Supporting Information

Seaver et al. 10.1073/pnas.1401329111

SI Text

PlantSEED Subsystems. PlantSEED organizes genomic data in a form of populated subsystems. A subsystem consists of a group of functional roles (isofunctional protein families) that jointly are involved in a specific biological process or a structural complex across a large set of genomes. The focal organism was *Arabidopsis*, and the evidence for the gene–reaction associations built into PlantSEED was taken from AraCyc, the Arabidopsis Information Resource (TAIR) (1), and the literature. For instance, the subsystem named "Alanine, serine, glycine metabolism in plants" incorporated the reactions and genes of six base AraCyc pathways involved in the biosynthesis of alanine, serine, and glycine. The relationship between the subsystems and the metabolic pathways that they incorporate is listed in detail in Dataset S4.

In the area of membrane energetics, subsystems are encoded to capture each individual polypeptide complex of the respiratory or photosynthetic electron chain, grouping all known structural components required for its function. Analogous complexes residing in different organelles are encoded in separate subsystems, e.g., F0F1-type ATP synthase in plants (plastidial) versus F0F1-type ATP synthase in plants (mitochondrial) and V-Type ATP synthase in plants (vacuolar).

In PlantSEED, individual complexes are not grouped into the linear pathways, such as "respiration" or "photosynthesis," with which they historically have been associated. Membrane potential and the redox state of each organelle are maintained by a complex network of multiple electron donor and electron acceptor reactions, and the relative contribution of these reactions shifts in response to growth conditions (2). The current state-of-the-art representation of these reactions in models and other on-line resources is far from being accurate or complete. For example, the thylakoid membrane harbors not only the canonical photosystem II, photosystem I, and cytochrome b6-f complexes but also NADH dehydrogenase-like complex, terminal plasmid oxidase, and ATP synthase (3). It is conceivable that photosynthetic and chlororespiration electron fluxes are intertwined in the thylakoid membrane (4), as they are in cyanobacteria (5). In additional, some metabolic pathways (e.g., carotenoid desaturation) are believed to be associated directly with the electron transport chain (6). Hence, encoding the chloroplast photosynthesis light reactions and respiration pathways in isolation [e.g., the photosynthesis light reactions (PWY-101) in AraCyc] is an oversimplification.

In PlantSEED an attempt was made to represent the mitochondrial and chloroplast electron transfer systems as a collection of membrane complexes and soluble components, each encoded in detail in a separate subsystem. We believe this approach (*i*) accounts for all known proteins/genes required to catalyze each reaction, (*ii*) explicitly conveys the fact that both organelles—mitochondria and chloroplasts—possess independent electron transfer chains and generate metabolic energy via chemiosmotic ATP synthesis, and (*iii*) constitutes an important, albeit small, step in the direction of a true and comprehensive encoding of this crucial area of plant physiology and metabolism. It allows "the parts to be strung together" in a more flexible and accurate way in future metabolic models and genome annotation projects.

PlantSEED Metabolic Reconstruction. ***Compartmentalization.*** The assignment of reactions to individual organelles and other distinct areas in a plant cell was made based on three sources of data: AraCyc, the Plant Proteomics Database (PPDB) (7), and a review of maize B-vitamin pathways (8). In every case, the data consisted of a set of genes whose protein products were determined to be targeted to different organelles and compartments. In the case of the PPDB, we used only localization data that were curated and thus used none of the predictions. It follows that the reactions catalyzed by these targeted proteins were assigned to the same compartment. In addition, AraCyc also contained compartmental data for individual reactions, which we used. There are nine compartments in the PlantSEED models; the numbers of genes and reactions assigned to different compartments are detailed in Table S2.

Media. For each metabolic reconstruction, we use different media for autotrophic and heterotrophic growth. Both media contain water, oxygen, carbon dioxide, ammonia, phosphate, sulfate, protons, nitrate, and magnesium. The heterotrophic medium contains sucrose, and the autotrophic medium contains light as an abstract compound.

High-quality maize biomass. Recent publications of metabolic models included basic biomass equations that generally were assumed to represent leaf tissue. The publications explored basic concepts of plant metabolism such as photosynthesis and phenylpropanoid biosynthesis, and, as such, very little consideration was placed on the diversity of compounds that a plant biosynthesizes. Here, anticipating that the exploration of plant metabolic models will proliferate, we describe a high-quality determination of their relative compositions as supported by an extensive literature search. The most recent publications of metabolic models derived much of the data either from the previous model (9) or from a small number of references (10). Here we describe a biomass that contains more cofactors, hormones, and fatty-acids, including detailed quantities, supported by almost 30 literature references, (Dataset S2). The following paragraphs briefly describe the biomass composition and identify the relevant references.

Amino acids. The fraction of the biomass attributable to protein is estimated to be 20% (11). To quantify how the individual amino acids, in turn, make up the fraction of the biomass fraction that is protein, we use the amino acid compositions of the small and large subunit of Rubisco, phosphoenolpyruvate carboxylase, and pyruvate phosphate dikinase, which respectively represent 35%, 8%, and 6% of the total protein in maize grown under optimal conditions (12). Water loss caused by the formation of the peptide bond was taken into account.

Nucleic acids. The biomass fraction attributable to nucleic acids is estimated to be 0.69% (11) and is split equally between RNA and DNA. The biomass fraction attributed to each nucleic acid in particular was calculated using the Guanine-Cytosine content published by Haberer et al. (13).

Carbohydrates. The biomass fraction attributable to carbohydrates is estimated to be 56.5%, the majority of which (80%) is cellulose and hemicellulose (11). For most of the monosaccharides, the biomass is calculated from the numbers directly available in ref. 11 and from an estimation of the fraction of which they make up the major polymers (14, 15). For galactose, glycerol, and sulfoquinovose, the biomass fraction is estimated using the numbers published for galactolipids, glycerolipids, and sulfolipids, respectively (see the discussion below of lipids and sterols). Finally, further evidence was used to deduce the biomass fraction of inositol (16) and galacturonate (17).

Phenolic compounds. The cell wall of maize is considered to be made up largely of two types of phenolic derivatives: *p*-coumaric acid and ferulic acid (18).

Vitamins and cofactors. We place more emphasize than the published metabolic models on biosynthetic pathways of cofactors as key components of metabolism and specify the biomass fraction assigned to each of the B vitamins and other cofactors with greater accuracy. The list of vitamins and cofactors includes biotin, NAD and derivatives, pyridoxal-5′-phosphate, FAD and FMN, CoA, phosphopantetheine, tetrahydrofolate and its derivatives, α-tocopherol, ascorbate, ubiquinone-9, and lipoic acid (19–30).

Pigments. We include three pigments in our biomass as important components of photosynthesis: chlorophyll, carotene, and lutein (26).

Lipids and sterols. For simplicity, the unsaturated and saturated fatty acids were limited to 16 and 18 carbons in length, respectively (31). We include three of the most common sterols—sitosterol, campesterol, and stigmasterol (32)—and also include phytosphingosine (33).

Hormones. Hormones are essential for plant development and are found in sufficient quantities to warrant inclusion in plant biomass. We include here the hormones indole-3-acetic acid and zeatin (34).

