frontiers in Plant Science

ORIGINAL RESEARCH
published: 10 March 2015
doi: 10.3389/fpls.2015.00142

Improved evidence-based genome-scale metabolic models for maize leaf, embryo, and endosperm

Samuel M. D. Seaver[1,2], *Louis M. T. Bradbury*[3,4], *Océane Frelin*[3], *Raphy Zarecki*[5], *Eytan Ruppin*[5], *Andrew D. Hanson*[3] *and Christopher S. Henry*[1,2]*

[1] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA, [2] Computation Institute, The University of Chicago, Chicago, IL, USA, [3] Horticultural Sciences Department, University of Florida, Gainesville, FL, USA, [4] Department of Biology, York College, City University of New York, New York, NY, USA, [5] Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

There is a growing demand for genome-scale metabolic reconstructions for plants, fueled by the need to understand the metabolic basis of crop yield and by progress in genome and transcriptome sequencing. Methods are also required to enable the interpretation of plant transcriptome data to study how cellular metabolic activity varies under different growth conditions or even within different organs, tissues, and developmental stages. Such methods depend extensively on the accuracy with which genes have been mapped to the biochemical reactions in the plant metabolic pathways. Errors in these mappings lead to metabolic reconstructions with an inflated number of reactions and possible generation of unreliable metabolic phenotype predictions. Here we introduce a new evidence-based genome-scale metabolic reconstruction of maize, with significant improvements in the quality of the gene-reaction associations included within our model. We also present a new approach for applying our model to predict active metabolic genes based on transcriptome data. This method includes a minimal set of reactions associated with low expression genes to enable activity of a maximum number of reactions associated with high expression genes. We apply this method to construct an organ-specific model for the maize leaf, and tissue specific models for maize embryo and endosperm cells. We validate our models using fluxomics data for the endosperm and embryo, demonstrating an improved capacity of our models to fit the available fluxomics data. All models are publicly available via the DOE Systems Biology Knowledgebase and PlantSEED, and our new method is generally applicable for analysis transcript profiles from any plant, paving the way for further *in silico* studies with a wide variety of plant genomes.

Keywords: systems biology, plant metabolism, transcriptomics, metabolic networks, flux balance analysis, *Zea mays*

OPEN ACCESS

Edited by:
Lee Sweetlove, University of Oxford, UK

Reviewed by:
Ján A. Miernyk, University of Missouri, USA
Camila Caldana, Brazilian Bioethanol Science and Technology Laboratory (CTBE), Brazil

***Correspondence:**
Christopher S. Henry, Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA
chrisshenry@gmail.com

Specialty section:
This article was submitted to Plant Systems and Synthetic Biology, a section of the journal Frontiers in Plant Science

Received: *16 October 2014*
Accepted: *22 February 2015*
Published: *10 March 2015*

Citation:
Seaver SMD, Bradbury LMT, Frelin O, Zarecki R, Ruppin E, Hanson AD and Henry CS (2015) Improved evidence-based genome-scale metabolic models for maize leaf, embryo, and endosperm. Front. Plant Sci. 6:142. doi: 10.3389/fpls.2015.00142

Introduction

The ability of a plant to grow and survive is linked to its metabolic network (Stitt et al., 2010), which indicates that a capacity to predict and understand plant metabolism will improve our understanding of plant response to changing environments and genetic perturbations (Mo et al., 2009;

Chang et al., 2011; Saha et al., 2011). Furthermore, the yield of a wide range of plant products is crucial to human society, particularly when inputs such as water are limited (Skirycz and Inze, 2010). Many classical biochemical and genetic experiments involve the elucidation of biological functions for individual gene products. However, many external and internal perturbations lead to systemic responses, and a systems-level understanding of plant metabolism is required to fully explain these system responses.

To build this systems-level understanding, several genome-scale metabolic reconstructions have recently been published for plant species (Poolman et al., 2009; de Oliveira Dal'molin et al., 2010a,b; Saha et al., 2011; Poolman et al., 2013). Each reconstruction consists of all reactions known to be catalyzed by one or more of the gene products in the plant genome. The methods employed to study these metabolic models, such as flux balance analysis (FBA), consider all reactions in the model when attempting to predict a biological phenotype, such as plant growth. Metabolic reconstructions are built from many data sources, notably public databases and individual publications. Reconstructions are validated by comparing the activity of well-characterized pathways *in silico* with biochemical evidence in the literature. Poolman et al. (2009) built the first genome-scale plant metabolic reconstruction, which could respire on heterotrophic media *in silico* and produce biomass components in proportions that matched *in vivo* observations. de Oliveira Dal'molin et al. (2010a) investigated autotrophic biosynthesis of plant biomass, showing that the model correctly predicted the reactions used for both photosynthesis and photorespiration. de Oliveira Dal'Molin et al. also developed a metabolic reconstruction of a C4 plant (de Oliveira Dal'molin et al., 2010b) containing plastidial reactions for photosynthesis. This reconstruction was shown to be capable of performing three known subtypes of C4 photosynthesis. In other work, Saha et al. (2011) show that genetic perturbations in the phenylpropanoid biosynthesis pathway could be simulated *in silico*, producing an impact on cell wall composition that compared favorably with experimental data from known maize mutants.

The validation approaches described above are based on a few well-known biochemical pathways, and involve large genome-scale metabolic reconstructions, built to provide a systems-level understanding of how a metabolic network behaves under certain conditions. For example, Schwender and Hay (2012) investigated how a metabolic reconstruction exhibited variation in reaction activity in response to variation in the biosynthetic demands of oil and protein as storage products in the plant embryo and were able to identify the utilization of a pathway within the network of reactions that was not yet characterized in the literature. Similarly, Töpfer et al. (2013) explored the means with which a set of pathways in a metabolic reconstruction responded to various conditions of light and temperature, showing, in one case, the preference for methylerythritol 4-phosphate pathway over the mevalonate pathway in isoprenoid biosynthesis, and also generating a new hypothesis for the role of homocysteine–cysteine conversion.

Genome-scale metabolic reconstructions are generated based on the annotation of all gene products in the full genome, and, thus, they include every reaction that can be catalyzed by the plant. However, a multi-cellular organism will activate different subsets of their genes in different organs, tissues, developmental stages, and environmental conditions. To be accurate, genome-scale metabolic reconstructions must represent the reduced metabolism that truly exists in cells of a specific type and in a specific condition. Most reconstructions mentioned previously were either intended to represent a leaf cell or the primary metabolism of a generic plant cell. Other metabolic reconstructions have been built to target specific tissues and organs, such as the seeds of barley (*Hordeum vulgare*; Grafahrend-Belau et al., 2009), and the embryos of oilseed rape (*Brassica napus*; Hay and Schwender, 2011a,b; Pilalis et al., 2011). Grafahrend-Belau et al. followed up their study of barley seeds by building manually curated metabolic reconstructions of barley stem and leaf, and integrating the three reconstructions into a single model (Grafahrend-Belau et al., 2013). Recently, several new approaches have emerged to integrate large-scale data (Baerenfaller et al., 2008) in an automated manner to either generate new condition-specific models (Mintz-Oron et al., 2012), or to constrain the behavior of individual reactions in a full genome-scale model to better reflect the behavior of specific organs or tissues (Töpfer et al., 2013).

The ongoing explosion in plant transcriptome sequencing, driven by advances in next-generation sequencing (NGS) and by the relative ease of sequencing a collection of cDNAs as opposed to predicting gene models in plant chromosomes (Ozsolak and Milos, 2011), means that many transcript profiles are now publicly available, and individual laboratories can afford to generate new transcript profiles for individual experiments. Indeed, Töpfer et al. used their own transcript profiles, which they generated from Arabidopsis rosettes (Töpfer et al., 2013). Several computational methods have been developed that are able to integrate transcript profiles with a metabolic reconstruction to produce improved predictions of reaction utilization and flux.

Töpfer et al. used E-flux (Colijn et al., 2009), which fits flux predictions based on gene expression data, but does not attempt to reduce a full genome model to a tissue or organ specific version. The Töpfer et al. work was focused on several primary and secondary metabolic pathways that are known to be active with the rosettes of Arabidopsis. Mintz-Oron et al. used the iMAT approach (Jerby et al., 2010; Zur et al., 2010), which generates aggregate models based on random sampling of fluxes to fit gene expression data. While this approach provides a more comprehensive account of the metabolic network, the extensive sampling can be cumbersome. An updated method eliminates the need for random sampling and thereby runs faster (Wang et al., 2012). This method searches for an optimal solution by iteratively activating each reaction whose associated genes have high expression, which means that the method still performed many optimizations. We have developed a new approach that requires far fewer optimization steps, allowing for transcriptome-based metabolic reconstructions to be formed from transcript profiles at a greater speed and with less complexity. We note here the introduction of the term transcriptome-based to reflect this class of model, which is based on fitting a genome-scale model to a select subset of gene expression data. The term tissue-specific is often used for models of this type. However, expression data often

does not capture the entire behavior or a tissue, nor does a single tissue necessarily reflect a single biological behavior (e.g., leaf tissue consists of several sub-cell types).

We demonstrate our approach for reconstruction of transcript-specific models with a new genome-scale metabolic reconstruction of maize. Our new genome-scale maize model includes three important enhancements over previously published models: (i) an expanded and improved biomass composition; (ii) improved gene-protein-reaction associations where low confidence gene-reaction mappings based on poor evidence or purely computational predictions have been removed; and (iii) improved compartmentalization of reactions to subcellular organelles based on a combination of literature evidence, curation, and gapfilling algorithms. The improved gene-reaction associations in our new model were critical to our use of maize transcript profiles (Davidson et al., 2011) to produce new transcriptome-based models of the leaf, embryo, and endosperm in maize. We applied our novel model reconstruction method to maximize the activity of reactions associated with high expression genes while removing as many reactions associated with low expression genes as possible. We also adjusted the biomass composition of our embryo and endosperm models to better fit the actual composition data for these tissues by curating data for individual components from a variety of literature sources. To test the accuracy of our models, we explored how well they replicate the flux profiles measured for central carbon metabolism in embryo and endosperm tissues (Alonso et al., 2010, 2011). This analysis demonstrates that our models have an improved fit between the fluxes generated *in silico* and the fluxes measured *in vivo*. All models produced from this work are available for download from the DOE Systems Biology Knowledgebase (http://kbase.us) and the PlantSEED resource (Seaver et al., 2014).

Materials and Methods

Biochemistry

We used the plant biochemistry database built for the PlantSEED project (Seaver et al., 2014). This database is notably built on KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2012) and MetaCyc (Caspi et al., 2012), which had been integrated using InChI (Heller et al., 2013) strings generated from mol files provided by both databases. The integration includes several plant biochemistry databases such as the BioCyc databases for *Arabidopsis thaliana* (*Arabidopsis*; AraCyc v11.5 Mueller et al., 2003; Zhang et al., 2010), and maize (MaizeCyc v2.2.2 Monaco et al., 2013 and CornCyc v4.0 Zhang et al., 2010), and several published metabolic models for *A. thaliana* (de Oliveira Dal'molin et al., 2010a,b, 2011; Saha et al., 2011; Mintz-Oron et al., 2012) and maize (de Oliveira Dal'molin et al., 2010b; Saha et al., 2011). The metabolic reconstructions we built depend on this integration, and the reactions for the respective *Arabidopsis* and maize metabolic reconstructions are thus drawn from this database.

