Downloaded from orbit.dtu.dk on: Aug 27, 2022

DTU Library

BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree

Norsigian, Charles J.; Pusarla, Neha; McConn, John Luke; Yurkovich, James T.; Dräger, Andreas; Palsson, Bernhard O.; King, Zachary

Published in:
Nucleic acids research

Link to article, DOI:
10.1093/nar/gkz1054

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Norsigian, C. J., Pusarla, N., McConn, J. L., Yurkovich, J. T., Dräger, A., Palsson, B. O., & King, Z. (2020). BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic acids research*, *48*(1), D402-D406. https://doi.org/10.1093/nar/gkz1054

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

D402–D406 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 7 November 2019*
doi: 10.1093/nar/gkz1054

BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree

Charles J. Norsigian [1], Neha Pusarla[1], John Luke McConn[1], James T. Yurkovich[2], Andreas Dräger [3,4,5], Bernhard O. Palsson [1,6,7] and Zachary King[1,*]

[1]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA, [2]Institute for Systems Biology, Seattle, WA 98109, USA, [3]Computational Systems Biology of Infection and Antimicrobial-Resistant Pathogens, Institute for Biomedical Informatics (IBMI), University of Tübingen, 72076 Tübingen, Germany, [4]Department of Computer Science, University of Tübingen, 72076 Tübingen, Germany, [5]German Center for Infection Research (DZIF), 72076 Tübingen, Germany, [6]Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA and [7]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, Building 220, 2800 Kongens Lyngby, Denmark

Received September 15, 2019; Revised October 21, 2019; Editorial Decision October 22, 2019; Accepted October 24, 2019

ABSTRACT

The BiGG Models knowledge base (http://bigg.ucsd.edu) is a centralized repository for high-quality genome-scale metabolic models. For the past 12 years, the website has allowed users to browse and search metabolic models. Within this update, we detail new content and features in the repository, continuing the original effort to connect each model to genome annotations and external databases as well as standardization of reactions and metabolites. We describe the addition of 31 new models that expand the portion of the phylogenetic tree covered by BiGG Models. We also describe new functionality for hosting multi-strain models, which have proven to be insightful in a variety of studies centered on comparisons of related strains. Finally, the models in the knowledge base have been benchmarked using Memote, a new community-developed validator for genome-scale models to demonstrate the improving quality and transparency of model content in BiGG Models.

INTRODUCTION

BiGG Models (http://bigg.ucsd.edu) was initially released in 2010 as a knowledge base of biochemically, genetically and genomically structured genome-scale metabolic network reconstructions, and the first release was followed by a complete redesign in 2016 (1,2). Since its initial release, the BiGG Models publications have been cited over 450 times (via Web of Science) and the website maintains a user base of ∼2000 monthly active users. BiGG Models is built around a workflow for standardizing models that is meant to verify and, in some cases, improve model quality. External studies have also indicated the high quality of models in BiGG. In one instance, the robustness of growth predictions for models in BiGG was demonstrated and used as a benchmark for a new collection of microbiome metabolic models (3). Another study on 'erroneous energy generating cycles'—a common issue in metabolic models—found that models in BiGG were less likely to have these undesirable cycles than models from other databases (4). A number of projects have used BiGG to automate reconstruction workflows and analyses (5–7).

With the BiGG Models 2020 update, we have included an additional 31 genome-scale metabolic models (GEMs) across four independent releases (versions 1.3–1.6), introduced the ability to download sets of multi-strain models that have been generated from a given base reconstruction page and continuously improved features with suggestions and contributions from the open source community. New content has increased the utility of the knowledge base for the community by expanding the number of organisms and metabolic processes represented. The BiGG Models architecture has been designed to enable these advances and continually improve the knowledge base.

KNOWLEDGE BASE CONTENT

BiGG Models continues to contain high-quality, manually curated GEMs collected from various publications. Quality control in BiGG Models begins with our requirement that all models undergo rigorous peer review before entry. We begin our import workflow with the exact model that was reported in a peer-reviewed publication, and the workflow is designed to improve the quality of annotations and standardization in the model, without making any changes

*To whom correspondence should be addressed. Tel: +1 517 320 0932; Email: zaking@ucsd.edu

© The Author(s) 2019. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/article-abstract/48/D1/D402/5614178 by guest on 30 April 2020

to the reaction content, parameterization or relationships (e.g. gene–reaction rules).

