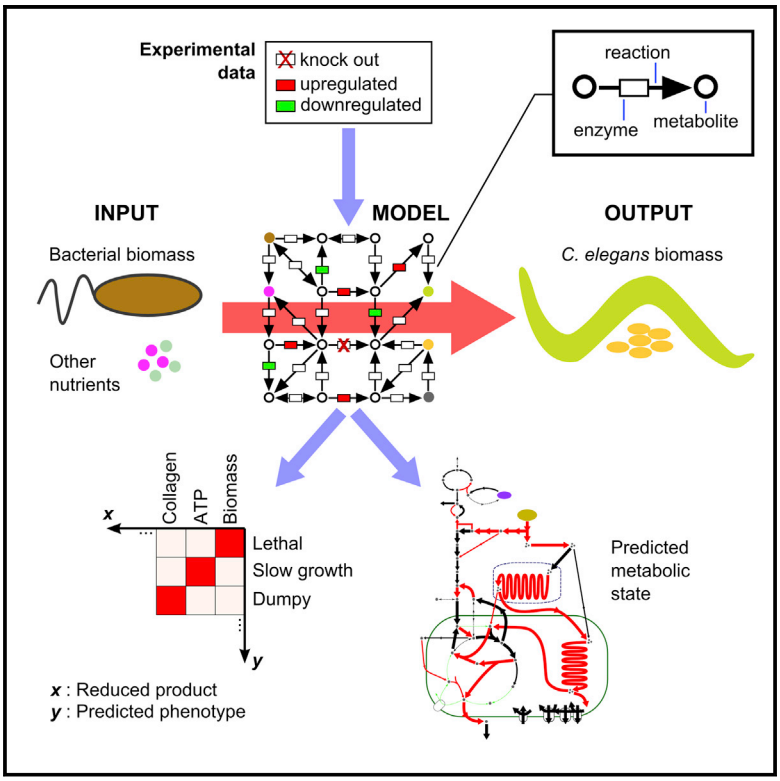
Article

Cell Systems

A *Caenorhabditis elegans* Genome-Scale Metabolic Network Model

Graphical Abstract



Highlights

- The metabolic network of *C. elegans* is reconstructed at the genome scale
- The network can mathematically convert bacterial diet into worm biomass
- *C. elegans* metabolic model is predictive of gene essentiality
- The metabolic network is integrated with the gene expression data of dauer animals

Authors

L. Safak Yilmaz, Albertha J.M. Walhout

Correspondence

lutfu.yilmaz@umassmed.edu (L.S.Y.), marian.walhout@umassmed.edu (A.J.M.W.)

In Brief

Yilmaz and Walhout present the first genome-scale reconstruction of the *C. elegans* metabolic network and show that this network can mathematically convert bacterial biomass into worm biomass and energy. The network can be integrated with gene expression and phenotypic data, demonstrating its predictive power.

Yilmaz & Walhout, 2016, Cell Systems *2*, 297–311
May 25, 2016
http://dx.doi.org/10.1016/j.cels.2016.04.012

Cell Systems
Article

A *Caenorhabditis elegans* Genome-Scale Metabolic Network Model

L. Safak Yilmaz[1,*] and Albertha J.M. Walhout[1,*]
[1]Programs in Systems Biology and Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA
*Correspondence: lutfu.yilmaz@umassmed.edu (L.S.Y.), marian.walhout@umassmed.edu (A.J.M.W.)
http://dx.doi.org/10.1016/j.cels.2016.04.012

SUMMARY

Caenorhabditis elegans is a powerful model to study metabolism and how it relates to nutrition, gene expression, and life history traits. However, while numerous experimental techniques that enable perturbation of its diet and gene function are available, a high-quality metabolic network model has been lacking. Here, we reconstruct an initial version of the *C. elegans* metabolic network. This network model contains 1,273 genes, 623 enzymes, and 1,985 metabolic reactions and is referred to as iCEL1273. Using flux balance analysis, we show that iCEL1273 is capable of representing the conversion of bacterial biomass into *C. elegans* biomass during growth and enables the predictions of gene essentiality and other phenotypes. In addition, we demonstrate that gene expression data can be integrated with the model by comparing metabolic rewiring in dauer animals versus growing larvae. iCEL1273 is available at a dedicated website (wormflux.umassmed.edu) and will enable the unraveling of the mechanisms by which different macro- and micronutrients contribute to the animal's physiology.

INTRODUCTION

The nematode *Caenorhabditis elegans* and its bacterial diet have been used as an interspecies system to gain insights into the connections between nutrients, genotype, and phenotype (Coolon et al., 2009; Gracida and Eckmann, 2013; MacNeil et al., 2013; Pang and Curran, 2014; Soukas et al., 2009; Watson et al., 2013, 2014). Different bacterial species or strains can be fed to the animal, and both *C. elegans* and its diet can be genetically manipulated (reviewed in Watson and Walhout, 2014; Yilmaz and Walhout, 2014). A main challenge now is to understand, at a systems level, how *C. elegans* responds to individual nutrients. Gaining such insights requires a high-quality model of both bacterial and *C. elegans* metabolic networks.

The metabolic network of an organism is the complete set of biochemical reactions in which metabolites are broken down and synthesized. It serves two major purposes: the generation of biomass for growth and reproduction, and the generation of energy to support cellular and organismal processes. Genome-scale metabolic network models have been used together with flux balance analysis (FBA) (O'Brien et al., 2015; Oberhardt et al., 2009) to calculate the steady-state conversion rates of compounds in every reaction of the network (i.e., reaction fluxes). Using a selected objective such as optimal growth or energy production, the calculated flux distribution predicts the metabolic state of the organism, given a set of constraints defined by nutritional or environmental conditions.

While metabolic networks have been reconstructed for a large number of bacteria and a few eukaryotic organisms (reviewed in O'Brien et al., 2015), no metabolic network model is available for *C. elegans*. Metabolic gene annotations are available in databases such as KEGG (Kanehisa et al., 2015) and are useful for pathway visualization. However, these annotations are remarkably incomplete and therefore most pathways are not capable of carrying flux. Thus, current databases do not provide a functional network structure that is suitable for FBA.

Here, we present the global reconstruction of the *C. elegans* metabolic network and its conversion into a mathematical model for use with FBA to generate mechanistic predictions and integrate additional data types (Figure 1A). We demonstrate that this model can simulate the conversion of bacterial diet into *C. elegans* biomass, predict effects of diet or genotypic manipulations on phenotypes, and can be integrated with gene expression data by mathematical modeling.

RESULTS

Overview of Reconstruction

We reconstructed the metabolic network of *C. elegans* using a modular pipeline that integrates multiple sources of information (Figure 1B). First, metabolic genes were annotated to establish gene-protein-reaction (GPR) associations (Thiele and Palsson, 2010), which were then used to manually reconstruct a template network in a pathway-by-pathway manner. Network gaps that prevented reactions from carrying flux were identified and filled. Reactions were localized to cytosol, mitochondria, or extracellular space for proper network compartmentalization. The resulting PRIME model (Figure 1B) was capable of producing *C. elegans* biomass from bacterial diet (Figure 1C). GPRs left out by the manual reconstruction process were exhaustively tested for flux carrying capacity in the PRIME model, and the ones that could add functionality to the network were re-incorporated. The resulting model includes 1,273 genes, 623 enzymes, and 1,985 metabolic reactions and was named iCEL1273. The components of iCEL1273 are presented in Tables S1, S2, S3, S4, and S5 (annotations, biomass compositions, reactions, compounds, and enzymes). The main steps of

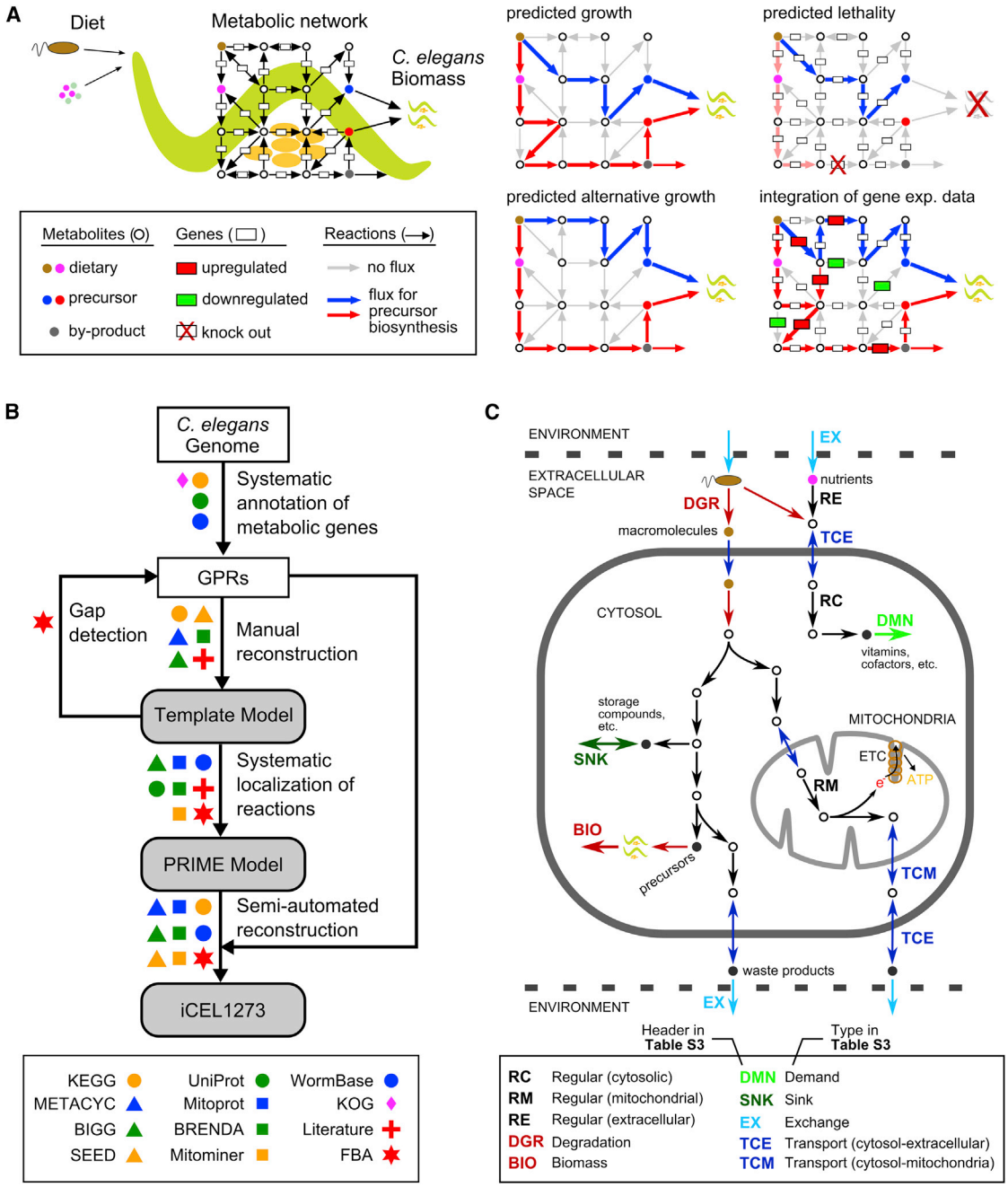


Figure 1. Overview of the *C. elegans* Metabolic Network Model and the Reconstruction Process

(A) Toy network representing the reconstructed *C. elegans* metabolic model. Two nutrients obtained from diet are used to synthesize two *C. elegans* biomass precursors with the excretion of one by-product as waste. The "predicted growth" indicates biomass production that can be achieved via indicated flux through the network (i.e., body growth or offspring). The "predicted alternative growth" depicts how the network can be rewired to use alternate pathways to achieve the same objective, as long as both precursors are successfully synthesized. The "predicted lethality" indicates genetic perturbations (e.g., knockout) that prevent biomass production due to the fatal disruption of flux. The "integration of exp. data" illustrates the incorporation of gene expression data that describe the up- and downregulation of genes encoding metabolic enzymes to deduce flux distribution under regulatory constraints.

(B) Pipeline of the *C. elegans* metabolic network reconstruction process. The top 12 resources are shown where used.

(C) Cartoon of the reconstructed network. The different types of reactions are indicated with their reaction ID headers and types provided in Table S3 (electron transport chain: ETC).

the reconstruction are presented below, followed by model validation. The details of the methods can be found in Supplemental Information.

Identification of *C. elegans* Metabolic Genes

To generate an initial list of *C. elegans* GPRs, we used the orthology system in KEGG (Kanehisa et al., 2015), which connects

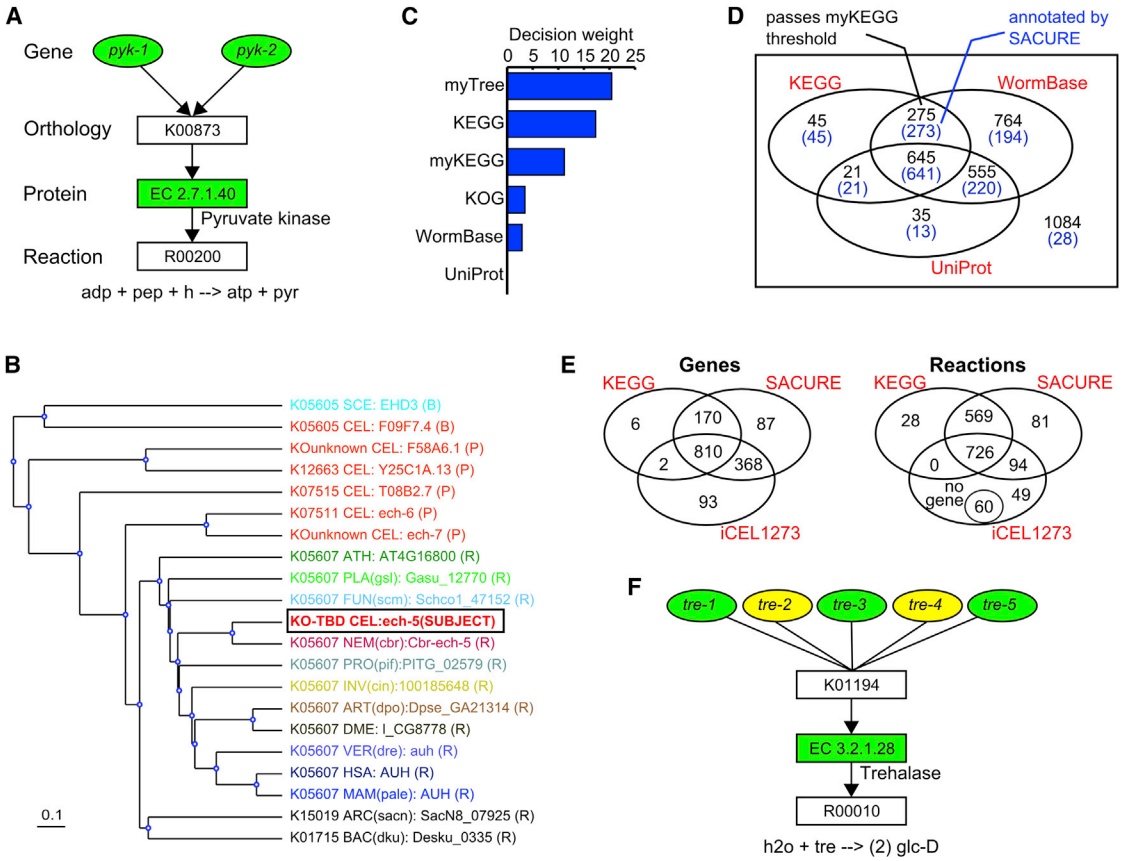


Figure 2. Annotation of *C. elegans* Metabolic Genes

(A) Example GPR association inferred via KO group.

(B) Example phylogenetic tree created by myTree that shows the relatedness of the *ech-5* gene (based on protein sequence) to genes from human, three model organisms, and representative organisms from ten taxonomic groups. The tree labels indicate KO group, taxonomy or model organism, organism name (if taxonomic group), and gene name, respectively. See Figure S1 for abbreviations and details.

