nature biotechnology

High-throughput generation, optimization and analysis of genome-scale metabolic models

Christopher S Henry[1], Matthew DeJongh[2], Aaron A Best[3], Paul M Frybarger[2,3], Ben Linsay[4] & Rick L Stevens[4,5]

Genome-scale metabolic models have proven to be valuable for predicting organism phenotypes from genotypes. Yet efforts to develop new models are failing to keep pace with genome sequencing. To address this problem, we introduce the Model SEED, a web-based resource for high-throughput generation, optimization and analysis of genome-scale metabolic models. The Model SEED integrates existing methods and introduces techniques to automate nearly every step of this process, taking ~48 h to reconstruct a metabolic model from an assembled genome sequence. We apply this resource to generate 130 genome-scale metabolic models representing a taxonomically diverse set of bacteria. Twenty-two of the models were validated against available gene essentiality and Biolog data, with the average model accuracy determined to be 66% before optimization and 87% after optimization.

Current sequencing technology is producing thousands of sequenced genomes each year, transforming the fields of genomics and bioinformatics and increasing demand for new tools that enable high-throughput generation of functioning genome-scale metabolic models. With a functioning genome-scale metabolic model, culture conditions can be predicted[1], phenotypes can be predicted and reconciled with experimental data[2], and poorly annotated regions of the metabolic network can be identified[3]. In short, genome-scale metabolic models are central to the use of sequence data to produce detailed and quantitative predictions of organism behavior. The process of reconstructing genome-scale metabolic models has been broken down into 96 steps[4], clearly outlining its complexity and explaining in part the slow pace of creation of new models. Here we introduce the Model SEED, a web-based resource (available at http://www.theseed.org/models/) designed to speed the creation of new metabolic models by automating most of these steps. Several steps, however, are not currently amenable to automation and must still be performed manually, which is why we designate the models we create as 'draft models'. We call this resource the Model SEED because it is built upon the foundation of accurate genome annotations provided by the SEED framework for annotation and analysis[5,6]. At the core of the Model SEED is a model reconstruction pipeline (**Fig. 1** and **Supplementary Fig. 1**), which integrates and augments technologies for genome annotation[5,6], construction of gene-protein-reaction (GPR) associations, generation of biomass reactions, reaction network assembly[7], thermodynamic analysis of reaction reversibility[8,9] and model optimization[2,9,10] to generate draft genome-scale metabolic models. Whereas existing automated reconstruction methodologies only address portions of the reconstruction process[7,10–13], the Model SEED is capable of generating functioning draft metabolic models of an organism starting from an assembled genome sequence. The integration of the Model SEED pipeline with the SEED framework also enables a tight coupling between genome annotation and metabolic reconstruction that is essential for the high-throughput generation of metabolic models.

Preliminary model reconstruction

We applied the Model SEED pipeline to generate draft models for a taxonomically diverse set of 130 bacterial organisms (**Fig. 2**). In the first step of the pipeline, the genome sequences for these 130 organisms were imported into the SEED using the RAST server (http://rast.nmpdr.org/)[6], which performs gene calling and annotation of genome sequences in ~24 h. Once a genome sequence has been annotated by RAST, users can utilize powerful tools for manual curation of annotations before proceeding with the subsequent steps in the pipeline. The pipeline continues with the 'preliminary reconstruction' step, which uses the RAST annotations to generate a preliminary model for each organism (Online Methods). These preliminary models consist of a reaction network complete with GPR associations, predicted Gibbs free energy of reaction values and an organism-specific biomass reaction including nonuniversal cofactors, lipids and cell wall components. Each preliminary model network includes all reactions associated with one or more enzymes encoded in the organism's genome as well as a set of spontaneous reactions that do not require enzymatic catalysis (**Supplementary Table 1**).

The GPR associations for each reaction in the network are generated based on the genome annotations and a mapping between biochemical reactions and the standardized functional roles assigned to genes during RAST annotation[7]. This mapping is used to differentiate between cases where protein products from multiple genes form a complex to catalyze a reaction, and cases where protein products from multiple genes can independently catalyze the same reaction (Online Methods). Although these GPR

[1]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA. [2]Department of Computer Science, Hope College, Holland, Michigan, USA. [3]Department of Biology, Hope College, Holland, Michigan, USA. [4]Computer Science Department and Computation Institute, University of Chicago, Chicago, Illinois, USA. [5]Computing, Environment, and Life Sciences Directorate, Argonne National Laboratory, Argonne, Illinois, USA. Correspondence should be addressed to C.S.H. (chenry@mcs.anl.gov).

Received 22 February; accepted 30 July; published online 29 August 2010; doi:10.1038/nbt.1672

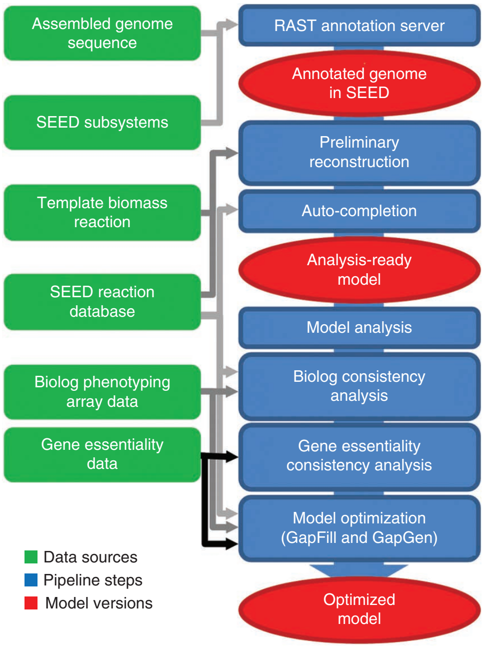

Figure 1 Model SEED genome-scale metabolic reconstruction pipeline. In the first step of the Model SEED pipeline, the assembled genome sequence is annotated by the RAST server and imported into the SEED analysis system. Next, a preliminary model is generated consisting of intracellular and transport reactions associated with genes on the basis of RAST annotations, spontaneous reactions and an organism-specific biomass reaction. In the auto-completion step of the pipeline, additional intracellular and transport reactions are added to create an analysis-ready model capable of simulating biomass production using only transportable nutrients. FBA is then used to generate phenotype predictions in the model analysis step. The final three steps of the pipeline involve the removal and addition of reactions from the model to fit Biolog phenotyping array data (when available) and gene essentiality data (when available) to produce an optimized model.

associations are generated based on well-curated annotations, they should be visually inspected to ensure accuracy.

