Supporting Information

Monk et al. 10.1073/pnas.1307797110

SI Text

Strain-Specific Model Building Procedure. All genomes were reannotated using the RAST server (1, 2). Reannotation led to 600 new genes being annotated (Dataset S1, worksheet 8). Genes that were annotated as pseudogenes in the original National Center for Biotechnology Information (NCBI) annotation were treated as pseudogenes, and the enzymatic function of the proteins was removed from the final models. A total of 567 metabolic pseudogenes were identified (Dataset S1, worksheet 7). Although automated reconstruction methods exist (3–5) and are powerful tools for creating draft metabolic reconstructions for a wide range of species, it has been shown that starting from the highly curated content of a closely related organism produces a more accurate model that the currently available automated methods (6). Therefore, gene sequences from six metabolic models for *Escherichia coli* K-12 MG1655 (7), *Salmonella typhimurium* LT2 (8), *Klebsiella pneumoniae* MGH 78578 (9), *Yersinia pestis* CO92 (10), *E. coli* W (11), and *E. coli* B REL606 (12) were used for identifying orthologs. The SEED Corresponding Genes tool was used to identify orthologs in each strain of *E. coli* (1). This tool identifies best bidirectional hits (BBHs) and accounts for gene context (13). A 70% identity (PID) cutoff was used for assigning orthologs. This cutoff was determined by generating curated models from each other. Encouragingly, at 70% sequence identity conservation with *E. coli* K-12, there is only a false-positive rate of about 2.5%, but more than 75% of the final manually curated *Salmonella* model is recovered (8). This low false-positive rate is preferred because we can rely on gap-filling approaches (14) to collect missing genes due to choosing a stringent cutoff. Genes that were missing orthologs in the original models were deleted from the model for the target strain. Additional reaction content was added from ModelSEED (3), KEGG (15, 16), and BIOCYC (4). All reactions added were manually curated according to a published protocol (14). MetaNetX (17) was used to standardize metabolites and reactions to Systems Biology Research Group (18) abbreviations. All genome sequences were downloaded from GenBank (19) on September 21, 2012. Gene names conform to the NCBI locus name according to the original annotation in GenBank.

DH1 FabZ Annotation Error. Investigating the genetic basis of auxotrophies also led to the discovery of sequencing and annotation errors in published genome sequences. Gap filling analyses predicted that *E. coli* DH1 is lacking the activity of 3-oxo-glutaryl-[ACP] methyl ester dehydrase coded for by the *fabZ* gene. This reaction catalyzes the start of the essential coenzyeme biotin's synthesis pathway (20). Bioinformatic analysis of the *E. coli* DH1 *fabZ* gene revealed that it is a pseudogene due to a single base pair deletion. PCR was used to amplify and sequence this gene in *E. coli* DH1. The PCR-derived sequence does not match that of *fabZ* in the annotated genome of *E. coli* DH1 available in GenBank (19) (accession no. CP001637). Therefore, this is likely the result of an error in the original sequencing and annotation of the genome. Thus, it was concluded that a predicted auxotrophy for biotin in *E. coli* DH1 was in fact due to a sequencing error in the annotation of this genome.

Unique Catabolic Capabilities of Each Strain. Growth in 654 different growth supporting conditions was simulated for all 55 genome-scale models (GEMs). All 55 strain models were able to produce biomass on 285 different conditions. More than 90% of strains were able to produce biomass on 510 different conditions. A large drop off in catabolic capabilities occurs in the remaining 144 growth conditions. These conditions represent the pan catabolic capabilities of the *E. coli* species and are the conditions discussed in the main text. Dataset S1, worksheet 5 shows all 654 growth conditions and the count of strains capable of producing biomass for each condition.

In Vivo Growth Comparisons. Although it may seem obvious that experimental results should be taken as a gold standard for a true growth phenotype, disagreements in growth/no growth of the same strain on the same substrates can differ between studies. Two studies (6, 21) examined growth phenotypes of diverse *E. coli* strains on different sole carbon sources. Three strains (*E. coli* K-12 MG1655, *E. coli* CFT073, and *E. coli* O157:H7 EDL933) and more than 50 carbon sources overlapped between the two studies. However, less than 50% of the growth predictions were in agreement. These studies used Biolog plates (Biolog), which measure the color change caused by reduction of a tetrazolium dye as a proxy for growth. The discrepancies could arise from a number of reasons including different thresholds for growth calling, different length of time allowed for the cells to grow, or even different inoculation volumes. For these reasons, this study performed all growth screens in-house where it was possible to standardize experimental procedures.