Compartments

An important aspect of plant metabolic models is the compartmentalization of reactions into plastids, mitochondria, and other organelles. To accurately capture this compartmentalization, we downloaded localization data for proteins from PPDB (Sun et al., 2009), SUBA (Tanz et al., 2013), AraCyc, MaizeCyc, and CornCyc. We systematically avoided any protein localizations generated solely via computational predictions. From PPDB, we only used data that the PPDB team had curated. From SUBA, we only used data from GFP experiments, which are more reliable than the data from mass spectrometry experiments. Finally, from AraCyc, MaizeCyc, and CornCyc, many reactions are localized according to biochemical support such as the histidine pathway in plastids (Ingle, 2011). Even if the genes associated with these pathways do not have localization data, we considered them to be localized if there was experimental evidence for the gene-reaction associations. Much of the localization data could only be applied directly to either of the two different species, and therefore we propagated the associations between *Arabidopsis* and maize by using the same conservative approach we applied to EnsemblCompara protein families in the PlantSEED project (Vilella et al., 2009; Kersey et al., 2014; Seaver et al., 2014).

Model Pathway-Gapfilling

A new gapfilling algorithm was applied during the reconstruction of all our plant genome-scale models. This algorithm provides a means of identifying the minimal set of reactions that must be made reversible or added to the model in order to activate as many gene-associated reactions in the model as possible. The constraints of the optimization problem resemble the constraints for existing classical gapfilling approaches (Satish Kumar et al., 2007; Kumar and Maranas, 2009).

$$N_{super} \bullet v = 0 \tag{1}$$

$$0 \leq v_i \leq 100 z_i \qquad i = 1, \ldots, r_{gapfill} \tag{2}$$

$$z_{for,i} + z_{rev,i} \leq 1 \qquad i = 1, \ldots, r_{gapfill} \tag{3}$$

$$-100 \leq v_{ex,i} \leq 100 \gamma_i \qquad i = 1, \ldots, m_{transported} \tag{4}$$

Equation (1) represents the mass balance constraints, where N_{super} is the matrix of stoichiometric coefficients through all reactions in our model plus all candidate reactions added from our biochemistry database, while $\boldsymbol{v}$ is the vector of fluxes through all model and database reactions represented in the N_{super} matrix. In these and all other constraints, reversible reactions have been decomposed into separate forward and backward component reactions to ensure that all fluxes are always positive. Equation (2) sets the bounds on the flux through reaction i, where v_i is the flux and z_i is a binary use variable equal to zero when the flux is zero and equal to one otherwise. Equation (3) ensures that the forward and backward components of the same reaction may not both be active at the same time; in our formulation, this constraint is the sole reason for using binary variables. Equation (4) establishes the growth conditions for the gapfilling analysis; metabolites present in the growth media (e.g., heterotrophic media or autotrophic media) have a γ_i of 1 in Equation (4). Otherwise γ_i is zero.

In addition to these standard constraints, we applied a new constraint that introduces a slack flux for all reactions found in the original un-gapfilled model:

$$v_{for,i} + v_{rev,i} + \delta_i \geq 0.01 \qquad i = 1, \ldots, r_{model} \tag{5}$$

Equation (5) states that the sum of the net flux through reaction i ($v_{for,i} + v_{rev,i}$) and the slack flux for reaction i (δ_i) must be greater than or equal to 0.01. As a result of this constraint, a reaction can only have a net flux of zero if the corresponding slack flux is 0.01. Thus, the slack flux is a variable used to identify reactions that carry no flux in the model. We utilize this new slack flux for this purpose in the objective function for our gapfilling.

Objective:

$$\text{Minimize} \sum_{i=1}^{r_{annotated}} a\left(\gamma_{activate,i}\delta_i\right) + \sum_{i=1}^{r_{gapfilling}} \left(\gamma_{gapfill,i}v_i\right) \quad (6)$$

This new objective function minimizes the sum of the slack fluxes associated with the reactions included in our original model while simultaneously minimizing the flux through all gapfilled reactions added to the model from our database. The purpose of this objective function is to maximize the number of gene-associated reactions that carry flux while minimizing the number of gapfilled reactions added to the model. This effectively gives precedence to the gene-associated reactions in our model. The activation coefficient, $\gamma_{activate,i}$, dictates the cost of leaving a gene-associated reaction inactive, while the gapfilling coefficient, $\gamma_{gapfill,i}$, dictates the cost of adding a gapfilled reaction to the model. In our gapfilling studies, we set $\gamma_{activate,i}$ equal to one for all gene-associated reactions, while we computed $\gamma_{gapfill,i}$ as described in our previous work (Henry et al., 2009, 2010).

We also used a scaling factor a in our objective function, which scales the cost of leaving some model reactions inactive against the cost of adding new reactions to the model from the database. We explored values for a ranging from 0.01 to 0.25, but we found only a small effect on the solutions produced. Generally, an a of 0.1 generated the most well-balanced gapfilling solutions.

In this gapfilling formulation, we utilize continuous linear flux variables in our objective function rather than the more typical binary variables (e.g., $z_{for,i}$ and $z_{rev,i}$) (Kumar et al., 2007). This adjustment reduced the compute time required to obtain a globally optimal solution by over 90% while having no appreciable impact on solutions obtained. This use of linear variables has been previously proposed in other published gapfilling algorithms, with detailed sensitivity analyses performed and similar results obtained (Latendresse, 2014). Thus, we do not repeat the sensitivity analysis here.

Transcriptome-Based Pathway-Gapfilling

Our method for producing transcriptome-based models builds on the pathway-gapfilling approach (see previous used during the reconstruction of our models. Our pathway-gapfilling approach attempts to maximize the number of number of active *gene-associated* reactions. This approach further refines the model toward a specific transcriptome by maximizing the activity of reactions associated with highly expressed genes while minimizing active reactions associated with minimally expressed genes. This formulation includes flexibility permitting high-expression reactions to remain "off" if activating them requires the function of too many low expression reactions, and vice versa.

The first step of this algorithm is to categorize every reaction in the model as either high expression or low expression. This is done by assigning an expression score, $E_{exp,i}$, to every gene-associated reaction i as follows:

$$E_{exp,i} = Max(C_{exp,i,j}) \qquad i = 1, \ldots, r \qquad j = 1, \ldots, c_i \quad (7)$$
$$C_{exp,j} = Min(P_{exp,j,k}) \qquad j = 1, \ldots, c_i \qquad k = 1, \ldots, p_j \quad (8)$$
$$P_{exp,k} = Max(G_{exp,k,l}) \qquad k = 1, \ldots, p_j \qquad l = 1, \ldots, g_k \quad (9)$$

In Equations (7)–(9), the reaction expression score, $E_{exp,i}$, is equal to the maximum of the complex expression scores, $C_{exp,i,j}$ for all c_i protein complexes catalyzing reaction i; the complex expression scores are equal to the minimum of the protein expression scores, $P_{exp,j,k}$, for all p_j protein subunits of each complex j; and the protein expression scores, are equal to the maximum of all gene expression scores, $G_{exp,k,l}$, associated with the g_k genes encoding each protein subunit. The gene expression score is equal to the normalized expression value of gene in the transcriptome being used as the basis to construct the model. In our analysis, the expression value of each gene was normalized by the median expression value for the same gene across all 37 conditions included in our data set, which included data from numerous organs, tissues, and growth conditions.

Reactions with an expression score falling below 0.2 were categorized as being "low expression." Biologically, a score of 0.2 means that the critical genes associated with the reaction are expressed at 20% of their average expression across all 37 conditions included in our transcriptomics data. This represents a conservative calling of "low expression" genes. We then applied the gapfilling algorithm as described in Equations (1)–(6) with two modifications: (i) the mass-balance constraints encoded by Equation (1) only included the stoichiometry of the reactions in the gapfilled full genome model (stoichiometry was not expanded to include the entire biochemistry database as done in full gapfilling); and (ii) the objective function was altered to maximize the high expression reaction activity while minimizing flux through low-expression reactions (Equation 10).

Objective:

$$Minimize \sum_{i=1}^{r_{high}} a\left(E_{exp\text{-}high,i}\delta_{high,i}\right) + \sum_{i=1}^{r_{low}} a\left(E_{exp\text{-}low,i}v_{low,i}\right) \quad (10)$$

Similar to our gapfilling formulation, this objective function minimizes the flux through the low expression reactions while also minimizing the slack fluxes associated with all high expression reactions. This maximizes the number of high expression reactions with a non-zero flux while setting the flux through as many low expression reactions as possible to zero. Again, we use a scaling factor a in our objective function, which scales the cost of leaving some high expression reactions inactive against the cost of activating some low expression reactions. We explored values for a ranging from 0.01 to 0.25, with only minimal effect on the solutions produced. We found an a of 0.1 generated the most well-balanced solutions.

Comparison with Estimated Fluxomics Data for Embryo and Endosperm

In order to calculate how well the metabolic model can match experimentally measured flux data for a list of specific reactions, we applied a QP where we minimized the distance between the predicted fluxes and the experimentally measured fluxes. The QP utilized the standard FBA constraints:

$$N_{model} \bullet v = 0 \tag{11}$$

$$v_{min,i} \leq v_i \leq 1000 \qquad i = 1, \ldots, r_{model} \tag{12}$$

$$-50 \leq v_{ex,i} \leq 50\gamma_i \qquad i = 1, \ldots, m_{transported} \tag{13}$$

Equation (11) represents our mass balance constraints, where N_{model} is the stoichiometry matrix for all model reactions and $\boldsymbol{v}$ is the vector of fluxes through all model reactions. Unlike our gapfilling formulation, in this study, reversible reactions were not decomposed. Equation (12) represents the bounds on the flux through each reaction, with the lower bound $v_{min,i}$ being zero if a reaction is irreversible and −1000 is a reaction is reversible. As in our gapfilling formulation, Equation (13) sets the bounds of uptake of nutrients from the environment.

In the quadratic objective function of our QP, we minimize the deviation of our predicted fluxes (v_i) from the experimentally measured fluxes ($v_{exp,i}$):

$$Minimize \sum_{i=1}^{r_{measured}} \left(v_{exp,i} - v_i\right)^2 \tag{14}$$

This approach is similar to that adopted by Lee et al. (2012), but by using QP, we find a single solution and avoid the iterative approach they describe. The calculations were done when the model was grown on heterotrophic media. After the minimal distance between experimental and model predicted fluxes was found via the QP problem as described above, we performed a Spearman correlation between the experimental flux values and the actual predicted flux values found by the solution when the model reached the minimal distance. The results in the form of the Spearman value and the *p*-value of the Spearman correlation are shown in **Table 2**.

Results

A High-Quality Evidence-Based Genome-Scale Metabolic Reconstruction of Maize

In order to generate a metabolic reconstruction based on available evidence, as described in the Materials and Methods Section, we started by building a full genome-scale metabolic reconstruction that integrated every reaction and gene-reaction association from all available resources. We then refined this model by removing the reactions and gene-reaction associations that did not have available support such as literature citation, human curation, or notation of presence in a specific compartment. We call this refined model an *Evidence-Based Model*. Here we described the process applied to complete this model refinement.

