

# High-throughput generation and optimization of genome-scale metabolic models

Christopher S. Henry,<sup>1</sup> Matthew DeJongh,<sup>2</sup> Aaron A. Best,<sup>3</sup> Paul M. Frybarger,<sup>2,3</sup> & Rick L. Stevens<sup>4</sup>

<sup>1</sup>*Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439*

<sup>2</sup>*Department of Computer Science, Hope College, 27 Graves Place, Holland, MI 49423*

<sup>3</sup>*Department of Biology, Hope College, 35 East 12th Street, Holland, MI 49423*

<sup>4</sup>*Computing, Environment and Life Sciences Directorate at Argonne National Laboratory and Computer Science Department and Computation Institute, University of Chicago, 1100 East 58<sup>th</sup> Street, Chicago, IL 60637*

Genome-scale metabolic models have proven to be crucial resources for translating detailed knowledge of thousands of distinct biochemical processes into global predictions of organism behavior. These models can be used to predict essential genes, organism phenotypes, organism response to mutations, and metabolic engineering strategies [1]. The models also serve as platforms for assessing and expanding knowledge of metabolism via an iterative cycle of experimentation, prediction, and reconciliation [2]. Despite these many applications, methods for creating genome-scale models are failing to keep pace with genome sequencing. In the past decade, 800+ prokaryotic genomes have been submitted to NCBI, but only 30 genome-scale models have been published [3]. To address this problem, we have developed the High-Throughput Genome-scale Metabolic Reconstruction (HT-GMR) pipeline, which rapidly generates predictive genome-scale metabolic models from prokaryotic genome sequences. This pipeline integrates numerous technologies for automating portions of the reconstruction process with minimal manual intervention, including genome annotation [4, 5], reaction network annotation and assembly [6], thermodynamic analysis to determine reaction reversibility [7, 8], and model optimization to fit experimental data [8-10]. We used the HT-GMR pipeline to generate 130 new genome-scale metabolic models and fit 22 of these models to available experimental data [11-18]. Gibbs free energy of reaction values were generated for 90% of the reactions in every model [7]. Any gaps preventing models from growing on known minimal media were identified and filled to enable the prediction of phenotypes and essential gene sets. Comparison of HT-GMR models to available published models reveals that in 16 of 19 cases, the HT-GMR models include more genes than their published counterparts. Validation of the 22 models with available growth phenotype data [11-18] reveals the models to have an average accuracy of 66% before optimization

**and 87% after optimization, which closely approaches the accuracy of available published models.**

Until now, the genome-scale metabolic reconstruction process has followed a “one genome at a time” paradigm, where years of manual effort have been expended to build the most comprehensive model possible for a single organism. However, this paradigm does not scale in a world where sequencing capacity is rising and sequencing cost is falling at exponential rates; high-throughput methods are needed for performing genome-scale metabolic reconstruction without sacrificing quality or accuracy. To meet this challenge, we have designed and implemented the High-Throughput Genome-scale Metabolic Reconstruction (HT-GMR) pipeline within the SEED framework for genome annotation and analysis [5]. The SEED framework addresses two fundamental needs for rapid generation of accurate genome-scale metabolic models: the need for high-quality, consistent underlying genome annotations, and the rapid annotation of newly sequenced genomes [4]. The SEED uses a *subsystems*-based approach to ensure high-quality annotation across sets of genes that are related by function (e.g., a metabolic process) or structure (e.g., a ribosomal complex). These subsystems are used to derive protein families that encode the core metabolic and non-metabolic functions of prokaryotic life [19]. Each SEED subsystem is maintained by an expert annotator who manually curates the protein families derived from the subsystem to ensure that they are consistently annotated across all sequenced genomes. Annotators also mine the scientific literature to improve the accuracy and detail of the subsystem annotations. The end result is the maintenance of a database of high-quality and up-to-date annotations that is continuously expanded to include new genome sequences.

