www.sciencemag.org/content/344/6183/510/suppl/DC1

Supplementary Materials for

Genomic Signatures of Specialized Metabolism in Plants

Lee Chae, Taehyong Kim, Ricardo Nilo-Poyanco, Seung Y. Rhee*

*Corresponding author. E-mail: srhee@carnegiescience.edu

Published 2 May 2014, *Science* **344**, 510 (2014)
DOI: 10.1126/science.1252076

This PDF file includes:

Materials and Methods

Figs. S1 to S9

Tables S1 to S4

Captions for data files S1 to S5

References

Other supplementary material for this manuscript includes:

Data files S1 to S5 (zipped archives)

Materials and Methods

Annotation of enzymes and metabolic network reconstruction

Whole proteome datasets for the following species were retrieved from Phytozome (www.phytozome.org) on the dates noted (table S1, fig. S1):

August 2, 2011
Chlamydomonas reinhardtii, GenBank accession number ABCN00000000 (*18*)
Glycine max (soybean), GenBank accession number ACUP00000000 (*19*)
Manihot esculenta (cassava), GenBank accession number PRJNA17471 (*20*)
Oryza sativa ssp. *japonica* (rice), GenBank accession number BABO00000000 (*21*)
Populus trichocarpa (poplar), GenBank accession number AARH00000000 (*22*)
Selaginella moellendorfii, GenBank accession number ADFJ00000000 (*23*)
Vitis vinifera (grapevine), GenBank accession number CAAP00000000 (*24*)

December 12, 2011
Sorghum bicolor, GenBank accession number ABXC00000000 (*25*)

July 25, 2013
Volvox carteri, GenBank accession number ACJH00000000 (*26*)
Coccomyxa subellipsoidea, GenBank accession number AGSI00000000 (*27*)
Micromonas pusilla CCMP1545, GenBank accession number ACCP00000000 (*28*)
Micromonas pusilla RCC299, GenBank accession number ACCO00000000 (*28*)
Ostreococcus lucimarinus, GenBank accession number GCA_000092065.1 (*29*)

Proteome datasets for *Zea mays* (maize, GenBank accession number AHID00000000) were downloaded from maizesequence.org (www.maizesequence.org) on August 5, 2011 (*30*), and for *Arabidopsis thaliana* (GenBank accession number GCA_000001735.1) from TAIR (www.arabidopsis.org) on August 2, 2011 (*31*).

We used CEGMA to confirm the quality of the genome annotations. CEGMA tests annotation quality by searching for the presence of 458 conserved, core eukaryotic proteins (*32*). The average CEGMA score for the species analyzed was 92.21% with grapevine having the lowest score (78.63%) (table S1). In comparison, the high confidence protein set for *Picea abies* (downloaded from ftp://congenie.org/congenie/fasta/GenePrediction/ on July 26, 2013) registered a 53.23% CEGMA score.

For a given plant species, protein sequences were submitted as individual queries against our in-house enzyme classification pipeline, E2P2 (Ensemble Enzyme Prediction Pipeline), which can be downloaded at https://dpb.carnegiescience.edu/labs/rhee-lab/software. The pipeline relies on homology transfer to annotate enzyme sequences in the form of full, four-part Enzyme Commission (EC) numbers, using single sequence (BLAST, E-value cutoff ≤ 1e-30, subset of SwissProt 15.3) and multiple sequence (Priam, November 2010; CatFam, version 2.0, 1% FDR profile library) models of enzymatic functions (*33–36*). The pipeline integrates predictions from the individual methods into a final set of annotations using an average weighted integration algorithm, where the weight of each prediction from each individual method was determined by averaging performance using the F_1-measure as a metric over 1,000 rounds of bootstrap testing (*37*) (fig. S2).

We trained and validated the pipeline using a dataset of 116,829 protein sequences from SwissProt 15.3 (*38*), where each sequence had evidence of existence at the protein or transcript level and whose functional annotation with Gene Ontology terms had an experiment-based evidence code (EXP, IDA, IPI, IMP, IGI, IEP). We then partitioned those sequences with four-part EC numbers and Gene Ontology IDs under 'catalytic activity' (GO:0003824) into a positive dataset of 25,562 enzyme sequences representing 2,406 distinct four-part EC numbers, and a dataset of 91,267 sequences representing non-enzymes (no EC designation partial or full, no GO IDs under 'catalytic activity,' no enzyme-related keywords, and not listed in the Enzyme database [http://enzyme.expasy.org]). We used a 0.632 bootstrap process (*39*) to assess the performance of the pipeline and its individual components over 1,000 rounds of testing, calculating precision, recall, and F_1-measure for each method on each EC number in each round of testing. Overall, the pipeline's average weighted integration scheme produced results of 86% precision, 87% recall, and 86% F_1-measure, compared to the individual methods whose results ranged from 54-81% precision, 57-82% recall, and 55-79% F_1-measure (fig. S3).

We also assessed the pipeline's performance on the whole genome of *Arabidopsis thaliana* (TAIR release version 10), using a gold-standard dataset of 1,300 manually curated enzyme annotations from AraCyc 7.0 (*40*) with full, four-part EC numbers and experimental evidence of function. The pipeline delivered performances of 77% precision, 67% recall, and 72% F_1-measure, which represents a marked improvement in coverage while maintaining precision compared to the previous annotation method used in building AraCyc 7.0 (75% precision, 15% recall, 25% F_1-measure)(fig. S4)(*40*).

Of the 472,176 query sequences from the 16 species, the pipeline classified a total of 84,361 as enzymes, representing 1,219 distinct four-part EC numbers. We then created individual Cyc databases for each species based on the annotated sequences using the Pathway Tools software suite (*41*). Pathway Tools matches annotated enzyme sets to metabolic reactions and their compounds based on a curated dataset of reference reactions in MetaCyc 15.0. Overall, 1,808 distinct enzymatic reactions were identified, based on the annotated enzyme sequences. A spreadsheet containing EC and reaction annotations for each species in the study is provided as file S1.

Functional classification of reactions

The foundation of the classification system is the Pathway Ontology provided by PlantCyc version 8.0 and MetaCyc version 15.0 (http://metacyc.org/META/class-tree?object=Pathways) (*40*, *41*). The ontology classifies metabolic pathways into a hierarchical system based on biological function reported in the literature, from which we used a set of 13 parent classes as our main functional categories: Amines and polyamines, Amino acids, Carbohydrates, Cofactors, Detoxification, Energy, Fatty acids and lipids, Hormones, Inorganic nutrients, Intermediate metabolism, Nucleotides, Redox, and Specialized metabolism. The Intermediate metabolism category contains pathways linking primary metabolic pathways to multiple downstream pathways that can represent several domains of metabolism. We also created an extra category, 'Other', to represent all annotations that do not fit into the 13 main classes. The functional classification assigned to a given pathway was transferred to the reactions, EC numbers, and proteins/genes associated with that pathway.

Once we performed the annotation process for the 16 species, we manually verified the classifications for all reactions involved in the study (file S2). In addition, we validated all reactions associated to specialized metabolism pathways using the literature (citations are provided in file S2).

In some instances, an EC number may catalyze a metabolic reaction that is not associated with any known pathway. These reactions are termed orphan reactions. In this case, we used the list of compounds that are consumed and produced by the reaction and their labeling in the MetaCyc Compound Ontology to see whether a reaction could be categorized into one of the 13 functional classes. In situations where all of the non-currency compounds associated to the reaction are clearly related to one of the functional classes, we assigned that reaction to that class. In cases where there is a mixture of compound types, we categorized the orphan reaction to the 'Other' category.

The functional classification of pathways allowed us to associate 1,493 of the predicted 1,808 reactions (82.6%) to at least one of the 13 functional classes. The remaining reactions largely represented large-molecule metabolism and reactions with no known pathways or compound classification and were bundled into the 'Other' category. In some cases, a reaction may be annotated to more than one pathway, and thus, carry more than one functional class assignment. Of the 1,808 reactions, 303 (16.8%) are labeled to more than one functional class. A spreadsheet of EC, reaction, pathway, and functional class assignments is provided as file S2.

To analyze the effects of gene clustering on specialized metabolism, we created a secondary classification of specialized metabolism. We relied on the PlantCyc and MetaCyc pathway descriptions to identify the compounds produced by the specialized metabolic pathways annotated in our data. We then grouped relevant compounds under three main classes of specialized metabolic compounds, as follows:

Nitrogen-containing specialized compounds
1. Alkaloids
2. Amides
3. Aminoglycosides
4. Benzoxazinoids
5. Betalains
6. Cyanogenic glucosides
7. Cyclic amides
8. Deazzapurines
9. Glucosinolates
10. Melanins
11. Non-protein amino acid derivatives
12. Organosulfurs

Phenypropanoid derivatives
1. Anthocyanins
2. Benzenoids
3. Benzoate derivatives

4. Cinnamates derivatives
5. Coumarins
6. Flavonoids
7. Lignins
8. Phenolic compounds
9. Prenylflavonoids
10. Stilbenes

Terpenoids
1. Monoterpenoids
2. Diterpenoids
3. Triterpenoids
4. Sesquiterpenoids

Other
1. Antibiotics
2. Fatty acid derivatives
3. Pentitol
4. Sugar acids

The classification of primary versus specialized metabolic processes has been a major challenge in the plant metabolism field. With the computational and genomic resources now available, including the genome-scale metabolic networks reported on here, the field is in a position to redefine how metabolic processes should be classified in light of the global, interconnected way one can now view pathways and their reactions and metabolites. One approach would be to create a broader and deeper classification system that offers a larger and more refined set of groupings to catalogue the diverse functionalities provided by plant metabolism. In addition, these classes ought to be organized within a hierarchical structure, to enable analyses at different levels of grouping, such as with gene ontologies or organismal taxonomies.