Carboxylic acids and other compounds. The plant biomass comprises many other compounds, and we note here a subset of these compounds for which a value is reported in the literature: *cis*-aconitate, citrate, malate, and oxaloacetate (11), lactate (35), and *S*-adenosylmethionine (36). Choline and ethanolamine were estimated from the values for phosphatidylcholine and phosphatidylethanolamine, respectively. Finally, the mineral content of the biomass was set at 5%, split evenly between potassium and chloride (11).

Analysis and Comparison of Plant Metabolic Models. Comparison of the subsystems annotations and models generated for the 10 reference genomes showed, as expected, high overall similarity, somewhat greater within eudicots and monocots than between these groups, with primary metabolism largely conserved (Table 1 and Fig. 4). The most complete annotation and model are for *Arabidopsis*, the best-documented plant genome and our template for subsystems design and curation, containing 2,098 protein families and 1,078 reactions. For comparison, some previously published *Arabidopsis* models include significantly more proteins and reactions (Table S1); however, only a fraction of their content is supported by experimental and/or genome evidence. For example, the *Arabidopsis* (iRS1597) and maize (iRS1563) models contain 1,798 and 1,985 reactions, respectively (9). However, only 42% of these reactions are supported by "evidence for the participation" [sic]. In contrast, the PlantSEED database aims to establish the highest possible accuracy of reaction–gene associations and to include in models only reactions supported by genome evidence, experimental evidence, or both. These criteria are essential prerequisites for the productive integration of gene expression, protein localization, and other -omics data into plant models and are vital for the advancement of plant metabolic modeling. We use as a benchmark the number of reactions in *Arabidopsis* that currently are associated with pathways and enzymes with experimental evidence—1,049 (37)—and strive to minimize, not inflate, the number of unsupported reactions in all PlantSEED models.

The conserved core of all of the plant genomes in our subsystems consists of 895 protein families and 746 reactions. Thus, 69% of the reactions in the *Arabidopsis* model are found in all 10 plant species. Furthermore, we find that 94% of all of the reactions curated for *Arabidopsis* are propagated to other plants with varying degrees of success. There is some taxon-specificity to this propagation, in that 90% of the reactions are propagated among eudicots and 80% are propagated among monocots. Among the reactions that do not propagate between *Arabidopsis* and other species (with the exception of *A. lyrata*) are several from glucosinolate biosynthesis. (The gapfilling process did restore many reactions and annotations that initially failed to propagate to additional plant genomes based on evidence of partially annotated pathways, but the glucosinolate biosynthesis pathway was not among the restored pathways.) This outcome validates our annotation propagation process, because glucosinolates are known to be confined to the order Brassicales (38).

We compared the gene–reaction associations for *Arabidopsis*, *Populus trichocarpa*, and maize in the PlantSEED database with those in the BioCyc databases AraCyc, PoplarCyc, and MaizeCyc (Dataset S4). The BioCyc databases include more gene–reaction associations than their PlantSEED counterparts, although our curation reveals that many of these additional associations are incorrectly mapped homologs. Relative to PlantSEED, an average of 4.5, 15.8, and 33.6 extra gene–reaction associations were present per pathway in AraCyc, PoplarCyc, and MaizeCyc, respectively. Overall, although PlantSEED may be less comprehensive than AraCyc, its data are substantially more precise.

Discussion of Pathway-Based Annotation and Gapfilling. Because of a long history of biochemical research, many metabolic pathways are well understood in terms of the reactions that occur and the compounds that are consumed or produced. In stark contrast, many enzymes that are responsible for the catalytic activity within a pathway have yet to be characterized. The missing enzymes are referred to as "pathway holes," and various methods have been used to fill such holes, notably one frequently used in the building of BioCyc databases (39). The entire annotation pipeline used in that process is pathway-based, and whole pathways are included in a newly generated BioCyc database if a gene is predicted to be associated with one or more enclosed reactions. However, many reactions are included that do not have an associated gene. For example, the list of pathways used for plants is maintained by Plant Metabolic Networks (www.plantcyc.org/about/savi_pipeline.faces). This lack of gene-associated reactions in turn leads to many pathway holes that enzymes can be predicted to fill. However, this leads to several problems, such as the inclusion of duplicate reactions from many closely related pathways with different annotation. For example, the reaction catalyzed by malate dehydrogenase [EC (Enzyme Commission) 1.1.1.37] has two instances; the first (MALATE-DEH-RXN) is associated with seven pathways, whereas the second instance (MAL-DEH-GLYOX-RXN) is associated with one pathway. None of the pathways overlap, but the two instances share some of the same genes, so that that both reactions could be added the BioCyc database for any newly added plant species. Pathway-based annotation also can lead to an overly aggressive addition of pathways to a reconstruction if the threshold for pathway evidence is too low. This occurs if a reaction is associated with many different pathways, and all of the unannotated reactions in each of these pathways are added to the new database. Conversely, the genes associated with the reaction do not necessarily guarantee the presence of all the associated pathways in the organism of study.

Compartmental Transport in Models. Although many reactions are localized to the compartments, many essential transport reactions are not included in the initial drafts of the metabolic models because the annotation of specific transporters is very limited. Even if transporters are annotated, the annotation frequently refers to multiple substrates. For example, the transporter class 2.A.1.1.14 from the Transporter Classification database (40) is described as a "Hexose:H^+ symporter." Representing this transporter in a metabolic model would require the use of multiple reactions, each describing a specific hexose. This process was performed manually by the groups that published the plant models that we integrated into the PlantSEED biochemistry. Therefore we have available the correct set of transporters for

the exploring gapfilling algorithm, and both the essential and pathway gapfilling accommodated the problem of transporter annotation by activating 58 localized enzymatic reactions by adding 20 transport reactions (Dataset S3). The list of transporters in the PlantSEED biochemistry and their origins are given in Dataset S1.

The Power of PlantSEED Annotation and Tools. The hypothetical protein family COG1836 (DUF92) was predicted to be a candidate for the missing role of phytyl-phosphate kinase by using plant-prokaryote cross-kingdom comparative genomics within SEED and PlantSEED. The evidence supporting this prediction is as follows:

i) Conserved gene clustering in prokaryotic genomes. The members of the COG1836 family cluster strongly with various genes of polyprenoid metabolism in Eubacteria and Archaea, including isopentenyl phosphate kinase, polyprenyl pyrophosphate synthetase (EC 2.5.1.-), geranylgeranyl diphosphate reductase (EC 1.3.1.83), phytol kinase (EC 2.7.1.-), and others. Typical examples are shown in Fig. 4*A*.

ii) Domain fusion events. Association with phytol kinase (EC 2.7.1.-) is especially informative: Members of the COG1836 family are colocated on the chromosome with homologs of phytol kinase (e.g., in *Anabaena* and *Nostoc*; Fig. 4*A*) and even fused with them into a single polypeptide, e.g., in the green sulfur bacteria *Chlorobium*, *Chlorobaculum*, *Pelodictyon*, and *Prosthecochloris* and in phototrophic Firmicutes such as *Heliobacterium modesticaldum*.

iii) Phylogenetic profiling. The members of the two families tend to co-occur in the same genomes, providing an additional clue indicating a potential functional association between the COG1836 family and phytol kinase. Both families are commonly present in photosynthetic organisms (oxygenic as well as anoxygenic) and extremophiles (including many Archaea), in which the need to salvage and recycle phytol (or a similar lipid moiety) might be especially pronounced.