Initial Reconstruction of Full Genome-Scale Metabolic Models

We built our initial genome-scale metabolic reconstructions for *Arabidopsis* and maize using all reactions and genes obtained from all available resources. The resources included KEGG, the respective BioCyc databases, and the respective published metabolic models for *Arabidopsis* and maize (de Oliveira Dal'molin et al., 2010a,b; Saha et al., 2011). The two initial reconstructions are named "Full" and were composed of 6399 total reactions for *Arabidopsis* and 6458 for maize (**Table 1**).

Although we used multiple sources, we note that every published metabolic model available was in turn derived from KEGG and the respective BioCyc database. These databases are dynamic and improved over time, and, as a consequence, the published models are considered outdated. We therefore did not fully integrate the published metabolic models with two important exceptions: transport reactions and organellar reactions. These two sets of reactions, with the exception of those present in the model generated by Mintz-Oron et al. (2012), were manually reviewed in order to ensure that intra-organellar metabolic networks were active. We therefore ensure that these reactions are included.

The most telling statistic in comparing the Full metabolic reconstructions for both species is that maize has many more gene-reaction associations. This is partly because maize has undergone a recent whole-genome duplication event (Schnable et al., 2009), thus creating many paralogs, and partly because, for the MaizeCyc and CornCyc databases, many gene-reaction associations were predicted, and thereby included many similar homologs.

For each metabolic reconstruction, we showed the number of reactions that came from each source in **Figure 1**. In the Evidenced models, most of the reactions originated from BioCyc databases because KEGG provides comparatively little literature evidence for gene-reaction associations. In contrast, there is significant overlap between the KEGG database and the metabolic models published by the Nielsen/Maranas groups. This is because those metabolic models were generated from KEGG alone. We also highlight the variation in the number of reactions, compartmentalized reactions, transport reactions and genes between our models and those in the literature in **Figure 2**. In the case of the number of reactions, compartmentalized reactions and transport reactions in the Full and Evidenced models for both species, we show that the models created in this work are larger than the published models, with the exception of the model published by Mintz-Oron et al. Our models are larger than other published models primarily due to the more comprehensive database of biochemistry and plant annotations from which we generate our models, as well as the inclusion of recent database updates in our new model. The model published by Mintz-Oron et al. is larger still generally because it was expanded to include many computationally predicted compartmentalized reactions and transporters.

Compartments

By using the protein localization data collected from various sources, we were able to confirm the presence of ∼2000 reactions in eight compartments (plastid, mitochondrion, peroxisome,

TABLE 1 | A list of metabolic models generated in our work and their statistics.

Species	Type/Organ/Tissue	Reactions	Compounds	Gene-reaction associations	Gapfilled reactions
Arabidopsis thaliana	Full	6399	6236	16,577	1073
Arabidopsis thaliana	Evidenced	2801	2864	4262	697
Zea mays	Full	6458	6250	35,226	979
Zea mays	Evidenced	2629	2634	5540	667
Zea mays	Evidenced/Leaf	2322	2635	4656	925
Zea mays	Evidenced/Embryo	2304	2636	4680	885
Zea mays	Evidenced/Endosperm	2280	2636	4602	920

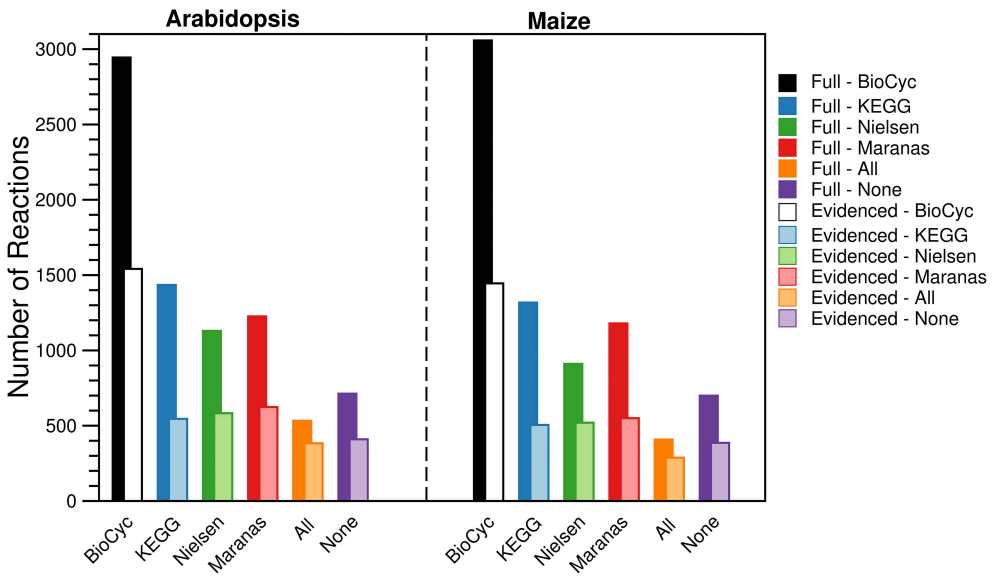

FIGURE 1 | Number of reactions in the Full and Evidenced metabolic reconstructions for *Arabidopsis* and maize. The bars represent the number of reactions shared with each of the four primary biochemical sources used to build the Full metabolic reconstruction. Reactions are counted multiple times if they are present in multiple compartments. The "All" category corresponds with reactions that are shared between all four sources, and the "None" category corresponds with reactions present in compartments that are not otherwise found in the primary sources due to protein localization evidence. The dominant source of reactions was the BioCyc databases, ~50% more reactions originated from AraCyc and MaizeCyc/CornCyc than from KEGG. In addition, the dominant source of evidence for gene-reaction evidence came from AraCyc, and as a result, far fewer reactions are shared between the Evidenced metabolic reconstructions and the published counterparts, which were originally derived from KEGG.

endoplasmic reticulum, nucleus, cell wall, vacuole, and Golgi body). We collected gene localization data for 12,398 *Arabidopsis* genes and 8737 maize genes for eight compartments in the metabolic reconstructions (see Materials and Methods), and we added reactions to the appropriate compartment whenever they were associated with a localized gene. We find that the gene localization data led to more than 700 reactions being placed in new locations that are not otherwise designated in the databases and published models used as sources; the "None" column in **Figure 1** indicates this. In the next Section, we highlight two reactions as an example of this. We show a breakdown of the number of reactions found in each compartment (**Figure 3**), and this highlights that the majority of the reactions are found in the plastid. Furthermore, we qualitatively examined the contribution of each database to the localization of reactions (**Figure 4**). The total number of reactions assigned to any compartment in the Full maize metabolic reconstruction by PPDB data is 1675, by GFP data is 1077, and by AraCyc data is 429. The PPDB data accounts for more reactions in the plastid, mitochondrion, and peroxisome, and the GFP data accounts for more reactions in the remaining compartments. Whilst there is some agreement between the sources, the number of reactions assigned to a compartment by PPDB or GFP alone is a validation of our decision to use multiple sources of evidence-based localization data.

Evidence for Gene-Reaction Associations

As stated above, we wish to refine our Full metabolic reconstructions to only contain reactions with reliable evidence for gene-reaction associations. Almost every gene-reaction association found in KEGG, and in any plant BioCyc databases that is

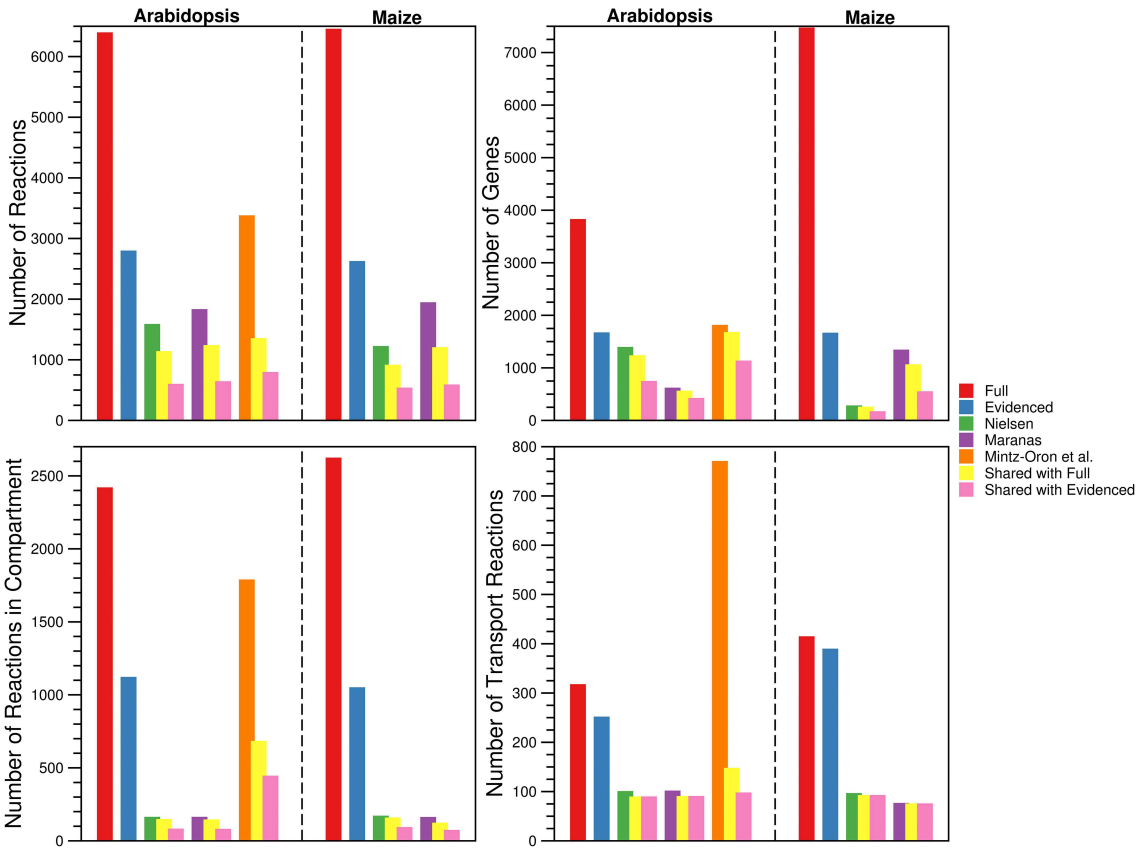

FIGURE 2 | Comparison of the number of reactions, genes, compartmentalized reactions, and transport reactions found in the Full and Evidenced metabolic reconstructions for Arabidopsis and Maize and the published metabolic models. For each of the published metabolic models, we also show the number of reactions, genes, compartmentalized reactions, and transport reactions that are shared with both the Full and Evidenced metabolic reconstructions. The number of reactions and genes in the Full metabolic reconstructions dwarf the numbers in the other models. The model generated by Mintz-Oron et al. was the first plant metabolic model to be published for which integration from more than biochemical source was performed, and as such, it has more reactions than the other published models. However, AraCyc has since gone through several expansions, which explains why so many more reactions are in the Full metabolic reconstructions (**Figure 1**). The high number of genes in the Full maize model is indicative of the number of paralogs for which computational predictions are made by multiple sources. Only 40% of the genes in iRS1563 are found in the Evidenced maize metabolic reconstruction. The Evidenced metabolic reconstructions contain over 1000 reactions that are found in other compartments (notably in the plastid, see **Figure 4**), which is approximately 10 times more than the number of compartmentalized reactions found in the models from the Nielsen and Maranas labs. The process of creating the metabolic model of Mintz-Oron et al. predicted many more compartmentalized and transport reactions than those found in the Evidenced metabolic reconstruction for Arabidopsis, but only 25% of the compartmentalized reactions and 13% of the transport reactions are found in the Evidenced metabolic reconstruction.

not AraCyc, are computationally predicted (Zhang et al., 2010; Nakaya et al., 2013; Kanehisa et al., 2014; Seaver et al., 2014). Additionally, in many of the cases, and this problem is particularly acute in plants, the set of computationally predicted genes associated with reactions may be homologous, but do not perform the same catalytic function (i.e., they are out-paralogs). The large number of gene-reaction associations in the Full metabolic reconstruction for maize highlights this problem because maize, as a species, had a recent whole-genome duplication leading to additional paralogs (Schnable et al., 2009). It is important to identify the correct gene-reaction associations, because the genes duplicated by whole-genome duplication in maize appear to be down-regulated (Schnable and Freeling, 2011; Schnable et al., 2011).