(C) Relative contribution of annotation resources to decision making for gene-KO connections in SACURE. The relative contribution was quantified as an overall weight based on logistic regression involving annotation variables and manual decisions. See Supplemental Information and Table S9 for details.

(D) Venn diagram illustrating candidate metabolic genes that pass an arbitrarily low myKEGG score threshold to be matched with a metabolic KO group. Only two of the 988 *C. elegans* metabolic genes previously annotated by KEGG were missed by this thresholding. The genes identified by SACURE are shown in parentheses.

(E) Venn diagrams illustrating final sets of genes and reactions in SACURE and iCEL1273 in relationship to KEGG. Only KEGG-based reactions are shown in the iCEL1273 set (i.e., custom reactions not found in KEGG database are not included) for comparison with the other two sets, which by definition only have KEGG reactions.

(F) Example of a metabolic gene family that was partially included in KEGG, but has been complemented by SACURE (genes annotated by both KEGG and SACURE: green and genes annotated by SACURE only: yellow).

annotated genes to one of ~17,000 KEGG orthology (KOs) groups representing genes with shared function throughout phylogeny. Of these, ~6,000 KO groups are first associated with enzymes designated by an enzyme commission (EC) number and then with metabolic reactions. For instance, *pyk-1* and *pyk-2* are both associated with KO group K00873 and EC 2.7.1.40, or pyruvate kinase, which catalyzes the conversion of phosphoenolpyruvate into pyruvate (Figure 2A). At the time of our analysis, KEGG had identified 988 *C. elegans* genes associated with 1,323 metabolic reactions (excluding signaling-related reactions).

To assess the completeness of KEGG annotations, we cross-referenced all *C. elegans* genes with metabolic enzyme information available in WormBase (Harris et al., 2013) and UniProt (UniProt Consortium, 2015). Specifically, we searched for enzyme names in the gene descriptions in WormBase and protein domain annotations in both WormBase and UniProt. This provided hundreds of additional candidate genes that were not annotated in KEGG, but could potentially be associated with metabolic enzymes and reactions. To determine which of these genes encode metabolic enzymes, we developed two auxiliary annotation data sets. The first, named myKEGG, was built by compiling all best-hit and reciprocal best-hit Smith-Waterman scores (based on protein sequence alignment) between each of the 20,519 *C. elegans* genes and genes from all 3,073 organisms incorporated in KEGG to yield an overall likelihood score for possible gene-KO group associations. The second, myTree, includes a phylogenetic tree for each *C. elegans* gene based on

protein sequence similarity (Figures 2B and S1). We used myTree as a visual aid to observe clustering of the query gene with other genes. In addition, we used an independent database of eukaryotic orthologous groups (designated as KOG) (Koonin et al., 2004), which provides a protein lineage based on seven model organisms, including *C. elegans*, but, like WormBase and UniProt, does not provide a direct connection between genes, enzymes, and reactions.

To connect potential metabolic genes to reactions, we developed a pipeline named SACURE (systematic annotation with manual curation and regression), which combines evidence from all six resources. To minimize false negatives, we started with a low myKEGG score threshold, resulting in 3,424 genes associated with metabolic KO groups. To rationalize accepting and rejecting gene-KO group associations and to standardize our annotations, we formulated the decision making process by machine learning. Weights were assigned to each resource depending on their contribution to annotation decisions (Figures 2C and S2). Overall, our GPR annotations were driven by clustering patterns in myTree, available KEGG annotation, and sequence similarity-based scores in myKEGG, while WormBase and KOG added support when these resources were not sufficient to confidently make a decision. UniProt did not contribute significantly, likely because it is redundant with the other sources (Figure 2D).

In total, SACURE identified 1,435 metabolic genes (Figures 2D, 2E, and S1; Table S1), 455 of which were missing in KEGG (31.7%). These genes brought in 175 metabolic reactions from KEGG for which no *C. elegans* gene was previously assigned (Figure 2E). Most of these genes (n = 343) complemented gene families that were only partially annotated in KEGG. For instance, three members of the trehalase (EC 3.2.1.28) family (*tre-1*, *tre-3*, and *tre-5*) were annotated in KEGG, and our annotation pipeline added two more (*tre-2* and *tre-4*), thus recovering the entire gene family as listed in WormBase (Harris et al., 2013) (Figure 2F).

Reconstruction of a Template *C. elegans* Metabolic Network: Pathway-by-Pathway Reconstruction and Gap Filling

The SACURE-annotated, KEGG-based reactions formed the backbone of our reconstruction. However, this collection does not provide a functional network that describes the conversion of nutrients into biomass and energy. First, not all *C. elegans* metabolic pathways are captured by reactions in the KEGG database. Examples include collagen, N-glycan, and iron-sulfur cluster biosynthesis. Such pathways were manually reconstructed using literature searches or MetaCyc (Caspi et al., 2014), yielding a total of 81 custom reactions (Table S3). In addition, many KEGG reactions were modified using *C. elegans*-relevant rather than generic compounds (e.g., cyclopropane fatty acids instead of long chain carboxylate), which resulted in 34 additional custom reactions.

Second, many pathways have gaps because of missing enzymes. SACURE filled 74 gaps in the KEGG template network, thus validating the computational annotations, which were made independent of gap analyses (Figure 3A). For instance, the *C. elegans* gene encoding methylglutaconyl-CoA hydratase was missing in KEGG, thus forming a gap in the leucine degradation pathway. Both myKEGG and myTree captured ECH-5 as a candidate for this enzyme (Figure 2B). Remaining gaps were iteratively detected by FBA (Figure 1B), first to find reactions that could not carry flux and then to identify potential rescue reactions (gap-fillers). Many gaps were manually filled by relaxing SACURE criteria or by inspecting homology with proteins from other organisms (lenient annotation, n = 72; Figure 3B). For instance, the tryptophan degradation pathway utilizes an arylformamidase, but this enzyme was annotated neither in KEGG nor by SACURE (Figure 3B). *afmd-1* was accepted as a gap-filler based on manual inspection of sequence homology. In 20 lenient annotations, multiple candidate genes could be linked to an enzyme, and the specific gene encoding the enzyme remains to be determined (TBD in the Gene column of Table S3). An additional 77 gap-filling reactions were annotated by manual curation based on the literature. For instance, *gob-1* has experimentally been determined to encode a trehalose-6-phosphatase (Kormish and McGhee, 2005) (Figure 3C).

Some gaps could not be filled by any of the above-mentioned methods. However, the corresponding reactions do need to be incorporated to enable network flux. In some cases, the gene encoding the relevant metabolic enzyme has not yet been identified. Two such enzymes are found in the carnitine biosynthesis pathway (Figure 3D). The first is a peptidase that degrades proteins harboring methylated lysine, and the second is an aldolase converting 3-hydroxytrimethyllysine into 4-trimethylammoniobutanal. While this conversion has been observed in mammals, little is known about the responsible enzymes (Vaz and Wanders, 2002). Both reactions were incorporated during gap filling to rescue the other four reactions in the carnitine biosynthesis pathway (Figure 3D), which is believed to be functional in *C. elegans* (Deusing et al., 2015). Finally, 39 gaps (37 transports and two metabolic conversions) were filled without association with any genes or uncharacterized enzymes, but based solely on FBA. For example, there is an annotated gluconokinase enzyme that is predicted to be functional only if gluconic acid can enter *C. elegans* cells. Therefore, we added a predicted gluconic acid transport reaction to the model (Figure 3E). The details of manual reconstruction process are provided in the comments and notes of Tables S1 and S3.

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

Our *C. elegans* metabolic network model is particularly aimed at converting bacterial diet (input) into worm biomass (output) and generating energy (Figure 1C). To enable network functionality that accurately reflects *C. elegans* metabolism, we added reactions for breakdown of bacteria and generation of worm biomass to the model. Bacterial biomass was based on the composition of an average *E. coli* cell (Neidhardt et al., 1990). In addition, we included specific information on the lipid composition of *E. coli* OP50 (Satouchi et al., 1993). *C. elegans* biomass composition has been only partially determined. We used specific information for *C. elegans* where available, including lipids (Brock et al., 2007; Brooks et al., 2009; Hutzell and Krusberg, 1982; Satouchi et al., 1993), trehalose (Miersch and Döring, 2012), and glycogen (Cooper and Van Gundy, 1970). To approximate the missing variables in nucleic acid and amino acid composition, we used the values previously established for yeast (Heavner

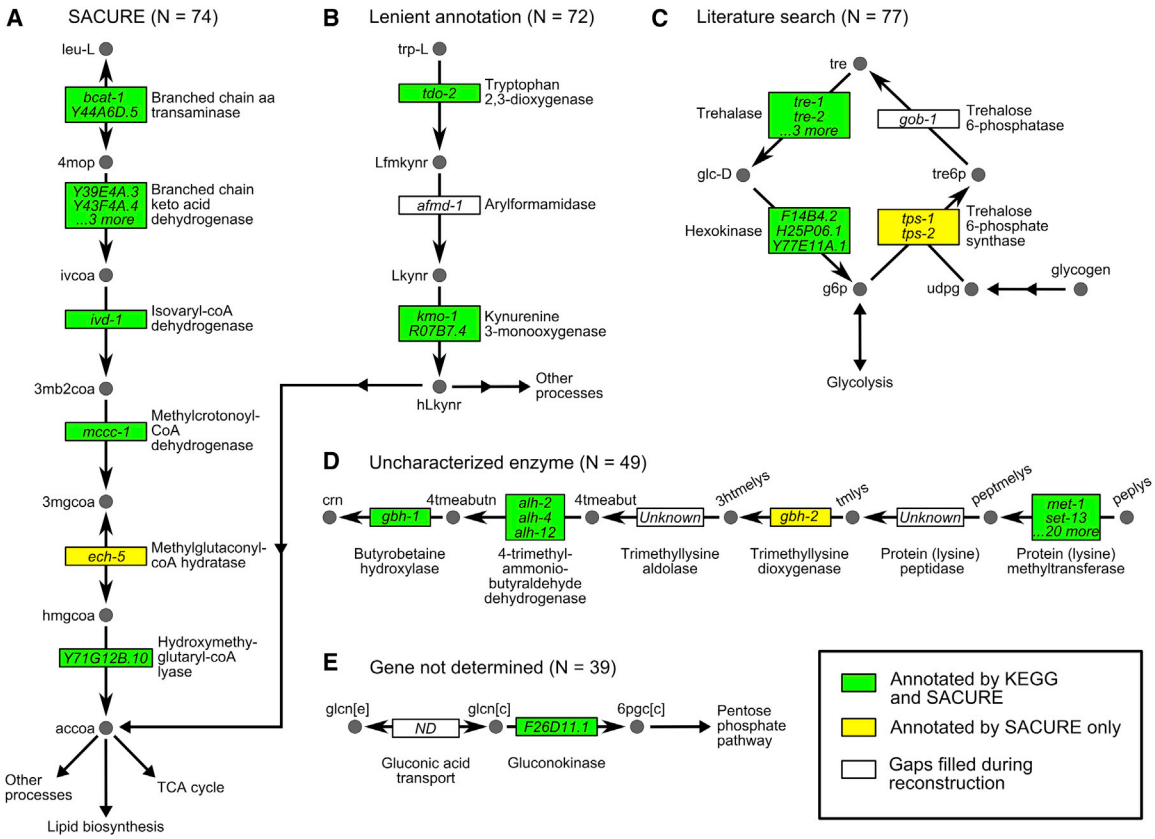


Figure 3. Gap-Filling Examples

(A–E) Example pathways with gaps. The different gap-filling methods are indicated as titles. The total number of reactions for each method is given as N. Up to three genes encoding the enzymes carrying out each reaction are shown in boxes with the enzyme names indicated. See Table S4 for compound abbreviations and names.

(A) Leucine degradation pathway in KEGG with a gap (yellow box) filled by SACURE.

(B) Tryptophan degradation pathway with a gap filled by lenient annotation.

(C) Trehalose production and degradation pathway with a gap filled by literature-based annotation.

(D) Carnitine biosynthesis with two gaps filled by adding two unknown enzymes for which no KO group is available for annotation.

(E) Potential gluconic acid degradation pathway with a gap filled by a transport reaction not annotated and not found in other eukaryotic models in BiGG.

et al., 2012). Overall, one bacterial and four *C. elegans* biomass compositions were generated, with the latter representing the output in four different modes of animal growth at different life stages (Table S2). To connect the defined input and output, we added 29 reactions for the degradation of bacterial biomass and 19 reactions for the assembly of precursors into *C. elegans* biomass (Figure 1C; Table S3).

Additional inputs and outputs include nutrients such as glucose and metabolites not consumable by biomass or energy generation, such as signaling molecules, which are synthesized or degraded by peripheral pathways. Flux to and from these pathways was driven by hundreds of transport and exchange (with the environment) reactions, as well as 82 demand and sink reactions (Thiele and Palsson, 2010) for the end products (Figure 1C; Table S3). Although transporters are generally uncharacterized in *C. elegans*, we included 17 known transporter proteins that carry 27 metabolites between the different compartments of the network (see below). We further assumed that metabolites transportable in yeast and human models (Schellenberger et al., 2010) are also transportable in *C. elegans* and incorporated additional transport reactions accordingly.

Finally, reactions obtained from KEGG are typically not curated for stoichiometry and reversibility (Feist et al., 2009). To define these parameters, we used evidence from BiGG (Schellenberger et al., 2010), MetaCyc (Caspi et al., 2014), SEED (Henry et al., 2010), and BRENDA (Chang et al., 2015), as well as literature curation.

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

Metabolism is highly compartmentalized in specific sub-cellular spaces such as the mitochondria. Since the precise sub-cellular localization is known for only few *C. elegans* proteins, we aimed at a minimal network and predicted reaction localization to three compartments: cytosol, mitochondria, and extracellular space. We employed multiple resources that use protein sequence, enzyme type, and reaction (Figure S3). In addition, we determined the flux carrying capacity of each reaction when localized to mitochondria, cytosol, or both.

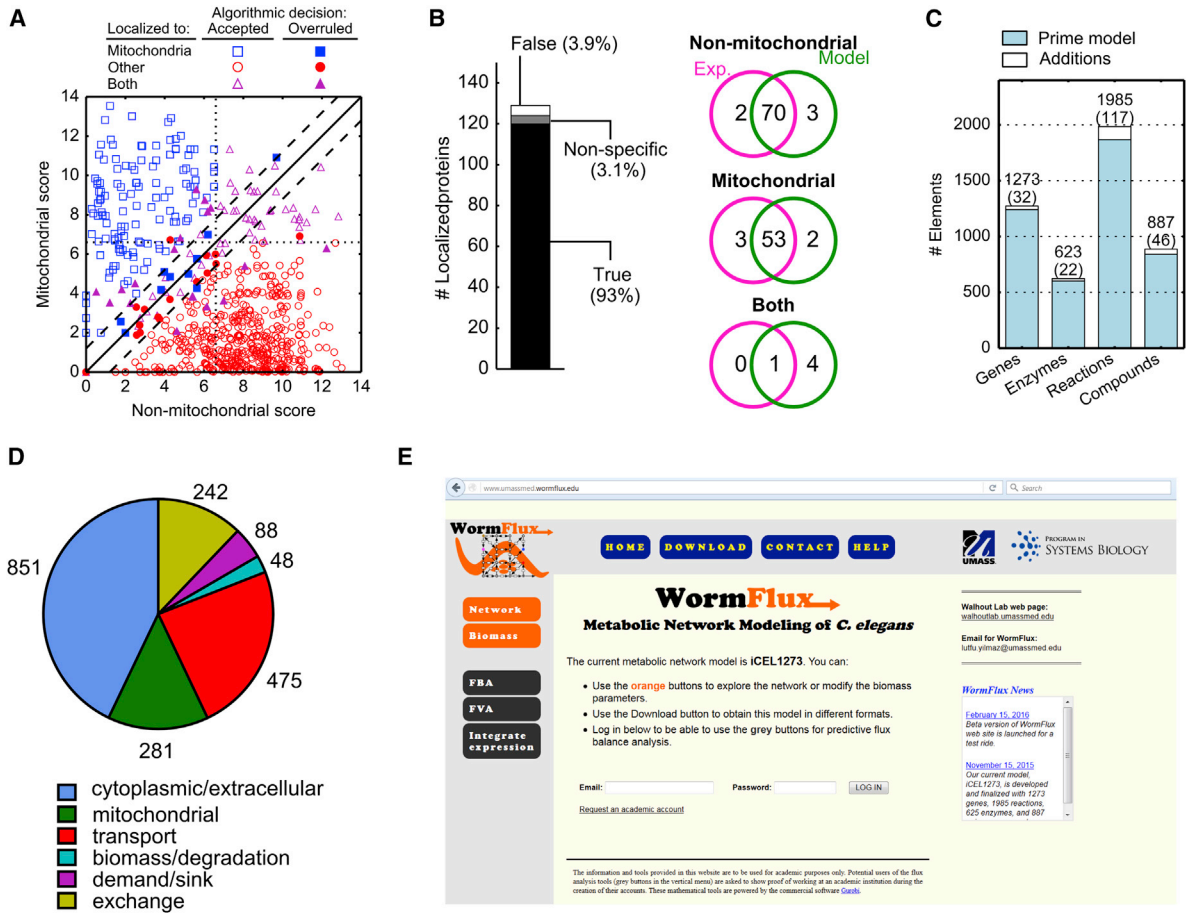


Figure 4. Reaction Localization and Key Statistics of Reconstructed Model

(A) Score-based assignment of reactions to mitochondria and other compartments (cytosolic or extracellular). An algorithmic compartment score greater than 6.2 (dotted lines) indicates automatic localization to that compartment. If two scores are within 1.2 of each other (dashed lines), the reaction is localized to both compartments. In other cases, the higher score determines the compartment. The manual decisions that violate these rules are indicated as closed shapes. The solid line shows identity.