SI Materials and Methods

Gap Filling. The COBRA implementation of the SMILEY algorithm (growMatch) (22) was used to predict sets of exchange and gap-filling reactions for models that were unable to simulate biomass in silico on M9 minimal media with glucose aerobically using flux balance analysis (FBA). The universal set of reactions used to fill gaps was the identified *E. coli* pan reactome discussed in the text. The Gurobi 5.0.0 mixed-integer linear programming solver was used (Gurobi Optimization) to implement SMILEY. When adding content to enable the strains to grow, exchange reactions indicating strain-specific auxotrophies were prioritized over adding new reactions without genetic evidence.

In Silico Growth Simulations. Each of the 55 metabolic network reconstructions were loaded into the COBRA Toolbox (23). M9 minimal media was simulated by setting a lower bound of −1,000 (allowing unlimited uptake) on the exchange reactions for Ca^{2+}, Cl^{-}, CO_2, Co^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, H^{+}, H_2O, K^{+}, Mg^{2+}, Mn^{2+}, MoO_4^{2-}, Na^{+}, Ni^{2+}, SeO_4^{2-}, SeO_3^{2}, and Zn^{2+}. A lower bound of −0.01 was placed on the cob(I)alamin exchange reaction. The default carbon source was glucose with a lower bound of −20, the default nitrogen source was NH_4^{-} with a lower bound of −1,000, the default phosphorous source was HPO_4^{2} with a default bound of −1,000, and the default sulfur source was $SO_4^{2}-$ with a default bound of −1,000 (Dataset S1, worksheet 5). To identify sole growth supporting carbon, nitrogen, phosphorous, and sulfur sources, each of these default compounds were removed from the media (lower bound set to 0) one at a time, and different compounds were added to determine if they supported growth. For aerobic simulations, O_2 was added with a lower bound of −20 and 0 was added for anaerobic simulations. For models with identified auxotrophies, the compound for which a strain was auxotrophic (Dataset S1, worksheet 6) was also added to the M9 minimal media for each simulation with a lower bound of −10. Model growth phenotypes were determined using FBA one at a time on each condition, with the core biomass reaction as the objective. Nutrient sources with growth rates

above zero were classified as growth supporting, whereas nutrient sources with growth rates of zero were classified as non–growth supporting. The Gurobi 5.0.0 linear programming solver (Gurobi Optimization) was used to perform FBA.

Heatmap and Phylogenetic Tree Construction. The binary results from the growth/no growth simulations for each strain were used to compute a correlation matrix based on dissimilarity indices calculated using the Jaccard method in the vegdist function of the Vegan R package. Ward's agglomerative clustering of the matrix of correlations was used to cluster the species using the hclust function of the Vegan R package and used to form a dendrogram. The heatmap was visualized using the gplots R package with values aligned based on the calculated dendrogram.

Decision Tree Construction. A decision tree (Fig. 3) was calculated based on growth/no growth values for each strain classified into their major pathotypes: InPec, ExPec, or commensal. The classification tree tool, part of the Orange Canvas software package (24), was used to calculate and display the decision tree using a Gini Index attribute selection criteria with no binarization and two minimum leaves for prepruning and $m = 2$ estimate for postpruning, with leaves of the same majority class being recursively merged.

Strains. Eleven strains of *E. coli* and one strain of *S. flexneri* were tested for their ability to grow on different carbon sources as part of this study. The 11 *E. coli strains* are SMS 3–5; CFT073; HS; DH1; UMN 026; K011; Sakai; ATCC 8739; 042; EDL933; and K12 MG1655. The *S. flexneri* strain was 2457T. *E. coli* 042 was a gift from Ian Henderson (Birmingham University, Birmingham, UK). All other strains were purchased from ATCC.