These observations are summarized in the PubSEED Subsystem COG1836 at http://pubseed.theseed.org/SubsysEditor.cgi?page=ShowSubsystem&subsystem=COG1836.

The members of the COG1836 family have no recognizable domains or motifs indicative of a specific function and currently are annotated merely as "hypothetical membrane proteins" DUF92 in all public on-line resources [including the National Center for Biotechnology Information, TAIR, the Kyoto Encyclopedia of Genes and Genomics (KEGG), and Phytozome] (Fig. 4*B*).

The PlantSEED Gateway. The genomes, metabolic models, and subsystems are accessible through the PlantSEED gateway: http://plantseed.theseed.org (Fig. S3). The front page of this gateway contains a summary table of reference plant genomes and models that allows users to download these data and access web views of the genomes and models. The genome viewer shows all plant genes in a tabular format. The annotation viewer for individual genes provides a comparative genomics view, showing similar genes and chromosomal regions in other plant genomes and in microbial genomes containing close homologs of a queried plant gene. The model viewer permits the selection and comparison of multiple plant models, enabling painting of multiple models in individual pathway maps and permitting side-by-side comparison in reaction, compound, and gene tables. The second page of the gateway shows the current list of PlantSEED subsystems, permitting access to a subsystem viewer showing the details of subsystem annotations in all reference plant genomes in a spreadsheet format, so one can pinpoint which species can perform which metabolic functions.

The individual genome pages of the PlantSEED database provide a unique "compare regions" view of plant genomes, permitting simultaneous viewing of a gene in a reference genome and its flanking genes along with a representative set of plant and bacterial homologs (Fig. 4 and Fig. S4). This view is an invaluable resource for plant annotation, because it allows a plant annotator attempting to predict the function of a plant gene to exploit the functional clustering that occurs in bacterial genomes.

The web gateway for PlantSEED is an essential component of this resource, because PlantSEED is not static. Over time, the number of reference genomes in PlantSEED will grow. The number of curated subsystems will grow also as ongoing curation integrates new literature, experimental data, and reference data. New nonmetabolic subsystems will be added to extend the scope of curated PlatntSEED annotations. Similarly, more reactions and compounds will be added to the biochemistry database, and this new chemistry will be integrated into subsystems and metabolic models. As new plant models are published, these models will be integrated into PlantSEED to facilitate comparison with the automated models generated by PlantSEED. Finally, over time, new model analysis tools will be integrated to support user-driven gapfilling analysis, flux balance analysis, and model-based analysis of transcriptomics, metabolomic, and proteomic datasets. Currently, many of these analytical tools are available already when PlantSEED models are exported in the Department of Energy Knowledgebase system (www.kbase.us). The PlantSEED includes tools to complete this export process seamlessly.

PlantSEED Biochemistry and Pathways. The PlantSEED biochemistry database was built on the ModelSEED database (41). The ModelSEED biochemistry consists of KEGG and several published microbial metabolic models. For the PlantSEED biochemistry, this database was expanded to include several BioCyc databases and several published plant metabolic models. The full list of external data sources included in this current release of the PlantSEED and their list of compounds and reactions are given in Table S1, and the lists of compounds and reactions are in Dataset S1.

The biochemistry database is developed in several stages, and sets of data sources are handled differently. For each stage, the compounds in a data source are matched with the core PlantSEED biochemistry first, and then the reactions are matched. Each reaction is matched based on the reagents included in the reaction, except for protons; therefore in building this database it is essential that care be taken in matching compounds. Because of the variety of synonyms for many small organic compounds, particularly those with a chiral center, matching compounds by name is no small feat. Instead, we build the core of our biochemistry database using the International Chemical Identifier (InChI) string format, a canonical representation of the compound's molecular structure that also includes stereochemistry and charge (42). The two data sources for which the largest number of mol files were available were KEGG (14,810 mol files downloaded on August 8/30/12) and MetaCyc (9,655 mol files downloaded with version 16.1), and therefore we built the core biochemistry based on these two databases.

To integrate the two core databases, an InChI string is generated for each mol file using the Marvin Beans software developed by ChemAxon (www.chemaxon.com/products/marvin/). If two InChI strings from each database match, then the two compounds are considered to be identical. Two aspects of chemical structure that are not rendered consistently in InChI strings hinder our ability to match them correctly: (*i*) protonation and (*ii*) radicals. To address these issues we (*i*) remove the proton sublayer (labeled 'p' in the InChI format), and (*ii*) append the string "rad" to any InChI string for a molecular structure that contains radicals. We ignore any mol files assigned to an abstracted compound "class" in MetaCyc (~220 mol files). In these cases, the mol files are used as representatives of the

class rather than as actual compounds, and we thus avoid matching a class with any compound in KEGG. Finally, before integrating the two databases, we consolidate the databases independently to find duplicate compounds and reactions in the databases.

When performing the integration of the two databases, we focus first on the core of compounds for which we are able to generate an InChI string. Then, for the remaining compounds with no predetermined molecular structure, we attempt an integration based on the compound names. A total of 5,116 compounds were found to be identical in KEGG (35% of KEGG compounds with mol structures) and MetaCyc (54% of MetaCyc compounds with mol structures). A further 157 compounds were matched based on name alone. The high percentage of integrated compounds (97%) being linked to canonical chemical structure is a strong incentive to use PlantSEED biochemistry for metabolic reconstructions.

Once compounds are integrated, the reactions are integrated based on reagents alone, avoiding any dependency on the wide variation exhibited by reaction names, with a single exception: We ignore protons. A total of 4,026 reactions are matched in KEGG (43% of reactions in KEGG) and MetaCyc (27% of reactions in MetaCyc). Of the matched reactions, 3,956 (98%) contain only reagents that are matched based on canonical InChI strings.

Special exceptions are made for two types of compounds in KEGG and MetaCyc, respectively, that hinder matching between the two databases. In the first case, for some sets of stereoisomers, KEGG has a "neutral" compound that represents a mixture of the stereoisomers. For example, KEGG contains the stereoisomers of glucose: α-D-glucose (C00267) and β-D-glucose (C00221) and also a stereochemically neutral version of D-glucose (C00031). Conversely, MetaCyc contains only α-D-glucose (ALPHA-GLUCOSE) and β-D-glucose (GLC). Integrating these two databases would lead to a match between the correct stereoisomers but also would include a third glucose compound, which is undesirable. We therefore use the stereochemistry encoded in the canonical InChI strings to detect compounds that have the same structure but with less or no recorded stereochemistry, and we exclude such compounds and their reactions from the database. Furthermore, we clone the excluded reactions and replace the neutral compounds with their stereoisomers.