(B) Computational determination of enzyme localization was validated against experimental localization data obtained from WormBase (Exp).

(C) The number of genes, enzymes, reactions, and compounds included in the PRIME model and additions from semi-automated reconstruction. The numbers at the top indicate elements in the final model, and the numbers in parentheses indicate additions.

(D) Distribution of reactions according to function and compartment.

(E) WormFlux provides an integrated platform for iCEL1273 access.

We rationalized the localization decisions by first calculating an overall score for the mitochondrial and non-mitochondrial localization of each reaction based on a weighted sum of evidence from different predictors. This score was then used to algorithmically decide whether a reaction is mitochondrial or not according to two thresholds (Figure 4A), which were based on the best agreement of predictions with the manually reconstructed template. Reactions were then re-localized to maximize the agreement between the decisions and predictions. All decisions were manually curated. For example, in many cases where a reaction was localized to both compartments due to low scores indicating lack of evidence (i.e., exceptions in lower-left quartile of Figure 4A), additional evidence from similar reactions catalyzed by the same protein were used. In the end, the localization of fewer than 10% of reactions was overruled manually (Figure 4A; Table S3).

We validated the predictions using experimentally determined localization of 130 proteins available in WormBase (Harris et al., 2013) (Table S6). One protein (ACO-1) was eliminated from this analysis as its annotation was in fact driven by the experimental annotation available in the used resources. For the remaining 129 proteins in this validation set, the sub-cellular localization was wrongly predicted for only five (3.9%) (Figure 4C). In addition, four proteins (3.1%) were predicted to localize to both mitochondria and cytosol. This analysis completed the reconstruction of the PRIME model (Figure 1B), with 1,868 reactions and 1,241 genes (Figure 4C).

Completion of Reconstruction by Semi-automated Expansion of the PRIME Model

The PRIME model can carry out the metabolic functions depicted in Figure 1C and forms the scaffold for additional reactions. Since the PRIME model was reconstructed in a pathway-by-pathway manner, many SACURE-annotated reactions that are not part of well-defined pathways in KEGG and MetaCyc, or those that

seem to be part of incomplete pathways were not incorporated. Remaining reactions may connect pathways, form alternative pathways that perform overlapping metabolic functions, or comprise isolated pathways (i.e., interconnected reactions disconnected from the network as a group). We performed an exhaustive computational analysis of whether the remaining reactions can support flux when added to the network.

Using the PRIME model, we tested the flux carrying capacity of 704 SACURE-annotated reactions that were left out during manual reconstruction. We also included helper reactions that could connect the annotated reactions to the network, which covered spontaneous reactions listed in KEGG, reactions associated with uncharacterized enzymes listed in KEGG, and transport reactions obtained from BiGG (yeast and human models). Additional (custom) transport reactions were provided for every compound to allow the reconstruction of isolated pathways with few inputs and outputs. Proper connection to the network was algorithmically defined as the ability of a reaction to carry flux, without need for a custom transport reaction that has no other function than rescuing this reaction. For instance, our annotations associated AMX-2 with the conversion of aminoacetone to methylglyoxal (Tables S1 and S3). However, while this reaction was recovered by SACURE, it was not included in the PRIME model since it was disconnected and did not form a gap, as no pathway was dependent on it. The semi-automated reconstruction added a spontaneous reaction that represents the degradation of L-2-Amino-3-oxobutanoate to aminoacetone to connect the AMX-2-catalyzed reaction to the network. Since L-2-Amino-3-oxobutanoate is a by-product of glycine and threonine breakdown, this additional reaction provided a lateral connection between amino acid metabolism and methylglyoxal detoxification.

The semi-automated pipeline identified 233 connected reactions, leaving 471 reactions disconnected. Most of the disconnected reactions (n = 297) are parallel reactions of existing enzymes in the model that we believe are either not relevant to *C. elegans* metabolism or are connected to the network with unknown pathways. Others are enzymes with functions that do not form complete pathways linked to the current network. We also manually curated the 233 connected reactions and eliminated the majority (n = 179) because they did not contribute to the functionality of the model (see details in Supplemental Information and Table S7).

Overall, the semi-automated procedure incorporated 117 additional reactions (56 annotated reactions and 61 helper reactions, mostly transport and exchange) and 32 genes to the model (Figure 4C; Table S3). All SACURE-annotated reactions that were excluded from the model are listed in Table S7, together with the 28 KEGG reactions that were not annotated by SACURE (Figure 2E). The final model contains 1,273 genes, 623 enzymes, 1,985 reactions, and 887 metabolites and is referred to as iCEL1273 (Figures 2E and 4C). The distribution of reactions into mitochondria and cytosol, as well as to different reaction types, is provided in Figure 4D.

WormFlux: a Dedicated Website for iCEL1273

iCEL1273 is available at a custom-made website called WormFlux (wormflux.umassmed.edu) (Figure 4E). WormFlux provides a searchable database with detailed descriptions of model elements and their annotations in gene, enzyme, reaction, compound, and pathway pages. To facilitate applications that require modifications in biomass, we provide a "Biomass" tool as part of WormFlux, which can take user-defined biomass parameters and adjusts bacterial degradation and worm biomass assembly pathways accordingly.

Validation of iCEL1273: Reproducing Observed Mass and Energy Balance during Two Stages of Life

To demonstrate that iCEL1273 adequately represents the conversion of bacterial diet into *C. elegans* biomass, we gathered information on bacterial intake, biomass production, O_2 consumption, and CO_2 release observed with *C. elegans* during L4 larval and reproductive young adult stages, where biomass production takes the form of growing body size and generation of progeny, respectively. Although precise bacterial ingestion rates are not available, the other rates are in relatively tight ranges. For each life stage, we constrained the model with experimental ranges observed for three of the four measured rates and determined the theoretical range for the fourth rate using flux variability analysis (FVA). The theoretical range overlapped with the experimental observation in every case (Figure 5A), showing that iCEL1273 can quantitatively explain *C. elegans* growth at quasi-steady states.

The wide predictive ranges reflect the variability of flux in the absence of additional constraints, as the model has a large solution space to freely perform conversions such as production of formic acid instead of CO_2 to waste carbon. This observed level of flexibility is desired since the actual productivity of the metabolism cannot be limited to the four fluxes considered. Most importantly, the model must be able to produce significant amounts of ATP in excess of the requirements for biomass production in order to meet growth-associated maintenance (GAM, linked to biomass production) and non-GAM (NGAM, independent of biomass production, including movement) costs. In addition to these requirements routinely used in metabolic network models, the *C. elegans* model needs to address the unknown energetic cost of digesting bacterial biomass. We calculated the additional ATP that iCEL1273 can produce to meet GAM (excluding polymerization reactions for which energetic cost is already a part of biomass assembly reactions), NGAM, and bacterial digestion, when constrained with the above mentioned experimental data. The achievable ranges of values for these requirements formed a finite volume (Figure 5B), where GAM and NGAM were consistent with previous models (e.g., Reed et al., 2003; Förster et al., 2003; Oh et al., 2007). Thus, iCEL1273 can satisfy input/output rates for two different modes of growth (body size and offspring) and generate energy for maintenance and digestion costs. We arbitrarily picked the center of mass of the tetrahedron in Figure 5B to determine the final costs in each of the three categories, which dictated the coefficients of ATP in the corresponding reactions (Table S3; NGAM: reaction RCC0005, GAM: BIO0010, and digestion: DGR0007).

Validation of iCEL1273: Gene Essentiality and Other Genotype-Phenotype Relationships

To test the ability of iCEL1273 to predict the outcome of genetic perturbations, we compared genes predicted to be essential by FBA to experimentally defined essential genes. FBA predicts

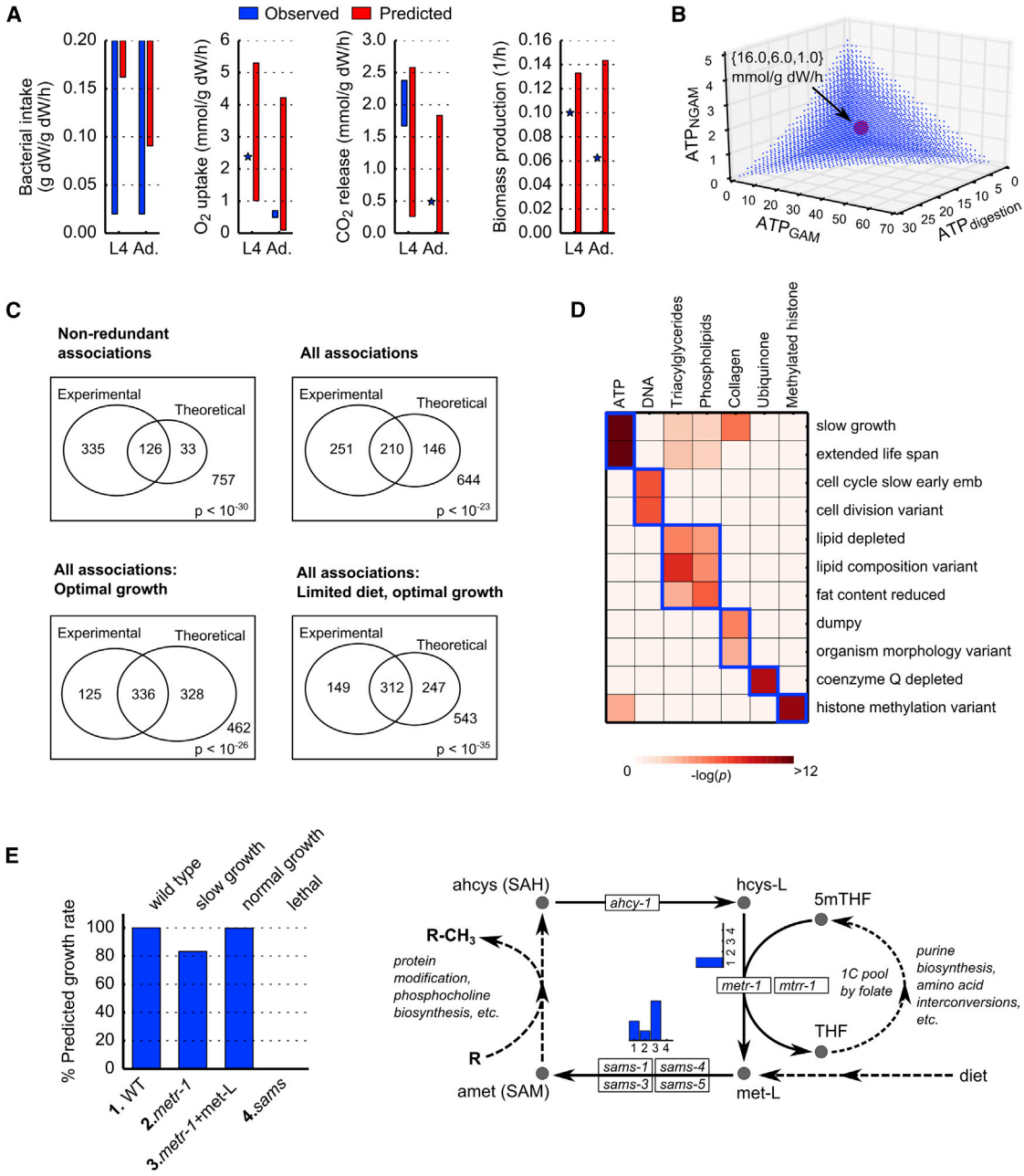


Figure 5. Model Validation

(A) Comparison of model-predicted flux ranges with observed production/consumption rates during growth in two stages of life (Table S10). When the model is constrained with three of the four experimental fluxes (blue bars if a range and stars if a single point), the predicted range for the fourth flux (red bars) overlaps with or covers the experimental value, meaning all experimental rates can be satisfied simultaneously. The analysis was performed for L4 and young adult (Ad.) stages.

(B) Predicted production of excess ATP to address unknown maintenance (GAM and NGAM) and digestion costs. All combinations shown by blue dots can be achieved while simultaneously satisfying all experimental constraints in (A). The center of mass of the tetrahedron was arbitrarily used to set the default values for each parameter in subsequent applications.

(C) Association of genes with experimental no-growth (lethal, larval lethal, larval arrest, embryonic lethal, embryonic arrest, and sterile) phenotypes with genes predicted by four different approaches as described in the main text. The statistical significance of associations is indicated by hypergeometric p values.

(D) Association of specific phenotypes with the predicted reduction in the production rates of different biomass precursors, specific metabolites, and ATP. The boxes indicate relationships that were expected based on phenotype description and its relevance to the produced metabolites.

(E) Mechanistic predictions of genotype-phenotype and diet-phenotype relationships in the methionine salvage pathway. The predicted growth rates for mutant animals with supplemented metabolites are compared to experimental observations indicated at the top (left). The methionine salvage pathway, together with the adjacent one-carbon pool by folate, is shown on the right. In this pathway, methionine synthase (METR-1) uses methylcobalamin (vitamin B12) as a cofactor. The inset bar graphs indicate fluxes (relative values) of corresponding reactions in the four cases shown on the left. For abbreviated compounds, see Table S4.

essentiality when a severe reduction in predicted biomass production is observed after all reactions that are non-redundantly associated with a gene are constrained to zero flux. Experimentally determined essential genes include those associated with lethal, growth arrest, or sterile phenotypes upon RNAi or mutation as reported in WormBase (Harris et al., 2013). Using a threshold of 50% for the reduction in biomass production efficiency, the model predicted 159 essential genes. About 80% of these genes are indeed essential (hypergeometric p value = 1.1×10^{-31}) (Figure 5C, top left Venn diagram). Biomass reduction thresholds between 30%–84% gave the same result, whereas we started underestimating or overestimating essentiality beyond these levels.

Of the 33 genes that were incorrectly predicted to be essential (Table S8), 19 are involved in the production of glycans and lipids, which are adjustable components of biomass. Different lipid and glycan compositions may support viability (Berninsone, 2006; Brock et al., 2007; Perez and Van Gilst, 2008; Zhang et al., 2003). However, the model uses a constant composition that is essential for biomass generation. Of the remaining 14 genes, five function in DNA polymerization, but may represent specific DNA processing activities that are not essential. For instance, the DNA polymerase *polh-1* is involved in DNA repair, which is not essential for viability (Harris et al., 2013). Four of the remaining nine genes, while not essential, do confer slow growth, sick, or reduced fecundity phenotypes. The last five genes were only tested by RNAi, and incomplete knockdown may explain the lack of essentiality. Indeed, the mutation of one of these genes, *idi-1*, results in larval arrest (Yochem et al., 2005), however this information was not yet present in WormBase (Harris et al., 2013).