We tackled this problem of over-annotation in two steps. First we included the gene-reaction associations for which there is evidence from two primary sources, AraCyc and PlantSEED (Mueller et al., 2003; Zhang et al., 2010; Seaver et al., 2014). The PathwayTools software enables users to assign evidence codes for gene-reaction associations, and in particular we were able to weed out all the gene-reaction associations where the evidence codes indicated that only a computational prediction was made. The PlantSEED project manually reviewed many of the gene-reaction associations found in AraCyc and elsewhere (Seaver et al., 2014), but also included many carefully reviewed in-paralogs (Sonnhammer and Koonin, 2002; Seaver et al., 2014), thus allowing us to include a greater number of gene-reaction associations in our metabolic reconstructions. By using these sources

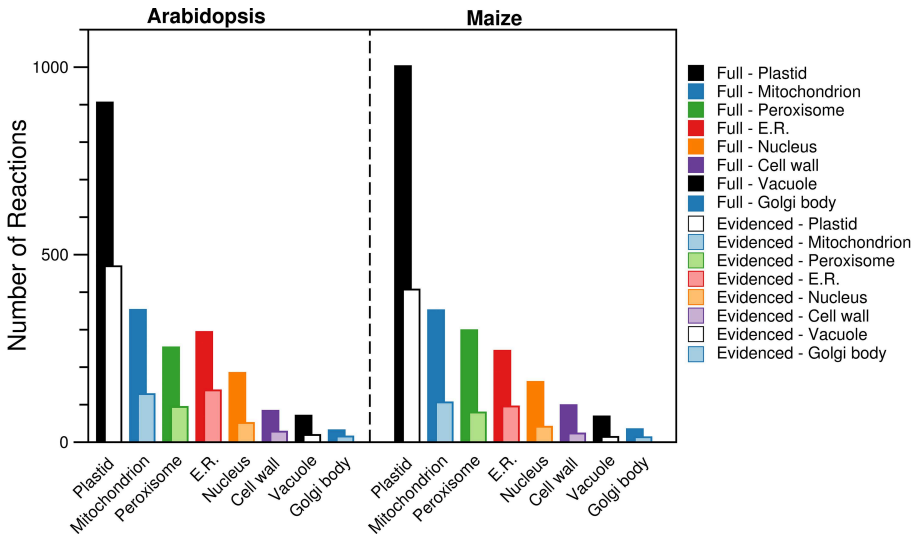

FIGURE 3 | Comparison of the number of reactions found in each of the compartments in the Full and Evidenced metabolic reconstructions for *Arabidopsis* and maize. Evidence from AraCyc, PPDB, and SUBA was used to assign each reaction to different compartments. The substantial difference in the number of reactions in each compartment for the Full and Evidenced models is a result of the large number of gene-reaction associations in the Full model, which in turn is a result of the many computational predictions used to make the associations. As such, the use of protein localization evidence to assign reactions to compartments is far more reliable with the use of evidence for the gene-reaction associations in the Evidenced model. ER, Endoplasmic reticulum.

of evidence, we produced an evidence-based metabolic reconstruction for *Arabidopsis* that contained only those reactions for which there was gene-reaction association evidence from AraCyc and PlantSEED, which we denote as "Evidenced." The Evidenced metabolic reconstruction for *Arabidopsis* is smaller, with 2801 reactions, and a smaller number of gene-reaction associations (**Table 1**). The number of reactions in the Evidenced metabolic reconstruction is 44% that of the Full metabolic reconstruction, but the number of gene-reaction associations is 26%, which is an indication of how many computational predictions are made for genes associated with reactions which otherwise have evidence for their associations with other genes.

In the second step of our model refinement, we considered the lack of evidence for any other species, given that much biochemical research in plants has been on *Arabidopsis* as a model organism. As a result, there exist only a tiny number of gene-reaction associations with evidence in MaizeCyc and CornCyc combined, and to create an Evidenced model for maize, one must consider propagating the gene-reaction associations from *Arabidopsis*. In order to avoid the pitfall of over-annotation, and yet create a reliable set of gene-reaction associations for maize, we used the same very conservative approach we applied to EnsemblCompara protein families in the PlantSEED project, described below (Vilella et al., 2009; Seaver et al., 2014). This approach greatly reduced the number of maize orthologs found in the same protein family as the *Arabidopsis* genes found in the Evidenced *Arabidopsis* metabolic reconstruction. In doing so, we are able to create an Evidenced metabolic reconstruction for maize by adding to the model only the reactions for which the associated genes have orthologs in the Evidenced metabolic reconstruction for *Arabidopsis*. The Evidenced metabolic reconstruction for maize has 2631 reactions and, ∼30,000 fewer gene-reaction associations than found in the Full metabolic reconstruction (∼84%; **Table 1**).

We highlight the utility of our approach with an example involving two reactions from the mevalonate pathway. Simkin et al. report, using YFP-fused constructs, that Phosphomevalonate kinase (PMK) and Mevalonate diphosphate decarboxylase (MVD) localize to the peroxisomes (Simkin et al., 2011). The complementary reactions for these two enzymes are found in AraCyc, MaizeCyc, and CornCyc, albeit without any localization data attached, and with experimental evidence only available for one enzyme in AraCyc. Thus, only one reaction (MVD) would be included in the Arabidopsis model and would only be cytosolic. The evidence for the gene-reaction associations is found in PlantSEED in the form of manual curation, and leads to both reactions being included in the Arabidopsis model. The results for the enzyme localization from Simkin et al. are found in SUBA, and the two reactions were therefore correctly added to the peroxisome in the Arabidopsis model. Finally, the use of EnsemblCompara protein families as described above leads to the correct maize genes being associated with the same reactions, and the reactions being thus added to the peroxisome in the maize model.

We generated a corresponding metabolic model for all four of our metabolic reconstructions by adding a biomass equation matching that used by the PlantSEED and containing more than 90 compounds. We also utilized a new pathway gapfilling method (see Materials and Methods) that attempts to generate biomass and simultaneously activate all reactions with associated genes. The pathway gapfilling recommended reactions to add to our models to produce biomass and improve the

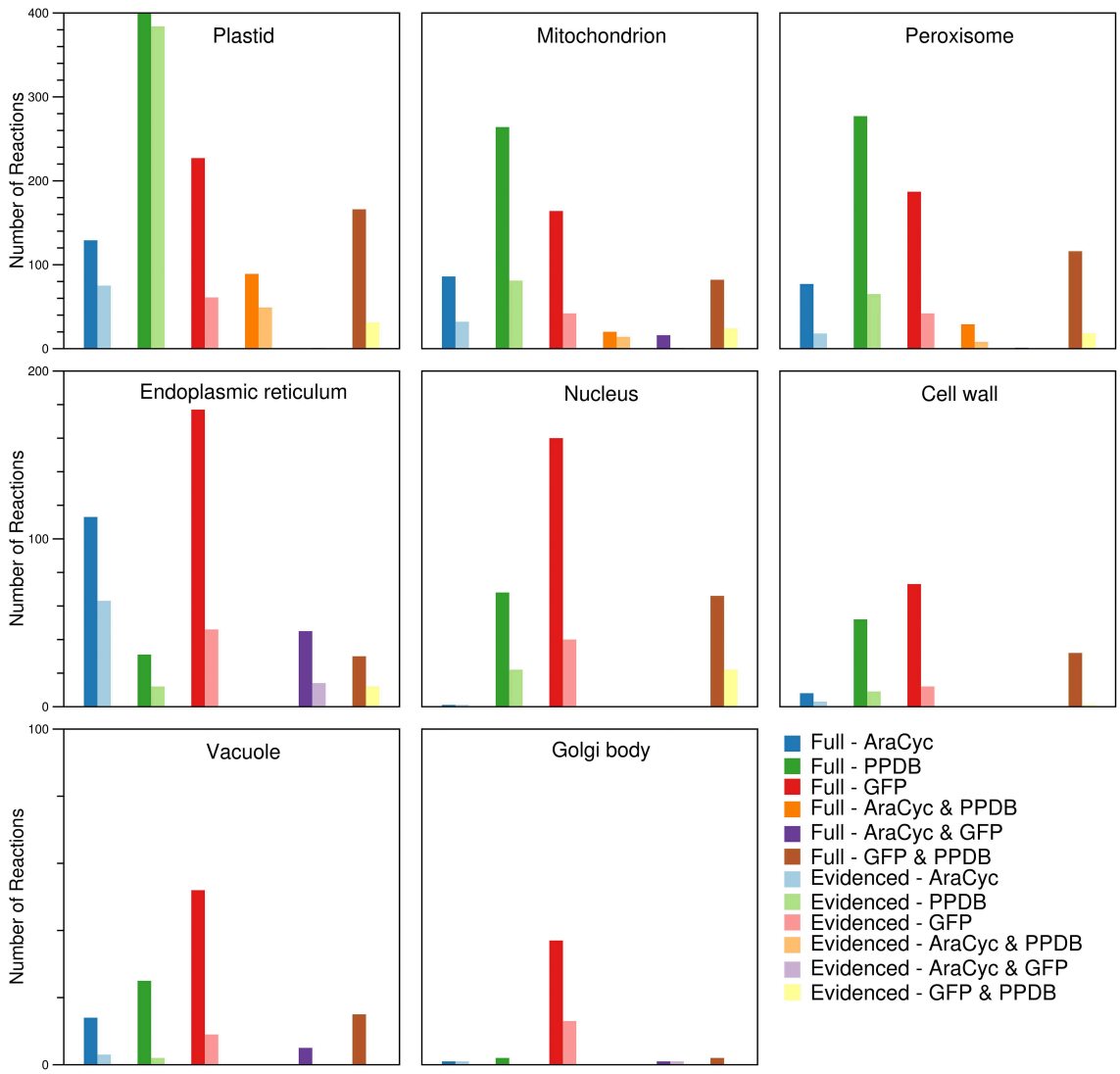

FIGURE 4 | Comparison of the sources responsible for the number of reactions found in each of the compartments in the Full and Evidenced metabolic reconstructions for maize. The assignment of reactions to compartments is done by using evidence from AraCyc, PPDB, and SUBA. In general, the number of reactions in different compartments appears to be mostly influenced by a single source. More reactions are assigned by AraCyc to the endoplasmic reticulum; by PPDB to the plastid, mitochondrion, and peroxisome; and by GFP experiments listed in SUBA to nucleus, cell wall, vacuole, and Golgi body. The pairing of PPDB and SUBA shares evidence more frequently in all compartments with the exception of the endoplasmic reticulum where there is more agreement between AraCyc and SUBA.

function of all the pathways included in the model. We tested our gapfilled models by simulating growth on heterotrophic media in the KBase environment before applying the transcript profiles.