The cofactor specificity of enzymes is also determined in the draft models based on genome annotations. For example, if an enzyme is known to use $NADP^+$ as an electron acceptor, then the functional role assigned to the associated gene will contain this information (e.g., "Non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (NADP) (EC 1.2.1.9)"); this functional role will subsequently be associated with a biochemical reaction that specifies $NADP^+$ as the cofactor. If the cofactor is unknown, then a standard cofactor is used (such as NAD^+). As annotation of cofactor specificity is often imprecise, cofactors should be visually inspected to ensure that the correct cofactors are used for the organism being modeled.

In metabolic models, biomass reactions are included to enable the simulation of cell growth and division via the simultaneous production of all small-molecule building blocks of biomass (e.g., amino acids, lipids, nucleotides and cofactors); the product of the biomass reaction is one gram of biomass, whereas the reactants are the constituent metabolites that combine to form one gram of biomass. During the preliminary reconstruction phase, the pipeline generates an organism-specific draft biomass reaction based on a reaction template (**Supplementary Table 2**). When a component of biomass is nonuniversal (e.g., cofactors, cell wall components), this template includes criteria specifying the metabolic subsystems and functional roles a genome must contain for the component to be added to the organism-specific biomass reaction. This template was tailored to produce nearly complete biomass reactions for the 130 demonstration organisms, but the nonuniversal portions of this template will require expansion to produce complete biomass reactions for some families of bacteria not included in the demonstration set (e.g., cyanobacteria). Biomass reactions must also include stoichiometric coefficients that indicate the relative abundance of each small molecule in the total biomass of an organism. Because the experimental data required to calculate these coefficients are not typically available, the Model SEED employs a set of rules to produce approximate coefficients for each biomass reaction (see Online Methods). These coefficients must be adjusted and fit to available experimental data before draft models may be used to produce quantitative predictions of organism growth rates. The approximate coefficients generated by the Model SEED are only sufficient for qualitative predictions of the conditions in which an organism will grow[14].

Automatic completion of model gaps

The models generated during the preliminary reconstruction usually contain gaps that prevent the production of one or more components of the biomass reaction. In the 'auto-completion' process, an optimization algorithm identifies the minimal set of reactions that must be added to each model to fill these gaps[10,15]. The reactions added during this process are different for each draft model, as metabolic requirements vary among organisms and different genome annotations contain different gaps. Reactions are selected from a comprehensive database of mass- and charge-balanced reactions standardized to aqueous conditions at neutral pH. This database combines all

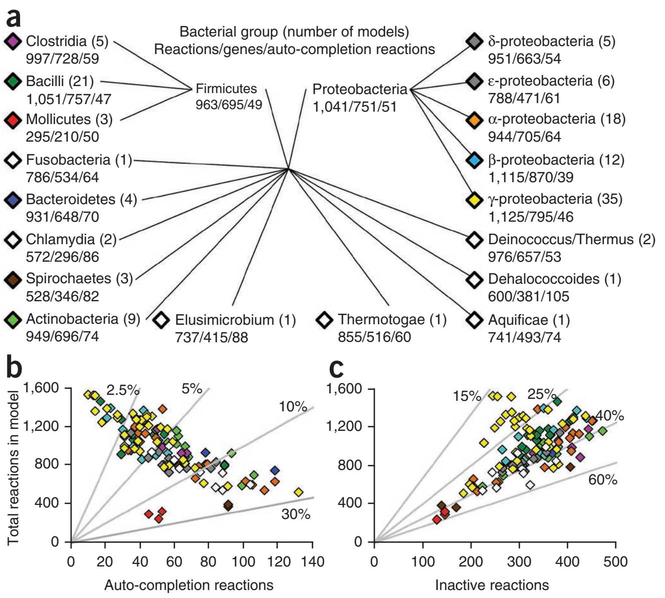

Figure 2 Properties of SEED models organized by taxonomy. (**a**) The 18 taxonomic groups containing the SEED models are displayed along with the number of models contained within each group and the average number of reactions, genes and auto-completion reactions included within the group models. The tree is arranged such that closely related taxonomic groups are co-localized. (**b**,**c**) Total number of reactions in each SEED model plotted against the number of reactions added during the auto-completion process (**b**) and the number of reactions that are inactive in FBA (**c**). Each point corresponds to a single SEED model, and the points are color coded by the taxonomic groups listed in **a**.

