Annotation of Metabolic Genes in *Caenorhabditis elegans* and Reconstruction of iCEL1273

1. Identification of *C. elegans* Metabolic Genes

To annotate metabolic genes, we used information from four databases (KEGG, WormBase, UniProt and a published list of eukaryotic orthology groups named KOGs (Koonin et al., 2004)) and two KEGG-based databases developed in this study (myKEGG and myTree). Each resource was used to predict the nearest KEGG orthology groups (KOs) for each gene in the *C. elegans* genome (a list of *C. elegans* genes encoding 20,519 proteins in KEGG). The predictions from different resources were both visually evaluated and converted to a numerical score for computational evaluations (**Table 1**). All predictions of gene-KO associations from all resources were combined using a custom pipeline called Systematic Annotation by manual CUration and Regression (SACURE) to give the final decision for each gene (*i.e.*, determination of the KO, enzyme, and reaction, if available, based on convincing evidence). The resources used in this procedure are explained below.

KEGG

Available annotations of *C. elegans* genes were collected from KEGG database (date: June, 2014). Finding gene-KO connections was straightforward with this dataset as KEGG-annotated genes are directly connected to KOs. For computational purposes, the score data for each gene was represented by 1 for KOs associated with the gene (typically only one KO) and 0 for the rest (**Table 1**).

WormBase

Protein domain annotations were obtained from Wormmart (version WS220) and concatenated with gene descriptions downloaded from the WormBase website (from gene Overview sections using html download option) (September, 2014) to make a WormBase text string for each gene. To match these annotations with KEGG KOs, names of all KOs and all enzymes were downloaded from KEGG. For each KO, a list of all alternative names were formed by combining KO names and names of enzymes associated with the KO. For each gene, annotation in WormBase was compared to all KO names using a word matching algorithm. This algorithm gave scores from 0 to 1 for a match between a WormBase text string and every KO name, thus defining the score for every potential gene-KO association. If all words in a KO name were not matched in the WormBase text string, the score was always zero. Otherwise, the score was increased by 0.5 for every perfect word match and reduced by 0.1 for each character interruption between words in the annotation. Final score was obtained by normalizing all KO scores for a gene with the highest scoring KO (hence scores varied from 0 to 1; **Table 1**).

UniProt

Protein names, family annotations, and EC numbers were downloaded from UniProt (date: October, 2014) (Bateman et al., 2015) for every protein-coding gene in *C. elegans*. Two scores were obtained (**Table 1**). First, protein name and family annotations were concatenated to make a UniProt annotation text and scored as described above for WormBase. Secondly, if an EC number was available, gene-KO associations were established with KOs related to the EC with a score of 1, while all other KOs were scored 0.

KOG

The identifier for all eukaryotic orthology groups (KOGs) from (Koonin et al., 2004) that included a *C. elegans* gene were obtained from Wormmart (version WS220). For each *C. elegans* gene in a KOG, the name of genes from up to six other organisms (*Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Encephalitozoon cuniculi*) in the same KOG were downloaded from the NCBI web page describing the KOG. Then these genes were cross referenced with KEGG to obtain KO associations if annotated and each KO connection established this way obtained a score of 1, while the rest of the KOs were scored 0 (**Table 1**).

myKEGG

To determine an overall protein sequence score for every potential gene-KO association, we used Smith-Waterman (SW) scores between each *C. elegans* gene in KEGG and best matching genes in up to 3,073 KEGG organisms (organisms that do not have a gene with a score of 100 or higher are not provided by KEGG as this score indicates that sequence similarity is not sufficient for a match). The SW table of each *C. elegans* gene was downloaded from KEGG for best hits (BH) and reciprocal best hits (RBH) (*i.e.*, two tables were obtained per gene). In addition to the best matching gene for each organism and the corresponding SW score, the SW tables indicate the KO to which the matching gene belongs, provided that the gene is successfully annotated by KEGG. Thus, when sorted with respect to a decreasing SW score, a visual inspection of these tables show the likely KO candidates for the query *C. elegans* gene based on which KOs are populated in highest scoring matches (*i.e.*, at the top rows of the sorted table; see **Figure 1A** for an example). To simplify the dataset and to minimize false positive identifications, we used an SW score threshold of 190; matches below this threshold were considered insignificant and removed from tables. This threshold was based on KEGG annotations, where we found only two metabolic genes that were associated with KOs with SW scores <190 (out of 988 total based on association with a metabolic reaction).

To translate our visual evaluation of SW tables into a computational algorithm, we devised a formula that scored KOs for each gene according to their relative proportion in top 10 (group A), top 100 (group B), and top 1000 (group C) best matching organisms (genes) in these tables. Given a candidate gene-KO association for a gene, the query KO was scored in each one of these groups and a combined score was obtained for the KO based on **Equation 1**, where, i indicates the group, w is the weight assigned to the group ($w_A = w_B = 0.45$, $w_C = 0.1$), c is a correction factor that is needed for tables with less than 10, 100 or 200 rows for the three respective groups ($c_A = N_A/10$, $c_B = (N_B\text{-}10)/90$, and $c_C = [\min(N_C,200)\text{-}100]/100$; N is the number of rows in the particular group), and s is the SW score. The last term in **Equation 1** indicates the sum of scores of matching genes annotated with the query KO as normalized by the total score from every KO in the group. Then, scores from BH and RBH tables were further weighed to get the final myKEGG score for the query KO according to **Equation 2**. In addition, a normalized myKEGG score was calculated, where the highest scoring KO for a given gene got a score of 1.0 (**Table 1**).

$$S_{table}^{KO} = \sum_{i=A,B,C} w_i c_i \frac{\sum_i s^{KO}}{\sum_i s^{all}} \quad (1)$$

$$S_{final}^{KO} = 0.8 S_{RBH}^{KO} + 0.2 S_{BH}^{KO} \quad (2)$$

myTree

As a final aid for annotation decisions, we created a phylogenetic tree for each gene based on protein sequences. Briefly, for a query gene that is to be annotated, we determined best matches in four other well-studied organisms (*H. sapiens*, *D. melanogaster*, *A. thaliana*, and *S. cerevisiae*) and best matches (with a KO annotation in KEGG) in any organism belonging to ten selected taxanomic groups (Bacteria, Archaea, Protists, Fungi, Plants, Invertebrates, Nematodes, Arthropods, Vertebrates, and Mammals). When best matches from the four species and from the taxonomic groups were the same due to taxonomic overlaps, we obtained the next best match in the taxonomic group to add to the tree. If the best match in any case was not a RBH, then the best reciprocal hit in *C. elegans* gene was also included in the tree. In addition, up to 5 potential paralogs of the query gene in *C. elegans* genome (top 5 matches) were used even if not captured as a reciprocal hit. In all matches, an SW score threshold of 200 was required, and when a match was not found, that organism, taxonomic group, or candidate paralog was excluded from the tree. The protein sequences of all available matches were downloaded from KEGG and aligned by MUSCLE (Edgar, 2004). MUSCLE was also used to create phylogenetic trees with the "maketree" function and resulting PHY file was converted to an SVG image using custom PYTHON scripts. An example is provided in **Figure 1B**.

