches (100% sequence identity and end-to-end length coverage) were kept. For a query sequence without a perfect match, the best hit was retained if the alignment identity was equal to or above 99% and the ratio of alignment over query length exceeded 95%. To eliminate uncertainty, all queries with more than one perfect hit were removed to generate the final identifier mapping (Supplemental Table S12). For Arabidopsis, we acquired a mapping table from ATTED-II. We limited our analyses to these four species because their enzyme coverage, with respect to identifier mapping, exceeded 50%. Subsequent analyses were performed only on the mapped genes. We applied the coexpression threshold derived from the original coexpression matrix described above to identify coexpressed gene pairs.

From all metabolic gene cluster predictions, we only retained clusters with at least one coexpressed enzyme pair (Supplemental Fig. S8). To further rank the retained clusters, we assigned a rank for each cluster:

$$r = 1 - p_{binomial} = 1 - \left(\binom{n}{k} p_{coex=1}^{k} * p_{coex=0}^{n-k} \right)$$

Here, k denotes the number of coexpressed metabolic gene pairs with a probability $p_{coex=1} = 0.01$ and $n - k$ is the number of noncoexpressed metabolic gene pairs, each given a probability of $p_{coex=0} = 0.99$. The rank of a cluster is defined as 1 minus the binomial probability of a metabolic gene cluster's composition of coexpressed and noncoexpressed enzyme pairs, without considering the order of their occurrence in the cluster. The lower this binomial probability, the less likely a cluster's composition has occurred by chance, and hence the higher its rank. Therefore, we reward the number of coexpressed metabolic gene pairs per cluster while penalizing the number of metabolic gene pairs without coexpression. In addition to disregarding the order of coexpressed enzyme pairs, the binomial coefficient is used to normalize the rank with respect to the cluster size (the total number of metabolic gene pairs).

Since each species has two coexpression data sets (microarray and RNA sequencing derived), two different ranks per cluster were obtained. Therefore, we selected our final rank as:

$$r_{Cluster} = max(r_{Microarray}, r_{RNA\text{-}Seq})$$

The biological rationale is that a gene cluster might be more active within the set of experimental conditions represented in only one of the data sets. Subsequently, we constructed a cumulative probability distribution (Supplemental Fig. S8) based on the binomial probabilities of all remaining metabolic gene clusters in order to decide on a maximum probability cutoff for a final

high-confidence set of gene cluster predictions. We selected $p_{binomial} = 0.01$ around the elbow of the cumulative distribution curve as the maximum binomial probability cutoff for a cluster to be considered within the high-confidence set (Supplemental Fig. S8). This is equivalent to a minimum rank of $r = 0.99$. Beyond the elbow point of the cumulative distribution curve, the number of lower ranked clusters increases drastically.

Cataloging Metabolic Gene Clusters Based on Signature Enzymes

Specialized metabolism was classified into subdomains based on the metabolites they produce or metabolize (Wink, 2010; Caspi et al., 2014). We examined four subdomains (terpenoids, phenylpropanoids, alkaloids, and polyketides) and compiled signature enzymes that produce the scaffold compounds of each subdomain (Supplemental Table S5). Scaffolds of terpenoids generally are produced from prenyl diphosphates, geranyl diphosphate, farnesyl diphosphate, and geranylgeranyl diphosphate, by terpene synthases (Wink, 2010). To compile the list of signature enzyme reactions, we extracted MetaCyc (Caspi et al., 2014) reactions producing terpenes from these three building blocks. Compounds of the second subdomain, phenylpropanoids, generally are derived from Phe by Phe ammonia lyase. The subsequent product trans-cinnamate is then further transformed into diverse classes of phenylpropanoids and derivatives, including the most studied classes flavonoids and stilbenes (Wink, 2010). To compile the list of signature enzymes for phenylpropanoids, we extracted MetaCyc reactions producing the scaffolds of major types of flavonoids (chalcone, flavanone, flavone, dihydroflavonol, flavonol, flavan-4-ol, flavan-3,4-diol, flavan-3-ol, anthocyanidin, and isoflavonoids) and stilbenes by consulting Wink (2010). Similarly, we extracted MetaCyc reactions producing the scaffolds of major types of the third subdomain alkaloids. Compounds of the subdomain plant polyketides generally are produced via the condensation of malonyl-CoA with another acyl-CoA by polyketide synthases. To compile the list for the polyketide subdomain, we extracted such reactions from MetaCyc. The reactions producing chalcone, stilbene, and acridone alkaloid catalyzed by chalcone synthase, stilbene synthase, and acridone synthase, respectively, also are polyketide synthase-type reactions. In this study, we classified them as signature enzymes of phenylpropanoids or alkaloids, not as polyketides, following the classification scheme in MetaCyc and Wink (2010). The compiled reactions (signature enzymes) were associated with four-part EC numbers or MetaCyc reaction identifiers if a complete four-part EC number was absent. Several reactions were removed from the signature enzymes list because they were catalyzed by enzymes whose sequence similarities were indistinguishable from those that catalyze different reactions (Supplemental Table S10).

