EcoSal Plus
Cellular and Molecular Biology of
E. coli, Salmonella, and the *Enterobacteriaceae*

DOMAIN 10 BIOINFORMATICS AND SYSTEMS BIOLOGY

The EcoCyc Database

PETER D. KARP,[1] WAI KIT ONG,[1] SUZANNE PALEY,[1] RICHARD BILLINGTON,[1] RON CASPI,[1] CAROL FULCHER,[1] ANAMIKA KOTHARI,[1] MARKUS KRUMMENACKER,[1] MARIO LATENDRESSE,[1] PETER E. MIDFORD,[1] PALLAVI SUBHRAVETI,[1] SOCORRO GAMA-CASTRO,[2] LUIS MUÑIZ-RASCADO,[2] CÉSAR BONAVIDES-MARTINEZ,[2] ALBERTO SANTOS-ZAVALETA,[2] AMANDA MACKIE,[3] JULIO COLLADO-VIDES,[2] INGRID M. KESELER,[1] AND IAN PAULSEN[3]

[1]Bioinformatics Research Group, SRI International, Menlo Park, CA 94025

[2]Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, A.P. 565-, Cuernavaca, Morelos 62100, México

[3]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW 2109, Australia

ABSTRACT EcoCyc is a bioinformatics database available at EcoCyc.org that describes the genome and the biochemical machinery of *Escherichia coli* K-12 MG1655. The long-term goal of the project is to describe the complete molecular catalog of the *E. coli* cell, as well as the functions of each of its molecular parts, to facilitate a system-level understanding of *E. coli*. EcoCyc is an electronic reference source for *E. coli* biologists and for biologists who work with related microorganisms. The database includes information pages on each *E. coli* gene product, metabolite, reaction, operon, and metabolic pathway. The database also includes information on *E. coli* gene essentiality and on nutrient conditions that do or do not support the growth of *E. coli*. The website and downloadable software contain tools for analysis of high-throughput data sets. In addition, a steady-state metabolic flux model is generated from each new version of EcoCyc and can be executed via EcoCyc.org. The model can predict metabolic flux rates, nutrient uptake rates, and growth rates for different gene knockouts and nutrient conditions. This review outlines the data content of EcoCyc and of the procedures by which this content is generated.

Received: 05 June 2018
Accepted: 18 September 2018
Posted: 12 November 2018
Supercedes previous posting at http://www.asmscience.org/content/journal/ecosalplus/10.1128/ecosalplus.ESP-0009-2013
Editor: Susan T. Lovett, Department of Biology, Brandeis University, Waltham, MA
Citation: EcoSal Plus 2018; doi:10.1128/ecosalplus.ESP-0006-2018.
Correspondence: Peter D. Karp: pkarp@AI.SRI.COM

doi:10.1128/ecosalplus.ESP-0006-2018

INTRODUCTION

The EcoCyc (pronounced "eeko-sike," as in "*eco*logy" and "en*cyc*lopedia") database describes the genome and the biochemical machinery of *Escherichia coli* K-12 MG1655. EcoCyc is an electronic reference source that is designed to accelerate the work of *E. coli* biologists and of researchers who work with related microorganisms. The project's long-term goal is describing the complete molecular catalog of the *E. coli* cell, as well as the functions of each of its molecular parts, to facilitate a system-level understanding of *E. coli*. A steady-state metabolic flux model for *E. coli* K-12 MG1655 is also available through the EcoCyc.org website; the model is generated from each new version of EcoCyc.

This review provides an overview of EcoCyc's data content and of the procedures by which these data enter EcoCyc. It also provides a quick introduction to commonly used features of the EcoCyc website. Resources available to assist users in comprehensively learning the website and downloadable software are listed in "How to Learn More," below.

EcoCyc occupies a central role in the ecosystem of *E. coli* data resources. EcoCyc has benefitted from long-term and relatively stable funding, resulting in a very deep collection of content due to extensive curation. Its data content is also quite broad, spanning not only the functions of *E. coli* gene products, but also the *E. coli* metabolic network and regulatory network. EcoSal has created hyperlinks from most *E. coli* gene names in its articles to the corresponding EcoCyc gene pages. In addition, EcoCyc is now responsible for generating updates to the RefSeq and Genbank entries for *E. coli* K-12 MG1655.

EcoCyc supports several different modes of interactive use via both the EcoCyc.org website and in conjunction with the downloadable Pathway Tools (1) software.

- Browsing Mode: EcoCyc is an encyclopedic reference providing information about the biological roles of *E. coli* genes, metabolites, and pathways. Visualization tools, such as a genome browser, metabolic map display, and regulatory network diagram, aid in the comprehension of these complex data.
- Analysis Mode: EcoCyc facilitates the analysis of high-throughput data such as gene expression and metabolomics data via tools for enrichment analysis, and for visualizing omics data on our Omics Dashboard, metabolic map diagram, complete genome diagram, and regulatory network diagram.
- Modeling Mode: The EcoCyc metabolic flux model can predict growth or no growth of wild-type and knockout *E. coli* strains under different nutrient conditions.

Users of EcoCyc fall into several different groups. Experimental biologists use EcoCyc as an encyclopedic reference on genes, pathways, and regulation, and they use its omics data analysis tools to analyze gene expression and metabolomics data. Examples of papers citing EcoCyc in the analysis of functional genomics data include references 2 through 9.

Because the EcoCyc data are structured within a sophisticated ontology (10, 11) that is amenable to computational analyses, EcoCyc enables scientists to ask computational questions spanning the entire genome of *E. coli*, the known metabolic network of *E. coli*, the known transport complement of *E. coli*, the known genetic regulatory network of *E. coli*, and combinations thereof. Past work includes the use of EcoCyc to develop methods for studying path lengths within metabolic networks (12–14), in studies relating protein structure to the metabolic network (15, 16), and in analysis of the *E. coli* regulatory network (17–20).

The development of many new bioinformatics methods requires high-quality, gold-standard data sets for the training and validation of those methods. EcoCyc has been used as a gold-standard data set for the development of genome-context methods for predicting gene function (21, 22), operon-prediction methods (23, 24), the prediction of promoters and transcription start sites (25–27), regulatory network reconstruction (28), and the prediction of functional and direct protein-protein interactions (29–32). The EcoCyc metabolic data have been used for studies concerning predicted metabolic networks and growth prediction (33, 34), and for model checking of a symbiotic bacterium's metabolic network (35).

Metabolic engineers alter microbes to produce biofuels, industrial chemicals, and pharmaceuticals; to degrade toxic pollutants; and to sequester carbon (36–38). Metabolic engineers who use *E. coli* as their host organism consult EcoCyc to aid in optimizing the production of an end product through a better understanding of the metabolic network and its regulation and to predict undesirable side effects of a metabolic alteration. Metabolic engineering studies using EcoCyc include references 39 through 44.

EcoCyc data are available for download in multiple file formats (see http://biocyc.org/download.shtml) and can be queried programmatically via web services (see http://biocyc.org/web-services.shtml). In general, each file contains all the database instances for a given database class (e.g., genes, reactions, pathways).

The Pathway Tools software that underlies EcoCyc (1) is not specific to *E. coli*, but rather has been applied to manage genomic and biochemical data for thousands of

other organisms. EcoCyc is part of the larger BioCyc collection of Pathway/Genome Databases (PGDBs) that, as of August 2018, contains databases for 13,000 sequenced microbes.

BRIEF INTRODUCTION TO THE EcoCyc WEBSITE

Here we outline some of the more commonly used aspects of the EcoCyc website in a manner geared toward new users. The most common operation within EcoCyc is to search for information about a given gene. To find a gene page within EcoCyc:

- Open EcoCyc.org in a web browser
- Enter a gene name (e.g., *trpA*) in the quick search box in the upper right (to the left of the "Quick Search" button)
- Click the "Gene Search" button

The resulting gene page will provide a variety of information; the types of information available vary tremendously depending on what is known for each gene. Experiment with clicking on the tabs just below the first table (e.g., "GO Terms," "Essentiality"); each tab provides an additional dimension of information. The sequence of the gene and of its product may be retrieved using the Operations menu on the right side of the page, such as the command "Nucleotide Sequence."