Carbon Source Testing. Concentrated stock solutions of D-(+)-glucose, lithium acetoacetate, deoxyribose, malic acid, D-(+)-melibiose, ferric citrate, and butyric acid were made by dissolving them in M9 minimal media. Ferric citrate required heat to dissolve. The stock solutions were then filter sterilized using Millipore Millex GP 0.22-μm membranes (Millipore), after which they were diluted with sterile M9 media to a final working concentration of 20 mM. Phenylacetaldehyde and phenethylamine were dissolved directly into M9 media at a 20 mM concentration before filter sterilization. The M9 medium contained (per liter): 6.8 g Na_2HPO_4; 3 g KH_2PO_4; 0.5 g NaCl; 1 g NH_4Cl; 2 mM $MgSO_4$; 0.1 mM $CaCl_2$; 4.2 mg $FeCl_3 \bullet 6H_2O$; 45 μg $ZnSO_4 \bullet 7H_2O$; 30 μg $CuCl_2 \bullet 2H_2O$; 30 μg $MnSO_4 \bullet H_2O$; 45 μg $CoCl_2 \bullet 6H_2O$; and 5.5 mg $Na_2EDTA \bullet 2H_2O$. All chemicals were sourced from Sigma-Aldrich. Two hundred microliters of the 10 growth media was then pipetted into each row of an untreated, flat bottom 96-well plate. As a negative control, we also included one row containing M9 media only; no carbon source had been added to this row.

The 11 strains of *E. coli* and *S. flexneri* 2a strain 2457T were then tested for growth on each of the carbon sources and the negative control sample. An overnight culture of each bacterium was diluted in M9 media to an OD_{600} value of ∼0.4. A 5-μL aliquot of each suspension was then inoculated into the designated wells of a 96-well plate. Growth was estimated by optical density at 48 h after inoculation. All OD_{600} measurements were made using a Molecular Devices Versamax plate reader. All tests were done in duplicate. Butyric acid and butane sulfonate did not support growth for any of the *E. coli* strains, including K-12 MG1655 for which the model predicted growth. This is likely because butyrate is toxic to *E. coli* cells at high concentrations such as those used in the growth screens.

1. Aziz RK, et al. (2012) SEED servers: High-performance access to the SEED genomes, annotations, and metabolic models. *PLoS ONE* 7(10):e48053.
2. Aziz RK, et al. (2008) The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics* 9:75.
3. Henry CS, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 28(9):977–982.
4. Caspi R, et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 36(Database issue):D623–D631.
5. Swainston N, Smallbone K, Mendes P, Kell D, Paton N (2011) The SuBliMinaL Toolbox: Automating steps in the reconstruction of metabolic networks. *J Integr Bioinform* 8(2):186.
6. Baumler DJ, Peplinski RG, Reed JL, Glasner JD, Perna NT (2011) The evolution of metabolic networks of E. coli. *BMC Syst Biol* 5:182.
7. Orth JD, et al. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. *Mol Syst Biol* 7:535.
8. Thiele I, et al. (2011) A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *BMC Syst Biol* 5(1):8.
9. Liao YC, et al. (2011) An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. *J Bacteriol* 193(7): 1710–1717.
10. Charusanti P, et al. (2011) An experimentally-supported genome-scale metabolic network reconstruction for Yersinia pestis CO92. *BMC Syst Biol* 5:163.
11. Archer CT, et al. (2011) The genome sequence of E. coli W (ATCC 9637): Comparative genome analysis and an improved genome-scale reconstruction of E. coli. *BMC Genomics* 12:9.
12. Yoon SH, et al. (2012) Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. *Genome Biol* 13(5):R37.
13. Binter E, et al. (2012) Grounding annotations in published literature with an emphasis on the functional roles used in metabolic models. *3 Biotech* 2(2):135–140.
14. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 5(1):93–121.
15. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40(Database issue): D109–D114.
16. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30.
17. Ganter M, Bernard T, Moretti S, Stelling J, Pagni M (2013) MetaNetX.org: A website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics* 29(6):815–816.
18. Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) BiGG: A Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 11:213.
19. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) GenBank. *Nucleic Acids Res* 33(Database issue):D34–D38.
20. Lin S, Hanson RE, Cronan JE (2010) Biotin synthesis begins by hijacking the fatty acid synthetic pathway. *Nat Chem Biol* 6(9):682–688.
21. Sabarly V, et al. (2011) The decoupling between genetic structure and metabolic phenotypes in Escherichia coli leads to continuous phenotypic diversity. *J Evol Biol* 24(7):1559–1571.
22. Reed JL, et al. (2006) Systems approach to refining genome annotation. *Proc Natl Acad Sci USA* 103(46):17480–17484.
23. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38(Database issue):D355–D360.
24. Demsar J, Zupan BLG, Curk T (2004) Orange: From experimental machine learning to interactive data mining. *Lecture Notes Comp Sci* 537–539.