Next, we focused on the 335 experimentally determined essential genes that were not correctly predicted by iCEL1273 (Figure 5C). Missing a large number of essential genes is expected for several reasons. First, our initial definition of essentiality assumed that redundant genes in GPR associations could fully replace each other. However, paralogs may be individually essential for viability when they function in separate physiological compartments. For instance, the pyruvate kinases PYK-1 and PYK-2 are expressed in muscle and intestine, respectively, and *pyk-1* is essential for viability, whereas *pyk-2* is not (Harris et al., 2013). To better capture these genes, we re-predicted essentiality, this time prohibiting functional replacement by paralogs in GPRs. This reduced the number of false negatives from 331 to 251 (Figure 5C, top right Venn diagram).

The second reason for missing essential genes is the underlying assumption that all reactions in the network can carry flux. However, many reactions are only conditionally active, as not all genes are expressed under all conditions. To address this issue, we computationally derived an optimal state of growth in the solution space, wherein maximum biomass production was achieved with minimum total flux in all reactions, a method known as parsimonious enzyme usage FBA (Lewis et al., 2010; Machado and Herrgård, 2014). Assuming that the protein cost of a flux is proportional to its magnitude, and also that cells are programmed to grow with minimal cost for the synthesis of the metabolic machinery, we used this specialized FBA approach to predict reactions and thus genes that participate in the optimal growth state and assumed that these genes are essential for viability. Importantly, optimized flux distribution of the entire network turned out to be an excellent predictor of gene essentiality by itself and further reduced the number of false negatives from 251 to 125 (Figure 5C, bottom left Venn diagram).

The third reason for missing essential genes is the assumption that all bacterial degradation products are available to *C. elegans* metabolism. We suspect this may not be the case for some nutrients, such as nucleotides coming from the degradation of bacterial nucleic acids. In addition, in the context of the animal's physiology, not all nutrients will be uniformly available in all tissues. As a result, some tissues may be dependent on specific metabolites such as trehalose, which is hypothesized to be an energy commodity in *C. elegans* (Braeckman et al., 2009; McElwee et al., 2006). We simulated an optimal growth state again by minimizing total flux when only amino acids, stored lipids, and trehalose were available as nutrients. This approach was again an excellent predictor of lethality (Figure 5C, bottom right Venn diagram) and captured 23 essential genes that were missed before (Table S8), including genes associated with nucleic acid biosynthesis (n = 15) and two of the five trehalases (*tre-1* and *tre-5*). Taken together, by using different types of nutrient and gene function simulations, the model correctly predicted 359 of 461 essential genes (77%) leaving 102 genes unexplained.

Next, we changed the objective of FBA to the maximization of demand reactions instead of biomass generation. The model predicted 45 of the remaining 102 genes to be essential for the production of vital molecules such as ubiquinone, methylated histones, and inositol phosphates, which may be required for the organism to grow, but are not included in the general biomass composition (Table S8). The remaining 57 essential genes that were not predicted by the model may be associated with the production of other metabolites not included in our demand list or may be essential under nutritional conditions that are yet to be explored.

Finally, we investigated the association between a set of 11 phenotypes and seven network objectives such as maximization of ATP or collagen production. For each phenotype, we found that genes predicted to be essential for the production of related metabolites are strongly associated with the gene sets reported in WormBase (Figure 5D). For instance, collagen is the main component of the *C. elegans* cuticle, and its biosynthetic production in the model is related to morphological phenotypes such as Dpy. Low efficiency of ATP generation is a good predictor of slow growth and also lifespan extension, as expected (Chin et al., 2014; Van Raamsdonk et al., 2010). It is important to note that the usage of WormBase during gene annotations did not create a testing bias for WormBase-derived phenotypes. The phenotypic descriptions are not specifically linked to the annotated functions with the exceptions of some genes associated with coenzyme Q depleted and histone methylation variant phenotypes, for which, the elimination of WormBase descriptions from SACURE input did not affect our predicted function. Altogether, these phenotypic relationships validate the pathways and GPR annotations comprising iCEL1273.

Validation of iCEL1273: Gene Essentiality and Genotype-Phenotype Relationships in Methionine Salvage Pathway

To demonstrate the utility of iCEL1273 for specific pathway analyses, we examined the methionine salvage pathway. At the center of this pathway is methionine synthase (MS, encoded by

metr-1), which uses vitamin B12 to convert homocysteine into methionine (Yilmaz and Walhout, 2014). Mutant *metr-1* animals cannot utilize vitamin B12 and exhibit a growth delay (Watson et al., 2014). In addition, supplementation of methionine partially rescues the *metr-1* mutant's developmental delay (Watson et al., 2014). In accordance with our experimental findings, simulated deletion of *metr-1* in iCEL1273 lowers the biomass production rate (Figure 5E) and growth is predicted to be restored when methionine uptake is allowed during FBA. Importantly, the model predicts that the growth reduction in *metr-1* mutants is mediated by reduced flux in the methionine salvage cycle rather than the connected folate pathway, which is consistent with our previous observations (Watson et al., 2014) (Figure 5E). The model also correctly predicts that the reaction that converts methionine to S-adenosylmethionine is essential. This example shows that iCEL1273 can provide mechanistic predictions for genotype-phenotype and diet-phenotype associations at the pathway level.

Case Study: Analysis of Dauer Metabolism by Integration of Gene Expression Data

Metabolic network flux can be rewired in response to environmental or physiological cues (Watson et al., 2015). One mechanism of network rewiring is by regulating metabolic gene expression. We tested the ability of iCEL1273 to predict metabolic network rewiring when *C. elegans* enters the dauer stage in response to adverse conditions. Regular *C. elegans* larval development is associated with a fast aerobic metabolism that builds large amounts of biomass in a short time, while dauer metabolism is characterized by a slow microaerobic metabolism that utilizes stored compounds such as fatty acids and glycogen as nutrients (Hu, 2007).

To predict metabolic rewiring in dauer versus growing animals, we used gene expression data from dauers and dauer recovery larvae (Wang and Kim, 2003). Altogether, 144 genes are upregulated and 241 are downregulated in dauer larvae ($p < 0.001$; Table S1). We used this gene expression data to identify two corresponding sets of reactions that are assumed to be in an on (n = 231) or off (n = 136) state in dauer, respectively. We then determined a flux distribution that best fits these reactions following a previously established integration method (Shlomi et al., 2008) (Figure 6A). We made two modifications to this method to devise an optimization strategy for our application. First, we performed flux fitting under three nutritional conditions: bacterial intake, usage of storage compounds (triacylglyceride and glycogen), or both. Second, we minimized total flux (absolute values) of reactions that were unrestrained during the fitting process (>83% of the network). Thus, we combined a gene expression integration method (Shlomi et al., 2008) and a purely predictive flux minimization method (Lewis et al., 2010; Machado and Herrgård, 2014) to find an optimal state where the network is most efficiently wired according to gene regulatory constraints. In addition, we back-calculated the number of genes that have consistent expression levels with the derived flux distribution (flux-compatible genes). We then evaluated optimization quality based on a high percentage of flux-compatible genes and a low sum of minimized flux in unrestrained reactions (inclusion of flux from fitted reactions in this sum did not change the conclusions; Figure S4).

For dauer larvae, both the highest number of flux-compatible genes and lowest sum of minimized flux were obtained when storage compounds were used as nutritional input (Figure 6B). This is in agreement with the fact that dauers do not eat and need to sustain their physiology by catabolizing stored energy sources (Hu, 2007). To model the network wiring of growing larvae, we reversed the up- and downregulated genes. In this case, the optimal fit was obtained when a bacterial diet was used as nutritional input, which agrees with the physiological reality of growing larvae (Figure 6B). Importantly, the model correctly predicted growth (biomass production) for growing larvae, but not for dauer larvae (Figure 6C). In addition, dauer larvae had predicted lower metabolic activity based on ATP production, O_2 consumption, and flux activity (Figure 6C). Because the integration approach is only semi-quantitative (e.g., expression levels are grouped into on and off states before flux fitting), the flux comparisons between the two states cannot be taken as a quantitative measure. Altogether, iCEL1273 correctly predicted, solely based on gene expression data, that the metabolism was adjusted for stored resources, low metabolic rate, and no biomass generation in dauer state and the use of bacterial diet, higher metabolic rate, and growth in the recovery state.

Compartmentalization of Dauer Metabolism

Several metabolic properties observed in the dauer larvae flux distribution (Figure 6A) are in agreement with known features of dauer metabolism. For instance, iCEL1273 correctly predicts a rewiring from the TCA cycle to the glyoxylate shunt, as well as a shift from oxidative phosphorylation to fermentation that results in the production of succinate from the reversal of the succinate-to-fumarate conversion (Braeckman et al., 2009; Burnell et al., 2005; Holt and Riddle, 2003). The predicted end product of this is propionate, which has been detected in the exometabolome of *C. elegans* exposed to anaerobic conditions (Butler et al., 2012). Another prediction from iCEL1273 is the concurrent production and utilization of trehalose, which is an important metabolite in *C. elegans* for energy production and desiccation prevention in dauer animals (Erkut et al., 2011).

We asked whether iCEL1273 could mechanistically explain the functionality of these rewired network properties under dauer conditions. We reasoned that the simultaneous production and consumption of trehalose might reflect distinct metabolic activities in different tissues, cells, or compartments. However, this could not be captured with whole-animal gene expression data. We modeled two hypothetical compartments, a microaerobic compartment that produces trehalose from stored fatty acids and an anaerobic compartment that uses trehalose as the sole energy source. Using FBA, we predicted flux distributions for maximum trehalose production and maximum energy generation in the respective compartments (Figure 7A). We found that the activation of the glyoxylate cycle results in greater trehalose production from fatty acids in the microaerobic compartment (Figure 7B). In the anaerobic compartment, the reversal of the TCA cycle to ferment trehalose all the way to succinate resulted in greater levels of ATP generation than when this pathway was blocked by limiting the flux through the fumarate reductase reaction (Figure 7C). The biological nature of each compartment is not yet known, but we hypothesize that they

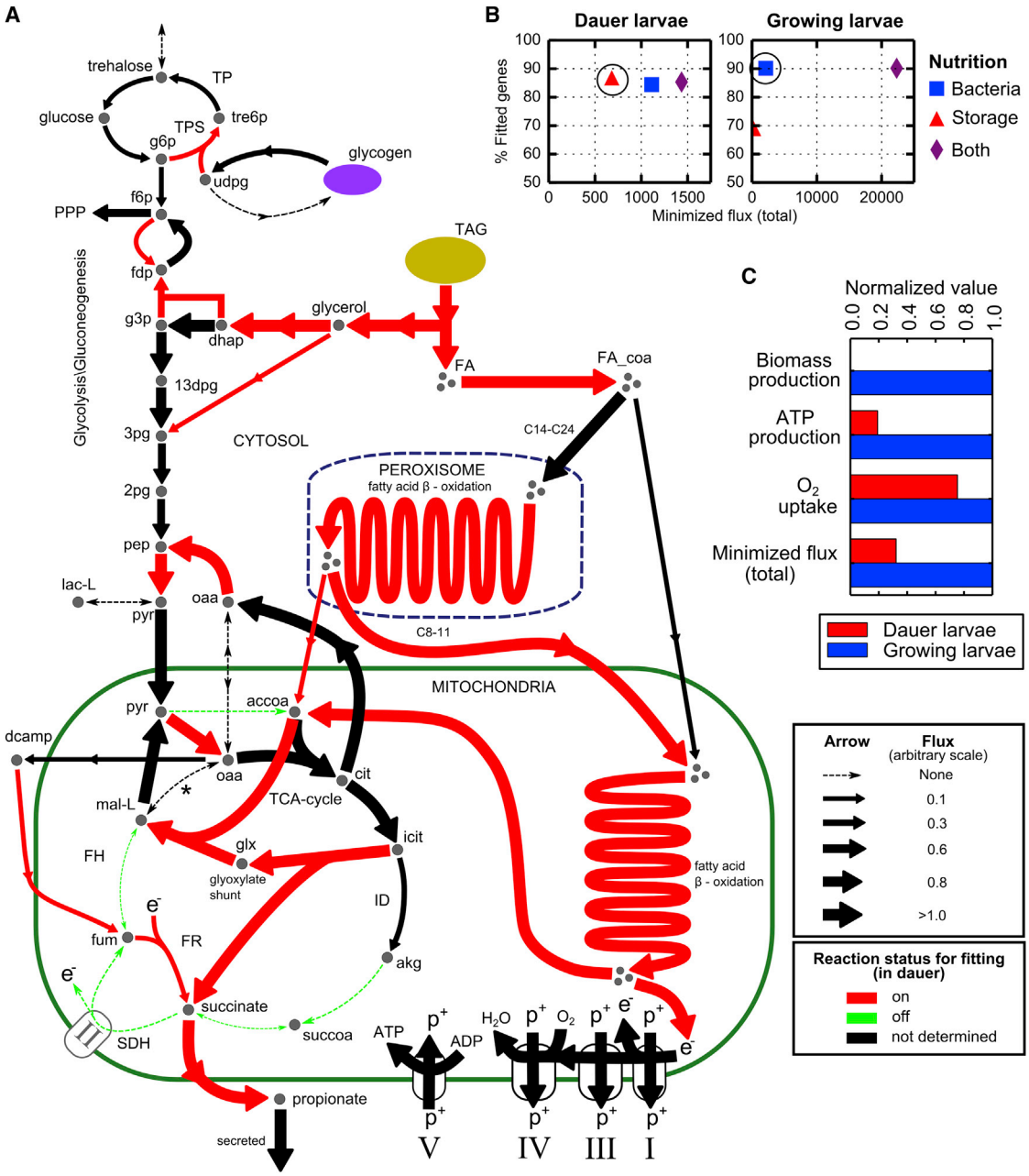


Figure 6. Integration of Gene Expression Data with iCEL1273

(A) Predicted optimal flux distribution in central carbon pathways in dauer animals that use stored compounds (triacylglycerides and glycogen) as nutrition. The red and green arrows indicate reactions to which the flux distribution was best-fitted based on gene expression. The black arrows indicate reactions with predicted incoming or outgoing fluxes with respect to best-fits. The flux imbalances observed in the figure are due to pathways not shown (e.g., about half of the incoming flux to 3 pg from 13 dpg and glycerol is diverted to amino acid metabolism [Table S3], hence the flux from 3 pg to 2 pg is less than the sum of fluxes producing 3 pg). The enzymes used in text are indicated by abbreviations. For abbreviated compounds, see Table S4.

(B) Optimization variables with three different nutritional conditions in two stages. A high percentage of genes with flux-compatible expression levels and a low sum of minimized fluxes (absolute values) are desired. The optimal states are circled.

(C) Summary of predicted dauer and growing larvae metabolism based on optimal states in (B). Each variable is normalized by itself (i.e., divided by the maximum of the two states).

may correspond to different tissues. Taken together, key properties of dauer metabolism can be predicted by iCEL1273, and our mechanistic predictions support the hypothesis that trehalose acts as a commodity metabolite (Braeckman et al., 2009).