Several other types of pages are available besides the gene page, including pathway pages, reaction pages, metabolite pages, transcription-unit pages, and growth-medium pages. One simple way to access these pages is by clicking on hyperlinks present in a gene page. For example, the first table in the *trpA* gene page provides links to associated reaction and pathway pages; both of those pages link to metabolite pages, which in turn link to all reactions and pathways in which that metabolite is a reactant, product, or regulator. A second way to access these pages is by doing a full quick search rather than the simpler gene search described earlier. For example, enter "pyruvate" into the quick search box and click "Quick Search" or press Enter. The result list is sorted by page type, and provides links to all objects within a page type whose common name or synonym includes "pyruvate." A number of other search tools are available that can help you describe what you are looking for much more precisely, in particular, the object-specific searches under the Search menu. For more information, see https://biocyc.org/PToolsWebsiteHowto.shtml#SearchHelp.

Some other commonly used tools are as follows. Most are intuitive to use after a little experimentation. Most of these tools are available for the other genomes in BioCyc.

- Genome browser: See menu item Genome → Genome Browser. A comparative version of the genome browser can be launched from a gene page to align regions of multiple genomes centered on orthologs of the gene; see the Operations menu on the right side of the gene page and use the command "Align in multi-genome browser."
- Launch a BLAST search against the *E. coli* genome using Search → BLAST Search. Searches are usually quite fast because a single genome is being searched.
- Full metabolic map diagram browser: See Metabolism → Cellular Overview. Search for genes, enzymes, and metabolites in this diagram using the Highlight commands within the Operations menu on the right side of this page. This diagram can be painted with transcriptomics, metabolomics, and other types of high-throughput data (much as traffic data are shown on Google Maps) using the command "Upload Data from File" in the Operations menu.
- Search for routes through the *E. coli* metabolic network using Metabolism → Metabolic Route Search.
- Regulatory network browser: See Genome → Regulatory Overview.
- The Omics Dashboard is a powerful tool for interpreting high-throughput data sets: See Analysis → Omics Dashboard, and see the Dashboard webinar at https://biocyc.org/webinar.shtml.
- SmartTables enable users to store, browse, and manipulate lists of objects (e.g., lists of genes or pathways). For example, with a few clicks you can define a SmartTable from a file containing a list of gene names, browse properties of the genes such as accession numbers and map positions, view the pathways in which the gene products function, and find the regulators of those genes. Learn more via the SmartTables webinar at https://biocyc.org/webinar.shtml.

OVERVIEW OF EcoCyc DATA CONTENT

The EcoCyc ontology includes a broad array of data types. Key to understanding the EcoCyc data and their presentation within the EcoCyc website and Pathway Tools is the notion of a database *class*, which describes a

specific type of data. For example, the class Genes provides the database definition of a gene, including the attributes (e.g., starting nucleotide position within the genome) and relationships (e.g., a gene and gene product are related by the Product relationship) of the class. Each specific gene within EcoCyc is stored in a single database *object*, or *frame*, that is an *instance* of the class Genes.

The types of data pages available through the EcoCyc website correspond loosely to the database classes within EcoCyc. For example, there is a gene class and a gene page, as well as a pathway class and a pathway page. However, one web page typically integrates information from multiple classes. For example, a pathway page integrates information from objects in the classes Pathways, Reactions, Genes, Proteins, and Chemicals.

Genome

EcoCyc contains the complete genome sequence of *E. coli* K-12 MG1655 (Genbank record version U00096.3) and describes the nucleotide position and function of all known protein-coding and RNA-coding *E. coli* genes and pseudogenes. Additional types of DNA and mRNA sites and regions present within EcoCyc (accessible by the website command Search → Search DNA or mRNA sites) are regulatory sites, origin of replication, phage attachment sites, transcription start sites, transcription factor binding sites, riboswitches, attenuators, terminators, transposons, and transcription units (database classes exist for each of these site types). Gene Ontology (GO) terms are assigned to gene products both by EcoCyc curators and by import of GO terms from UniProt (45). EcoCyc data on the essentiality of *E. coli* genes are described in "Essential Gene Information" (see below).

Proteome

EcoCyc describes all known monomers and multimeric protein complexes of *E. coli*. EcoCyc contains extensive annotation of the features of *E. coli* proteins (assigned by EcoCyc curators and imported from UniProt) such as phosphorylation sites, metal ion binding sites, and enzyme active sites. Relevant classes within EcoCyc include Polypeptides and Protein Complexes. Extensive information is encoded about metabolic enzymes, including their cofactors, activators, inhibitors, kinetic parameters, and subunit structure.

RNAome

EcoCyc describes all known RNAs (with the exception of messenger RNAs, which are not represented in EcoCyc) and protein-RNA complexes of *E. coli*. Relevant classes within EcoCyc include RNAs, rRNAs, tRNAs, and regulatory RNAs.

Regulation

EcoCyc contains the most complete description of the regulatory network of any organism. It covers *E. coli* operons, transcription start sites, transcription factors, transcription factor binding sites, attenuators, riboswitches, and small-RNA regulators, as well as substrate-level regulation of *E. coli* enzymes. Each molecular regulatory interaction is described as an instance of class Regulation, whose subclasses describe different types of regulation.

Metabolism

EcoCyc describes all known metabolic and signal-transduction pathways of *E. coli*; each pathway, each metabolic reaction, and each metabolite are encoded as separate database objects.

Membrane Transporters

EcoCyc annotates *E. coli* membrane transport proteins and the reactions that they mediate. The depiction of transport reactions aims to accurately represent the mechanisms by which *E. coli* moves both endogenous and exogenous substances across the inner and outer membrane. Transport proteins are classified according to International Union of Biochemistry and Molecular Biology recommendations (46) and relevant classes include a primary active transporter; an electrochemical potential-driven transporter, pores and channels; and a phosphotransferase system protein.

Gene Essentiality

EcoCyc includes several gene essentiality data sets. For more details, see section on "Essential Gene Information."

Growth Observations

EcoCyc integrates data on the growth of *E. coli* under many different growth media, as described in "Conditions of *E. coli* Growth and Nongrowth."

Database Links

EcoCyc contains extensive links to other biological databases containing protein and nucleic acid sequence data, bibliographic data, protein structures, and descriptions of different *E. coli* strains.

LITERATURE-BASED CURATION

Curation is the process of manually refining and updating a bioinformatics database. The EcoCyc project uses a *literature-based curation* approach in which database updates are based on evidence in the experimental literature. As of April 2018, EcoCyc version 22.0 encodes information from 35,000 publications. A staff of three full-time curators updates the annotation of the *E. coli* genome on an ongoing basis. The person-decades of curation effort applied to EcoCyc results in deeper and more accurate database content than uncurated resources such as KEGG (47) and PATRIC (48). For example, KEGG and PATRIC lack the extensive minireview summaries present within EcoCyc gene and pathway pages. KEGG also lacks the enzyme activator, inhibitor, and kinetics data provided in EcoCyc gene pages.

Curators collect gene, protein, pathway, and compound names and synonyms. They classify genes and gene products by using the Gene Ontology (49) and MultiFun (50) ontologies, and they classify pathways within the Pathway Tools pathway ontology. From the experimental literature, curators capture protein complex components and the stoichiometry of these subunits, cellular localization of polypeptides and protein complexes, experimentally determined protein molecular weights; enzyme activities and any enzyme prosthetic groups, cofactors, activators, or inhibitors. Operon structure and gene regulation information are encoded.

Curators author minireview summaries with extensive citations. Within the summaries for proteins, RNAs, pathways, and operons, curators add additional information not otherwise captured in the highly structured database fields of EcoCyc. For example, curators use the free-text summary sections to describe the overall function of a gene product, the phenotypes caused by mutation, depletion, or overproduction of each gene product; any known genetic interactions; protein domain architecture and structural studies; the similarity to other proteins; or any functional complementation experiments that have been described. Summaries can also be used to note cases in which the published reports present contradictory results. In such cases, both viewpoints will be presented with proper attribution. This approach strives to ensure that no information is lost.