While visual inspection of phylogenetic trees was very important for annotation decisions, conversion of these evaluations into an algorithm was necessary for SACURE. Thus we obtained two scores that quantitatively defined the information found in these trees. First was a cluster score to define the relatedness of the query gene to KOs in the same lineage in the tree. Proportion of each KO assigned to genes sharing the same lineage with the query gene (*i.e.*, branching from the same node plus up to two prior nodes on the tree) was calculated. Starting from the lowest node, for every node up to the third node in a row that covers both the query gene and the evaluated KO, proportion of the KO was multiplied by 1/3 and added to a score sum for the KO. Thus, only KOs that shared the lowest node with the query gene could get a score different than 0. Unannotated genes (*i.e.*, genes without a KO association) were included in the calculation of these proportions. The cluster score gets a maximum value of 1 (**Table 1**) (*i.e.*, when all three lowest nodes covering the query gene are dominated by one KO). The other tree score was based on the entire tree, where the cumulative similarity score of each KO in the tree (*i.e.*, the sum of reciprocals of distance from query gene for every gene associated with that KO) was calculated and the resulting values were normalized by the average of two highest scoring KOs (regular normalization by maximum score was avoided to reward the highest score only to KOs that totally dominated the trees). Unknown KOs for unannotated genes were all included in the scoring as a single KO. This method yielded a

cluster score between 0 and 2 (**Table 1**). See **Figure 1B** for an example for tree and cluster scores.

Systematic Annotation by Manual Curation and Regression (SACURE)

We annotated metabolic genes in *C. elegans* by manual curation reinforced by an algorithm that verified and rationalized our decision-making process based on the variables and scores described above. First, a set of candidate metabolic genes was determined based on association with a metabolic KO that required a minimum myKEGG score of 0.0004 and additional evidence in at least in one of the four external databases used (KEGG, WormBase, UniProt, and KOG). A metabolic KO was defined as any KO that is linked to an enzyme or a reaction in KEGG database. The small threshold for myKEGG was set to minimize the number of false negatives so that manual curation was feasible. Only two metabolic genes annotated by KEGG were missed at this threshold; higher thresholds increased this number and were therefore avoided. The resulting set had 2,850 candidate protein sequences with evidence for association with at least one enzyme or reaction in KEGG database.

Potential gene-KO associations were manually inspected based on the evidence from different resources. After an initial evaluation, we started training a logistic regression function, which determined the weights of each annotation resource in the decision-making process. The input of the function was scores from all resources (**Table 1**) for all possible gene-KO associations, and the output was a probability value (P) of accepting an association, with a probability greater than 0.5 indicating an acceptance and one lower than this value indicating a rejection (see **Figure 2A** for the final function). This function was best fitted to the manual decisions using the *mnrfit* routine of MATLAB (version R2014a) (The MathWorks, Inc., Natick, MA). We then checked how the output of this function fitted to the manual decisions. Misfits resulted in one of two actions before the next step: (i) some decisions were wrong or inconsistent with the rest of the decisions because of human errors and these were corrected; or (ii) some decisions could not be captured by the logistic function because evidence in some of the resources was not adequately interpretable by our scoring algorithms (most frequently, tree scores were underestimated when a tree was dominated by *C. elegans* paralogs [**Figure 1C**]) and these decisions were separated from the evaluation list as irregulars (see below). Then, logistic function fitting was repeated with the remaining regular decisions, and this process was iteratively continued, until 2,353 genes remained in the regular set with 1,704 manually accepted gene-KO associations in 1,704 genes, 13,763 manually rejected associations in all genes, and only 11 misfits to algorithmic decisions. The weights of the final logistic function for each resource are shown in **Table 1**.

We used the trained logistic function to divide our annotation decisions into two categories: regular (with a defined formula based on the calculated weights) and irregular (based on an exception that overrules this formula), thereby rationalizing all of our decisions with some defined basis. In addition, we divided our decisions into three confidence levels based on the *p*-values from logistic function and whether the association was grouped as regular or irregular: (1) low confidence, regular with $0.5 < P \leq 0.9$ or irregular; (2) medium confidence, regular with $0.9 < P \leq 0.99$; and (3) high confidence, regular with $0.99 < P$.

To establish the final set of annotated metabolic genes and reactions for metabolic network reconstruction, we first modified the definition of metabolic KOs and enzymes. We removed 37 enzymes, as these were associated with functions such as protein kinases or ubiquitin modifications, and were therefore not relevant to the design of our metabolic model. We also added 91 new KOs to the list of metabolic orthology groups as their connections to KEGG enzymes or reactions were not clear in the database links and were to be established manually (*e.g.*, K02272 is a KO associated with cytochrome c oxidase subunit 7c, but the association with the corresponding enzyme EC 1.9.3.1 was not available in KEGG). Gene associations with these additional KOs were evaluated with the trained logistic function followed by manual curation, adding 109 genes to the regular set. After all these changes, the number of accepted gene-KO associations was 1,182 in our regular set and 180 in our irregular set. Out of 180 irregular decisions, 32 were changed to regular as the final logistic function actually captured these decisions (this was not the case initially as they were not captured by earlier versions of the model during training, and were therefore categorized as irregular). An additional set of 9 gene-KO associations were found among the set of genes with a high myKEGG score but no evidence from databases (ignored during manual evaluations) with the help of the trained logistic function. These additions were manually confirmed as well. Finally, for a set of 64 genes, we indirectly established connections to metabolic reactions although these genes could not be associated with any KOs directly. Specifically, we incorporated genes for which all candidate KOs (or enzymes) overlapped in a set of reactions. On the overall, we obtained 1,435 SACURE-annotated genes distributed into different confidence categories as shown in **Figure 2B**. Some of the reactions in **Figure S2B** were generic reactions and some were repeated (i.e. two reaction IDs in KEGG indicated the same biochemical reaction). We removed most generic reactions (those with specific versions available in the database) and kept only one of each of the repeated reaction pairs in the rest of the analysis, which resulted in a reduction of 81 reactions from the annotation set.

Validation of SACURE

To check if the trained logistic function robustly captured our regular decisions, we performed leave-one-out cross validation. Testing one gene-KO decision at a time in 3,408 cases (all 1,704 accepted associations and as many rejected associations that were randomly picked), we first removed a decision, then refitted the function to the remaining decisions, predicted the decision that was left out, and compared this prediction to the original decision. Out of 3,408 tests, and excluding the 11 misfits, only 4 decisions originally picked by the logistic function became wrong during cross validation (0.1% error rate). This cross validation test proves that the trained logistic function (**Figure 2A**) captures our regular manual decisions.

We further evaluated the predictive power of the trained and validated logistic function in retrospect, by comparing algorithmic decisions with conclusions from SACURE. In total, SACURE pipeline yielded curated decisions for 2,972 genes including both core metabolic genes and others associated with signaling reactions. Logistic function decisions for 174 (5.9%) of these 2,972 genes resulted in false negatives (algorithmic null association was manually overruled by a positive gene-KO association) and 28 (0.9%) false positives (algorithmic decision was manually rejected). The low disagreement rate (6.8%) between manual and algorithmic decisions indicates

that vast majority of the annotations made in this study are based on an annotation formula, as represented by the weights of the logistic function (Table 1).

During the reconstruction process, 185 genes were re-annotated to complement gene-reaction associations in a network context (see below). Among these, 147 annotations were missed by SACURE, which makes about 12% of the model genes and 7% of all curated genes in this study. Although more annotations are certainly needed for a more complete picture of *C. elegans* metabolism, the fact that 88% of genes that make a mathematically functional global-scale network model came from this annotation pipeline also validates the approach taken in this study.