The HT-GMR pipeline extends the SEED’s “many genomes in parallel” approach to all of the remaining steps in the genome-scale metabolic reconstruction process: preliminary reconstruction, model completion, consistency analysis, and model

optimization (described below; see Fig. 1). We applied the HT-GMR pipeline to generate new genome-scale metabolic models for 130 organisms spread across 13 bacterial divisions (these can be viewed and downloaded from <http://www.theseed.org/models/>). We selected organisms for this initial study based on annotation quality (measured by the fraction of the genes in the genome that are included in SEED subsystems), pathogenicity, industrial applicability, and availability of a published model for the organism (to allow for comparison).

Model generation begins with the *preliminary reconstruction* step of the HT-GMR pipeline, which uses the SEED annotations to produce a preliminary genome-scale model for each of the 130 organisms. Each preliminary model consists of a network of metabolic and transport reactions, gene-protein-reaction associations, and an organism-specific biomass reaction. To support the preliminary reconstruction step, we mapped biochemical reactions to functional roles in subsystems to enable the automated assembly of gene-protein-reaction networks [6], and created a template biomass reaction that is used in the HT-GMR pipeline to generate a distinct biomass reaction for each organism (Supplementary Table S1).

Each preliminary model produced by the pipeline undergoes an *auto-completion* step to fill any gaps in the reaction network that prevent the production of one or more small molecule building blocks in the biomass reaction (e.g., amino acids, nucleotides, and cofactors). Reactions are added from a comprehensive database of approximately 12,000 spontaneous reactions, enzymatic reactions, and trans-membrane transport reactions maintained as a part of the SEED (<http://www.theseed.org/reactions/>). This database combines all the biochemistry contained in the KEGG [20, 21] and 13 published genome-scale metabolic models [10, 12, 15, 16, 22-30] into a single, non-redundant set. The auto-completion step ensures that every HT-GMR model is capable of simulating cell growth. It also produces a list of metabolic functions predicted to be

missing from the genome annotations in the SEED (Supplementary Table S2). On average, 56 reactions were added to the 130 HT-GMR models during the auto-completion step (Fig. 2c), increasing average size of the models to 965 reactions. As expected, the most well studied bacterium, *Escherichia coli K12*, required the fewest auto-completion reactions (10 reactions) in order for biomass to be produced. This was in spite of the fact that the auto-completion for *E. coli* was performed while simulating growth in glucose minimal media. In contrast, the model of the endosymbiont *Buchnera aphidicola* required the most auto-completion reactions (132 reactions). Many of the reactions added to *B. aphidicola* were transporters for essential metabolites that *B. aphidicola* cannot produce biosynthetically. Some of the remaining intracellular auto-completion reactions represent metabolic functions that are predicted to be missing from the *B. aphidicola* annotations. However, many also represent metabolic functions that are provided to *B. aphidicola* by its host (e.g., the lipopolysaccharide biosynthesis pathways) [31]. In general, the 23 HT-GMR models associated with obligate-intracellular organisms required more reactions (74 on average) to be added during the auto-completion step, a result of the dependency of these organisms on metabolic functions provided by their hosts. The functions added to the obligate-intracellular organisms during the auto-completion step provide useful insights into the nature of the symbiotic or parasitic relationship these organisms have with their hosts.

The auto-completion results also enable the identification of the regions of the metabolic network where the gaps in the genome annotations for the model organisms appear to be more prevalent (i.e., *Cell Wall and Capsule Biosynthesis* pathways, *Cofactors, Vitamins, and Prosthetic Group Biosynthesis* pathways). On average, 21 (15%) of the 141 reactions associated with the *Cell Wall and Capsule Biosynthesis* pathways in the HT-GMR models were added during the auto-completion process, meaning they are enzymatic reactions required for biomass production that have no gene associated with them. The *Cell Wall and Capsule Biosynthesis* reactions added