EcoCyc entries are generally updated when new literature becomes available. Regular PubMed searches are used to generate lists of potentially curatable publications, which are then evaluated and prioritized for curation. Papers containing newly identified functions of gene products, as well as substantial advances in understanding the functions of known gene products, are given the highest priority for curation. Because the Pathway Tools software continues to evolve and to enable the addition of new data types, older entries are also being updated in a systematic fashion (e.g., each enzyme in a metabolic pathway) as time allows.

The transcriptional regulatory information in EcoCyc and RegulonDB is curated by the group of Dr. Julio Collado-Vides at the Universidad Nacional Autónoma de México (UNAM). Both databases include the same data content on transcriptional regulation of gene expression. The actual data curation occurs within EcoCyc, and the information is periodically propagated to RegulonDB (51).

STATISTICS ON EcoCyc CONTENT

Tables 1, 2, 3, and 4 present statistics on EcoCyc content. The listed numbers are current as of version 22.0, released in April 2018. Most changes in the numbers in these tables reflect changes in our knowledge of *E. coli*, but some changes are due to reconceptualization of data, reorganization of data (example: splitting a single meta-

Table 1 Genes and gene products in EcoCyc

Genes/proteins[a]	v17.5 (October 2013)	v22.0 (April 2018)
Genes	4,501	4,501
Protein-coding genes	4,284	4,278
tRNA genes	86	89
rRNA genes	22	22
Regulatory RNA genes	41	58
Other RNA genes	56	36
Pseudogenes	133	186
Protein complexes	995	1,064
Heteromultimeric protein complexes	290	303
Homomultimeric protein complexes	705	766
Protein features	23,114	35,381
Enzymes (excluding transporters)	1,245	1,327
Transporters	267	284

[a]Protein features are annotations of protein sites and regions such as enzyme active sites, metal ion binding sites, and transmembrane domains. A small number of insertion sequence elements are included in the count of genes but are not included in the subcategories of genes.

Table 2 Gene annotation status in EcoCyc

Gene annotation status	v17.5 (October 2013)	v22.0 (April 2018)
Genes of known or predicted molecular function[a]	3,127	3,237
Genes of known molecular function	2,710	2,866
Genes of predicted molecular function	417	371
Genes of unknown molecular function	1,374	1,264
Citations	25,406	35,024
Textbook equivalent pages of summaries	2,537	3,049

[a]Genes of known molecular function have experimental evidence for their assigned function, whereas genes of predicted molecular function have had their function predicted computationally.

bolic pathway into two smaller pathways to reflect evolutionary conservation more accurately), or changing definitions within the software. Many of these statistics can be computed using tools under the Analysis → Comparative Analysis website menu.

CONDITIONS OF *E. COLI* GROWTH AND NONGROWTH

As of 2011, EcoCyc incorporates media that have been shown experimentally to support or not support growth of both wild-type and knockout strains of *E. coli* K-12.

Table 3 Reactions, compounds, and pathways in EcoCyc

Reactions/metabolites/pathways	v17.5 (October 2013)	v22.0 (April 2018)
Metabolic reactions	1,443	1,989
Transport (including electron transfer) reactions	379	509
Pathways	401	431
Small-molecule metabolism base pathways	291	318
Signaling pathways	29	31
Superpathways[a]	81	82
Metabolites	2,466	2,846
Metabolites that are substrates of enzyme-catalyzed reactions	1,331	1,101
Metabolites that are physiological enzyme regulators	121	116
Metabolites that are cofactors or prosthetic groups	56	45
Transported metabolites	274	355

[a]Superpathways are sets of base metabolic pathways connected via shared substrates.

Table 4 Regulation-related objects and interactions in EcoCyc

Transcriptional/translational regulation	v17.5 (October 2013)	v22.0 (April 2018)
Transcription units	4,510	3,560
Transcription start sites	3,777	3,857
Terminators	259	307
Transcription factors	194	210
Transcription factor binding sites	2,773	2,944
Instances of regulation of transcription initiation[a]	3,293	3,462
Instances of regulation by transcriptional attenuation	20	24
Instances of regulation of translation	146	246

[a]Each member of "Instances of Regulation of Transcription Initiation" describes a single regulatory interaction between a transcription factor and its binding site.

This work has two goals. First, a comprehensive encyclopedia of *E. coli* growth conditions will be assembled for experimentalists. The spectrum of environmental conditions supporting the growth of a bacterium is among its most important phenotypic traits. We cannot expect to understand the functions of all genes in an organism unless we understand the full range of the environments in which the cell can grow. Second, a comprehensive collection of *E. coli* growth media will drive more accurate systems biology modeling of *E. coli*. The larger the set of growth media against which these computational models are validated, the more accurate and comprehensive that the models will be.

EcoCyc captures approximately 50 published media that are used by *E. coli* laboratories; growth data are provided for some of these media. EcoCyc also records the results of high-throughput experiments using Biolog Phenotype Microarrays (PMs) that measure cell respiration as a sensitive indicator of microbial growth (52). The commercially available PM system for microorganisms provides a comprehensive set of phenotype tests for 379 nutrient conditions, including information on the ability to metabolize 190 carbon (C) compounds, 95 nitrogen (N) compounds, 59 phosphorus (P) compounds, and 35 sulfur (S) compounds. EcoCyc currently includes five sets of PM data from the following sources:

- B. Bochner and X. Lei, personal communication, 2012.

  Strain: *E. coli* K-12 BW30270 (*rph*+ [RNase PH] derivative of MG1655; the strains also show a PyrE

deficiency. Found to be *fnr*$^+$ as well, according to K. A. Datsenko and B. L. Wanner, unpublished results.)

This data set includes aerobic growth observations for the full complement of C, N, P, and S compounds that are included in the PM system plus growth observations for 95 C sources under anaerobic conditions.

- "Genome Scale Reconstruction of a *Salmonella* Metabolic Model," AbuOun et al., 2009 (53).

  Strain: *E. coli K*-12 MG1655 (American Type Culture Collection 700926)

  This data set includes growth observations for the full complement of C, N, P, and S compounds under aerobic conditions. Bacteria were pregrown on LB agar before the inoculation of Biolog plates and incubation at 37°C for 26 h. The Omnilog instrument (a specialized incubator plus reader) was used for data collection and analysis.
- "The Evolution of Metabolic Networks of *E. coli*," Baumler et al., 2011 (54).

  Strain: *E. coli K-12* MG1655

  This data set consists of growth observations for 95 C compounds under aerobic and anaerobic conditions. Bacteria were pregrown on Biolog Universal Growth Agar plus sheep blood (BUG-S) before the inoculation of Biolog plates and incubation at 37°C. Growth was monitored by measuring optical density at 600 nm with readings taken at 3, 6, 12, 24, and 48 h (D. Baumler, personal communication).
- "Addition of *Escherichia coli* K-12 growth-observation and gene essentiality data to the EcoCyc database," Mackie et al., 2013 (55).

  Strain: *E. coli* K-12 MG1655 (Coli Genetic Stock Center 7740).

  This data set consists of growth observations for the full complement of C, N, P, and S compounds under aerobic conditions. Bacteria were pregrown on either LB or R2A agar before inoculation of Biolog plates and incubation at 37°C for 48 h. The Omnilog instrument was used for data collection and analysis.
- "Comparative Multi-Omics Systems Analysis of *Escherichia coli* strains B and K-12," Yoon et al., 2012 (56).

  Strain: *E. coli* K-12 MG1655

  This data set consists of growth observations for the full complement of C, N, P, and S compounds under aerobic conditions. Bacteria were pregrown on BUG-S agar before the inoculation of Biolog plates and incubation at 37°C for 48 h. The Omnilog instrument was used for data collection and analysis.

