A protocol for generating a high-quality genome-scale metabolic reconstruction

Ines Thiele[1,2] & Bernhard Ø Palsson[1]

[1]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. [2]Current address: Center for Systems Biology, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavik, Iceland. Correspondence should be addressed to B.Ø.P. (palsson@ucsd.edu).

Published online 7 January 2010; doi:10.1038/nprot.2009.203

Network reconstructions are a common denominator in systems biology. Bottom–up metabolic network reconstructions have been developed over the last 10 years. These reconstructions represent structured knowledge bases that abstract pertinent information on the biochemical transformations taking place within specific target organisms. The conversion of a reconstruction into a mathematical format facilitates a myriad of computational biological studies, including evaluation of network content, hypothesis testing and generation, analysis of phenotypic characteristics and metabolic engineering. To date, genome-scale metabolic reconstructions for more than 30 organisms have been published and this number is expected to increase rapidly. However, these reconstructions differ in quality and coverage that may minimize their predictive potential and use as knowledge bases. Here we present a comprehensive protocol describing each step necessary to build a high-quality genome-scale metabolic reconstruction, as well as the common trials and tribulations. Therefore, this protocol provides a helpful manual for all stages of the reconstruction process.

INTRODUCTION

Metabolic network reconstruction has become an indispensable tool for studying the systems biology of metabolism[1–7]. The number of organisms for which metabolic reconstructions have been created is increasing at a pace similar to whole genome sequencing. However, the quality of metabolic reconstructions differs considerably, which is partially caused by varying amounts of available data for the target organisms and also by a missing standard operating procedure that describes the reconstruction process in detail. This protocol details a procedure by which a quality-controlled quality-assured reconstruction can be built to ensure high quality and comparability between reconstructions. In particular, the protocol points out data that are necessary for the reconstruction process and that should accompany reconstructions. Moreover, standard tests are presented, which are necessary to verify functionality and applicability of reconstruction-derived metabolic models. Finally, this protocol presents strategies to debug non- or malfunctioning models. Although the reconstruction process has been reviewed conceptually by numerous groups[8–11] and a good general overview of the necessary data and steps is available, no detailed description of the reconstruction, debugging and iterative validation process has been published. This protocol seeks to make this process explicit and generally available.

The presented protocol describes the procedure necessary to reconstruct metabolic networks intended to be used for computational modeling, including the constraint-based reconstruction and analysis (COBRA) approach[11,12] (see **Box 1** for definition). These network reconstructions, and *in silico* models, are created in a bottom–up manner based on genomic and bibliomic data and thus represent a biochemical, genetic and genomic (BiGG) knowledge base for the target organism[9]. These BiGG reconstructions can be converted into mathematical models and their systems and physiological properties can be determined. For example, they can be used to simulate the maximal growth of a cell in a given environmental condition using flux-balance analysis (FBA)[13,14]. In contrast, the generation of networks derived from top-down approaches (high-throughput data-based interference of component interactions) is not discussed here, as they do not generally result in functional, mathematical models.

The metabolic reconstruction process described herein is usually very labor and time intensive, spanning from 6 months for well-studied, medium-sized bacterial genomes, to 2 years (and six people) for the metabolic reconstruction of human metabolism[15]. Often, the reconstruction process is iterative, as demonstrated by the metabolic network of *Escherichia coli*, whose reconstruction has been expanded and refined over the last 19 years[7]. As the number of reconstructed organisms increases, the need to find automated, or at least semi-automated, ways to reconstruct metabolic networks straight from the genome annotation is growing. Despite the growing experience and knowledge, to date, we are still not able to completely automatically reconstruct high-quality metabolic networks that can be used as predictive models. Recent reviews highlight current problems with genome annotations and databases, which make automated reconstructions challenging and thus they require manual evaluation[8,9]. Organism-specific features, such as substrate and cofactor utilization of enzymes, intracellular pH and reaction directionality remain problematic and thus require manual evaluation. However, some organism-specific databases and approaches exist, which can be used for automation. We describe here the manual reconstruction process in detail.

A limited number of software tools and packages are available (freely and commercially), which aim at assisting and facilitating the reconstruction process (**Table 1**). This protocol can, in principle, be combined with those reconstruction tools. For generality, we present the entire procedure using a spreadsheet, namely Excel workbook (Microsoft), and a numeric computation and visualization software package, namely Matlab (Mathwork, Natwick, MA, USA). Free spreadsheets (e.g., OpenOffice and Google Docs) could be used instead of the listed spreadsheet. Alternatively, MySQL databases may be used, as they are very helpful in structuring and tracking data. Matlab was also used to encode the COBRA Toolbox,

TABLE 1 | Data sources frequently used for metabolic reconstructions.

Name	Link	Comment
Genome databases		
Comprehensive Microbial Resource (CMR)	http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi	
Genomes OnLine Database (GOLD)	http://www.genomesonline.org/	
TIGR	http://www.tigr.org/db.shtml	
NCBI Entrez Gene	http://www.ncbi.nlm.nih.gov/sites/entrez	
SEED database[32]	http://theseed.uchicago.edu/FIG/index.cgi	Comparative genomics tool
Biochemical databases		
KEGG[41]	http://www.genome.jp/kegg/	
BRENDA[42]	http://www.brenda-enzymes.info/	
Transport DB[89]	http://www.membranetransport.org/	
PubChem[86]	http://pubchem.ncbi.nlm.nih.gov/	
Transport Classification Database (TCDB)	http://www.tcdb.org/	TCDB is a curated database of factual information from over 10,000 published references
pK_a Plugin	http://www.chemaxon.com/product/pka.html	Free for academic users
pK_a DB	http://www.acdlabs.com/products/phys_chem_lab/pka/	Commercial software package to determine acid–base ionization/dissociation constant, pK_a
Organism-specific databases		
Ecocyc[43]	http://ecocyc.org/	*Escherichia coli* database
PyloriGene[37]	http://genolist.pasteur.fr/PyloriGene	*Helicobacter pylori* database
Gene Cards	http://www.genecards.org/	Human gene database
Protein localization databases		
PSORT[47]	http://www.psort.org/psortb/	Support vector machine (SVM) based
PA-SUB[48]	http://www.cs.ualberta.ca/~bioinfo/PA/Sub/	Proteome Analyst Specialized Subcellular Localization Server (SVM based)
Bio-numbers		
CyberCell Database (CCDB)[88]	http://redpoll.pharmacy.ualberta.ca/CCDB/cgi-bin/STAT_NEW.cgi	
B10NUMB3R5	http://bionumbers.hms.harvard.edu/	
Available reconstruction software packages		
Simpheny	http://www.genomatica.com/technology/technologySuite.html	Commercial software
COBRA simulation environments		
CellNetAnalyzer[90]/ FluxAnalyzer[91]	http://www.mpi-magdeburg.mpg.de/projects/cna/cna.html	Matlab is required
COBRA Toolbox[16]	http://systemsbiology.ucsd.edu/Downloads/Cobra_Toolbox	Matlab is required
FluxExplorer[92]		
MetaFluxNet[93,94]	http://mbel.kaist.ac.kr/lab/mfn/	Standalone package

which is a suite of COBRA functions commonly used for simulations[16]. This Toolbox was extended to facilitate the reconstruction, debugging and manual curation process described herein.

This protocol describes in detail the process to generate metabolic reconstructions applicable for representatives of all domains of life. The process of reconstructing prokaryotic and eukaryotic metabolic networks is, in principle, identical, although eukaryote reconstructions are more challenging because of the size of genomes, coverage of knowledge and the multitude of cellular compartments. Specific properties and pitfalls are highlighted.

The described reconstruction and debugging process requires organism-specific information. The minimum information includes the genome sequence, from which key metabolic functions can be obtained, and physiological data, such as growth conditions, which allow the comparison of model prediction to refine the network's content. In general, the more information about physiology, biochemistry and genetics is available for the target organism, the better the predictive capacity of the models. This property becomes obvious considering that the network evaluation and validation process relies on comparing predicted phenotypes (e.g., growth rate) with experimental observations. Additional cellular objectives (other than maximal growth rate) may be compared with experimental data but they are not detailed in this protocol[15,17–20].

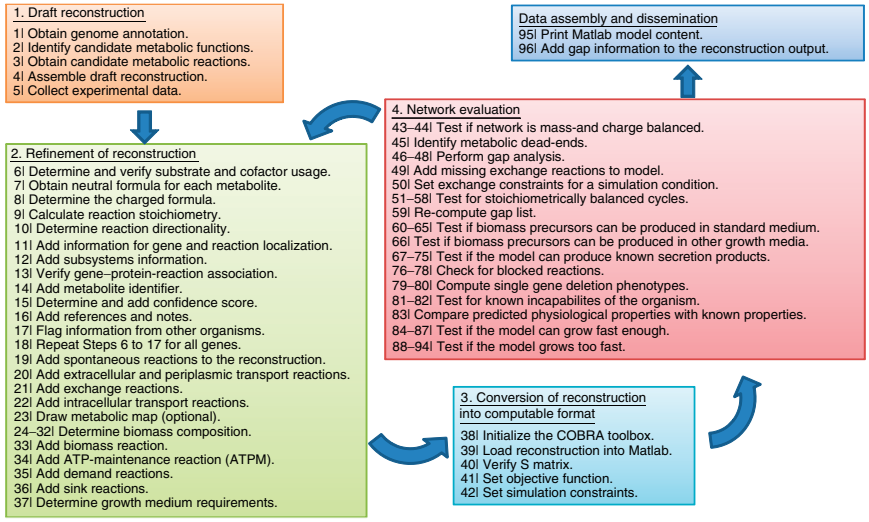


Figure 1 | Overview of the procedure to iteratively reconstruct metabolic networks. In particular, Stages 2–4 are continuously iterated until model predictions are similar to the phenotypic characteristics of the target organism and/or all experimental data for comparison are exhausted.

Although this protocol presents the reconstruction process in terms of metabolic networks, the same approach can, and has been, applied for reconstructing signaling[21,22] and transcription/translation networks[23]. Regulatory networks have not yet been constructed in a fully stoichiometric manner, although a pseudo-stoichiometric approach has been proposed[24,25]. The reconstruction process for these networks is not as well established as for metabolic networks and is thus still subject to active research.

A myriad of data sources are used during the reconstruction process rendering metabolic network reconstructions as knowledge bases, which summarize and structure the available BiGG knowledge about the target organism. Frequently used organism-unspecific, and some of the organism-specific, resources are listed in **Table 1**. It should be noted that the quality and wealth of organism-specific information will directly affect the quality and coverage of the metabolic reconstruction. Great resources are organism-specific books that have been published for a growing number of organisms[26–29]. In cases where organism-specific information is scarce, data from phylogenic neighbors may be of great help. It is important to ensure that, in cases where the reconstruction relies extensively on relative information, the overall behavior of the model matches the target organism. This assurance can be achieved by carefully comparing the predictions with experimental and physiological data, such as growth conditions, secretion products and knockout phenotypes.

The resulting knowledge bases can be queried, used for mapping experimental data (e.g., gene expression, proteomic, fluxomic and metabolomic data) and converted into a mathematical format to investigate metabolic capabilities and generate new biological hypotheses. The multitude of possible applications of BiGG knowledge bases distinguishes them from automated efforts. By introducing standards in content and format with this protocol it will soon be possible to compare metabolic reconstructions between different organisms, which will further enhance our understanding of the evolutionary processes and may provide a complementary approach to comparative genomics.

Experimental design

The metabolic network reconstruction process described herein consists of four major stages followed by its prospective use in Stage 5 (**Fig. 1**). The order of steps in the different stages is a recommendation and may be altered within each stage, and with some limitations between stages, as long as they are completed. The quality of the reconstruction is generally ensured by carrying out all the steps.

Stage 1: Creating a draft reconstruction. It is to be noted that the creation of a draft reconstruction and the manual reconstruction refinement (next stage) may be combined for bacterial reconstructions with the main emphasis on reconstruction refinement.

The first stage consists of the generation of a draft reconstruction based on the genome annotation of the target organism and biochemical databases. This draft reconstruction, or automated reconstruction, is thus a collection of genome-encoded metabolic functions, some of which may be falsely included even though others are missing (e.g., because of missing, wrong or incomplete annotations). Software tools such as Pathway tools[30] or metaSHARK[31] can be used for the generation of draft reconstruction, but they do not replace the manual curation.

Genome annotation (Step 1): Genomic information is important to unambiguously define the gene properties with respect to the organism's genome, as well as to allow data mapping (e.g., gene expression) in subsequent studies. As the draft reconstruction, and to some extent the curated reconstruction, relies mainly

on the genome annotation, it is important to download the most recent version available to ensure that updates and corrections since the genome's original publication are accounted for. Thus, the quality and reliability of the genome annotation is crucial to the quality of reconstruction. It should be noted that the manual reconstruction refinement tries to identify low-confidence gene annotations by retrieving further, experimental evidence for the presence of a gene product and its metabolic function. The reconstruction assembly and refinement may also require re-annotation of genes, but the procedure is not further discussed in this protocol. Please refer to the available work and reviews[32–36]. Furthermore, in some cases, the genome-sequencing group created an organism-specific database (e.g., for *Helicobacter pylori*[37] and *E. coli*[38]), which is very valuable during the reconstruction process. **Table 1** lists some of the commonly used databases for genome annotations.

Candidate metabolic functions (Step 2): To obtain the draft reconstruction, one can automatically retrieve metabolic genes from the genome annotation by using, e.g., key words or gene ontology (GO) catergories[39] (see **Supplementary Fig. 1**). Metabolic reactions catalyzed by the identified gene products can be connected with the draft reconstruction by using enzyme commission (E.C.) numbers[40] and biochemical reaction databases, e.g., KEGG[41] and BRENDA[42]. It is to be noted that this first stage aims to obtain a list of candidates that will not necessarily be complete or comprehensive. Many false positives may be present in the list. For example, proteins involved in DNA methylation or rRNA modification also have E.C. numbers, but their functions are normally not considered in metabolic reconstructions. Another example involves kinases that may be involved in signal transfer reactions or annotated as 'histidine kinase-like', and thus, no specific function can be derived from this annotation. A more targeted query for metabolic annotations could be designed to reduce the number of false positives but it does not replace manual curation.

Stage 2: Manual reconstruction refinement. In this stage, the entire draft reconstruction will be re-evaluated and refined. For each gene and reaction entry, two questions will be asked: (1) Should this entry be here? (2) Is there an entry missing to connect the entry with the remainder of the network?

The second stage of the reconstruction process concentrates on curation and refinement of the network content. We highlight in this protocol parts that need special attention. In particular, the metabolic functions and reactions collected in the draft reconstruction are individually evaluated against organism-specific literature (and expert opinion). This manual evaluation is important since (1) not all annotations have a high confidence score (e.g., low e-value) and (2) biochemical databases are mostly organism unspecific, listing enzyme activities found in various organisms, not all of which may be present in the target organism (**Fig. 2**). Including organism-unspecific reactions may affect the predictive behavior of the resulting models. Furthermore, information about biomass composition, maintenance parameters and growth conditions are collected in this stage, which will provide a basis for the simulations in Stages 3 and 4.