during the auto-completion process belong primarily to three SEED subsystems: *LOS Core Oligosaccharide Biosynthesis* (Gram negative), *Teichoic and Lipoteichoic Acids Biosynthesis* (Gram positive), and *KDO2-Lipid A Biosynthesis* (Gram negative). In general, the Gram negative cell wall biosynthesis pathways involved more auto-completion reactions than did the Gram positive pathways. Similarly, 10 (5%) of the 194 reactions associated with the *Cofactors, Vitamins, and Prosthetic Group Biosynthesis* pathways in each HT-GMR model were added during the auto-completion step. These belong primarily to three SEED subsystems: *Ubiquinone Biosynthesis, Menaquinone and Phylloquinone Biosynthesis*, and *Thiamin Biosynthesis*. Overall, while an average of 56 reactions were added to the HT-GMR models during the auto-completion step, over 31 of these reactions are associated with only six subsystems in the SEED. This indicates that the quality and completeness of the SEED annotations can be significantly improved by targeting these six subsystems for additional experimental work and manual curation. Because the preliminary reconstruction and auto-completion steps in the HT-GMR pipeline are fully automated, any improvements made to the SEED annotations can be rapidly integrated into new versions of the HT-GMR models.

We call the models that are generated by the preliminary reconstruction and auto-completion steps *analysis-ready* because they can simulate the production of biomass from transportable nutrients. Hence, these models can be used with flux balance analysis tools [1, 32] to predict gene essentiality, organism growth conditions, organism phenotypes, and overall organism response to environmental and genetic manipulations. An analysis-ready model was produced for all 130 genomes processed by the HT-GMR pipeline. On average, the analysis-ready models include 965 reactions (Fig. 2a), 688 genes (Fig. 2b), and 876 metabolites. Models vary significantly in size, from 243 reactions and 193 genes (*Onion yellows phytoplasma*) to 1396 reactions and 1586 genes (*Burkholderia xenovorans*).

The remaining steps of the HT-GMR pipeline involve the optimization of the analysis-ready models to better fit any available experimental growth phenotype data. Experimental data including Biolog phenotyping arrays [11-16] and gene essentiality datasets [17, 18] are available for 22 of the organisms for which models were constructed, and these data were used to validate and optimize the analysis-ready HT-GMR models for these 22 organisms (Fig. 3). The analysis-ready HT-GMR models had an average predictive accuracy of 60% for the Biolog data, 72% for the essentiality data, and 66% overall (blue bars in Fig. 3). These accuracies are low compared with the accuracy typical for published genome-scale models, which is around 90%. However, the extreme accuracy of published genome-scale models is largely a product of an iterative process of manual refinement to better fit available experimental data [2]. We implemented three steps in the HT-GMR pipeline to replicate this manual refining process in high-throughput: *Biolog consistency analysis*, *gene essentiality consistency analysis* and *model optimization*.

Consistency analysis of the 14 HT-GMR models with available Biolog data revealed an average of 69 nutrients that were lacking transport reactions in each model but were known to be metabolized based on the experimental data. Transport reactions were added to the models for each of these nutrients, which resulted in a 10.1% improvement in the accuracy of the Biolog phenotype predictions (Fig 3a) and a 4.9% improvement in overall accuracy. The large number of transport reactions added during the Biolog consistency analysis indicates that identification of the genes associated with transport of Biolog nutrients remains an open problem in annotation. However, the large increase in accuracy that results from the addition of transport reactions during the Biolog consistency analysis indicates that the intracellular pathways required to metabolize the Biolog nutrients are typically captured in the SEED annotations. The addition of new transport reactions during the Biolog consistency analysis had a negligible effect on the essentiality prediction accuracy (Fig 3b), with only two models

(*E. coli* K12 and *P. aeruginosa* PAOI) showing any improvement at all. This indicates that the transport reactions most essential for capturing how an organism grows and interacts with its environment in complex media (where most essentiality experiments are conducted) are well-captured by existing annotations or the auto-completion algorithm. Application of the gene essentiality consistency analysis algorithm to the 14 HT-GMR models with available essentiality data resulted in the identification and correction of 202 annotations that were inconsistent with available essentiality data (see Methods and Supplementary Table S5). This improved the average accuracy of the gene essentiality predictions to 78% (red bars in Fig. 3).

The model optimization step of the HT-GMR pipeline, which is a modified version of the GrowMatch algorithm [9], takes place in two stages: *GapFill* and *GapGen* (see Methods). Application of the GapFill stage resulted in the addition of an average of 15 new reactions, while the reversibility constraints were relaxed on an average of 5 existing reactions (Supplementary Table S6). These changes improved overall model accuracy to 83% on average (green bars in Fig. 3). As with the auto-completion, the GapFill stage of the model optimization step generated numerous predictions of metabolic functions that are missing from the current genome annotations, which will be useful to the ongoing efforts to improve these annotations.