Transcriptome-Based Metabolic Reconstructions of Maize

Maize Transcriptomics

We built transcriptome-based metabolic reconstructions of maize, derived directly from the gapfilled genome-scale Evidenced metabolic model, such that each transcriptome-based model will be a subset of the Evidenced metabolic model. To generate these transcriptome-based metabolic reconstructions, we used RNA-Seq data collated at qTeller (http://qteller.com/, downloaded on 02/04/2014). The data consists of 37 experiments from nine sources, covering a range of cells, tissues, organs, and conditions. As an initial exploration of how the transcript profiles may affect a transcriptome-based model, we computed, for each of the datasets, and at 10 different thresholds, the number of reactions in the genome-scale Full and Evidenced metabolic models for maize that would be active in the organ or tissue, and conditions from which the transcript profiles were retrieved (**Figure 5**). The threshold was applied to the reaction expression scores (Equations 7–9), and as the threshold increases,

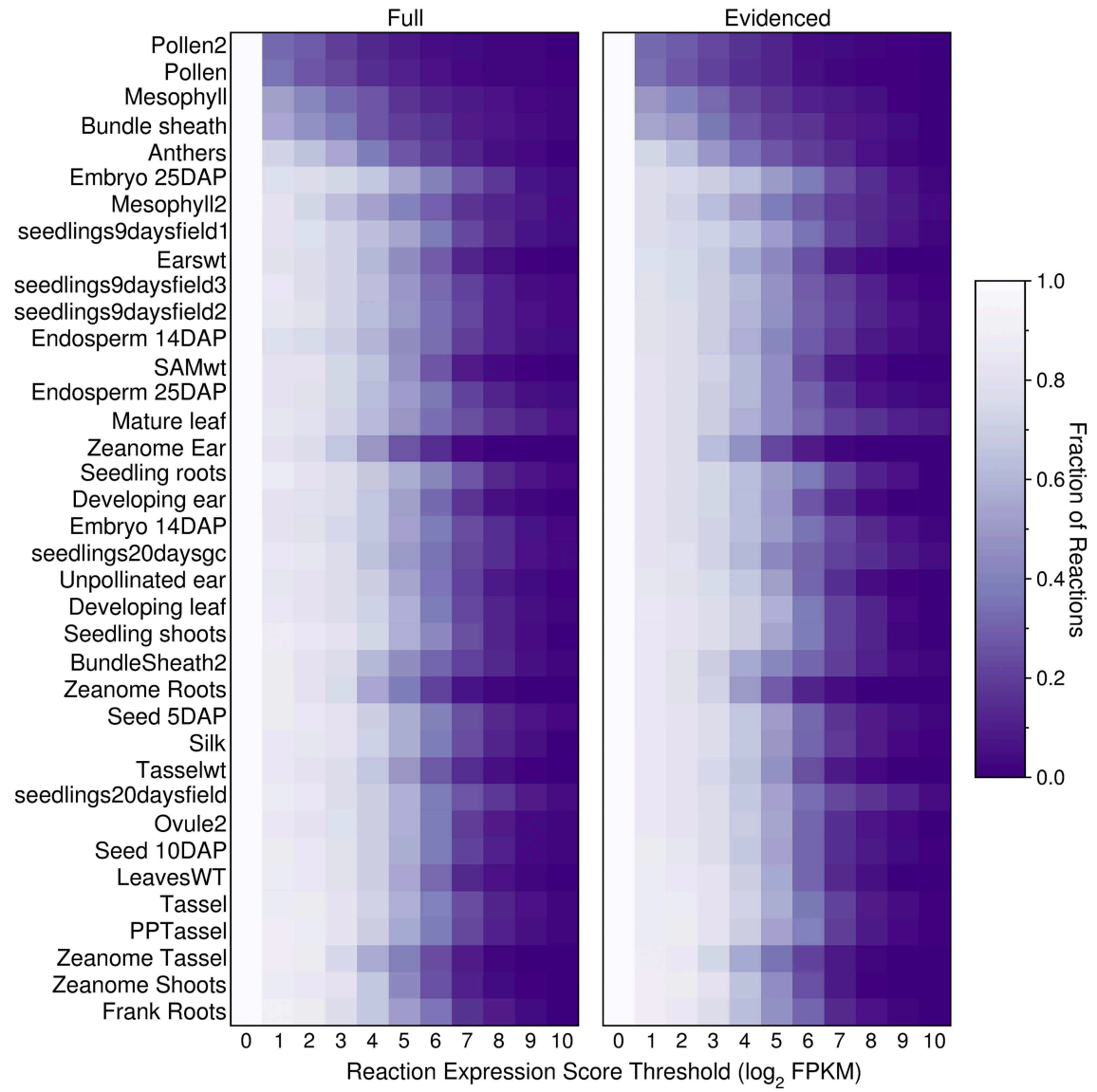

FIGURE 5 | Fraction of active reactions in the Full and Evidenced metabolic reconstructions for maize at different thresholds. An expression score is computed for each reaction (Equations 7–9) using maize transcript profiles from qTeller (http://qteller.com). The transcript profiles are ordered by the sizes of the resulting metabolic reconstruction when the threshold applied is one, thus the smaller reconstructions with fewer active reactions are positioned at the top of the figure. Although the Evidenced metabolic reconstruction has half the number of reactions found in the Full metabolic reconstruction, both models appear to shrink at similar rates when increasing the threshold. Two sets of tissues, in general, have more inactive reactions at lower thresholds: (1) reproductive tissues, such as pollen and anthers, as well as tissues consisting of single cell types such as mesophyll and bundle sheath, and (2) tissues which originated from the Zeanome project.

the number of reactions that would be active in the resulting metabolic reconstruction decreases. The results show that the smallest metabolic reconstructions are derived either from data from specific cell types (mesophyll and bundle sheath) or highly reproductive tissues (pollen and anthers); the other tissues and organs with larger reconstructions encompassed multiple cell types and in general, up to a threshold of four, show little difference in the sizes of the resulting metabolic network. Furthermore, qualitatively, it appears that the relative change in the network sizes is similar across organs and tissues in both the Full and Evidenced metabolic models. Finally, using several of the transcript profiles from the same source appears to consistently result in metabolic networks that are relatively smaller, notably those from the Zeanome dataset (http://www.ncbi.nlm.nih.gov/Traces/sra/?study=SRP011480), which is an important reminder that, when performing *in silico* experiments using transcript profiles, one must ensure they come from the same source.

To investigate further, we explored how the threshold creates gaps in the primary metabolism of transcriptome-based models. We aggregated the various pathways under nine different categories of primary metabolism as defined by the PlantSEED project (Seaver et al., 2014) and we explored how these pathways shrink in size as the threshold is increased (**Figures 6**, **7**). Overall, within each pathway category, a similar pattern is observed where the sex organs and single-cell transcript profiles result in the smaller metabolic model, and for all transcript profiles, there appears to be a similar decrease in the sizes of the pathways. However, it is notable that this pattern varies from category to category and from organ to organ or tissue to tissue. Many essential reactions that may be necessary for a derived metabolic model to operate may be inactivated by the use of a simple expression threshold. For instance, within almost every category, there are reactions for which the computed expression score is zero or constitutively low (**Figure 7**), but the reactions are essential. Reactions in the "Fatty acids" category would appear to be the most impervious to the use of a low threshold as many, if not all, of the reactions appear to exhibit a medium to high expression score across most organs and tissues. A notable example is the set of reaction expression scores computed from the transcriptome labeled Embryo_25DAP (25 days after pollination), which matches our understanding of the embryo typically being rich in lipids. As it is therefore not reasonable to use a simplistic approach to generate transcriptome-based metabolic models, we thus develop a novel method for applying the gene expression levels in transcript profiles directly to the genome-scale metabolic model (see Materials and Methods). The method attempts to activate every reaction that is associated with a highly expressed gene whilst minimizing activity of reactions associated with minimally expressed genes. The results of the generation of these models from transcript profiles using this method are found in Section Generating the Transcriptome-Based Metabolic Models. However, first we address the derivation of new biomass compositions to represent the leaf, endosperm and embryo tissues.

High-Quality Maize Biomass Equation for Leaf, Endosperm and Embryo Tissue

One use for the metabolic models we build is to predict the biosynthesis of plant biomass components. This is done by creating a specialized biomass composition reaction that contains each of the biomass components in relative proportions, and by "maximizing" biomass production when simulating growth in the metabolic model. All of the prior published metabolic models for plants have assumed a basic biomass composition that contained mostly primary metabolites. Little emphasis was placed on the diversity of compounds that a plant biosynthesizes. For our transcriptome-based metabolic models, we aim to distinguish between the functions of the models by providing a high-quality biomass composition reaction representing the organ or tissue from which the modeled transcriptomes were collected. We constructed these reaction based on an extensive literature search. Here we describe a biomass that contains more cofactors and fatty acids, supported by almost 30 literature references, including detailed quantifications. The following paragraphs briefly described the biomass composition along with the relevant references.

Amino acids

The biomass fraction attributable to protein is estimated to be 8 and 11.6% of dry weight in endosperm and embryo, respectively (Ingle et al., 1965). To quantify the relative contribution of each amino acid in the endosperm, the total amino acid context determined experimentally by Misra et al. (1972) was used with two exceptions. Firstly, the cysteine content was doubled as the reported value concerned cystine. Secondly, the glutamate:glutamine and aspartate:asparagine ratios were deduced from the composition of mature Zein proteins (Wu et al., 2012) to estimate their individual contribution. For composition of amino acids in embryo, the sequences of two globulins were used, which account for 20% of total embryo protein (Belanger and Kriz, 1989; Wallace and Kriz, 1991). Water loss due to formation of the peptide bond was taken into account.

Nucleic acids

The biomass fraction attributable to DNA was reported to be 0.038 and 0.015% in endosperm and embryo, respectively, while that attributable to RNA was reported to be 0.3 and 0.1% in endosperm and embryo, respectively (Ingle et al., 1965). The biomass fraction attributed to each nucleotide was estimated using published GC content (Haberer et al., 2005).

Carbohydrates

The endosperm biomass fraction attributable to carbohydrates was calculated to be about 90% of dry mass (Ingle et al., 1965; Alonso et al., 2011). Of this carbohydrate fraction, 77.6% is starch, 16.6% is cell walls (Alonso et al., 2011) and the remaining 5.8% is free sucrose, fructose, and glucose (Ingle et al., 1965). The reported composition of endosperm cell walls (Dewitt et al., 1999) was used to calculate the quantities of the majority of the monosaccharides. The embryo biomass fraction attributable to carbohydrates is calculated to be 58.5% (Rolletschek et al., 2005; Alonso et al., 2010). Of this carbohydrate fraction, 49.6% is starch, 42.7% is cell walls (Alonso et al., 2010) and the remaining 7.7% is free sucrose, fructose, and glucose (Rolletschek et al., 2005). The reported composition of cell walls (McCann et al., 2007) was used to calculate the quantities of the majority of the monosaccharides and the ratio of monosaccharides found in the leaf (Penning de Vries et al., 1974) was used to calculate ribose, glucuronate, and galacturonate content.