Metabolic network reconstruction

We extracted reaction and compound information from each species-specific database, and converted the data into bi-directional, reaction-centric metabolic networks where nodes represent reactions and edges (links between nodes) represent compounds shared between two nodes (*7*). A compound was connected to a reaction if the compound is either a substrate or product of the reaction, as curated in MetaCyc 15.0. We removed 24 currency compounds from the networks, as listed below (*42*). In total, the 16 metabolic networks contained 1,621 distinct reaction nodes.

Currency compounds
1. Proton
2. Water
3. Oxygen
4. $NADP^+$
5. NADPH

6. ATP
7. Diphosphate
8. Carbon dioxide
9. Phosphate
10. ADP
11. Coenzyme A
12. UDP
13. NAD^+
14. NADH
15. AMP
16. Ammonia
17. Hydrogen peroxide
18. Oxidized electron acceptor
19. Reduced electron acceptor
20. 3-5-ADP
21. GDP
22. Carbon monoxide
23. GTP
24. FAD

Reaction node similarity and hierarchical clustering

We calculated the similarity of reaction node sets between all-pairwise combinations of species using the Jaccard index as a similarity metric (*43*). The resulting similarity value represents the percentage of all unique nodes in the two sets that are found in common between the two species. We calculated distances among species by subtracting the Jaccard index from 1 and assembling results into a distance matrix. We performed average-linkage hierarchical clustering on the distance matrix using the heatmap.2 function in the R gplots package. We calculated clustering support over 1,000 rounds of bootstrap data using the pvclust package in R (*44*). Similar to bootstrap analysis of phylogenetic trees, pvclust assesses uncertainty in a hierarchical clustering by repeating the clustering process over many random samples of the input data. The Approximately Unbiased (AU) p-values represent the strength of support for the clusters in the hierarchically clustered data.

Statistical analyses

All statistical analyses were performed using R, including phyper (hypergeometric), kruskal.test (Kruskal-Wallis), chisq.test (X^2), t.test (Student's t-test), and wilcox.test (Wilcoxon Rank Sum) from the stats package; ks.boot (Kolmogorov-Smirnov) from the matching package; and z.test (Z test) from the BSDA package.

Enzyme expansion rates

For a given reaction class (represented by a unique, four-part Enzyme Commission number) found among two or more species, we plotted the $\log_{10}$ of the number of protein sequences associated with that reaction class against the $\log_{10}$ of the total number of proteins in the genome, for each of the species. We then performed a least-squares linear regression on the variables using the lm function in the R stats package, and extracted the slope of the linear regression. In the case of a logarithmic transformation of both variables, the estimated slope of the regression represents a power-law scaling exponent, which indicates how the number of enzymes associated with a reaction class has expanded with respect to changes in the total number of proteins (*45*). Scaling exponents near 1 indicate a reaction class whose associated protein inventories have expanded linearly along with all the proteins in the genome. Exponents near 0 indicate minimal changes in reaction inventory size compared to changes in proteome size.

Gene duplication data

Arabidopsis, soybean, and sorghum loci retained after whole-genome-duplication were downloaded from the Plant Genome Duplication Database (PGDD) on December 23, 2011 (Arabidopsis) and October 17, 2013 (soybean and sorghum) (*46*). These species were chosen because the PGDD data were derived using the same genome annotation versions as those used in the local duplication (LD) analysis in this study. Locally duplicated (LD) genes were identified by searching the Arabidopsis, soybean, and sorghum genomes using the following criteria: all LD genes must be 1) associated with the same four-part EC number, 2) separated by no more than a 10-gene interval, and 3) within 100kb from its nearest duplicate (*47*, *48*).

Metabolic gene cluster identification and co-expression

Metabolic gene clusters in the analyzed genomes represent a heterogeneous mix of enzymatic functions according to these parameters: 1) all genes in a cluster must be associated with a four-part EC number, 2) more than one distinct EC number must be represented in the cluster, and 3) all genes in a cluster must be contiguously located on the same chromosome. Clusters that consisted entirely of a tandemly duplicated metabolic gene were removed from analysis. To generate a background estimation of clustering, we randomly redistributed EC annotations among all genes in each genome, scanned the shuffled genomes for metabolic gene clusters, and recorded the number and size of the identified clusters for each species. We repeated the process 1,000 times per species to estimate the final background distribution of cluster sizes.

Pearson-based co-expression correlation values for all gene-gene comparisons were taken from the Arabidopsis gene expression database ATTED-II (http://atted.jp)(*16*). The dataset consisted of 1,388 microarray slides comprising 58 individual experimental treatments, including biotic and abiotic stress, hormones, and light, as well as tissue-specific/development-related samples. We calculated co-expression values for four types of gene sets: 1) clustered metabolic genes, 2) genes in metabolic pathways, 3) random groups of genes, and 4) genes located contiguously along Arabidopsis chromosomes (neighboring genes). For the neighboring gene set, we used a sliding-window approach where we calculated co-expression for all contiguous genes residing within windows of size 3-10 and 14 along the Arabidopsis genome. The window sizes were chosen to match the distribution of cluster sizes found in Arabidopsis (Fig. 3A).

To calculate the co-expression value for each cluster or pathway, we performed the following:

1) We divided the genes annotated to the cluster or pathway according to the reactions they encode.

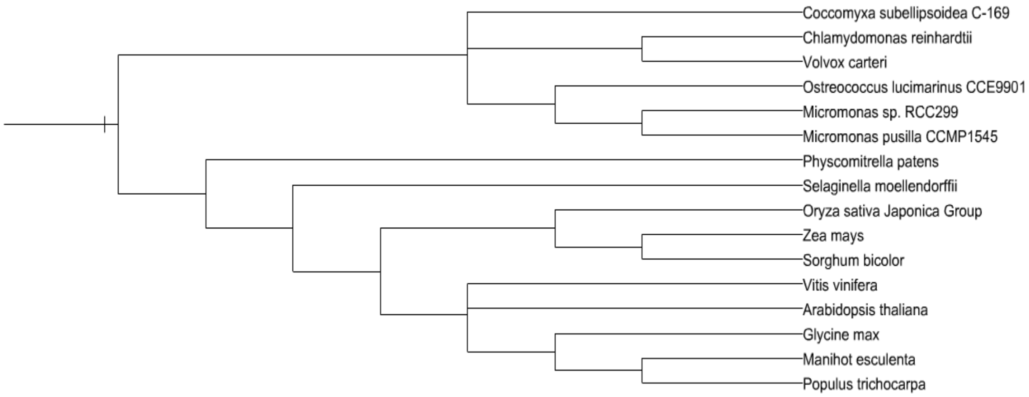
2) For the resulting set of reactions, we identified all possible combinations of genes so that each gene in the combination represents a different reaction in the set and no reaction is represented by more than one gene.

3) For each combination of genes, we extracted the pairwise co-expression values for all genes in the combination across all experimental treatments and recorded the average of the maximum co-expression values for the gene combination.

4) The final co-expression value for the cluster or pathway represents the highest co-expression value found among all of the gene combinations.