To load a model into BiGG, first each model is aligned to the shared namespace of reactions and metabolites across all models. When identifiers can be improved automatically (e.g. by finding a universal reaction based on the reactants), the workflow does this automatically; in other cases, non-matching identifiers are left as is to ensure that model content does not change. Next, genome annotations are loaded into the database for each model, providing explicit links between metabolic reactions and genes. When adding content to the BiGG Models database, manual efforts are made to ensure that each metabolite identifier follows the specified naming convention, each reaction contains a unique identifier and gene–reaction rules are properly represented in valid Boolean logic. When obvious errors are identified (typos, duplicate metabolites), these are corrected manually, with feedback from the model authors. The coalescence of genome annotation information, with external database links, and reaction, metabolite, and gene information from peer-reviewed models drives the quality of the knowledge base.

To ensure that model content (the reaction connectivity, gene–reaction rules and parameters that affect model predictions) has not changed from the peer-reviewed version presented in the original publication, an internal testing suite runs 18 tests for each model, for a total of >1900 tests. For example, tests ensure that reaction, metabolite, and gene counts have not changed, that all reactions that were mass balanced in the published model are still balanced and that genes have mapped to genome annotations correctly. An additional 36 tests are included to spot-check bugs and edge cases that have appeared during previous builds of BiGG Models. The full test suite is available in the source code (https://github.com/SBRG/bigg_models/blob/master/bigg_models/tests).

In the 2016 release of BiGG Models, there were 77 GEMs; with this update, we detail 31 additional models, covering release versions 1.3–1.6 (http://bigg.ucsd.edu/updates), and bringing the total to 108 GEMs (8–13). Genome annotations for each model (where possible) are downloaded from the National Center for Biotechnology reference sequence database (14) and linked to the corresponding GEM. Notable additions are the Recon3D, iCHOv1 and iML1515 (15–17) for the human metabolic network, Chinese hamster ovary cell and *Escherichia coli* K-12 MG1655, respectively. BiGG Models continues to host gold-standard models within a shared knowledge base of biological reactions and metabolites. We also demonstrate that the new GEMs valuably expand the portion of the reactome encapsulated by the knowledge base. The number of unique reactions represented in the database more than doubled from 11,459 in the 2016 version to 28,302. Likewise, the number of unique metabolites has more than doubled from 4,040 to 9,088. In addition to expanding the number of metabolic processes within the database, we sought to evaluate the diversity of reaction presence among GEMs within the database. Reaction presence or absence of the shared namespace was identified for every representative GEM, and this matrix was subject to multiple correspondence analysis (Figure 1). Notably, this analysis shows that new models within the update exist at the edge of each cluster demonstrating that the new content is increasing the level of dissimilarity among GEM reaction content. This separation among models conveys that the metabolic space within BiGG Models is moving past representations of shared common pathways and incorporating an increasing amount of organism-specific biochemical capabilities.

This update also includes multi-strain models, a recent development within the metabolic modeling community. We define multi-strain models as those generated via the ability to extend the content contained within a gold-standard reconstruction to related strains of interest. This technique has proven insightful in a number of studies for comparative analysis of strains (18–24). Thus, we have included a means for the hosting of the draft strain-specific models generated within these studies on BiGG Models. Each strain-specific model is available to download within a zip folder from the page of the base reconstruction used to generate the strain-specific models. The GEMs of iCN718, iYL1228 and STM_v1_0 (18,25,26) each contain datasets of multi-strain models linked from their reconstruction pages within BiGG Models. Identifiers in multi-strain reconstructions are inherently BiGG Models compliant as they have been generated through the use of a hosted model. These multi-strain models have demonstrated value in comparative simulation to identify key differences among the strains of a species and they all represent starting points toward manually curated reconstructions for each strain should the proper steps be undertaken (27).