In the second case, PathwayTools allows the use of compound classes in BioCyc databases. A compound class is an abstract representation of a particular collection of compounds. For example, the compound class Ubiquinones represents four different instances of a ubiquinone. A compound class cannot be assigned any chemical structure and therefore cannot be matched with a KEGG compound or even included in any metabolic model. However, many MetaCyc reactions contain compound classes, indicating that a variety of reagents and products may be used in a single reaction. In most cases, these reactions are not usable for a metabolic model. We therefore do not allow any compound class in any BioCyc database to match any other compound. In addition, for every compound class, we iterate through every instance of each compound class found within a single reaction and generate new stoichiometrically balanced reactions containing the appropriate instance. An example is NADH dehydrogenase (NADH-DEHYDROG-A-RXN) which contains the compound classes Ubiquinols and Ubiquinones. This reaction is cloned twice, to include either ubiquinol-8 and ubiquinone-8 or ubiquinol-9 and ubiquinone-9 as the appropriate pair of instances from their respective classes.

In metabolic flux analysis (MFA) a global mass-balance constraint must be applied, so every reaction that is included in an MFA model must be balanced consistently in terms of stoichiometry and charge. Although the formula of a compound is defined within a canonical InChI string, identical compounds originating from different databases may vary in their protonation state as recorded in their separate mol files. To balance the protons in every reaction and also the corresponding charge effectively, all compounds with a chemical structure are charged to a pH of 7 using Marvin Beans. Several groups of compounds are not charged effectively in this manner, namely compounds that have multiple components, compounds that have an R group (or another letter representing unknown structure), and compounds that are polymeric (in that they contain an indeterminate number of structural repeats). We use the fully protonated state for these compounds, where possible.

Finally, all reactions, when imported, are set by default to be thermodynamically reversible. To improve the solution computed for any MFA model so that it closely represents what happens in vivo, we compute whether each reaction is thermodynamically feasible in both directions. We use the group contribution method (43, 44) to compute the Gibbs free energy of reaction along with a set of heuristics previously designed to improve a model's accuracy (45) to determine the thermodynamic feasibility of a reaction's direction.

1. Lamesch P, et al. (2012) The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res* 40(Database issue):D1202–D1210.
2. Cooper GM Hausman RE (2013) *The Cell: A Molecular Approach* (Sinauer Associates, Sunderland, MA) 6th Ed.
3. Aluru MR, Rodermel SR (2004) Control of chloroplast redox by the IMMUTANS terminal oxidase. *Physiol Plant* 120(1):4–11.
4. Shikanai T (2007) Cyclic electron transport around photosystem I: Genetic approaches. *Annu Rev Plant Biol* 58:199–217.
5. Battchikova N, Eisenhut M, Aro EM (2011) Cyanobacterial NDH-1 complexes: Novel insights and remaining puzzles. *Biochim Biophys Acta* 1807(8):935–944.
6. Carol P, Kuntz M (2001) A plastid terminal oxidase comes to light: Implications for carotenoid biosynthesis and chlororespiration. *Trends Plant Sci* 6(1):31–36.
7. Sun Q, et al. (2009) PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res* 37(Database issue):D969–D974.
8. Gerdes S, et al. (2012) Plant B vitamin pathways and their compartmentation: A guide for the perplexed. *J Exp Bot* 63(15):5379–5395.
9. Saha R, Suthers PF, Maranas CD (2011) Zea mays iRS1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE* 6(7):e21784.
10. Mintz-Oron S, et al. (2012) Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proc Natl Acad Sci USA* 109(1):339–344.
11. Penning de Vries FW, Brunsting AH, van Laar HH (1974) Products, requirements and efficiency of biosynthesis: A quantitative approach. *J Theor Biol* 45(2):339–377.
12. Sugiyama T, Mizuno M, Hayashi M (1984) Partitioning of nitrogen among ribulose-1,5-bisphosphate carboxylase/oxygenase, phosphoenolpyruvate carboxylase, and pyruvate orthophosphate dikinase as related to biomass productivity in maize seedlings. *Plant Physiol* 75(3):665–669.
13. Haberer G, et al. (2005) Structure and architecture of the maize genome. *Plant Physiol* 139(4):1612–1624.
14. Buchala AJ, Wilkie KCB (1973) Uronic acid residues in the total hemicelluloses of oats. *Phytochemistry* 12(3):655–659.
15. Buchala AJ, Wilkie KC (1970) Non-endospermic hemicellulosic beta-glucans from cereals. *Naturwissenschaften* 57(10):496.
16. Gibbins LN, Norris FW (1963) Vitamins in germination. Distribution of inositol during the germination of the dwarf bean, Phaseolus vulgaris. *Biochem J* 86:64–67.
17. Wiethölter N, et al. (2003) Isolation and characterisation of the homogalacturonan from type II cell walls of the commelinoid monocot wheat using HF-solvolysis. *Carbohydr Res* 338(5):423–431.
18. Abedon BG, Hatfield RD, Tracy WF (2006) Cell wall composition in juvenile and adult leaves of maize (Zea mays L.). *J Agric Food Chem* 54(11):3896–3900.
19. Baldet P, Alban C, Axiotis S, Douce R (1993) Localization of free and bound biotin in cells from green pea leaves. *Arch Biochem Biophys* 303(1):67–73.
20. Hampp R, Goller M, Füllgraf H (1984) Determination of compartmented metabolite pools by a combination of rapid fractionation of oat mesophyll protoplasts and enzymic cycling. *Plant Physiol* 75(4):1017–1021.
21. Huang S, Zeng H, Zhang J, Wei S, Huang L (2011) Interconversions of different forms of vitamin B6 in tobacco plants. *Phytochemistry* 72(17):2124–2129.
22. Cataldi TRI, Nardiello D, Carrara V, Ciriello R, Benedetto GED (2003) Assessment of riboflavin and flavin content in common food samples by capillary electrophoresis with laser-induced fluorescence detection. *Food Chem* 82(2):309–314.
23. Tumaney AW, Ohlrogge JB, Pollard M (2004) Acetyl coenzyme A concentrations in plant tissues. *J Plant Physiol* 161(4):485–488.
24. Battey JF, Ohlrogge JB (1990) Evolutionary and tissue-specific control of expression of multiple acyl-carrier protein isoforms in plants and bacteria. *Planta* 180(3):352–360.
25. Mehrshahi P, et al. (2010) Functional analysis of folate polyglutamylation and its essential role in plant metabolism and development. *Plant J* 64(2):267–279.
26. Leipner J, Fracheboud Y, Stamp P (2008) Acclimation by suboptimal growth temperature diminishes photooxidative damage in maize leaves. *Plant Cell Environ* 20(3):366–372.