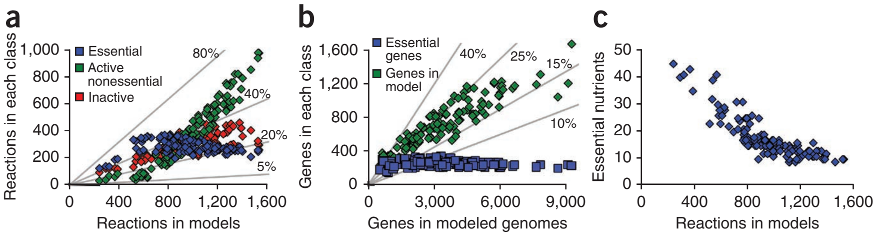

Figure 3 Properties of SEED models predicted using FBA. (**a**) FBA was used to classify each reaction in each model as essential (blue), nonessential but capable of carrying flux (green) or incapable of carrying flux (red), and the number of reactions in each class was plotted against the total number of reactions in each SEED model. (**b**) Gene essentiality in the SEED models, as predicted using FBA, compared with the total number of genes included in the genome of each modeled organism. Number of essential genes, blue; number of genes in model, green. Lines indicate the percentage of total model reactions (**a**) or organism genes (**b**) that was captured in each region of the plots. (**c**) Number of essential nutrients that are required for growth of each SEED model, as predicted by FBA, compared with the total number of reactions for each model.

the biochemistry contained in the KEGG[16,17] and 13 published genome-scale metabolic models[10,18–29] into a single, nonredundant set. Because this database is standardized at neutral pH, model reactions may require adjustment when intracellular conditions deviate significantly from the standard. The auto-completion process ensures that every SEED model is capable of simulating cell growth, and it produces a list of metabolic functions predicted to be missing from the genome annotations (**Supplementary Table 3**).

When applied to our set of 130 demonstration organisms, the auto-completion process added an average of 56 reactions to each model (**Supplementary Table 3**). In general, the number of reactions added during auto-completion increased as the total number of reactions in the model decreased (**Fig. 2b**). One explanation for this trend is that many of the smaller models are associated with endosymbiotic or pathogenic organisms, which depend upon host cells to perform many metabolic functions. As a result, these organisms import many essential metabolites rather than synthesizing them *de novo*, and poorly annotated transporters are often missing from preliminary reconstructions. For example, our model of the endosymbiont *Buchnera aphidicola*, which consisted of only 517 reactions before auto-completion, required the largest number of auto-completion reactions (132 reactions). Many of the reactions added involve the transport of essential metabolites for which biosynthesis pathways appear to be lacking. Some of the intracellular reactions added represent metabolic functions that are predicted to be missing from the *B. aphidicola* annotations. However, most represent metabolic functions that are provided to *B. aphidicola* by its host (e.g., lipopolysaccharide biosynthesis pathways)[30]. This result demonstrates how functions added during the auto-completion process suggest hypotheses about metabolic interactions between obligate intracellular organisms and their hosts.

The auto-completion results also enable the identification of regions of the metabolic network where gaps in the genome annotations for the 130 organisms appear to be most prevalent. Over 50% of the reactions added to the SEED models during the auto-completion process are associated with metabolic processes involved in either cofactor biosynthesis (ubiquinone biosynthesis, menaquinone and phylloquinone biosynthesis and thiamin biosynthesis) or cell wall biosynthesis (LOS core oligosaccharide biosynthesis, teichoic and lipoteichoic acids biosynthesis and KDO2-lipid A biosynthesis). This explains the notable exception, involving the three mollicute models (red points in **Fig. 2b**), to the inverse relationship between the number of auto-completion reactions and the model size. Because these mollicutes lack a cell wall, none of the cell wall biosynthesis reactions were added during the auto-completion process.

As a case study for how auto-completion results can drive the improvement of genome annotations, we performed a directed search to identify genes responsible for a reaction added to the *Mycobacterium tuberculosis* H37Rv model during the auto-completion process (namely, *2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase (EC 4.2.99.20)*). This reaction performs an essential step of the menaquinone biosynthesis pathway. In our directed search, all genes annotated with this reaction in other genomes were identified, and BLASTP was used to search for homologs for these genes in the *M. tuberculosis* genome. One such gene, *Rcas_1310* in *R. castenholzi*, was found to be homologous with *Rv0554* in *M. tuberculosis* with an e-value of 2.2×10^{-8}. The *Rv0554* gene clusters on the *M. tuberculosis* genome with six other genes involved in the menaquinone biosynthesis pathway, lending additional confidence to this functional assignment. To our knowledge, this is the first time the *4.2.99.20* activity has been associated with a gene in *M. tuberculosis*, and this association fills an important gap in a required metabolic pathway.

Model analysis

We call the draft models generated by the preliminary reconstruction and auto-completion processes 'analysis-ready' because they can simulate the production of biomass from transportable nutrients. On average, our 130 demonstration models include 965 reactions (**Fig. 3a**), 688 genes (**Fig. 3b**) and 876 metabolites. In the 'model analysis' step of the Model SEED pipeline, flux variability analysis (FVA)[31] is used to classify the reactions in the SEED models as essential, active or inactive (**Figs. 2c** and **3a**). Reactions classified as inactive cannot carry flux during simulated growth and are indicative of gaps in the metabolic network where additional manual curation is required. In the 130 SEED models, the average fraction of inactive reactions is 31.7% (**Fig. 2c**); not surprisingly, this is larger than the fraction of inactive reactions typically found in manually refined published models (16%). The remaining reactions that are not inactive in the SEED models are classified as either essential (if they must carry flux for growth to occur) or active (if they can carry flux but aren't essential for growth). The smaller SEED models tend to have fewer essential reactions (**Fig. 3a**), which is likely a result of metabolites being imported rather than synthesized and of biomass reactions involving fewer cofactors.