VALIDATION OF MODELS WITH MEMOTE

BiGG Models now links to the model validation tool, Memote, which evaluates and scores GEMs with a set of community-maintained tests (28). Consistent with the efforts in BiGG Models to maximize the value of metabolic models, evaluation with Memote provides a means to quantify model quality. Quality, in this case, indicates that GEMs adhere to established standards such as consistent identification of model components and biologically feasible results under varied growth conditions. This standardized approach to model validation ensures the quality of BiGG Models content and provides a benchmark for continued improvement.

Both the original 77 GEMs included in the 2016 release of BiGG and the 31 GEMs included in this update were evaluated with Memote (Figure 2). Largely due to improved gene, metabolite, and reaction annotations, the average Memote score of JSON-formatted models increased from 40% to 58%, while that of the SBML-formatted (29–31) models advanced from 66% to 73% (Supplementary Table S1). While these scores represent significant improvements, ongoing database annotation efforts will be necessary to maximize Memote scores for models in BiGG. Memote does not currently support testing of MATLAB-formatted models; however, BiGG generates MATLAB-formatted models using the same data sources as the JSON-formatted files, so equivalent model content is present. These results highlight the value of BiGG Models as a knowledge base of GEMs, and scoring its content with Memote reinforces its effort to

Downloaded from https://academic.oup.com/nar/article-abstract/48/D1/D402/5614178 by guest on 30 April 2020