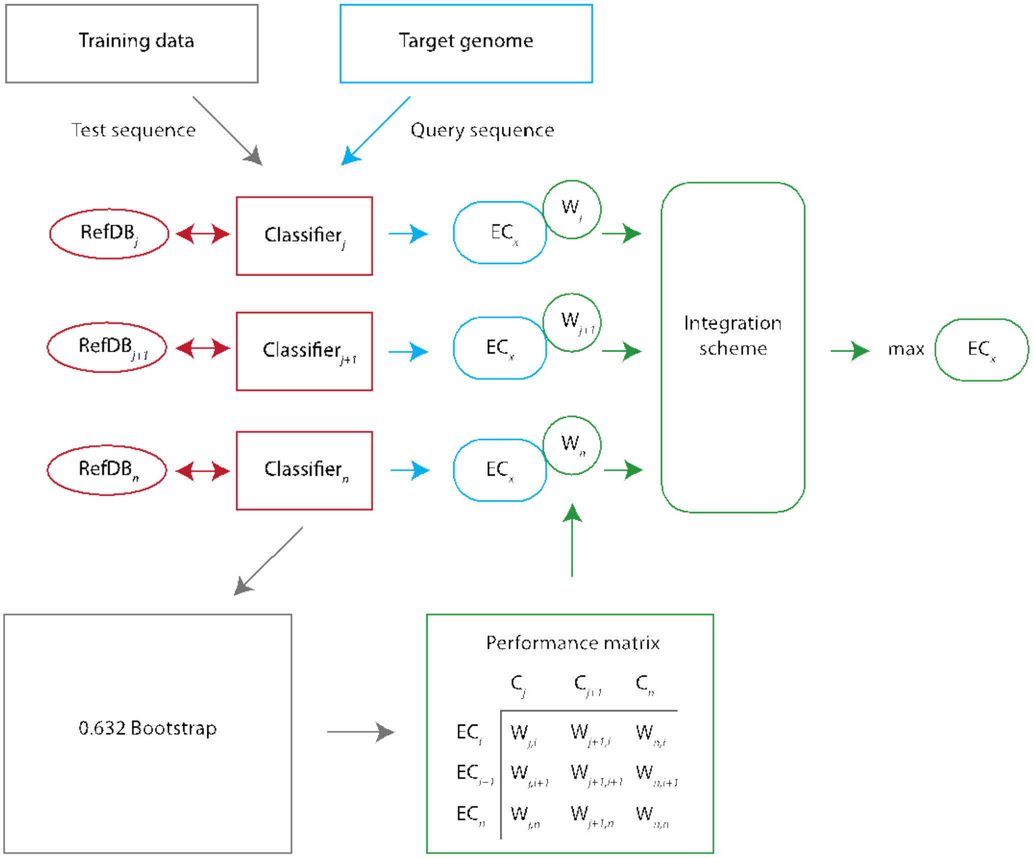
To generate the co-expression value for a random or neighboring gene set, we extracted the maximum co-expression value found across all samples for each pairwise combination of genes in the set. We then averaged the maximum values (as with Step 3 for the cluster and pathway gene sets) to calculate the final co-expression value for the gene set. We compared the distributions of co-expression values between the background sets (random and neighboring) and the test sets (specialized metabolism clusters and pathways and non-specialized metabolism clusters and pathways) using the Wilcoxon Rank Sum test. In the cluster example shown in fig. S9, the background control consisted of the distribution of maximum co-expression correlation values for all pairwise comparisons within each cluster.

Fig. S1.


Coccomyxa subellipsoidea C-169
Chlamydomonas reinhardtii
Volvox carteri
Ostreococcus lucimarinus CCE9901
Micromonas sp. RCC299
Micromonas pusilla CCMP1545
Physcomitrella patens
Selaginella moellendorffii
Oryza sativa Japonica Group
Zea mays
Sorghum bicolor
Vitis vinifera
Arabidopsis thaliana
Glycine max
Manihot esculenta
Populus trichocarpa

Sixteen species from the green plant lineage. The tree shown here depict the phylogenetic relationships among the 16 species used in the analysis. The tree was generated by the Interactive Tree of Life (http://itol.embl.de) and represents a pruned version of the NCBI taxonomy (*49*).

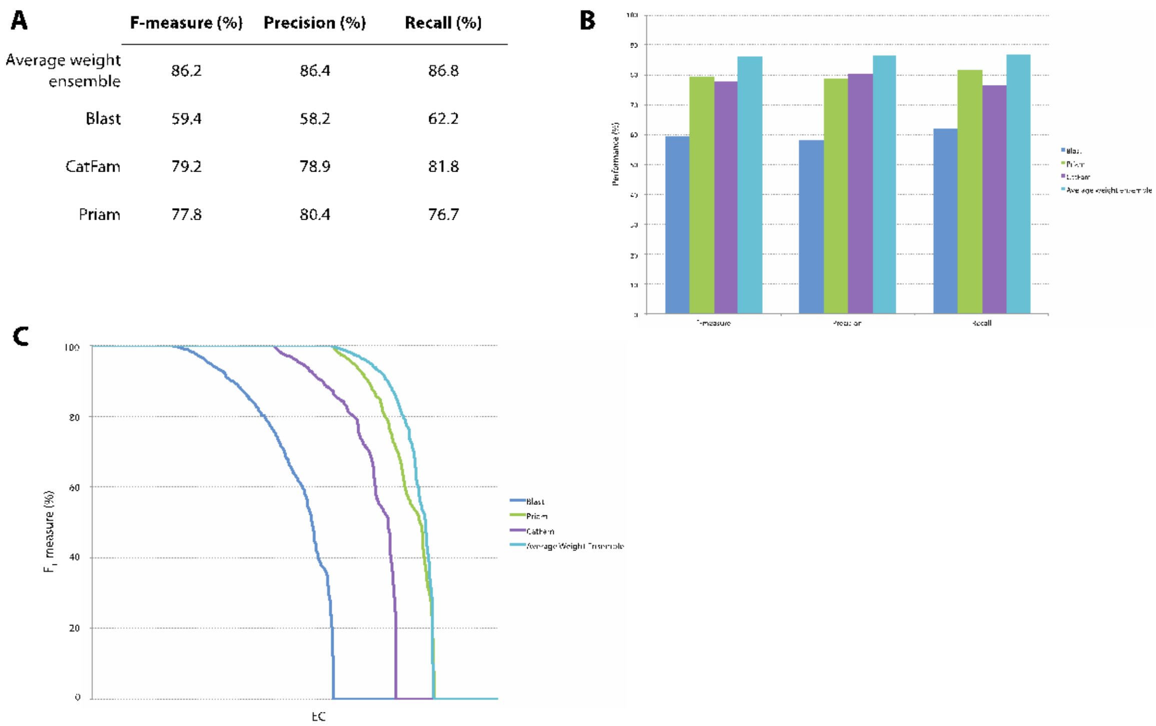
Fig. S2.



	C_j	C_{j+1}	C_n
EC$_i$	$W_{j,i}$	$W_{j+1,i}$	$W_{n,i}$
EC$_{i-1}$	$W_{j,i+1}$	$W_{j+1,j+1}$	$W_{n,i+1}$
EC$_n$	$W_{j,n}$	$W_{j+1,n}$	$W_{n,n}$

Schematic of the Ensemble Enzyme Prediction Pipeline (E2P2). Gray arrows and outlines indicate the processes involved in learning the performance weights stored in the performance matrix. All performance weights were estimated over 1,000 rounds of 0.632 bootstrap testing (*39*). The training data is based on a subset of SwissProt 15.3, where the existence and functional annotation of each sequence is inferred from experimental evidence (*38*). The total training dataset contains 25,562 sequences representing 2,406 four-part EC numbers and 91,267 sequences labeled as non-enzymes (*36*). Blue arrows and outlines indicate the processes involved in base-level enzyme functional predictions. Query sequences from the target genome are submitted to the individual classifiers (red). Each classifier returns a predicted EC number (or no EC), based on searches against its reference database. Next, weights are appended to each prediction using data learned from the training procedure, where a weight represents the performance of that classifier on that EC number. All weighted predictions are then processed by the integration algorithm. Green arrows and outlines represent the ensemble classification process (*37*). All weighted predictions are processed by an integration algorithm to produce a final prediction. In this case, the decision is based on an average weight function where the EC number with the highest average weight among the predictions is selected as the final prediction.

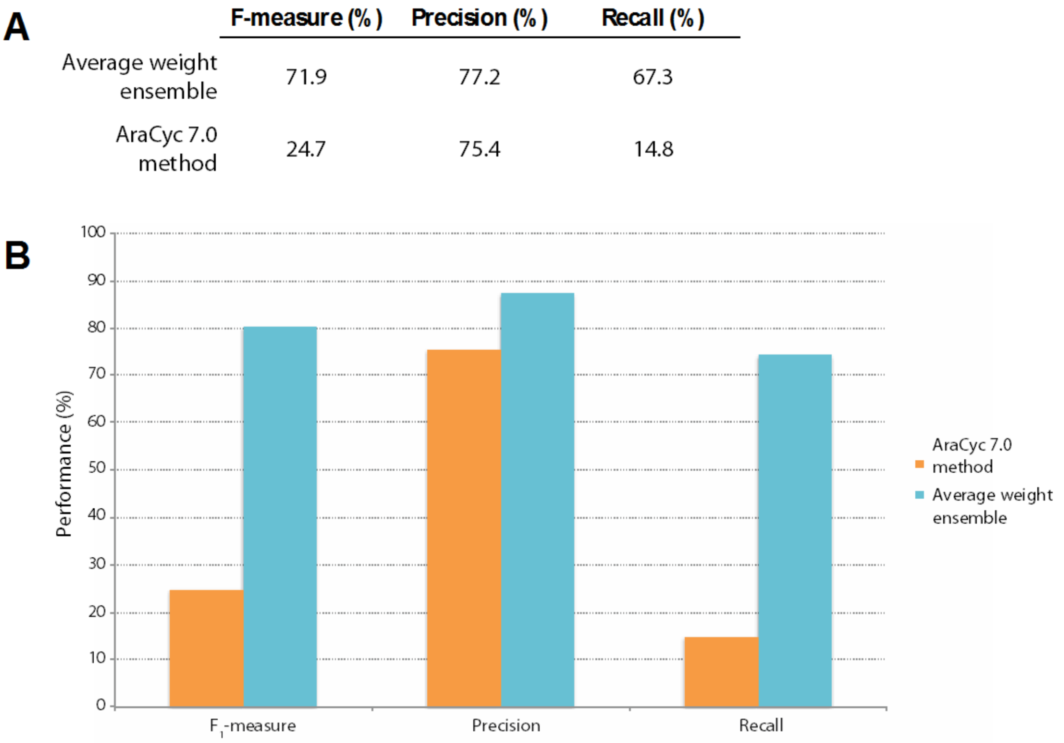
Fig. S3.