Flux balance analysis (FBA) is also used in the model analysis process to predict the essential genes in the SEED models. Despite wide variations in the genome sizes of our demonstration organisms, the number of essential metabolic genes remained relatively constant around an average value of 237 (**Fig. 3b**). This result implies that bacteria with larger genomes do not maintain redundant copies of essential genes to improve robustness. This conclusion is supported by previous studies[32], which reveal that larger genomes include a greater fraction of genes encoding secondary metabolic functions, transcriptional control and signaling mechanisms to improve versatility. Although the number of predicted essential genes remained relatively constant across all models, the specific reactions associated with these genes varied substantially. Only 47 reactions were associated with essential genes in nearly every model analyzed,

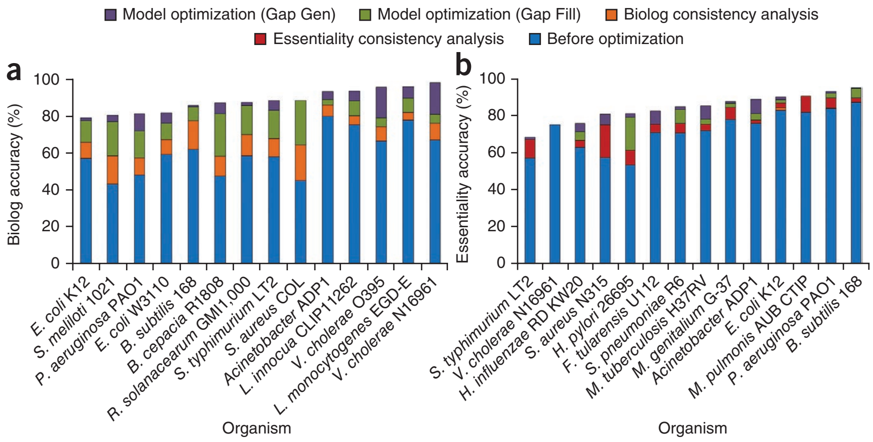

Figure 4 Accuracy of models generated by the Model SEED pipeline. (**a**,**b**) The accuracy of the SEED models in predicting Biolog phenotyping array data (**a**) and gene essentiality data (**b**) steadily improved during the model-refining steps of the pipeline. Before optimization (blue bars), the SEED models had an average overall accuracy of 66%; this increased to 71% after the Biolog consistency analysis (orange bars), to 75% after the gene essentiality consistency analysis (red bars), to 83% after the GapFill stage of the model optimization (green bars) and to 88% after the GapGen stage of the model optimization (purple bars). The gene essentiality consistency analysis affected only the GPR associations in the models, so it did not affect the accuracy of the Biolog phenotyping array predictions.

whereas 740 reactions were associated with essential genes in fewer than ten models analyzed (**Supplementary Table 4**).

Flux balance analysis is also used in the model analysis step to identify the nutrients that are essential for growth in each SEED model. In general, the number of essential nutrients decreases as the number of reactions in the models increases (**Fig. 3c**). Although defined growth conditions are unknown for many of the modeled organisms, this analysis reveals a wide range of predicted nutrient requirements. These predictions are invaluable to efforts to culture these organisms in defined media conditions[1]. All predictions generated from the SEED models are available on the Model SEED website.

Comparison with existing models and phenotype data

Biolog phenotyping arrays[18,20,21,33–35] and gene essentiality data sets[36,37] are available for 22 of the 130 demonstration organisms, and these data sets were used to validate and optimize the models for these organisms (**Fig. 4**). After the auto-completion process, the models had an average predictive accuracy of 60% for Biolog data, 72% for essentiality data and 66% overall (blue bars in **Fig. 4**). A modified version of the Growmatch algorithm[2] was included in the Model SEED pipeline to identify and correct the possible errors in the models that cause the incorrect predictions. This model optimization process consists of four steps: (i) Biolog consistency analysis to identify missing transport reactions; (ii) gene essentiality consistency analysis to identify conflicts between GPR relationships and essentiality data; (iii) gap filling to identify overconstrained or missing reactions; and (iv) gap generation to address underconstrained or extra reactions. These four optimization steps improved the average accuracy of the SEED models to 89% for Biolog data, 85% for essentiality data and 87% overall (**Fig. 4**, **Supplementary Methods** and **Supplementary Tables 5–7**). No genome-scale metabolic models have been published for eight of the organisms with available Biolog data and four of the organisms with available gene essentiality data. Nonetheless, the draft models of these organisms are as accurate as the draft models of organisms for which published models do exist (**Table 1**).

Genome-scale models have already been published for 19 of the organisms selected for metabolic reconstruction by the Model SEED pipeline[10,18–29]. Comparison of SEED models with their published counterparts shows that, on average, 86% of the genes in the published models are also included in the SEED models (**Supplementary Methods** and **Supplementary Table 8**). Most genes found exclusively in the published models were not included in the SEED models because either the functions assigned to these genes in the SEED are inconsistent with the reactions mapped to them in the published models or the functions are not specific enough to allow for mapping to explicit reactions. One example of additional content included in the SEED models that was not included in the published models is the sedoheptulose bisphosphate bypass in *Escherichia coli*. This bypass, exclusively in the SEED *E. coli* model, converts D-erythrose

Table 1 Prediction accuracy of SEED models

Organism	Published model exists	Biolog accuracy (%)		Essentiality accuracy (%)	
		Original	Optimized	Original	Optimized
B. cepacia R1808	No	47.5	87.3	–	–
E. coli W3110	No	59.3	81.8	–	–
F. tularensis U112	No	–	–	70.9	82.5
L. innocua CLIP11262	No	75.5	93.8	–	–
L. monocytogenes EGD	No	77.8	96.0	–	–
M. pulmonis AUB CTIP	No	–	–	81.8	90.5
R. solanacearum GMI	No	58.6	87.7	–	–
S. meliloti 1021	No	43.2	80.6	–	–
S. pneumoniae R6	No	–	–	70.6	84.8
V. cholerae N16961	No	67.1	98.2	75.0	75.0
V. cholerae O395	No	66.5	95.7	–	–
New model average	No	61.9	90.1	74.6	83.2
Acinetobacter ADP1	Yes	80.0	93.3	75.7	88.8
B. subtilis 168	Yes	62.0	86.0	87.2	95.0
E. coli K12	Yes	57.1	79.3	82.7	89.9
H. influenzae RD KW20	Yes	–	–	62.9	75.7
H. pylori 26695	Yes	–	–	53.2	80.9
M. genitalium G-37	Yes	–	–	77.7	87.5
M. tuberculosis H37RV	Yes	–	–	71.9	85.1
P. aeruginosa PAO1	Yes	48.1	81.5	83.9	92.9
S. aureus COL	Yes	45.2	88.7	–	–
S. aureus N315	Yes	–	–	57.3	80.6
S. typhimurium LT2	Yes	58.0	88.6	57.0	68.2
Models with published counterpart average	Yes	58.4	86.2	71.0	84.5

Empty elements in the table indicate a lack of Biolog or essentiality data for the corresponding organism.