Reconstruction assembly: It is generally recommended to refine and assemble the curated reconstruction in a pathway-by-pathway manner, starting from the canonical pathways. Peripheral pathways and reactions/gene products without clear pathway assignment are added in a later step. This approach has the advantage that reactions are evaluated within their metabolic context and missing gene annotations can readily be identified, thus facilitating gap analysis and debugging in Stage 4. However, this approach will also result in the identification of additional reactions that are not in the

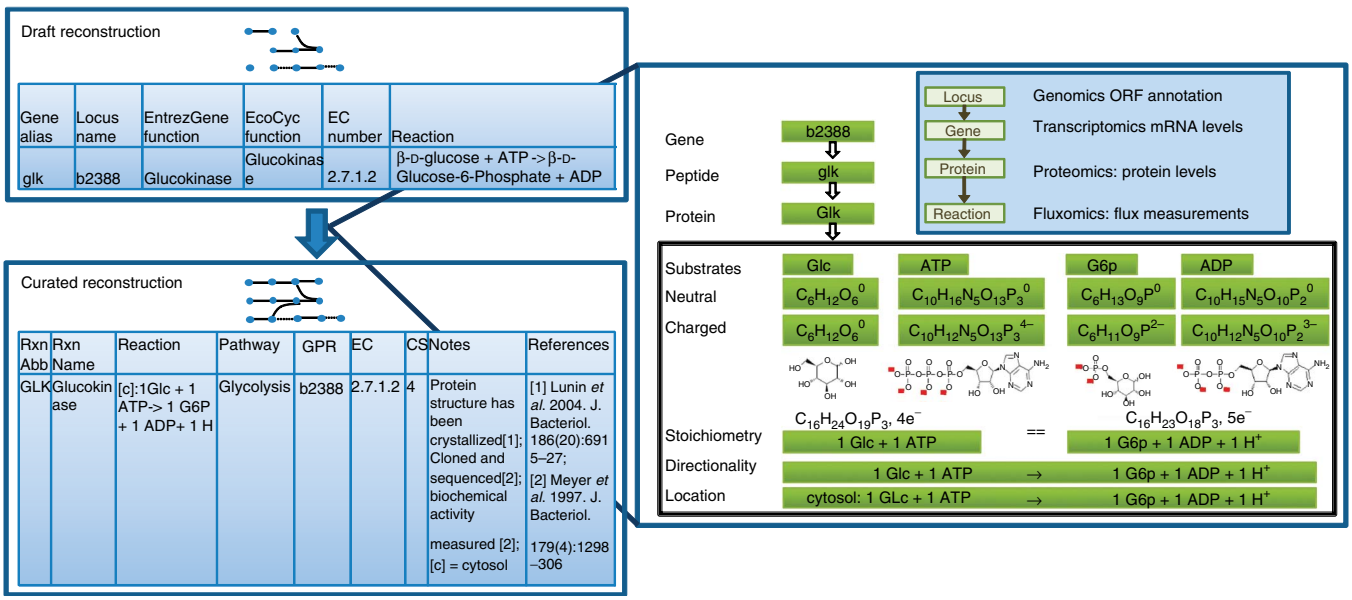


Figure 2 | Refinement of reconstruction content. The draft reconstruction is converted into a curated reconstruction by re-evaluation of the content. In particular, the metabolic reactions, obtained from biochemical databases or the literature, need to be tested for mass and charge balancing. Many resources omit protons and water. Furthermore, adjusting metabolites to a particular pH may change their charged formulae and thus may require correction of the network reaction. For instance, the reaction catalyzed by the glucokinase, which was obtained from KEGG[86], is not mass and charge balanced when charged metabolite formula at pH 7.2 is considered. The right-hand side (RHS) is missing an H^+ and the charge is unbalanced. Adding a proton to the RHS balances both sides of the equation in terms of protons and electrons/charge. Glc, D-glucose; G6P, D-glucose-6-phosphate; ATP, adenosine triphosphate; ADP, adenosine diphosphate; H^+, proton; CS, confidence score.

pathway, which is presently under investigation. One can choose to only include the main reaction(s) associated with the considered pathway. The remaining reactions should be noted so that, if necessary, they can be retrieved readily.

Verification of metabolic functions (Step 6): The draft reconstruction identified a set of metabolic genes and functions that are thought to be present in the target organism. Owing to potential errors or incomplete genome annotation, the presence of the annotated gene and its function should be supported using experimental data or literature.

Use of phylogenetically close organisms (Step 6): If no organism-specific information can be found in the literature, information from phylogenetically close organisms can be used but should be marked as such. If enzyme-associated reactions are included purely based on gene annotation, they should receive the lowest confidence score (**Table 2**). In the case of problems during subsequent simulations, these low confidence reactions can easily be identified.

Generic reaction terms (Step 6): In some cases, it is appropriate to exclude certain reactions from being added to in the reconstruction. Reactions containing generic terms, such as protein, DNA, electron acceptor, and so on, should not be included, as they are not specific enough and normally serve in databases as space holders until more knowledge and biochemical evidence become available.

Substrate and cofactor usage (Step 6): Substrate and cofactor specificity of enzymes may differ between organisms. Organism-unspecific databases, such as KEGG[41] and BRENDA[42], list all possible transformations of an enzyme that have been identified in any organism. In addition, BRENDA lists organism-specific information along with relevant references and kinetic parameters. As a rule of thumb, one can assume that enzymes, which have only one reaction associated in, e.g., KEGG[41], do not require organism refinement for substrate and cofactor usage. However, enzymes that are associated with multiple reactions, with varying substrates and/or cofactors, require manual refinement. Information about substrate and cofactor utilization can be obtained from organism-specific biochemical studies and may also be listed in organism-specific databases (e.g., Ecocyc[43]). This part of the curation process can be very time consuming and laborious, as it may be difficult to find the necessary information. Often, this part requires intensive literature search. It is important to pay great attention as false inclusion of substrates or cofactors can greatly change the *in silico* behavior (i.e., predictive potential) of the reconstruction.

Charged formula for each metabolite (Steps 7 and 8): In databases, metabolites are generally listed with their uncharged formula. In contrast, in medium and in cells, many metabolites are protonated or deprotonated. The protonation state, and thus, the charged formula, depends on the pH of interest. Often metabolic networks are reconstructed assuming an intracellular pH of 7.2. However, the intracellular pH of bacterial cells may vary depending on, e.g., environmental conditions. Also, the pH of organelles may be different, e.g., peroxisome and lysosome. The protonated formula is calculated based on the pK_a value of the functional groups (**Fig. 3**). Software packages, such as Pipeline Pilot and pK_a DB, can predict the pK_a values for a given compound (**Table 1**). **Figure 3** shows some examples of charged molecules and their pK_a values.

Reaction stoichiometry (Step 9): Once the charged formula is obtained for each metabolite, the reaction stoichiometry can be determined by counting different elements on the left- and right-hand side of the reaction (**Fig. 2**). Addition of protons and water may be required in this step, as some databases and many biochemical textbooks omit these molecules from the reactions. It is therefore necessary to balance every element and charge on both sides of the reaction. This step is easy for many central metabolic reactions but may become challenging for more complex reactions. It should be noted that unbalanced reactions may lead to the synthesis of protons or energy (ATP) out of nothing (see also **Fig. 4** for examples).

Reaction directionality (Step 10): Biochemical data for the target organism are very important for determination of reaction directionality but may not be available. New approaches such as the estimation of standard Gibbs free energy of formation ($\Delta_f G'°$) and of reaction ($\Delta_r G'°$) in a biochemical system are available[44,45]. The $\Delta_f G'°$ and $\Delta_r G'°$ can be obtained for most KEGG[41] reactions from Web GCM[44]. Another approach combines thermodynamic information with network topology and heuristic rules to assign reaction directionality[46]. Biochemical textbooks may also report reaction directionalities. In addition, one can use the following rules of thumb: (1) reactions involving transfer of phosphate

TABLE 2 | Confidence scoring system currently employed for metabolic reconstructions.

Evidence type	Confidence score	Examples
Biochemical data	4	Direct evidence for gene product function and biochemical reaction: protein purification, biochemical assays, experimentally solved protein structures and comparative gene-expression studies (e.g., Chhabra *et al.*[95])
Genetic data	3	Direct and indirect evidence for gene function: knockout characterization, knock-in characterization and overexpression
Physiological data	2	Indirect evidence for biochemical reactions based on physiological data: secretion products or defined medium components serve as evidence for transport and metabolic reactions
Sequence data	2	Evidence for gene function: genome annotation and SEED annotation[32]
Modeling data	1	No evidence is available, but reaction is required for modeling. The included function is a hypothesis and needs experimental verification. The reaction mechanism may be different from the included reaction(s)
Not evaluated	0	

from ATP to an acceptor molecule should be irreversible (with the exception of the ATP synthetase, which is known to occur in reverse direction); and (2) reactions involving quinones are generally irreversible.

It is to be noted that assigning the wrong direction to a reaction may have significant impact on the model's performance. In general, one should leave a reaction reversible if no information is available and the aforementioned rules of thumb do not apply. However, models with too many reversible reactions (too loose constraints) may have the so-called futile cycle that can overcome the proton gradient by freely exchanging metabolites and protons across compartments. Therefore, assigning the correct reversibility to transport reactions is especially important (see below).

Information for gene and reaction localization (Step 11): This information may also be difficult to obtain. The compartments that have been considered in various metabolic reconstructions are listed in **Supplementary Table 1**. Algorithms such as PSORT[47] and PASUB[48] can be used to predict the cellular localization of proteins based on nucleotide or amino acid sequences. A recently published protocol describes the use of internet-accessible tools to predict the subcellular location of eukaryotic and prokaryotic proteins[49]. High-throughput experimental approaches are available to locate individual proteins, including immunofluorescence[50] and GFP tagging of individual proteins[51]. In the absence of appropriate data, proteins should be assumed to reside in the cytosol. Incorrect assignment of the reaction location can lead to additional gaps in the metabolic network and misrepresentation of the network properties, especially if intracellular transport reactions need to be added without further evidence.

Gene–protein-reaction (GPR) associations (Step 13): The genome annotation often provides information about the GPR association, i.e., it indicates which gene has what function (**Fig. 5**). The verification and refinement necessary in this step includes determining: (i) if the functional protein is a heteromeric enzyme complex; (ii) if the enzyme (complex) can carry out more than one reaction and (iii) if more than one protein can carry out the same function (i.e., isozymes exist). For the first case (i), the genome annotation often has refined information, e.g., 'protein X, catalytic subunit'—which indicates that there is at least one more subunit needed for the function of the protein complex. Furthermore, KEGG[41] lists subunits in some cases. Often, a more comprehensive database and/or literature search is required. Also, the protein-complex composition may differ between organisms. The second case (ii) can also be identified from biochemical databases and/or literature. Multitasking of enzymes may also differ between

Molecule/group	Acid	Base	pK_a
Acetic acid	$H_3C-C(=O)-OH$	$H_3C-C(=O)-O^-$ + H^+	4.76
Carboxyl group	$R-CH(NH_3^+)-C(=O)-OH$	$R-CH(NH_3^+)-C(=O)-O^-$ + H^+	1.8–2.4
Ammonium	NH_4^+	NH_3 + H^+	9.25
Amino group	$R-CH(NH_3^+)-C(=O)-O^-$	$R-CH(NH_2)-C(=O)-O^-$ + H^+	8–11
Bicarbonate	$HO-C(=O)-OH$	$HO-C(=O)-O^-$ + H^+	3.77
	$HO-C(=O)-O^-$	$O^-C(=O)-O^-$ + H^+	10.2
Glycine	$NH_3^+-CH_2-C(=O)-OH$	$NH_3^+-CH_2-C(=O)-O^-$ + H^+	2.34
	$NH_3^+-CH_2-C(=O)-O^-$	$NH_2-CH_2-C(=O)-O^-$ + H^+	9.6
Phosphoric acid	$HO-P(=O)(OH)-OH$	$HO-P(=O)(OH)-O^-$ + H^+	2.14
	$HO-P(=O)(OH)-O^-$	$HO-P(=O)(O^-)-O^-$ + H^+	6.86
	$HO-P(=O)(O^-)-O^-$	$O^-P(=O)(O^-)-O^-$ + H^+	12.4

Figure 3 | List of functional groups, their charge formula and the corresponding pK_a.

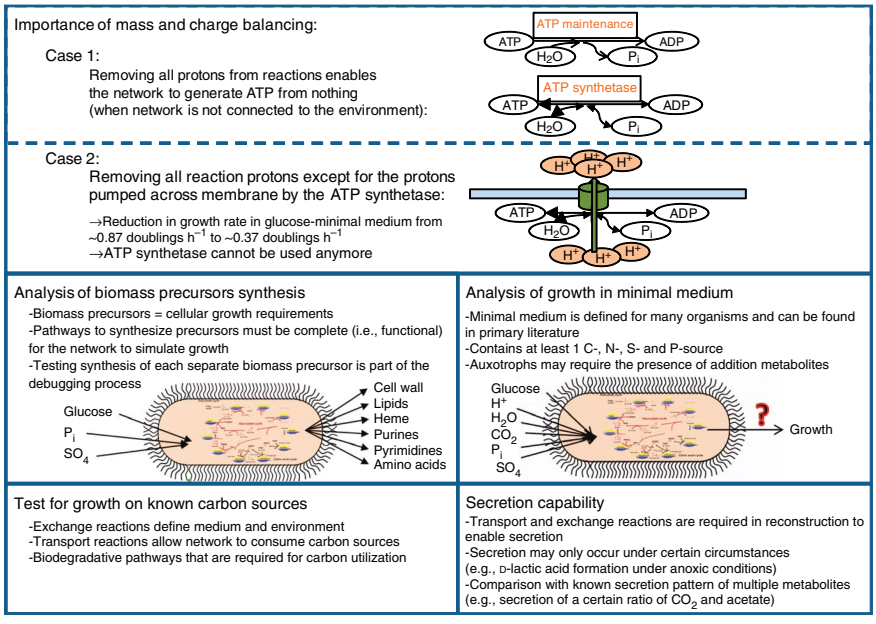


Figure 4 | Examples of network evaluation. The network evaluation and debugging stage (Stage 4) includes various quality-controlled quality-assured (QC/QA) tests, some of which are illustrated in this figure. For instance, mass and charge balancing of network reactions is crucial to ensure similar properties of the model and the cell or organism. A standard test for most metabolic reconstructions is to verify that each biomass precursor, which makes up a new cell, can be produced by the model in different growth conditions (e.g., minimal medium, different carbon sources and so on). Other QC/QA tests may include the capability to secrete certain metabolites given a particular growth condition. At its end, the models will have similar properties as the cell and error cases can be used to systematically refine the models and thus the reconstruction content.

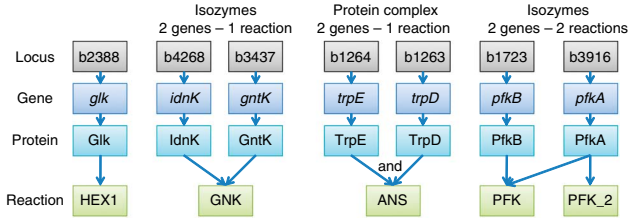
organisms. It is to be noted that mistakes or misassignments in the GPR associations will change results of *in silico* gene deletion studies. However, discrepancies between *in silico* and *in vivo* results can be used to refine knowledge and reconstructions (see Steps 79 and 80).

Linear pathways, such as fatty acid oxidation, have often been combined into few lumped reactions. The genes associated with these reactions are all required, with the exception of isozymes. Subsequently, the GPR association should reflect the requirement for all genes within the lumped reaction by using the Boolean rule AND.

Metabolite identifiers (Step 14): Metabolite identifiers are necessary to enable the use of reconstructions for high-throughput data mapping (e.g., metabolomic or fluxomic data) and for comparison of network content with other metabolic reconstructions. Therefore, metabolites and reactions need to be recognizable by other scientists and by software tools. Each metabolite should be associated with at least one of the following identifiers: ChEBI[52], KEGG[41] and PubChem[53]. In many cases, having one of the identifiers is sufficient to automatically obtain the other two identifiers. Furthermore, database-independent representations of metabolites such as SMILES[54] and InCHI strings[55,56] are also helpful when associated with each metabolite. These representations represent the exact chemical structure of compounds. In addition, collecting Molfiles (MDL file format, http://www.symyx.com/), which hold information about the atoms, bonds, connectivity and coordinates of a molecule, will be very useful, e.g., if the online software for pK_a determination is being used (see Step 10 for details).

Confidence scoring system (Step 15): The confidence score provides a fast way of assessing the amount of information available for a metabolic function, pathway or the entire reconstruction[15,57]. Every network reaction is associated with a confidence score reflecting the information and evidence currently available. The confidence score ranges from 0 to 4, where 0 is the lowest and 4 is the highest evidence score (**Table 2**). It should be noted that multiple information types result in a cumulative confidence score. For example, a confidence score of 4 may represent physiological and sequence evidence.

Spontaneous reactions (Step 19): An excerpt of typical spontaneous reactions included in metabolic reconstructions is listed in **Supplementary Table 2**. Note that only those spontaneous reactions should be added that have at least one metabolite connecting them to the rest of the reconstruction. This is to avoid too many dead-end metabolites caused by spontaneous reactions. In more recent reconstructions, spontaneous reactions have been associated with an artificial gene (*s0001*) and protein (S0001). By doing so, reaction and gene essentiality studies are easier to analyze. Furthermore, this artificial GPR association makes it easy to distinguish between spontaneous and orphan reactions, i.e., reactions without known gene.



Reaction abbreviation	Reaction name	E. C.number	GPR
HEX1	Hexokinase (D-glucose:ATP)	2.7.1.1	(b2388)
GNK	Gluconokinase	2.7.1.12	(b3437) or (b4268)
ANS	Anthranilate synthase	4.1.3.27	(b1264) and (b1263)
PFK	Phosphofructokinase	2.7.1.11	(b1723) or (b3916)
PFK_2	Phosphofructokinase (2)	2.7.1.11	(b3916)

Figure 5 | Gene–protein-reaction (GPR) associations. Examples of GPR associations and their representation in Boolean format are shown for *Escherichia coli*.