Availability and potential applications of SACURE

The annotation database obtained for the *C. elegans* genome is available at WormFlux, with 3,018 curated decisions (including those mentioned above plus curations made during the reconstruction process) and 17,326 non-curated decisions, the latter set showing purely algorithmic results for mostly non-metabolic genes. The low predictive error rate mentioned above may or may not be valid for the non-metabolic gene set, as the training of the decision function was carried out by metabolic genes, so non-curated decisions should be used with care. The approach developed in this study may also be useful for annotation of metabolic genes in other genomes found in KEGG, by replacing WormBase descriptions with other organism databases, or by using a different set of descriptive annotation resources (note that one of the current resources, KOG, is limited to only 6 other organisms). Either way, the logistic function would need to be retrained by manual curation as the current rules (weights) cannot be generalized to other genomes (*e.g.*, due to differential levels of completion in KEGG database, different annotation sources, etc.). The computational tools used in SACURE (myKEGG, myTree, and word-matching algorithms) are not standalone applications as they are dependent on KEGG for SW tables (myKEGG and myTree), MUSCLE for sequence alignment (myTree), and text input from descriptive databases for enzyme name matching (word-matching algorithms). Our customized codes used in this pipeline are available for potential users upon request.

2. Reconstruction of a Template *C. elegans* Metabolic Network: Biomass, Transport, and Demand/Sink Reactions

Degradation of bacterial biomass

The degradation of bacterial biomass is represented by Degradation-type reactions in with DGR header (29 reactions in total). All products of degradation are made exportable, which means that the model is not constrained to using a constant proportion of different materials and can waste food in excess. Importantly, degradation was established such that 1 unit of bacterial intake (reaction EXC0001) amounts to 1 g of material in standard flux units (mmoles/g dW/h, where dW denotes the dry weight of *C. elegans* used in flux normalization).

The coefficients in the degradation reactions are a function of the composition and formulation of different components of the bacterial biomass. This biomass composition was based on that of *E. coli* in (Neidhardt et al., 1990) except for phospholipids and the soluble component. Phospholipid composition was adjusted to the OP50 strain (standard

diet of *C. elegans*) according to (Satouchi et al., 1993). Only essential metabolites (required by biomass assembly or demand reactions) were included in the soluble component. The fraction of most of these compounds in the overall biomass was based on *E. coli* metabolome database (ECMDB) (Guo et al., 2013) except for vitamin B6 components (approximated based on (Dempsey, 1971)), iron-related compounds (approximated based on (Matzanke et al., 1989)), and coenzyme A, which was set arbitrarily since the concentration given in ECMDB exceeded the limit for the proportion of the entire soluble component in bacterial biomass.

Assembly of C. elegans biomass

The assembly of *C. elegans* biomass was represented by Biomass-type reactions with the BIO header (19 reactions in total). Four different biomass reactions (biomass reaction is defined as the final step of an assembly) were used to represent four different forms of animal biomass mainly depending on the absence/presence of DNA (to address cell division) and storage compounds (triacylglycerides [TAG], glycogen, and trehalose). These are BIO0100 (no DNA, with storage), BIO0101 (no DNA, no storage), BIO0102 (with both DNA and storage), and BIO0103 (with DNA, no storage). In addition, collagen proteins, major components of *C. elegans* cuticle, were not included in BIO0102. Thus, BIO0102 was designed to represent the biomass assembly in germline to make embryos, whilst BIO0100 and BIO0101 represented body mass with and without storage, and BIO0103 represented progeny assembly inside the eggs. The metabolite coefficients in these reactions as well as other assembly reactions are a function of the composition and formulation of different components of the *C. elegans* biomass. The fraction of macromolecules (proteins, DNA, RNA, TAG, etc.) was first determined for the complete biomass (with both DNA and storage), and then, these fractions were recalculated by making one or both of these two components zero and increasing the rest proportionally.

Since the biomass composition of *C. elegans* has not been studied systematically, we collected information on different biomass components from various studies and developed an approximate composition. This constant composition was used in all analyses as a first approximation, although many components of biomass may be varied in different stages of life. Overall fraction of total lipids was based on (Hutzell and Krusberg, 1982), whilst the ratio of phospholipids to TAG was approximated as 1 based on (Brock et al., 2007; Brooks et al., 2009). Glycogen content was obtained from (Cooper and Vangundy, 1970). Trehalose fraction was approximated as 1% based on (Miersch and Doring, 2012). Glycans of *C. elegans* are represented with N-linked glycans and chitin in the model. While no quantitative information was found for these components, O-linked glycans are reported to make approximately 1% of biomass in (Hanover et al., 2005). We assumed a fraction of 2% for total glycans, equally divided between the representative forms chitin and N-linked glycans. For other variables that were not available in the literature, we used the biomass composition of yeast based on (Forster et al., 2003) as a first approximation. These variables included the amino acid composition of proteins, the overall fractions of DNA, RNA, and ash (*i.e.*, the proportion that was not represented by any metabolite in the biomass reaction), and the relative ratio of the four bases in RNA. The proportions of the four bases in DNA were determined based on the GC% of *C. elegans* genome, approximated as 35%. The remaining portion of biomass after all of the above estimations was assumed to be made of proteins. Protein mass was

divided into mitochondrial, cytosolic and collagen components which were assumed to make 20%, 70%, and 10% of total protein, respectively. Inclusion of the mitochondrial component was necessary to link the separate mitochondrial protein biosynthesis pathway to the biomass assembly. The collagen component was included since collagens form a significant proportion of the cuticle and have a specific, predictable amino acid composition, which was based on 21 major collagens according to (Page and Johnstone, 2007).

The lipid composition of *C. elegans* biomass was further detailed using relatively precise reports from the literature. The macro composition of phospholipids (phosphatidylcholine, sphingomyelin, ether-lipids etc.) was based on (Satouchi et al., 1993). Fatty acid compositions in phospholipids and TAG were based on (Brock et al., 2007) with two exceptions. First fatty acids with chain length greater than 20 carbons, which were rarely detectable in analytical studies (Reis et al., 2011), were represented in the model by a 24-carbon chain molecule assumed to make only 1% of total fatty acids. Second, the mass ratio of cyclic fatty acid cis-11,12-methyleneoctadecanoic acid in TAG was reduced from 0.17 to a symbolic 0.0001, as the only source for cyclic fatty acids is the bacterial diet and the original ratio made this compound limiting for growth based on stored lipids. This limitation was considered as non-realistic since animals can adjust the composition of TAG as evident from the variation of composition in different studies (Brock et al., 2007; Perez and Van Gilst, 2008).

The energetic cost of polymerization reactions that form proteins, DNA, and RNA was determined according to (Neidhardt et al., 1990) and included in the coefficients of ATP or GTP consumed in these reactions.

Transport

Since the identity of metabolite transporters is generally not known in *C. elegans*, we derived most (80%) of the transport reactions from yeast (Forster et al., 2003) and human (Duarte et al., 2007) metabolic models in BiGG (Schellenberger et al., 2010). First a collection of all transport reactions in these two models was formed. Then compounds in the *C elegans* model were cross-referenced with those in BiGG. This process was straightforward for most compounds as we used the BiGG nomenclature in the naming of our compounds. Other compounds in *C. elegans* were matched with their counterparts in BiGG if available (*e.g.*, dedolp [dehydrodolichol diphosphate] in the *C. elegans* model matches dedolp_L and dedolp_U in the human model, which are the liver and uterine homologs of this metabolite, respectively). Using the transport collection and compound matches, the corresponding transport reactions were determined for every compound in the *C. elegans* model. All organelles in BiGG transport reactions, except for mitochondria, were converted to cytosol, since organelle compartmentalization is not made in iCEL1273 except for mitochondria. The simplest form of available transport was incorporated for each compound (*e.g.*, reversible ammonium transport between cells and extracellular space is coupled with sodium, calcium, chloride, or proton transport in the human model, but these reactions were rejected and a simpler reaction that reversibly transports just ammonium was incorporated from the yeast model). Importantly, protons involved in all incorporated reactions were eliminated, as the inclusion of protons in mitochondrial transport reactions resulted in an artificially large ATP synthesis ability.