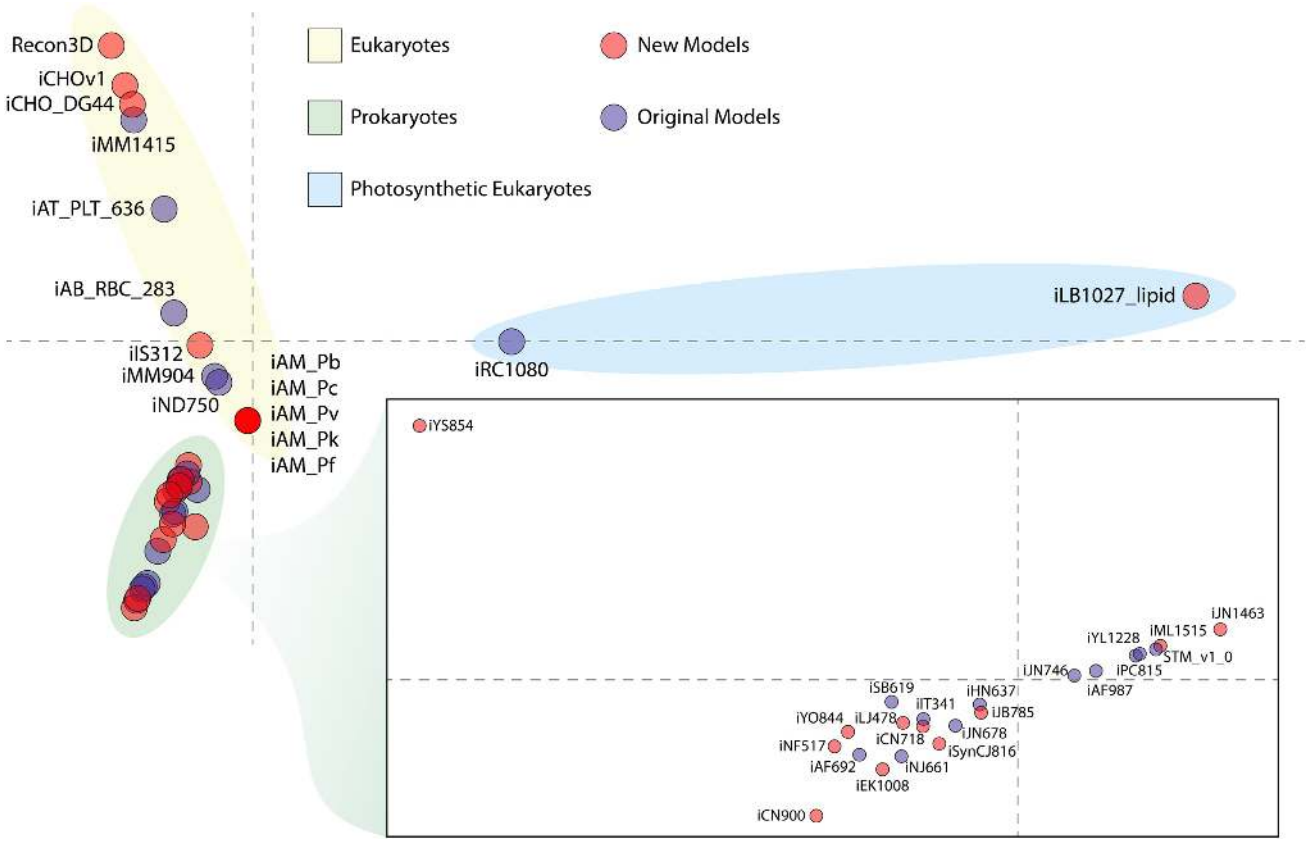

Figure 1. Multiple correspondence analysis of the reaction presence or absence within each model clusters models according to eukaryotic (yellow ellipse), prokaryotic (green ellipse and inset) and photosynthetic eukaryotes (blue ellipse) within metabolic reaction space. Dimension 1 (x-axis) explained 14.5% of the variance; dimension 2 (y-axis) explained 14.2%. Further, a number of the models newly introduced within this update (red circles) are found at edges of the MCA plot, indicating that within these two dimensions, they contribute to additional diversity in reaction content compared to the previous release. For this analysis, iML1515 was used as a representative *E. coli* model and iIS312 as representative for *Trypanosoma cruzi*.