DISCUSSION

Metabolic network models serve both as knowledge bases and predictive tools (Andersen et al., 2008; O'Brien et al., 2015;

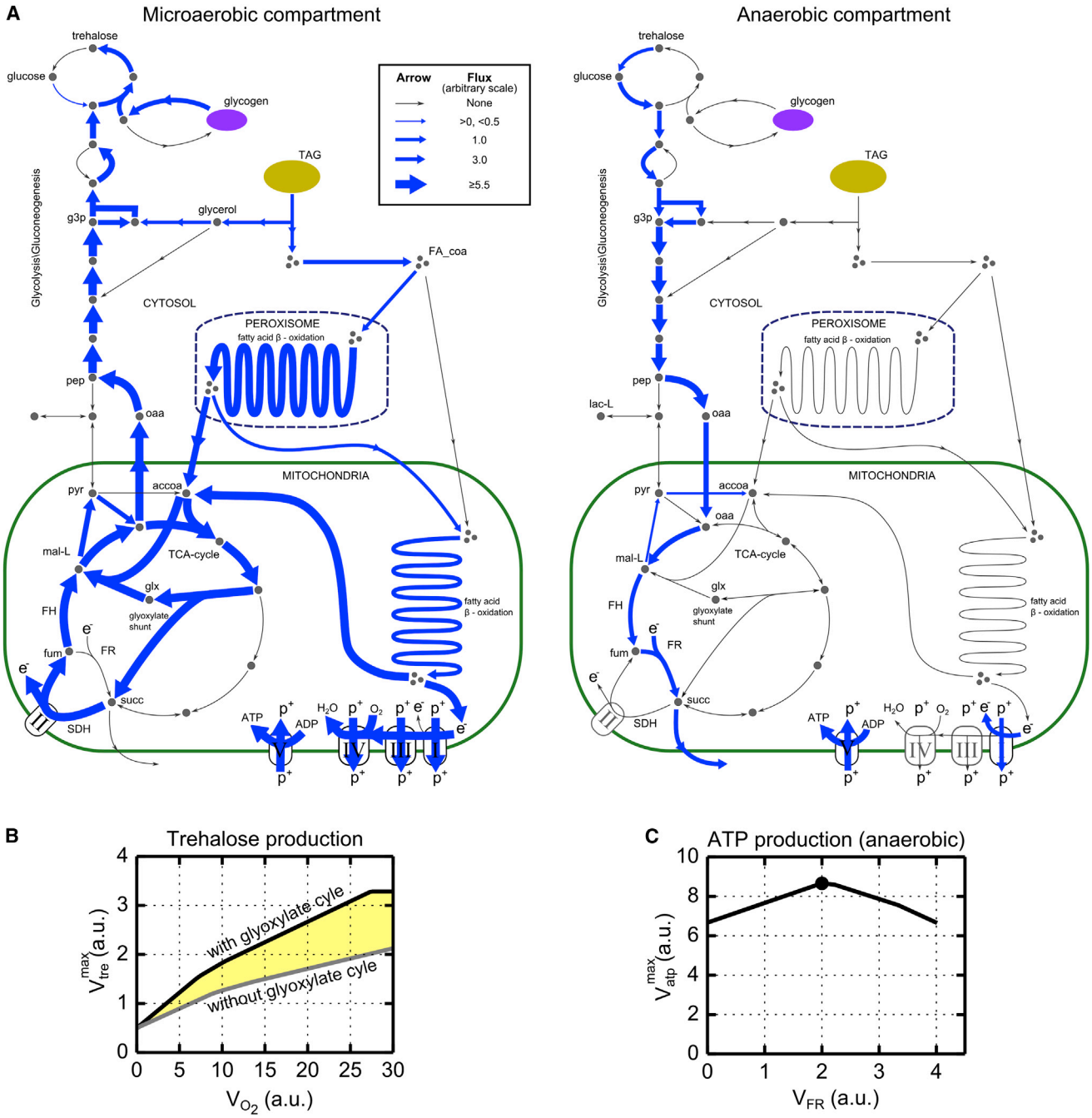


Figure 7. Mechanistic Analysis of Dauer Metabolism Using iCEL1273

(A) Two hypothetical compartments proposed to explain the key metabolic properties of dauer larvae (Figure 6A). The predicted fluxes are shown.

(B) Analysis of maximum trehalose production in the microaerobic compartment in (A) as a function of oxygen availability. The black and gray curves were obtained with and without the glyoxylate pathway (by constraining the relevant reactions to carry zero flux), respectively. The yellow region indicates the additional amount of trehalose that can be generated from the glyoxylate pathway.

(C) Sensitivity of ATP generation in the anaerobic compartment to the flux in fumarate reductase (FR) reaction. The circle indicates the optimal point where energy is maximized.

Oberhardt et al., 2009; Shlomi et al., 2008). The annotation database of iCEL1273 (wormflux.umassmed.edu), and its predictive power shown by multiple validation tests make it a suitable metabolic model for *C. elegans*.

The metabolic network of *C. elegans* has not been systematically studied before. To make a comprehensive list of metabolic genes and reactions in this organism, we developed and validated SACURE as an objective annotation pipeline (Supplemental

Information) and used the outcome in network reconstruction. Additional annotations came from the modeling-based reconstruction. It is important to note that model validation tests carried out in this study verify both the network structure and the annotations, as gene-reaction associations were extensively used in these tests. However, metabolic annotations are not yet complete for *C. elegans*. In the future, it is likely that new annotated pathways can be incorporated into iCEL1273, by connecting them to the central network or diet at one end and to worm biomass or demand/sink reactions at the other (Figure 1C). We also envision that additional compartmentalization can be incorporated into the model when sufficient experimental evidence is collected for protein localization in *C. elegans*. Thus, iCEL1273 can be considered a minimal global-scale metabolic model that is expected to evolve as more information is obtained. For future additions, a good starting point would be the annotated reactions in Table S7 that were excluded from the model.

iCEL1273 can be further refined when more precise information of both *C. elegans* and bacterial biomass composition becomes available. We approximated *C. elegans* biomass parameters that were not yet measured using data from yeast. Further, the variability of biomass composition in different stages of life and in different tissues of the animal will be considered in the future. Finally, a major goal will be to incorporate different bacterial compositions for different *C. elegans* diets. The Biomass tool of WormFlux as well as exchange, demand, and sink reactions readily available in the network can be used to control the dietary input and biomass output of iCEL1273.

With increasingly accurate descriptions input, output, and experimental constraints, FBA will become a powerful predictive tool to explore metabolic network properties and functionality at a systems level. However, it is important to note that FBA cannot be used to predict metabolite concentrations or to make a dynamic simulation of animal growth because of the steady-state assumption used. These limitations are offset by constraint-based approaches that allow data integration for a given environmental condition. As we showed for dauer metabolism, gene expression data can be used to constrain the network and correctly capture the relevant metabolic state. Similar methods can be applied to tissue-derived expression data to derive tissue-specific metabolic networks and states (Shlomi et al., 2008; Jerby et al., 2010). Furthermore, methods have recently been developed for integrating metabolomics data with metabolic networks (Töpfer et al., 2015), which may open exciting new opportunities for studying *C. elegans* metabolism.

EXPERIMENTAL PROCEDURES

Constraint-Based FBA

All mathematical modeling procedures used were based on FBA (Orth et al., 2010). Briefly, the main idea of FBA is to satisfy mass balance at every node (metabolite) of the metabolic network simultaneously. Assuming steady state, total flux in and out of each node (i.e., the difference in production and consumption rates of compounds) equals zero, which is represented by Equation 1, where, S is the stoichiometry matrix of reaction coefficients (dimensions n × m; n = number of compounds and m = number of reactions), v is the flux vector (m × 1), and 0 is the vector indicating zero sum of fluxes at each node (n × 1).

$$\boldsymbol{S}\cdot\boldsymbol{v} = 0. \qquad \text{(Equation 1)}$$

The solution to Equation 1 alone is the null space of S. To obtain a biologically meaningful solution in this space, flux values are first constrained based on thermodynamic and other relevant information by Equation 2. Typically, these constraints include reaction reversibility rules (e.g., the flux of an irreversible reaction can only take positive values; $0 < v < 1{,}000$, with 1,000 used as an arbitrary upper limit that represents infinity) and known or prescribed uptake/secretion rates in exchange reactions (e.g., for an uptake reaction that is set to provide up to 1 unit of a metabolite to the system: $-1 \leq v \leq 0$). Allowed uptake rates characterize the specific input (diet, oxygen, etc.) of a particular solution.

$$v_1^{\min} \leq v_i < v_i^{\max} \text{ for } i = \{1, 2, \ldots, m\}. \qquad \text{(Equation 2)}$$

In addition, a biological objective is defined for maximizing or minimizing a set of fluxes as shown in Equation 3. For instance, to predict the metabolism of optimal growth, the constant for the flux of biomass reaction is set at 1 ($c_{\text{biomass rxn}} = 1$) and the rest of the reactions at 0. Maximization of the objective then yields maximum possible biomass production, i.e., growth rate.

$$obj\ f = \sum_{i=1}^{n} c_i v_i. \qquad \text{(Equation 3)}$$

Equations 1–3 are solved together as a linear programming problem using a specialized solver. The solver used in this study was Gurobi Optimizer version 6 (Gurobi Optimization). FBA can be modified to carry out different applications of metabolic network modeling. For instance, the variability of the objective function as a function of a particular flux can be calculated to perform sensitivity analyses as in Figures 7B and 7C and gene expression data can be integrated using mixed integer linear programming (Shlomi et al., 2008) as in Figure 6. Details of different variants of FBA used in this study are provided in Supplemental Information.

Additional Methods

Details of methods used in every subsection of Results are available in Supplemental Information, following the same sub-section titles for convenience.

Model Availability

iCEL1273 can be downloaded from WormFlux in different formats including text, MS Excel, and SBML (Hucka et al., 2003).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and ten tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.04.012.

AUTHOR CONTRIBUTIONS

L.S.Y. and A.J.M.W. designed the study. L.S.Y. developed the model, performed the validation tests, and created the web application under the supervision of A.J.M.W. L.S.Y. and A.J.M.W. wrote the manuscript.

ACKNOWLEDGMENTS

We thank Emma Watson, Lucie Kozlowskie, Lesley MacNeil, and Juan Fuxman Bass for discussions and critical reading of the manuscript. This work was supported by NIH grant R21GM108045 to L.S.Y. and A.J.M.W.

Received: November 19, 2015
Revised: March 8, 2016
Accepted: April 15, 2016
Published: May 19, 2016

REFERENCES

Andersen, M.R., Nielsen, M.L., and Nielsen, J. (2008). Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. Mol. Syst. Biol. *4*, 178.

Berninsone, P.M. (2006). Carbohydrates and glycosylation. WormBook *18*, 1–22.

Braeckman, B.P., Houthoofd, K., and Vanfleteren, J.R. (2009). Intermediary metabolism. In WormBook: the Online Review of *C elegans* Biology, The *C. elegans* Research Community, eds., http://dx.doi.org/10.1895/wormbook.1.146.1, http://www.wormbook.org/chapters/www_intermetabolism/intermetabolism.html.

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*, e7545.

Burnell, A.M., Houthoofd, K., O'Hanlon, K., and Vanfleteren, J.R. (2005). Alternate metabolism during the dauer stage of the nematode *Caenorhabditis elegans*. Exp. Gerontol. *40*, 850–856.

Butler, J.A., Mishur, R.J., Bokov, A.F., Hakala, K.W., Weintraub, S.T., and Rea, S.L. (2012). Profiling the anaerobic response of *C. elegans* using GC-MS. PLoS ONE *7*, e46140.

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.

Chang, A., Schomburg, I., Placzek, S., Jeske, L., Ulbrich, M., Xiao, M., Sensen, C.W., and Schomburg, D. (2015). BRENDA in 2015: exciting developments in its 25th year of existence. Nucleic Acids Res. *43*, D439–D446.

Chin, R.M., Fu, X., Pai, M.Y., Vergnes, L., Hwang, H., Deng, G., Diep, S., Lomenick, B., Meli, V.S., Monsalve, G.C., et al. (2014). The metabolite α-ketoglutarate extends lifespan by inhibiting ATP synthase and TOR. Nature *510*, 397–401.

Coolon, J.D., Jones, K.L., Todd, T.C., Carr, B.C., and Herman, M.A. (2009). *Caenorhabditis elegans* genomic response to soil bacteria predicts environment-specific genetic effects on life history traits. PLoS Genet. *5*, e1000503.

Cooper, A.F., and Van Gundy, S.D. (1970). Metabolism of glycogen and neutral lipids by *Aphelenchus-Avenae* and *Caenorhabditis-Sp* in aerobic, microaerobic and anearobic environments. J. Nematol. *2*, 305–315.

Deusing, D.J., Beyrer, M., Fitzenberger, E., and Wenzel, U. (2015). Carnitine protects the nematode *Caenorhabditis elegans* from glucose-induced reduction of survival depending on the nuclear hormone receptor DAF-12. Biochem. Biophys. Res. Commun. *460*, 747–752.

Erkut, C., Penkov, S., Khesbak, H., Vorkel, D., Verbavatz, J.M., Fahmy, K., and Kurzchalia, T.V. (2011). Trehalose renders the dauer larva of *Caenorhabditis elegans* resistant to extreme desiccation. Curr. Biol. *21*, 1331–1336.

Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L., and Palsson, B.O. (2009). Reconstruction of biochemical networks in microorganisms. Nat. Rev. Microbiol. *7*, 129–143.

Förster, 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.

Gracida, X., and Eckmann, C.R. (2013). Fertility and germline stem cell maintenance under different diets requires nhr-114/HNF4 in *C. elegans*. Curr. Biol. *23*, 607–613.

Harris, T.W., Baran, J., Bieri, T., Cabunoc, A., Chan, J., Chen, W.J., Davis, P., Done, J., Grove, C., Howe, K., et al. (2013). WormBase 2014: new views of curated biology. Nucleic Acids Res. *42*, D789–793.

Heavner, B.D., Smallbone, K., Barker, B., Mendes, P., and Walker, L.P. (2012). Yeast 5 - an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. BMC Syst. Biol. *6*, 55.

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.

Holt, S.J., and Riddle, D.L. (2003). SAGE surveys *C. elegans* carbohydrate metabolism: evidence for an anaerobic shift in the long-lived dauer larva. Mech. Ageing Dev. *124*, 779–800.

Hu, P.J. (2007). Dauer. In WormBook, The *C. elegans* Research Community, eds., http://dx.doi.org/10.1895/wormbook.1.144.1, http://www.wormbook.org/chapters/www_dauer/dauer.html.

Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., et al.; SBML Forum (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics *19*, 524–531.

Hutzell, P.A., and Krusberg, L.R. (1982). Fatty acid compositions cise fatty acid composition observed in wild-type worms of *Caenorhabditis elegans* and *C. briggsae*. Comp. Biochem. Physiol. *73B*, 517–520.

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.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2015). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457–D462.

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 Biol. *5*, R7.

Kormish, J.D., and McGhee, J.D. (2005). The *C. elegans* lethal gut-obstructed *gob-1* gene is trehalose-6-phosphate phosphatase. Dev. Biol. *287*, 35–47.

Lewis, N.E., Hixson, K.K., Conrad, T.M., Lerman, J.A., Charusanti, P., Polpitiya, A.D., Adkins, J.N., Schramm, G., Purvine, S.O., Lopez-Ferrer, D., et al. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. Mol. Syst. Biol. *6*, 390.

Machado, D., and Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput. Biol. *10*, e1003580.

MacNeil, L.T., Watson, E., Arda, H.E., Zhu, L.J., and Walhout, A.J.M. (2013). Diet-induced developmental acceleration independent of TOR and insulin in *C. elegans*. Cell *153*, 240–252.

McElwee, J.J., Schuster, E., Blanc, E., Thornton, J., and Gems, D. (2006). Diapause-associated metabolic traits reiterated in long-lived daf-2 mutants in the nematode *Caenorhabditis elegans*. Mech. Ageing Dev. *127*, 458–472.

Miersch, C., and Döring, F. (2012). Sex differences in carbohydrate metabolism are linked to gene expression in *Caenorhabditis elegans*. PLoS ONE *7*, e44748.

Neidhardt, F.C., Ingraham, J.L., and Schaechter, M. (1990). Physiology of the Bacterial Cell: a Molecular Approach (Sunderland, Mass: Sinauer Associates).

O'Brien, E.J., Monk, J.M., and Palsson, B.O. (2015). Using genome-scale models to predict biological capabilities. Cell *161*, 971–987.

Oberhardt, M.A., Palsson, B.O., and Papin, J.A. (2009). Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. *5*, 320.

Oh, Y.K., Palsson, B.O., Park, S.M., Schilling, C.H., and Mahadevan, R. (2007). Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. J. Biol. Chem. *282*, 28791–28799.