Dataset S1. Model information in .xlsx format

Dataset S1

Worksheet 1: Strain-specific reconstruction information. A total of 55 strain-specific reconstructions were created. Species include *Escherichia coli, Shigella boydii, Shigella dysenteriae, Shigella flexneri*, and *Shigella sonnei*. *E. coli* pathotypes modeled are enterohemorrhagic *E. coli* (EHEC), enteroinvasive *E. coli* (EIEC), uropathogenic *E. coli* (UPEC), enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), avian pathogenic *E. coli* (APEC), adherent invasive *E. coli* (AIEC), and general extraintestinal pathogenic *E. coli* (ExPEC). Commensal *E. coli* strains comprise the greatest fraction of models with 19 models, followed by EHEC strains (8) and UPEC strains (7). Worksheet 2: Pan *E. coli* reactions. All reaction information associated with the 2,501 reactions in the pan reactome plus the 386 exchange reactions required for substrates to enter or leave the cell including reaction abbreviation, name, formula, and EC numbers. Reactions are assigned to metabolic systems and subsystems according to Orth et al. (7). Reactions have also been cross-referenced to major biochemical databases such as ModelSEED (3), KEGG (15, 16), and BIOCYC (17). Worksheet 3: Pan *E. coli* metabolites. All metabolite information associated with the 2,041 metabolites in the pan reactome including metabolite abbreviation, name, compartment, charge, formula, inchi and smile string, molecular weight, and CAS number. Metabolites have also been cross-referenced to major biochemical databases such as ModelSEED (3), KEGG (15, 16), and BIOCYC (17). Worksheet 4: Reaction presence and GPRs for each strain. Each strain's model specific gene protein reaction (GPR) association for reactions present in the pan reactome. Blank cells indicate that the model of the strain does not encode an enzyme catalyzing the reaction. Worksheet 5: In silico growth screens. Results of growth simulations for each of the 55 strains on 654 different growth supporting carbon, nitrogen, sulfur, and phosphorous sources. All simulations were performed in both aerobic and anaerobic conditions. Each condition is named with the element it is testing as well as whether the simulation was conducted in aerobic or anaerobic conditions. In silico predicted growth rate is presented in units of $hours^{-1}$. Worksheet 6: In silico M9 minimal media formulation and strain-specific auxotrophies. In silico formulation of M9 minimal media used to perform the growth screens in each different condition. This table also includes the specific compound that was removed to test each different carbon, nitrogen, phosphorous, and sulfur source. It also includes the specific exchange reactions that were opened in models of auxotrophic strains to allow the models to support growth on M9 minimal media. Worksheet 7: Strain-specific metabolic pseudogenes. A list of the 567 annotated metabolic pseudogenes based on NCBI annotations for each strain. The table also lists each pseudogene's closest identity gene in *E. coli* K-12 MG1655, as well as the percentage identity match. The function of all annotated metabolic pseudogenes was removed from each strain specific model. Worksheet 8: Strain-specific newly annotated metabolic genes. A table of the 600 newly annotated metabolic genes that were determined by reannotation using the RAST framework. These genes were not originally annotated in the NCBI annotations. Also shown is the percentage identity of amino acids between the newly annotated gene and its homolog in *E. coli* K-12 MG1655. Worksheet 9: Unique catabolic capabilities of each strain. The major difference driving nutrients discussed in the main text. Each nutrient source indicates if it can be used as a sole carbon or nitrogen source as well as if it can be catabolized in aerobic/anaerobic conditions. The count of strains indicates the number of *E. coli* and *Shigella* strains that can use each nutrient as a sole source of the designated essential element.

Dataset S2. Zip file of all 55 models in SBML format

Dataset S2

Models are available at the public BIGG database (19) (http://bigg.ucsd.edu).