This was caused by thermodynamically infeasible loops that involved the transport of interconvertible metabolites and provided a net flux of protons out of mitochondria. The transport of protons to and from mitochondria is limited in iCEL1273 to the electron transport chain and ATP synthase to allow stoichiometric calculations of ATP generation. Potential contributions from other transport reactions cannot be described accurately and this uncertainty is currently considered as part of maintenance costs (see below in section 6). All BiGG-related transport reactions are indicated in reaction comments.

For a subset of metabolites, 99 transport reactions were added but not automatically incorporated from BiGG. These included known transporters (*e.g.*, HGR-1 for heme transport), unknown ones that carry out transport reactions predicted to be present with high confidence (*e.g.*, N-acetylglucosamine uptake is inserted as a transport reaction since this compound is part of the axenic medium for *C. elegans* (Lu and Goetsch, 1993)), and gap fillers.

All compounds that are localized to extracellular space (*i.e.*, involved in at least one transport reaction between cytosol and extracellular space compartments) are drained or imported by exchange reactions, to allow mass balance during FBA. Exchange reactions are used for controlling the input and output of the model by flux constraints to define the conditions tested (see below). These reactions are indicated as exchange-type with EX header.

Demand/sink reactions

Endpoint metabolites that are biologically functional without further conversion by metabolic reactions are drained by demand reactions to allow mass balance during their production. These metabolites include signaling molecules (*e.g.*, phosphoinositols), vitamins (*e.g.*, cobalamin [vitamin B12]), cofactors (*e.g.*, coenzyme A), modified proteins (*e.g.*, methylated histones), and others (*e.g.*, glutaurine). Reactions that drain certain endpoint metabolites are made reversible since these metabolites can also be degraded when available. Reversible reactions that both provide and consume endpoint metabolites are called sink reactions (Thiele and Palsson, 2010). Examples include sink reactions for storage compounds (*e.g.*, trehalose) and other metabolites that may be degraded and used in different forms if available (*e.g.*, methylated histones can be demethylated). The difference between demand/sink reactions and exchange reactions is that the endpoint compounds do not need to be transported, as they are used, stored or consumed where they are made available. As with exchange reactions, demand and sink reactions are used to control the input and output of the model for specific tests (see below). Demand and sink reactions are indicated as Demand-type and Sink-type with headers DMN and SNK.

Reaction reversibility and stoichiometry

To decide whether a reaction is reversible or irreversible, we used the information about the direction of the reaction in BiGG, MetaCyc (Caspi et al., 2014), SEED (Aziz et al., 2008; Henry et al., 2010), and Brenda (Schomburg et al., 2004). Three cases were possible regarding reaction directionality: reversible, irreversible in the assumed forward direction, irreversible in the reverse direction to what is assumed. Since databases did not always agree on reaction directionality, we calculated a cumulative score for each case of directionality for a reaction by adding individual scores from the different resources. The

individual scores were 1 or 0 for reports in SEED and MetaCyc since for a given reaction there was at most one matching reaction in each of these databases. For BiGG and Brenda, the directionality scores were defined as the proportion of reports supporting each case, since there were typically multiple matches. In addition to direct matches in Brenda, which was not frequently available, overall reversibility score for the enzyme associated with the reaction was also considered as another Brenda score. These individual scores were summed for each case of directionality. If the score of the best case was higher than the next by >80%, that case was selected. If not, or if the highest score was <0.5 for any case, the reaction was made reversible (*i.e.*, a low overall score meant lack of sufficient data for a decision, which lead to an assumption of a reversible reaction). Exceptions were made in the decision process in multiple cases such as when one database gave more convincing evidence than others (*e.g.*, when multiple experimental reports are available in Brenda for the direction of a reaction), when the information regarding reversibility was found in literature, or when reversibility could be based on similar reactions in the absence of data for the specific reaction in question. All reversibility exceptions are indicated in reaction comments.

Stoichiometry of a reaction was determined according to the following data in a priority order (*i.e.*, the first method that provided an answer determined the stoichimetry): (1) stoichiometry of matching reactions in BiGG, (2) stoichiometry of the matching reaction in MetaCyc, (3) stoichiometry reported in literature. If none of these sources had the information sought, we determined stoichiometry based on mass and charge balance. To determine molar weight for mass balance, compound formulas were obtained from KEGG, MetaCyc, or BiGG. For charge balance, compound charges were obtained from BiGG if available, or were based on other methods as indicated in metabolite comments. Exceptional cases in stoichiometric decisions were rare and are also indicated in comments.

3. PRIME Model: Systematic Localization of *C. elegans* Metabolic Reactions

Reactions were divided into three compartments: mitochondria, cytosol, and extracellular space. The localization of biomass, demand, transport and exchange reactions was straightforward based on their definition (*e.g.*, a demand reaction is localized to the compartment where the drained compound is present). The locations of the other reactions, which are the core set of reactions in the model and are designated as “regular” category (reactions with header R), were systematically determined based on seven resources and FBA (**Figure 3A**).We first used our procedure to decide whether each reaction should be localized to mitochondria or not. Non-mitochondrial reactions were then further localized to extracellular space or cytosol manually. Since only three non-mitochondrial reactions were localized to extracellular space, the main task of this procedure was to decide between mitochondrial and cytosolic localization for every regular reaction.

The resources used in systematic localization provided evidence at different levels (**Figure 3A**). Four of the localization resources predicted the targeting of proteins encoded by the genes in reaction GPR to mitochondria, cytosol, or other organelle. Localization to other organelles was equivalent to localization to cytosol in the model.

Brenda was used as a resource to collect non-specific information regarding the localization of the general enzyme associated with the reaction (*e.g.*, EC 2.4.2.30). BiGG models and FBA provided evidence for the localization of the reaction itself. Each resource was used to obtain a cytosolic and a mitochondrial score from 0 to 1. These scores were then multiplied by weights (depending on the resource, **Figure 3A**) and summed to get a final score on each compartment. The cumulative scores for each compartment were used to decide on reaction localization (see below). Data was derived from these resources as follows:

Mitoprot: This tool was used to calculate the probability (P_m) that a protein is targeted to mitochondria based on the N-terminal sequence (Claros and Vincens, 1996). While the P_m value defined the mitochondrial score, the corresponding cytosolic score was 1-P_m. Protein sequences were obtained from WormBase. When multiple isoforms were available for the product of the same gene, scores were calculated for each isoform, and the maximum scores were used in each compartment. Since Mitoprot provided a direct prediction based on specific protein sequence, we valued this resource with a relatively high weight of 2 for scores <0.95, and an even larger weight of 4 for scores ≥ 0.95 (indicative of 95% confidence).