In addition, we compiled reactions catalyzed by tailoring enzymes. EC numbers of typical tailoring enzymes include EC 2.4.* (glycosyltransferases), EC 2.1.1.* (methyltransferases), and EC 2.3.1.* (acyltransferases). Cytochrome P450s and 2ODD generally are considered to be tailoring enzymes, but they also can act as signature enzymes, such as the few listed as phenylpropanoid signature enzymes (Supplemental Table S5) and in several known clusters that make cyanogenic glucosides (Boycheva et al., 2014; Nützmann and Osbourn, 2014). Therefore, cytochrome P450s and 2ODDs were classified separately from the other, more typical tailoring reactions (Supplemental Table S5). We removed reactions from the list of tailoring reactions if they were already identified to be signature enzymes.

Genes in the predicted clusters were classified subsequently as signature or tailoring based on the reaction annotation scheme above. Each cluster was then classified based on the classification of its component genes (Supplemental Table S4).

Characterization of Arabidopsis Gene Clusters

Single gene transposition data were obtained from Woodhouse et al. (2011; Supplemental Table S9). Genomic locations of gene clusters were obtained from the BIOMART tool of Phytozome (Goodstein et al., 2012) or Ensembl Plants (Kersey et al., 2014). Arabidopsis transposons and their genomic locations were downloaded from TAIR (https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_transposable_elements/TAIR10_Transposable_Elements.txt). Transposons were associated with genes if they were found within 1,000 bp upstream or downstream of the genes. All statistical tests of enrichment were performed using hypergeometric tests in R (phyper).

Examination of the Evolution and Conservation of Previously Reported Plant Metabolic Gene Clusters

We collected cluster conservation information from the literature (Supplemental Table S11). When such information was not available, we performed the following analysis. For each cluster of species X, we searched for ortholog information for each one of the clustered pathway enzymes in a closely related species Y. We used two resources to retrieve orthologs, Ensembl Plants (Kersey et al., 2016) and PLAZA (Van Bel et al., 2012). If species Y was not included in the ortholog analysis in either of those resources, the best hit by BLASTP in species Y was chosen as the ortholog of the pathway enzyme. A cluster was considered to be conserved in species Y if (1) orthologs of at least two enzymes annotated to at least two reactions were found in species Y and (2) the orthologs were located in close proximity (less than 573 kb, the size that contains 99.9% of all predicted clusters) on a chromosome.

For any of the reported functional clusters, we examined the presence of LD among the enzyme-coding genes. Five clusters contained at least one group of genes that are local duplicates of each other (LD group). An LD group was considered to be conserved in species Y if (1) orthologs of at least two genes of the LD group were found in species Y and (2) the orthologs were separated by no more than a 10-gene interval and within 100 kb, criteria for LD used by Hanada et al. (2008).

Several of the reported functional clusters were not included in this analysis because genomes of their closely related species are not available.