Table 2 Example uses of the Model SEED resource

Research question	Unique capability of Model SEED	Insights or results generated
What are the essential genes in my newly sequenced organism?	Functioning draft models enable essential genes to be predicted.	357 correctly predicted essential genes in four microbes not previously modeled, and 30,316 essential genes predicted in all models
What defined culture conditions will my organism grow in?	Functioning metabolic models enable culture conditions to be predicted.	1,391 Biolog growth conditions correctly predicted in eight microbes not previously modeled, and essential nutrients predicted for all models
What are some global trends in microbial metabolic behavior?	Functioning draft models for many diverse microbes enable the exploration of such trends.	**Figures 2** and **3** show global trends in gene essentiality, reaction activity, essential nutrients and annotation gaps.
How accurate are the annotations for my organism of interest?	Functioning models convert annotations into predictions of experimentally observable phenotypes.	**Figure 4** shows the accuracy of models generated from annotations for 22 organisms based on comparison with experimentally observed phenotypes.
What are the knowledge gaps in genome annotation in general?	Recurring annotation gaps can be identified by comparing gaps found in every model.	Cofactor biosynthesis and cell wall biosynthesis account for 50% of annotation gaps found (**Supplementary Tables 2** and **6**).
What alternative pathways are present in an organism's metabolic reaction network?	Comprehensive reaction database, functional role mappings and updated annotations enable identification of alternative pathways.	Sedoheptulose bisphosphate bypass identified in the pentose phosphate pathway of *E. coli*, which is unique to the SEED *E. coli* model. Bypass in *E. coli* experimentally confirmed in ref. 38.
How can I identify and fill the gaps in my genome annotations?	Directed searches may be performed for functions added during model auto-completion and optimization.	Auto-completion process identified EC 4.2.99.20 as missing in *M. tuberculosis*, and a directed search identified peg.554 (Rv0554) as a candidate for this function.

4-phosphate and dihydroxyacetone phosphate to D-sedoheptulose 7-phosphate in the pentose phosphate pathway. It has been experimentally demonstrated to exist in transaldolase-deficient *E. coli* mutants[38] and is associated with the secondary activities of two glycolytic enzymes (6-phosphofructokinase and fructose-bisphosphate aldolase).

Manual curation

When comparing the steps in the Model SEED pipeline (**Fig. 1**) with the steps outlined in the published metabolic reconstruction protocol[4], we found that the pipeline replicates 73 of the first 82 steps in the protocol. The preliminary reconstruction step of the pipeline automates most of the first 42 steps of the protocol. The only steps missing are experimental data collection, assigning gene and reaction localization (mostly for eukaryotic models), addition of intracellular transport reactions (SEED models only include cytosol and extracellular compartments), determination of biomass reaction coefficients and loading models into the COBRA toolbox. The auto-completion and model analysis portions of the Model SEED pipeline automate all of the protocol steps 43–66 and 67–80, respectively, with the only exception being reconnection of inactive reactions. The Model SEED does not attempt to reconnect inactive reactions because this requires manual curation to differentiate the inactive reactions that are a result of misannotation from those that should be reconnected. The model optimization process implemented in the Model SEED corresponds with steps 81–82 of the published protocol.

The models produced by the Model SEED still require some manual curation before they can match most published models in quality and accuracy. We have included a tutorial on this curation process within the Model SEED website and in the **Supplementary Methods**. The infrastructure provided in the Model SEED facilitates this curation process by providing a functioning draft model with testable predictions, enabling validation of models with experimental data and supporting comparison of models in the Model SEED database (including many published models). We are also developing tools to directly support the iterative refinement of draft models within the Model SEED website.

DISCUSSION

Here we demonstrate the Model SEED as a resource for the generation, optimization and analysis of draft genome-scale metabolic models for 130 taxonomically diverse bacteria. Unlike existing resources such as KEGG[16,17] or MetaCyc[39] that focus on cataloging gene functions, metabolic reactions and pathways, the Model SEED produces functioning metabolic models that not only describe what pathways are present but also predict how those pathways are used by each organism. These unique capabilities make the Model SEED a valuable resource for numerous applications in biology (**Table 2**). The model validation (**Fig. 4b** and **Supplementary Table 9**) and large-scale gene essentiality predictions (**Fig. 3a**) demonstrate that SEED models can correctly identify many essential metabolic genes. The Biolog prediction validation (**Fig. 4a**) and essential nutrient predictions (**Fig. 3c**) demonstrate that culture conditions can be predicted. The model optimization (**Fig. 4**) and auto-completion results (**Fig. 2**) show how the Model SEED is useful as a means of assessing annotation quality. And the global trends in model predictions and statistics (**Figs. 2** and **3**) demonstrate an ability to study universal trends in microbial behavior. By providing biologists with a means of rapidly producing a functioning draft metabolic model for an organism with the click of a button, the Model SEED makes genome-scale metabolic models more accessible to the wider scientific community. The Model SEED also enables the rapid rebuilding of models to integrate improved annotations and new experimental data. Rapid update of genome-scale metabolic models is essential for keeping up with the emergence of new high-throughput experimental data sets and for enabling researchers worldwide to benefit from new discoveries in organism metabolism.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by the US Department of Energy under contract DE-ACO2-06CH11357, by the National Institute of Allergy and Infectious Diseases under contract HHSN266200400042C and by the National Science Foundation under grants MCB-0745100, CCF-0829929 and DBI-0850546. M.D. was also supported by the Argonne National Laboratory Guest Faculty Program and the United States Fulbright Scholarship Program. We acknowledge the SEED annotators and development team for producing the annotations and computational infrastructure that make this work possible. We thank R. Overbeek, V. Vonstein, R. Olson, T. Disz, S. Devoid, F. Xia and T. Paczian for assistance with the use of the SEED genome annotation and analysis tools.