Mitominer: This database provides experimental and theoretical evidence for mitochondrial localization of genes in twelve eukaryotic species including five metazoans. Since *C. elegans* is not part of this database, we scored genes in our reconstruction based on their potential orthologs in Mitominer. An ortholog was defined as a reciprocal best hit in KEGG SW score tables (see above, section 1). The orthologs were cross-referenced with gene names in a Mitominer reference table that lists proteins with evidence for mitochondrial localization, mostly based on fluorescence assays and proteomics analyses. The Mitominer score for cytosol (S_{cyt}) was then based on the ratio of orthologs (in the twelve Mitominer organisms) that had no hits in the evidence table. Mitochondrial score (S_{mit}) was calculated as a function of two variables: (1) the ratio of hits in the Mitominer database (1- S_{cyt}) and (2) the evidence available for the ortholog with the strongest evidence of mitochondrial targeting. The equation for this score is $S_{mit} = 0.5E + 0.5(1-S_{cyt})$, where E is the highest evidence score in all orthologs. The evidence score was calculated as $E = 0.8exp + 0.2thr$, where *exp* stands for the strength of the experimental and *thr* for that of the theoretical evidence provided. To define the strength of the evidence score, we differentially weighed fluorescence-based and mass-spec-based (proteomics) reports from tests in the organism carrying the orthologous protein. If there were more than 1 fluorescence-based reports, or more than 7 mass-spec reports, *exp* was given a value of 1. If only one of these two types of evidence was available with less than or equal to these thresholds (1 and 7, respectively), then *exp* = 0.5. If both types of evidence was available in any number of reports, *exp* was given a value of 1. The strength of the theoretical score (*thr*) was defined as the ratio of theoretical predictors that predicted mitochondrial targeting of the orthologous protein sequence. The total number of predictors was 5. The overall Mitominer score was given a relative weight of 1.5 in the total localization score (**Figure 3A**) as it was not directly based on *C. elegans* genes, but it integrated experimental information about homologous genes from multiple other eukaryotes.

UniProt and Organelle Database: Available information on the subcellular localization *C. elegans* proteins was downloaded from UniProt (Bateman et al., 2015) and Organelle Database (Wiwatwattana and Kumar, 2005). Mitochondrial and non-mitochondrial scores were defined as 0 or 1 depending on the absence or presence of each compartment in the reported information (all non-mitochondrial localizations were considered as the cytosolic compartment). These scores were given a low weight (**Figure 3A**) since there was no information for vast majority of proteins in both databases, and since the existing information was mainly based on theoretical predictions (not related to Mitoprot).

Brenda: Protein localization information was collected from Brenda for all enzymes in the model (only eukaryotic reports were evaluated). For each enzyme, the proportion of the number of reports that indicate enzyme localization to mitochondria determined the mitochondrial score and the proportion of the rest of the localization reports determined the cytosolic score. However, if one of the reports was directly based on *C. elegans* proteins, the score was made 1.0 for the corresponding location. The weight of Brenda score was set at 1 (**Figure 3A**) as this analysis was based on indirect associations based on the generic enzyme, without assessment of homology.

BiGG: This database includes reactions from the metabolic network models of two eukaryotes, human and yeast, for which subcellular localizations in the corresponding model are indicated. Each reaction in the *C. elegans* model was first searched in these models. If no matches were found, both compartments (mitochondrial or non-mitochondrial) were given 0 score. If matches were found, the score of a compartment was increased by 0.5 for the occurrence of the reaction in that compartment in each organism. For example, if the mitochondrial version of a reaction was found in the yeast network but not in the human network, the mitochondrial score would be 0.5. If the reaction was found in the cytosol of the yeast network and the peroxisome of the human network, the non-mitochondrial score would be 1.0. BiGG scores were given a medium weight (**Figure 3A**) since these eukaryotic models reflect systematic reconstructions in two well studied eukaryotes, although this information is also not direct.

FBA: The localization of a reaction to mitochondria or cytosol was also scored based on the capacity of the reaction to carry flux in either compartment. Three tests were performed for each reaction in the model, by localizing the reaction to mitochondria, cytosol, and both compartments. In each test, maximum flux that the reaction could take was calculated as described above (see section 2). If this flux was not zero in a compartment in any one of these tests, that compartment was scored 1. The weight of this score was 2 (**Figure 3A**), reflecting the fact that flux carrying capacity provides a direct prediction for the correct localization in modeling terms. In addition, for each of the three tests above, maximum biomass production and maximum energy generation were calculated, by using the biomass drain (BIO0010) and ATP-maintenance (RCC0005) reactions as the maximized objective, respectively. If the localization of the reaction to a particular compartment increased one or both of these values compared to otherwise, then the score of that compartment was changed to 4 as a bonus (**Figure 3A**). If localization to both compartments was necessary for the increase in biomass or energy production, then both compartments received this bonus score.

Reaction localization was based on cumulative evidence from the resources defined above. An overall score was calculated for mitochondrial and non-mitochondrial compartmentalization of each reaction by summing the scores multiplied by the corresponding weights (**Figure 3A**). For reactions that were associated with multiple genes or enzymes, the maximum gene- and enzyme-level scores were used for each compartment. The range of the overall score was from 0 (no evidence for the compartment scored or no data) to 14 (consistently perfect scores for the compartment). To algorithmically decide the location of reactions from overall scores, two thresholds were determined, which we designate as τ_1 and τ_2. A reaction was localized to a compartment either if the cumulative score passed τ_1 for that compartment or if the score of that compartment was above the score of the other compartment by more than τ_2. If the two compartment scores were within τ_2 of each other, the reaction was localized to both. These thresholds were set at optimal values of $\tau_1 = 6.2$ and $\tau_2 = 1.2$, which maximized the agreement between the localizations in the template model and algorithmic decisions. Since the template model was manually reconstructed, reaction localization was based mainly on pathways, gap-filling criteria, and a manual evaluation of evidence in the above defined resources. The disagreements between the computational decisions and manual localizations were then resolved by either re-localizing reactions or setting exceptions that overruled these scores. This procedure was carried out iteratively, since FBA-based scores changed when reaction localizations were changed. When no more changes were observed in computational decisions, all reactions were localized to mitochondrial and non-mitochondrial compartments on a rational basis, either as algorithmically explained by the cumulative scores or as decided by an exception rule. All exceptions for protein localization are explained in reaction comments.

Finally, reactions that were associated with multiple genes and localized to both cytosol and mitochondria were further examined to divide the GPR into the two compartments. The genes (proteins) associated with such reactions were localized based on overall scores from the four resources yielding evidence at the gene level (**Figure 3A**). Scores were manually evaluated, and for each gene, the compartment that was clearly ahead in cumulative score was selected. If scores were close or if both were low, the gene (protein) was localized to both compartments. Exceptional cases are indicated in reaction comments. With the reaction and protein re-localizations, the reconstruction of the prime model was completed.

Validation of subcellular localization in the Prime model

To validate reaction and protein localization in the prime model, experimental protein localization data was downloaded from WormBase. Specifically, IDA (inferred from direct assay) reports for cellular component in the gene ontology section were used. IDA protein locations were available for proteins encoded by 132 genes in the prime model. Locations of these genes in the prime model were determined based on the locations of the reactions they are associated with.

We first checked whether the experimental information was a part of the decision-making in some of these genes, mainly since UniProt, Organelle Database, and Brenda reports may cover available experimental data. For only one gene (*aco-1*) did this information affect both the score from either of these resources and the algorithmic conclusion based on total score. Therefore, this gene was excluded from the validation

analysis. In addition, the predictions for the location of two of the remaining 131 genes were correct, but not used in the model due to technical restrictions in the model design. One of these genes is *vha-8*, which encodes a vacuolar ATPase, but is localized to mitochondria as there is only one ATPase in the model. The other one is *acs-2*, which encodes an acyl coA synthetase, an important component of phospholipids biosynthesis. However, we avoided the inclusion of separate, mitochondrial pathways for the biosynthesis of mitochondrial phospholipids, and lumped all related genes in cytosolic pathways to make a cytosolic phospholipid that represented all phospholipids in the biomass. Both *vha-8* and *acs-2* were excluded from validation analysis.