Intracellular transport reactions (Step 22): When multicompartment networks are constructed, intracellular transport reactions need to be added for all the metabolites that are supposed to 'move' between compartments. Intracellular transport systems are not very well studied and many of these are not annotated in the genome. Finding experimental data is often not easy. A general approach should be to minimize the number of intracellular transport reactions to the ones that really need to be there. If too many transport reactions are added in a reconstruction, they can cause cycles (futile cycles or Type III pathways). This is a common problem in reconstructions with multiple compartments. For the directionality of intracellular transport reactions, one should consider the nature of the pathway in the compartment. For instance, if the pathway is biosynthetic, it is very likely that (i) the precursor(s) is only imported, (ii) the product(s) of the pathway is only exported from the compartment and (iii) intermediates are not transported at all. Another issue is the transport mechanism. Many transport reactions are in symport or antiport with protons, cations or other metabolites. However, not much information is available for intracellular transporters, but the used mechanism may affect the predictive potential of the model. To minimize the error and increase consistency, one can adopt the intracellular transport mechanism from a corresponding transport reaction from extracellular/periplasmic space to cytoplasm if it is known (and it is not an ABC transport reaction); otherwise (facilitated) diffusion reaction may be assumed as the mechanism. In any case, these reactions should receive a low confidence score (1 for modeling purpose) to enable easy identification (**Table 2**), as well as a note and references describing where the mechanism was taken from.

Identification of missing functions: The refinement stage of the reconstruction process is also an ideal point to identify missing functions in the draft reconstruction. Using KEGG[41] maps, e.g., one can analyze the metabolic 'environment' of the reaction(s) under inspection. If the genome annotation of the target organism is present in KEGG[41], one can highlight the genes on the map. This gives an estimate of the 'connectivity' of the reaction with its metabolic surrounding (**Supplementary Fig. 2**). Missing reactions/functions may become apparent for which experimental/annotation evidence should be collected (see also gap analysis). Creating organism-specific maps, using specific drawing software, is of great use for identifying missing functions as well as for network evaluation and debugging.

Biomass composition (Steps 24–33): The biomass reaction accounts for all known biomass constituents and their fractional contributions to the overall cellular biomass (**Table 3**). A detailed biomass composition of the target organism needs to be determined experimentally for cells growing in log phase[58–60]. However, it may not be possible to obtain a detailed biomass composition for the target organism. In this case, one can estimate the

TABLE 3 | Chemical composition of cells.

Cellular component	Cellular content %(wt/wt)
Protein	55
RNA	20.5
DNA	3.1
Lipids	9.1
Lipopolysaccharides	3.4
Peptidoglycan	2.5
Glycogen	2.5
Polyamines	0.4
Other	3.5
Total	100.00

Listed here is the cellular content of *E. coli* taken from Neidhardt *et al.*[64].

relative fraction of each precursor from the genome (e.g., by using the Comprehensive Microbial Resource (CMR) database, **Table 1**). Note that we do not suggest taking the RNA composition from *E. coli* rather than estimating it using organism-specific genome data. One reason is that the number of rRNA operons, which contain rRNA and tRNA molecules, can differ significantly between organisms. For instance, *E. coli* has seven rRNA operons per genome[61], whereas *Mycoplasma capricolum* has two[62] and *Halobacterium cutirubrum* has only one rRNA operon[63].

In comparison to other biomass precursors, it is slightly more difficult to determine the lipid composition of the cell. The contribution of fatty acids and phospholipids needs to be determined from experiments. Note that compounds, such as phospholipids, can consist of many different fatty acids (different chain length, saturated and unsaturated). Available experimental data often report the average composition of these compounds, listing the measured fraction of fatty acids with different chain length and saturation status. Thus, the model compounds will not represent all possible combinations but only average compounds that are consistent with the experimental data individual.

The composition of the biomass reaction has an important role for *in silico* gene deletion experiments. If a biomass precursor is not accounted for in the biomass reactions, the synthesis reactions may not be required for growth (i.e., it is nonessential). Therefore, associated genes may not be essential. Subsequently, the presence or absence of a metabolite in the biomass reaction may affect the *in silico* essentiality of reactions and their associated gene(s). In contrast, the fractional contribution of each precursor has a minor role for gene and reaction essentiality studies. When one wishes to predict the optimal growth rate accurately, the fractional distribution of each compound has an important role. The unit of the biomass reaction is h^{-1}, as all biomass precursor fractions are converted to mmol g_{DW}^{-1}. Therefore, the biomass reaction sums the mole fraction of each precursor necessary to produce 1 g dry weight of cells.

Growth-associated ATP maintenance reaction (GAM) (Step 32): The GAM reaction accounts for the energy (in the form of ATP) necessary to replicate a cell, e.g., for macromolecular synthesis (e.g., proteins, DNA and RNA). The GAM is best determined in chemostat growth experiments (see also **Fig. 6**). Alternatively, if experimental data is not available, the GAM can be estimated by determining the energy required for macromolecular synthesis. Therefore, the total amount of macromolecule (protein, DNA and RNA) is determined from databases or other resources. Neidhardt *et al.*[64] list the amount of phosphate bonds necessary to synthesize a macromolecule, which is then multiplied with the total amount of the macromolecule. These phosphate bonds are accounted for by adding ATP hydrolysis to the biomass reaction (x ATP + x H_2O → x ADP + x P_i + x H^+, where x is the number of required phosphate bonds). Note that this estimate will be too low, as other growth-associated cellular processes also require ATP.

Non-GAM reactions (NGAM) (Step 34): More recent reconstructions include an ATP hydrolysis reaction (1 ATP + 1 H_2O → 1 ADP + 1 P_i + 1 H^+), which represents NGAM requirements of the cell to maintain, e.g., turgor pressure[65]. The value for the reaction rate can be estimated from growth experiments. For example, based on such measurements, the reaction flux rate was constrained to 8.39 mmol g_{DW}^{-1} h^{-1} in the *E. coli* metabolic model[65] (**Fig. 6**).

Demand reactions (Step 35): Demand reactions are unbalanced network reactions that allow the accumulation of a compound, which otherwise is not allowed in steady-state models because of mass-balancing requirements (i.e., in steady state the sum of influx equals the sum of efflux for each metabolite) (**Fig. 7**). Most of the demand reactions will be added in the gap-filling process (Steps 46–48). At this stage, demand functions should only be added for compounds that are known to be produced by the organism, e.g., certain cofactors, lipopolysaccharide and antigens, but (i) for which no information is available about their fractional distribution to the biomass or (ii) which may only be produced in some environmental conditions. By including a demand reaction for a particular metabolite one can turn otherwise blocked reactions (cannot carry flux) into active reactions (can carry flux). In general, metabolic reconstructions contain only few demand reactions. However, during the debugging- and network-evaluation process (Stage 4), demand reactions may temporarily be added to the model to test or verify certain metabolic functions. They will be removed from the model before versioning.

Sink reactions (Step 36): Sink reactions are similar to demand reactions but are defined to be reversible and thus provide the

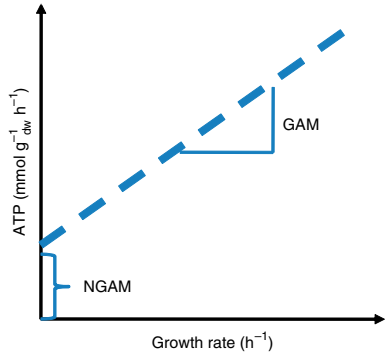


Figure 6 | Growth-associated maintentance (GAM) and non-GAM (NGAM). The best way to obtain accurate information regarding GAM and NGAM is by plotting growth data obtained from chemostat growth experiments. GAM and NGAM can be directly read from the plot.

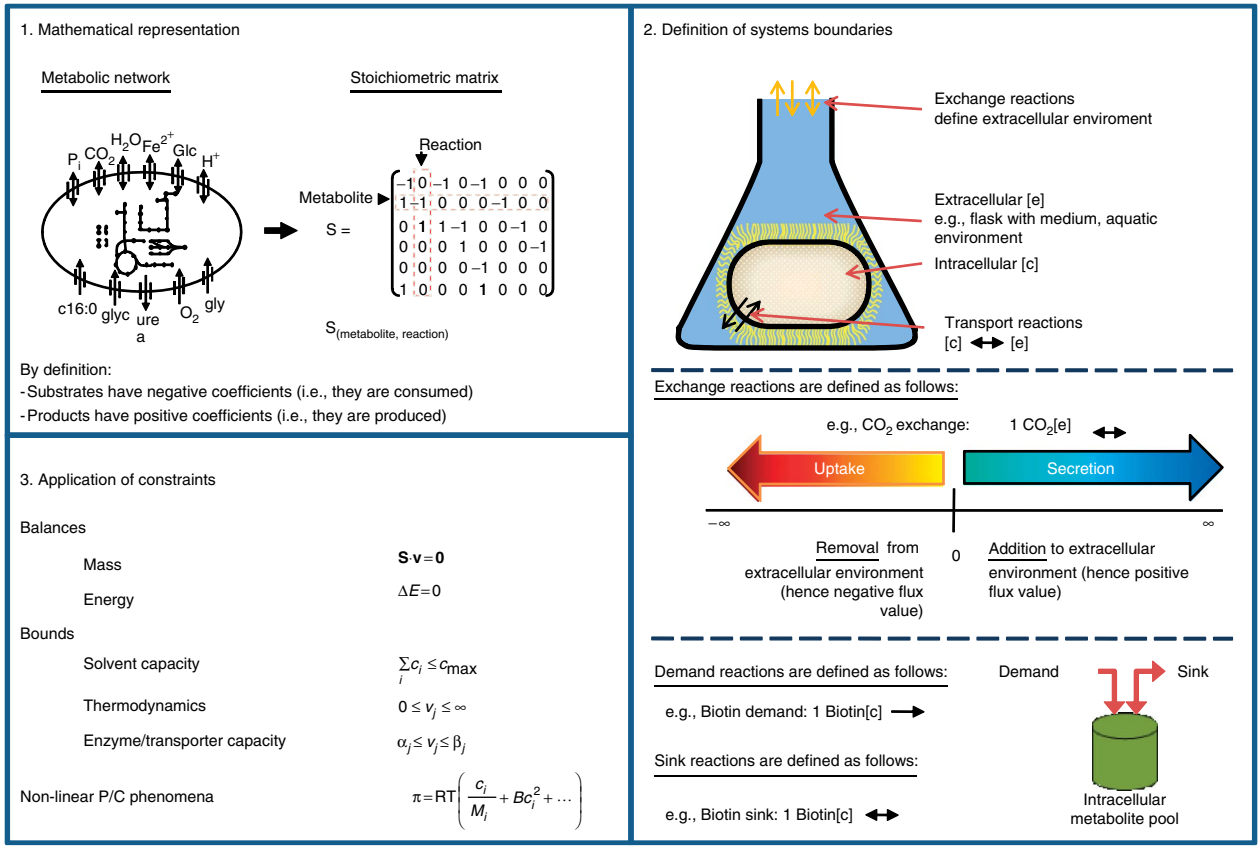


Figure 7 | Conversion of reconstruction into a condition-specific model. This conversion requires three main steps. (1) The first step involves the mathematical representation by a stoichiometric matrix, S, of the network reaction list. The columns of S correspond to the network reactions, whereas the rows represent the network metabolites. The substrates in a reaction are defined to have a negative coefficient, whereas the products have a positive value. The metabolites participating in a reaction have a nonzero entry in the S matrix. (2) Now that the reconstruction is in a computer-readable format, the systems boundaries need to be defined. In particular, this means that for all metabolites that can be consumed or secreted by the target organism, a so-called exchange reaction needs to be added to the reconstruction. The exchange reactions can be employed in later simulation to define environmental conditions (e.g., carbon source). (3) As a last step, constraints will be added to the reconstruction, thus rendering it to a condition-specific model. Mass conservation is a basic physical law. All steady states can be thus described by S.v = 0, where v is a vector of reaction fluxes. Adding further constraints such as thermodynamics (reaction directionality), enzyme capacity or regulation (i.e., presence or absence of an enzyme) to the model will lead to a smaller, more confined set of feasible steady-states flux solutions.

network with metabolites (see **Fig. 7** for examples). These sink reactions are of great use for compounds that are produced by nonmetabolic cellular processes but that need to be metabolized. Adding too many sink reactions may enable the model to grow without any resources in the medium. Therefore, sink reactions have to be added with care. As for demand reactions, sink reactions are mostly used during the debugging process. They help in identifying the origin of a problem (e.g., why a metabolite cannot be produced). These sink reactions are functionally replaced by filling the identified gap.

Growth medium requirements (Step 37): Information about growth-enabling media is of great help in the following two stages. Thus, if possible, it should be collected before the conversion and debugging stage. The following information should be collected: (1) Which metabolites are present? (2) Are there any auxotrophies? (3) The definition of a base medium composition, e.g., water, protons, ions and so on. (4) Information about rich medium composition. This data will be crucial for simulations and network evaluation. If uptake or secretion rates are available, they should also be documented and collected. Although this step is easy for the experimentalist, researchers who cannot grow the target organism have to identify growth requirements from the literature (or genome annotation). In some cases, research studies describe minimal, defined or rich medium composition. In other cases, the culturing conditions reported in some experimental study must be sufficient.

Stage 3: Conversion from reconstruction to mathematical model. In the third stage, the reconstruction is converted into a mathematical format and condition-specific models are defined. This stage can mostly be automated. Moreover, systems boundaries are defined, converting the general reconstruction into a condition-specific model. It is to be noted that the initial model may differ in scope and boundaries to the final model, which is obtained after multiple iterations of validation and refinement and is used to simulate phenotypic behavior in a prospective manner. **Figure 7** illustrates the conversion of a reconstruction into mathematical format.

Simulation constraints (Step 42): Using the functions in the COBRA Toolbox, it is very easy to change reaction constraints, but sometimes it is difficult to keep track of all the changes. In fact, one of the most common reasons for errors in simulation is that

TABLE 4 | General error modes in metabolic networks.

Error mode	Action
Wrong reaction constraints	Check reaction constraints if they are applied correctly
Missing transport reactions	Add transport reactions
Missing exchange reactions	Add exchange reactions
Cofactor cannot be consumed or produced	Follow **Figure 13**
Shuttling of compounds across compartment	Adjust reversibility of transport reactions

reaction constraints are not set correctly (**Table 4**). Therefore, it is important to have an expectation of the results before running a simulation to avoid erroneous conclusions. It is recommended that the constraints are checked by copying the model reaction abbreviations as well as lower and upper bounds into a spreadsheet. For most models, this is the easiest way to see where problems are with the constraints. Similarly, copying calculated solution(s) into a spreadsheet is very helpful.

Stage 4: Network evaluation = 'Debugging mode'. The fourth stage in the reconstruction process consists of network verification, evaluation and validation. Common error modes in metabolic reconstructions are listed in **Table 4**. The metabolic model created in the third step is tested, among other things, for its ability to synthesize biomass precursors (such as amino acids, nucleotides triphosphates and lipids). This evaluation generally leads to the identification of missing metabolic functions in the reconstruction, so-called network gaps, which are added by partially repeating Stages 2 and 3. Thus, the reconstruction process is an iterative procedure. An important issue is to decide when to stop the iterative process and call a reconstruction 'finished'. This decision is normally based on the definition of the scope and purpose of the reconstruction.

Metabolic dead end (Step 45): At this point, the first iteration of the manual curated reconstruction is finished. It is expected that the network contain a significant number of gaps, i.e., missing reactions and functions. We recommend carrying out a first gap analysis at this stage of the reconstruction process, as it will ease subsequent computation and reduce the number of 'bugs' in the model. Comparing dead-end metabolites identified in this step with the curated reaction list generated in Stage 2 will accelerate the debugging process.

Candidate reactions for gap filling (Steps 46 and 47): This step will require an intensive literature search and may include re-annotation of a genome to find candidate genes and reactions to fill the gap (see **Table 1** and **Supplementary Table 3** for some example tools). KEGG[41] maps, biochemical textbooks or other available biochemical maps can be used to identify the metabolic 'environment' of the dead-end metabolite. If the genome annotation of the target organism is present in KEGG[41], one can highlight the dead-end metabolite on the map (**Supplementary Fig. 2**). This context analysis may give an indication of which enzyme(s) may be able to produce or synthesize the dead-end metabolite and thus provide a good starting point for literature and/or genome search.

Gap filling is a tricky business. In some cases, a gap should be filled to ensure that the model is functional, i.e., biomass precursor synthesis or a certain physiological function can be simulated. In other cases, filling a gap may enable the model to carry out a function that the target organism is not able to do (see **Fig. 8** for some examples). In general, if no information supports the existence of a particular gap reaction, the gap should only be filled if it is required for the model's functionality. In such cases, the confidence score should be set to 1, which corresponds to 'modeling purpose' only, and allows retrieving these low-confidence reactions readily, if desired. Earlier, we highlighted that enzymes, which are listed in biochemical databases to catalyze multiple reactions, should be included in the reconstruction with care and that it should be noted whether evidence for all the reactions could be found. Some of the identified dead-end metabolites will originate from such secondary reactions of these 'multitasking' enzymes. Closing these gaps may affect the predictive potential of the reconstruction; therefore, the only gaps that should be filled are those that are required for network functionality (e.g., biomass precursor synthesis) or which have supporting data. It should be kept in mind that adding new reactions to the network may cause new gaps. Therefore, when adding reactions make sure that all the metabolites are connected to the network.