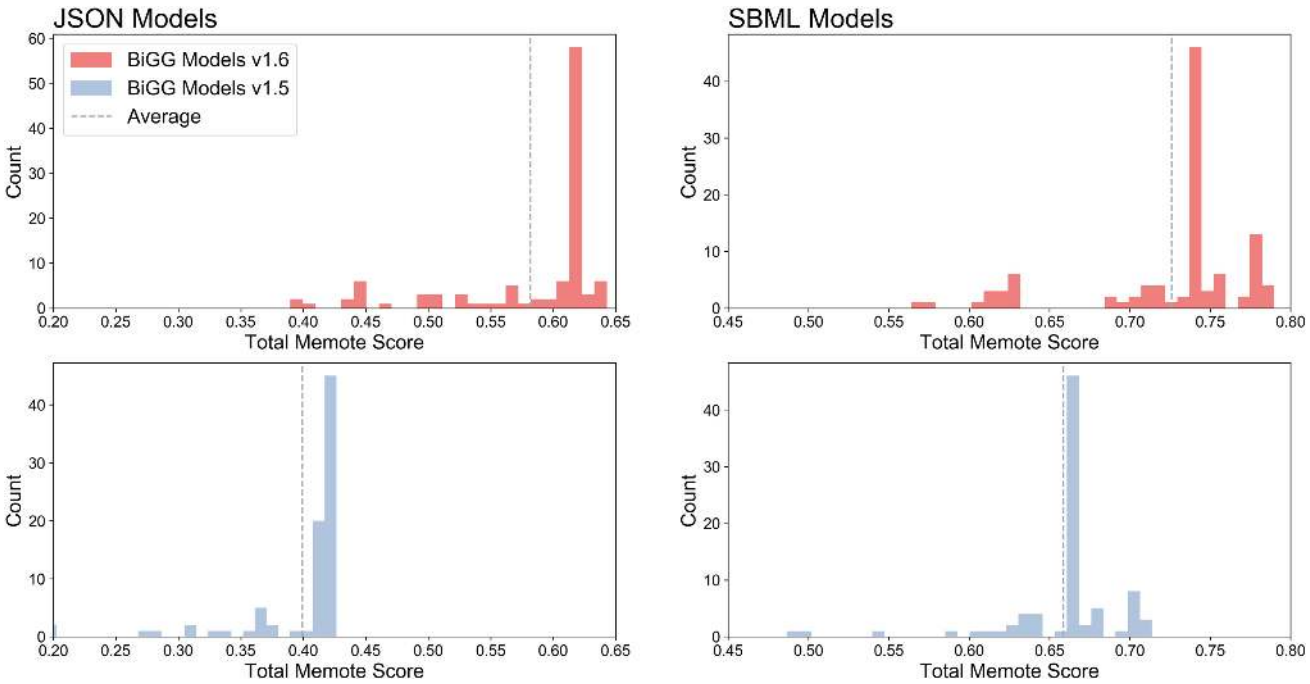

Figure 2. The latest update has resulted in improved Memote annotation scores for both JSON and SBML model formats. See Supplementary Table S1 for detailed score information for each model.

Downloaded from https://academic.oup.com/nar/article-abstract/48/D1/D402/5614178 by guest on 30 April 2020

provide access to GEMs with thorough and consistent standards.

ADDITIONAL FEATURES AND IMPROVEMENTS

Regular improvements are made to BiGG Models that have made the knowledge base faster, easier to use and better for analysis. Filters are now provided during search to filter out multi-strain reconstructions in the search results (see the toggle titled 'Exclude multi-strain models from search'). Gene and protein sequences are now included directly in the database and available by API. A new advanced search feature allows users to identify all gene and protein sequences for any universal BiGG reaction (see 'Find sequences for BiGG Models reaction' on the advanced search page).

A new 'universal' model was added for download on the Data Access page; this model provides all reactions and metabolites from BiGG in a single COBRA-compatible JSON file, so users can rapidly add BiGG content to their own computational workflows using COBRA tools. Namespace downloads on the Data Access page have also been extended to include old and deprecated identifiers. External database links are regularly updated with the latest information from MetaNetX (32). Many manual improvements have been made to annotations, including better gene mapping for yeast models. SBML downloads have improved through regular updates to the ModelPolisher project (https://github.com/draeger-lab/ModelPolisher).

Since the 2016 release of BiGG Models, the website has been deployed on a new server to dramatically improve speed when searching and browsing. Finally, bugs and suggestions are collected on GitHub (https://github.com/SBRG/bigg_models), and this has led to continuous and transparent improvements to the site by the BiGG Models team.

CONCLUSION

BiGG Models continues to be a widely used and well-maintained platform for integrating, sharing and standardizing GEMs. The updated knowledge base integrates the metabolic knowledge for 108 GEMs, as well as including the content for 515 draft strain-specific models across three organisms, all available within the knowledge base. BiGG Models is free for academic use and continues to extend the content within the knowledge base. Further, all source code continues to be available on GitHub to enable submission of potential bugs. The development of BiGG Models continues to evolve with the needs of the research community, introducing multi-strain models and validation through Memote testing. Future BiGG Models releases will continue to be shaped by the feedback from users.

DATA AVAILABILITY AND REQUIREMENTS

BiGG Models is freely available online for academic and non-profit use at http://bigg.ucsd.edu, under the BiGG License described at http://bigg.ucsd.edu/license. While the content of BiGG is restricted to academic and non-profit use to protect intellectual property claims, the source code is open source and available to all users under the MIT license at https://github.com/SBRG/bigg_models. Installation of an independent system requires Python 3.5 and PostgreSQL 9.4 or later.

We encourage community members to submit their model content to BiGG Models, and the website includes a section that describes the minimum requirements for inclusion in BiGG and the process for submitting a new model: http://bigg.ucsd.edu/about These requirements reflect the quality standards set by BiGG Models: identifier standardization for reactions and metabolites, links to genome annotations and peer-reviewed publication as the primary means of verifying model quality.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Novo Nordisk Fonden [NNF10CC1016517].
Conflict of interest statement. None declared.