The results of validation with the remaining 129 genes are shown in **Figure 3B** in comparison with the performance of our gene-based predictors. While Mitoprot showed an excellent performance by itself, both error rate and nonspecific matching were tripled with this tool compared to the metabolic model. The Mitominer-based predictor developed in this study was the next best and had a reasonable error rate of about 15% despite its indirect capture of evidence based on gene orthology. UniProt and Organelle Database clearly had poor coverage compared to other tools, although error rates were low or moderate.

4. Completion of Reconstruction by Semi-Automated Expansion of the Prime Model

To explore the possibility of connecting the rest of the SACURE-annotated reactions (704 reactions that were not incorporated during pathway-by-pathway manual reconstruction; hereafter referred to as the query set) to the prime model, we used a semi-automated reconstruction pipeline. This procedure had the following steps:

1) Reversibility and localization of the reactions in the query set were determined based on multiple resources as explained above (sections 3 and 4, respectively). As an exception, experimental data in WormBase (section 4) was directly incorporated for these reactions when available, overruling other evidence.

2) Prime model reactions, query set, reactions of uncharacterized enzymes in KEGG, spontaneous reactions in KEGG, and BiGG transport reactions (human and yeast models; see section 3) were merged to form a unified reaction network.

3) Reactions that were disconnected in the unified network at both ends were eliminated right away, as these reactions would never be useful in our connectivity criteria (see below). Then, additional transport reactions were incorporated for every compound in the query set that was not transportable by BiGG transport reactions. The final network had a total of 8,679 reactions and was converted to a mathematical model for FBA.

4) FBA was combined with mixed integer linear programming (MILP, see section 7 below) (Shlomi et al., 2008) to maximize the number of query set reactions that carried flux while minimizing the number of additional (not BiGG-based) transport reactions that carried flux. Reactions that could not carry flux in this step were eliminated as they needed more than one transport reaction to be connected to the network. It is important to note that the optimization technique used in this step mathematically captures all reactions that are connected to the network (*i.e.*,

that can carry flux) based on our criteria (*i.e.*, not dependent on a specific transport reaction with no other use). This property was verified by test cases.

5) Further FBA analyses were carried out to determine the dependence of the remaining query set reactions (that could carry flux) on reactions other than those in the prime model (*i.e.*, a reaction is dependent on another if it cannot carry flux when the other reaction is constrained to zero flux). Query set reactions that were dependent on additional transport reactions which had no other function (*i.e.*, no other query reaction depended on them) were eliminated. Auxiliary reactions (reactions of uncharacterized enzymes, spontaneous reactions, all transport reactions) that did not have any function (*i.e.*, no query set reactions were dependent on them) were also eliminated.

6) The remaining reactions in the query set (N=233) are connected to the network. As a final step, these reactions were manually examined to decide which ones are to be incorporated into the model.

Most of the reactions from step 5 (77%) were rejected during manual curation, since they did not add any new function to the model. For instance, R00572 is a KEGG reaction for pyruvate kinase (associated with *pyk-1* and *pyk-2* in SACURE) that uses CTP in the conversion of phosphoenolpyruvate to pyruvate. This conversion is represented in the model with an ATP-based reaction (RC00200). Since ATP and CTP are interconvertible (RC00570), the addition of R00572 does not add any function to the model except for artificially increasing the number of reactions. Therefore this reaction, as well as three other forms of the same conversion using other nucleoside triphosphates (GTP, UTP, ITP), were not incorporated into the model.

REFERENCES

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. Bmc Genomics *9*, 75.

Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Ar-Ganiska, J., Bely, B., Bingley, M., *et al.* (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-D212.

Brock, T.J., Browse, J., and Watts, J.L. (2007). Fatty acid desaturation and the regulation of adiposity in Caenorhabditis elegans. Genetics *176*, 865-875.

Brooks, K.K., Liang, B., and Watts, J.L. (2009). The Influence of Bacterial Diet on Fat Storage in *C. elegans*. Plos One *4*.

Byerly, L., Cassada, R.C., and Russell, R.L. (1976). The life cycle of the nematode *Caenorhabditis elegans*. I. Wild-type growth and reproduction. Dev Biol *51*, 23-33.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., *et al.* (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res *42*, D459-D471.

Claros, M.G., and Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem *241*, 779-786.

Cooper, A.F., and Vangundy, S.D. (1970). Metabolism of Glycogen and Neutral Lipids by Aphelenchus-Avenae and Caenorhabditis-Sp in Aerobic, Microaerobic, and Anaerobic Environments. J Nematol *2*, 305-&.

Dempsey, W.B. (1971). Role of Vitamin-B6 Biosynthetic Rate in Study of Vitamin-B6 Synthesis in Escherichia-Coli. J Bacteriol *108*, 1001-&.

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B.O. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. P Natl Acad Sci USA *104*, 1777-1782.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Ferris, H., Lau, S., and Venette, R. (1995). Population Energetics of Bacterial-Feeding Nematodes - Respiration and Metabolic Rates Based on Co2 Production. Soil Biol Biochem *27*, 319-330.

Ferris, H., Venette, R.C., and Lau, S.S. (1997). Population energetics of bacterial-feeding nematodes: Carbon and nitrogen budgets. Soil Biol Biochem *29*, 1183-1194.

Forster, J., Famili, I., Fu, P., Palsson, B.O., and Nielsen, J. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res *13*, 244-253.

Guo, A.C., Jewison, T., Wilson, M., Liu, Y.F., Knox, C., Djoumbou, Y., Lo, P., Mandal, R., Krishnamurthy, R., and Wishart, D.S. (2013). ECMDB: The E-coli Metabolome Database. Nucleic Acids Res *41*, D625-D630.

Hanover, J.A., Forsythe, M.E., Hennessey, P.T., Brodigan, T.M., Love, D.C., Ashwell, G., and Krause, M. (2005). A Caenorhabditis elegans model of insulin resistance: Altered macronutrient storage and dauer formation in an OGT-1 knockout. P Natl Acad Sci USA *102*, 11266-11271.

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 *28*, 977-982.

Hirsh, D., Oppenheim, D., and Klass, M. (1976). Development of Reproductive-System of *Caenorhabditis-elegans*. Dev Biol *49*, 200-219.

Hutzell, P.A., and Krusberg, L.R. (1982). Fatty-Acid Compositions of *Caenorhabditis-elegans* and *Caenorhabditis-briggsae*. Comp Biochem Phys B *73*, 517-520.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., *et al.* (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome biology *5*, R7.

Lu, N.C., and Goetsch, K.M. (1993). Carbohydrate Requirement of *Caenorhabditis-Elegans* and the Final Development of a Chemically-Defined Medium. Nematologica *39*, 303-311.

Matzanke, B.F., Muller, G.I., Bill, E., and Trautwein, A.X. (1989). Iron-Metabolism of Escherichia-Coli Studied by Mossbauer-Spectroscopy and Biochemical Methods. Eur J Biochem *183*, 371-379.

McGhee, J.D. (2007). The C. elegans intestine. WormBook : the online review of *C elegans* biology, 1-36.

Miersch, C., and Doring, F. (2012). Sex Differences in Carbohydrate Metabolism Are Linked to Gene Expression in Caenorhabditis elegans. Plos One *7*.

Neidhardt, F.C., Ingraham, J.L., and Schaechter, M. (1990). Physiology of the bacterial cell : a molecular approach (Sunderland, Mass.: Sinauer Associates).

Page, A.P., and Johnstone, I.L. (2007). The cuticle. WormBook : the online review of *C elegans* biology, 1-15.