Orth, J.D., Thiele, I., and Palsson, B.O. (2010). What is flux balance analysis? Nat. Biotechnol. *28*, 245–248.

Pang, S., and Curran, S.P. (2014). Adaptive capacity to bacterial diet modulates aging in *C. elegans*. Cell Metab. *19*, 221–231.

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

Reed, J.L., Vo, T.D., Schilling, C.H., and Palsson, B.O. (2003). An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). Genome Biol. *4*, R54.

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*, 213.

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

Soukas, A.A., Kane, E.A., Carr, C.E., Melo, J.A., and Ruvkun, G. (2009). Rictor/TORC2 regulates fat metabolism, feeding, growth, and life span in *Caenorhabditis elegans*. Genes Dev. *23*, 496–511.

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

Töpfer, N., Kleessen, S., and Nikoloski, Z. (2015). Integration of metabolomics data into metabolic networks. Front. Plant Sci. *6*, 49.

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.

Van Raamsdonk, J.M., Meng, Y., Camp, D., Yang, W., Jia, X., Bénard, C., and Hekimi, S. (2010). Decreased energy metabolism extends life span in *Caenorhabditis elegans* without reducing oxidative damage. Genetics *185*, 559–571.

Vaz, F.M., and Wanders, R.J. (2002). Carnitine biosynthesis in mammals. Biochem. J. *361*, 417–429.

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

Watson, E., and Walhout, A.J. (2014). *Caenorhabditis elegans* metabolic gene regulatory networks govern the cellular economy. Trends Endocrinol. Metab. *25*, 502–508.

Watson, E., MacNeil, L.T., Arda, H.E., Zhu, L.J., and Walhout, A.J.M. (2013). Integration of metabolic and gene regulatory networks modulates the *C. elegans* dietary response. Cell *153*, 253–266.

Watson, E., MacNeil, L.T., Ritter, A.D., Yilmaz, L.S., Rosebrock, A.P., Caudy, A.A., and Walhout, A.J.M. (2014). Interspecies systems biology uncovers metabolites affecting *C. elegans* gene expression and life history traits. Cell *156*, 759–770.

Watson, E., Yilmaz, L.S., and Walhout, A.J.M. (2015). Understanding metabolic regulation at a systems level: metabolite sensing, mathematical predictions and model organisms. Annu. Rev. Genet. *49*, 553–575.

Yilmaz, L.S., and Walhout, A.J.M. (2014). Worms, bacteria, and micronutrients: an elegant model of our diet. Trends Genet. *30*, 496–503.

Yochem, J., Hall, D.H., Bell, L.R., Hedgecock, E.M., and Herman, R.K. (2005). Isopentenyl-diphosphate isomerase is essential for viability of *Caenorhabditis elegans*. Mol. Genet. Genomics *273*, 158–166.

Zhang, W., Cao, P., Chen, S., Spence, A.M., Zhu, S., Staudacher, E., and Schachter, H. (2003). Synthesis of paucimannose N-glycans by *Caenorhabditis elegans* requires prior actions of UDP-N-acetyl-D-glucosamine:alpha-3-D-mannoside beta1,2-N-acetylglucosaminyltransferase I, alpha3,6-mannosidase II and a specific membrane-bound beta-N-acetylglucosaminidase. Biochem. J. *372*, 53–64.

Cell Systems, Volume *2*

Supplemental Information

A *Caenorhabditis elegans* Genome-Scale Metabolic Network Model

L. Safak Yilmaz and Albertha J.M. Walhout

SUPPLEMENTAL FIGURES

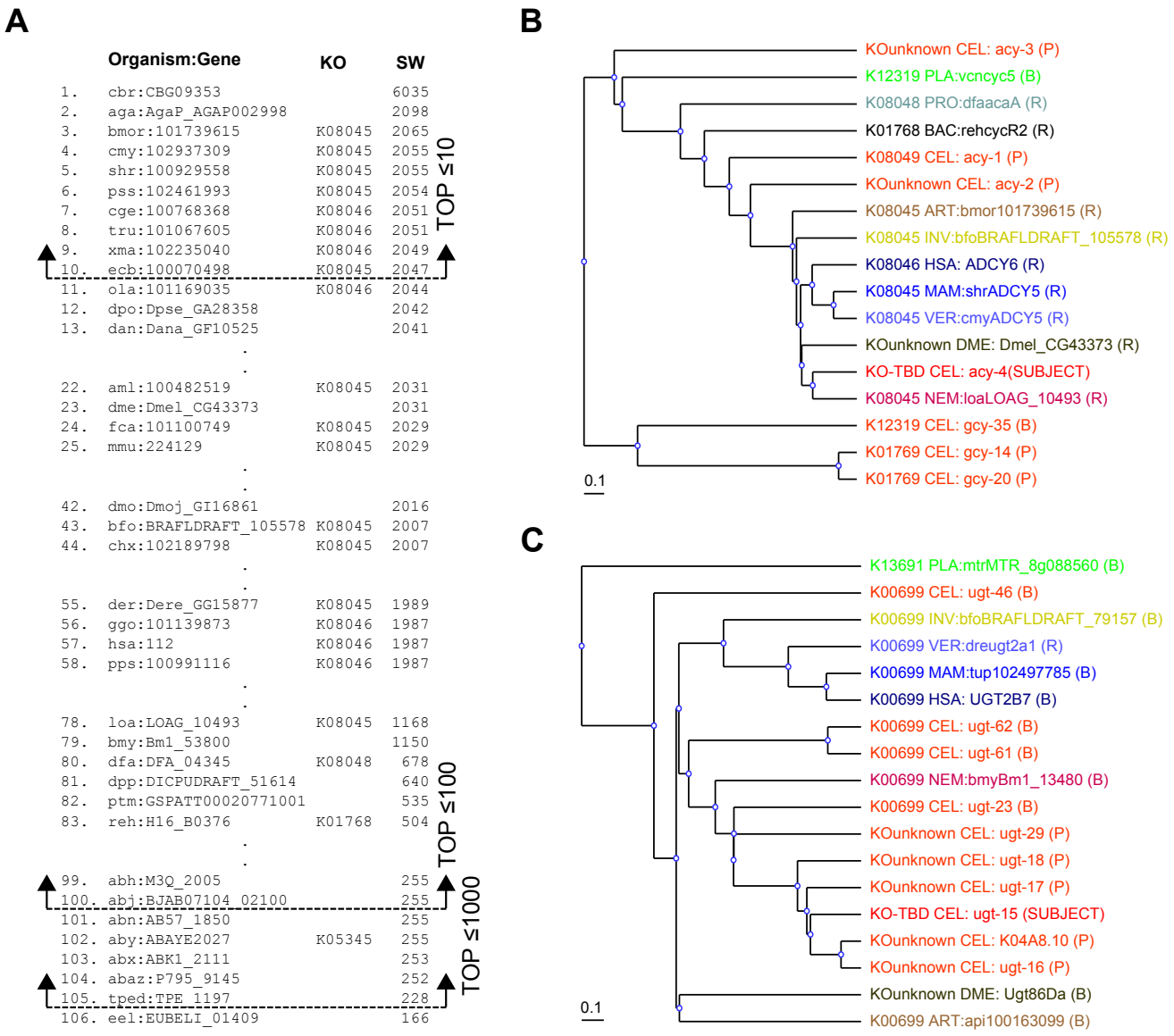


Figure S1, related to Figure 2. **Illustration of myKEGG scoring algorithm and examples of custom phylogenetic trees. (A-B)** Evaluation of *C. elegans* gene *acy-4* with myKEGG **(A)** and myTree **(B)**. The final decision for this gene is an association with K08045 (adenylate cyclase) at low confidence. **(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 S1**) and 0.59 and 0.29 overall (**Equation S2**), 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 in this figure and **Figure 2B**: 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 for sequences from other organisms indicates genes introduced as best matches (B) or reciprocal best hits (R). Parenthetical information for *C. elegans* sequences indicates genes introduced as paralogues (P) or as a reciprocal best match to one of the other organisms in the tree (B). Organism abbreviations are from KEGG.

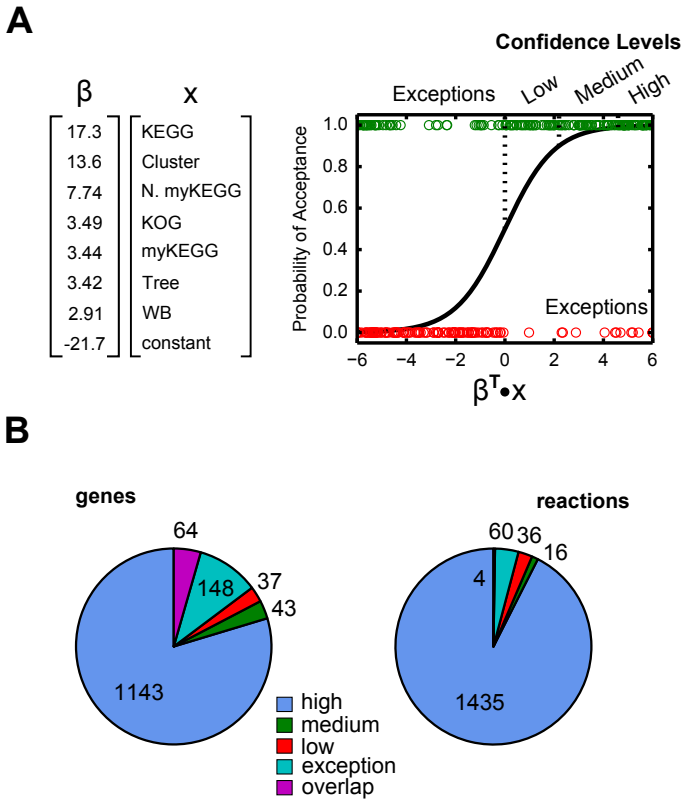


Figure S2, related to Figure 2. **Logistic regression function used in SACURE and the distribution of annotated genes and reactions with regard to confidence levels. (A)** The weights of contributing variables in the trained logistic function (on the left, see **Table S9** for the variables; the constant 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 (see also **Table S1**) 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.

.

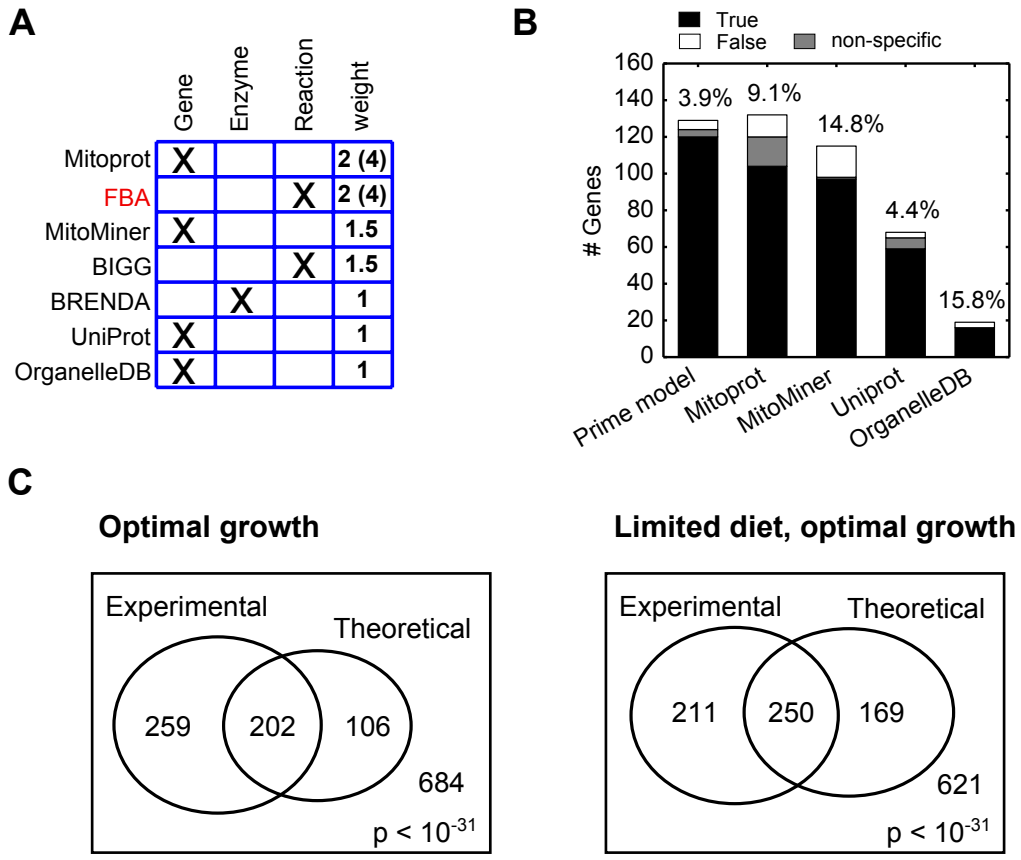


Figure S3, related to Figure 4 and Figure 5. Predictors for reaction localization scoring algorithm of Figures 4A and 4B and reproduction of optimal growth relationships in Figure 5B using non-redundant gene-reaction associations only. (A) Resources used for evaluating reaction localization to mitochondria 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 (**Figure 4A**). 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 (**Table S8**). Prime model column is the same as that in **Figure 4B**. 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). See **Table S6** for details. **(C)** Association of genes with experimental no-growth (lethal, larval lethal, larval arrest, embryonic lethal, embryonic arrest, and sterile) phenotypes with genes predicted by two different approaches also shown in **Figure 5B**. The difference from corresponding Venn diagrams in **Figure 5B** is that, only non-redundant gene-reaction associations are considered, as opposed to all associations. Statistical significance of associations is indicated by hypergeometric *p*-values.

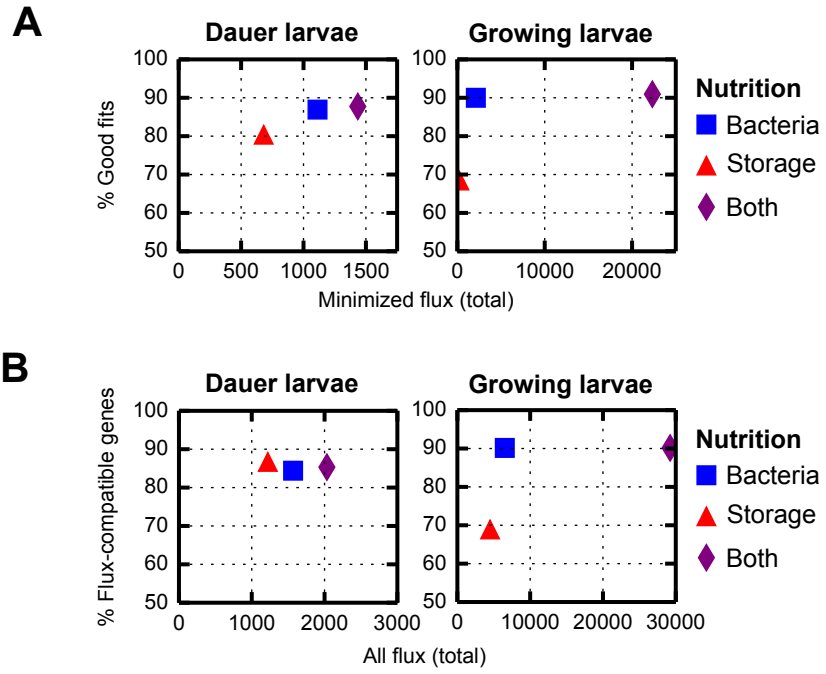


Figure S4, related to Figure 6. Detailed evaluation of fitting quality during the integration of gene expression data. (A) Percentage of reactions (with on or off activity states) that fit the flux distribution is plotted against the sum of minimized flux (absolute flux values of reactions with an undetermined state of activity). **(B)** Percentage of genes that have a regulation state compatible with the flux distribution (upregulated and active or downregulated and inactive) is plotted against the sum of all flux (absolute values) after flux minimization for reactions with an undetermined state of activity.

SUPPLEMENTAL TABLES

Table S9. Related to Figure 2. 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 S10. Related to Figure 5. 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.