Availability of Data and Materials

Project name, E2P2 v3.0; project home page, https://dpb.carnegiescience.edu/labs/rhee-lab/software; operating system, 64-bit Linux; programming language, Python; other requirements, Java 1.6.1 or higher (included in the package); license, GNU GPLv3 (the package includes BLAST and PRIAM software, which have their own licenses); restrictions, no specific restrictions. Please follow the restrictions of BLAST and PRIAM.

Project name, SAVI v3.02; project home page, https://dpb.carnegiescience.edu/labs/rhee-lab/software; operating systems, 64-bit Linux and 64-bit Windows 7 or higher; programming language, Java; other requirements, Java SDK 1.6 or higher; license: GNU GPLv3; restrictions, none.

Project name, PlantClusterFinder v1.0; project home page, https://dpb.carnegiescience.edu/labs/rhee-lab/software; operating system, 64-bit Linux; programming language, MATLAB, Java; other requirements, Java SDK 1.6 or higher; license, GNU GPLv3.

The metabolic databases generated in this study (PlantCyc, AraCyc, BarleyCyc, BrachypodiumCyc, CassavaCyc, ChineseCabbageCyc, ChlamyCyc, CornCyc, GrapeCyc, MossCyc, OryzaCyc, PapayaCyc, PoplarCyc, PotatoCyc, SelaginellaCyc, SetariaCyc, SorghumBicolorCyc, SoyCyc, SpirodelaCyc, SwitchgrassCyc, TomatoCyc, WheatACyc, and WheatDCyc) are available from the PMN project Web site (www.plantcyc.org). All the cluster data are available as additional files.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. The RPSD and performance of E2P2.

Supplemental Figure S2. Scheme and effect of the SAVI pipeline.

Supplemental Figure S3. Overview of metabolic pathway database content.

Supplemental Figure S4. Overview of the metabolic gene cluster identification pipeline PlantClusterFinder.

Supplemental Figure S5. Prevalence of metabolic genes in clusters.

Supplemental Figure S6. Number of metabolic gene clusters per 1,000 genes in genomes.

Supplemental Figure S7. Metabolic gene cluster size distribution.

Supplemental Figure S8. Pipeline for high-confidence metabolic gene cluster prediction.

Supplemental Figure S9. Enrichment of metabolic domains in LD or WGD metabolic genes.

Supplemental Figure S10. Three-part EC level performance of E2P2 compared with EF class level performance.

Supplemental Figure S11. Predicted cluster example for the Glc degradation pathway.

Supplemental Figure S12. Predicted cluster examples for the mitochondrial electron chain.

Supplemental Table S1. Data sources.

Supplemental Table S2. Public plant pathway databases.

Supplemental Table S3. Information compiled by this study about sequencing gaps of different sizes in the genome annotation files of the 17 plants and one algal species.

Supplemental Table S4. List of predicted metabolic gene clusters of each species with gene and cluster classifications labeled.

Supplemental Table S5. List of experimentally characterized plant metabolic gene clusters, validation of cluster prediction, and classification of signature, cytochrome P450, and tailoring enzymes.

Supplemental Table S6. Experimental descriptions for all gene expression data sets that were used by ATTED-II to construct the gene coexpression data sets.

Supplemental Table S7. List of high-confidence gene clusters.

Supplemental Table S8. Metabolic domain classification for all base pathways and reactions used in this study.

Supplemental Table S9. Enrichment tests of transpositions in clustered and nonclustered, metabolic and nonmetabolic genes.

Supplemental Table S10. Analysis of potential misannotations between similar EF classes, and investigation of false-positive annotations in Arabidopsis.

Supplemental Table S11. Examination of the evolution and conservation of previously reported plant metabolic gene clusters involved in specialized metabolite biosynthesis.

Supplemental Table S12. Summary of ATTED identifier mappings in each species.