Stoichiometrically balanced cycles (SBCs) (Steps 51–59): SBC, or Type III-extreme pathways[66], are formed by internal network reactions and can carry fluxes despite closed exchange reactions (closed system). Examples for simple or more complex Type III pathways in metabolic networks can be found elsewhere[67,68]. These SBCs are artifacts of metabolic reconstructions due to insufficient constraints (e.g., thermodynamic constraints and regulatory constraints). Recent efforts have concentrated on dealing with these SBCs[67]. It should be noted that SBCs are not futile cycles. This protocol shows how to identify SBCs and even highlights some possible approaches to eliminate them. However, no systematic, universally valid approach has yet been developed to eliminate SBCs. For practical purposes, in simulation one can use the 'min norm' option for the linear programming (LP) solver, which will minimize the sum of the squares of fluxes and thus will return an optimal solution without the net flux around SBCs.

The following steps will test if the model can or cannot grow. This means that we will test for qualitative behavior but not focus on the correctness of predicted growth rates.

Biomass precursor production (Steps 60–66): The composition of the biomass reaction was determined in Stage 2. It is recommended to test for the model's ability to produce each individual biomass component in standard medium condition (e.g., minimal medium M9 supplemented with D-glucose) (**Fig. 4**). This sequential approach will facilitate the debugging process and will make it easier to find causes of error. It is very likely that these tests will lead to the addition of further reactions by repeating steps listed in the second stage. Furthermore, this step may lead to the addition of reactions for which no experimental evidence and candidate genes can be identified. These reactions should be marked with the tag 'modeling purposes' only (confidence score of 1). Care must be taken with such reactions, as too many of them may change the overall properties of the network (in this or other simulation conditions). Moreover, the overall performance of the model in standard

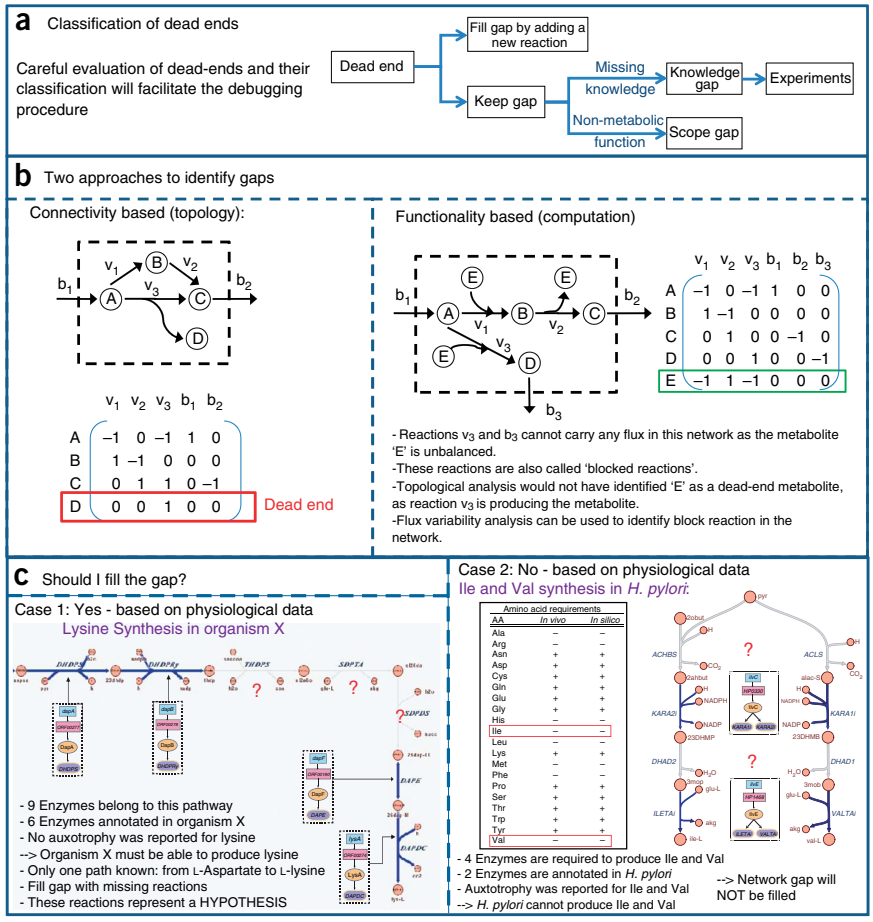


Figure 8 | Gap analysis. The gap analysis includes the identification and the tentative filling of network gaps. (**a**) Although many dead-end metabolites that create network gaps can be connected to the network by re-evaluating genomic and experimental data, some dead-end metabolites will remain in the refined, curated reconstruction. These dead-end metabolites can be categorized into two groups, depending on the type of reactions that could connect them to the remaining network: knowledge gaps and scope gaps. The knowledge gaps represent the missing biochemical knowledge for the target organism. In contrast, the scope gaps include reactions and cellular processes, which are currently not accounted for in the metabolic reconstruction (e.g., DNA methylation). (**b**) There are at least two approaches to identify gaps in the reconstruction. In the connectivity-based approach, one can count the nonzero entries in each row of the stoichiometric (S) matrix and identify those metabolites, which are only produced or consumed. In the example, metabolite D is only produced by reaction v_3 and the S matrix contains only one entry in the row corresponding to metabolite D. A second approach is based on model functionality; in this approach the model capability to carry flux through every network reaction is tested. This approach identifies blocked reactions, which are directly or indirectly associated with one or more dead-end metabolites. In the shown example, one would not identify metabolite E as a dead-end metabolite with the connectivity-based approach, as it is produced and consumed in the network. However, testing for flux through reactions containing E will show that reaction v_3 and b_3 cannot carry any flux in this model. (**c**) Two sample cases are shown that address the question of filling a gap or not.

medium condition is determined and, in some cases, corrected. This step needs great care, as there may be many possible ways of filling a gap.

Subsequently, the capability to produce biomass precursors needs to be tested in other growth media. Therefore, the correctness of the network content is evaluated with respect to all the known growth conditions of the target organism. This includes all the known carbon, nitrogen, sulfur and phosphorus sources. Physiological information is of great value to determine all growth conditions. For example, Gutnick *et al.*[69] have tested about 600 compounds and have found that 100 can serve as carbon or nitrogen sources for *Salmonella typhimurium*. The model should be able to produce biomass in the majority of these instances. However, not all the known conditions may be reproduced by the model—this is not a problem, as it represents a starting point for experimental studies to identify missing metabolic functions. Nevertheless, great attention should be given to collecting and documenting those cases and thus enabling other researchers to pursue them.

By-product secretion (Step 70): If such information is available, it can be used to further refine the model. The first question is whether the model can produce the secretion product(s) from a given substrate, whereas a subsequent question could be if a specific ratio of by-product secretion is correct. Classical biochemical studies often reported measured secretion products given in a certain carbon source (e.g., Schroeder *et al.*[70]). This information is very helpful to compare the phenotypic traits of the model with those of the target organism.

Blocked reactions (Steps 76–78): Reactions that cannot carry any flux in any simulation conditions are called blocked reactions. These reactions are directly or indirectly associated with dead-end metabolites, which cannot be balanced and give rise to the so-called blocked compounds[71]. It is good to be aware of those reactions, especially, if one expects different results in a simulation (e.g., false-negative analysis of single-gene deletion). In the early phase of the debugging stage, the reconstruction can contain many blocked reactions that one might decide to fill if the supporting information is available or if they are required for the overall function of the network. Targeted use of sink and demand reactions around a pathway of blocked reactions will facilitate the identification of the source problem. Other blocked reactions may remain if the terminal dead-end metabolite is beyond the scope of the metabolic reconstruction or no information and evidence for filling the gap is available. The easiest way to determine blocked reactions is by carrying out flux variability analysis[72,73].

Single-gene-deletion phenotypes (Steps 79 and 80): Analysis of false-positive and false-negative predictions will help to further refine the network content if the information is available or provides a basis for experimental studies otherwise (**Fig. 9**). Numerous

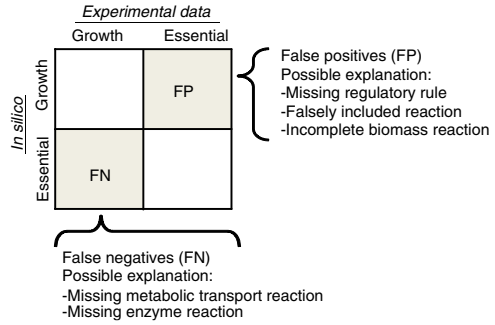


Figure 9 | *In silico* gene essentiality study as network evaluation tool. Although agreement of gene essentiality between experimental and *in silico* data is very helpful to validate the reconstruction content and model setup, analysis of inconsistencies will enable the discovery of new biological knowledge.

reconstructions relied on phenotyping data (e.g., biolog data), or gene essentiality data, to improve the network content and thus the predictive potential[74,75].

Known incapabilities (Steps 81 and 82): So far we compared whether the model can reproduce growth on a certain substrate, secrete a particular by-product and so on. In this step, it should be tested if known incapabilities of the organism can also be reproduced by the model. For example, *H. pylori* is known to be an auxotroph for certain amino acids, subsequently, their lack in the medium should abolish *in silico* growth[76]. It is important to use those 'negative' data (incapabilities) and to correct for errors. Error cases can be removed by analyzing the confidence score associated with the reactions along the pathway. In the example of *H. pylori*, this would be the biosynthetic reactions leading to amino acid synthesis[76] (**Fig. 8**). In a more algorithmic approach, a single-reaction deletion study can be carried out and the results can be analyzed in terms of which deletions disable growth. This smaller subset of reactions needs to be evaluated manually. Note that the deletion of a single function may not be sufficient when alternate pathways exist in the network. Missing incapabilities may not only be caused by falsely added reactions in the metabolic network but may be a consequence of missing regulatory information. Literature may provide the necessary data.

Comparison of predicted physiological properties with known properties (Step 83): The model should also be tested for known capabilities, besides the aforementioned growth performance and secretion capability. For instance, this test can include known carbon splits in central metabolic pathways, as observed with a recently published *Pseudomonas putida* network[57]. The P/O ratio was investigated for *Methanosarcina barkeri*[77], *Saccharomyces cerevisiae*[78] and compared with the known growth data. Many more examples exist and the suite of necessary tests depends on the available data as well as the properties of the network.

Quantitative evaluation of growth rate (Steps 84–94): Too slow growth means that at least one precursor of the biomass function cannot be synthesized sufficiently. This implies that the model's biomass production is carbon, nitrogen, oxygen, sulfur or phosphate limited. As there are generally less active uptake reactions for a particular element than biomass precursors, it is faster to test if any of the medium components are growth limiting. If the biomass reaction value increases when the uptake of reaction flux is increased, it means that this compound is limiting. This gives you a hint as to where in the network something must be missing or constraining. Further analysis of shadow prices and reduced costs, which are associated with the LP solution, can be of great help to identify metabolites or reactions that limit the rate of biomass. For example, the *P. putida* network[57] was not able to grow as fast as reported experimentally *in silico* when toluene was used as a carbon source. *In silico* analysis suggested that oxygen is rate limiting and that more oxygen-efficient reactions are missing in the network. Whether this discrepancy can be resolved by iterative network refinement depends on the specific case, and thus, no general solution can be proposed. As in the case of *P. putida*'s oxygen restriction, such error cases can lead to further experimental investigation that will ultimately increase our biological insight and the reconstruction's quality.

When the predicted growth rate is higher than expected, many explanations are possible. (1) The optimization for growth assumes that microbial cells maximize their growth. However, many other objective functions are possible and may be more appropriate depending on the experimental setup and growth conditions of the target organism[6,18–20,79–82]. (2) The GAM, which is a part of the biomass reaction, may be estimated wrongly and needs adjustment. (3) It can indicate that constraints are missing or incorrect (e.g., NGAM, missing regulation). (4) Falsely included reactions increase growth rate. Knowledge about the model and the expected flux map is crucial for identifying these errors. Proton shuttling reactions may be present that circumvent the ATP synthetase (e.g., because of a futile cycle). Note that this is only the case in aerobic growth conditions. Such shuttling reactions may be enabled by many reversible transport reactions. Reactions associated with such loops can be readily identified (see Steps 51–59). Also, looking at the flux through the reactions of oxidative phosphorylation may indicate if they are used under aerobic condition or not. Alternatively, one can investigate if there is one reaction that enables the model to grow too fast. In this case, a single-reaction deletion study will push one toward the right solution. Another approach could be to investigate the directionality of network reactions. As indicated earlier, reaction directionality may have a role in the fast growth rates. Therefore, improving reaction directionality assignments may be helpful. Make sure that only those reactions that are known to produce ATP are allowed for ATP synthesis, whereas all other reactions are set irreversible (ATP utilization). Similarly, reactions using quinones as electron acceptors should not run reversibly. This might cause problems and may allow circumventing the electron transport chain. These examples are very specific to a model and problem, and no general rule for corrections can be proposed.

Stage 5: Prospective use. Once the necessary content and desired *in silico* capability is reached, one can start to use the reconstruction in a prospective manner, which represents a fifth step in the reconstruction process that is not addressed here.

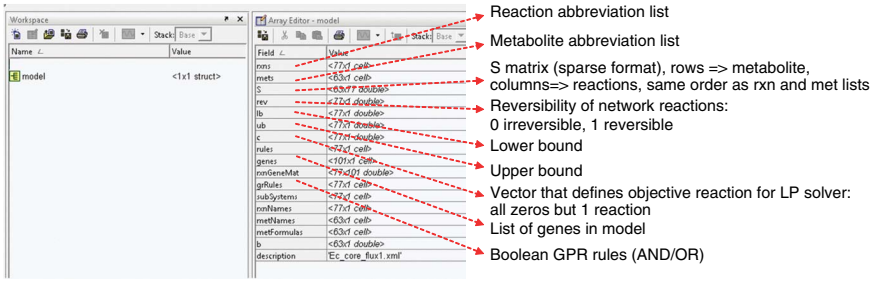


Figure 10 | Components of the model structure in Matlab. The reconstruction is imported into Matlab (Step 39). The entire reconstruction content is stored in a structure array. The screen shot illustrates the main fields contained in the model structure. The information is stored in subarrays in these fields. Note that the order of the reactions and metabolites corresponds to the order of columns and rows in the stoichiometric (S) matrix, respectively.

MATERIALS

EQUIPMENT

- A standard personal computer that can run Matlab.
- Matlab, version 6.0 or above (Mathwork), a numerical computation and visualization software.
- COBRA Toolbox (version 1.3.4 or above) is provided at http://systemsbiology.ucsd.edu/downloads/COBRAToolbox
- The SBML Toolbox for Matlab, which allows reading models in SBML format http://sbml.org/Software/SBMLToolbox
- An LP solver. Multiple solvers are currently supported by the COBRA Toolbox:
  (i) glpk (freeware): http://www.gnu.org/software/glpk/
  (ii) LINDO (LINDO Systems) Matlab API (commercial)
  (iii) CPLEX (ILOG) through the Tomlab (Tomlab Optimization) optimization environment (commercial, but best LP solver available)
  (iv) Mosek (MOSEK ApS) (commercial)
- Extreme pathway software package, X3, provided at http://systemsbiology.ucsd.edu/downloads/Extreme_Pathway_Analysis
- Excel (Microsoft, http://office.microsoft.com/en-us/excel/default.aspx) or similar database programs can be used for collecting reconstruction information.

EQUIPMENT SETUP

COBRA Toolbox The COBRA Toolbox[16] should be downloaded and copied in a local folder on the user's computer. Extract the .zip file. After opening Matlab, a path should be set to the local folder, containing the COBRA Toolbox (Matlab → File → Set Path → Add with Subfolder, choose the corresponding folder and save). All working files (SBML and xls files) should also be stored in the local folder to allow access to reconstruction and models. A full documentation of the COBRA Toolbox can be found in the 'doc' subfolder within the main Toolbox folder, which has all help files as html files. Furthermore, help for Matlab and COBRA Toolbox functions can be accessed through Matlab's 'help' facility by typing 'help function_name' on Matlab command line. See also Becker *et al.*[16].

SBML Toolbox Comprehensive documentation on SBML, the file format, and model setup, can be found at the official SBML website (http://sbml.org/documents/, level 2 version 1). The SBML file describing the model has to include at least the following information: stoichiometry of each reaction, upper/lower bounds of each reaction, and objective function coefficients for each reaction. In addition, gene-reaction associations can be added to the 'Notes' section.

Spreadsheet The first two reconstruction steps are illustrated in this protocol using spreadsheets. It is important that the order of the columns in the spreadsheet match the example given in **Supplementary Methods 2**.