Other Supplemental Table Legends

Table S1. Related to Figures 2 and 6. Annotation of metabolic genes.

Table S2. Related to Experimental Procedures. Biomass compositions of *C. elegans* and the bacterial diet.

Table S3. Related to Figures 4 and 6. Reactions in iCEL1273.

Table S4. Related to Figure 4. Compounds in iCEL1273.

Table S5. Related to Figure 4. Enzymes in iCEL1273.

Table S6. Related to Figure 4. Validation of gene localization in PRIME model.

Table S7. Related to Figure 2 and Experimental Procedures. Annotated reactions excluded from iCEL1273.

Table S8. Related to Figure 5. Phenotypic predictions.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

The details of methods followed in each subsection of Results (main text) are presented below with the matching titles.

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 S9**). 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 S9**).

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 S9**).

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 S9**). 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 S9**).

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 S1A** 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 S1**, 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 S1** 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 S2**. In addition, a normalized myKEGG score was calculated, where the highest scoring KO for a given gene got a score of 1.0 (**Table S9**).

$$S_{table}^{KO} = \sum_{i=A,B,C} w_i c_i \frac{\sum_i s^{KO}}{\sum_i s^{all}} \quad \textbf{(S1)}$$

$$S_{final}^{KO} = 0.8 S_{RBH}^{KO} + 0.2 S_{BH}^{KO} \quad \textbf{(S2)}$$

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 S1B**.

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 S9**) (*i.e.*, when all three lowest nodes covering the query gene are dominated by one KO, as in **Figure 2B** in the main text). 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 S9**). See **Figure S1B** 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 S9**) 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 S2A** 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 S1C**]) 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 S9**. In **Figure 2C** of the main text, the weights in **Table S9** were grouped for

resources that yielded two scores (*e.g.*, myTree has two different scores) by multiplying each weight by the maximum possible score in each category.

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 S2B**. All final gene-reaction associations in SACURE are shown in **Table S1**, with regular associations indicated as logistic regression - based, irregular ones as exceptions (the types of exceptional rules are also indicated), and overlap-based ones as intersections (the types of intersections are also indicated). Some of the reactions in **Table S1** and **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 S2A**) 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 S9).

During the reconstruction process, 185 genes were re-annotated (**Table S1**) to complement gene-reaction associations in a network context (see below). Among these, 147 annotations were missed by SACURE (**Table S1**), 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: Pathway-by-Pathway Reconstruction and Gap Filling

Using GPRs from SACURE, a template metabolic network of *C. elegans* was reconstructed in a pathway-by-pathway manner, following pathway definitions in KEGG, MetaCyc, and the literature. Pathways for each reaction are indicated in **Table S3**, together with references in comments section when the pathway was not directly adopted from KEGG. Although obvious pathway gaps were generally detected and fixed during manual reconstruction (see main text for different types of gap-filling approaches), flux balance analysis (FBA) was needed at advanced stages of the network to identify additional gaps. Presence of network gaps is evident from the lack of flux carrying capacity of a subset of reactions. To determine if a reaction can carry any flux in the forward direction, FBA was performed with the objective function that maximizes the flux in the reaction. For reversible reactions, flux carrying in the reverse direction was additionally tested by minimizing the flux as the objective, since the flux values are negative in the reverse direction. If the maximum flux for an irreversible reaction is zero or both maximum and minimum fluxes for a reversible reaction are zero, then a network gap prevents flux flow. In these tests, to make sure flux carrying capacity was not limited by the diet, all possible nutrients were supplied to the model by setting the lower boundary of all exchange and sink reactions to a negative value.

For each zero-flux reaction detected, potential gap-filling reactions were identified by a careful analysis of the network and flux flow. When the gap could not be found manually, shadow prices were calculated for the zero-flux reaction as a useful aid in finding the potential gaps. A shadow price is the change in objective function when a metabolite deviates from steady state hence being produced or consumed (Reznik et al., 2013). Thus, it identifies metabolites that limit the objective function (*i.e.*, the flux in the reaction examined). Potential rescue reactions are those that produce or degrade these metabolites. If these reactions were not available in the GPR set from SACURE, reaction annotations were revisited. Specifically, lenient annotations, literature searches, and uncharacterized enzymes were tested (see main text). When these approaches did not annotate any rescue reactions, the zero-flux reaction was rescued by the transport of a reactant or a product. Reactions that could not be rescued by a single transport reaction

were considered as disconnected and left out of the network until the last step of reconstruction where nearly exhaustive gap-filling was performed (see section 5 below).

Since the tracking of all gap-filler reactions is difficult during manual reconstruction (*e.g.*, when many reactions are inserted as a whole pathway from the literature, which ones are filling gaps and which ones are forming alternative pathways are not clear), gaps in **Figure 3** were retrospectively defined once iCEL1273 was fully reconstructed. To determine if a query reaction in the model is a gap-filler, first the flux of this reaction was constrained to zero, and then the flux carrying capacity of all of the rest of the reactions was calculated. If a subset of the rest of the reactions lost flux carrying capacity under this constraint, then the query reaction is a gap-filler, unless it is annotated by KEGG alone or it is an automatically incorporated BiGG transport reaction (see below in section 3). The type of annotation carried out to identify the gap-filler reaction determines the category to which the reaction belongs, according to the categories in **Figure 3**. All these gap-fillers are also indicated in **Table S3**.

3. 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 **Table S3** 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 EXC0050) 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, which are indicated in **Table S2**. 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 in **Table S3** 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, which are indicated in **Table S2**. 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 (**Table S4**). 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 the comments section of **Table S3**.

For a subset of metabolites, 99 transport reactions were added but not automatically incorporated from BiGG (**Table S3**, Transport-type reactions without the indication of a BiGG transport in comments). 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 (*e.g.* gluconic acid transport, **Figure 3E**). The basis for each of these transport reactions is indicated in the comments column of **Table S3**.

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 in **Table S3**.

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 in **Table S3**.

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 the comments column of **Table S3**.

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 **Table S4**, comments section. Exceptional cases in stoichiometric decisions were rare and are indicated in **Table S3**, comments section.

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

Reactions were divided into three compartments: mitochondria, cytosol, and extra-cellular 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 in **Table S3** (reactions with header R), were systematically determined based on seven resources and FBA (**Figure S3A**).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 S3A**). 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 S3A**) 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 S3A**) 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 S3A**) 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 S3A**) 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 S3A**) 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 S3A**), 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, **Table S3**) 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 S3A**). 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 S3A**). 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 (**Figure 4A**). All exceptions for protein localization are explained in the comments column of **Table S3**.

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 S3A**). 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 the comments column of **Table S3**. With the reaction and protein re-localizations, the reconstruction of the prime model was completed (**Figure 1B**, **Table S3**).

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 (PhosphoL, **Table S4**) 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 4B** in the main text. This result is repeated here in comparison with the performance of our gene-based predictors (**Figure S3B**). 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.

5. 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, **Table S3**). Since ATP and CTP are interconvertible (RC00570, **Table S3**), 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. All SACURE-annotated, excluded reactions are listed in **Table S7**. The reasons for exclusion are indicated in the comments column. Reactions eliminated at steps 3-4 above are indicated as disconnected.

6. Validation of iCEL1273: Reproducing Observed Mass and Energy Balance during Two Stages of Life

To test if iCEL1273 can reproduce observed production/consumption rates related to growth (biomass production as body mass or progeny, bacterial intake, and respiration rates), we obtained relatively accurate measurements of these variables at different stages of life from the literature. Since predictions with FBA assume a steady state condition, we looked for rates measured in short time intervals where animals can be assumed to be in a quasi-steady state condition as a first approximation. Based on multiple studies, we were able to determine approximate rates for two stages: the L4 larval stage during post-embryonic development and the young adult stage where egg laying begins (**Table S10**). Since these rates were not reported in units

usable in FBA, we first carried out unit conversions. Specifically, we converted the amount of bacteria consumed and biomass produced from numbers (of bacteria and eggs) or length (worm body) to grams, and converted different O_2 consumption and CO_2 release rates to mmoles of gas per hour per gram dry weight of worm biomass. Also in respiration studies, we found either O_2 consumption or CO_2 production rates, but not both, so one had to be derived from the other. To perform all these calculations, we used the following assumptions:

i. Wet weight of a worm is related to its length by the relationship in (Ferris et al., 1995).
ii. Wet/dry weight ratio is 3.4 (Neidhardt et al., 1990) (based on bacteria, assumed to be valid for worms also).
iii. Wet weight/ protein weight ratio in worms is 5 based on (Van Voorhies, 2002).
iv. Respiration quotient in *C. elegans* (CO_2 released / O_2 consumed) varied between 0.7 and 1, according to (Van Voorhies, 2002).
v. Wet weight of a worm egg is 0.035 μg based on (Byerly et al., 1976).
vi. Wet weight of a bacterium (*E. coli*) is 0.95 pg based on (Neidhardt et al., 1990).

Using the length-mass relationship mentioned in (i), a length-based growth curve provided in (Hirsh et al., 1976) was translated into a weight curve, and the biomass production rate was calculated for the L4 stage using **Equation S3**, wherein, μ is the growth rate, *W* is weight and Δt is the length of L4 stage which is ~12 hours. This equation is a good approximation for the average production rate as the growth curve is nearly linear during the L4 stage. Average respiration rates were obtained for the L4 stage from (Vanfleteren and DeVreese, 1996) using assumptions (ii), (iii), and (iv). For the young adult stage, both egg laying rates and respiration rates were obtained from the same study (Van Voorhies and Ward, 1999). The reported egg laying rate was ~3.75 eggs per worm per hour, which resulted in 0.065 gram eggs per gram worm per hour when converted using (v) and assuming a wet weight of 2 μg for a young adult based on the length data in (Hirsh et al., 1976) and the assumption (i). Hourly respiration rates were calculated using assumptions (ii) and (iv). As for bacterial consumption rates, we could not find carefully measured data, except that egg-laying adults are known to consume several millions of bacteria per day (Hirsh et al., 1976; McGhee, 2007) and the average consumption per day during the entire life span of *C. elegans* is in the same order of magnitude (Ferris et al., 1997). We therefore assumed that an adult animal consumes 1 to 10 million bacteria per day and converted this range to 0.02 to 0.2 g bacteria / g worm / hour based on (vi) and again assuming a wet weight of 2 μg per worm. Since the biomass production rate was even larger in the L4 stage (**Table S10**, note that this value is normalized by the worm weight and actually L4 worms produce less biomass per day on an absolute scale), we believe this is a rough but reasonable range that covers both stages. When converted back to numbers, this range corresponds to 125 thousand – 1.25 million bacteria per day for L4 worms as their body mass is much smaller than adults (~0.25 μg at middle L4 stage as compared to ~2 μ g in adult stage).

$$\mu = \frac{1}{\Delta t}\frac{W_{endL4} - W_{beginL4}}{W_{midL4}} \qquad \text{(S3)}$$

The rates in **Table S10** were used to test the performance of iCEL1273 when constrained by observations. Each rate in **Table S10** can be predicted by one or more reactions in the model. These reactions are EXC0050 for bacterial intake, EX00007 for oxygen exchange, EX00011 (CO_2 exchange) and EX00288 (bicarbonate exchange) for CO_2 release (the sum of flux in these reactions predicts total release), BIO0100 for biomass production in L4 stage, and BIO0102 for biomass production during egg-laying adult stage (see section 3 above for the specificity of biomass reactions; all reactions can be found in **Table S3**). Four flux variability analysis tests were performed for each stage of life. In these tests, reactions related to one of the four rates were set as the objective function, all the other reactions were constrained with observed rates (in the case of CO_2 release rate, the sum of EX00011 and EX00288 was constrained), and two separate FBA solutions were obtained for maximizing and minimizing the objective function. The minimum and maximum values obtained determined the predicted range for the rate tested. These tests could result in a failure in two ways. First, no solutions may be obtained, hence indicating that the model cannot simultaneously meet the requirements in the three constrained rates, irrespective of the rate that is

being predicted. Second, the model may predict a range that is outside the experimental range of the tested rate. In the case of a success, the model predicts a range that overlaps with the experimental range of the tested rate. As shown in Figure 5A, the model passed all tests by predicting a range consistent with the experimental observation.

To derive feasible maintenance and digestion costs for *C. elegans*, the model was constrained by an achievable range of values for non-growth associated maintenance (NGAM), growth associated maintenance (GAM), and digestion costs. NGAM cost was imposed by constraining the flux of reaction RCC0005 (**Table S3**) to the selected value. To impose GAM and digestion costs, the coefficients of ATP in reactions BIO0010 and DGR0007 were set according to the selected values, respectively. Then, for each stage of life, FBA was performed with the objective of minimizing bacterial intake while constraining the model by the three other observed rates (**Table S10**). In the case of a success, the minimum bacterial intake in the solution is less than the upper limit (0.2 g bacteria / g worm / hour in both stages, **Table S10**), which means the model can satisfy all experimental criteria with the imposed energetic costs. In the case of a failure, the solution is greater than the upper limit for bacterial intake and the model cannot meet the energetic requirements with the selected costs. We varied the selected costs exhaustively by gridding the 3D space for NGAM (at steps of 0.1 mmoles ATP / g dW / h), GAM (steps of 1 unit), and digestion costs (steps of 1 unit). Figure 5B shows all values that resulted in a success, and therefore, each data point shows a combination of energetic costs that can be met by iCEL1273. The center of mass of this gridded space was arbitrarily selected as the final values for these costs.

7. Validation of iCEL1273: Gene Essentiality and Genotype-Phenotype Relationships

To test the ability of iCEL1273 to predict specific phenotypes that may be related to metabolic functions, we downloaded all observed phenotypes from WormBase, reported in the Phenotypes section. For each gene, this data includes observations from experiments in which the expression of the gene is perturbed either by mutation or by RNAi on wild type animals. Observations with the rare case of mutation experiments on a genetic background (*i.e.*, with multiple mutations) were removed from the dataset to reduce complexity.

Experimentally determined essential genes

Genes associated with the following phenotypes were considered essential: lethal, larval lethal, L1 lethal, early larval lethal, larval arrest, L1 arrest, early larval arrest, embryonic lethal, embryonic arrest, and sterile.

Predicted essential genes

Four methods were used to predict essential genes, which varied based on the definition of essentiality and diet, as explained in the following.

<u>*Non-redundant associations with regular diet*</u>: Perturbation of a gene was represented in the model by constraining the reactions that are non-redundantly associated with the gene to a flux value of zero. A non-redundant association occurs in two types of GPR. In the first type, the query gene is the only gene in the GPR, and in the second, the query gene is connected to others with an AND logic, as a subunit of a protein, as a component of a protein complex, or as an enzyme in a merged reaction involving multiple enzymes. The diet was defined as bacteria, arbitrarily provided at a rate of -0.32 g/g dW/h in the EXC0050 reaction (the negative sign indicates uptake by convention). Most other exchange reactions were constrained to positive values to allow secretion while inhibiting uptake of other nutrients. Exceptions included oxygen, water, phosphate, and cholesterol, for which the lower limit of exchange was set at a negative value to allow uptake. This uptake was limited to 0.01 mmoles/g dW/hour for cholesterol, which was required in trace amounts for a small set of reactions, and was made practically limitless for others to avoid limiting the model with these nutrients (up to 1000 mmoles/g dW/hour). The objective in FBA was defined as the maximization of biomass production by either reaction BIO0100 or BIO0103 which together account for all biomass components (see section 3 above). The maximum biomass production with each reaction was then calculated to find the growth rate upon perturbation (B_p). This value was compared to growth rate obtained when no constraints were applied (B_o, represents wild type). Reduction in growth rate due to perturbation is given by $(B_o\text{-}B_p)/B_o$. Predicted essentiality was defined as the case when the perturbation of

the query gene caused a value of 0.5 or greater (*i.e.*, >50% reduction in growth rate) in at least one of the objective functions (i.e., BIO0100 or BIO0103). Other thresholds were tested and the strength of associations between predicted and experimentally determined essential genes (based on hypergeometric *p*-value) did not change in the range 0.3-0.84.