ACKNOWLEDGMENTS

We thank C.-H. You for helping with generating the reference enzyme sequence database; C. Yu for providing CatFam profile training source code; T. Sen for initial analysis of the reference enzyme sequence database; M. Schaeffer, J. Gardiner, and L. Harper for participating in maize pathway curation; R. Caspi and H. Foerster for plant pathway curation in MetaCyc; P. May for helping to migrate ChlamyCyc to our website; P. Karp and S. Paley for Pathway Tools support; D. Priamurskiy, V. Dwaraka, T. Tran, and C. Johansen for help with data retrieval and entry; C.R. Buell, F. Cheng, C.J. Dill, D.M. Goodstein, R.D. Hayes, T. Itoh, G.J. King, L.A. Mueller, S. Prochnik, S. Rounsley, S. Saha, H. Sakai, J. Schmutz, P.S. Schnable, J. Stein, X. Wang, and D. Ware for help in clarifying how sequencing gaps are annotated in genome assembly files; and B. Muller and G. Huntress for systems support.

Received December 20, 2016; accepted February 21, 2017; published February 22, 2017.

LITERATURE CITED

Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK, et al (2012) KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant Cell Physiol **53:** e1

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

Aoki Y, Okamura Y, Tadaka S, Kinoshita K, Obayashi T (2016) ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. Plant Cell Physiol **57:** e5

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815

Bednarek P, Osbourn A (2009) Plant-microbe interactions: chemical diversity in plant defense. Science **324:** 746–748

Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, Bridges S, Burgess S, Buza T, McCarthy F, et al (2013) Gene Ontology annotations and resources. Nucleic Acids Res **41:** D530–D535

Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, Pujar A, Leto J, Gosselin J, Mueller LA (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Res **39:** D1149–D1155

Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, Osbourn A (2015) Investigation of terpene diversification across multiple sequenced plant genomes. Proc Natl Acad Sci USA **112:** E81–E88

Boycheva S, Daviet L, Wolfender JL, Fitzpatrick TB (2014) The rise of operon-like gene clusters in plants. Trends Plant Sci **19:** 447–459

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics **10:** 421

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, et al (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res **42:** D459–D471

Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, et al (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res **40:** D742–D753

Caspi R, Dreher K, Karp PD (2013) The challenge of constructing, classifying, and representing metabolic pathways. FEMS Microbiol Lett **345:** 85–93

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

Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, et al (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell **158:** 412–421

Claudel-Renard C, Chevalet C, Faraut T, Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res **31:** 6633–6639

Dale JM, Popescu L, Karp PD (2010) Machine learning methods for metabolic pathway prediction. BMC Bioinformatics **11:** 15

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

Dharmawardhana P, Ren L, Amarasinghe V, Monaco M, Thomason J, Ravenscroft D, McCouch S, Ware D, Jaiswal P (2013) A genome scale metabolic network for rice and accompanying analysis of tryptophan, auxin and serotonin biosynthesis regulation under biotic stress. Rice (N Y) **6:** 15

Dutartre L, Hilliou F, Feyereisen R (2012) Phylogenomics of the benzoxazinoid biosynthetic pathway of Poaceae: gene duplications and origin of the Bx cluster. BMC Evol Biol **12:** 64

Ehrlich PR, Raven PH (1964) Butterflies and plants: a study in coevolution. Evolution **18:** 586–608

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res **30:** 1575–1584

Farnsworth NR (1988) Screening plants for new medicines. *In* EO Wilson, ed, Biodiversity. National Academy of Sciences/Smithsonian Institution, Washington, DC, pp 83–97

Field B, Fiston-Lavier AS, Kemen A, Geisler K, Quesneville H, Osbourn AE (2011) Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proc Natl Acad Sci USA **108:** 16116–16121

Fondi M, Emiliani G, Fani R (2009) Origin and evolution of operons and metabolic pathways. Res Microbiol **160:** 502–512

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol **60:** 433–453

Gogna N, Hamid N, Dorai K (2015) Metabolomic profiling of the phytomedicinal constituents of Carica papaya L. leaves and seeds by ^{1}H NMR spectroscopy and multivariate statistical analysis. J Pharm Biomed Anal **115:** 74–85

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res **40:** D1178–D1186

Grotewold E (2005) Plant metabolic diversity: a regulatory perspective. Trends Plant Sci **10:** 57–62