For both endosperm and embryo, the galactose, glycerol, and sulfoquinovose biomass fraction was estimated using values for galactolipids, glycerolipids, and sulfolipids, respectively (see the Section Lipids and Sterols). Finally, further evidence was used to deduce the biomass fraction of inositol (Teas, 1954).

Phenolic compounds

The cell wall of maize is considered to contain two main types of phenolic derivatives: p-coumaric acid and ferulic acid (Assabgui et al., 1993; Saulnier et al., 1995).

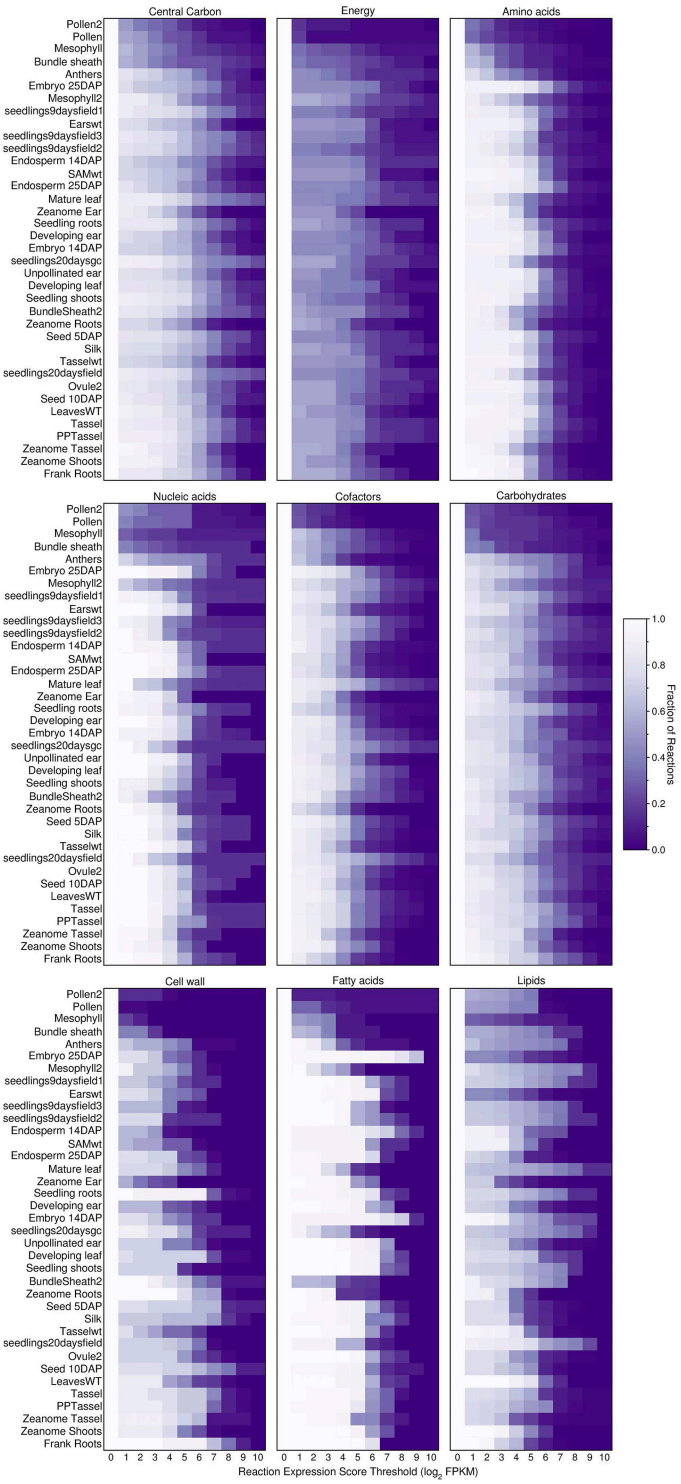

FIGURE 6 | Fraction of active reactions involved in different categories of plant primary metabolism at different thresholds. An expression score is computed for each reaction (Equations 7–9) using maize transcript profiles from qTeller (http://qteller.com). The results shown here are for the Evidenced metabolic reconstruction for maize. The figure indicates that between categories of primary metabolism and from tissue to tissue, the fraction of active reactions exhibits substantial variation. Some tissues have a high fraction of reactions active at a high threshold within certain categories, for example, within the tissue sample named "Embryo_25DAP" (25 days after pollination) and within the category of Fatty acids. This result reflects a known biological function of the embryo, as a store of lipids. The high degree of variation in the number of active reactions at different thresholds in plant primary metabolism is a strong indication that using a single gene expression threshold across an entire metabolic reconstruction may produce undesired results.

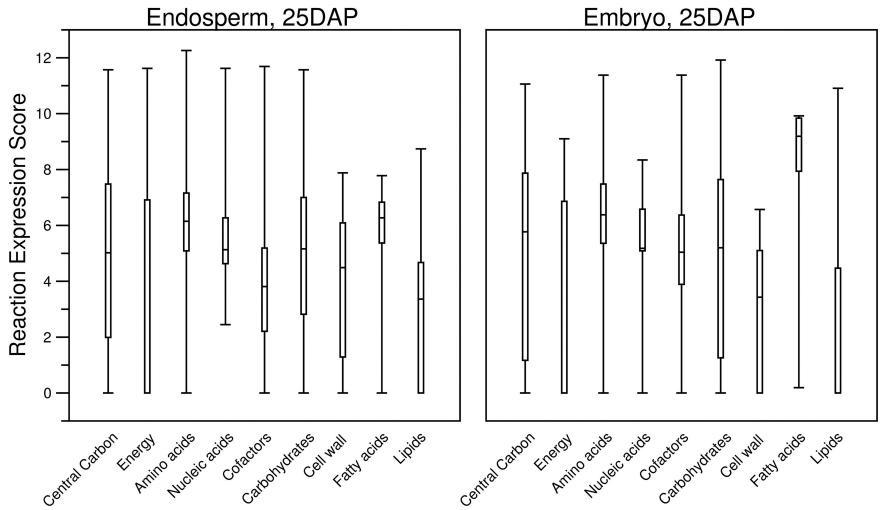

FIGURE 7 | Boxplots describing the distribution of computed reaction expression scores (Equations 7–9) from the transcript profiles of two tissues and within the different categories of plant primary metabolism. Almost every category contained at least one reaction with a reaction expression score of zero. Furthermore, for the "Energy" and "Lipids" categories, more than half of the reaction expression scores are zero. It can be seen that the median reaction expression scores for "Embryo, 25DAP" are higher, which supports the observation made for this tissue in the previous figure (**Figure 6**). Additionally, the lower quartiles of the reaction expression scores in the "Carbohydrates" and "Cell wall" categories are higher for "Endosperm, 25DAP." Both of these categories include pathways involved in sugar metabolism, and this supports the known biological function of the endosperm as a storage of starch.

Vitamins and cofactors

As key components of metabolism, we emphasized biosynthetic pathways of cofactors more than the published metabolic models, and specified the biomass fraction assigned to each of the B vitamins and other cofactors with greater accuracy. The list of vitamins and cofactors included biotin, thiamin diphosphate, NAD and derivatives, FAD and FMN, coenzyme A, 4-phosphopantetheine, tetrahydrofolate and its derivatives, α-tocopherol, ascorbate, ubiquinone-9, lipoic acid, heme, and pyridoxal-5′-phosphate (Cameron and Teas, 1948; Teas, 1954; Giri et al., 1960; Ingle et al., 1965; Metz et al., 1970; Weber, 1987; Battey and Ohlrogge, 1990; Shannon et al., 1996; Szal et al., 2003; Tumaney et al., 2004; Shi et al., 2005; Drozak and Romanowska, 2006; Hu et al., 2006; Naqvi et al., 2009; Perez-Lopez et al., 2010; Richter et al., 2010; Enami et al., 2011; Spielbauer et al., 2013; Seaver et al., 2014).

Pigments

Two pigments were included in our endosperm and embryo biomass: β-carotene, and lutein (Weber, 1987).

Lipids and sterols

Lipids represent 1.5 and 32.6% of the biomass of endosperm and embryo, respectively (Weber, 1979). The biomass composition of fatty acids and sitosterol, campesterol, stigmasterol, and phytosphingosine in this study were based on those reported by Weber (1979). Galactose, glycerol, and sulfoquinovose content were also calculated based on the lipid composition.

Carboxylic acids and other compounds

Many other compounds compose plant biomass, and we included here a list of a subset of these for which a value is reported in the literature: *cis*-aconitate, citrate, malate, oxaloacetate lactate (Skogerson et al., 2010; Rolletschek et al., 2011), and S-adenosylmethionine (Apelbaum and Yang, 1981). 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 (Penning de Vries et al., 1974).

Generating the Transcriptome-Based Metabolic Models

We used the novel transcriptome-based gapfilling approach (see Materials and Methods) along with three separate transcript profiles to generate metabolic models that are specific to the leaf, endosperm and embryo, which are named "Leaves 20-day old seedling – field," "Endosperm 25 days after pollination," and "Embryo 25 days after pollination" (Davidson et al., 2011) http://www.ncbi.nlm.nih.gov/bioproject/80041). We used these three transcript profiles in particular because they came from the same experiments and, therefore, were processed in a similar manner.

We applied the three transcriptome profiles separately to the Evidenced metabolic model of maize (see Materials and methods Section) to generate three separate metabolic models that can grow in heterotrophic media. All three metabolic models contained an average of 2302 reactions (**Table 1**), which is 88% of the number of reactions in the Evidenced model, and there are 2153 reactions that are found in all three of them. By comparison, the final compartmentalized model created by Mintz-Oron et al. (2012) for *Arabidopsis* has 3508 reactions and the resulting tissue-specific models generated from their work has on average 2848 reactions, which is 81% of the reactions in their full model.

TABLE 2 | Comparison of prior published model of maize with the models generated by this work using the percentage of blocked reactions and the spearman rank correlation coefficient when using fluxomics data (*p*-value in parentheses).

Type/Tissue	Blocked reactions (%)	Endosperm	Embryo
iRS1563*	53	0.69 (2.3×10^{-3})	0.46 (7.5×10^{-2})
Full	30	0.99 (9.4×10^{-51})	0.99 (7.1×10^{-54})
Evidenced	21	0.99 (4.7×10^{-34})	0.83 (1.7×10^{-10})
Evidenced/Endosperm	16	0.99 (1.3×10^{-29})	n/a
Evidenced/Embryo	16	n/a	0.83 (8.6×10^{-10})

**Saha et al. (2011).*

This result indicates that our approach, while using a full set of gene expression data for the maize transcript profiles to generate smaller models, results in models whose sizes are similar to other work on generating organ and tissue-specific models for a plant.

Comparison of Fluxes in Tissue-Specific Metabolic Reconstructions to Fluxomics Data

We have described a process that generates and refines metabolic models in three steps, generating metabolic models at each step. We can now show how these metabolic models not only compare with fluxomics data, but how that comparison improves at each step, resulting in transcriptome-based models with the closest fit to the original fluxomics data.