AUTHOR CONTRIBUTIONS

C.S.H. developed and operated the Model SEED resource and the related algorithms with assistance from all other authors. All authors participated in curation and testing of the SEED models. M.D. contributed significantly to reaction network annotation, model testing and model curation. A.A.B. contributed significantly to template biomass reaction design, model curation and selection of organisms for reconstruction. P.M.F. contributed to reaction network annotation. R.L.S. contributed to model curation and the design of pipeline and algorithms. B.L. contributed to Model SEED web interface development. All authors contributed to the writing and revision of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Yus, E. *et al.* Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326**, 1263–1268 (2009).
2. Kumar, V.S. & Maranas, C.D. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput. Biol.* **5**, e1000308 (2009).
3. 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).
4. Thiele, I. & Palsson, B. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121 (2010).
5. Overbeek, R., Disz, T. & Stevens, R. The SEED: A peer-to-peer environment for genome annotation. *Commun. ACM* **47**, 46–51 (2004).
6. Aziz, R.K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
7. DeJongh, M. *et al.* Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics* **8**, 139 (2007).
8. 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).
9. Henry, C.S., Zinner, J., Cohoon, M. & Stevens, R. *i*Bsu1103: a new genome scale metabolic model of *B. subtilis* based on SEED annotations. *Genome Biol.* **10**, R69 (2009).
10. Suthers, P.F. *et al.* A genome-scale metabolic reconstruction of *Mycoplasma genitalium, i*PS189. *PLOS Comput. Biol.* **5**, e1000285 (2009).
11. 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).
12. Tsoka, S., Simon, D. & Ouzounis, C.A. Automated metabolic reconstruction for *Methanococcus jannaschii*. *Archaea* **1**, 223–229 (2004).
13. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).
14. Pramanik, J. & Keasling, J.D. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnol. Bioeng.* **60**, 230–238 (1998).
15. Satish Kumar, V., Dasika, M.S. & Maranas, C.D. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**, 212 (2007).
16. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
17. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
18. 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).
19. Reed, J.L., Vo, T.D., Schilling, C.H. & Palsson, B.O. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biol.* **4**, R54 (2003).
20. Durot, M. *et al.* Iterative reconstruction of a global metabolic model of *Acinetobacter baylyi* ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst. Biol.* **2**, 85 (2008).
21. 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).
22. Goelzer, A. *et al.* Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Syst. Biol.* **2**, 20 (2008).
23. Schilling, C.H. *et al.* Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**, 4582–4593 (2002).
24. Oliveira, A.P., Nielsen, J. & Forster, J. Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol.* **5**, 39 (2005).
25. Feist, A.M., Scholten, J.C., Palsson, B.O., Brockman, F.J. & Ideker, T. Modeling methanogenesis with a genome-scale metabolic reconstruction of. *Methanosarcina barkeri*. *Mol. Syst. Biol.* **2**, 2006 0004 (2006).
26. Jamshidi, N. & Palsson, B.O. Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain *i*NJ661 and proposing alternative drug targets. *BMC Syst. Biol.* **1**, 26 (2007).
27. Nogales, J., Palsson, B.O. & Thiele, I. A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: *i*JN746 as a cell factory. *BMC Syst. Biol.* **2**, 79 (2008).
28. Duarte, N.C., Herrgard, M.J. & Palsson, B.O. Reconstruction and validation of *Saccharomyces cerevisiae i*ND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**, 1298–1309 (2004).
29. Becker, S.A. & Palsson, B.O. Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol.* **5**, 8 (2005).
30. Douglas, A.E. Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera*. *Annu. Rev. Entomol.* **43**, 17–37 (1998).
31. Mahadevan, R. & Schilling, C.H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276 (2003).
32. Konstantinidis, K.T. & Tiedje, J.M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. USA* **101**, 3160–3165 (2004).
33. von Eiff, C. *et al.* Phenotype microarray profiling of *Staphylococcus aureus* menD and hemB mutants with the small-colony-variant phenotype. *J. Bacteriol.* **188**, 687–693 (2006).
34. Bochner, B.R. Global phenotypic characterization of bacteria. *FEMS Microbiol. Rev.* **33**, 191–205 (2009).
35. Keymer, D.P., Miller, M.C., Schoolnik, G.K. & Boehm, A.B. Genomic and phenotypic diversity of coastal *Vibrio cholerae* strains is linked to environmental factors. *Appl. Environ. Microbiol.* **73**, 3705–3714 (2007).
36. Gerdes, S. *et al.* Essential genes on metabolic maps. *Curr. Opin. Biotechnol.* **17**, 448–456 (2006).
37. Zhang, R., Ou, H.Y. & Zhang, C.T. DEG: a database of essential genes. *Nucleic Acids Res.* **32**, D271–D272 (2004).
38. Nakahigashi, K. *et al.* Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol. Syst. Biol.* **5**, 306 (2009).
39. Karp, P.D., Riley, M., Paley, S.M. & Pellegrini-Toole, A. The MetaCyc Database. *Nucleic Acids Res.* **30**, 59–61 (2002).

ONLINE METHODS

The Model SEED pipeline consists of seven consecutively applied steps (Fig. 1): (i) annotation; (ii) preliminary reconstruction; (iii) auto-completion; (iv) FBA analysis; (v) Biolog consistency analysis; (vi) gene essentiality consistency analysis; and (vii) reaction network optimization. Each of these steps is described in detail below.