Gunatilaka AL (2008) Natural Products in Plants: Chemical Diversity. Wiley Encyclopedia of Chemical Biology. John Wiley & Sons, New York

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol **148:** 993–1003

Hanson AD, Henry CS, Fiehn O, de Crécy-Lagard V (2016) Metabolite damage and metabolite damage control in plants. Annu Rev Plant Biol **67:** 131–152

Harper L, Gardiner J, Andorf C, Lawrence CJ (2016) MaizeGDB: the Maize Genetics and Genomics Database. Methods Mol Biol **1374:** 187–202

Hoffmeister D, Keller NP (2007) Natural products of filamentous fungi: enzymes, genes, and their regulation. Nat Prod Rep **24:** 393–416

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet **43:** 476–481

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

IUBMB (1992) Enzyme Nomenclature. Academic Press, San Diego

Jacquemin J, Ammiraju JS, Haberer G, Billheimer DD, Yu Y, Liu LC, Rivera LF, Mayer K, Chen M, Wing RA (2014) Fifteen million years of evolution in the Oryza genus shows extensive gene family expansion. Mol Plant **7:** 642–656

Jermy T (1984) Evolution of insect/host plant relationships. Am Nat **124:** 609–630

Jones CG, Firn RD (1991) On the evolution of plant secondary chemical diversity. Philos Trans R Soc Lond B Biol Sci **333:** 273–280

Jung S, Ficklin SP, Lee T, Cheng CH, Blenda A, Zheng P, Yu J, Bombarely A, Cho I, Ru S, et al (2014) The Genome Database for Rosaceae (GDR): year 10 update. Nucleic Acids Res **42:** D1237–D1244

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res **42:** D199–D205

Karp PD, Latendresse M, Caspi R (2011) The pathway tools pathway prediction algorithm. Stand Genomic Sci **5:** 424–429

Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C, et al (2016) Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res **44:** D574–D580

Kersey PJ, Allen JE, Christensen M, Davis P, Falin LJ, Grabmueller C, Hughes DS, Humphrey J, Kerhornou A, Khobova J, et al (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res **42:** D546–D552

Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, Fedorova ND (2010) SMURF: genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol **47:** 736–741

Lee TH, Tang H, Wang X, Paterson AH (2013) PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res **41:** D1152–D1158

Li G, Köllner TG, Yin Y, Jiang Y, Chen H, Xu Y, Gershenzon J, Pichersky E, Chen F (2012) Nonseed plant *Selaginella moellendorffi* [corrected] has both seed plant and microbial types of terpene synthases. Proc Natl Acad Sci USA **109:** 14711–14715

Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, et al (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature **402:** 761–768

Matsuba Y, Nguyen TT, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, Schäfer P, Kudrna D, Wing RA, Bolger AM, et al (2013) Evolution of a complex locus for terpene biosynthesis in *Solanum*. Plant Cell **25:** 2022–2036

Mehrotra S, Goyal V (2014) Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. Genomics Proteomics Bioinformatics **12:** 164–171

Mintz-Oron S, Mandel T, Rogachev I, Feldberg L, Lotan O, Yativ M, Wang Z, Jetter R, Venger I, Adato A, et al (2008) Gene expression and metabolism in tomato fruit surface tissues. Plant Physiol **147:** 823–851

Miyamoto K, Fujita M, Shenton MR, Akashi S, Sugawara C, Sakai A, Horie K, Hasegawa M, Kawaide H, Mitsuhashi W, et al (2016) Evolutionary trajectory of phytoalexin biosynthetic gene clusters in rice. Plant J **87:** 293–304

Moghe GD, Last RL (2015) Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. Plant Physiol **169:** 1512–1523

Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, Amarasinghe V, Youens-Clark K, Thomason J, Preece J, et al (2014) Gramene 2013: comparative plant genomics resources. Nucleic Acids Res **42:** D1193–D1199

Morgat A, Axelsen KB, Lombardot T, Alcántara R, Aimo L, Zerara M, Niknejad A, Belda E, Hyka-Nouspikel N, Coudert E, et al (2015) Updates in Rhea: a manually curated resource of biochemical reactions. Nucleic Acids Res **43:** D459–D464