Perez, C.L., and Van Gilst, M.R. (2008). A C-13 isotope labeling strategy reveals the influence of insulin signaling on lipogenesis in *C-elegans*. Cell Metab *8*, 266-274.

Reis, R.J.S., Xu, L.L., Lee, H., Chae, M., Thaden, J.J., Bharill, P., Tazearslan, C., Siegel, E., Alla, R., Zimniak, P., *et al.* (2011). Modulation of lipid biosynthesis contributes to stress resistance and longevity of *C. elegans* mutants. Aging-Us *3*, 125-147.

Reznik, E., Mehta, P., and Segre, D. (2013). Flux imbalance analysis and the sensitivity of cellular growth to changes in metabolite pools. PLoS computational biology *9*, e1003195.

Satouchi, K., Hirano, K., Sakaguchi, M., Takehara, H., and Matsuura, F. (1993). Phospholipids from the Free-Living Nematode *Caenorhabditis-Elegans*. Lipids *28*, 837-840.

Schellenberger, J., Park, J.O., Conrad, T.M., and Palsson, B.O. (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. Bmc Bioinformatics *11*.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004). BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res *32*, D431-D433.

Shlomi, T., Cabili, M.N., Herrgard, M.J., Palsson, B.O., and Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. Nat Biotechnol *26*, 1003-1010.

Thiele, I., and Palsson, B.O. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc *5*, 93-121.

Van Voorhies, W.A. (2002). The influence of metabolic rate on longevity in the nematode *Caenorhabditis elegans*. Aging Cell *1*, 91-101.

Van Voorhies, W.A., and Ward, S. (1999). Genetic and environmental conditions that increase longevity in *Caenorhabditis elegans* decrease metabolic rate. P Natl Acad Sci USA *96*, 11399-11403.

Vanfleteren, J.R., and DeVreese, A. (1996). Rate of aerobic metabolism and superoxide production rate potential in the nematode *Caenorhabditis elegans*. J Exp Zool *274*, 93-100.

Wang, J., and Kim, S.K. (2003). Global analysis of dauer gene expression in *Caenorhabditis elegans*. Development *130*, 1621-1634.

Wiwatwattana, N., and Kumar, A. (2005). Organelle DB: a cross-species database of protein localization and function. Nucleic Acids Res *33*, D598-D604.

FIGURES

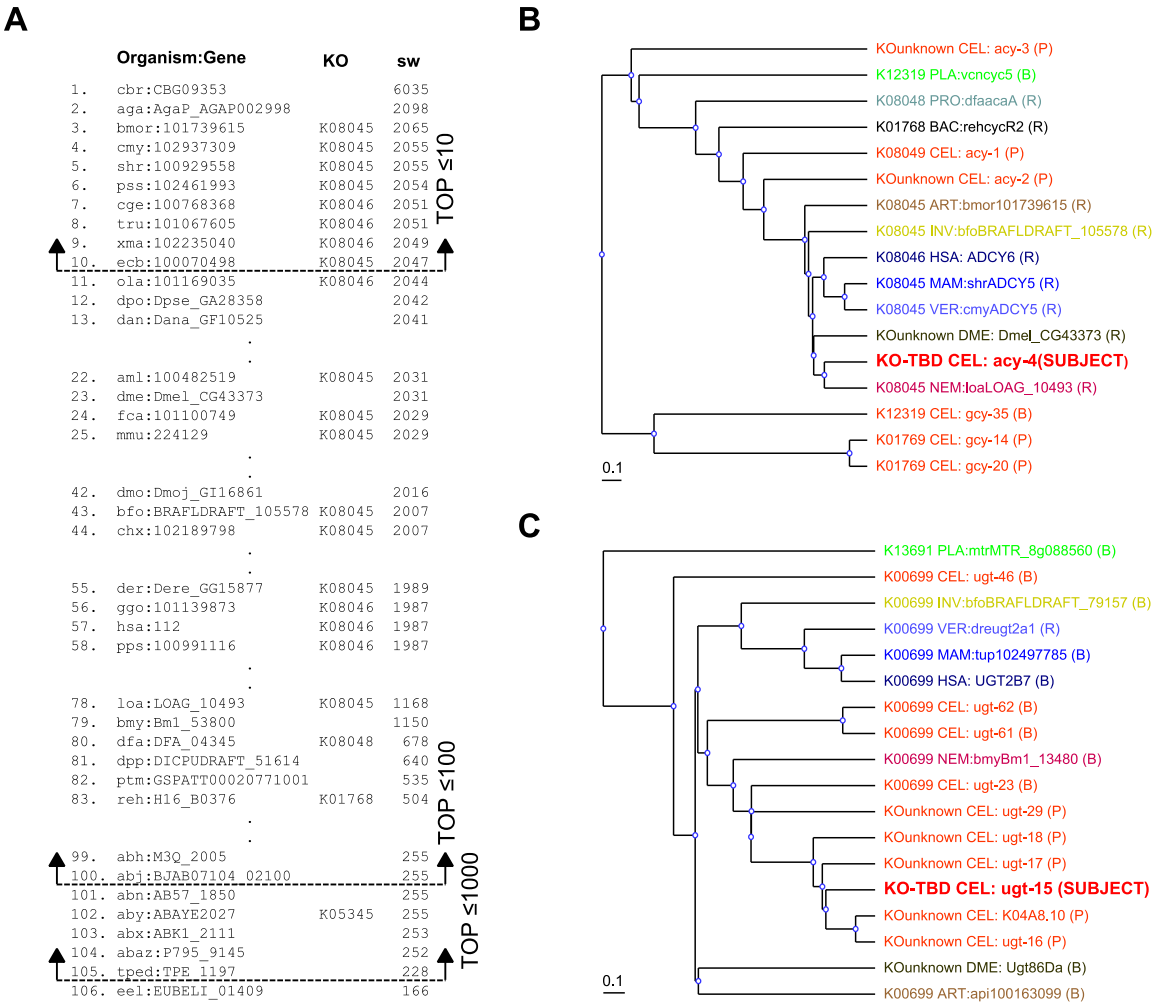

Figure 1. **(A-B)** Evaluation of *C. elegans* gene *acy-4* with myKEGG **(A)** and myTree **(B)**. The final decision for this gene is an association with K08045 (adenylate cyclase) at low confidence. **(A)** Reciprocal best hits (RBH) table showing related genes in other organisms in descending order of Smith-Waterman alignment scores in KEGG (*s*). Organism and gene pairs are designated according to KEGG nomenclature. Top 1000 rows are truncated at 105 genes due to the minimum threshold of *s*=200 for a significant match. Accordingly, the correction factor ((105-100)/100 = 0.05) makes the contribution of the top 1000 group to overall score insignificant. The most likely KO associations are obtained for K08045 and K08046 with myKEGG scores of 0.59 and 0.28 in the RBH table (**Equation 1**) and 0.59 and 0.29 overall (**Equation 2**), respectively. **(B)** Phylogenetic tree shows clustering around K08045, with a cluster score of 0.33 for this KO. Tree scores are 1.53 and 0.23 for K08045 and K08046, respectively. **(C)** A myTree example that needed manual curation. *C. elegans* gene *ugt-15* was manually associated with K00699 (glucuronosyltransferase). Due to the domination of the tree by *C. elegans* paralogs, myTree scores underestimated the strength of the visible clustering pattern (clustering score is 0 and tree score is 0.35 for association with K00699). Tree abbreviations: CEL, *C. elegans*; HAS, *H. sapiens*; DME, *D. melanogaster*; ATH, *A. thaliana*; SCE, *S. cerevisiae*; BAC, bacteria; ARC, archaea; PRO, protists; FUN, fungi; PLA, plants; INV, invertebrates; NEM, nematodes; ART, arthropods; VER, vertebrates; MAM, mammals. Parenthetical information indicates genes introduced as best matches (B) or reciprocal best hits (R) in other organisms, and introduced as paralogues (P) or reciprocal best match to one of the organisms in the tree (B) in *C. elegans*. Organism abbreviations are from KEGG.