A	F-measure (%)	Precision (%)	Recall (%)
Average weight ensemble	86.2	86.4	86.8
Blast	59.4	58.2	62.2
CatFam	79.2	78.9	81.8
Priam	77.8	80.4	76.7

Performance of the Ensemble Enzyme Prediction Pipeline (E2P2). (A, B) Performance of ensemble algorithm and base-level methods for all EC numbers in 1,000 rounds of 0.632 bootstrap testing on a curated data set of 116,829 proteins (*6*). (C) F_1-measure performance of ensemble algorithm and base-level methods, sorted by decreasing performance across all EC numbers.

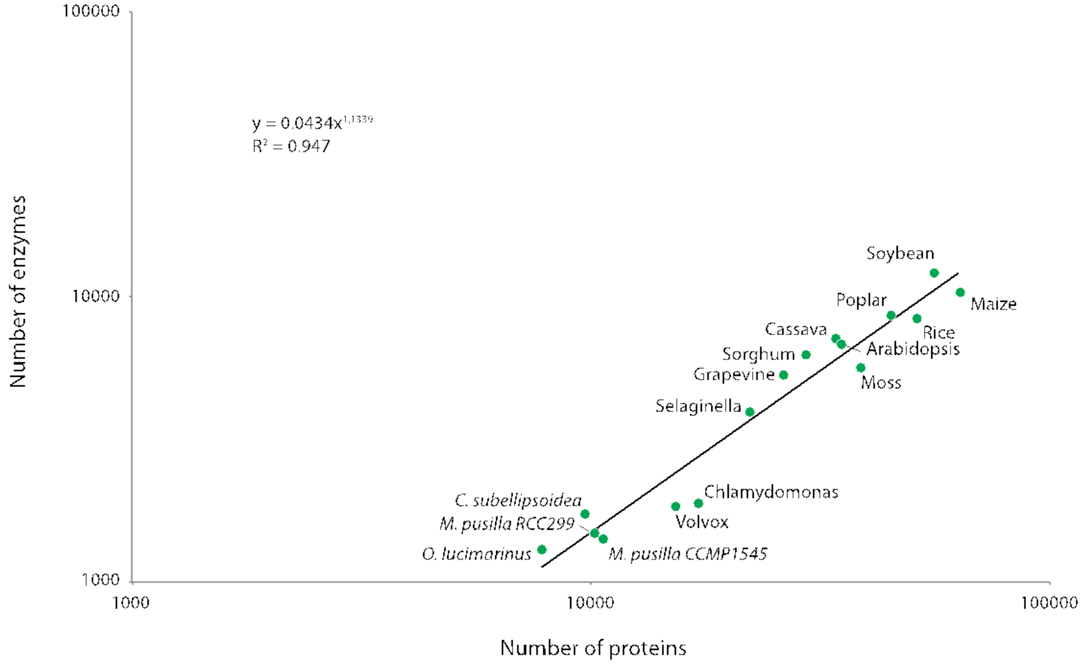
Fig. S4.



	F-measure (%)	Precision (%)	Recall (%)
Average weight ensemble	71.9	77.2	67.3
AraCyc 7.0 method	24.7	75.4	14.8

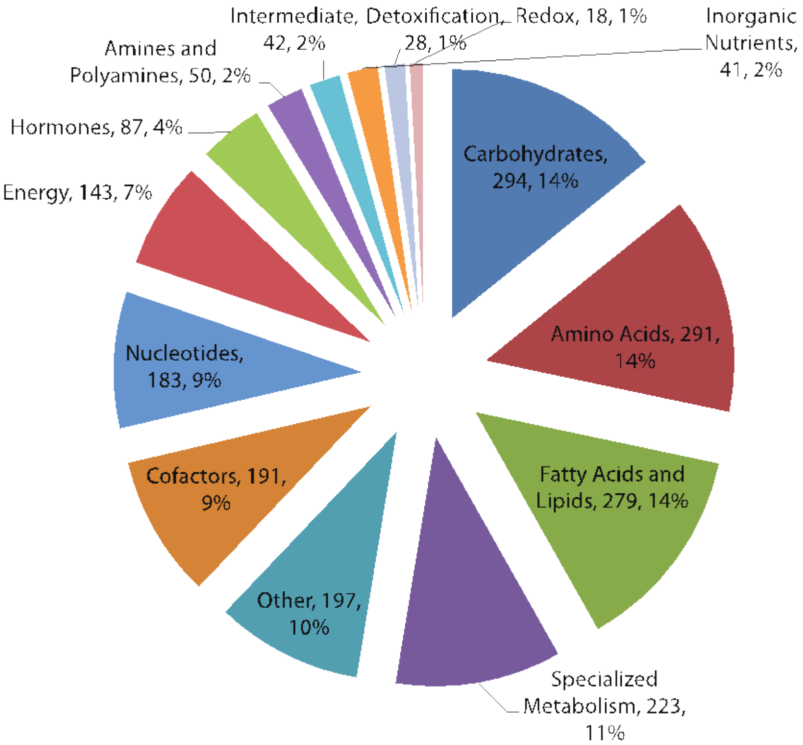
Validation of the Ensemble Enzyme Prediction Pipeline (E2P2) on AraCyc data. (A, B) Performance of ensemble algorithm versus enzyme annotation method used to predict four-part EC numbers in AraCyc version 7.0 (*40*). Data set consisted of 1,300 manually curated Arabidopsis enzyme sequences with four-part EC numbers and experimental evidence of function.

Fig. S5.



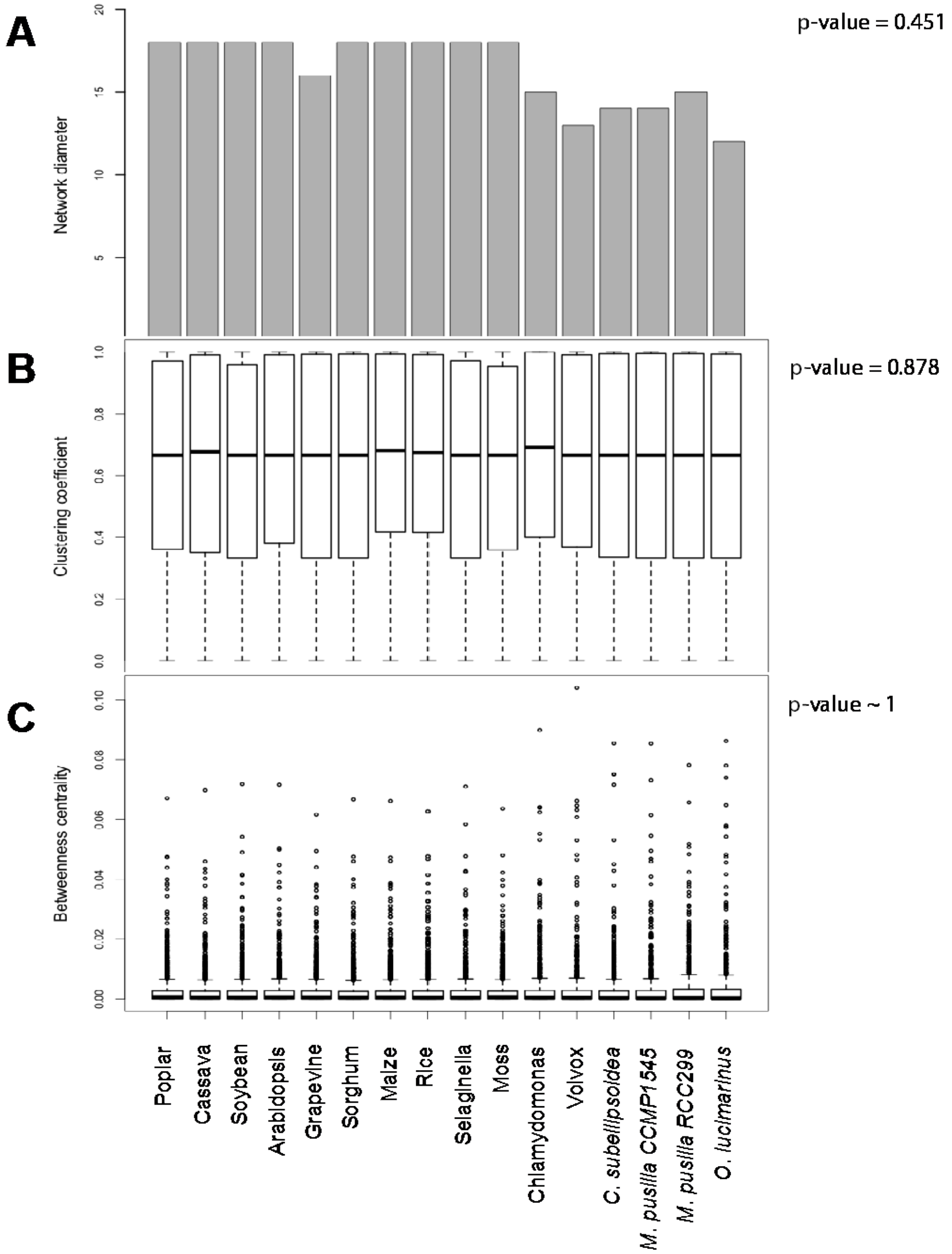
Growth of the enzyme complement as a function of total number of proteins. The scaling exponent of 1.1339 indicates that the number of enzymes has increased linearly with respect to proteome size for the species analyzed.