Application of the GapGen stage of the model optimization step resulted in the removal of an average of 5 reactions, while reversibility constraints were tightened on an average of 8 reactions (Supplementary Table S7). These changes further improved overall model accuracy to 88% on average (purple bars in Fig. 3). Details of the exact reactions added and removed from each of the 22 HT-GMR models during the GapFill and GapGen stages are provided in Supplementary Tables S6 and S7, along with data on the explicit model predictions that were corrected with the addition or deletion of each reaction. In general, the GapFill and GapGen algorithms produced a more

significant improvement on the Biolog phenotyping array prediction accuracy than on gene essentiality prediction accuracy (Fig. 3).

Genome-scale models have already been published for 19 of the organisms selected for reconstruction by the HT-GMR pipeline [10, 12, 15, 16, 22-30] (Table 1). We compared the HT-GMR models for these 19 organisms with their published counterparts to determine how effectively the HT-GMR pipeline reproduces the results of the manual reconstruction process. A comparison of the number of reactions included in the models reveals that the HT-GMR models are significantly larger than their published counterparts with only one exception (*iAF1260*) (Table 1). In the HT-GMR models, many pathways that are typically lumped together in published models (e.g., fatty acid biosynthesis) were expanded to enable thermodynamic analysis, which is one of the reasons that the HT-GMR models contain more reactions. Comparison of the number of genes represented in the HT-GMR models and the published models is a better measure of relative complexity and completeness of the models because both models are derived from the same set of genes. In 16 of the 19 cases we examined, the HT-GMR models include more genes than their published counterparts (Table 1), indicating that the HT-GMR models (and the annotations they are derived from) tend to be more complete than their published counterparts. The improved coverage of the genomes by the HT-GMR models exemplifies the completeness of the SEED's subsystems-based annotations and the effectiveness of the HT-GMR pipeline in producing genome-scale models with a size and detail that matches and often exceeds the manual reconstruction process.

We conducted a detailed comparison for nine of the HT-GMR models to determine the extent to which the reactions and genes in these models overlap with the reactions and genes in the published models (Supplementary Table S3). The comparison reveals that the published models and the HT-GMR models share an average of 567

reactions, which amounts to 66% of the reactions in the published models and 42% of the reactions in the HT-GMR models. The gene comparison reveals that the published models and the HT-GMR models share an average of 505 genes, which amounts to 82% of the genes in the published models and 42% of the genes in the HT-GMR models. These results demonstrate that the published models and the HT-GMR models both contain a significant amount of exclusive content. However, the HT-GMR models contain a much larger fraction of reactions and genes that are not represented in the published models. Most of the exclusive genes in the published models are not involved in any SEED subsystems and are assigned to generic functional roles in the SEED (e.g., amino acid biosynthesis, amino acid permease). To capture these genes in the HT-GMR models, we will need to seek evidence to explain why specific reactions were assigned to these genes in the published models. The exclusive genes in the HT-GMR models are associated with specific metabolic functions involved in a wide range of subsystems in categories such as *Amino Acid Biosynthesis*, *Vitamin and Cofactor Biosynthesis*, and *Trans-membrane Transport*. These results indicate that published models may be significantly improved by reconciling them with the HT-GMR models.

Here we have applied the first high-throughput genome-scale metabolic reconstruction pipeline to the generation and optimization of 130 new genome-scale metabolic models; and we have constructed an online resource for downloading, viewing, comparing, and analyzing these new models (<http://www.theseed.org/models/>). While we have limited this initial application of the HT-GMR pipeline to 130 of the most well studied and annotated genomes in the SEED, the technology can rapidly be scaled up to produce functioning genome-scale metabolic models for all available prokaryotic genome sequences. In tandem with the SEED's rapid annotation service (the RAST [4]), the HT-GMR pipeline is also capable of producing a new model from any prokaryotic genome sequence in approximately five days. While the accuracy of the models will decline significantly for organisms that are poorly covered by the current