Mosteller F, Tukey JW (1968) Data analysis, including statistics. *In* G Lindzey, E Aronson, eds, Handbook of Social Psychology, Vol 2. Addison-Wesley, Reading, MA, pp 80–203

Nützmann HW, Huang A, Osbourn A (2016) Plant metabolic clusters: from genetics to genomics. New Phytol **211:** 771–789

Nützmann HW, Osbourn A (2014) Gene clustering in plant specialized metabolism. Curr Opin Biotechnol **26:** 91–99

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

Osterman A, Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol **7:** 238–251

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23:** 1061–1067

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA **101:** 9903–9908

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

Qi X, Bakht S, Leggett 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

Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics **23:** 401–407

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

Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, Munaretto C, Ulas S, Stelzer M, Grote A, et al (2013) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res **41:** D764–D772

Seaver SM, Gerdes S, Frelin O, Lerma-Ortiz C, Bradbury LM, Zallot R, Hasnain G, Niehaus TD, El Yacoubi B, Pasternak S, et al (2014) High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. Proc Natl Acad Sci USA **111:** 9645–9650

Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J (2001) Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc Natl Acad Sci USA **98:** 5099–5103

Takos AM, Knudsen C, Lai D, Kannangara R, Mikkelsen L, Motawia MS, Olsen CE, Sato S, Tabata S, Jørgensen K, et al (2011) Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in *Lotus japonicus* and suggests the repeated evolution of this chemical defence pathway. Plant J **68:** 273–286

Tieman D, Zeigler M, Schmelz E, Taylor MG, Rushing S, Jones JB, Klee HJ (2010) Functional analysis of a tomato salicylic acid methyl transferase and its role in synthesis of the flavor volatile methyl salicylate. Plant J **62:** 113–123

UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) **2011:** bar009

Urbanczyk-Wochniak E, Sumner LW (2007) MedicCyc: a biochemical pathway database for *Medicago truncatula*. Bioinformatics **23:** 1418–1423

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

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol **158:** 590–600

Van Moerkercke A, Fabris M, Pollier J, Baart GJ, Rombauts S, Hasnain G, Rischer H, Memelink J, Oksman-Caldentey KM, Goossens A (2013) CathaCyc, a metabolic pathway database built from *Catharanthus roseus* RNA-Seq data. Plant Cell Physiol **54:** 673–685

Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, et al (2015) antiSMASH 3.0: a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res **43:** W237–W243

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

Weng JK (2014) The evolutionary paths towards complexity: a metabolic perspective. New Phytol **201:** 1141–1149

Weng JK, Philippe RN, Noel JP (2012) The rise of chemodiversity in plants. Science **336:** 1667–1670

Wink M (2010) Biochemistry of Plant Secondary Metabolism. Wiley-Blackwell, Chichester, UK

Wisecaver JH, Rokas A (2015) Fungal metabolic gene clusters: caravans traveling across genomes and environments. Front Microbiol **6:** 161

Wisecaver JH, Slot JC, Rokas A (2014) The evolution of fungal metabolic pathways. PLoS Genet **10:** e1004816

Woodhouse MR, Tang H, Freeling M (2011) Different gene families in *Arabidopsis thaliana* transposed in different epochs and at different frequencies throughout the rosids. Plant Cell **23:** 4241–4253

Yang YW, Lai KN, Tai PY, Li WH (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. J Mol Evol **48:** 597–604

Yu C, Zavaljevski N, Desai V, Reifman J (2009) Genome-wide enzyme annotation with precision control: catalytic families (CatFam) databases. Proteins **74:** 449–460

Zhang P, Dreher K, Karthikeyan A, Chi A, Pujar A, Caspi R, Karp P, Kirkup V, Latendresse M, Lee C, et al (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. Plant Physiol **153:** 1479–1491

Zi J, Mafu S, Peters RJ (2014) To gibberellins and beyond! Surveying the evolution of (di)terpenoid metabolism. Annu Rev Plant Biol **65:** 259–286