27. Chen Z, Young TE, Ling J, Chang SC, Gallie DR (2003) Increasing vitamin C content of plants through enhanced ascorbate recycling. *Proc Natl Acad Sci USA* 100(6): 3525–3530.
28. Hu WH, et al. (2006) Different effects of chilling on respiration in leaves and roots of cucumber (Cucumis sativus). *Plant Physiol Biochem* 44(11-12):837–843.
29. Drozak A, Romanowska E (2006) Acclimation of mesophyll and bundle sheath chloroplasts of maize to different irradiances during growth. *Biochim Biophys Acta* 1757(11):1539–1546.
30. Pérez-López U, et al. (2010) Lipoic acid and redox status in barley plants subjected to salinity and elevated CO2. *Physiol Plant* 139(3):256–268.
31. Leech RM, Rumsby MG, Thomson WW (1973) Plastid differentiation, acyl lipid, and fatty acid changes in developing green maize leaves. *Plant Physiol* 52(3):240–245.
32. Jennings PH, Zscheile FP, Brannaman BL (1970) Sterol changes in maize leaves infected with Helminthosporium carbonum. *Plant Physiol* 45(5):634–635.
33. Zitomer NC, Glenn AE, Bacon CW, Riley RT (2008) A single extraction method for the analysis by liquid chromatography/tandem mass spectrometry of fumonisins and biomarkers of disrupted sphingolipid metabolism in tissues of maize seedlings. *Anal Bioanal Chem* 391(6):2257–2263.
34. Wang C, Yang A, Yin H, Zhang J (2008) Influence of water stress on endogenous hormone contents and cell damage of maize seedlings. *J Integr Plant Biol* 50(4):427–434.
35. Sheng M, Tang M, Zhang F, Huang Y (2011) Influence of arbuscular mycorrhiza on organic solutes in maize leaves under salt stress. *Mycorrhiza* 21(5):423–430.
36. Apelbaum A, Yang SF (1981) Biosynthesis of stress ethylene induced by water deficit. *Plant Physiol* 68(3):594–596.
37. Mueller LA, Zhang P, Rhee SY (2003) AraCyc: A biochemical pathway database for Arabidopsis. *Plant Physiol* 132(2):453–460.
38. Sønderby IE, Geu-Flores F, Halkier BA (2010) Biosynthesis of glucosinolates—gene discovery and beyond. *Trends Plant Sci* 15(5):283–290.
39. Green ML, Karp PD (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5:76.
40. Saier MH, Jr., Reddy VS, Tamang DG, Västermark A (2014) The transporter classification database. *Nucleic Acids Res* 42(Database issue):D251–D258.
41. Henry CS, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 28(9):977–982.
42. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I (2013) InChI - the worldwide chemical structure identifier standard. *J Cheminform* 5(1):7.
43. Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V (2006) Genome-scale thermodynamic analysis of Escherichia coli metabolism. *Biophys J* 90(4): 1453–1461.
44. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys J* 95(3): 1487–1499.
45. Henry CS, Zinner JF, Cohoon MP, Stevens RL (2009) iBsu1103: A new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. *Genome Biol* 10(6):R69.

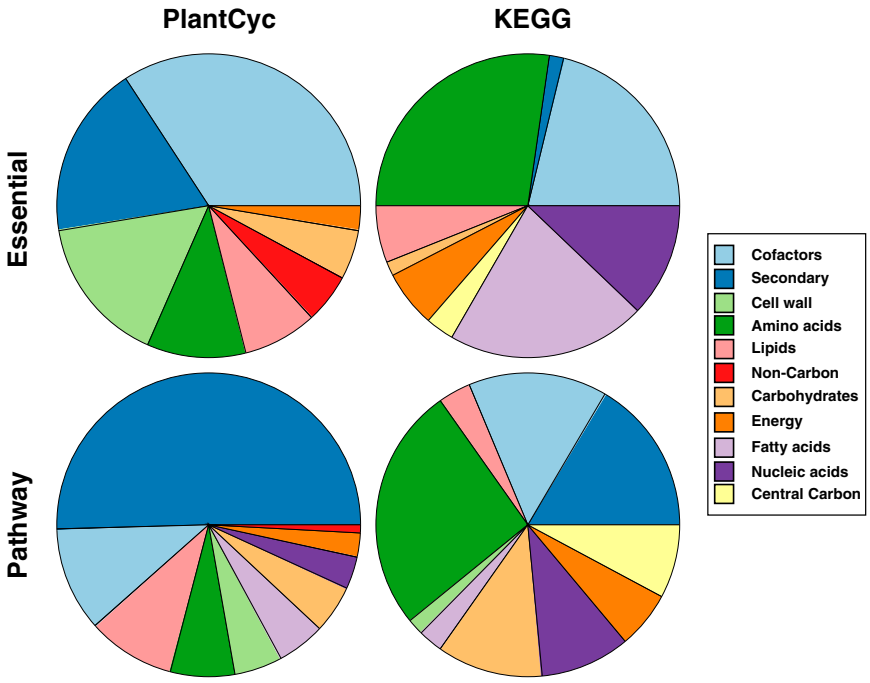

Fig. S1. Distribution of groups of pathways that needed gapfilling in PlantSEED models. A total of 202 reactions were added to PlantSEED metabolic reconstructions as a result of essential gapfilling (*Top row*), and 365 reactions were added to PlantSEED metabolic reconstructions as a result of pathway gapfilling (*Bottom row*). There is significant overlap between the pathways found in PlantCyc and KEGG, and some gapfilled reactions can be found only in one and not in the other. The three groups of pathways that needed the most gapfilling were secondary metabolism (26%), amino acid metabolism (18%), and cofactor metabolism (17%). These three groups account for more than half of all gapfilled reactions.