set of SEED subsystems, the availability of preliminary functioning genome-scale metabolic models for these organisms will significantly benefit efforts to improve their annotations. Additionally, in the process of constructing these models we have generated hundreds of new predictions of functions believed to be missing from the current annotations (Supplementary Tables S2 and S6). As microbiologists test for the presence of the missing functions proposed in this work, we will update our models and rerun the auto-completion and model optimization steps to replace predictions that are inconsistent with their findings. This iterative process of model improvement has been highly successful when applied to *E. coli* [2], but each cycle of the process has previously required years because of the time needed to absorb new data and publish an updated model. The HT-GMR pipeline will enable us to rebuild all of the HT-GMR models on a monthly or even biweekly basis. Rapid update of genome-scale models is essential for keeping up with the emergence of new high-throughput experimental data sets and for enabling researchers worldwide to rapidly benefit from new discoveries in organism metabolism.

### **Methods Summary**

The HT-GMR pipeline uses a semi-automated process to produce optimized genome-scale metabolic models from annotated genomes in the SEED framework for genome annotation and analysis. The pipeline consists of five consecutively applied steps (Fig. 1): (1) assembly of a preliminary reconstruction from a SEED annotated genome; (2) auto-completion of the preliminary reconstruction to fill any gaps in the reaction network that prevent the production of biomass from transportable nutrients; (3) Biolog consistency analysis to identify metabolized nutrients that are lacking transport reactions in the model; (4) gene essentiality consistency analysis to identify cases where gene-protein-reaction relationships in the models are inconsistent with available gene essentiality data; and (5) model optimization to add or remove a minimal

set of reactions from the model to improve the accuracy of model predictions. Each of these five steps is described in detail in the Additional Methods section. Flux Balance Analysis (FBA) is used with the HT-GMR models to simulate organism growth in a variety of environments and with a variety of genetic modifications. Simulation results are compared with available Biolog phenotyping array data and gene essentiality data to assess model accuracy.

## References

1. Feist AM, Palsson BO: **The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli.** *Nat Biotechnol* 2008, **26**:659-667.
2. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO: **Integrating high-throughput and computational data elucidates bacterial networks.** *Nature* 2004, **429**:92-96.
3. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BØ: **Reconstruction of Biochemical Networks in Microbial Organisms.** *Nat Rev Microbiol* 2009, **7**:129-143.
4. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9**:75.
5. Overbeek R, Disz T, Stevens R: **The SEED: A peer-to-peer environment for genome annotation.** *Communications of the Acm* 2004, **47**:46-51.
6. DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, Best A: **Toward the automated generation of genome-scale metabolic networks in the SEED.** *BMC Bioinformatics* 2007, **8**:139.
7. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V: **Group contribution method for thermodynamic analysis of complex metabolic networks.** *Biophys J* 2008, **95**:1487-1499.
8. Henry CS, Zinner J, Cohoon M, Stevens R: **iBsu1103: a new genome scale metabolic model of B. subtilis based on SEED annotations.** *Genome Biol* 2009, **10**:R69.
9. Kumar VS, Maranas CD: **GrowMatch: an automated method for reconciling in silico/in vivo growth predictions.** *PLoS Comput Biol* 2009, **5**:e1000308.
10. Suthers PF, Dasika MS, Kumar VS, Denisov G, Glass JI, Maranas CD: **A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189.** *PLoS Comput Biol* 2009, **5**:e1000285.
11. von Eiff C, McNamara P, Becker K, Bates D, Lei XH, Ziman M, Bochner BR, Peters G, Proctor RA: **Phenotype microarray profiling of Staphylococcus**