Data on growth conditions can be accessed from the EcoCyc website by invoking the menu command Search → Growth Media and then clicking on the button "All Growth Media for this Organism." Individual media are shown in the initial table; PM data are shown in the following tables. The coloring of each box indicates the degree of growth observed under that condition. Three levels of growth are recorded: no growth, low growth, and growth (see legend that indicates the colors associated with each level of growth). Click on any growth medium to request a page describing its composition and to see genes that are essential or not essential for growth under that condition.

ESSENTIAL GENE INFORMATION

When essentiality data are available for a given gene, the EcoCyc gene page includes a table of the growth media under which that gene has been found to be either essential or not essential for growth. Clicking on the growth medium will navigate to a growth-medium page that lists all essentiality information for that growth medium.

As of 2011, EcoCyc incorporates five large-scale data sets on gene essentiality in *E. coli*. Gene essentiality information is useful for

- Predicting antibiotic targets for pathogenic bacteria.
- Guiding the design of minimal genomes.
- Validating genome-scale metabolic flux models. Model predictions can be compared with the experimental data recorded in EcoCyc to assess model accuracy.
- Providing clues regarding the functions of genes of unknown function, when essentiality varies depending on conditions of growth.

EcoCyc incorporates data on essentiality from the following publications:

- "Experimental Determination and System Level Analysis of Essential Genes in *Escherichia coli*, MG1655," Gerdes et al. (57).

  Strain: *E. coli* K-12 MG1655 (F^- λ^- *ilvG rfb-50 rph-1*)

  This study used a genetic footprinting technique with a Tn*5*-based transposome system and reported

unambiguous assessment of approximately 87% of *E. coli* open reading frames (ORFs) for essentiality. Six hundred twenty-six genes were identified as essential for aerobic growth in rich media, while 3,126 genes were dispensable. Note that the inability to obtain an insertion mutant by using this system may in some cases be a reflection of the nontargeted nature of transposon insertion rather than a reflection of gene essentiality. For this and other technical reasons, 327 genes were classified in this study as ambiguous with regard to essentiality.

- "Construction of *Escherichia coli* K-12 In-Frame, Single-Gene Knockout Mutants: The Keio Collection," Baba et al. (58) and corrections (59)

  Strain: *E. coli* K-12 BW25113 [*rpoS(Am) rph-1* λ^- *rrnB3* Δ*lacZ4787 hsdR514* Δ*(araBAD)567* Δ*(rhaBAD) 568 rph-1*]

  This study created 3,985 in-frame, single-gene deletion mutants by using the lambda RED recombinase system. Three hundred three genes were unable to be disrupted and were predicted to be essential for growth in rich media at 37°C. Note that, in some cases, there were secondary impacts from single-gene deletions, such as compensating suppressor mutations. There were also errors in some of the mutants described in this paper, which were later corrected (59). This study also profiled the growth of the mutants in minimal glucose MOPS (morpholinepropanesulfonic acid) media to identify genes that are conditionally essential under these conditions.

- "Experimental and Computational Assessment of Conditionally Essential Genes in *Escherichia coli*," Joyce et al. (60)

  Strain: *E. coli* K-12 BW25113 [*rpoS(Am) rph-1* λ^- *rrnB3* Δ*lacZ4787 hsdR514* Δ*(araBAD)567* Δ*(rhaBAD) 568 rph-1*]

  This study used the Keio collection of single-gene knockout mutants and profiled them for growth on glycerol-supplemented minimal medium. One hundred nineteen genes were identified as essential for growth on glycerol. They also combined these observations with those made by Baba et al. (58) regarding the conditional essentiality of the mutants when grown on glucose-supplemented minimal media and were thus able to identify a conserved conditionally essential core of 94 genes that are required for *E. coli* K-12 to grow under minimal nutritional supplementation but are not essential for growth under rich conditions.

- "A Genome-Scale Metabolic Reconstruction for *Escherichia coli* K-12 MG1655 that Accounts for 1260 ORFs and Thermodynamic Information," Feist et al. (61)

  This publication used the experimental data regarding conditional gene essentiality from Joyce et al. (60) and from Baba et al. (58) and compared these data with the computationally predicted essential genes in their genome-scale metabolic reconstruction of *E. coli*. This data set is included in EcoCyc to facilitate the benchmarking of computational predictions of essentiality from the EcoCyc model with computations from the model of Feist et al. (61).

- "Multicopy Suppression Underpins Metabolic Evolvability," Patrick et al. 2007 (62)

  Strain: *E. coli* K-12 BW25113 [*rpoS(Am) rph-1* λ^- *rrnB3* Δ*lacZ4787 hsdR514* Δ*(araBAD)567* Δ*(rhaBAD) 568 rph-1*]

  This study used the conditionally essential gene sets identified by Baba et al. (58) and Joyce et al. (60) and tested them for their ability to form colonies on glucose M9 agar. They identified 107 genes that were conditionally essential under these conditions.

E. coli Experimental Literature

The EcoCyc.org website contains a searchable corpus of 50,000 experimental articles regarding *E. coli*, including articles cited by EcoCyc, and many others. For 38,000 of the articles, the full text is available and searchable along with the title and abstract; for the remaining articles, only the title and/or abstract are available and searchable. To search these articles, see Search → Search Full-text Articles.

EcoCyc METABOLIC FLUX MODEL

Constraint-based modeling techniques such as flux balance analysis (FBA) (63) are useful and widely used tools to study genome-scale metabolic networks. These network models contain all the metabolic reactions and associated genes for a given organism. Such models have found applications in six major categories, including metabolic engineering, model-driven discovery, cellular-phenotype prediction, analysis of biological network properties, studies of evolutionary processes, and, more

recently, modeling interspecies interactions (61, 64). SRI's application of the EcoCyc metabolic model will focus on the prediction of the following two types of cellular phenotypes: (i) growth of *E. coli* simulated under different nutrient conditions; the predicted growth rate, nutrient uptake rates, and rates of product secretion were compared with experimental results and (ii) growth of *E. coli* simulated under different gene knockouts; the resulting growth versus no-growth predictions were compared with experimental observations.

EcoCyc Metabolic Model Updates

The EcoCyc metabolic model draws its reactions from the EcoCyc database that is continuously being updated by curators. A new version of the EcoCyc database is released three times a year. The EcoCyc model serves as a check on the curation work performed because executing a model provides a system-wide check on the metabolism database content. Many curation anomalies such as in assignment of reaction directions or association of genes to reactions have been corrected by checking the model. With each release, the metabolic model is verified and updated as necessary. This approach ensures that the EcoCyc metabolic model reflects the current knowledge of *E. coli* more closely.

The EcoCyc metabolic model is distinct from the *E. coli* models developed by the Palsson group (61, 65–67), but they have much in common because EcoCyc and the iAF1260 models were partially unified in 2007 (61) and both groups consult the other's work when updating their models.

The biomass reaction is an integral component of an FBA model. It contains a list of all biosynthetic products that the model must synthesize to represent cellular growth (e.g., amino acids, nucleoside triphosphates). The more compounds are present in the biomass reaction, the more comprehensively the model is mimicking cellular metabolism. The "core" and "expanded" biomass reactions of the EcoCyc metabolic model ("EcoCyc-22.05") have been updated using the "core" and "wild-type" biomass reactions of the *E. coli* model recently published by Monk et al. ("iML1515") (67) as references. The core biomass reaction contains 93 precursors to essential cellular components of *E. coli*, while the expanded biomass contains 33 additional precursors to all the typical components found in wild-type *E. coli* (see **Supplemental Tables ST1** and **ST2**). The biomass reactions for the EcoCyc model were further modified according to our research findings and experimental data on gene essentiality. The lipids component for the expanded biomass reaction for EcoCyc was completely reworked based on the work by Oursel et al. (68). However, overall mass fractions of different biomass groups such as amino acids, DNA, RNA, lipids, etc. are kept the same as done in iML1515.