Model SEED reconstruction pipeline: preliminary reconstruction. The Model SEED reconstruction pipeline produces analysis-ready genome-scale metabolic models starting with high-quality genome annotation in the context of the SEED framework, and optimizes them when phenotype and gene essentiality data are available (**Fig. 1**). In the first step of this pipeline, the assembled genome sequence is annotated by the RAST server and imported into the SEED. In the second step, a preliminary metabolic model is constructed consisting of (i) the spontaneous reactions, enzymatic reactions and transport reactions that make up an organism's metabolism; (ii) the set of GPR relationships that describe how reaction activity depends upon an organism's genes; and (iii) a biomass reaction that describes the essential small-molecule building blocks of the organism. Enzymatic intracellular and transmembrane transport reactions are included in the preliminary model if one or more of the functional roles associated with these reactions in the SEED (http://www.theseed.org/models/) have been assigned to one or more of the genes in the annotated genome. The functional role-to-reaction mappings in the SEED are used to construct the GPR relationships that encode how genes work together to form the protein complexes that catalyze enzymatic reactions. Additionally, if neighboring nonhomologous genes are associated with the same reaction in a model, the protein products for these genes are also assumed to function together in a single enzyme complex. These GPR relationships are essential for correctly predicting the impact of gene knockout on organism viability and behavior by using a genome-scale model. The biomass reaction in the preliminary model is assembled based on the template biomass reaction in the SEED (**Supplementary Table 2**), which was constructed from a curation of the biomass reactions included in 19 existing genome-scale metabolic models[10,18–29]. The template biomass reaction includes 83 small-molecule reactants, 39 of which are universal building blocks included in the biomass reaction of every organism (e.g., nucleotides for RNA and amino acids for protein). The remaining 44 reactants are included in a subset of the biomass reactions based on specific criteria that must be satisfied by evidence available in the annotated genome. These criteria include cell wall type (Gram positive, Gram negative, other) and subsystem variant codes that indicate specifically how an organism implements certain metabolic functions.

The stoichiometric coefficients in biomass reactions typically indicate the relative abundance of each small-molecule building block in an organism's biomass. Model SEED uses the following rules to generate stoichiometric coefficients that very roughly approximate the relative abundance of biomass components in each modeled organism: (i) relative abundances for amino acids, nucleotides, protein, DNA, RNA and cofactors are based on measured values in *E. coli*[18] for gram-negative organisms and *Bacillus subtilis*[21] for gram-positive organisms; (ii) a growth-associated ATP maintenance of 60 mmol per gram biomass per hour is assumed, which is approximately the value used in genome-scale models published to date; (iii) all cofactors are assumed to be present in equal mass; and (iv) the net mass of all biomass components sums to one gram.

Model SEED reconstruction pipeline: auto-completion. The preliminary metabolic models assembled during the second step of the Model SEED pipeline typically contain gaps in their reaction networks that prevent the production of one or more essential building blocks in the biomass reaction. As a result of these gaps, preliminary models are incapable of simulating cell growth under any conditions. In the third step of the Model SEED pipeline, these gaps are identified and eliminated through a process called auto-completion. In the auto-completion process, an optimization is performed to identify the minimal set of new reactions that must be added to the preliminary model to enable the production of biomass in the minimal confirmed growth medium for the modeled organism (**Supplementary Table 3**). If the minimal confirmed growth medium for an organism is unknown, any transportable metabolite is allowed to be consumed from the medium during the auto-completion process. The reactions added during the auto-completion process are selected from a comprehensive database of spontaneous reactions, enzymatic reactions and trans-membrane transport reactions maintained as a part of the SEED. This database consists of ~12,000 reactions and 15,044 compounds, and it combines all the biochemistry contained in the KEGG[16,17] and 13 published genome-scale metabolic models[10,18–29] into a single, nonredundant set. Often the gaps in the reaction network of a preliminary model may be filled by many different distinct sets of reactions. Equation (1) shows the novel objective function used in the auto-completion optimization to select for the set of reactions that represents the best possible hypothesis of what is actually missing from the genome annotations.

$$\text{Minimize}\sum_{i=0}^{R}\left(1 + P_{T,i} + P_{K,i} + P_{SS,i} + P_{F,i} - f_{SS,i} - f_{p,i}\right)z_i \qquad (1)$$

In this objective function, z_i is a binary variable created for any reaction not currently included in the model. Separate z_i variables are created for the forward and reverse directions of each reaction, and if a reaction included in the model is irreversible, a z_i variable is introduced for the direction of the reaction not included in the model. Thus auto-completion solutions also involve making some existing reactions in the model reversible.

$P_{T,i}$ is a penalty on the addition of transport reactions during the auto-completion process. This penalty equals 4 for transport reactions involving compounds in the biomass reaction, 2 for all other transport reactions and 0 for intracellular reactions. This penalty ensures that completion of intracellular biosynthesis pathways is favored over the addition of transport reactions. $P_{K,i}$ is a penalty favoring addition of KEGG reactions. This penalty equals 0 for KEGG reactions and 2 for non-KEGG reactions. Addition of KEGG reactions is favored to avoid the addition of simplified lumped reactions included in many existing models. $P_{SS,i}$ is a penalty favoring the addition of reactions mapped to SEED functional roles and subsystems. This penalty equals 0 if the reaction is mapped to at least one functional role in a SEED subsystem, 1 if the reaction is mapped to at least one functional role not found in a subsystem and 3 if the reaction is not mapped to any functional roles. Reactions mapped to SEED functional roles and subsystems are favored because these reactions take part in the core pathways of metabolism and represent the most well-curated portion of the known biochemistry. $P_{f,i}$ is a penalty on the addition of reactions proceeding in a thermodynamically unfavorable direction. This penalty equals 0 if a reaction is proceeding in a favorable direction[9], and $5 + 0.1(\Delta_r G'^{\circ} - 10$ kcal/mol) if the reaction is proceeding in an unfavorable direction, where $\Delta_r G'^{\circ}$ is the estimated Gibbs free energy change of reaction[8]. If $\Delta_r G'^{\circ}$ cannot be calculated, $P_{f,i}$ equals 6 for unfavorable reactions. $f_{ss,i}$ is a bonus applied to reactions involved in subsystems already well represented in the preliminary model. $f_{ss,i}$ is equal to the number of reactions in the preliminary model associated with the subsystem over the total number of reactions in the database associated with the subsystem. Similarly, $f_{p,i}$ is a bonus applied to reactions involved in short linear pathways (called scenarios[7]) already well represented in the preliminary model. $f_{p,i}$ is equal to the number of reactions in the preliminary model associated with the scenario over the total number of reactions in the database associated with the scenario.