Fig. S6.



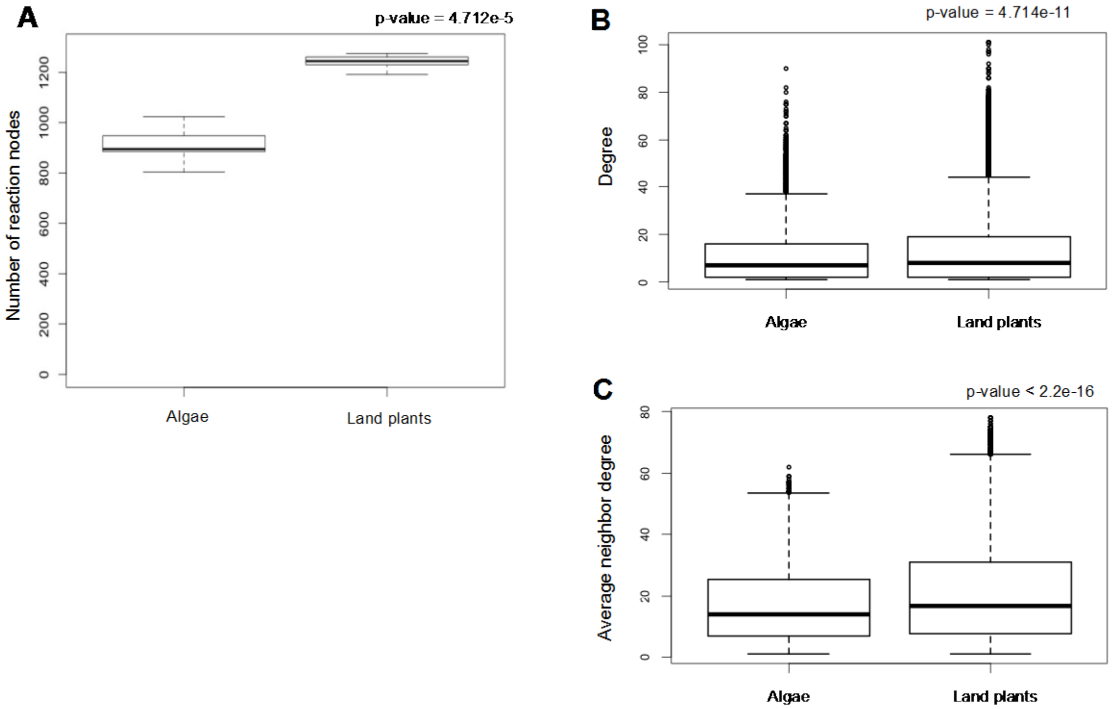
Functional composition of the union of metabolic network reactions among all 16 species. Functional classes were annotated according to the protocol described in Materials and Methods (*6*).

Fig. S7.



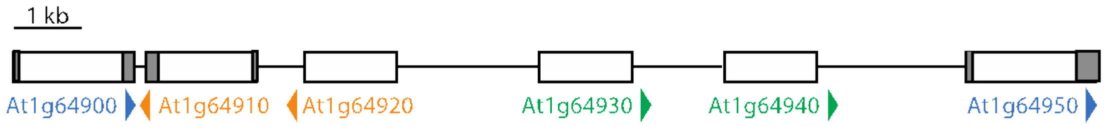
Comparison of network structure among the species. Network diameter is a measure of network size. Clustering coefficient and betweenness centrality are measures of network density. Differences in these network features were tested using the Kruskal-Wallis test.

Fig. S8.



Differences in the level of network connectivity between land plants and algae. The average number of reaction nodes in the land plant networks is significantly greater than that found in the algal networks (Student's t-test) (A). The increased number of reaction nodes, given the similarity of size and density across all of the networks, has resulted in greater connectivity in the plants networks, as seen in comparisons of the degree distributions (B) and average neighbor degrees (C) (Kruskal-Wallis test). Note: A node's degree is the number of nodes to which it is directly connected, while a node's average neighbor degree is the average degree of its immediate neighboring nodes. The degree distribution and average neighbor degree distribution describe the values of these measures for all nodes in a network.

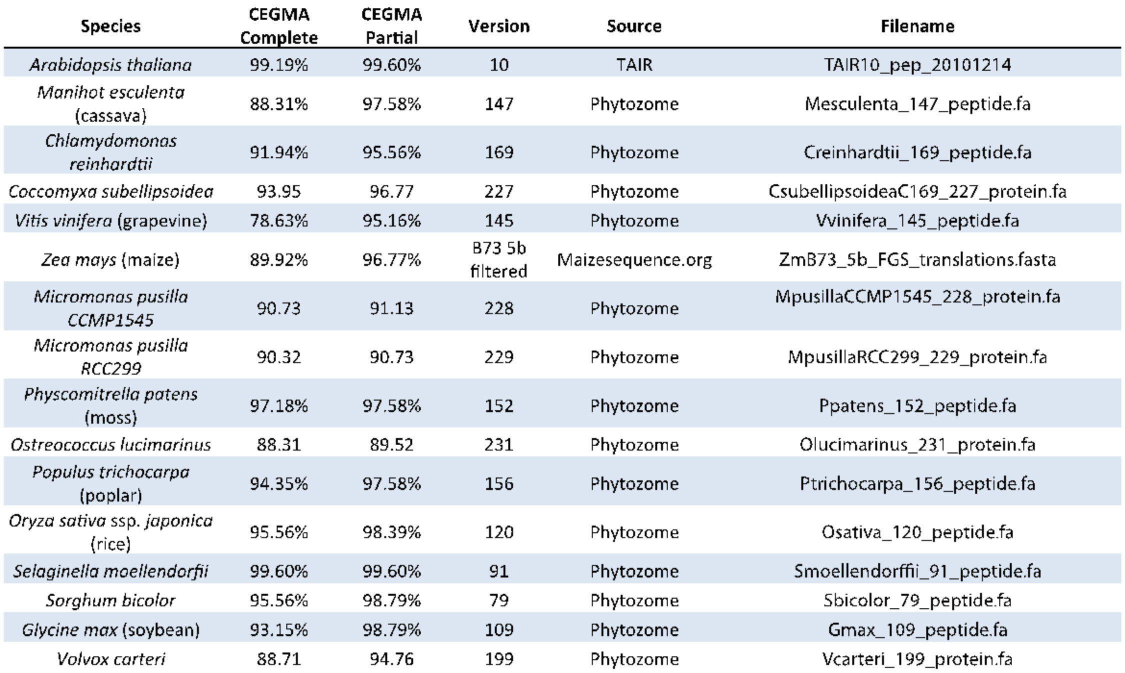
Fig. S9.



Gene pairs		Condition (co-expression)
At1g64900	At1g64930	Light (0.472), biotic stress (0.4385)
At1g64920	At1g64930	Tissue (0.4706)
At1g64920	At1g64940	Light (0.5941)

Cluster example. Cluster 104 contains six genes annotated to three types of monooxygenation/hydroxylation and glycosylation reactions referenced in MetaCyc (*24*). The reference reactions are assigned to separate pathways that act downstream of 2*S*-naringenin production in the flavonoid biosynthesis pathway and are related to the metabolism of anthocyanins, flavonoids, and isoflavones. However, the classes of compounds associated with these reactions are quite broad. Furthermore, the combination of monooxygenation/hydroxylation and glycosylation has been observed for several known plant specialized metabolic pathways, where hydroxylation followed by glycosylation serves to stabilize a highly reactive compound or render toxic ones safe for storage (*50*). Given that the existence of genistein, daidzein, and other isoflavones in Arabidopsis is disputed (*51*), it is possible that the currently uncharacterized genes in cluster 104 may be involved in the synthesis of novel metabolites. Intriguingly, we identified a number of instances in which a monooxygenase/hydroxylase-related member of cluster 104 co-expressed with a glycosylation-related member, perhaps in formulating a novel metabolic response (mean co-expression of randomized clusters = 0.261; standard deviation of co-expression from randomized clusters = 0.209, see Materials and Methods).