REFERENCES

1. Schellenberger,J., Park,J.O., Conrad,T.M. and Palsson,B.Ø. (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, **11**, 213.
2. King,Z.A., Lu,J., Dräger,A., Miller,P., Federowicz,S., Lerman,J.A., Ebrahim,A., Palsson,B.O. and Lewis,N.E. (2016) BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.*, **44**, D515–D522.
3. Babaei,P., Shoaie,S., Ji,B. and Nielsen,J. (2018) Challenges in modeling the human gut microbiome. *Nat. Biotechnol.*, **36**, 682–686.
4. Fritzemeier,C.J., Hartleb,D., Szappanos,B., Papp,B. and Lercher,M.J. (2017) Erroneous energy-generating cycles in published genome scale metabolic networks: identification and removal. *PLoS Comput. Biol.*, **13**, e1005494.
5. Chan,S.H.J., Cai,J., Wang,L., Simons-Senftle,M.N. and Maranas,C.D. (2017) Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. *Bioinformatics*, **33**, 3603–3609.
6. Machado,D., Andrejev,S., Tramontano,M. and Patil,K.R. (2018) Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.*, **46**, 7542–7553.
7. Xavier,J.C., Patil,K.R. and Rocha,I. (2017) Integration of biomass formulations of genome-scale metabolic models with experimental data reveals universally essential cofactors in prokaryotes. *Metab. Eng.*, **39**, 200–208.
8. Broddrick,J.T., Rubin,B.E., Welkie,D.G., Du,N., Mih,N., Diamond,S., Lee,J.J., Golden,S.S. and Palsson,B.O. (2016) Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E8344–E8353.
9. Levering,J., Broddrick,J., Dupont,C.L., Peers,G., Beeri,K., Mayers,J., Gallina,A.A., Allen,A.E., Palsson,B.O. and Zengler,K. (2016) Genome-scale model reveals metabolic basis of biomass partitioning in a model diatom. *PLoS One*, **11**, e0155038.
10. Calmels,C., McCann,A., Malphettes,L. and Andersen,M.R. (2019) Application of a curated genome-scale metabolic model of CHO DG44 to an industrial fed-batch process. *Metab. Eng.*, **51**, 9–19.
11. Monk,J.M., Koza,A., Campodonico,M.A., Machado,D., Seoane,J.M., Palsson,B.O., Herrgård,M.J. and Feist,A.M. (2016) Multi-omics quantification of species variation of *Escherichia coli* links molecular features with strain phenotypes. *Cell Syst.*, **3**, 238–251.
12. Seif,Y., Monk,J.M., Mih,N., Tsunemoto,H., Poudel,S., Zuniga,C., Broddrick,J., Zengler,K. and Palsson,B.O. (2019) A computational knowledge-base elucidates the response of *Staphylococcus aureus* to different media types. *PLoS Comput. Biol.*, **15**, e1006644.

Downloaded from https://academic.oup.com/nar/article-abstract/48/D1/D402/5614178 by guest on 30 April 2020