Variables The imported model from the spreadsheets is contained in a model structure (see **Fig. 10** for details on this structure). All functions in the COBRA Toolbox access the information stored in the model structure. The values computed by the COBRA Toolbox are fluxes, which represent reaction rates for all model reactions. The units for fluxes used throughout this protocol are mmol g_{DW}^{-1} h^{-1}, where g_{DW} is the dry weight of the cell in grams.

Installation The Matlab software, SBML Toolbox and one or more of the suggested LP solvers should be installed following the instructions of the software providers. Note that the SBML Toolbox and the LP solver also need to be accessible in the Matlab path (see above). Sample installation instructions for the lp_solve LP solver on Windows can be found in Becker *et al.*[16]. The SBML Toolbox is downloaded and installed. The installation instructions are to be followed. Choose 'libsbml' in the dialog field. Once installed, open Matlab and type 'install'. If you get an error with 'libsbml' (when opening Matlab again), go to 'setpath' and add the folder 'libsbml' with subfolders.

The COBRA Toolbox is initiated by typing in the Matlab command window:
(1) changeCobraSolver(solverName); where 'solverName' is, e.g., 'lp_solve'
(2) initCobraToolbox;

▲ **CRITICAL STEP** SBML Toolbox and the LP solver should be tested for functionality following the software provider's instructions before attempting to use the COBRA Toolbox.

X3 X3 is the software package used to determine stoichiometrically unbalanced cycles or Type III pathways. X3.exe needs to be placed and extracted in the local folder. Help can be accessed by opening the DOS command line, changing to the local folder and typing X3 –h. The extreme pathway tool will be called from Matlab by the COBRA Toolbox.

KEGG Many steps of the protocol have been illustrated using KEGG[41] because it is freely accessible and very helpful for the illustrated pathway-by-pathway reconstruction process. However, one has to keep in mind three properties of KEGG[41]: (1) It is NOT organism-specific data; hence, not all reactions associated with an enzyme may be catalyzed by the enzyme of the target organism, and (2) KEGG[41] may not update the genome annotation of the target organism on a regular basis; hence, the information may be outdated and need a 'second opinion' from another more recent resource. (3) Not all reactions in the KEGG[41] database are mass and charge balanced, as they omit protons and water molecules, although the KEGG database is continuously updated and improved[83,84].

PROCEDURE

Stage 1: Creating a draft reconstruction ● TIMING Days to 1 week

1| *Obtain genome annotation.* The genome annotation can be obtained from various sources, including sequencing centers (e.g., TIGR) and the National Center for Biotechnology Information (NCBI) depository. The following information should be retrieved for each gene: genome position, coding region, strand, locus name, alias, gene function (i.e., current annotation) and protein classification (e.g., E.C. number[40]).

▲ **CRITICAL STEP** In eukaryotic organisms, information regarding alternate transcripts must also be collected, as different splice forms may have distinct function or cellular localization.

2| *Identify candidate metabolic functions*. This step is straightforward once the genome annotation has been obtained. Different approaches can be applied to collect candidate metabolic functions including searching for E.C. numbers (complete and partial)[40] and for metabolic terms (e.g., dehydrogenase, kinase and so on) (**Supplementary Fig. 1**). If GO[39] or cluster of orthologous groups of proteins[85] information is obtained with the genome annotation, they can be used as well to find metabolic enzymes.

3| *Obtain candidate metabolic reactions for these functions* (e.g., from KEGG[41]). Use comprehensive reaction databases such as KEGG[41], BRENDA[42] and publically available reconstructions, as a resource to combine the gene functions with metabolic reactions.

4| *Assemble draft reconstruction*. Collect all candidate metabolic genes and their potential reactions in a spreadsheet. This spreadsheet will serve as a starting point for the manual curation process (see **Fig. 2** and **Supplementary Data 1**, for an example).

5| *Collect experimental data*. ● **TIMING** **Ongoing throughout the reconstruction process.**
The manual curation process relies heavily on experimental, organism-specific information. All possible information needs to be retrieved. The following steps will include reviewing of scientific literature to collect information listed in **Table 5**. Alternatively, additional experimental data can be generated by growing and measuring various metabolic capabilities and properties of the target organism.

Stage 2: Manual reconstruction refinement ● **TIMING** **Months to 1 year**
6| *Determine and verify substrate and cofactor usage*. Use primary literature, and to a lesser extend KEGG[41] and BRENDA[42], to determine and verify substrate and cofactor specificity of the enzyme in the target organism. As a rule of thumb, one can assume that enzymes, which have only one reaction associated, e.g., in KEGG[41] do not require organism refinement.
! **CAUTION** Often only biochemical data can reveal the correct cofactor and substrate, as binding sites may not be distinguishable in gene sequence from related metabolites.

7| *Obtain a neutral formula for each metabolite in the reaction*. The neutral formula can be readily obtained from various resources, including KEGG[41], BRENDA[42] and PubChem[86]. Although PubChem[86] is more comprehensive, KEGG[41] is certainly the most accessible resource, especially when KEGG[41] is used for obtaining the reactions.
! **CAUTION** Check that the formula is correct (i.e., verify with other databases and textbooks).

8| *Determine the charged formula for each metabolite in the reaction*. Retrieve the molecular structure for each metabolite if it has not been done in Step 7. Determine the charged formula (e.g., for pH 7.2) based on the pK_a value of the functional groups (**Fig. 3**). This can also be done using software packages such as Pipeline Pilot, and pK_a DB can predict pK_a values for a given compound (**Table 1**).

9| *Calculate reaction stoichiometry*. Count every element and the charge on each side of the equation. On each side, the same number of elements and charge must be present. It may be necessary to add protons and water to the reaction. This step is easy for many central metabolic reactions but may become challenging for more complex reactions.

10| *Determine reaction directionality*. Use biochemical data and literature if available. Alternatively, the standard $\Delta_f G'°$ and of $\Delta_f G'°$ can be calculated based on group contribution theory for most KEGG[41] reactions from Web GCM[44,45]. If data on reaction of interest are not available, the following rule of thumb may be applied: (1) reactions involving transfer of phosphate from ATP to an acceptor molecule should be irreversible (with the exception of the ATP synthetase, which is known to occur in reverse); and (2) reactions involving quinones are generally irreversible.

11| *Add information for gene and reaction localization*. This information may be difficult to obtain from primary literature. The use of algorithms such as PSORT[47] and PASUB[48] can be considered if no experimental data are available.
▲ **CRITICAL STEP** In the absence of appropriate data, proteins should be assumed to reside in the cytosol.

12| *Add subsystem information to the reaction*. This information will be of great help for the debugging and network evaluation work. The subsystem assignment can be done based on, e.g., biochemical textbooks or KEGG[41] maps. Note that a reaction or an enzyme can appear in multiple KEGG[41] maps; therefore, the subsystem should reflect its primary function.

13| *Verify GPR association*. Determine if the functional protein is a heteromeric enzyme complex, if the enzyme (complex) can carry out more than one reaction and if more than one protein can carry out the same functions (i.e., isozymes exist). Organism-specific databases and primary literature can be used to obtain this information.
▲ **CRITICAL STEP** Mistakes or misassignments in the GPR associations will change the results of *in silico* gene-deletion studies.

TABLE 5 | List of experimental data commonly used for reconstruction, modeling and network evaluation.

Data type	Purpose	Literature	Data-bases	Genome	Growth experi-ments	Pheno-typing	Protein structures	Comparative genomics[a]	Micro-array data	Proteomic data	Exometabo-lomic data	Metabo-lomic data	Flucomic data	Single-gene deletion	Biochemi-cal essays
Gene function	Reconstruction refinement	X	X	X				X							X
Protein function	Reconstruction refinement	X	X	X			X	X							X
Reaction mechanism	Reconstruction refinement	X	X	X											X
Growth media	Transport, simulations	X			X	X					X	X			
Carbon sources	Transport reactions, simulations	X			X	X					X	X			
Gene/ protein presence/ absence	Condition-specific models, cell-type models		X						X	X				X	
Reaction constraints	Simulations				X	X			X	X	X		X		
Network evaluation	Debugging				X	X		X			X	X	X	X	
Gap filling	Debugging	X	X	X		X			X	X					X
Cofactor/ substrate specificity	Reconstruction refinement	X	(X)				X								X
Reaction directionality	Reconstruction refinement	X	(X)												X

[a]Comparative genomics can be done using, e.g., SEED[32].

14| *Add metabolite identifier.* Associate each metabolite with at least one of the following identifiers: ChEBI[52], KEGG[41] and PubChem[53]. In addition, associate database-independent representations of metabolites such as SMILES[54] and InCHI strings[55,56] with each metabolite.

15| *Determine and add the confidence score.* The proposed confidence score system listed in **Table 2** should be used.

16| Flag those reactions for which information from other organisms was used.

17| *Add references and notes based on experimental information.* In Steps 6–13 many organism-specific experimental data are collected that must be associated with the reconstruction in the form of references and notes. This allows other users of the reconstruction to easily retrace the evidence and supporting material for reaction and gene inclusion.

18| *Repeat Steps 6–17 for all those genes that were identified in the draft reconstruction.* These steps are to be repeated for metabolic functions identified from bibliomic sources during the reconstruction process.

19| *Add spontaneous reactions to the reconstruction.* The biochemical literature and databases (e.g., KEGG[41] and BRENDA[42]) are to be used to identify candidate spontaneous reactions that are to be included. Only include those reactions, which have at least one metabolite present in the reconstruction to minimize the number of dead ends. Associate the spontaneous reactions with an artificial gene (*s0001*) and protein (S0001).

20| *Add extracellular and periplasmic transport reactions to the reconstruction.* This addition is done based on experimental data. The rule here is that for every metabolite that is known to be taken up from the medium or that is known to be secreted into the medium, a transport reaction should exist (from extracellular space to periplasm and from periplasm to cytoplasm). The transport reactions for metabolites that can diffuse through the membranes must be included. Small, hydrophilic compounds can diffuse through the outer membrane[87].

21| *Add exchange reactions to the reconstruction.* Exchange reactions need to be added for all extracellular metabolites. The exchange reactions represent the systems boundaries (**Fig. 7**).

22| *Add intracellular transport reactions to the reconstruction* (for multicompartment reconstructions only). Use biochemical and physiological information; however, finding experimental data is often not easy. Only include intracellular transport reactions that really need to be there to avoid futile cycles, or Type III pathways.

23| *Draw metabolic map* (*optional*). If appropriate drawing software is available, the creation of organism-specific maps is very useful for gap analysis, network evaluation and data mapping.

Determine biomass composition ● TIMING Days to weeks

24| *Determine the chemical composition of the cell, i.e., protein, RNA, DNA, lipids, and cofactor content* (see also **Table 3** and **Supplementary Fig. 3a**). This information can be retrieved from experimental data or primary literature.

25| Determine the amino acid content either experimentally (option A) or by estimation (option B).

(A) Experimental determination of amino acid content

(i) Obtain data for each amino acid.

(B) Estimation of amino acid composition from genome information (e.g., use CMR database (**Table 1**))

(i) The amino acid content can be determined by selecting the Genome Tools tab, followed by Analysis Tools and finally Codon Usage.

26| The molar percentage and molecular weight of each amino acid must be used to calculate the weight per mol protein. Add the individual amino acid values to give a total molecular weight of the protein content. Subsequently, calculate the weight percent per amino acid. Then multiply the calculated weight percent by the cellular content percentage of the macromolecule and divide by the molecular weight of the individual monomer (**Fig. 11** and **Supplementary Fig. 3b**).

27| Determine the nucleotide content either experimentally (option A) or by estimation (option B).

(A) Experimental determination of the nucleotide content

(i) Obtain data for each deoxynucleotide triphosphate (dATP, dCTP, dGTP and dTTP) and each nucleotide triphosphate (ATP, CTP, GTP and UTP).

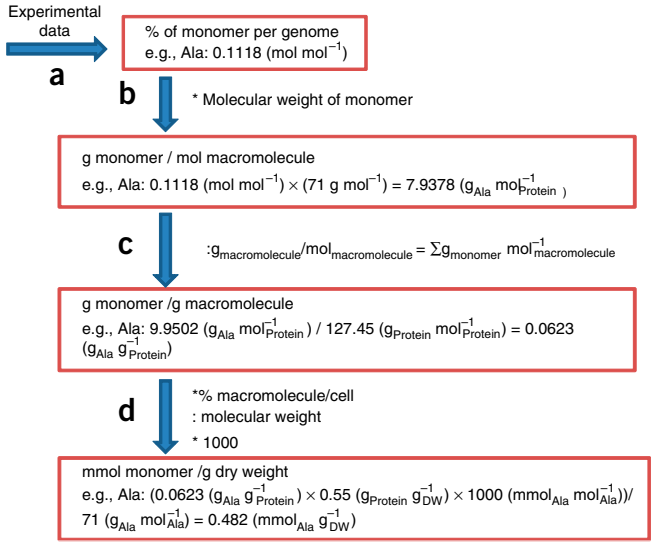


Figure 11 | Flow chart to calculate the fractional contribution of a precursor to the biomass reaction. This approach can be used for amino acids, nucleotide triphosphates (ATP, GTP, CTP and UTP) and deoxynucleotide triphosphates (dATP, dGTP, dCTP and dTTP). The steps are illustrated for L-alanine (Ala). (**a**) The fractional contribution of alanine to the proteome is obtained from experimental data or estimated from genome sequence. (**b**) To convert the molar percentage into weight of alanine per mole protein, the molar percentage is multiplied by the molecular weight of alanine. Note that the polymerization of amino acid leads to the loss of a water molecule, which needs to be considered when calculating the molecular weight. Once the weight of amino acid per mole protein is obtained for all amino acids, they are summed to obtain the weight of protein per mole protein. (**c**) The weight of alanine per mole protein is converted into weight alanine per weight protein by multiplying with the sum of all amino acids' weight. (**d**) Finally, the weight of alanine is multiplied by the cellular content of protein (see **Fig. 13a**) and divided by its molecular weight to obtain the mole alanine per cell dry weight. Multiplying this molar contribution by a factor of 1,000 will result in a final unit of mmol alanine per gram of dry weight.

(B) Estimation of nucleotide composition from genome information

(i) For example, use CMR database (**Table 1**). From the Genome Tools tab, select Summary Information, followed by DNA Molecule Info. The number of each dNTP (i.e., dATP, dCTP, dGTP and dTTP) present in the genome is listed on the summary page.

(ii) To determine the RNA composition of the cell, the codon usage that was accessed for the amino acid content in Step 25 can be used. It must be remembered that RNA incorporates U instead of T; therefore, the codon usage needs to be read with every T replaced by a U.

(iii) Tabulate the frequency of each nucleotide.

28| Calculate the fractional distribution of each nucleotide to the biomass composition by repeating Step 26.

29| *Determine the lipid content*. Determine the contributions from fatty acids and phospholipids. Therefore, (i) determine the average molecular weight of a fatty acid in the cell by incorporating the average fatty acid composition of the cell (requires experimental data, e.g., from literature). (ii) The average molecular weight of each fatty acid must be used and (iii) add the weight contributions of each fatty acid to determine the average molecular weight for the fatty acid chain. (iv) Use this weight to calculate the average molecular weight of various lipids within the cell. Carry out such a computation by adding the molecular weight of the core structure of the molecule and the molecular weight of the fatty acids attached to the core structure based on the average molecular weight of one fatty acid that was determined above. (v) The molar percentages of the three major phospholipids, phosphatidylethanolamine, phosphatidylglycerol and cardiolipin, may be found in the literature. (vi) Then determine the phospholipid contributions to the biomass function (**Supplementary Fig. 3c**).

30| *Determine the content of the soluble pool* (*polyamines and vitamins and cofactors*). The soluble pool contains, e.g., spermidine, coenzyme A and folic acid (see **Supplementary Table 4** for a more comprehensive list). Use **Figure 12** as a template to determine the composition of the soluble pool for the target organism and to calculate the fractional distribution to the biomass reaction.

31| *Determine the ion content*. The calculation of the molar fraction of the ions is illustrated in **Supplementary Table 5**. It assumes that concentration data are available or can be estimated for each ion. Information about the ion content can be obtained from different resources, including primary literature and databases (e.g., CyberCell Database[88]). Convert the reported concentration (c_i) for each ion species i into mM. Add all the ion species (total ion concentration, c_{total}). Calculate the molar fraction (f_i) of each ion species i by dividing c_i with c_{total}:

$$f_i = \frac{c_i}{c_{total}} \quad \text{where} \quad c_{total} = \sum c_i$$

32| *Determine GAM*. Experimental data should be used to determine the GAM. Alternatively, part of GAM can be estimated by the energy required for macromolecular synthesis, e.g., proteins. **Figure 13** illustrates how to calculate the GAM using the total amount (mmol) of macromolecule (protein, DNA and RNA) and known amount of phosphate bonds necessary to synthesize the macromolecules. Note that this estimate will be too low, as other growth-associated cellular processes also require ATP.