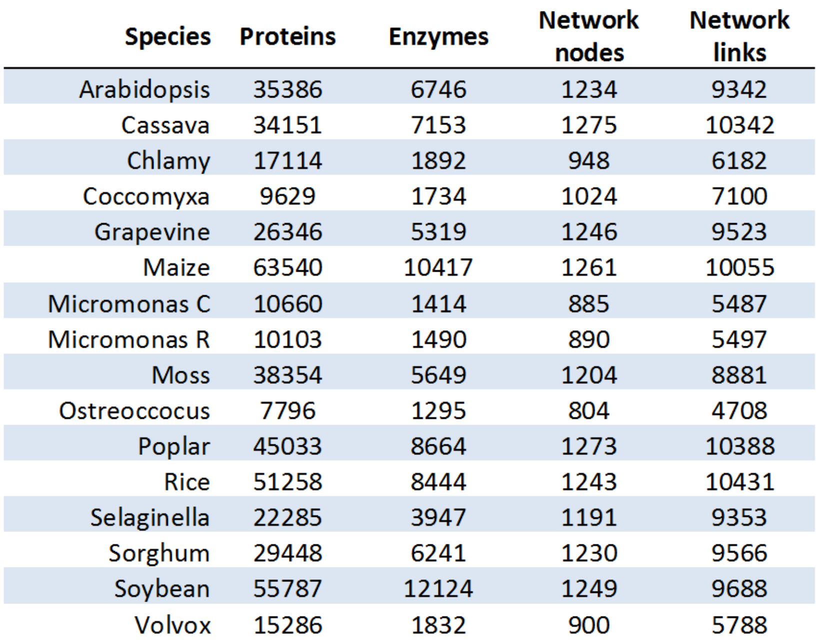
Table S1.



Species	CEGMA Complete	CEGMA Partial	Version	Source	Filename
Arabidopsis thaliana	99.19%	99.60%	10	TAIR	TAIR10_pep_20101214
Manihot esculenta (cassava)	88.31%	97.58%	147	Phytozome	Mesculenta_147_peptide.fa
Chlamydomonas reinhardtii	91.94%	95.56%	169	Phytozome	Creinhardtii_169_peptide.fa
Coccomyxa subellipsoidea	93.95	96.77	227	Phytozome	CsubellipsoideaC169_227_protein.fa
Vitis vinifera (grapevine)	78.63%	95.16%	145	Phytozome	Vvinifera_145_peptide.fa
Zea mays (maize)	89.92%	96.77%	B73 5b filtered	Maizesequence.org	ZmB73_5b_FGS_translations.fasta
Micromonas pusilla CCMP1545	90.73	91.13	228	Phytozome	MpusillaCCMP1545_228_protein.fa
Micromonas pusilla RCC299	90.32	90.73	229	Phytozome	MpusillaRCC299_229_protein.fa
Physcomitrella patens (moss)	97.18%	97.58%	152	Phytozome	Ppatens_152_peptide.fa
Ostreococcus lucimarinus	88.31	89.52	231	Phytozome	Olucimarinus_231_protein.fa
Populus trichocarpa (poplar)	94.35%	97.58%	156	Phytozome	Ptrichocarpa_156_peptide.fa
Oryza sativa ssp. *japonica* (rice)	95.56%	98.39%	120	Phytozome	Osativa_120_peptide.fa
Selaginella moellendorfii	99.60%	99.60%	91	Phytozome	Smoellendorffii_91_peptide.fa
Sorghum bicolor	95.56%	98.79%	79	Phytozome	Sbicolor_79_peptide.fa
Glycine max (soybean)	93.15%	98.79%	109	Phytozome	Gmax_109_peptide.fa
Volvox carteri	88.71	94.76	199	Phytozome	Vcarteri_199_protein.fa

Genome data sources and annotation quality check. CEGMA tests genome annotation quality by searching for the presence of 458 conserved, core eukaryotic proteins (*32*). The CEGMA Complete score represents the percentage of the 458 test proteins found in the target species that meet a minimum alignment length threshold of 70%. The CEGMA Partial score represents the percentage of the 458 test proteins that do not meet the 70% alignment length threshold but still exceed a pre-computed minimum alignment score.

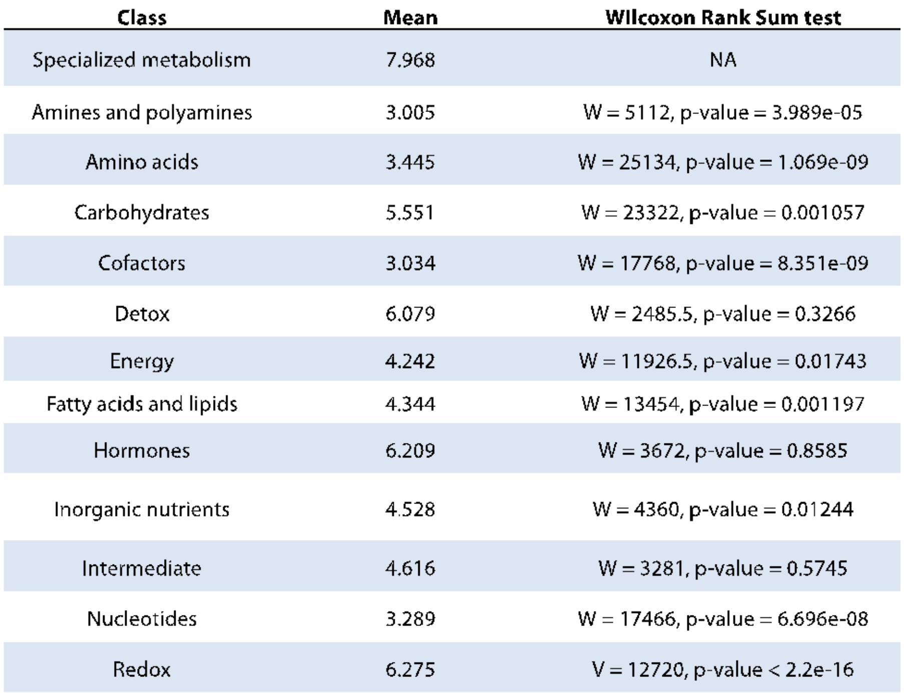
Table S2.



Species	Proteins	Enzymes	Network nodes	Network links
Arabidopsis	35386	6746	1234	9342
Cassava	34151	7153	1275	10342
Chlamy	17114	1892	948	6182
Coccomyxa	9629	1734	1024	7100
Grapevine	26346	5319	1246	9523
Maize	63540	10417	1261	10055
Micromonas C	10660	1414	885	5487
Micromonas R	10103	1490	890	5497
Moss	38354	5649	1204	8881
Ostreoccocus	7796	1295	804	4708
Poplar	45033	8664	1273	10388
Rice	51258	8444	1243	10431
Selaginella	22285	3947	1191	9353
Sorghum	29448	6241	1230	9566
Soybean	55787	12124	1249	9688
Volvox	15286	1832	900	5788

Summary of inferred enzymes and metabolic networks. Protein figures represent the total number of gene coding sequences including splice variants, as annotated at Phytozome (www.phytozome.org). Enzyme numbers are the subset of proteins assigned to four-digit Enzyme Commission numbers by an in-house functional annotation pipeline shown in fig. S2. Nodes in the metabolic networks are the set of unique reactions associated to each species' enzyme inventory. Network links are the number of compounds connecting two nodes in the network.

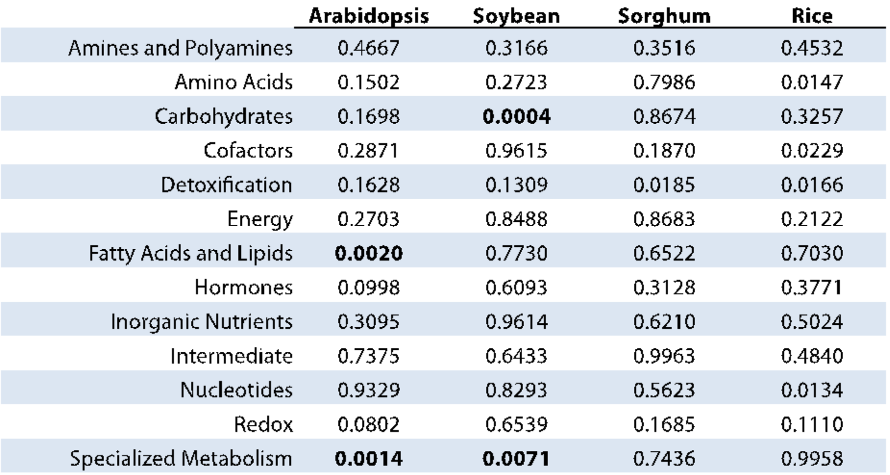
Table S3.



Class	Mean	WIlcoxon Rank Sum test
Specialized metabolism	7.968	NA
Amines and polyamines	3.005	W = 5112, p-value = 3.989e-05
Amino acids	3.445	W = 25134, p-value = 1.069e-09
Carbohydrates	5.551	W = 23322, p-value = 0.001057
Cofactors	3.034	W = 17768, p-value = 8.351e-09
Detox	6.079	W = 2485.5, p-value = 0.3266
Energy	4.242	W = 11926.5, p-value = 0.01743
Fatty acids and lipids	4.344	W = 13454, p-value = 0.001197
Hormones	6.209	W = 3672, p-value = 0.8585
Inorganic nutrients	4.528	W = 4360, p-value = 0.01244
Intermediate	4.616	W = 3281, p-value = 0.5745
Nucleotides	3.289	W = 17466, p-value = 6.696e-08
Redox	6.275	V = 12720, p-value < 2.2e-16

Comparisons of the average mean size of EC-based enzyme families belonging to each functional class versus specialized metabolism. Differences in the distribution of family sizes were assessed statistically using the Wilcoxon Rank Sum test.

Table S4.