The auto-completion objective is combined with the following set of constraints to form a complete mixed integer linear optimization problem (MILP), which may then be solved directly using the CPLEX 11.1 optimization package typically in a few hours and nearly always in <24 h:

$$N_{Super} \bullet v = 0 \qquad (2)$$

$$0 \le v_i \le 1{,}000 z_i \; i = 1,\ldots,r \qquad (3)$$

$$v_{bio} > 10^{-3}\ \text{g/g CDW h} \qquad (4)$$

In the auto-completion optimization, equation (2) represents the mass balance constraints that enforce the quasi-steady-state assumption of FBA, N_{Super} is the stoichiometric matrix for the superset of KEGG/model reactions with reversible reactions decomposed into separate forward and backward components and v is the vector of fluxes through the superset reactions.

Equation (3) enforces the bounds on the reaction fluxes (v_i) and the values of the reaction use variables (z_i). Equation (4) forces the flux through the biomass reaction, v_{bio}, to a nonzero value, ensuring that the z_i variables associated with the reactions needed to enable model growth are set to 1 during the optimization. Once the auto-completion optimization has produced a set of z_i and v_i values that optimally satisfy all constraints, the reactions with z_i values equal to 1 are added to the preliminary model to produce an analysis-ready model. The abbreviation CDW in the units of equation (4) stands for cell dry weight.

Model SEED reconstruction pipeline: analysis-ready model optimization. The remaining steps of the Model SEED pipeline involve the optimization of the analysis-ready model to better fit any experimental growth phenotype data that are available. Because these steps of the pipeline require data for fitting, they can be applied only to those organisms for which experimental data exist. The first optimization step of the pipeline, called Biolog consistency analysis, is performed only for organisms with available Biolog phenotyping array data[34]. In this step, the list of nutrients for which transport reactions exist in the model is compared against the list of nutrients the organism is known to metabolize based on available Biolog phenotyping array data. If no transport reaction exists in the model for a nutrient that is known to be metabolized, the transport reaction associated with the nutrient is added to the model. Because the transport reactions added in this process are not associated with a gene, it is impossible to discern the specific mechanism used to drive the movement of the nutrient across the cell membrane. Therefore, every transport reaction added to the SEED models during the Biolog consistency analysis follows a mechanism of proton symport for negative ions and proton antiport for positive ions.

The second optimization step of the pipeline, called gene essentiality consistency analysis, is performed only for organisms with available gene essentiality data. In this step, the data are used to identify and correct errors in annotations and GPR relationships included in the analysis-ready model. An algorithm is used to automatically search for instances of inconsistency between model annotations and available gene essentiality data. Three types of inconsistency are examined during the consistency analysis: (i) identical functional roles are assigned to an essential gene and one or more nonessential genes, (ii) identical functional roles are assigned to multiple essential genes without indicating that the protein products of these genes form a complex and (iii) one or more essential genes and one or more nonessential genes are all annotated to encode portions of the same protein complex. Once inconsistent annotations are identified, they are grouped by associated metabolic function, and a variety of annotation corrections are automatically proposed. Proposed corrections are then manually reviewed for implementation in the model.

The third optimization step in the pipeline, called model optimization, involves using the GrowMatch algorithm[2] with additional global optimization steps as described[9]. The model optimization proceeds in two stages: (i) GapFill to correct errors in the model that prevent growth *in silico* when growth is observed *in vivo* (false-negative predictions) and (ii) GapGen to correct errors in the model that allow growth *in silico* when growth is not observed *in vivo* (false-positive predictions). In the GapFill stage, a series of mixed integer linear optimization problems (MILPs) is solved to produce a set of possible solutions. Each solution represents a minimal set of modifications to the model reaction network that results in a maximal reduction in false-positive predictions. The modifications proposed by the GapFill algorithm include the addition of new reactions to the model reaction network or switching an existing reaction from being irreversible to being reversible. The most physiologically reasonable solution is then manually identified for implementation in the refined model.

The GapGen stage of the model optimization is similar to the GapFill stage in that a series of MILPs is solved to produce a small number of solutions, one of which is manually selected for implementation to maximally reduce prediction errors. In the GapGen stage, however, false-positive predictions are eliminated, and reactions are made irreversible or removed entirely rather than being added. The GapGen stage of the model optimization provides a valuable means of identifying reactions in the models that were underconstrained by the reversibility prediction method used.

Model validation using FBA. FBA is first used in the Model SEED pipeline to verify that every model produced by the pipeline is ready for analysis, by confirming that the model is capable of simulating biomass production in the minimal defined growth medium for the modeled organism. If no minimal defined growth medium is known for the organism, FBA is used to ensure that the model is capable of simulating biomass production using only nutrients for which transmembrane transport reactions exist in the model.

In the assessment and optimization of the SEED models, FBA is used to calculate the maximum possible growth *in silico* for every experimental condition with available data. Model accuracy is assessed by determining that fraction of experimental conditions where the growth predicted *in silico* and growth observed *in vivo* are either both zero or both nonzero.

doi:10.1038/nbt.1672