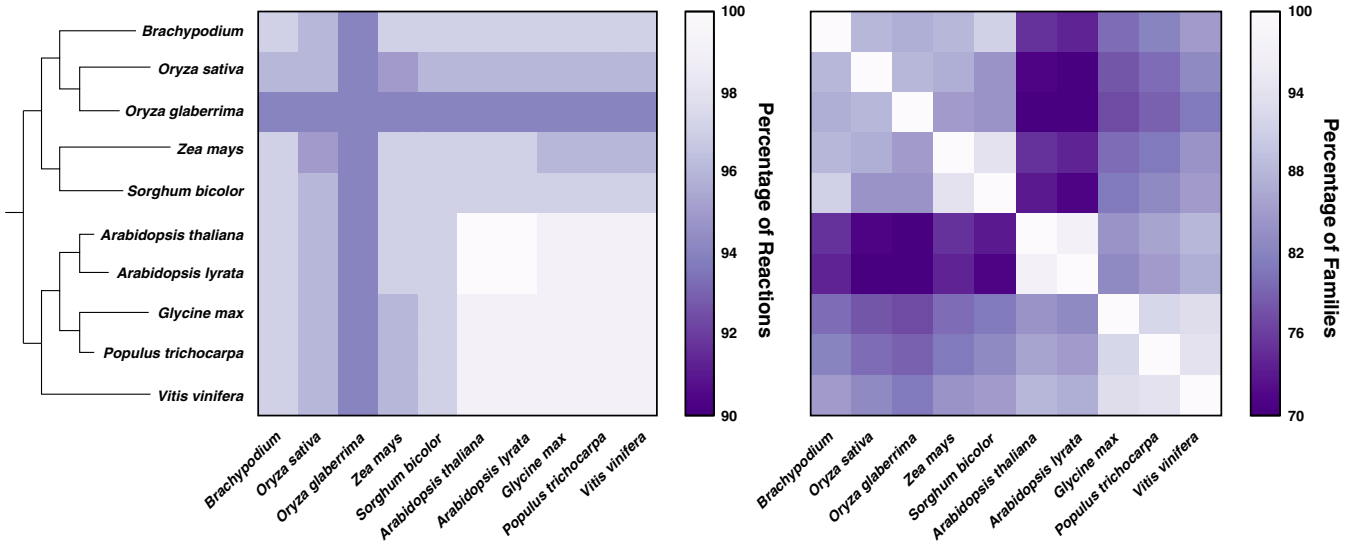

Fig. S2. Comparison of draft metabolic reconstructions in PlantSEED. (*Left*) A heatmap showing the success with which reactions found in *Arabidopsis* are propagated to other species. The species are in the order used in the cladogram. The average percentage of reactions propagated from *Arabidopsis* to other species is 98%. This percentage differs between eudicots (99.6%) (*Lower Right Quadrant*) and monocots (96%) (*Upper Right Quadrant*). Among the pathways that were not propagated from *Arabidopsis* to any monocot are glucosinolate biosynthetic pathways (see main text). (*Right*) A heat map showing the percentage of protein families within which two pairs of species share orthologs. *Arabidopsis* consistently shares fewer genes in orthologous families with monocots (74%), particularly *Oryza glaberrima* (70%). This statistic, when compared side-by-side with that of *Oryza sativa* (74%) suggests that the genome for *O. glaberrima* is of lesser quality.

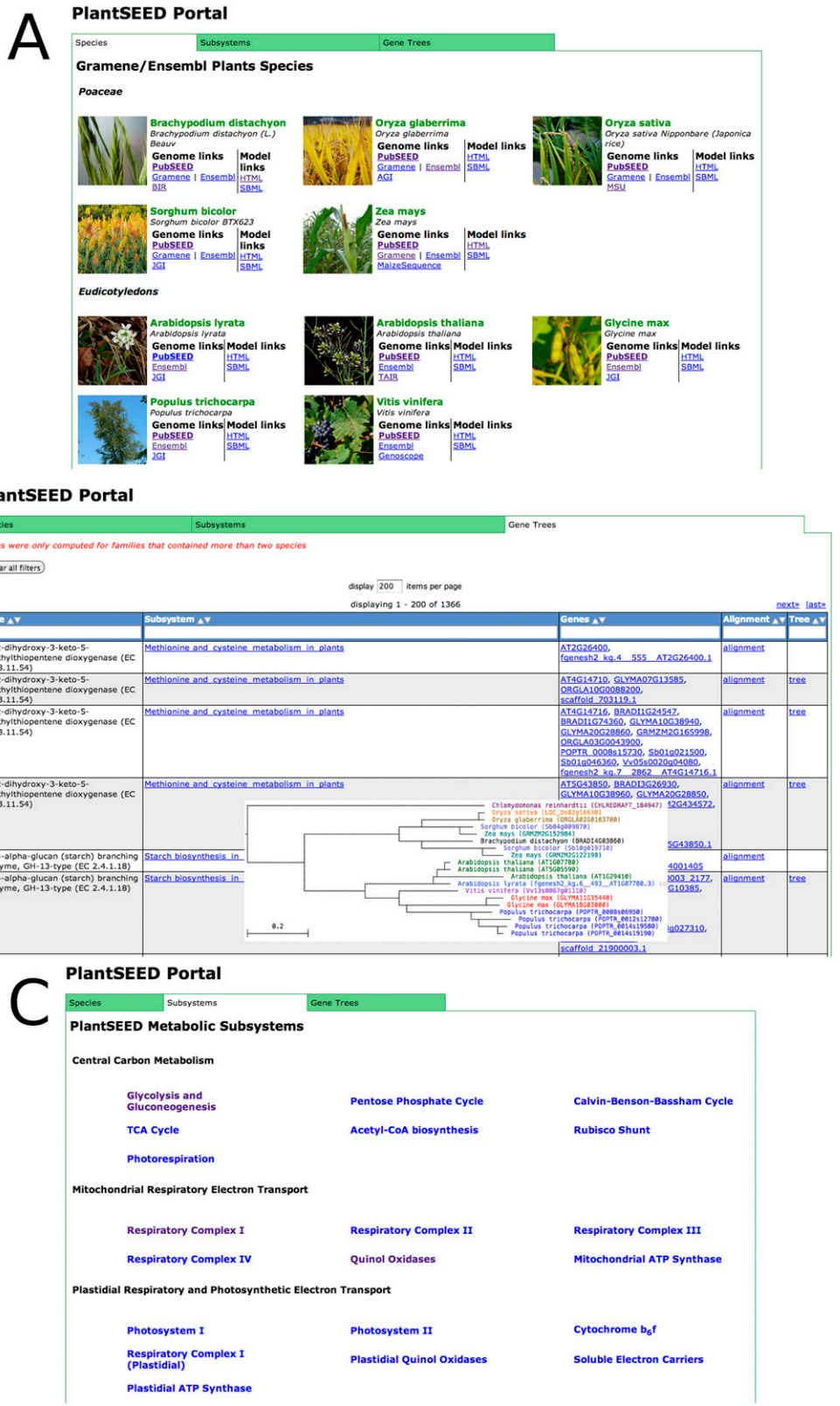

Fig. S3. The PlantSEED website. (*A*) The 10 reference genomes in the PlantSEED are shown on the front page of the PlantSEED website. Each genome is linked to various resources, including within PubSEED. (*B*) The list of PlantSEED subsystems, with links to each subsystem, and its corresponding spreadsheet of genomes and functional roles. (*C*) The list of genes assigned to each functional role in a subsystem, showing the protein families used, with links to their corresponding alignment and gene trees.