The number of biomass metabolites that the model synthesizes, and the number of metabolic reactions that are active (that carry non-zero flux) when synthesizing those biomass metabolites can be considered to define the scope of a metabolic model; the larger these quantities are, the larger a fraction of the cell's metabolic network is present in the model. Table 5 compares those quantities for three versions of the EcoCyc model under glucose

Table 5 Comparison of the number of biomass metabolites and the number of reactions carrying flux for EcoCyc-17.5, EcoCyc-18.0, and EcoCyc-22.05 under glucose aerobic and anaerobic conditions

Glucose	EcoCyc-17.5[a]	EcoCyc-18.0	EcoCyc-22.05
Aerobic			
No. of biomass metabolites (core biomass)	56	71	93
No. of reactions carrying flux (core biomass)	385	387	445
No. of biomass metabolites (expanded biomass)	N/A	82	126
No. of reactions carrying flux (expanded biomass)	N/A	412	567
Anaerobic			
No. of biomass metabolites (core biomass)	56	69	90
No. of reactions carrying flux (core biomass)	379	371	422
No. of biomass metabolites (expanded biomass)	N/A	78	123
No. of reactions carrying flux (expanded biomass)	N/A	380	552

[a]N/A, not available.

aerobic and anaerobic conditions. The big increase in number of biomass metabolites in EcoCyc-22.05 versus EcoCyc-18.0 is due to the inclusion of 20 amino acids as part of the protein group with the appropriate biomass coefficients while treating the other 20 compounds associated with the tRNA charging pathway as a separate biomass requirement based on gene essentiality. Molybdate, copper ion, and ferric ion were also added to the soluble pool group in the biomass reactions. See **Supplemental Methods** and **Supplemental Tables ST1** and **ST2** for details about the new biomass reactions.

EcoCyc Metabolic Model Access

There are two ways to run the EcoCyc metabolic model: the model can be run online at EcoCyc.org (see Metabolism → Run Metabolic Model) or on a computer by downloading and installing a Pathway Tools software configuration that includes EcoCyc and invoking the MetaFlux (69) modeling component of Pathway Tools (1) (see Tools → MetaFlux). When running the model via the web, users only have access to "solving mode," which is the ability to perform FBA. By creating a copy of one of the provided example files, a user can make any modification to the nutrient, secretion, reaction, or biomass metabolites sets, as well as upper and lower bounds on any flux before executing the new model. When running the model via the installed Pathway Tools software, users have access to "development mode" and "knockout mode" that can perform gap-filling and computational gene knockout experiments, respectively, in addition to "solving mode." In brief, gap filling is a procedure to identify reactions to add to a model to fill gaps in the metabolic network (70). Computational gene knockout experiment here refers to the procedure of simulating the effect of a gene knockout for all genes in the model.

EcoCyc provides several example files describing invocations of the metabolic model under different nutrient conditions. The .fba files are named according to the carbon source ("cs") and terminal electron acceptor ("tea") used, and are shown below:

1. cs-glucose-tea-oxygen
2. cs-glucose-tea-none
3. cs-glycerol-3p-tea-fumarate
4. cs-lactose-tea-none
5. cs-oleate-tea-nitrate
6. cs-tryptophan-tea-none

Each successful FBA run generates a log file (.log) containing all the reactions included in the model and a solution file (.sol) containing the flux distribution of the FBA run. On both the web version and the desktop version, there are buttons to paint the solution of an FBA run onto the cellular overview (71) and the omics dashboard (72) (see **Supplemental Figures S1** and **S2**). Please see Chapter 9 of the Pathway Tools Website User's Guide (http://biocyc.org/PToolsWebsiteHowto.shtml) for the web version or Chapter 8 of the Pathway Tools User's Guide for the desktop version for more information about running metabolic models. For the desktop version, the example FBA files can be found within the installed Pathway Tools directory tree at pathway-tools/aic-export/pgdbs/biocyc/ecocyc/{VERSION}/data/fba/fba-examples/.

To obtain version 22.05 of the EcoCyc model, execute the Pathway Tools license agreement (see http://biocyc.org/download.shtml), then follow the download instructions you will receive by email, and download the version 22.05 configuration listed on the download page.

EcoCyc Metabolic Model Validation

To ensure the EcoCyc metabolic model predictions are still reliable, two validation steps were performed: (i) a comparison of the predicted extracellular fluxes of the latest EcoCyc model, EcoCyc-22.05, a previously published EcoCyc model ("EcoCyc-17.5"), and the iML1515 model against experimental chemostat data under glucose aerobic (73) and anaerobic (74, 75) conditions; (ii) a comparison of the predicted gene essentiality of EcoCyc-22.05_BW with single-gene knockout mutant experimental data under glucose and glycerol aerobic conditions.

Extracellular flux prediction comparison

MetaFlux was used to perform FBA on the EcoCyc-based models EcoCyc-17.5 and EcoCyc-22.05 while COBRApy (76) was used for iML1515 (see **Supplemental Methods**). Flux predictions were obtained for aerobic and anaerobic growth on glucose. In all cases, the uptake rate of glucose is set to an upper bound reflecting experimental uptake rates (i.e., 3.008 mmol/gDW/h for aerobic glucose and 10 mmol/gDW/h for anaerobic glucose, where gDW stands for gram dry weight). Oxygen uptake rate is set to 0 mmol/gDW/h under the anaerobic condition. All other nutrient sources are constrained to an arbitrarily high upper bound (3000 mmol/gDW/h in

Table 6 Comparison of experimental aerobic glucose-limited chemostat growth data with EcoCyc-17.5, EcoCyc-22.05, and iML1515 model predictions

Aerobic fluxes[a]	Experimental	EcoCyc-17.5	EcoCyc-22.05	iML1515
Growth rate (1/h)	0.300	0.272	0.256	0.246
Glucose uptake (mmol/gDW/h)	3.008	3.008	3.008	3.008
O_2 uptake	7.413	4.472	6.842	7.436
NH_4 uptake	2.367	3.026	2.735	2.654
Sulfate uptake		0.068	0.064	0.062
Phosphate uptake		0.288	0.254	0.237
CO_2 production	7.385	5.480	7.485	7.961
H_2O production		13.026	14.177	14.451
H^+ production		2.582	2.313	2.258

[a]Growth rate is in units of h^{-1}. Metabolite uptake and production rates are in units of mmol/gDW/h where gDW stands for gram dry weight. Experimental data are calculated based on data from Kayser et al. (73).

MetaFlux and 1000 mmol/gDW/h in COBRApy). Additionally, all models here were simulated using the core version of the biomass reaction. These results are given in Tables 6 and 7 along with the experimental results from glucose-limited chemostat experiments on *E. coli* K-12 strains.

For the glucose aerobic condition (see Table 6), the extracellular flux results for EcoCyc-22.05 matched the experimental results better than did EcoCyc-17.5; however, the predicted growth rate is worse. For the glucose anaerobic condition (see Table 7), all extracellular flux results for EcoCyc-22.05 were further from the experimental results when compared with EcoCyc-17.5, with the exception of ethanol production. The prediction for growth rate using the new EcoCyc model is now much worse. However, for both conditions, the overall behavior of EcoCyc-22.05 and iML1515 was very similar. This is expected since their biomass reactions are similar. It is also believed that the new, larger growth-associated maintenance (GAM) and non–growth-associated maintenance (NGAM) terms used in EcoCyc-22.05 and obtained from iML1515 are the main reason for the reduced growth rates for both conditions.