	Arabidopsis	Soybean	Sorghum	Rice
Amines and Polyamines	0.4667	0.3166	0.3516	0.4532
Amino Acids	0.1502	0.2723	0.7986	0.0147
Carbohydrates	0.1698	**0.0004**	0.8674	0.3257
Cofactors	0.2871	0.9615	0.1870	0.0229
Detoxification	0.1628	0.1309	0.0185	0.0166
Energy	0.2703	0.8488	0.8683	0.2122
Fatty Acids and Lipids	**0.0020**	0.7730	0.6522	0.7030
Hormones	0.0998	0.6093	0.3128	0.3771
Inorganic Nutrients	0.3095	0.9614	0.6210	0.5024
Intermediate	0.7375	0.6433	0.9963	0.4840
Nucleotides	0.9329	0.8293	0.5623	0.0134
Redox	0.0802	0.6539	0.1685	0.1110
Specialized Metabolism	**0.0014**	**0.0071**	0.7436	0.9958

Enrichment of functional domains in four plant species. The table lists p-values associated with the enrichment of functional domains in clustered gene sets for four species. P-values were calculated using the hypergeometric test. P-values ≤ 0.01 in bold.

Additional Data Table S1 (separate file)

Sequence IDs, Enzyme Commission classes, and MetaCyc reactions annotated to the protein datasets for the 16 species in the study.

Additional Data Table S2 (separate file)

Mapping files for all reactions and pathways in the study, including common names, Enzyme Commission classes, functional classifications, and literature citations, where relevant.

Additional Data Table S3 (separate file)

Reactions unique to each lineage grouping, as well as their functional classifications and associated pathways.

Additional Data Table S4 (separate file)

Gene IDs and classifications for all genes involved in the local duplication and whole-genome duplication analyses.

Additional Data Table S5 (separate file)

Gene IDs and Enzyme Commission classes for all genes found clustered in their respective species, organized by cluster.

References

1. B. M. Schmidt, D. M. Ribnicky, P. E. Lipsky, I. Raskin, Revisiting the ancient concept of botanical therapeutics. *Nat. Chem. Biol.* **3**, 360–366 (2007). doi:10.1038/nchembio0707-360 Medline

2. P. R. Ehrlich, P. H. Raven, Butterflies and plants: A study in coevolution. *Evolution* **18**, 586–608 (1964). doi:10.2307/2406212

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

4. A. L. Schilmiller, E. Pichersky, R. L. Last, Taming the hydra of specialized metabolism: How systems biology and comparative approaches are revolutionizing plant biochemistry. *Curr. Opin. Plant Biol.* **15**, 338–344 (2012). doi:10.1016/j.pbi.2011.12.005 Medline

5. J.-K. Weng, R. N. Philippe, J. P. Noel, The rise of chemodiversity in plants. *Science* **336**, 1667–1670 (2012). doi:10.1126/science.1217411 Medline

6. See supplementary materials on *Science* Online.

7. A.-L. Barabási, Z. N. Oltvai, Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004). doi:10.1038/nrg1272 Medline

8. B. C. M. van Wijk, C. J. Stam, A. Daffertshofer, Comparing brain networks of different size and connectivity density using graph theory. *PLoS ONE* **5**, e13701 (2010). doi:10.1371/journal.pone.0013701 Medline

9. B. D. Mishler, Deep phylogenetic relationships among "plants" and their implications for classification. *Taxon* **49**, 661–683 (2000). doi:10.2307/1223970

10. J.-K. Weng, J. P. Noel, Chemodiversity in *Selaginella*: A reference system for parallel and convergent metabolic evolution in terrestrial plants. *Front. Plant Sci.* **4**, 119 (2013). doi:10.3389/fpls.2013.00119 Medline

11. D. Nelson, D. Werck-Reichhart, A P450-centric view of plant evolution. *Plant J.* **66**, 194–211 (2011). doi:10.1111/j.1365-313X.2011.04529.x Medline

12. K. Yonekura-Sakakibara, K. Hanada, An evolutionary view of functional diversity in family 1 glycosyltransferases. *Plant J.* **66**, 182–193 (2011). doi:10.1111/j.1365-313X.2011.04493.x Medline

13. R. C. Moore, M. D. Purugganan, The evolutionary dynamics of plant duplicate genes. *Curr. Opin. Plant Biol.* **8**, 122–128 (2005). doi:10.1016/j.pbi.2004.12.001 Medline

14. D. J. Kliebenstein, A role for gene duplication and natural variation of gene expression in the evolution of metabolism. *PLOS ONE* **3**, e1838 (2008). doi:10.1371/journal.pone.0001838

15. H. Y. Chu, E. Wegel, A. Osbourn, From hormones to secondary metabolism: The emergence of metabolic gene clusters in plants. *Plant J.* **66**, 66–79 (2011). doi:10.1111/j.1365-313X.2011.04503.x Medline

16. T. Obayashi, K. Nishida, K. Kasahara, K. Kinoshita, ATTED-II updates: Condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant Cell Physiol.* **52**, 213–219 (2011). doi:10.1093/pcp/pcq203 Medline

17. K. Yonekura-Sakakibara, K. Saito, Functional genomics for plant natural product biosynthesis. *Nat. Prod. Rep.* **26**, 1466–1487 (2009). doi:10.1039/b817077k Medline

18. S. S. Merchant, S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin, L. Maréchal-Drouard, W. F. Marshall, L. H. Qu, D. R. Nelson, A. A. Sanderfoot, M. H. Spalding, V. V. Kapitonov, Q. Ren, P. Ferris, E. Lindquist, H. Shapiro, S. M. Lucas, J. Grimwood, J. Schmutz, P. Cardol, H. Cerutti, G. Chanfreau, C. L. Chen, V. Cognat, M. T. Croft, R. Dent, S. Dutcher, E. Fernández, H. Fukuzawa, D. González-Ballester, D. González-Halphen, A. Hallmann, M. Hanikenne, M. Hippler, W. Inwood, K. Jabbari, M. Kalanon, R. Kuras, P. A. Lefebvre, S. D. Lemaire, A. V. Lobanov, M. Lohr, A. Manuell, I. Meier, L. Mets, M. Mittag, T. Mittelmeier, J. V. Moroney, J. Moseley, C. Napoli, A. M. Nedelcu, K. Niyogi, S. V. Novoselov, I. T. Paulsen, G. Pazour, S. Purton, J. P. Ral, D. M. Riaño-Pachón, W. Riekhof, L. Rymarquis, M. Schroda, D. Stern, J. Umen, R. Willows, N. Wilson, S. L. Zimmer, J. Allmer, J. Balk, K. Bisova, C. J. Chen, M. Elias, K. Gendler, C. Hauser, M. R. Lamb, H. Ledford, J. C. Long, J. Minagawa, M. D. Page, J. Pan, W. Pootakham, S.

Roje, A. Rose, E. Stahlberg, A. M. Terauchi, P. Yang, S. Ball, C. Bowler, C. L. Dieckmann, V. N. Gladyshev, P. Green, R. Jorgensen, S. Mayfield, B. Mueller-Roeber, S. Rajamani, R. T. Sayre, P. Brokstein, I. Dubchak, D. Goodstein, L. Hornick, Y. W. Huang, J. Jhaveri, Y. Luo, D. Martínez, W. C. Ngau, B. Otillar, A. Poliakov, A. Porter, L. Szajkowski, G. Werner, K. Zhou, I. V. Grigoriev, D. S. Rokhsar, A. R. Grossman, The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–250 (2007). doi:10.1126/science.1143609 Medline

19. J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X. C. Zhang, K. Shinozaki, H. T. Nguyen, R. A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker, S. A. Jackson, Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010). doi:10.1038/nature08670 Medline

20. S. Prochnik, P. R. Marri, B. Desany, P. D. Rabinowicz, C. Kodira, M. Mohiuddin, F. Rodriguez, C. Fauquet, J. Tohme, T. Harkins, D. S. Rokhsar, S. Rounsley, The cassava genome: Current progress, future directions. *Trop. Plant Biol.* **5**, 88–94 (2012). doi:10.1007/s12042-011-9088-z Medline

21. S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, C. R. Buell, The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res.* **35**, D883–D887 (2007). doi:10.1093/nar/gkl976 Medline

22. G. A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G. L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Déjardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze,

Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjärvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J. C. Leplé, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouzé, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C. J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer, D. Rokhsar, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006). doi:10.1126/science.1128691 Medline