All associations with regular diet: This method is the same as the previous one, except that gene perturbation was represented by the elimination (constraining to zero flux) of all reactions associated with the query gene, irrespective of the type of association. Thus, if the query gene is in a set of paralogs in the GPR of a reaction (*i.e.*, connected with an OR logic), the reaction is eliminated in this method unlike the previous method, which assumed that paralogs are redundant in function.

Optimal growth with regular diet: As different from previous methods, no gene perturbations were simulated with this method. First FBA was used to calculate the maximum growth rate with bacterial diet as described above (identical to B_o above). Importantly, this solution is not unique in flux distribution, *i.e.*, there is a large solution space with alternative flux distributions using alternative pathways to yield the same growth rate. We defined optimal growth as the solution with minimum total flux, when the maximum growth rate was not altered. To find this solution, we performed a second FBA by modifying the model such that each reversible reaction was divided into two reactions, one defining the forward direction and the other the reverse direction. Then all reactions were constrained to a positive flux, and the biomass production rate (flux of BIO0010) was constrained to *Bo*. FBA was performed with the objective of minimizing the sum of flux in all reactions (*i.e.*, the objective function covered all reactions). The calculated flux distribution was then mapped from the modified model back to the original model, such that, for each reaction that is reversible in the original model, the flux in the reverse direction in the modified model was subtracted from the flux in the forward direction to obtain the corresponding flux in the original model. Irreversible reactions were mapped without modification. The flux distribution obtained with this procedure determined the predicted optimal growth state. Essential genes were defined as genes that are active during optimal growth, *i.e.*, associated with reactions that carry flux in this state. For **Figure 5B**, all types of GPR associations were considered, as in the previous method. As a comparison, **Figure S3C** presents the case when only non-redundant associations are considered.

Optimal growth with modified diet: This method is the same as the previous one except that the diet was modified. Bacterial intake rate was set to zero. Instead the model was provided with all 20 amino acids, trehalose and triacylglycerides (TAG). For amino acid uptake, the corresponding exchange reactions were constrained to a minimum flux of -0.01 mmoles/g dW/h (*i.e.*, this amount of each amino acid was made available to the model). For trehalose and triacylglycerides, the corresponding sink reactions (SNK0013 and SNK0014, respectively) were constrained to -1.0 and -0.23 mmoles/g dW/h, respectively. These numbers were arranged to allow the generation of same amounts of energy as the bacterial diet above (calculated with FBA by setting the flux of RCC0005, ATP consumption reaction, as the objective to be maximized). In addition, the soluble component of the bacterial diet was provided with a maximum uptake rate of 0.01 mmoles/g dW/h, as this portion of the bacterial diet contains essential nutrients (heme iron, molybdenum, etc.) for biomass production. The results of this method were also presented for all associations (**Figure 5B**) and non-redundant associations (**Figure S3C**).

Phenotypes and gene essentiality results are listed in **Table S8** for every gene in iCEL1273. Predicted essentiality results from the methods described above are indicated in the corresponding columns with descriptive titles.

Metabolic products of genes

To provide a potential explanation for false negatives from the above four methods (unpredicted essential genes) and also to clearly identify the role of genes predicted to be essential, we determined the relationship between genes and a set of key products. These products were metabolites drained by demand and sink reactions (vitamins, cofactors, modified proteins, signaling compounds, etc.; see section 3), biomass precursors (DNA, RNA, phospholipids, etc.), and energy (ATP). In essence, we evaluated how the production capacity of the model changed upon the perturbation of each gene for this set of products. The approach is the same as the method titled “all associations with regular diet” in the previous subsection, except that, instead of maximizing the biomass production, this method maximized the production of a

metabolite or ATP generation. If the tested product was a metabolite consumable by a demand/sink reaction, the flux of that reaction was maximized as the objective. If the metabolite was a biomass precursor, an artificial demand reaction was temporarily created and the same method was applied using this reaction. For energy generation, the objective was the maximization of the flux in reaction RCC0005, which converts synthesized ATP back to ADP. For each query gene, this method was repeated for all products one by one. If the maximum production rate decreased by more than 50% upon the perturbation of the query gene, this gene was considered as essential for the tested product. The metabolites found by this method for each gene are listed in the theoretical products column of **Table S8**.

Other phenotypic predictions

In addition to growth-related phenotypes, we searched other metabolism-related phenotypes reported in WormBase for iCEL1273 genes. We found 10 well-defined, relatively frequent phenotypes (associated with ≥ 5 genes) for which we had a related product in the model that could allow a phenotypic prediction. We added to this list the slow growth phenotype to represent non-lethal but growth-related phenotypes that are frequently encountered. The expected relationships were as follows (*phenotypes* – metabolic products that are related): *Slow growth/extended life span* – ATP, *cell cycle slow early emb/cell division variant* – DNA, *lipid depleted/lipid composition variant/fat content reduced* – TAG/phospholipids, *dumpy/organism morphology variant* – collagen, *coenzyme Q depleted* – ubiquinone, *histone methylation variant* – methylated histone. We assigned genes to each one of the products using the perturbation-response method described in the previous subsection, except that the threshold for the decrease in production rate that was used to associate genes with products was not taken only at 50%, but eight different thresholds from 0.1% to 99.9% were applied. Then we compared the predicted gene-product relationships with experimental gene-phenotype associations, as was done with gene essentiality (**Figure 5C**). A hypergeometric *p*-value < 0.001 was considered as a statistically validated association. As can be seen in **Figure 5D**, where the lowest *p*-values obtained from different thresholds are shown, all expected relationships were significant. As a side note, only two of these relationships became insignificant when a single threshold of 50% was used in this analysis (*lipid depleted* - phospholipids, *organism morphology variant* - collagen). The most significant associations obtained in this analysis are presented in **Table S8**, in the column named selected phenotype associations.

8. Validation of iCEL1273: Gene Essentiality and Genotype-Phenotype Relationships in Methionine Salvage Pathway

The method used in this section was described in the previous section. See subsection Predicted essential genes (FBA methods for all associations with regular diet).

9. Case Study: Analysis of Dauer Metabolism Using Gene Expression Data

Microarray-based gene expression data was obtained from a study that compares dauer and growing larvae (Wang and Kim, 2003). The analysis of metabolic states based on this dataset was carried out in two steps. First, gene expression was translated into reaction activity based on GPRs, and a best-fit flux distribution to this activity map was derived. This step gives the maximum number of active and inactive reactions that can be fitted by a predicted flux distribution. Second, while keeping the number of flux-fitted reactions constant, an optimal flux distribution with minimum total flux in other reactions (*i.e.*, with undetermined activity) was obtained. In the following, the first two subsections describe these two methods. In the third subsection, we explain the final optimization procedure that applies this analysis to different conditions of nutrient availability for dauer and growing larvae to define the final metabolic state for each stage of life.

Integration of gene expression data using mixed integer linear programming (MILP)

To identify a flux distribution that best fits the gene expression data, first, we determined genes that were clearly upregulated and downregulated based on a *p*-value cutoff at 0.001. Then, these expression states were integrated with the network model using the mixed integer linear programming (MILP) approach described in (Shlomi et al., 2008). This method requires a set of reactions that ideally carry flux in the optimized solution and another set of reactions that do not (referred to as highly expressed and lowly

expressed reactions in the original publication, respectively). We refer to these sets as reactions with on (R_{ON}) and off (R_{OFF}) states respectively. The objective is to find a flux distribution that maximizes the total number of R_{OFF} reactions with zero flux and R_{ON} reactions with a flux exceeding a threshold. Gene expression was translated into reaction status using the logical operators in GPRs (Shlomi et al., 2008). For instance, if an upregulated gene is associated with a reaction as a single gene or as connected to other genes with an OR logic (*i.e.*, if it is one of the paralogs), then the reaction is assigned to the on state; if it is connected to other genes with an AND logic (*e.g.*, if it is a protein subunit), whether the reaction is on or off depends on the other genes. If a downregulated gene is related to a reaction as a single gene or as connected to the others with an AND logic, then the reaction is assigned to the off state. Reactions that were not categorized as on or off have an undetermined status since transcriptional regulation is not sufficient to draw a conclusion. The activity of these floating reactions is to be predicted by the fitting method.

In terms of MILP, the goal of maximizing the agreement between reaction status and flux distribution is represented by the maximization objective in **Equation S4**, which is fulfilled while satisfying **Equations S5-S9** simultaneously. In **Equation S4** *y*-values are integer variables used in MILP that take values of 0 or 1. Equations **S5** and **S6** are the main equations of constraint-based FBA (see main text, Experimental Procedures), where *S* is the stoichiometry matrix, *v* denotes the vector of reaction fluxes, and $v_{\min}$ and $v_{\max}$ indicate the lower and upper bounds of these fluxes. **Equation S7** represents the formulation that encourages zero flux in reactions in the off category. Since a *y*-value of 1 constrains the flux of a reaction to 0 in this expression (both upper and lower bounds become zero), making a lowly expressed reaction carry zero-flux is equal to adding 1 to the objective in **Equation S4**, which is to be maximized. If this cannot be made (*i.e.*, the reaction has to carry flux in the best-fit condition when all other constraints are imposed), then the *y*-value is set at 0 (hence not contributing to the objective), which converts **Equation S7** into the original flux constraint from **Equation S6**. Similarly, **Equation S8** represents encouraging a positive flux in the forward direction in reactions of the on category. A *y*-value of 1 is equivalent to having a flux equal to or greater than a minimum value set beforehand (ε). For reversible reactions, **Equation S9** is also implemented, this time rewarding a negative flux less than or equal to $-\varepsilon$. We arbitrarily used ε=0.1 in this study.

$$\max\left(\sum_{i\in R_{ON}}(y_i^f + y_i^r) + \sum_{i\in R_{OFF}} y_i\right); y_i^f, y_i^r, y_i \in \{0,1\} \qquad \text{S4}$$

s.t.

$$S \cdot v = 0 \qquad \text{S5}$$

$$v_{\min} < v < v_{\max} \qquad \text{S6}$$

$$v_{\min,i}(1 - y_i) \le v_i \le v_{\max,i}(1 - y_i); i \in R_{OFF} \qquad \text{S7}$$

$$v_i \ge v_{\min,i} - y_i^f (v_{\min,i} - \varepsilon); i \in R_{ON} \qquad \text{S8}$$

$$v_i \le v_{\max,i} - y_i^r (v_{\max,i} + \varepsilon); i \in R_{ON} \qquad \text{S9}$$

Flux minimization

The solution from the previous subsection provides the maximized objective, *i.e.*, the maximum number of reactions with fluxes in agreement with the assigned on or off status (hereafter referred to as fitted reactions). The solution also includes a flux distribution, which is not unique, as different distributions can

satisfy the same objective. We further optimized this solution by a flux minimization approach. This method was the same as flux minimization described in section 7 above (Predicted essential genes; optimal growth with regular diet) with three exceptions. First, the constraint that was held constant during minimization was not the growth rate but was the number of fitted reactions (*i.e.*, the value in **Equation S4**). Second, **Equations S7-S9** were also implemented during FBA to meet this constraint (**Equations S5** and **S6** are implemented by default). Third, instead of minimizing flux from all reactions, we minimized only reactions with undetermined status (not in on or off category), which are unrestrained during flux fitting and make a large portion of the network (>83% of all reactions in the network in all cases). The technical reason for not including reactions that belong to on or off category in flux minimization was because the fluxes in these reactions are controlled by the integration method (**Equations S7-S9**), which has to meet the constraint on the number of fitted reactions, at the same time as flux minimization.

Optimization for different conditions

We applied the integration and flux minimization method above to two different stages of life (dauer and growing larvae) using three different sources of nutrients (bacteria, storage compounds, and both). A total of 395 genes in iCEL1273 were regulated at the selected *p*-value cutoff (<0.001), with 241 downregulated and 144 upregulated in the dauer stage (numbers are reversed for the growing larvae stage). In each stage, reactions were assigned to on or off states accordingly as described above. A drawback of our approach is that, reactions in the off category (reactions associated with downregulated genes) are forced to have no flux. However, downregulation does not necessarily mean switching to an off mode; it may be only indicative of a decrease in flux. Reactions associated with the electron transport chain (ETC) and ATP synthase are probably of this kind due to their important role in energy generation, and were therefore excluded from the fitting, although many genes associated with these reactions were downregulated in the dauer stage. Indeed, the relative value of fluxes in predicted distributions for dauer and growing larvae were consistent with a reduction in flux in the dauer stage (see below; **Figure 6C**, O_2 usage and ATP production), thereby verifying our assumption in retrospect. The final numbers of reactions in the on and off categories were 231 and 136 for the dauer, and 230 and 90 for the growth stage, respectively.

The nutritional conditions were established using the exchange reaction for bacteria (EXC0050) and sink reactions for glycogen (SNK0012) and TAG (SNK0014). The exchange of oxygen, water, phosphorus, and cholesterol were also allowed. The rates of all these exchange reactions were lower bound by -1000 units of flux in order not to limit the model by any nutrients, except for cholesterol, which was given in small amounts (lower bound by -1 unit) so that its degradation could not compete with other nutrients as a source of energy or carbon, while supporting flux in pathways depending on cholesterol.

The results of data integration and flux minimization for each of the six tests (two stages X three nutrient conditions) are shown in **Figure S4A**. The quality of fitting was initially based on a high number of fitted reactions and a low sum of minimized fluxes. For both stages, both the percentage of fitted reactions and minimized flux sum increased as the nutritional supply became richer starting from only storage and changing to bacteria and then to both storage and bacteria. This is expected as more nutrients can support flux in more reactions. However, storage compounds provided a competitive fit at the lowest sum of minimized flux, and bacterial diet for the growth stage had a good balance between the two quality parameters. In addition, the number of fitted reactions does not necessarily reflect the agreement between gene expression and flux distribution, as the relationship between genes and reactions is not one on one. We therefore calculated the number of genes that have expression levels compatible with the fluxes. This compatibility required an upregulated gene to be active (associated with reactions carrying flux) and a downregulated gene to be inactive (only associated with zero flux reactions) in the flux distribution (ETC and ATP synthetase genes were excluded as mentioned above). Given the GPR associations, 218 genes in the dauer stage and 193 genes in the growth stage dictated reaction status as on or off and therefore could be related to reaction fluxes. Among these sets of genes, the percentages of flux-compatible genes are shown for all tests in **Figure 6B**. From these plots we reached a clear conclusion about optimal nutrient conditions for each stage (storage compounds for dauer and bacteria for growing larvae). Using the sum of flux from all reactions instead of the sum of minimized flux in non-regulated reactions did not change this conclusion (**Figure S4B**).

Once the optimal points were defined for each stage, key variables plotted in **Figure 6C** were derived from these flux distributions. Biomass production, ATP synthesis, and O_2 uptake rates correspond to the absolute flux values from reactions BIO0010, RMC0004, and EX00007, respectively. Importantly, the flux in BIO0010 for growing larvae came from BIO0101, which is the biomass synthesis reaction for body mass. As another side note, using the sum of minimized reactions as in **Figure 6C** or the sum of all fluxes (**Figure S4B**) did not change the conclusions about overall metabolic activity in the two stages.

10. Compartmentalization of Dauer Metabolism

FBA was carried out using two different sets of nutritional input and objective function in two compartments. In the microaerobic compartment, TAG, glycogen, and oxygen were provided to the model by corresponding exchange or sink reactions as described in the previous section, and the objective was set to maximize the trehalose output via exchange reaction EX01083. TAG and glycogen consumptions were arbitrarily limited to 1 unit of flux each. Oxygen uptake was set at a maximum value of 15 units, which made sure this nutrient was limiting (**Figure 7B**). In the anaerobic compartment, trehalose uptake was set at a maximum of 1 flux unit as the only energy source and energy generation was maximized using the reaction that consumes ATP (RCC0005) as the objective function.

Sensitivity analyses in **Figures 7B** and **7C** were carried out by constraining the flux of the variable in the *x* axis and maximizing the variable in the *y* axis. The grey curve in **Figure 7B** was derived by constraining the flux in glyoxylate cycle reactions to zero.

SUPPLEMENTAL 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.