Gene essentiality prediction accuracy

Gene essentiality analysis is a useful way to interrogate and validate a genome-scale metabolic network model. In

Table 7 Comparison of experimental anaerobic glucose-limited chemostat growth data with EcoCyc-17.5, EcoCyc-22.05, and iML1515 model predictions

Anaerobic fluxes	Experimental	EcoCyc-17.5	EcoCyc-22.05	iML1515
Growth rate (1/h)[a]	0.3	0.24	0.163	0.158
Glucose uptake (mmol/gDW/h)	10.0	10.00	10.000	10.000
O_2 uptake	0.0	0.00	0.000	0.000
NH_4 uptake		2.52	1.738	1.701
Sulfate uptake		0.06	0.041	0.040
Phosphate uptake		0.23	0.161	0.152
CO_2 production		0.04	0.055	−0.058
H_2O production		−1.92	−4.034	−4.591
H^+ production	27.8	27.50	28.223	28.612
Acetate production	7.5	8.29	8.823	8.831
Formate production	11.3	17.37	17.930	18.228
Succinate production	1.2	0.00	0.000	0.053
Ethanol production	8.7	8.13	8.503	8.746

[a]Growth rate is in units of h^{-1}. Metabolite uptake and production rates are in units of mmol/gDW/h where gDW stands for gram dry weight. Experimental data from Belaich et al. (74) via Varma et al. (75).

a gene essentiality analysis, every gene associated with a valid reaction in a model is knocked out one at a time, and FBA is performed on the resulting model. There are two ways a gene can be associated with a reaction in EcoCyc: genes whose products are enzymes that catalyze a reaction, or genes whose products are substrates of a reaction (e.g., acyl-carrier protein). A gene knockout can be simulated using a metabolic model by removing any reactions that are associated with the gene to be knocked out that do not have other isozymes or alternative substrates associated with it. Note that genes that are present in the database but not associated with any reaction of small-molecule metabolism (e.g., transcription factors, RNA polymerases, or ribosome subunits) are omitted from this analysis.

Here, the gene essentiality analysis performed on a previously published version of the EcoCyc metabolic model, specifically EcoCyc-18.0 (77), was repeated on EcoCyc-22.05_BW. The suffix "_BW" after the model name here indicates that additional reactions were removed from the model to represent genes lost in the *E. coli* BW25113 strain (relative to MG1655) which is used in the gene essentiality experiments (see **Supplemental Methods** for list of reactions removed). The two experimental essentiality data sets used to validate the models include (i) the deletion study by Baba et al. (58) as updated by Yamamoto et al. (59) that tested the Keio collection of *E. coli* BW25113 single-gene deletion strains for growth on LB rich media and MOPS minimal media with 0.4% glucose, and (ii) the deletion study by Joyce et al. (60) that tested the same Keio collection of *E. coli* BW25113 single-gene knockout mutants for growth on M9 minimal medium with 1% glycerol. See **Supplemental Methods** for additional details on how the experimental data are categorized as essential and nonessential gene deletions.

As done previously, single-gene knockout simulations were performed on the 1,671 genes in EcoCyc-22.05_BW on aerobic glucose and glycerol conditions to see if each simulated mutant is capable of producing the core and expanded biomass metabolite sets. A gene is considered essential if the FBA simulation resulted in no growth and nonessential if it resulted in a positive growth rate. These results were compared with the corresponding experimental results, and any mispredictions found were addressed with additional literature-based manual curation of EcoCyc. The final essentiality prediction results after curation are summarized in Table 8 for glucose and Table 9 for glycerol. In both tables, false positives are when an FBA simulation predicts growth yet experiment shows no growth; false negatives are when an FBA simulation predicts no growth yet experiment shows growth. The overall prediction accuracy for the aerobic glucose condition using the core biomass reaction is 94.3% (compared with 95.2% for EcoCyc-18.0), and the accuracy for aerobic glycerol condition is 93.2% (compared with 94.5% for EcoCyc-18.0). See **Supplemental Table ST3** for a full list of genes considered and their experimental and computationally predicted essentiality results.

Table 8 Comparison of experimental gene essentiality results with computational EcoCyc-22.05_BW results for aerobic growth on MOPS medium with 0.4% glucose

KO growth on glucose (sim/exp)[a]	Core biomass	Expanded biomass
True positive (growth/growth)	1,377	1,330
False positive (growth/no growth)	70	66
False negative (no growth/growth)	25	72
True negative (no growth/no growth)	199	203
Overall accuracy	94.3%	91.7%

[a]KO, knockout; sim, simulation; exp, experiment.

The overall prediction accuracy in the latest EcoCyc metabolic model is slightly reduced compared with the previously published model, primarily due to two broad issues: database changes and software changes. For database changes, curators are constantly updating the EcoCyc database, which may include addition/deletion of reactions/genes, modification of gene to protein to reaction associations, changes to the compound ontology, and modification of reaction direction. Any of these changes may affect FBA simulation results. As for software changes, the Pathway Tools software is also constantly being improved, and sometimes this may involve changes to how data are structured and represented. For

Table 9 Comparison of experimental gene essentiality results with computational EcoCyc-22.05_BW results for aerobic growth on MOPS medium with 1% glycerol

KO growth on glycerol (sim/exp)[a]	Core biomass	Expanded biomass
True positive (growth/growth)	1,364	1,315
False positive (growth/no growth)	76	73
False negative (no growth/growth)	38	87
True negative (no growth/no growth)	187	190
Overall accuracy	93.2%	90.4%

[a]KO, knockout; sim, simulation; exp, experiment.

example, since the previous EcoCyc model was published, the concept of overall reaction and subreactions was introduced where subreactions are a more detailed breakdown of the enzymatic and nonenzymatic steps of the overall reaction. This change led to the misprediction of several mutants because some genes were not correctly associated with their overall reaction or subreaction.

More generally, model predictions can differ from experimental measurements owing to a number of reasons, including the operation of additional, unmodeled reactions and metabolites; existing reactions operating in a different fashion from the model (e.g., the model contains a "perfect" respiratory electron-transfer chain without the possibility of reactive oxygen-species generation); the presence of regulation or of product inhibition that deactivates reactions or limits their throughput; and differences in optimization objective functions depending on the specified feed source.

leuB and *glmS* are specific examples of genes for which additional curated information in the newer model resulted in mispredictions:

leuB (b0073), whose product is the enzyme that catalyzes the 3-isopropylmalate dehydrogenase reaction (3-ISOPROPYLMALDEHYDROG-RXN), was identified as a false-positive result (i.e., simulation predicted the gene to be nonessential, but it was categorized as essential experimentally). Simulation predicted it to be nonessential because there is an isozyme associated with this reaction, specifically *dmlA* (b1800). However, under glucose aerobic conditions, the expression of *dmlA* is repressed and thus would not be able to serve as an isozyme for this essential reaction in reality. This sort of regulatory behavior is not currently captured by the EcoCyc metabolic model, and, thus, the model is unable to make the correct prediction.

glmS (b3729), whose product is the enzyme that catalyzes the L-glutamine–D-fructose 6-phosphate aminotransferase reaction (L-GLN-FRUCT-6-P-AMINOTRANS-RXN), was identified as a false-positive result. This result is mainly due to the glucosamine-6-phosphate deaminase reaction (GLUCOSAMINE-6-P-DEAMIN-RXN) being marked as reversible. According to the gene summary page for *nagB* (b0678), "Overexpression of *nagB* can suppress the defect of a *glmS* mutant, showing that glucosamine-6-phosphate deaminase can act in the biosynthetic direction, supplying glucosamine-6-phosphate for the synthesis of UDP-GlcNAc." Since the model is currently unable to account for reaction directions based on specific conditions, GLUCOSEAMINE-6-P_DEAMIN-RXN is treated as reversible, and, thus, L-GLN-FRUCT-6-P-AMINOTRANS-RXN is no longer essential.

UPDATE FREQUENCY

The EcoCyc.org and BioCyc.org websites and downloadable files are updated three times per year. A faster, more powerful version of EcoCyc that you can install locally on your computer (Macintosh, PC/Windows, PC/Linux) is released semiannually.

DATA SOURCES INCORPORATED INTO EcoCyc

EcoCyc includes data imported from the following bioinformatics databases. In most cases, the data are reimported once or twice per year. It is noted that many literature references within EcoCyc were obtained from PubMed.

UniProt Features

UniProt protein features (the UniProt KB term is sequence annotations) from the complete proteome of *E. coli* K-12 MG1655 in SwissProt are imported into EcoCyc for every EcoCyc release. All protein features with experimental or nonexperimental evidence qualifiers are imported, with the exception of the following types: turn, helix, beta strand, and coiled-coil. The chain type is only imported if it does not span the entire length of the protein. Examples of imported feature types include catalytic domains, phosphorylation sites, and metal ion binding sites. Citations associated with UniProt protein features are imported if they include an associated PubMed ID.