33| *Compile and add biomass reaction to the reconstruction*. In this step, all precursors are assembled in one single reaction, the biomass reaction, which is then added to the reaction list of the reconstruction. Add GAM to biomass reaction as follows: x ATP + x H_2O → x ADP + x P_i + x H^+, where x is the number of required phosphate bonds.

▲ **CRITICAL STEP** It is to be noted that some metabolites might be produced. For instance, in the *E. coli* biomass reaction, proton (H^+), orthophosphate (P_i) and some other metabolites are produced[65]. These metabolites originate mainly from the growth-associated ATP hydrolysis (Step 32).

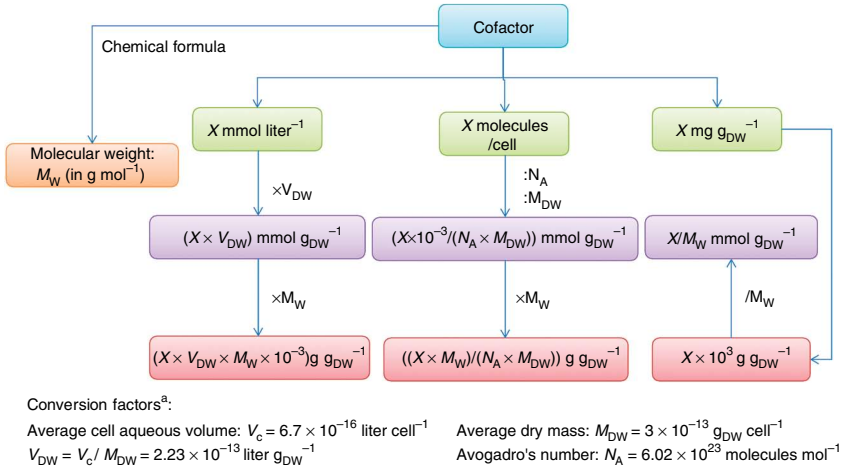


Figure 12 | Determination of the content of soluble pool. Depending on the available information from literature, measurements or database entries the conversion into mmol g_{DW}^{-1} and g g_{DW}^{-1} is shown. The value in the purple box corresponds to the stoichiometric coefficient in the biomass reactions for the precursor. [a]Information was obtained from Cybercell Database (CCDB; see **Table 1** for the link)[75].

34| *Add NGAM*. Add the following reaction to the reconstruction reaction list: 1 ATP + 1 H_2O → 1 ADP + 1 P_i + 1 H^+.

35| *Add demand reactions to the reconstruction*. Add demand functions for compounds that are known to be produced by the organism, e.g., certain cofactors, lipopolysaccharide and antigens, but (i) for which no information is available about their fractional distributions to the biomass or (ii) which may only be produced in some environmental conditions.

36| *Add sink reactions to the reconstruction*. Sink reactions are of great use for compounds that are produced by nonmetabolic cellular processes but need to be metabolized.

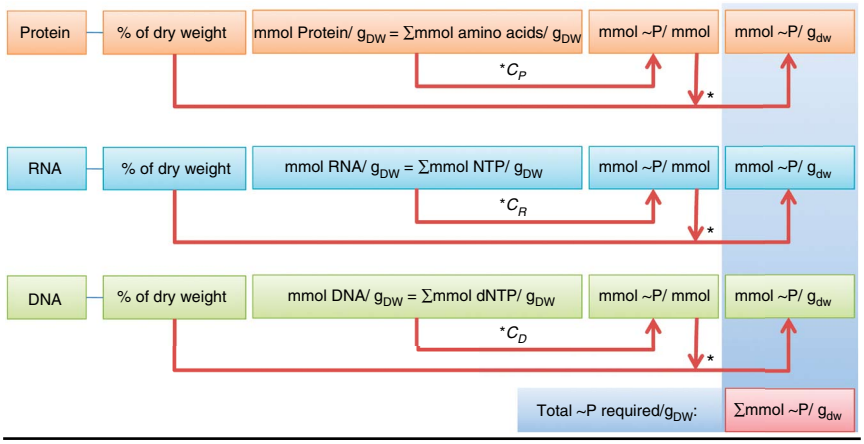
▲ **CRITICAL STEP** Adding too many sink reactions may enable the model to grow without any resources in the medium. Therefore, sink reactions have to be added with care.

37| *Determine growth medium requirements*. Use experimental data and primary literature to retrieve essential nutrients and defined medium composition. Compile a list of growth requirements.

Stage 3: Conversion from reconstruction to mathematical model ● TIMING Days to 1 week

38| *Initialize the COBRA Toolbox*. Install Matlab, the required Toolboxes (SBML Toolbox and COBRA Toolbox) and an LP solver[16]. Start Matlab as described in the installation instruction. Within Matlab, change to the main working directory. Initiate the COBRA Toolbox by entering the command initCobraToolbox in the Matlab command line. Note that the default LP solver can be changed by editing the initCobraToolbox script or at any time during a Matlab session by using the changeCobraSolver function included in the Toolbox.

a Biosynthetic cost: required energy (in ~P) per cellular content of macromolecules:



b

	wt %	Total mmol	mmol ~P/mmol	Total
Protein	0.563	5.197	C_P = 4.324	22.472
DNA	0.031	0.101	C_D = 1.365	0.138
RNA	0.21	0.649	C_R = 0.406	0.264
			Total	22.873

Growth-associated maintenance:
Hydrolysis of 22.873 mmol ATP g_{DW}^{-1}
Added to biomass reaction:
x ATP + x H2O → x ADP + x P_i + x H,
where x is 22.873

Figure 13 | Determination of growth-associated maintenance (GAM) cost. (**a**) Calculation of GAM cost. (**b**) Sample calculation for *Escherichia coli*[65]. The energy necessary for the synthesis of the macromolecules from the building blocks were obtained from **Tables 5** and **6** of Chapter 3 in Neidhardt *et al.*[64]. The coefficient c_p, c_D and c_R were obtained calculating the total energy necessary for the macromolecules divided by the total number of building blocks (see Neidhardt *et al.*[64]).

A list of frequently used COBRA Toolbox functions is given in **Supplementary Table 6**. See also the *Nature Protocol* on the COBRA Toolbox for details on initializing, testing and using the Toolbox[16].
? TROUBLESHOOTING

39| *Load reconstruction into Matlab.* Save the reaction list in a spreadsheet with the same order of columns as shown in **Supplementary Methods 2** ('RxnFileName'). A second file containing metabolite information needs to be saved as well ('MetFileName'). The following COBRA Toolbox function should be used to read the reconstruction into Matlab:

```
model = xls2model(RxnFileName, MetFileName);
```

The loaded metabolic model is stored in a structure named 'model' in Matlab. This structure contains all the information about the reconstruction in different fields of the structure. **Figure 10** provides a description of the individual fields and their content.
? TROUBLESHOOTING

40| *Verify S matrix.* Use spy(matrix) to verify the structure of the imported S matrix. This visualization should be repeated when reactions are added to the reconstruction to ensure that they are connected to the network.

41| *Set objective function.* Use the following COBRA Toolbox function to set the objective function of the model:

```
model = changeObjective(model, rxnNameList, objectiveCoeff)
```

The reaction(s) that should be set as the objective function is given by 'rxnNameList'. It will receive a corresponding coefficient 'objectiveCoeff'. This means that a single reaction or a linear combination of multiple reactions can be chosen as the objective function.
▲ **CRITICAL STEP** The COBRA Toolbox is set up in a way that the coefficient(s) for the objective function has to be a positive number. When minimizing, the input option to the COBRA Toolbox function optimizeCBmodel.m can be set to 'min'. The default option of the 'optimizeCBmodel' function is maximizing ('max') (see **Supplementary Table 6**).

42| *Set simulation constraints.* Use the following function to set the constraints of the model:

```
model = changeRxnBounds(model, rxnNameList, value, boundType)
```

The list of reactions for which the bounds should be changed is given by 'rxnNameList', whereas an array contains the new boundary reaction rates ('value'). This type of bound can be set to lower bound ('l') or upper bound ('u'). Alternatively, both bounds can be changed ('b'). Use the following command to list all constrained reactions that are greater than a minimal value ('MinInf') and smaller than a maximal value ('MaxInf'):

```
PrintConstraints(model, MinInf, MaxInf)
```

In addition, there is a function available that lists all reactions and their flux values in a solution ('fluxData'):

```
printFluxVector(model, fluxData)
```

Stage 4: Network evaluation = 'Debugging mode' ● TIMING Weeks to months
43| *Test if the network is mass and charge balanced. Check for stoichiometrically unbalanced reactions.* All, or a subset, of the network reactions can be given as input ('RxnList') along with the model structure ('model'):

```
[UnbalancedRxns] = CheckMassChargeBalance(model, RxnList)
```

In case of unbalanced reactions, the function returns a structure containing the name of the unbalanced reaction and which elements are unbalanced ('UnbalancedRxns').

44| *Evaluate stoichiometrically unbalanced reactions.* Looking at the reaction equations and the charged formula for each metabolite will help to identify ways to balance the reactions. Normally, there are two common errors causing unbalanced reactions: Missing proton and/or water or the stoichiometric coefficient of at least one metabolite is wrong. If it is the latter error, repeat Step 9. If a proton as substrate is missing, then a proton donor may be necessary (e.g., NADH, NADPH). This will require a literature search to identify a candidate proton donor. If a water molecule is missing, it should be kept in mind that after adding water to the equation the proton and oxygen will need to be balanced again.
! CAUTION A few network reactions are always unbalanced. These reactions include the biomass reaction, demand, sink and exchange reactions.

45| Identify metabolic dead ends. Use

```
[Gaps] = AnalyzeGaps(model)
```

to identify the gaps. The function will return a list of all metabolites ('Gaps') that are only produced ('Product') or consumed ('Substrate') in the network. Dead-end metabolites that are caused by reversible reactions will be listed as 'Substrate_Product' in the 'Gaps' list. Copy this gap list into an excel sheet, where information and references can be easily added for each dead-end metabolite.

46| *Identify candidate reactions to fill gaps.* Use primary literature and genome annotation tools to find candidate genes and reactions to fill the gap (see **Table 1** for some example tools). Also, use KEGG[41] maps, biochemical textbooks or other available biochemical maps to identify the metabolic 'environment' of the dead-end metabolite. If the genome annotation of the target organism is present in KEGG[41], one can highlight the dead-end metabolite on the map. This may give an indication of which enzyme(s) may be able to produce or synthesize the dead-end metabolite and thus provide a good starting point for literature and/or genome search.

47| *Add gap reactions to the reconstruction.* If experimental and/or annotation data support gap reactions or they are needed for modeling purposes, the reaction(s) should be added to the reconstruction by repeating Steps 6–17.
▲ **CRITICAL STEP** Keep in mind that adding new reactions to the network may cause new gaps. Therefore, when adding reactions one should make sure that all the metabolites are connected to the network. Repeat Step 45, if necessary.

48| *Add notes and references to dead-end metabolites.* Each dead-end metabolite should be documented. The note for the remaining dead-end metabolites should distinguish between *knowledge* and *scope* gap for future reference (**Fig. 8a**).
▲ **CRITICAL STEP** The more detailed and carefully the gap-filling steps are done (Steps 46–48) the easier and faster the debugging process will be.

49| *Add missing exchange reactions to model.* The gap-filling process may have resulted in the inclusion of further transport reactions. Thus, exchange reactions need to be added to the reconstruction. Repeat Step 21.

50| *Set exchange constraints for a simulation condition.* Determine an environmental condition, in which most network evaluation tests should be carried out initially ('*standard condition*'). Use

```
model=changeRxnBounds(model,rxnNameList,value,boundType)
```

to set the constraints. Reactions whose bounds should be changed are listed in 'rxnNameList'. The new value for each reaction is contained in the array 'value'. Finally, the type of constraint has to be defined in the list 'boundType'. The possible types are: 'l' for lower bound, 'u' for upper bound and 'b' if both reaction bounds should be set to the specified value.

Test for stoichiometrically balanced cycles or Type III pathways (*optional*)

51| *Test for Type III pathways.* Therefore, use the following function:

```
TestForTypeIIIPathways(model,ListExch);
```

A list of indices of the exchange reactions in the S matrix ('ListExch') has to be provided to the function. These exchange reactions will be set to zero and then the flux variability of the closed model is calculated. This function requires that X3.exe is in the working directory. The function will return files if there are Type III pathways in the model.
? TROUBLESHOOTING

52| *Analyze the output if Type III pathways are found.* If Type III pathways have been identified, there are two output files: one file ('ModelTestTypeIII_myT3.txt') has all Type III pathways as a matrix, wherein the rows are the different pathways and the columns correspond to the network reaction (in the same order as given in 'ModelTestTypeIII_myRxnMet.txt'). Note that the extreme pathway package converts network reactions into elementary reactions (i.e., irreversible reactions). A second file ('ModelTestTypeIII_myT3_Sprs.txt') contains the Type III pathways in a sparse format, which is easier to analyze by hand.

53| *Identify Type III pathways.* Note that reversible reactions form Type III pathways as well. In general, one is looking for Type III pathways that contain three or more reactions. It is possible that multiple complicated Type III pathways may exist in the model. Listing the corresponding reaction formulas or even drawing a map might be helpful to understand how the reactions form the loop(s).

54| Analyze directionality of each reaction participating in a Type III pathway. Re-investigate the thermodynamic information if available (Step 10).

55| Analyze if any reaction participating in a Type III pathway may be falsely included in the reconstruction by reviewing the supporting evidence.

56| If none of the reactions or reaction directions can be corrected based on experimental or thermodynamic information, you can try to iteratively limit the directionality of the loop reactions. A more elaborate procedure has been described elsewhere[67].

57| Adjust the directionality for all those reactions identified in Steps 54–56, note the change and reasons.

58| After eliminating a reaction direction or a deletion of a reaction, repeat the Type III pathway analysis. Also, make sure that the removal of directionality or reaction does not affect the model's growth capabilities.
▲ **CRITICAL STEP** Keep in mind that such a change to the network is a hypothesis and may cause problems under different simulation conditions (e.g., environmental conditions).

59| *Recompute gap list.*

```
[Gaps] = AnalyzeGaps(model).
```

Again, the list 'Gaps' contains remaining gaps in the network. It will be helpful to have an overview of the remaining dead-end metabolites.

Test if biomass precursors can be produced in standard medium (set in Step 42)

60| Obtain the list of biomass components:

```
[BiomassComponent,BiomassFraction]=PrintBiomass(model,BiomassNumber)
```

where the biomass reaction index is provided with 'BiomassNumber'. The function returns all the biomass components ('BiomassComponent') and their corresponding fractions in the array 'BiomassFraction'. It also prints the results in the command window.

61| Add demand function for each biomass precursor ('metaboliteNameList'):

```
[modelNew,rxnNames]=addDemandReaction(model,metaboliteNameList);
```

Note that 'metaboliteNameList' should be identical to 'BiomassComponent', obtained in Step 60. A new model is returned ('modelNew'), which has additional demand reactions for every precursor whose reaction abbreviations are listed in 'rxnNames'.

62| *For each biomass component, perform the following test:* Change objective function to the demand function ('rxnName'):

```
modelNew=changeObjective function(modelNew,rxnName);
```

63| Maximize ('max') for new objective function (Demand function)

```
FBAsolution=optimizeCbModel(modelNew,'max');
```

The structure 'FBAsolution' contains the optimal solution vector ('FBAsolution.x') and also the value for the objective reaction ('FBAsolution.obj'). If it is Case 1, the model can produce biomass component (FBAsolution.obj >0), proceed with the next biomass component. If it is Case 2, the model cannot produce biomass component (FBAsolution.obj = 0). Follow Steps 64 and 65.

64| Identify reactions that are mainly responsible for synthesizing the biomass component.

65| For each of these reactions, follow the wire diagram given in **Figure 14**.

66| Test if biomass precursors can be produced in other growth media. Repeat Steps 60–65.

Test if the model can produce known secretion products

67| Collect a list of known secretion products and medium conditions.

68| *Set the constraints to the desired medium condition* (e.g., minimal medium + carbon source). For changing the constraints, use the following function:

```
model=changeRxnBounds(model,rxnNameList,value,boundType)
```

Reactions whose bounds should be changed are listed in 'rxnNameList'. The new value for each reaction is contained in the array 'value'. Finally, define the type of constraint in the list 'boundType'. The possible types are: 'l' for lower bound, 'u' for upper bound and 'b' if both reaction bounds should be set to the specified value. If the model shall be required to grow in addition to producing the by-product, set the lower bound (boundType = 'l') of the biomass reaction ('rxnNameList') to the corresponding value ('value').

```
model=changeRxnBounds(model,rxnNameList,value,boundType);
```

69| Change the objective function to the exchange reaction of your secretion product:

```
model=changeObjective(model,rxnNameList,objectiveCoeff)
```

The reaction(s) that should be set as the objective function is given by 'rxnNameList'. They will receive a corresponding coefficient 'objectiveCoeff'.