13. Abdel-Haleem,A.M., Hefzi,H., Mineta,K., Gao,X., Gojobori,T., Palsson,B.O., Lewis,N.E. and Jamshidi,N. (2018) Functional interrogation of *Plasmodium* genus metabolism identifies species- and stage-specific differences in nutrient essentiality and drug targeting. *PLoS Comput. Biol.*, **14**, e1005895.
14. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.
15. Brunk,E., Sahoo,S., Zielinski,D.C., Altunkaya,A., Dräger,A., Mih,N., Gatto,F., Nilsson,A., Preciat Gonzalez,G.A., Aurich,M.K. *et al.* (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.*, **36**, 272–281.
16. Monk,J.M., Lloyd,C.J., Brunk,E., Mih,N., Sastry,A., King,Z., Takeuchi,R., Nomura,W., Zhang,Z., Mori,H. *et al.* (2017) iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.*, **35**, 904–908.
17. Hefzi,H., Ang,K.S., Hanscho,M., Bordbar,A., Ruckerbauer,D., Lakshmanan,M., Orellana,C.A., Baycin-Hizal,D., Huang,Y., Ley,D. *et al.* (2016) A consensus genome-scale reconstruction of Chinese hamster ovary cell metabolism. *Cell Syst.*, **3**, 434–443.
18. Norsigian,C.J., Kavvas,E., Seif,Y., Palsson,B.O. and Monk,J.M. (2018) iCN718, an updated and improved genome-scale metabolic network reconstruction of *Acinetobacter baumannii* AYE. *Front. Genet.*, **9**, 121.
19. Norsigian,C.J., Attia,H., Szubin,R., Yassin,A.S., Palsson,B.Ø., Aziz,R.K. and Monk,J.M. (2019) Comparative genome-scale metabolic modeling of metallo-beta-lactamase-producing multidrug-resistant *Klebsiella pneumoniae* clinical isolates. *Front. Cell Infect. Microbiol.*, **9**, 161.
20. Seif,Y., Kavvas,E., Lachance,J.-C., Yurkovich,J.T., Nuccio,S.-P., Fang,X., Catoiu,E., Raffatellu,M., Palsson,B.O. and Monk,J.M. (2018) Genome-scale metabolic reconstructions of multiple *Salmonella* strains reveal serovar-specific metabolic traits. *Nat. Commun.*, **9**, 3771.
21. Fouts,D.E., Matthias,M.A., Adhikarla,H., Adler,B., Amorim-Santos,L., Berg,D.E., Bulach,D., Buschiazzo,A., Chang,Y.-F., Galloway,R.L. *et al.* (2016) What makes a bacterial species pathogenic? Comparative genomic analysis of the genus *Leptospira*. *PLoS Negl. Trop. Dis.*, **10**, e0004403.
22. Fang,X., Monk,J.M., Mih,N., Du,B., Sastry,A.V., Kavvas,E., Seif,Y., Smarr,L. and Palsson,B.O. (2018) *Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. *BMC Syst. Biol.*, **12**, 66.
23. Bosi,E., Monk,J.M., Aziz,R.K., Fondi,M., Nizet,V. and Palsson,B.Ø. (2016) Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E3801–E3809.
24. Monk,J.M., Charusanti,P., Aziz,R.K., Lerman,J.A., Premyodhin,N., Orth,J.D., Feist,A.M. and Palsson,B.Ø. (2013) Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 20338–20343.
25. Thiele,I., Hyduke,D.R., Steeb,B., Fankam,G., Allen,D.K., Bazzani,S., Charusanti,P., Chen,F.-C., Fleming,R.M.T., Hsiung,C.A. *et al.* (2011) A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella* Typhimurium LT2. *BMC Syst. Biol.*, **5**, 8.
26. Liao,Y.-C., Huang,T.-W., Chen,F.-C., Charusanti,P., Hong,J.S.J., Chang,H.-Y., Tsai,S.-F., Palsson,B.O. and Hsiung,C.A. (2011) An experimentally validated genome-scale metabolic reconstruction of *Klebsiella pneumoniae* MGH 78578, iYL1228. *J. Bacteriol.*, **193**, 1710–1717.
27. Thiele,I. and Palsson,B.Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.
28. Lieven,C., Beber,M.E., Olivier,B.G., Bergmann,F.T., Chauhan,S., Correia,K. and Others (2018) Memote: a community driven effort towards a standardized genome-scale metabolic model test suite. bioRxiv doi: https://doi.org/10.1101/350991, 21 June 2018, preprint: not peer reviewed.
29. Hucka,M., Bergmann,F.T., Hoops,S., Keating,S.M., Sahle,S., Schaff,J.C., Smith,L.P. and Wilkinson,D.J. (2015) The Systems Biology Markup Language (SBML): language specification for level 3 version 1 core. *J. Integr. Bioinform.*, **12**, 266.
30. Olivier,B.G. and Bergmann,F.T. (2018) SBML Level 3 Package: Flux Balance Constraints version 2. *J. Integr. Bioinform.*, **15**, 660–690.
31. Hucka,M. and Smith,L.P. (2016) SBML Level 3 package: Groups, Version 1 Release 1. *J. Integr. Bioinform.*, **13**, 290.
32. Moretti,S., Martin,O., Van Du Tran,T., Bridge,A., Morgat,A. and Pagni,M. (2015) MetaNetX/MNXref—reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.*, **44**, D523–D526.

Downloaded from https://academic.oup.com/nar/article-abstract/48/D1/D402/5614178 by guest on 30 April 2020