- aureus menD and hemB mutants with the small-colony-variant phenotype.** *J Bacteriol* 2006, **188**:687-693.
12. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R: **Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data.** *J Biol Chem* 2007, **282**:28791-28799.
  13. Bochner BR: **Global phenotypic characterization of bacteria.** *Fems Microbiol Rev* 2009, **33**:191-205.
  14. Keymer DP, Miller MC, Schoolnik GK, Boehm AB: **Genomic and phenotypic diversity of coastal Vibrio cholerae strains is linked to environmental factors.** *Appl Environ Microbiol* 2007, **73**:3705-3714.
  15. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1261 ORFs and thermodynamic information.** *Mol Syst Biol* 2007, **3**:121.
  16. Durot M, Le Fevre F, de Berardinis V, Kreimeyer A, Vallenet D, Combe C, Smidtas S, Salanoubat M, Weissenbach J, Schachter V: **Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data.** *BMC Syst Biol* 2008, **2**:85.
  17. Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, Osterman A: **Essential genes on metabolic maps.** *Curr Opin Biotechnol* 2006, **17**:448-456.
  18. Zhang R, Ou HY, Zhang CT: **DEG: a database of essential genes.** *Nucleic Acids Res* 2004, **32**:D271-272.
  19. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V *et al*: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Research* 2005, **33**:5691-5702.
  20. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Research* 2000, **28**:27-30.
  21. Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet.** *Nucleic Acids Research* 2002, **30**:42-46.
  22. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).** *Genome Biol* 2003, **4**:1-12.
  23. Goelzer A, Bekkal Brikci F, Martin-Verstraete I, Noirot P, Bessieres P, Aymerich S, Fromion V: **Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis.** *BMC Syst Biol* 2008, **2**:20.
  24. Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BO: **Genome-scale metabolic model of Helicobacter pylori 26695.** *Journal of Bacteriology* 2002, **184**:4582-4593.
  25. Oliveira AP, Nielsen J, Forster J: **Modeling Lactococcus lactis using a genome-scale flux model.** *BMC Microbiol* 2005, **5**:39.

26. Feist AM, Scholten JC, Palsson BO, Brockman FJ, Ideker T: **Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri***. *Mol Syst Biol* 2006, **2**:2006 0004.
27. Jamshidi N, Palsson BO: **Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain iNJ661 and proposing alternative drug targets**. *BMC Syst Biol* 2007, **1**:26.
28. Nogales J, Palsson BO, Thiele I: **A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory**. *BMC Syst Biol* 2008, **2**:79.
29. Duarte NC, Herrgard MJ, Palsson BO: **Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model**. *Genome Research* 2004, **14**:1298-1309.
30. Becker SA, Palsson BO: **Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation**. *BMC Microbiol* 2005, **5**:8.
31. Douglas AE: **Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera***. *Annu Rev Entomol* 1998, **43**:17-37.
32. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox**. *Nat Protoc* 2007, **2**:727-738.
33. Lee J, Yun H, Feist AM, Palsson BO, Lee SY: **Genome-scale reconstruction and in silico analysis of the *Clostridium acetobutylicum* ATCC 824 metabolic network**. *Appl Microbiol Biotechnol* 2008, **80**:849-862.
34. Mahadevan R, Bond DR, Butler JE, Esteve-Nunez A, Coppi MV, Palsson BO, Schilling CH, Lovley DR: **Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling**. *Appl Environ Microbiol* 2006, **72**:1558-1568.
35. Schilling CH, Palsson BO: **Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis**. *Journal of Theoretical Biology* 2000, **203**:249-283.
36. Thiele I, Vo TD, Price ND, Palsson BO: **Expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants**. *J Bacteriol* 2005, **187**:5818-5830.
37. Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, Smid EJ: **Analysis of growth of *Lactobacillus plantarum* WCFS1 on a complex medium using a genome-scale metabolic model**. *J Biol Chem* 2006, **281**:40041-40048.
38. Kim TY, Kim HU, Park JM, Song H, Kim JS, Lee SY: **Genome-scale analysis of *Mannheimia succiniciproducens* metabolism**. *Biotechnol Bioeng* 2007, **97**:657-671.
39. Baart GJ, Zomer B, de Haan A, van der Pol LA, Beuvery EC, Tramper J, Martens DE: **Modeling *Neisseria meningitidis* metabolism: from genome to metabolic fluxes**. *Genome Biol* 2007, **8**:R136.
40. Mazumdar V, Snitkin ES, Amar S, Segre D: **Metabolic network model of a human oral pathogen**. *J Bacteriol* 2009, **191**:74-90.