70| *Maximize ('max') for the new objective function* (as a secretion is expected to have a positive flux value, see **Fig. 7**):

```
FBAsolution=optimizeCBModel(model,'max');
```

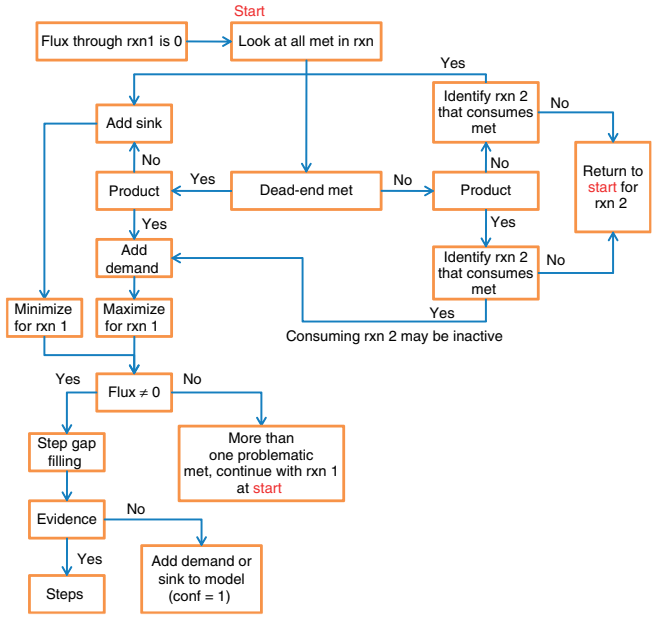


Figure 14 | Flow chart on debugging network reactions that cannot carry flux. 'rxn' stands for reaction; 'conf' stands for confidence score; and 'met' stands for metabolite.

If the product can be produced (FBAsolution.obj >0), proceed with the next by-product. If the product cannot be produced (FBAsolution.obj = 0), the corresponding pathway is missing or incomplete, and thus, gap analysis must be performed (Steps 45–49).

Test if the model can produce a certain ratio of two secretion products

71| *Set the constraints to the desired medium condition* (e.g., minimal medium + carbon source). For changing the constraints, use the following function:

```
model=changeRxnBounds(model,rxnNameList,value,boundType)
```

72| *Verify that both by-products can be produced independently*. Repeat Steps 69 and 70.

73| Add a row to the S matrix (see **Fig. 8b** for an example of an S matrix) to couple the by-product secretion reactions:

```
modelNew=AddRatioReaction (model,ListOfRxns,RatioCoeff)
```

The two reactions that should be set to a certain ratio are listed in 'ListOfRxns'. Their ratio is given in 'RatioCoeff' by listing the corresponding coefficients in this array. For example, 1:2 is given as [1 2]. If the model is required to grow while producing the by-product, then set the lower bound of the biomass reaction to a corresponding value.

```
model=changeRxnBounds(model,rxnNameList,value,boundType);
```

74| Change the objective function to the exchange reaction of one of your secretion products:

```
model=changeObjective(model,rxnNameList,objectiveCoeff)
```

75| *Maximize for the new objective function* (as a secretion is expected to have a positive flux value, see **Fig. 7**):

```
FBAsolution=optimizeCBModel(model,'max');
```

If the product can be produced (FBAsolution.obj >0), the second by-product can be produced in the defined ratio. If the product cannot be produced (FBAsolution.obj = 0, or problem is infeasible), i.e., the ratio cannot be matched. The debugging is less straightforward in this case as multiple reasons may apply. One very likely reason is that the organism (or cell) in

the experimental condition under which the ratio was determined did not grow optimally. If in Step 73 a lower bound is set on the growth rate it may cause the discrepancy (because of competition for, e.g., carbons in by-products and biomass reaction). The bound could be set lower. Alternatively, some more elaborate tools that are currently not in the COBRA Toolbox can be used to identify missing genes/reactions (**Supplementary Table 3**).

Check for blocked reactions

76| Change simulation conditions to rich medium or open all exchange reactions:

```
model=changeRxnBounds(model,rxnNameList,value,boundType)
```

Note that the value of the exchange reactions ('rxnNameList') does not matter, as this step is testing a qualitative not a quantitative property. Therefore, one can set the value to −infinity (e.g., −1,000) and +infinity (e.g., +1,000). As we are changing upper and lower bound the boundType is 'b'.

77| *Run analysis for blocked reactions*. The function returns a list of blocked reactions ('BlockedReactions').

```
BlockedReactions=FindBlockedReaction(model)
```

78| *Connect reaction to remaining network* (*optional*). This depends on the function of the blocked reaction. Follow the diagram in **Figure 14**.

Compute single-gene deletion phenotypes

79| *Compute single-gene deletion phenotypes*. Use the following function in the COBRA Toolbox:

```
[grRatio,grRateKO,grRateWT]=singleGeneDeletion(model,method,geneList)
```

This function allows the use of different methods ('method') for optimization, e.g., FBA, minimization of metabolic adjustment (MOMA)[6] or linear MOMA[16]. The list of genes that shall be deleted is given by 'geneList'. If no gene list is given or the string is empty, all genes in the reconstruction will be deleted and tested for growth capabilities of the knockout mutant. The function calculates the growth rate of the wild-type strain ('grRateWT') of each deletion strain ('grRateKO'), as well as the relative growth rate ratios ('grRatio').

80| *Compare with experimental data*. The evaluation of inconsistencies will lead to further reconstruction refinement (**Fig. 9**). Repeat the gap analysis as necessary (Steps 45–49).

Test for known incapabilities of the organism

81| *Set simulation condition*. Change objective function. Test for incapability by maximizing for objective function. If incapable, no solution or zero flux should be returned.

82| Use single-reaction deletion to identify candidate reactions that enable the model's capability despite known incapability:

```
[grRatio,grRateKO,grRateWT,hasEffect,delRxns,fluxSolution]=singleGeneDeletion(model);
```

This smaller subset of reactions needs to be manually evaluated. Note that the deletion of a single function may not be sufficient when alternate pathways exist in the network.

▲ **CRITICAL STEP** Missing incapabilities may not only be caused by falsely added reactions in the metabolic network but may be a consequence of missing regulatory information. Literature may provide the necessary data.

Test if the model can predict the correct growth rate or other quantitative properties

83| *Compare the predicted physiological properties with the known properties*. Use the suite of functions in the COBRA Toolbox along with experimental data (e.g., phenotypic, physiological and genetic data).

Test if the model can grow fast enough

84| Optimize for biomass reaction in different medium conditions and compare with experimental data. If the model does not grow at all, follow option A. If the model does not grow fast enough, follow option B.

(A) If the model does not grow at all

(i) Check your boundary constraints. If these are correct, it is possible that the simulated condition does not support growth (compare with experimental data) or the network is incomplete. In the latter case, return to Steps 45–49 to identify the missing links in the network.

(B) If the model does not grow fast enough

(i) Check your boundary constraints. If these are correct, the possibilities of error modes are quite numerous. It is advised to verify the constraints applied to the model (e.g., reaction directionality). Use the function that lists all constrained reactions that are greater than a minimal value ('MinInf') and smaller than a maximal value ('MaxInf'):

```
PrintConstraints(model,MinInf,MaxInf);
```

85| Test if any of the medium components are growth limiting. If so, increase the uptake rate ('value') of one substrate ('rxnNameList') at a time by using:

```
model=changeRxnBounds (model,rxnNameList,value,boundType)
```

and setting the bound type to lower bound 'l' ('boundType')

86| *Maximize for biomass.* If the biomass reaction value increases, it means that this compound is limiting. This gives you a hint as to where in the network something must be missing.

87| Determine the reduced cost associated with network reactions when optimizing for objective function. Use

```
FBAsolution=optimizeCbModel(model,osenseStr,primalOnlyFlag)
```

Set primalOnlyFlag to 'false' to get the reduced cost returned with the optimal solution (FBAsolution.w). When maximizing the objective function 'osenseStr' will be 'max', whereas minimization is defined by 'min'. Find the reactions with the lowest reduced cost values. Increase flux through those reactions, if possible, by removing upper bounds. This will lead to increased flux through the objective reaction.

Test if the model grows too fast

88| Optimize for biomass reaction in different medium conditions and compare with experimental data.

89| *Verify that the model constraints are set as intended.* Use the function that lists all the constrained reactions that are greater than a minimal value ('MinInf') and smaller than a maximal value ('MaxInf'):

```
PrintConstraints(model,MinInf,MaxInf);
```

Carry out one or more of the following tests to identify possible errors in the network

90| Verify that all fractions and precursors in the biomass reaction are consistent with the present knowledge. This may include that the GAM in the biomass reaction is not correct.

91| Identify shuttling reactions, e.g., proton shuttling, by repeating Steps 51–58. Thereby, one is looking for reactions associated with loops.

92| Re-investigate the thermodynamic information associated with the network reaction, i.e., reaction directionality, supporting evidence and uncertainty associated with thermodynamic data.

93| Use single-reaction deletion to identify single reactions that may enable the model to grow too fast. Use the following function by setting the 'method' to 'FBA', and the 'rxnList' should contain one or more reactions that are to be deleted. If all network reactions are to be tested, then 'rxnList' does not need to be defined:

```
[grRatio,grRateKO,grRateWT]=singleRxnDeletion(model,method,rxnList)
```

The function will return the wild-type growth rate ('grRateW'), the growth rate of the reaction-deleted network ('grRateKO') and the relative growth rate ratio ('grRatio'). However, it is most likely that multiple reactions contribute to this observation, and thus, they are not identified by this method.

94| *Reduced cost*. The reduced cost analysis can be used to identify those reactions that can reduce the growth rate (positive cost value). Use:

```
FBAsolution=optimizeCbModel(model,osenseStr,primalOnlyFlag)
```

Set primalOnlyFlag to 'false' to get the reduced cost returned with the optimal solution (FBAsolution.w). When maximizing the objective function 'osenseStr' will be 'max', whereas minimization is defined by 'min'.

▲ **CRITICAL STEP** Changes to the model may be condition specific and should be well documented.

▲ CRITICAL STEP An unconstrained NGAM reaction can change the model prediction in some cases. For example, if the computed growth rate of the model is too high, check the flux value through the NGAM reaction in the optimal solution.

Data assembly and dissemination ● TIMING **Days to weeks**

95| *Print Matlab model content*. Make the final reconstruction available to the research community in at least two formats: (1) as a spreadsheet containing all information collected during the reconstruction process (as shown in **Supplementary Methods 2**); and (2) in SBML format, which is a transportable format of the models and can be used with other modeling tools. To export the reconstruction from Matlab into Excel format, use:

writeCBmodel(model, format, FileName) where 'format' is 'xls'

To export a model in SBML format, use the same function but change the format to 'sbml'. The output file name is defined by 'FileName'.

▲ CRITICAL STEP It should be noted that the SBML format will not contain all identifiers, references and notes. It is therefore crucial to distribute the reconstruction in a different format. Ideally, the reconstruction content is made available through a web page, such as BiGG (http://bigg.ucsd.edu), which facilitates queries.

96| *Add gap information to the reconstruction output*. In Steps 45–48 information regarding the remaining and resolved network gaps was collected. These should be associated with the output of the final reconstruction (e.g., in Excel format).

● TIMING

The timing of the entire reconstruction process depends on the properties of the target organism (prokaryote versus eukaryote, genome size), the quality of the genome annotation and the availability of experimental data. The timing listed below represents an average and can be used to plan the different stages. All COBRA Toolbox functions described in this protocol finish with a couple of seconds to a few hours on a newer personal computer (Intel Core 2 Duo 6600 2.4 GHz with 4 Gb of memory running Windows Vista).

Steps 1–4 (Stage 1), Draft reconstruction: days to a week

Step 5 (Stage 1), Collect experimental data: ongoing throughout the reconstruction process

Steps 6–23 (Stage 2), Manual reconstruction refinement: months to a year (if debugging and gap filling is done along the way)

Steps 24–36 (Stage 2), Determine biomass composition: days to weeks, depending on data availability

Step 37 (Stage 2), Determine growth medium requirements: days to weeks, depending on data availability

Steps 38–42 (Stage 3), Conversion from reconstruction to mathematical model: days to a week

Steps 43–94 (Stage 4), Network evaluation = 'Debugging mode': week to months

Steps 95–96, Data assembly and dissemination: days to weeks, depending on how much and in which format data were collected.

TABLE 6 | Extract of reconstructions and their key properties that were constructed in accordance with this protocol.

Organism	Strain	Genes	Version	GR	Mets	Rxns	Comp	Ref
Bacillus subtilis		4,225	model_v3	844	988	1,020	2 (c,e)	75
Escherichia coli	K12 MG1655	4,405	iAF1260	1,260	1,039	2,077	3 (c,e,p)	65
Helicobacter pylori	26695	1,632	iIT341	341	485	476	2 (c,e)	76
Pseudomonas putida	KT2440	5,350	iNJ746	746	911	950	3 (c,p,e)	57
Pseudomonas putida	KT2440	5,350	iJP815	815	886	877	2 (c,e)	96
Pseudomonas aeruginosa	PA01	5,640	iMO1056	1,056	760	883	2 (c,e)	97
Mycoplasma genitalium	G-37	521	iPS189	189	274	262	2 (c,e)	98
Lactobacillus plantarum	WCFS1	3,009		721	531	643	2 (c,e)	73
Streptomyces coelicolor	A3(2)	8,042		700	500	700	2 (c,e)	99
Leishmania major	Friedlin	8,370	iAC560	560	1,101	1,112	8 (a,f,y,c,e,m,r,n)	100
Saccharomyces cerevisiae	Sc288	6,183	iMM904	904	713	1,412	8 (c,e,m,x,n,r,v,g)	101
Homo sapiens		28,783	Recon 1	1,496	2,766	3,311	8 (c,e,m,x,n,r,v,g)	15

Abbreviations: Comp, compartments; GR, genes in reconstruction; Mets, metabolites; Ref, reference; Rnxs, reactions. Please refer to **Supplementary Table 1** for compartment abbreviations.

A complete list of reconstructions, constructed in part or in full in accordance with this protocol, can be found at http://gcrg.ucsd.edu/In_Silico_Organisms/Other_Organisms. This website is continuously updated.

? TROUBLESHOOTING

Step 38: See installation instructions of the COBRA Toolbox[16] for details on how to install and set Matlab, SBML and COBRA Toolbox.

Step 39: The script may fail during the loading of the model from the xls files. Check:

whether the headers are correct (**Supplementary Methods 2**).

that all necessary information is available.

the metabolic reaction is written correctly → example; if there are multiple spaces in the reaction, the script does not work.

Separator for left-hand side and right-hand side can be -->, ->, <==>, <=>

Mixing numbers and strings can cause problems as well. See Ecoli_core.xls as an example on how the input file should look.

Step 51: It should be made sure that the directory in which one is working is the same wherein the X3.exe script was copied to. The .expa file produced by the function must be in the same directory as X3.exe.

ANTICIPATED RESULTS

This protocol will result in a reconstruction that covers most of the known metabolic information of the target organism and represents a knowledge database. This reconstruction can be used as a resource for information (query tool), high-throughput data mapping (context for content) and a starting point for mathematical models. **Table 6** lists a subset of published reconstructions that were constructed based on the presented protocol.

To facilitate the use of the presented COBRA Toolbox commands (Steps 43–95), we listed examples of their use in **Supplementary Methods 1**.

BOX 1 | GLOSSARY

Bibliome: A bibliome is a collection of primary and review literature as well as textbooks.

Biochemical, Genetic and Genomic (BiGG) knowledge base: A BiGG knowledge base is a genome-scale reconstruction, which incorporates in a structured manner genomic, proteomic, biochemical and physiological information of a particular organism or cell.

Biomass reaction: The biomass reaction lumps all known biomass precursors and their fractional distribution to a cell into one network reaction.

Blocked reactions: Network reactions that cannot carry any flux in any simulation condition are called blocked reactions. Generally, these blocked reactions are caused by missing links in the network.

Constraint-based reconstruction and analysis (COBRA): COBRA is a modeling approach in which manually curated, stoichiometric network reconstructions are constructed. Subsequently, models can be obtained and analyzed by applying equality and inequality constraints and by computing functional states. Constraints include mass conservation and thermodynamics (for directionality), as well as constraints reflecting experimental conditions and regulatory constraints.

Dead-end metabolite: A dead-end metabolite is only produced or consumed in the network.

Demand reaction: When the consumption reaction(s) of a metabolite is not known or outside the scope of the reconstruction it can be represented by this unbalanced, intracellular reaction (e.g., 1 A -->).

Exchange reactions: These reactions are unbalanced, extra-organism reactions that represent the supply to or removal of metabolites from the extra-organism "space". (See **Fig. 7**.)