23. J. A. Banks, T. Nishiyama, M. Hasebe, J. L. Bowman, M. Gribskov, C. dePamphilis, V. A. Albert, N. Aono, T. Aoyama, B. A. Ambrose, N. W. Ashton, M. J. Axtell, E. Barker, M. S. Barker, J. L. Bennetzen, N. D. Bonawitz, C. Chapple, C. Cheng, L. G. Correa, M. Dacre, J. DeBarry, I. Dreyer, M. Elias, E. M. Engstrom, M. Estelle, L. Feng, C. Finet, S. K. Floyd, W. B. Frommer, T. Fujita, L. Gramzow, M. Gutensohn, J. Harholt, M. Hattori, A. Heyl, T. Hirai, Y. Hiwatashi, M. Ishikawa, M. Iwata, K. G. Karol, B. Koehler, U. Kolukisaoglu, M. Kubo, T. Kurata, S. Lalonde, K. Li, Y. Li, A. Litt, E. Lyons, G. Manning, T. Maruyama, T. P. Michael, K. Mikami, S. Miyazaki, S. Morinaga, T. Murata, B. Mueller-Roeber, D. R. Nelson, M. Obara, Y. Oguri, R. G. Olmstead, N. Onodera, B. L. Petersen, B. Pils, M. Prigge, S. A. Rensing, D. M. Riaño-Pachón, A. W. Roberts, Y. Sato, H. V. Scheller, B. Schulz, C. Schulz, E. V. Shakirov, N. Shibagaki, N. Shinohara, D. E. Shippen, I. Sørensen, R. Sotooka, N. Sugimoto, M. Sugita, N. Sumikawa, M. Tanurdzic, G. Theissen, P. Ulvskov, S. Wakazuki, J. K. Weng, W. W. Willats, D. Wipf, P. G. Wolf, L. Yang, A. D. Zimmer, Q. Zhu, T. Mitros, U. Hellsten, D. Loqué, R. Otillar, A. Salamov, J. Schmutz, H. Shapiro, E. Lindquist, S. Lucas, D. Rokhsar, I. V. Grigoriev, The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960–963 (2011). doi:10.1126/science.1203810 Medline

24. O. Jaillon, J. M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E.

Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pè, G. Valle, M. Morgante, M. Caboche, A. F. Adam-Blondon, J. Weissenbach, F. Quétier, P. Wincker; French-Italian Public Consortium for Grapevine Genome Characterization, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007). doi:10.1038/nature06148 Medline

25. A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V. Grigoriev, E. Lyons, C. A. Maher, M. Martis, A. Narechania, R. P. Otillar, B. W. Penning, A. A. Salamov, Y. Wang, L. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C. McCann, R. Ming, D. G. Peterson, D. Mehboob-ur-Rahman, P. Ware, K. F. Westhoff, J. Mayer, D. S. Messing, Rokhsar, The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009). doi:10.1038/nature07723 Medline

26. S. E. Prochnik, J. Umen, A. M. Nedelcu, A. Hallmann, S. M. Miller, I. Nishii, P. Ferris, A. Kuo, T. Mitros, L. K. Fritz-Laylin, U. Hellsten, J. Chapman, O. Simakov, S. A. Rensing, A. Terry, J. Pangilinan, V. Kapitonov, J. Jurka, A. Salamov, H. Shapiro, J. Schmutz, J. Grimwood, E. Lindquist, S. Lucas, I. V. Grigoriev, R. Schmitt, D. Kirk, D. S. Rokhsar, Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* **329**, 223–226 (2010). doi:10.1126/science.1188800 Medline

27. G. Blanc, I. Agarkova, J. Grimwood, A. Kuo, A. Brueggeman, D. D. Dunigan, J. Gurnon, I. Ladunga, E. Lindquist, S. Lucas, J. Pangilinan, T. Pröschold, A. Salamov, J. Schmutz, D. Weeks, T. Yamada, A. Lomsadze, M. Borodovsky, J. M. Claverie, I. V. Grigoriev, J. L. Van Etten, The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* **13**, R39 (2012). doi:10.1186/gb-2012-13-5-r39 Medline

28. A. Z. Worden, J. H. Lee, T. Mock, P. Rouzé, M. P. Simmons, A. L. Aerts, A. E. Allen, M. L. Cuvelier, E. Derelle, M. V. Everett, E. Foulon, J. Grimwood, H. Gundlach, B. Henrissat, C. Napoli, S. M. McDonald, M. S. Parker, S. Rombauts, A. Salamov, P. Von Dassow, J. H. Badger, P. M. Coutinho, E. Demir, I. Dubchak, C. Gentemann, W. Eikrem, J. E. Gready, U. John, W. Lanier, E. A. Lindquist, S. Lucas, K. F. Mayer, H. Moreau, F. Not, R. Otillar, O. Panaud, J. Pangilinan, I. Paulsen, B. Piegu, A. Poliakov, S. Robbens, J. Schmutz, E. Toulza, T. Wyss, A. Zelensky, K. Zhou, E. V. Armbrust, D. Bhattacharya, U. W. Goodenough, Y. Van de Peer, I. V. Grigoriev, Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**, 268–272 (2009). doi:10.1126/science.1167222 Medline

29. B. Palenik, J. Grimwood, A. Aerts, P. Rouzé, A. Salamov, N. Putnam, C. Dupont, R. Jorgensen, E. Derelle, S. Rombauts, K. Zhou, R. Otillar, S. S. Merchant, S. Podell, T. Gaasterland, C. Napoli, K. Gendler, A. Manuell, V. Tai, O. Vallon, G. Piganeau, S. Jancek, M. Heijde, K. Jabbari, C. Bowler, M. Lohr, S. Robbens, G. Werner, I. Dubchak, G. J. Pazour, Q. Ren, I. Paulsen, C. Delwiche, J. Schmutz, D. Rokhsar, Y. Van de Peer, H. Moreau, I. V. Grigoriev, The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7705–7710 (2007). doi:10.1073/pnas.0611046104 Medline

30. P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga,

M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C. T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A. P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J. M. Chia, J. M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, R. K. Wilson, The B73 maize genome: Complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009). doi:10.1126/science.1178534 Medline

31. D. Swarbreck, C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang, E. Huala, The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014 (2008). doi:10.1093/nar/gkm965 Medline

32. G. Parra, K. Bradnam, I. Korf, CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007). doi:10.1093/bioinformatics/btm071 Medline

33. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990). Medline

34. C. Claudel-Renard, C. Chevalet, T. Faraut, D. Kahn, Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* **31**, 6633–6639 (2003). doi:10.1093/nar/gkg847 Medline

35. C. Yu, N. Zavaljevski, V. Desai, J. Reifman, Genome-wide enzyme annotation with precision control: Catalytic families (CatFam) databases. *Proteins Struct. Funct. Bioinform.* **74**, 449–460 (2009). doi:10.1002/prot.22167 Medline

36. K. Tipton, S. Boyce, History of the enzyme nomenclature system. *Bioinformatics* **16**, 34–40 (2000). doi:10.1093/bioinformatics/16.1.34 Medline

37. T. G. Dietterich, in *Multiple Classifier Systems* (Springer-Verlag, Berlin, 2000), pp. 1–15.

38. UniProt Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40**, D71–D75 (2012). doi:10.1093/nar/gkr981 Medline

39. B. Efron, Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Stat. Assoc.* **78**, 316–331 (1983). doi:10.1080/01621459.1983.10477973

40. P. Zhang, K. Dreher, A. Karthikeyan, A. Chi, A. Pujar, R. Caspi, P. Karp, V. Kirkup, M. Latendresse, C. Lee, L. A. Mueller, R. Muller, S. Y. Rhee, 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 (2010). doi:10.1104/pp.110.157396 Medline

41. R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, P. D. Karp, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **40**, D742–D753 (2012). doi:10.1093/nar/gkr1014 Medline

42. K. D. Verkhedkar, K. Raman, N. R. Chandra, S. Vishveshwara, Metabolome based reaction graphs of *M. tuberculosis* and *M. leprae*: A comparative network analysis. *PLOS ONE* **2**, e881 (2007). doi:10.1371/journal.pone.0000881 Medline

43. J. I. Fuxman Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, A. J. Walhout, Using networks to measure similarity between genes: Association index selection. *Nat. Methods* **10**, 1169–1176 (2013). doi:10.1038/nmeth.2728 Medline

44. R. Suzuki, H. Shimodaira, Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006). doi:10.1093/bioinformatics/btl117 Medline

45. S. S. Stevens, On the psychophysical law. *Psychol. Rev.* **64**, 153–181 (1957). doi:10.1037/h0046162 Medline

46. T.-H. Lee, H. Tang, X. Wang, A. H. Paterson, PGDD: A database of gene and genome duplication in plants. *Nucleic Acids Res.* **41**, D1152–D1158 (2013). 10.1093/nar/gks1104 Medline

47. C. Rizzon, L. Ponger, B. S. Gaut, Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLOS Comput. Biol.* **2**, e115 (2006). doi:10.1371/journal.pcbi.0020115 Medline

48. K. Hanada, C. Zou, M. D. Lehti-Shiu, K. Shinozaki, S.-H. Shiu, Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* **148**, 993–1003 (2008). doi:10.1104/pp.108.122457 Medline

49. I. Letunic, P. Bork, Interactive Tree Of Life v2: Online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39** (suppl.), W475–W478 (2011). doi:10.1093/nar/gkr201 Medline

50. M. Frey, K. Schullehner, R. Dick, A. Fiesselmann, A. Gierl, Benzoxazinoid biosynthesis, a model for evolution of secondary metabolic pathways in plants. *Phytochemistry* **70**, 1645–1651 (2009). doi:10.1016/j.phytochem.2009.05.012 Medline

51. O. Lapčík, Isoflavonoids in non-leguminous taxa: A rarity or a rule? *Phytochemistry* **68**, 2909–2916 (2007). doi:10.1016/j.phytochem.2007.08.006 Medline