The experimental data we used were fluxes for central carbon metabolism estimated using ^{14}C labeling in two different tissues, the embryo and endosperm (Alonso et al., 2010, 2011). The reactions from these two studies were matched to the reactions in the models, and we used the approach described in Section Comparison with Estimated Fluxomics Data for Embryo and Endosperm to fit the fluxes within the models to the experimentally determined fluxes. We report the Spearman correlation and its *p*-value in **Table 2**, showing that the correlation is high for both transcriptome-based models. This result indicates that the central carbon metabolism of the models generated in this work is able to perform as observed in the original tissues. The reactions used here have a median expression score of 6.60 and 7.17 in the embryo and endosperm transcriptomics dataset, respectively, but the lowest expression score is ∼1.2 for both tissues. This last statement in turn exemplifies the importance of our approach, in ensuring that reactions with a low expression score are still included in model generated from a transcript profile if considered to be essential for the metabolic functioning of the organ or tissue.

Discussion

In this manuscript, we created a total of seven metabolic reconstructions for two species (see Supplementary Material). In succession, we created two Full metabolic reconstructions for *Arabidopsis* and maize, comprised of many possible sources of plant biochemistry reconciled into single large networks. These Full models also included many predicted gene-reaction associations, a subset for which we found evidence either in the literature or via human inference, and we used these to create a more reliable metabolic reconstruction for the two species. Finally, via use of a novel, simple and fast organ and tissue-specific pathway gapfilling method, along with well-curated biomass for the leaf, endosperm and embryo, we generated three metabolic models specific for these organ and tissues. The evidence that we used, for both the genes whose products catalyze the reactions and the localization of gene products in different compartments, is comprehensive and reliable.

Our approach allows us to create relatively large metabolic reconstructions that compare favorably to the prior published metabolic models, albeit with a smaller set of gene-reaction associations. This enables us to apply transcriptome data with a high degree of confidence. The approach is validated by the fact that the embryo and endosperm models retained nearly every reaction of central carbon metabolism. This was done both by the body of evidence available for the gene-reaction associations, and the pathway gapfilling method which included reactions with a low expression score, but were essential to the models. Finally, it was shown that the same models can be active and able to replicate the activity observed in published experimental fluxomics datasets. To date, we believe we are the first to apply such wide-ranging body of evidence to the generation of large-scale metabolic reconstructions.

All of our work was carried out through the DOE Systems Biology Knowledgebase (KBase; http://kbase.us/), an open software and data platform that aim to enable researchers to predict and ultimately design biological function. The data is publicly available within KBase workspaces named "Maize_Tissue_Models" (https://narrative.kbase.us/functional-site/#/ws/objects/Maize_Tissue_Models) and also via the PlantSEED website (http://plantseed.theseed.org). The KBase software environment allows researchers to copy the individual metabolic models and to explore the models using the suite of modeling tools available.

Acknowledgments

This work was supported by National Science Foundation Grant Number IOS-1025398, by an endowment from the C.V. Griffin Sr. Foundation, and by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy under Contract Number DE-ACO2-06CH11357, as part of the DOE Systems Biology Knowledgebase.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2015.00142/abstract

Data Sheet 1.ZIP

Zip file of metabolic models in SBML format. All seven metabolic models, the Full metabolic and Evidenced metabolic models for *Arabidopsis* and maize and the three tissue-specific metabolic models for maize leaf, endosperm, and embryo are available for download in SBML format.

Data Sheet 2.XLSX

Spreadsheet of metabolic models. An Excel spreadsheet containing details for the seven metabolic models and also containing details of their relationships with prior published models, and the reaction-gene and gene-compartment evidence.

References

Alonso, A. P., Dale, V. L., and Shachar-Hill, Y. (2010). Understanding fatty acid synthesis in developing maize embryos using metabolic flux analysis. *Metab. Eng.* 12, 488–497. doi: 10.1016/j.ymben.2010.04.002

Alonso, A. P., Val, D. L., and Shachar-Hill, Y. (2011). Central metabolic fluxes in the endosperm of developing maize seeds and their implications for metabolic engineering. *Metab. Eng.* 13, 96–107. doi: 10.1016/j.ymben.2010.10.002

Apelbaum, A., and Yang, S. F. (1981). Biosynthesis of stress ethylene induced by water deficit. *Plant Physiol.* 68, 594–596. doi: 10.1104/pp.68.3.594

Assabgui, R. A., Reid, L. M., Hamilton, R. I., and Arnason, J. T. (1993). Correlation of kernel (E)-ferulic acid content of maize with resistance to *Fusarium graminearum*. *Phytopathology* 83, 949–953. doi: 10.1094/Phyto-83-949

Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., et al. (2008). Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* 320, 938–941. doi: 10.1126/science.1157956

Battey, J. F., and Ohlrogge, J. B. (1990). Evolutionary and tissue-specific control of expression of multiple acyl-carrier protein isoforms in plants and bacteria. *Planta* 180, 352–360. doi: 10.1007/BF01160390

Belanger, F. C., and Kriz, A. L. (1989). Molecular characterization of the major maize embryo globulin encoded by the glb1 gene. *Plant Physiol.* 91, 636–643. doi: 10.1104/pp.91.2.636

Cameron, J. W., and Teas, H. J. (1948). The relation between nicotinic acid and carbohydrates in a series of maize endosperm genotypes. *Proc. Natl. Acad. Sci. U.S.A.* 34, 390–398. doi: 10.1073/pnas.34.8.390

Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., et al. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 40, D742–D753. doi: 10.1093/nar/gkr1014

Chang, R. L., Ghamsari, L., Manichaikul, A., Hom, E. F., Balaji, S., Fu, W., et al. (2011). Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol. Syst. Biol.* 7:518. doi: 10.1038/msb.2011.52

Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., et al. (2009). Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Comput. Biol.* 5:e1000489. doi: 10.1371/journal.pcbi.1000489

Davidson, R. M., Hansey, C. N., Gowda, M., Childs, K. L., Lin, H., Vaillancourt, B., et al. (2011). Utility of RNA sequencing for analysis of maize reproductive transcriptomes. *Plant Genome* 4, 191–203. doi: 10.3835/plantgenome2011.05.0015

de Oliveira Dal'molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010a). AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol.* 152, 579–589. doi: 10.1104/pp.109.148817

de Oliveira Dal'molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010b). C4GEM, a genome-scale metabolic model to study C-4 plant metabolism. *Plant Physiol.* 154, 1871–1885. doi: 10.1104/pp.110.166488

de Oliveira Dal'molin, C. G., Quek, L. E., Palfreyman, R. W., and Nielsen, L. K. (2011). AlgaGEM–a genome-scale metabolic reconstruction of algae based on the *Chlamydomonas reinhardtii* genome. *BMC Genomics* 12(Suppl. 4):S5. doi: 10.1186/1471-2164-12-S4-S5

Dewitt, G., Richards, J., Mohnen, D., and Jones, A. M. (1999). Comparative compositional analysis of walls with two different morphologies: archetypical versus transfer-cell-like. *Protoplasma* 209, 238–245. doi: 10.1007/BF01453452

Drozak, A., and Romanowska, E. (2006). Acclimation of mesophyll and bundle sheath chloroplasts of maize to different irradiances during growth. *Biochim. Biophys. Acta* 1757, 1539–1546. doi: 10.1016/j.bbabio.2006.09.001

Enami, K., Ozawa, T., Motohashi, N., Nakamura, M., Tanaka, K., and Hanaoka, M. (2011). Plastid-to-nucleus retrograde signals are essential for the expression of nuclear starch biosynthesis genes during amyloplast differentiation in tobacco BY-2 cultured cells. *Plant Physiol.* 157, 518–530. doi: 10.1104/pp.111.178897

Giri, K. V., Rao, N. A., Cama, H. R., and Kumar, S. A. (1960). Studies on flavinadenine dinucleotide-synthesizing enzyme in plants. *Biochem. J.* 75, 381–386.

Grafahrend-Belau, E., Junker, A., Eschenroder, A., Muller, J., Schreiber, F., and Junker, B. H. (2013). Multiscale metabolic modeling: dynamic flux balance analysis on a whole-plant scale. *Plant Physiol.* 163, 637–647. doi: 10.1104/pp.113.224006

Grafahrend-Belau, E., Schreiber, F., Koschutzki, D., and Junker, B. H. (2009). Flux balance analysis of barley seeds: a computational approach to study systemic properties of central metabolism. *Plant Physiol.* 149, 585–598. doi: 10.1104/pp.108.129635

Haberer, G., Young, S., Bharti, A. K., Gundlach, H., Raymond, C., Fuks, G., et al. (2005). Structure and architecture of the maize genome. *Plant Physiol.* 139, 1612–1624. doi: 10.1104/pp.105.068718

Hay, J., and Schwender, J. (2011a). Computational analysis of storage synthesis in developing *Brassica napus* L. (oilseed rape) embryos: flux variability analysis in relation to (1)(3)C metabolic flux analysis. *Plant J.* 67, 513–525. doi: 10.1111/j.1365-313X.2011.04611.x

Hay, J., and Schwender, J. (2011b). Metabolic network reconstruction and flux variability analysis of storage synthesis in developing oilseed rape (*Brassica napus* L.) embryos. *Plant J.* 67, 526–541. doi: 10.1111/j.1365-313X.2011.04613.x

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. (2013). InChI–the worldwide chemical structure identifier standard. *J. Cheminform.* 5:7. doi: 10.1186/1758-2946-5-7

Henry, C. S., Dejongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization, and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 1672, 1–6. doi: 10.1038/nbt.1672

Henry, C. S., Zinner, J., Cohoon, M., and Stevens, R. (2009). *i*Bsu1103: a new genome scale metabolic model of *B. subtilis* based on SEED annotations. *Genome Biol.* 10:R69. doi: 10.1186/gb-2009-10-6-r69

Hu, W. H., Shi, K., Song, X. S., Xia, X. J., Zhou, Y. H., and Yu, J. Q. (2006). Different effects of chilling on respiration in leaves and roots of cucumber (*Cucumis sativus*). *Plant Physiol. Biochem.* 44, 837–843. doi: 10.1016/j.plaphy.2006.10.016

Ingle, J., Beitz, D., and Hageman, R. H. (1965). Changes in composition during development and maturation of maize seeds. *Plant Physiol.* 40, 835–839. doi: 10.1104/pp.40.5.835

Ingle, R. A. (2011). Histidine biosynthesis. *Arabidopsis Book* 9:e0141. doi: 10.1199/tab.0141

Jerby, L., Shlomi, T., and Ruppin, E. (2010). Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol. Syst. Biol.* 6, 401. doi: 10.1038/msb.2010.56

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi: 10.1093/nar/gkt1076

Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., et al. (2014). Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.* 42, D546–D552. doi: 10.1093/nar/gkt979

Kumar, V. S., Dasika, M. S., and Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. *BMC Bioinform.* 8:212. doi: 10.1186/1471-2105-8-212

Kumar, V. S., and Maranas, C. D. (2009). GrowMatch: an automated method for reconciling *in silico*/*in vivo* growth predictions. *PLoS Comput. Biol.* 5:e1000308. doi: 10.1371/journal.pcbi.1000308

Latendresse, M. (2014). Efficiently gap-filling reaction networks. *BMC Bioinform.* 15:225. doi: 10.1186/1471-2105-15-225

Lee, D., Smallbone, K., Dunn, W. B., Murabito, E., Winder, C. L., Kell, D. B., et al. (2012). Improving metabolic flux predictions using absolute gene expression data. *BMC Syst. Biol.* 6:73. doi: 10.1186/1752-0509-6-73

McCann, M. C., Defernez, M., Urbanowicz, B. R., Tewari, J. C., Langewisch, T., Olek, A., et al. (2007). Neural network analyses of infrared spectra for classifying cell wall architectures. *Plant Physiol.* 143, 1314–1326. doi: 10.1104/pp.106.093054

Metz, J., Lurie, A., and Konidaris, M. (1970). A note on the folate content of uncooked maize. *S. Afr. Med. J.* 44, 539–541.