The SEED Viewer
SEED Viewer version 2.0
Welcome to the SEED Viewer - a read-only browser of the curated SEED data.
For more information about The SEED please visit theSEED.org.
For daily updates on SEED activity visit the Daily SEED
»Navigate »Organism »Comparative Tools »Feature »Feature Tools »Help find login
Annotation Overview for fig|3702.11.peg.11370 in Arabidopsis thaliana:
GTP cyclohydrolase I (EC 3.5.4.16) type 1
[to old protein page]
current assignment GTP cyclohydrolase I (EC 3.5.4.16) type 1
comment cytosolic (bioinformatic evidence in Arabidopsis)
taxonomy id 3702
internal links genome browser | feature evidence | sequence
annotation history show
CDD link show cdd
alignments and trees 3 alignments and trees
edit functional role GTP cyclohydrolase I (EC 3.5.4.16) type 1
aliases show
EC Number 3.5.4.16
contig 3 (23,459,830bp)
ACH [?] show essentially identical genes
run tool Psi-Blast run tool
propagation lock Unlocked Toggle lock
This feature is part of a subsystem
• In cak folate its role is GTP cyclohydrolase I (EC 3.5.4.16) type 1. However, the functionality of this subsystem has not yet been classified for this organism.
• In Queuosine-Archaeosine Biosynthesis its role is GTP cyclohydrolase I (EC 3.5.4.16) type 1. However, the functionality of this subsystem has not yet been classified for this organism.
• In Folate biosynthesis in plants its role is GTP cyclohydrolase I (EC 3.5.4.16) type 1.
• In Test - Folate its role is GTP cyclohydrolase I (EC 3.5.4.16) type 1. However, the functionality of this subsystem has not yet been classified for this organism.
• In Que-2 its role is GTP cyclohydrolase I (EC 3.5.4.16) type 1. However, the functionality of this subsystem has not yet been classified for this organism.
• In Folate Biosynthesis its role is GTP cyclohydrolase I (EC 3.5.4.16) type 1. However, the functionality of this subsystem has not yet been classified for this organism.
Compare Regions For fig|3702.11.peg.11370
The chromosomal region of the focus gene (top) is compared with four similar organisms. The graphic is centered on the focus gene, which is red and numbered 1. Sets of genes with similar sequence are grouped with the same number and color. Genes whose relative position is conserved in at least four other species are functionally coupled and share gray background boxes. The size of the region and the number of genomes may be reset. Click on any arrow in the display to refocus the comparison on that gene. The focus gene always points to the right, even if it is located on the minus strand.
Display options Regular Advanced
Region Size (bp) 16000
Number of Regions 100
<< < draw > >>
Visual Region Information
Tabular Region Information
Sequences
update with selected
uncheck all
A. thaliana
A. lyrata subsp
P. trichocarpa
G. max
S. bicolor
B. distachyon
Z. mays
O. glaberrima
M. roseus P3M
M. aquaeolei VT
C. phaeobactero
C. parvum NCIB
H. sp. PCC 7418
C. sp. PCC 7103
C. psychrerythr
A. sublithincol
M. purpuratum 9
S. sp. PCC 6803
S. subsalsa PCC
M. repens PCC 1
C. sp. PCC 9212
F. bacterium HT
P. saltans DSM
C. epipsammum P
S. sp. PCC 7002
A. nitrofigilis
L. sp. PCC 8106
m. bacterium HP
T. marina 5811
L. pneumophila
A. platensis NI
F. sp. PCC 9339
C. raciborskii
N. sp. PCC 7120
O. nigro-viridi
P. sp. PCC 7327
H. elongata DSM
C. burnetii Dug
G. limnaea DSM
M. ruestringens
K. algicida OT-
T. azollae 0708
T. mobilis 8321
P. vibrioformis
N. nodulosa PCC
F. taffensis DS
R. marinus DSM
M. vaginatus FG
L. blandensis M
A. cylindrica P
C. aponinum PCC
R. biformata HT
S. canadensis D
C. reinhardtii
M. aeruginosa N
P. fluorescens
C. sp CCY 0110
N. defluvii
N. oceani ATCC
L. wolffii str.
D. salina PCC 8
M. hydrothermal
A. sp. MP5ACTX8
A. albertimagni
T. scotoductus
C. difficile NA

Fig. S4. An example of cross-kingdom comparative genomics within PlantSEED. The website shown here can be accessed directly via http://pubseed.theseed.org/?page=Annotation&feature=fig|3702.11.peg.11370&number_of_regions=100. The large number of bacterial homologs, represented by the aligned red arrows, all share the same or similar annotation as the *Arabidopsis* homolog. When the plant gene in question has little or no annotation, the web page for that gene can be used to explore the annotation in the corresponding bacterial homologs.

Table S1. Sources of biochemistry integrated into PlantSEED

Database or model	Species	Owner (reference)	Date or version	No. of compounds	No. of reactions
KEGG	N/A	KEGG (1, 2)	8/30/12	17,227	11,403
MetaCyc	N/A	PathwayTools (3)	16.1	12,324	14,893
EcoCyc	*Escherichia coli*	PathwayTools (3)	16.1	3,355	2,411
PlantCyc	N/A	Plant Metabolic Network (4, 5)	7.0	3,919	4,436
AraCyc	*Arabidopsis thaliana*	Plant Metabolic Network (5)	10.0	4,229	4,008
PoplarCyc	*Populus trichocarpa*	Plant Metabolic Network (4)	4.0	2,318	2,301
SoyCyc	*Glycine max*	Plant Metabolic Network	2.0	2,796	2,657
ChlamyCyc	*Chlamydomonas reinhardtii*	Plant Metabolic Network	3/03/10	1,718	1,875
BrachyCyc	*Brachypodium distachyon*	Gramene (6, 7)	2.0	2,348	3,104
MaizeCyc	*Zea mays*	Gramene (8)	2.0	2,219	2,358
RiceCyc	*Oryza sativa*	Gramene (9)	2.0.1	1,859	1,955
SorghumCyc	*Sorghum bicolor*	Gramene (6, 7)	1.0.1	1,773	1,873
iAF1260	*Escherichia coli*	Feist et al. (2007) (10)		1,041	2,064
iAF692	*Methanosarcina barkeri*	Feist et al. (2006) (11)		562	613
iIN800	*Saccharomyces cerevisiae*	Nookaew et al. (2008) (12)		683	1,053
iJR904	*Escherichia coli*	Reed et al. (2003) (13)		629	921
iMA945	*Salmonella spp.*	AbuOun et al. (2009) (14)		1032	1,960
iMM904	*Saccharomyces cerevisiae*	Mo et al. (2009) (15)		712	1,401
iRR1083	*Salmonella typhimurium* LT2	Raghunathan et al. (2009) (16)		759	1,086
iSB619	*Staphylococcus aureus* N315	Becker & Palsson (2005) (17)		614	639
iSO783	*Shewanella oneidensis* MR-1	Pinchuk et al. (2010) (18)		634	774
iAbaylyiv4	*Acinetobacter baylyi* ADP1	Durot et al. (2008) (19)		699	867
	*Bacillus subtilis**	Goelzer et al. (2008) (20)		475	504
iGT196	*Buchnera aphidicola*	Thomas et al. (2009) (21)		740	210
iIT341	*Helicobacter pylori*	Thiele et al. (2005) (22)		411	473
iJN746	*Pseudomonas putida* KT2440	Nogales et al. (2008) (23)		706	915
iMO1056	*Pseudomonas aeruginosa* PAO1	Oberhardt et al. (2008) (24)		750	864
iND750	*Saccharomyces cerevisiae*	Duarte et al. (2004) (25)		650	1,038
iNJ661	*Mycobacterium tuberculosis* H37Rv	Jamshidi & Palsson (2007) (26)		761	951
iPS189	*Mycoplasma genitalium*	Suthers et al. (2009) (27)		277	262
iRS1563	*Zea mays*	Saha et al. (2011) (28)		1,812	1,949
iRS1597	*Arabidopsis thaliana*	Saha et al. (2011) (28)		1,759	1,837
iYO844	*Bacillus subtilis*	Oh et al. (2007) (29)		776	1,016
	*Chlamydomonas reinhardtii**	Boyle & Morgan (2009) (30)		266	485
	*Chlamydomonas reinhardtii**	Manichaikul et al. (2009) (31)		124	238
	*Chlamydomonas reinhardtii**	Chang et al. (2011) (32)		1,164	2,084
C4GEM	*Zea mays*	de Oliveira Dal'Molin et al. (2010) (33)		1,207	1,227
	*Arabidopsis thaliana**	Mintz-Oron et al. (2012) (34)		1,181	3,382
	*Arabidopsis thaliana**	Poolman et al. (2009) (35)		1,224	1,354
AraGEM	*Arabidopsis thaliana*	de Oliveira Dal'Molin et al. (2010) (36)		1,546	1,590
AlgaGEM	*Chlamydomonas reinhardtii*	de Oliveira Dal'Molin et al. (2011) (37)		1,662	1,713

The numbers of compounds and reactions are listed for each source after integration and may not reflect the numbers seen in the literature. In addition, the number of reactions includes compartmentalized reactions, which may be duplicates.
*No unique identifier for these models was described in the literature.

1. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30.
2. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40(Database issue):D109–D114.
3. Caspi R, et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 40(Database issue):D742–D753.
4. Zhang P, 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(4):1479–1491.
5. Mueller LA, Zhang P, Rhee SY (2003) AraCyc: A biochemical pathway database for Arabidopsis. *Plant Physiol* 132(2):453–460.
6. Jaiswal P (2011) Gramene database: A hub for comparative plant genomics. *Methods Mol Biol* 678:247–275.
7. Youens-Clark K, et al. (2011) Gramene database in 2010: Updates and extensions. *Nucleic Acids Res* 39(Database issue):D1085–D1094.
8. Monaco MK, et al. (2013) Maize metabolic network construction and transcriptome analysis. *Plant Gen* 6(1):1–12.
9. Dharmawardhana P, et al. (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(1):15.
10. Feist AM, et al. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:121.
11. Feist AM, Scholten JC, Palsson BO, Brockman FJ, Ideker T (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol Syst Biol* 2(1).
12. Nookaew I, et al. (2008) The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: A scaffold to query lipid metabolism. *BMC Syst Biol* 2(1):71.
13. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol* 4(9):R54.
14. AbuOun M, et al. (2009) Genome scale reconstruction of a Salmonella metabolic model: Comparison of similarity and differences with a commensal Escherichia coli strain. *J Biol Chem* 284(43):29480–29488.
15. Mo ML, Palsson BO, Herrgård MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* 3:37.
16. Raghunathan A, Reed J, Shin S, Palsson B, Daefler S (2009) Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction. *BMC Syst Biol* 3:38.

17. Becker SA, Palsson BO (2005) Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: An initial draft to the two-dimensional annotation. *BMC Microbiol* 5:8.
18. Pinchuk GE, et al. (2010) Constraint-based model of Shewanella oneidensis MR-1 metabolism: A tool for data analysis and hypothesis generation. *PLOS Comput Biol* 6(6):e1000822.
19. Durot M, et al. (2008) Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst Biol* 2:85.
20. Goelzer A, et al. (2008) Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis. *BMC Syst Biol* 2:20.
21. Thomas GH, et al. (2009) A fragile metabolic network adapted for cooperation in the symbiotic bacterium Buchnera aphidicola. *BMC Syst Biol* 3:24.
22. Thiele I, Vo TD, Price ND, Palsson BO (2005) Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): An in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol* 187(16):5818–5830.
23. Nogales J, Palsson BO, Thiele I (2008) A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory. *BMC Syst Biol* 2:79.
24. Oberhardt MA, Puchałka J, Fryer KE, Martins dos Santos VA, Papin JA (2008) Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1. *J Bacteriol* 190(8):2790–2803.
25. Duarte NC, Herrgård MJ, Palsson BO (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* 14(7):1298–1309.
26. Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Syst Biol* 1:26.
27. Suthers PF, et al. (2009) A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189. *PLOS Comput Biol* 5(2):e1000285.
28. Saha R, Suthers PF, Maranas CD (2011) Zea mays iRS1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE* 6(7):e21784.
29. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R (2007) Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* 282(39):28791–28799.
30. Boyle NR, Morgan JA (2009) Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii. *BMC Syst Biol* 3:4.
31. Manichaikul A, et al. (2009) Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat Methods* 6(8):589–592.
32. Chang RL, et al. (2011) Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol Syst Biol* 7:518.
33. de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010) C4GEM, a genome-scale metabolic model to study C4 plant metabolism. *Plant Physiol* 154(4):1871–1885.
34. Mintz-Oron S, et al. (2012) Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proc Natl Acad Sci USA* 109 (1):339–344.
35. Poolman MG, Miguet L, Sweetlove LJ, Fell DA (2009) A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant Physiol* 151(3):1570–1581.
36. de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol* 152(2):579–589.
37. de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW, Nielsen LK (2011) AlgaGEM—a genome-scale metabolic reconstruction of algae based on the Chlamydomonas reinhardtii genome. *BMC Genomics* 12(Suppl 4):S5.

Table S2. Enumeration of compartmentalized genes and reactions in PlantSEED

Compartment	PlantSEED ID	No. genes	No. reactions
Cytosol	c	1,576	1,529
Plastid	d	660	148
Mitochondria	m	314	40
Endoplasmic reticulum	r	155	40
Peroxisome	x	92	23
Extracellular	e	63	2
Cell wall	w	60	1
Nucleus	n	48	1
Vacuole	v	46	10

Table S3. Genes and reactions in the FAD biosynthetic pathways of AraCyc and PlantSEED

Role*	EC number	AraCyc reactions	AraCyc genes[†,‡]	PlantSEED reactions	PlantSEED genes[†]
GTPCH2	3.5.4.25	GTP-CYCLOHYDRO-II-RXN	**AT2G22450**	rxn00300	**AT5G59750**
			AT5G59750		**AT5G64300**
			AT5G64300		
PyrD	3.5.4.26	RIBOFLAVINSYNDEAM-RXN	**AT4G20960**	rxn02475	**AT4G20960**
			AT3G47390		
PyrR	1.1.1.193	RIBOFLAVINSYNREDUC-RXN	**AT3G47390**	rxn02474	**AT3G47390**
PyrP	—	RIBOPHOSPHAT-RXN		rxn05039	
DHBPS	4.1.99.12	DIOHBUTANONEPSYN-RXN	**AT2G22450**	rxn05040	**AT2G22450**
			AT5G59750		**AT5G64300**
			AT5G64300		
DMRLS	2.5.1.78	LUMAZINESYN-RXN	**AT2G44050**	rxn03080	**AT2G44050**
RSA	2.5.1.9	RIBOFLAVIN-SYN-RXN	**AT2G20690**	rxn00048	**AT2G20690**
			AT2G44050		
			AT3G03260		
RK	2.7.1.26	RIBOFLAVINKIN-RXN	AT1G56500	rxn00392	**AT4G21470**
			AT2G38740		
			AT3G48420		
			AT4G11570		
			AT4G21470		
			AT4G25840		
			AT4G39970		
			AT5G57440		
FMNse	3.1.3.-			rxn00391	*AT1G79790*
FMNAT	2.7.7.2	FADSYN-RXN	AT1G56500	rxn00122	**AT5G03430**
			AT2G38740		*AT5G08340*
			AT3G48420		*AT5G23330*
			AT4G11570		
			AT4G21470		
			AT4G25840		
			AT4G39970		
			AT5G03430		
			AT5G57440		
FAD-Ppse	3.6.1.18			rxn00121	*AT2G42070*

*The full names of these roles can be seen in the subsystem "Riboflavin, FMN and FAD biosynthesis in plants."
[†]Genes shared between the two sources are shown in bold, and genes uniquely curated for PlantSEED are shown in italics.
[‡]The AraCyc gene–reaction associations listed here were taken from AraCyc version 10.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)
Dataset S4 (XLSX)
Dataset S5 (XLSX)