41. Oberhardt MA, Puchalka J, Fryer KE, Martins dos Santos VA, Papin JA: **Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1.** *J Bacteriol* 2008, **190**:2790-2803.
42. Resendis-Antonio O, Reed JL, Encarnacion S, Collado-Vides J, Palsson BO: **Metabolic reconstruction and modeling of nitrogen fixation in *Rhizobium etli*.** *PLoS Comput Biol* 2007, **3**:1887-1895.
43. Borodina I, Krabben P, Nielsen J: **Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism.** *Genome Res* 2005, **15**:820-829.

**Supplementary Information** accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

### **Acknowledgements**

This work was supported by the U.S. 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 grant MCB-0745100. We acknowledge the SEED annotators and development team for producing the annotations and computational infrastructure that make this work possible. We thank Ross Overbeek, Veronica Vonstein, Folker Meyer, Robert Olson, Terry Disz, and Mike Kubal for assistance with the use of the SEED genome annotation and analysis tools.

### **Author contributions**

CSH developed and operated the HT-GMR pipeline and the related algorithms with assistance from all other authors. All authors participated in curation and testing of the HT-GMR models. MD contributed significantly to reaction network annotation, model testing, and model curation. AAB contributed significantly to template biomass reaction design, model curation, and selection of organisms for reconstruction. PMF contributed to reaction network annotation. RLS contributed to model curation and the design of pipeline and algorithms. All authors contributed to the writing and revision of the manuscript.

### **Author information**

Reprints and permissions information is available at [npg.nature.com/reprintsandpermissions](http://npg.nature.com/reprintsandpermissions).

Correspondence and requests for materials should be addressed to Christopher Henry

([chenry@mcs.anl.gov](mailto:chenry@mcs.anl.gov)).

## Tables

**Table 1. Comparison with published models**

Organism name	Published model	Published reactions	SEED Reactions	Published genes	SEED genes
<i>Acinetobacter</i>	iAbaylyiv4 [16]	868	1196	775	785
<i>B. subtilis</i>	iYO844 [12]	1020	1463	844	1041
<i>C. acetobutylicum</i>	iJL432 [33]	502	989	432	721
<i>E. coli</i>	iAF1260 [15]	2013	1529	1261	1083
<i>G. sulfurreducens</i>	iRM588 [34]	523	721	588	468
<i>H. influenzae</i>	iCS400 [35]	461	969	400	575
<i>H. pylori</i>	iIT341 [36]	476	731	341	421
<i>L. plantarum</i>	iBT721 [37]	643	908	721	699
<i>L. lactis</i>	iAO358 [25]	621	965	358	646
<i>M. succiniciproducens</i>	iTK425 [38]	686	1048	425	659
<i>M. tuberculosis</i>	iNJ661 [27]	939	1021	661	728
<i>M. genitalium</i>	iPS189 [10]	264	294	189	214
<i>N. meningitidis</i>	iGB555 [39]	496	903	555	560
<i>P. gingivalis</i>	iVM679 [40]	679	744	0*	399
<i>P. aeruginosa</i>	iMO1056 [41]	883	1386	1056	1094
<i>P. putida</i>	iNJ746 [28]	950	1261	746	1053
<i>R. etli</i>	iOR363 [42]	387	1264	363	1242
<i>S. aureus</i>	iSB619 [30]	641	1115	619	770
<i>S. coelicolor</i>	iIB700 [43]	700	1159	700	987

\*genes were not associated with reactions in this model.

## Figures legends

Figure 1. High-throughput Genome-scale Metabolic Reconstruction Pipeline. In the first step of the HT-GMR pipeline, a preliminary model is assembled consisting of intracellular and transport reactions associated with genes on the basis of the SEED annotations, spontaneous reactions, and a distinct biomass reaction assembled from the template biomass reaction. In the auto-completion step of the pipeline, additional intracellular and transport reactions are added to the preliminary model to generate an analysis-ready model capable of simulating biomass production using only transportable nutrients. The final three steps of the HT-GMR pipeline involve the removal/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.