Extreme pathways (ExPa's): ExPa's are a unique and minimal set of flux vectors, which lie at the edges of the bounded null space. Biochemically meaningful steady-state solutions can be obtained by non-negative linear combinations of ExPa's.

Flux-balance analysis (FBA): FBA is a formalism that defines the metabolic network as a linear programming optimization problem. The main constraints in FBA are imposed by the steady-state mass conservation of metabolites.

Futile cycles: Stoichiometrically unbalanced cycles, which are associated with energy consumption.

Gene–protein-reaction (GPR) association: A GPR association connects genes, proteins and reactions in a logical relationship (AND, OR).

Genome-scale model (GEM): A GEM is derived from a GENRE by converting it into a mathematical form (i.e., an *in silico* model) and by assessing its phenotypic properties computationally.

Genome-scale network reconstruction (GENRE): A GENRE is formed based on an organism-specific BiGG knowledge base. A GENRE is a collection of biochemical transformation derived from the genome annotation and the bibliome of the target organism. A network GENRE is unique to an organism, as is its genome.

Flux variability analysis (FVA): FVA is a frequently used computational tool for investigating more global capabilities under a given simulation condition (e.g., network redundancy). Therefore, every network reaction will be chosen as an objective function, and the minimal and maximal possible flux value through the reaction is determined by minimizing and maximizing the objective function.

Linear programming (LP): LP is an optimization technique, in which a linear objective function is optimized (i.e., minimized or maximized) subject to linear equality and inequality constraints.

Network gap: A network gap is a missing reaction or function in the network that can connect one or more dead-end metabolites with the remainder of the network.

(*continued*)

BOX 1 | GLOSSARY (CONTINUED)

Objective function: An objective function is a network reaction, or a linear combination of network reactions, for which a linear programming problem is optimized.
Sink reaction: When the synthesis reaction(s) of a metabolite is not known or outside the scope of the reconstruction, its discharge can be represented by this unbalanced, intracellular reaction (e.g., 1 A <-->).
P/O ratio: This ratio represents the number of ATP molecules (P), which are formed per oxygen atom (O), that are consumed during respiration.
Reduced cost: A parameter associated with linear programming. It can be used to investigate properties associated with the calculated optimal solution. Each network reaction has reduced cost values associated, which represents the amount the objective value would increase if the flux through the reaction would be increased by 1 U. Note that by definition only negative reduced cost values can increase the objective value.
Type III extreme pathway: These stoichiometric balanced cycles (SBC) are a subset of ExPa's that are only composed of intracellular reactions, i.e., that all exchange reactions (i.e., systems boundaries) have zero flux.

Note: Supplementary information is available via the HTML version of this article.

ACKNOWLEDGMENTS We would like to acknowledge R.M.T. Fleming, A. Feist and N. Jamshidi for their valuable discussions. We thank M. Abrahams, S.A. Becker and F.-C. Cheng for reading the paper. We also thank S. Burning for preparing the biomass reaction manual, as well as A. Bordbar and R.M.T. Fleming for providing Matlab code. I.T. was supported by National Institutes of Health (NIH) grant R01 GM057089.

AUTHOR CONTRIBUTION: I.T. and B.Ø.P. designed the concept and wrote the paper. I.T. developed protocol.

Published online at http://www.natureprotocols.com.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions.

1. Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z.N. & Barabasi, A.L. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**, 839–843 (2004).
2. Thiele, I., Price, N.D., Vo, T.D. & Palsson, B.O. Candidate metabolic network states in human mitochondria: impact of diabetes, ischemia and diet. *J. Biol. Chem.* **280**, 11683–11695 (2005).
3. Pal, C. *et al.* Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667–670 (2006).
4. Barrett, C.L., Herring, C.D., Reed, J.L. & Palsson, B.O. The global transcriptional regulatory network for metabolism in *Escherichia coli* attains few dominant functional states. *Proc. Natl. Acad. Sci. USA* **102**, 19103–19108 (2005).
5. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. & Palsson, B.O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96 (2004).
6. Segre, D., Vitkup, D. & Church, G.M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* **99**, 15112–15117 (2002).
7. Feist, A.M. & Palsson, B.O. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat. Biotechnol.* **26**, 659–667 (2008).
8. Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L. & Palsson, B.O. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7**, 129–143 (2009).
9. Reed, J.L., Famili, I., Thiele, I. & Palsson, B.O. Towards multidimensional genome annotation. *Nat. Rev.ws Genet.* **7**, 130–141 (2006).
10. Notebaart, R.A., van Enckevort, F.H., Francke, C., Siezen, R.J. & Teusink, B. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics* **7**, 296 (2006).
11. Durot, M., Bourguignon, P.Y. & Schachter, V. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol. Rev.* **33**, 164–190 (2009).
12. Price, N.D., Papin, J.A., Schilling, C.H. & Palsson, B. Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol.* **21**, 162–169 (2003).
13. Schilling, C.H., Edwards, J.S., Letscher, D. & Palsson, B.O. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol. Bioeng.* **71**, 286–306 (2000).
14. Varma, A. & Palsson, B.O. Metabolic flux balancing: basic concepts, scientific and practical use. *Nat. Biotechnol.* **12**, 994–998 (1994).
15. Duarte, N.C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. USA* **104**, 1777–1782 (2007).
16. Becker, S.A. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protoc.* **2**, 727–738 (2007).
17. Savinell, J.M. & Palsson, B.O. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *J. Theor. Biol.* **154**, 421–454 (1992).
18. Burgard, A.P. & Maranas, C.D. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol. Bioeng.* **82**, 670–677 (2003).
19. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* **3**, 1–15 (2007).
20. Gianchandani, E.P., Oberhardt, M.A., Burgard, A.P., Maranas, C.D. & Papin, J.A. Predicting biological system objectives *de novo* from internal state measurements. *BMC Bioinformatics* **9**, 43 (2008).
21. Papin, J.A. & Palsson, B.O. The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys. J.* **87**, 37–46 (2004).
22. Li, F., Thiele, I., Jamshidi, N. & Palsson, B.O. Identification of potential pathway mediation targets in Toll-like receptor signaling. *PLoS Comput. Biol.* **5**, e1000292 (2009).
23. Thiele, I., Jamshidi, N., Fleming, R.M. & Palsson, B.O. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.* **5**, e1000312 (2009).
24. Gianchandani, E.P., Papin, J.A., Price, N.D., Joyce, A.R. & Palsson, B.O. Matrix formalism to describe functional states of transcriptional regulatory systems. *PLoS Comput. Biol.* **2**, e101 (2006).
25. Gianchandani, E.P., Joyce, A.R., Palsson, B.O. & Papin, J.A. Functional States of the genome-scale Escherichia coli transcriptional regulatory system. *PLoS Comput. Biol.* **5**, e1000403 (2009).
26. Mobley, H.L.T., Mendz, G.L. & Hazell, S.L. *Helicobacter pylori* (ASM Press, Washington, D.C., 2001).
27. Neidhardt, F.C. (ed.) *Escherichia coli and Salmonella: Cellular and Molecular Biology* 2nd edn. (ASM Press, Washington, D.C., 1996).
28. Dickinson, J.R. & Schweizer, M. *The Metabolism and Molecular Physiology of Saccharomyces cerevisiae* 2nd edn. (Taylor & Francis Ltd, London, Philadelphia, 2004).
29. Ramos, J.L. *Pseudomonas* (Academic/Plenum Publishers, New York Kluwer, 2004).
30. Karp, P.D., Paley, S. & Romero, P. The pathway tools software. *Bioinformatics (Oxford, England)* **18** (Suppl 1): S225–S232 (2002).
31. Pinney, J.W., Shirley, M.W., McConkey, G.A. & Westhead, D.R. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. *Nucleic Acids Res.* **33**, 1399–1409 (2005).
32. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
33. Stein, L. Genome annotation: from sequence to biology. *Nat. Rev. Genet.* **2**, 493–503 (2001).
34. Aziz, R.K. *et al.* The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

35. Overbeek, R., Bartels, D., Vonstein, V. & Meyer, F. Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem. Rev.* **107**, 3431–3447 (2007).
36. Manichaikul, A. *et al.* Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat. Methods* **6**, 589–592 (2009).
37. Boneca, I.G. *et al.* A revised annotation and comparative analysis of Helicobacter pylori genomes. *Nucleic Acids Res.* **31**, 1704–1714 (2003).
38. Karp, P.D. *et al.* Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Res.* **35**, 7577–7590 (2007).
39. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
40. (NC-IUBMB), N.C.o.t.I.U.o.B.a.M.B. *Enzyme Nomenclature* 6th edn. (Academic Press, San Diego, California, 1992).
41. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354–D357 (2006).
42. Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I. & Schomburg, D. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.* **35**, D511–D514 (2007).
43. Karp, P.D. *et al.* The EcoCyc database. *Nucleic Acids Res.* **30**, 56–58 (2002).
44. Jankowski, M.D., Henry, C.S., Broadbelt, L.J. & Hatzimanikatis, V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* **95**, 1487–1499 (2008).
45. Fleming, R.M.T., Thiele, I. & Nasheuer, H.P. Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to *Escherichia coli*. *Biophys. Chem.* **145**, 47–56 (2009).
46. Kümmel, A., Panke, S. & Heinemann, M. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* **7**, 1–12 (2006).
47. Gardy, J.L. *et al.* PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics (Oxford, England)* **21**, 617–623 (2005).
48. Lu, Z. *et al.* Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics (Oxford, England)* **20**, 547–556 (2004).
49. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**, 953–971 (2007).
50. Ross-Macdonald, P. *et al.* Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418 (1999).
51. Huh, W.K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
52. Brooksbank, C., Cameron, G. & Thornton, J. The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.* **33**, D46–D53 (2005).
53. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**, D5–D12 (2007).
54. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
55. Coles, S.J., Day, N.E., Murray-Rust, P., Rzepa, H.S. & Zhang, Y. Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem.* **3**, 1832–1834 (2005).
56. Williams, A.J. Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today* **13**, 502–506 (2008).
57. Nogales, J., Palsson, B.O. & Thiele, I. A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory. *BMC Syst Biol* **2**, 79 (2008).
58. Izard, J. & Limberger, R.J. Rapid screening method for quantitation of bacterial cell lipids from whole cells. *J Microbiol Methods* **55**, 411–418 (2003).
59. Benthin, S., Nielsen, J. & Villadsen, J. A simple and reliable method for the determination of cellular RNA content. *Biotechnol. Tech.* **5**, 39–42 (1991).
60. Herbert, D., Phipps, P.J. & Strange, R.E. Chemical analysis of microbial cells. *Methods Microbiol.* **5**, 209–344 (1971).
61. Lindahl, L. & Zengel, J.M. Ribosomal genes in Escherichia coli. *Annu. Rev. Genet.* **20**, 297–326 (1986).
62. Sawada, M., Osawa, S., Kobayashi, H., Hori, H. & Muto, A. The number of ribosomal RNA genes in Mycoplasma capricolum. *Mol. Gen. Genet.* **182**, 502–504 (1981).
63. Hui, I. & Dennis, P.P. Characterization of the ribosomal RNA gene clusters in Halobacterium cutirubrum. *J. Biol. Chem.* **260**, 899–906 (1985).
64. Neidhardt, F.C., Ingraham, J.L. & Schaechter, M. *Physiology of the Bacterial Cell: A Molecular Approach* (Sinauer Associates, Sunderland, MA, USA, 1990).
65. Feist, A.M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**, 121 (2007).
66. Schilling, C.H., Letscher, D. & Palsson, B.O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248 (2000).
67. Price, N.D., Thiele, I. & Palsson, B.O. Candidate states of *Helicobacter pylori*'s genome-scale metabolic network upon application of loop law thermodynamic constraints. *Biophys. J.* **90**, 3919–3928 (2006).
68. Palsson, B.O. *Systems Biology: Properties of Reconstructed Networks* (Cambridge University Press, New York, 2006).
69. Gutnick, D., Calvo, J.M., Klopotowski, T. & Ames, B.N. Compounds which serve as the sole source of carbon or nitrogen for Salmonella typhimurium LT-2. *J. Bacteriol.* **100**, 215–219 (1969).
70. Schroeder, C., Selig, M. & Schoenheit, P. Glucose fermentation to acetate, CO 2 and H 2 in the anaerobic hyperthermophilic eubacterium Thermotoga maritima: involvement of the Embden–Meyerhof pathway. *Arch. Microbiol.* **161**, 460–470 (1994).
71. Satish Kumar, V., Dasika, M.S. & Maranas, C.D. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**, 212 (2007).
72. Reed, J.L. & Palsson, B.O. Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res.* **14**, 1797–1805 (2004).
73. Teusink, B. *et al.* Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. *J. Biol. Chem.* **281**, 40041–40048 (2006).
74. Reed, J.L. *et al.* Systems approach to refining genome annotation. *Proc. Natl. Acad. Sci. USA* **103**, 17480–17484 (2006).
75. Oh, Y.K., Palsson, B.O., Park, S.M., Schilling, C.H. & Mahadevan, R. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J. Biol. Chem.* **282**, 28791–28799 (2007).
76. Thiele, I., Vo, T.D., Price, N.D. & Palsson, B. An expanded metabolic reconstruction of Helicobacter pylori (*i*IT341 GSM/GPR): an *in silico* genome-scale characterization of single and double deletion mutants. *J. Bacteriol.* **187**, 5818–5830 (2005).
77. Feist, A.M., Scholten, J.C.M., Palsson, B.O., Brockman, F.J. & Ideker, T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol. Syst. Biol.* **2**, 1–14 (2006).
78. Famili, I., Forster, J., Nielsen, J. & Palsson, B.O. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. USA* **100**, 13134–13139 (2003).
79. Knorr, A.L., Jain, R. & Srivastava, R. Bayesian-based selection of metabolic objective functions. *Bioinformatics* (Oxford, England) **23**, 351–357 (2007).
80. Holzhutter, H.G. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.* **271**, 2905–2922 (2004).
81. Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. USA* **102**, 7695–7700 (2005).
82. Schuster, S., Pfeiffer, T. & Fell, D.A. Is maximization of molar yield in metabolic networks favoured by evolution? *J. Theor. Biol* **252**, 497–504 (2008).
83. Ott, M.A. & Vriend, G. Correcting ligands, metabolites, and pathways. *BMC Bioinformatics* **7**, 517 (2006).
84. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2008).
85. Tatusov, R.L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
86. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13–D21 (2008).
87. Jarlier, V. & Nikaido, H. Mycobacterial cell wall: structure and role in natural resistance to antibiotics. *FEMS Microbiol. Lett.* **123**, 11–18 (1994).
88. Sundararaj, S. *et al.* The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. *Nucleic Acids Res.* **32**, D293–D295 (2004).
89. Ren, Q., Chen, K. & Paulsen, I.T. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res.* **35**, D274–D279 (2007).
90. Klamt, S., Saez-Rodriguez, J. & Gilles, E.D. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst. Biol.* **1**, 2 (2007).
91. Klamt, S., Stelling, J., Ginkel, M. & Gilles, E.D. FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics (Oxford, England)* **19**, 261–269 (2003).

92. Luo, R.Y., Liao, S., Zeng, S.Q., Li, Y.X. & Luo, Q.M. FluxExplorer: a general platform for modeling and analyses of metabolic networks based on stoichiometry. *Chin. Sci. Bull.* **51**, 689–696 (2006).
93. Lee, D.Y., Yun, H., Park, S. & Lee, S.Y. MetaFluxNet: the management of metabolic reaction information and quantitative metabolic flux analysis. *Bioinformatics (Oxford, England)* **19**, 2144–2146 (2003).
94. Lee, S.Y. *et al.* Systems-level analysis of genome-scale *in silico* metabolic models using MetaFluxNet. *Biotechnol. Bioproc. Eng.* **10**, 425–431 (2005).
95. Chhabra, S.R. *et al.* Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium Thermotoga maritima. *J. Biol. Chem.* **278**, 7540–7552 (2003).
96. Puchalka, J. *et al.* Genome-scale reconstruction and analysis of the Pseudomonas putida KT2440 metabolic network facilitates applications in biotechnology. *PLoS Comput. Biol.* **4**, e1000210 (2008).
97. Oberhardt, M.A., Puchalka, J., Fryer, K.E., Martins dos Santos, V.A. & Papin, J.A. Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1. *J. Bacteriol.* **190**, 2790–2803 (2008).
98. Suthers, P.F. *et al.* A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189. *PLoS Comput. Biol.* **5**, e1000285 (2009).
99. Borodina, I., Krabben, P. & Nielsen, J. Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome Res.* **15**, 820–829 (2005).
100. Chavali, A.K., Whittemore, J.D., Eddy, J.A., Williams, K.T. & Papin, J.A. Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major. *Mol. Syst. Biol.* **4**, 177 (2008).
101. Mo, M.L., Palsson, B.O. & Herrgard, M.J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* **3**, 37 (2009).