Mintz-Oron, S., Meir, S., Malitsky, S., Ruppin, E., Aharoni, A., and Shlomi, T. (2012). Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proc. Natl. Acad. Sci. U.S.A.* 109, 339–344. doi: 10.1073/pnas.1100358109

Misra, P. S., Jambunathan, R., Mertz, E. T., Glover, D. V., Barbosa, H. M., and McWhirter, K. S. (1972). Endosperm protein synthesis in maize mutants with increased lysine content. *Science* 176, 1425–1427. doi: 10.1126/science.176.4042.1425

Mo, M. L., Palsson, B. O., and Herrgard, M. J. (2009). Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* 3:37. doi: 10.1186/1752-0509-3-37

Monaco, M. K., Sen, T. Z., Dharmawardhana, P. D., Ren, L., Schaeffer, M., Naithani, S., et al. (2013). Maize metabolic network construction and transcriptome analysis. *Plant Gen.* 6. doi: 10.3835/plantgenome2012.09.0025. Available online at: https://www.crops.org/publications/citation-manager/tpg/6/1/plantgenome2012.09.0025

Mueller, L. A., Zhang, P., and Rhee, S. Y. (2003). AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol.* 132, 453–460. doi: 10.1104/pp.102.017236

Nakaya, A., Katayama, T., Itoh, M., Hiranuka, K., Kawashima, S., Moriya, Y., et al. (2013). KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Res.* 41, D353–D357. doi: 10.1093/nar/gks1239

Naqvi, S., Zhu, C., Farre, G., Ramessar, K., Bassie, L., Breitenbach, J., et al. (2009). Transgenic multivitamin corn through biofortification of endosperm with three vitamins representing three distinct metabolic pathways. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7762–7767. doi: 10.1073/pnas.0901412106

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98. doi: 10.1038/nrg2934

Penning de Vries, F. W., Brunsting, A. H., and van Laar, H. H. (1974). Products, requirements and efficiency of biosynthesis: a quantitative approach. *J. Theor. Biol.* 45, 339–377. doi: 10.1016/0022-5193(74)90119-2

Perez-Lopez, U., Robredo, A., Lacuesta, M., Sgherri, C., Mena-Petite, A., Navari-Izzo, F., et al. (2010). Lipoic acid and redox status in barley plants subjected to salinity and elevated CO2. *Physiol. Plant.* 139, 256–268. doi: 10.1111/j.1399-3054.2010.01361.x

Pilalis, E., Chatziioannou, A., Thomasset, B., and Kolisis, F. (2011). An *in silico* compartmentalized metabolic model of *Brassica napus* enables the systemic study of regulatory aspects of plant central metabolism. *Biotechnol. Bioeng.* 108, 1673–1682. doi: 10.1002/bit.23107

Poolman, M. G., Kundu, S., Shaw, R., and Fell, D. A. (2013). Responses to light intensity in a genome-scale model of rice metabolism. *Plant Physiol.* 162, 1060–1072. doi: 10.1104/pp.113.216762

Poolman, M. G., Miguet, L., Sweetlove, L. J., and Fell, D. A. (2009). A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant Physiol.* 151, 1570–1581. doi: 10.1104/pp.109.141267

Richter, A., Peter, E., Pors, Y., Lorenzen, S., Grimm, B., and Czarnecki, O. (2010). Rapid dark repression of 5-aminolevulinic acid synthesis in green barley leaves. *Plant Cell Physiol.* 51, 670–681. doi: 10.1093/pcp/pcq047

Rolletschek, H., Koch, K., Wobus, U., and Borisjuk, L. (2005). Positional cues for the starch/lipid balance in maize kernels and resource partitioning to the embryo. *Plant J.* 42, 69–83. doi: 10.1111/j.1365-313X.2005.02352.x

Rolletschek, H., Melkus, G., Grafahrend-Belau, E., Fuchs, J., Heinzel, N., Schreiber, F., et al. (2011). Combined noninvasive imaging and modeling approaches reveal metabolic compartmentation in the barley endosperm. *Plant Cell* 23, 3041–3054. doi: 10.1105/tpc.111.087015

Saha, R., Suthers, P. F., and Maranas, C. D. (2011). Zea mays iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE* 6:e21784. doi: 10.1371/journal.pone.0021784

Satish Kumar, V., Dasika, M. S., and Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. *BMC Bioinform.* 8:212. doi: 10.1186/1471-2105-8-212

Saulnier, L., Marot, C., Chanilaud, E., and Thibault, J.-F. (1995). Cell wall polysaccharide interactions in maize bran. *Carbohydr. Polym.* 26, 279–287. doi: 10.1016/0144-8617(95)00020-8

Schnable, J. C., and Freeling, M. (2011). Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PLoS ONE* 6:e17855. doi: 10.1371/journal.pone.0017855

Schnable, J. C., Springer, N. M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4069–4074. doi: 10.1073/pnas.1101368108

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Schwender, J., and Hay, J. O. (2012). Predictive modeling of biomass component tradeoffs in *Brassica napus* developing oilseeds based on *in silico* manipulation of storage metabolism. *Plant Physiol.* 160, 1218–1236. doi: 10.1104/pp.112.203927

Seaver, S. M., Gerdes, S., Frelin, O., Lerma-Ortiz, C., Bradbury, L. M., Zallot, R., et al. (2014). High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. *Proc. Natl. Acad. Sci. U.S.A.* 111, 9645–9650. doi: 10.1073/pnas.1401329111

Shannon, J. C., Pien, F. M., and Liu, K. C. (1996). Nucleotides and nucleotide sugars in developing maize endosperms (synthesis of ADP-glucose in brittle-1). *Plant Physiol.* 110, 835–843.

Shi, J., Wang, H., Hazebroek, J., Ertl, D. S., and Harp, T. (2005). The maize low-phytic acid 3 encodes a myo-inositol kinase that plays a role in phytic acid biosynthesis in developing seeds. *Plant J.* 42, 708–719. doi: 10.1111/j.1365-313X.2005.02412.x

Simkin, A. J., Guirimand, G., Papon, N., Courdavault, V., Thabet, I., Ginis, O., et al. (2011). Peroxisomal localisation of the final steps of the mevalonic acid pathway in planta. *Planta* 234, 903–914. doi: 10.1007/s00425-011-1444-6

Skirycz, A., and Inze, D. (2010). More from less: plant growth under limited water. *Curr. Opin. Biotechnol.* 21, 197–203. doi: 10.1016/j.copbio.2010.03.002

Skogerson, K., Harrigan, G. G., Reynolds, T. L., Halls, S. C., Ruebelt, M., Iandolino, A., et al. (2010). Impact of genetics and environment on the metabolite composition of maize grain. *J. Agric. Food Chem.* 58, 3600–3610. doi: 10.1021/jf903705y

Sonnhammer, E. L., and Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 18, 619–620. doi: 10.1016/S0168-9525(02)02793-2

Spielbauer, G., Li, L., Romisch-Margl, L., Do, P. T., Fouquet, R., Fernie, A. R., et al. (2013). Chloroplast-localized 6-phosphogluconate dehydrogenase is critical for maize endosperm starch accumulation. *J. Exp. Bot.* 64, 2231–2242. doi: 10.1093/jxb/ert082

Stitt, M., Sulpice, R., and Keurentjes, J. (2010). Metabolic networks: how to identify key components in the regulation of metabolism and growth. *Plant Physiol.* 152, 428–444. doi: 10.1104/pp.109.150821

Sun, Q., Zybailov, B., Majeran, W., Friso, G., Olinares, P. D., and van Wijk, K. J. (2009). PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res.* 37, D969–D974. doi: 10.1093/nar/gkn654

Szal, B., Jolivet, Y., Hasenfratz-Sauder, M.-P., Dizengremel, P., and Rychter, A. M. (2003). Oxygen concentration regulates alternative oxidase expression in barley roots during hypoxia and post−hypoxia. *Physiol. Plant.* 119, 494–502. doi: 10.1046/j.1399-3054.2003.00161.x

Tanz, S. K., Castleden, I., Hooper, C. M., Vacher, M., Small, I., and Millar, H. A. (2013). SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis. *Nucleic Acids Res.* 41, D1185–D1191. doi: 10.1093/nar/gks1151

Teas, H. J. (1954). B vitamins in starchy and sugary maize endosperms. *Plant Physiol.* 29, 190–194. doi: 10.1104/pp.29.2.190

Töpfer, N., Caldana, C., Grimbs, S., Willmitzer, L., Fernie, A. R., and Nikoloski, Z. (2013). Integration of genome-scale modeling and transcript profiling reveals metabolic pathways underlying light and temperature acclimation in Arabidopsis. *Plant Cell* 25, 1197–1211. doi: 10.1105/tpc.112.108852

Tumaney, A. W., Ohlrogge, J. B., and Pollard, M. (2004). Acetyl coenzyme A concentrations in plant tissues. *J. Plant Physiol.* 161, 485–488. doi: 10.1078/0176-1617-01258

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19, 327–335. doi: 10.1101/gr.073585.107

Wallace, N. H., and Kriz, A. L. (1991). Nucleotide sequence of a cDNA clone corresponding to the maize globulin-2 gene. *Plant Physiol.* 95, 973–975. doi: 10.1104/pp.95.3.973

Wang, Y., Eddy, J. A., and Price, N. D. (2012). Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst. Biol.* 6:153. doi: 10.1186/1752-0509-6-153

Weber, E. J. (1979). The lipids of corn germ and endosperm. *J. Am. Oil Chem. Soc.* 56, 637–641. doi: 10.1007/BF02679340

Weber, E. J. (1987). Carotenoids and tocols of corn grain determined by HPLC. *J. Am. Oil Chem. Soc.* 64, 1129–1134. doi: 10.1007/BF02612988

Wu, Y., Wang, W., and Messing, J. (2012). Balancing of sulfur storage in maize seed. *BMC Plant Biol.* 12:77. doi: 10.1186/1471-2229-12-77

Zhang, P., Dreher, K., Karthikeyan, A., Chi, A., Pujar, A., Caspi, R., et al. (2010). Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.* 153, 1479–1491. doi: 10.1104/pp.110.157396

Zur, H., Ruppin, E., and Shlomi, T. (2010). iMAT: an integrative metabolic analysis tool. *Bioinformatics* 26, 3140–3142. doi: 10.1093/bioinformatics/btq602

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.