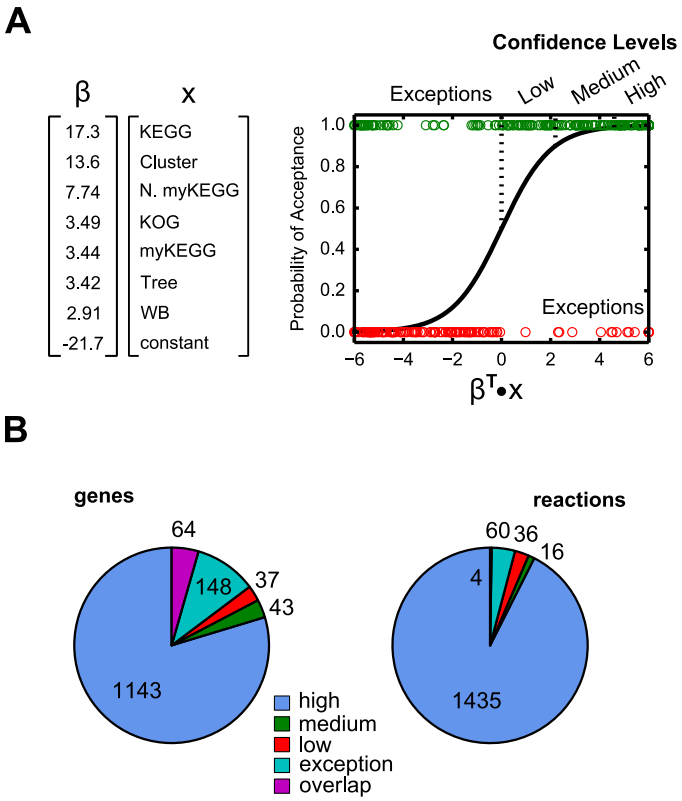

Figure 2. **(A)** The weights of contributing variables in the trained logistic function (on the left, see **Table 1** for the variables; the constant of the logistic function is also shown), and the agreement between this function and all SACURE annotations (on the right). Green and red circles indicate accepted and rejected gene-KO associations, respectively. Confidence intervals are defined as stated in Supplemental Experimental Procedures. Acceptances below a probability of 0.5 and rejections above this value show cases where manual decisions overruled the predictions of the logistic function. **(B)** SACURE-annotated genes and reactions according to confidence levels and exception rules including the derivation of gene-reaction relationships based on overlapping reactions in candidate KOs or enzymes (see Supplemental Experimental Procedures). Most genes and reactions were annotated as consistent with the logistic regression function (high, medium, and low confidence). For reactions associated with multiple genes, highest confidence was used.

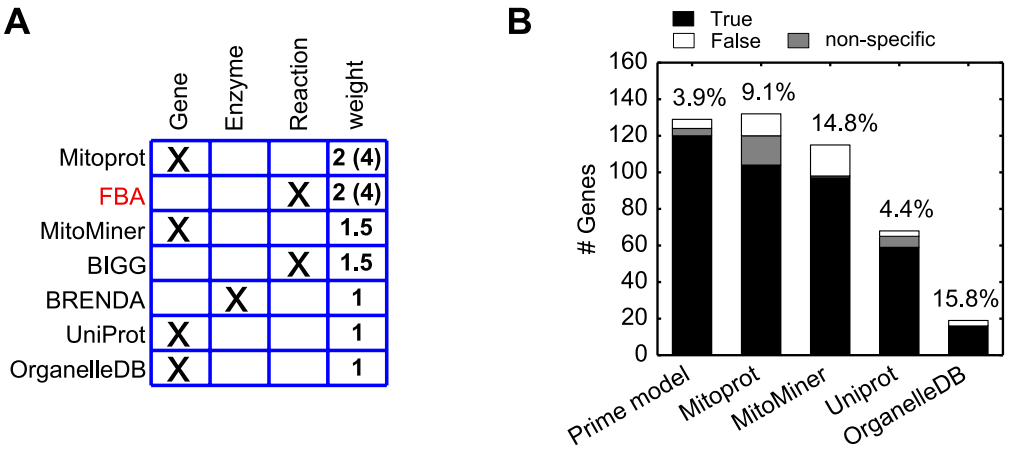

Figure 3. **(A)** Resources used for evaluating reaction localization to mitochondria or other compartments. Gene, enzyme, and reaction indicate at which level the predictor works. Gene-level predictions evaluate the targeting of proteins encoded by the genes in reaction GPR to mitochondria or other compartments. The enzyme level predictor evaluates the localization of the general enzyme in GPR in the Brenda database. Reaction level predictors localize the reaction. Each predictor gives a score from 0 to 1 for each compartment (mitochondrial and non-mitochondrial). These scores are multiplied with the indicated weights and summed to obtain a cumulative evidence score, which is then used for decision-making. Weights in parentheses indicate a bonus awarded when an exceptional score is achieved (see Supplemental Experimental Procedures). **(B)** Comparison of the accuracy of reaction localization by the prime model (*i.e.*, based on the pipeline indicated in section 4) and by individual gene-level predictors. Predictions are tested against the experimental validation set. Predictions by gene-level predictors were based on a score threshold of 0.5 (out of 1.0) to assign a protein to a particular compartment (mitochondria or other).

TABLES

Table 1. Predictors used for the annotation of metabolic genes[a].

Predictor	Input	Method	Assignment	Output	Weight
KEGG	KO	Direct	KO	{0,1}	17.3
Cluster Score	Phylogenetic tree	Lineage algorithm	KO	[0,1]	13.6
Normalized myKEGG score	SW tables	Equation S2, normalized	KO	[0,1]	7.74
myKEGG score	SW tables	Equation S2	KO	[0,1]	3.44
KOG	KOG, SW tables	Indirect	EC	{0,1}	3.49
Tree score	Phylogenetic tree	Tree algorithm	KO	[0,2)	3.42
WormBase description	Text, protein domains	Word matching	EC	[0,1]	2.91
UniProt description[b]	Text, protein families	Word matching	EC	[0,1]	0.00
UniProt EC[b]	EC	Direct	EC	{0,1}	0.00

[a]Abbreviations: EC, Enzyme Commission number, KO, KEGG Orthology; KOG, orthology groups based on (Koonin et al., 2004); SW, Smith-Waterman alignment.
[b]UniProt scores were rejected by the model as they were associated with small weights and zeroing these weights did not change algorithmic decisions.

Table 2. Validation of iCEL1273 with observed consumption/production rates.

	L4 Stage		Adult Stage (3 days)	
Constraint	**Observed range**	**Model range**	**Observed range**	**Model range**[a]
Bacterial uptake (g dW/g dW/h)	0.02-0.2	0.16-*unb*[a]	0.02-0.2	0.09-*unb*[a]
O_2 uptake (mmol/g dW/h)	2.4	1.1-5.3	0.49-0.70	0.10-4.2
CO_2 release (mmol/g dW/h)	1.7-2.4	0.26-2.6	0.49	0.0-1.8
Biomass production (1/h)	0.100	0-0.133	0.065	0-0.144

[a] Unbound since excess bacterial material can be excreted as waste product.