Gene Ontology

Gene Ontology (GO) and its applications are described in more detail in reference 78. Since the summer of 2008, a file containing all *E. coli* K-12 GO term annotations called gene_association.ecocyc has been generated periodically, that may be obtained from the Gene Ontology Consortium website.

GO annotation is a standard part of EcoCyc's manual literature-based curation process. The GO annotations are added to the database objects that represent the functional gene products or multimers, not directly to the

gene objects. This approach models the biology more accurately because it indicates exactly which form of the gene product has the specified GO function. In parallel, manual annotation of *E. coli* genes with GO is ongoing at EcoliWiki. On a regular basis, the GO annotations are merged. The latest UniProt and EcoliWiki annotations are imported into EcoCyc. Because the GO Consortium does not accept electronic annotations as part of the gene association file if the annotations are more than 1 year old, these UniProt annotations are reimported into EcoCyc on a regular basis.

EcoCyc incorporates many electronic and experimental GO term annotations of *E. coli* K-12 gene products obtained from the "UniProt [multispecies] GO Annotations @ EBI" file downloaded from the Gene Ontology Consortium. When this import was first performed in 2007, approximately 30,000 new IEA ("Inferred from Electronic Annotation") GO term assignments were added to EcoCyc, along with approximately 1,000 assignments with experimental evidence codes including assignments from high-throughput protein interaction studies.

A gene association file is generated from each EcoCyc release. This file is sent to Drs. J. Hu and D. Siegele at Texas A&M for further processing. They incorporate annotations made in the EcoliWiki wiki-based community annotation system since the last EcoCyc update to the file, along with annotations containing qualifiers (mainly contributes to) not supported by EcoCyc. Only those annotations that are complete by GO Consortium standards are extracted from EcoliWiki; incomplete annotations are left with the hope that community members will eventually complete them. Hu and Siegele run the GO Consortium validation scripts and deposit the file with the GO Consortium via their Concurrent Versioning System.

GenBank

The GenBank record U00096, produced by the Blattner laboratory in October 1997, was the source of the original *E. coli* MG1655 genome sequence and annotation incorporated by EcoCyc.

A corrected nucleotide sequence was deposited in GenBank as U00096.2 in 2004, and the revised sequence was incorporated into EcoCyc as of version 8.6 (November 2004). The revised genome annotation published in reference 79 was incorporated into EcoCyc in version 10.0 (March 2006).

A second update to the nucleotide sequence was deposited in GenBank as U00096.3 in 2013. This update reflected corrections to precisely represent the sequence of the *E. coli* K-12 MG1655 strain deposited in culture collections as the sequenced strain. The revised sequence was incorporated into EcoCyc as of version 20.0 (May 2016).

RefSeq Collaboration

EcoCyc is involved in a collaboration to update the genome annotation of the GenBank (U00096.3) and RefSeq (NC_000913.3) entries for *E. coli* K-12 MG1655 on an ongoing basis. The primary collaborators currently include EcoCyc, UniProtKB/Swiss-Prot, Guy Plunkett, and NCBI. The collaborators routinely share their data and resolve data conflicts. The updates of gene names, gene positions, and gene product names are shared among all partners. The updated RefSeq and GenBank entries will be generated using software that extracts all data in these entries (e.g., gene names and nucleotide coordinates, protein names) from EcoCyc. Updated GenBank and RefSeq entries are expected to be created in late 2018, and at regular intervals thereafter.

MetaCyc

The EcoCyc and MetaCyc databases exchange data as part of the release processes for both databases. The updates that have occurred to enzymes, genes, pathways, reactions, and metabolites are exchanged between the databases based on automated comparisons of update dates to ensure that the latest information and corrections are propagated between the databases.

EcoCyc ACCESSION NUMBERS

This section summarizes the accession numbers present within EcoCyc. All EcoCyc objects (e.g., genes, metabolites, pathways, DNA sites) have unique identifiers. In some cases, the identifiers for a given type of object follow more than one naming convention, usually because the convention was changed at a later time but previously assigned identifiers were kept to minimize disruption to users. Accession numbers assigned by other databases are also incorporated whenever possible to facilitate interoperability.

Gene Accession Numbers

Three systems of accession numbers are typically available for genes within EcoCyc. Any of these accession numbers may be used when querying EcoCyc genes "by name" and in the website Quick Search, and may be used

to identify genes in EcoCyc transcriptomics data import tools such as the Omics Dashboard.

- EcoCyc ID: The EcoCyc project assigns unique identifiers to each gene that, for historical reasons, are of variable syntax, and are of the form "Gnnnn," "EGnnnnn," or "G0-nnnnn." EcoCyc IDs are stored as the frame unique identifier of the EcoCyc gene object.
- B-numbers: Originally assigned by the Blattner laboratory as part of the *E. coli* genome project, the b-number identifiers are of the form "bnnnn." B-numbers were originally assigned sequentially along the genome. When a gene object is removed from the genome because of a decision that insufficient evidence for the existence of that gene is available, that b-number is retired and is not reused. When new genes are added to the genome, they are assigned the next highest available b-number. Thus, b-numbers are no longer purely sequential along the genome. B-numbers are stored in the EcoCyc gene slot Accession-1.
- ECK numbers: ECK numbers were assigned to the *E. coli* K-12 MG1655 and W3110 genomes in 2005 in an attempt to provide shared accession numbers for genes common to the two genomes (79). ECK numbers are stored in the EcoCyc gene slot Accession-2. For only the first 18 or so genes in the *E. coli* K-12 MG1655 genome (starting at the origin of replication where the nucleotide coordinate is zero) are the b-number and ECK number the same number; for subsequent genes, the numbers have diverged. That is, mokC=b0018=ECK0018, but for numbers after 0018, the b and ECK accession numbers for each gene differ.

OTHER *E. COLI* AND *SHIGELLA* PATHWAY/GENOME DATABASES IN BioCyc

BioCyc version 22.0 (2018) includes PGDBs for 524 *E. coli* and 25 *Shigella* strains. Most of these PGDBs were generated computationally and lack the extensive manual literature-based curation of the EcoCyc K-12 database. To select a given genome for querying in the BioCyc website, click on the words "change organism database" under the Quick Search and Gene Search buttons in the upper right corner of most EcoCyc web pages.

Two of these PGDBs have undergone additional curation: the BioCyc PGDBs for strains *E. coli* W3110 and for *E. coli* B str. REL606. Both strains underwent a computational annotation-normalization procedure in which gene names, product names, heteromultimeric protein complexes, and Gene Ontology terms were propagated from EcoCyc to their orthologous genes in these other two strains. The orthologs were computed by SRI as bidirectional best-BLAST hits with additional manual review and curation. The propagation was performed under the assumption that genome-annotation pipelines typically introduce syntactically large but semantically insignificant variation in the naming of genes and gene products. In addition, *E. coli* B str. REL606 underwent literature-based curation by SRI to incorporate experimental information regarding the genes and pathways present in this strain but not in the EcoCyc strain MG1655.

WE ENCOURAGE YOUR FEEDBACK

Feedback from the scientific community has proved invaluable to improving EcoCyc during its many years of development. We strongly encourage your comments and suggestions for improvements in all areas, including:

- The database content of EcoCyc
- The presentation of information within the EcoCyc website
- The analysis tools provided in conjunction with EcoCyc
- The performance of the EcoCyc website

If you see an error or omission within EcoCyc, please report it by using the "Provide Feedback" button near the top of every data page. Please email suggestions or questions to BioCyc support at biocyc-support@ai.sri.com.

During every EcoCyc release, we email a summary of new developments to our BioCyc users mailing list and post to the BioCyc Facebook and Twitter feeds. To subscribe to the mailing list, please see http://biocyc.org/subscribe.shtml.