Figure 2. Statistics for analysis-ready HT-GMR models. The number of reactions (a) and genes (b) included in the 130 analysis-ready models follows a roughly bell-shaped distribution (blue bars). However, the distribution of reactions added by the auto-completion algorithm (c) is distinctively skewed to the left, indicating that a small number of models require a significantly larger number of auto-completion reactions.

Figure 3. Accuracy of models generated by the HT-GMR pipeline. The accuracy of the HT-GMR models in predicting Biolog phenotyping array data (A) and gene essentiality data (B) steadily improved during the model refining steps of the pipeline. Prior to optimization, the HT-GMR models had an average overall accuracy of 66% (blue bars); this increased to 71% after the Biolog consistency analysis (orange bars); 75% after the gene essentiality consistency analysis (red bars); 83% after the GapFill stage of the model optimization (green bars); and 88% after the GapGen stage of the model optimization (purple bars). The

gene essentiality consistency analysis impacted only the gene-protein-reaction associations in the models, so it had no effect on the Biolog phenotyping array prediction accuracy.

## **Additional Methods**

### **HT-GMR pipeline: preliminary reconstruction**

The HT-GMR pipeline produces analysis-ready genome-scale metabolic models starting from the high-quality genome annotations produced and maintained within the SEED framework, and optimizes them when phenotype and gene essentiality data are available (Fig.1). In the first step of this pipeline, a preliminary metabolic model is constructed consisting of (1) the spontaneous reactions, enzymatic reactions, and transport reactions that make up an organism's metabolism; (2) the set of gene-protein-reaction (GPR) relationships that describe how reaction activity depends upon an organism's genes; and (3) a biomass reaction that describes the essential small molecule building blocks of the organism. Enzymatic intracellular and trans-membrane 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/reactions/>) 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. 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 S1), which was constructed from a curation of the biomass reactions included in 19 existing genome-scale metabolic models [10, 12, 15, 16, 22-30]. 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.

### **HT-GMR pipeline: auto-completion**

The preliminary metabolic models assembled during the first step of the HT-GMR 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 second step of the HT-GMR 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 modelled organism (Supplementary Table S2). 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 (<http://www.theseed.org/reactions/>). This database consists of approximately 12,000 reactions and 15,044 compounds, and it combines all the biochemistry contained in the KEGG [20, 21] and 13 published genome-scale metabolic models [10, 12, 15, 16, 22-30] into a single, non-redundant set. Often the gaps in the reaction network of a preliminary model may be filled by many different distinct sets of reactions. The objective function of the auto-completion optimization is parameterized to favor the selection of the set of reactions that represents the best possible hypothesis of what is actually missing from the genome annotations. In this objective function, the addition of transport reactions is penalized so the completion of intracellular biosynthesis pathways is favored over the addition of

transport reactions. The addition of transport reactions for small molecules included in the biomass reaction is penalized more heavily. Addition of reactions proceeding in a thermodynamically unfavorable direction is also penalized to avoid auto-completion solutions that involve thermodynamically infeasible pathways. Addition of reactions associated with functional roles or subsystems in the SEED is favored because these reactions take part in the core pathways of metabolism and represent the most well curated portion of the known biochemistry. Addition of reactions associated with subsystems and pathways that are already heavily represented in the annotated genome is particularly favored because these reactions are more likely to be filling gaps in the genome annotations. Once the auto-completion optimization produces a set of reactions that optimally satisfy these criteria, the reactions are added to the preliminary model to produce an *analysis-ready* model.

### **HT-GMR pipeline: analysis-ready model optimization**

The remaining steps of the HT-GMR pipeline involve the optimization of the analysis-ready model to better fit any experimental growth phenotype data that is 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 [13]. 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.

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 developed by Kumar and colleagues [9] with additional global optimization steps described in detail in our previous work [8]. 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 under-constrained by the reversibility prediction method used.

### **Model validation using flux balance analysis**

Flux balance analysis (FBA) is first used in the HT-GMR pipeline to verify that every model produced by the pipeline is analysis-ready, by confirming that the model is capable of simulating biomass production in the minimal defined growth medium for the modelled 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 trans-membrane transport reactions exist in the model.

In the assesment 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.

### **The following government license should be removed before publication.**

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce,

prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.