HOW TO LEARN MORE

- How to use a Pathway Tools website such as EcoCyc
- Instructional videos on how to use EcoCyc
- Pathway/Genome Database Concepts Guide
- Publications on EcoCyc: References (77, 80–94)

- BioCyc User's Guide
- MetaCyc User's Guide
- Guide to the Pathway Tools Schema
- How to download the Pathway Tools software and organism flat-file databases

HOW TO CITE EcoCyc

Please cite EcoCyc in publications that benefit from the use of the EcoCyc database or website. Please cite EcoCyc as the most recent *Nucleic Acids Research* Database issue article, currently: Keseler et al. 2017, *Nucleic Acids Res* **45:** D543–D550.

ACKNOWLEDGMENTS

Monica Riley led the curation of EcoCyc for many years, from its inception. Her efforts created the content for the first organism-scale metabolic database. John Ingraham and Robert Gunsalus were valued advisors to EcoCyc for many years. We thank the scientists who have contributed corrections and suggestions to EcoCyc over the years, and we thank the scientists who have served on the EcoCyc Steering Committee for their many valuable suggestions. Many contributors to EcoCyc are listed on the EcoCyc credits page.

The development of EcoCyc is funded by NIH grants GM77678 and GM71962 from the NIH National Institute of General Medical Sciences.

REFERENCES

1. Karp PD, Latendresse M, Paley SM, Krummenacker M, Ong QD, Billington R, Kothari A, Weaver D, Lee T, Subhraveti P, Spaulding A, Fulcher C, Keseler IM, Caspi R. 2016. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform* **17:**877–890 http://dx.doi.org/10.1093/bib/bbv079.

2. Kim KS, Lee S, Ryu CM. 2013. Interspecific bacterial sensing through airborne signals modulates locomotion and drug resistance. *Nat Commun* **4:**1809 http://dx.doi.org/10.1038/ncomms2789.

3. Bower JM, Gordon-Raagas HB, Mulvey MA. 2009. Conditioning of uropathogenic *Escherichia coli* for enhanced colonization of host. *Infect Immun* **77:**2104–2112 http://dx.doi.org/10.1128/IAI.01200-08.

4. Rhodius V, Van Dyk TK, Gross C, LaRossa RA. 2002. Impact of genomic technologies on studies of bacterial gene expression. *Annu Rev Microbiol* **56:**599–624 http://dx.doi.org/10.1146/annurev.micro.56.012302.160925.

5. Gonzalez R, Tao H, Purvis JE, York SW, Shanmugam KT, Ingram LO. 2003. Gene array-based identification of changes that contribute to ethanol tolerance in ethanologenic *Escherichia coli*: comparison of KO11 (parent) to LY01 (resistant mutant). *Biotechnol Prog* **19:**612–623 http://dx.doi.org/10.1021/bp025658q.

6. Taoka M, Yamauchi Y, Shinkawa T, Kaji H, Motohashi W, Nakayama H, Takahashi N, Isobe T. 2004. Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. *Mol Cell Proteomics* **3:**780–787 http://dx.doi.org/10.1074/mcp.M400030-MCP200.

7. Goswami M, Narayana Rao AVSS. 2018. Transcriptome profiling reveals interplay of multifaceted stress response in *Escherichia coli* on exposure to glutathione and ciprofloxacin. *mSystems* **3:**e00001–e00018 http://dx.doi.org/10.1128/mSystems.00001-18.

8. Piazza I, Kochanowski K, Cappelletti V, Fuhrer T, Noor E, Sauer U, Picotti P. 2018. A map of protein-metabolite interactions reveals principles of chemical communication. *Cell* **172:**358–372.e23 http://dx.doi.org/10.1016/j.cell.2017.12.006.

9. Sharma P, Haycocks JRJ, Middlemiss AD, Kettles RA, Sellars LE, Ricci V, Piddock LJV, Grainger DC. 2017. The multiple antibiotic resistance operon of enteric bacteria controls DNA repair and outer membrane integrity. *Nat Commun* **8:**1444 http://dx.doi.org/10.1038/s41467-017-01405-7.

10. Karp PD. 2000. An ontology for biological function based on molecular interactions. *Bioinformatics* **16:**269–285 http://dx.doi.org/10.1093/bioinformatics/16.3.269.

11. Miller SJ. 2013. Introduction to Ontology Concepts and Terminology. Presentation. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2013/paper/download/140/105.

12. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. 2002. Hierarchical organization of modularity in metabolic networks. *Science* **297:**1551–1555 http://dx.doi.org/10.1126/science.1073374.

13. Simeonidis E, Rison SC, Thornton JM, Bogle ID, Papageorgiou LG. 2003. Analysis of metabolic networks using a pathway distance metric through linear programming. *Metab Eng* **5:**211–219 http://dx.doi.org/10.1016/S1096-7176(03)00043-0.

14. Arita M. 2004. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA* **101:**1543–1547 http://dx.doi.org/10.1073/pnas.0306458101.

15. Jardine O, Gough J, Chothia C, Teichmann SA. 2002. Comparison of the small molecule metabolic enzymes of *Escherichia coli* and *Saccharomyces cerevisiae*. *Genome Res* **12:**916–929 http://dx.doi.org/10.1101/gr.228002.

16. Rison SC, Thornton JM. 2002. Pathway evolution, structurally speaking. *Curr Opin Struct Biol* **12:**374–382 http://dx.doi.org/10.1016/S0959-440X(02)00331-7.

17. Ma HW, Kumar B, Ditges U, Gunzer F, Buer J, Zeng AP. 2004. An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res* **32:**6643–6649 http://dx.doi.org/10.1093/nar/gkh1009.

18. Shen-Orr SS, Milo R, Mangan S, Alon U. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* **31:**64–68 http://dx.doi.org/10.1038/ng881.

19. Perez-Acle T, Fuenzalida I, Martin AJM, Santibañez R, Avaria R, Bernardin A, Bustos AM, Garrido D, Dushoff J, Liu JH. 2018. Stochastic simulation of multiscale complex systems with PISKaS: A rule-based approach. *Biochem Biophys Res Commun* **498:**342–351 http://dx.doi.org/10.1016/j.bbrc.2017.11.138.

20. Fang X, Sastry A, Mih N, Kim D, Tan J, Yurkovich JT, Lloyd CJ, Gao Y, Yang L, Palsson BO. 2017. Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities. *Proc Natl Acad Sci USA* **114:**10286–10291 http://dx.doi.org/10.1073/pnas.1702581114.

21. Karimpour-Fard A, Leach SM, Gill RT, Hunter LE. 2008. Predicting protein linkages in bacteria: which method is best depends on task. *BMC Bioinformatics* **9:**397 http://dx.doi.org/10.1186/1471-2105-9-397.

22. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. 2004. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5:**R35 http://dx.doi.org/10.1186/gb-2004-5-5-r35.

23. Price MN, Huang KH, Alm EJ, Arkin AP. 2005. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* **33:**880–892 http://dx.doi.org/10.1093/nar/gki232.

24. Steinhauser D, Junker BH, Luedemann A, Selbig J, Kopka J. 2004. Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics* **20:**1928–1939 http://dx.doi.org/10.1093/bioinformatics/bth182.

25. Burden S, Lin YX, Zhang R. 2005. Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. *Bioinformatics* **21:**601–607 http://dx.doi.org/10.1093/bioinformatics/bti047.

26. Gordon L, Chervonenkis AY, Gammerman AJ, Shahmuradov IA, Solovyev VV. 2003. Sequence alignment kernel for recognition of promoter regions. *Bioinformatics* **19:**1964–1971 http://dx.doi.org/10.1093/bioinformatics/btg265.

27. Umarov RK, Solovyev VV. 2017. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One* **12:**e0171410 http://dx.doi.org/10.1371/journal.pone.0171410.

28. Fu Y, Jarboe LR, Dickerson JA. 2011. Reconstructing genome-wide regulatory network of *E. coli* using transcriptome data and predicted transcription factor activities. *BMC Bioinformatics* **12:**233 http://dx.doi.org/10.1186/1471-2105-12-233.

29. Watanabe RL, Morett E, Vallejo EE. 2008. Inferring modules of functionally interacting proteins using the Bond Energy Algorithm. *BMC Bioinformatics* **9:**285 http://dx.doi.org/10.1186/1471-2105-9-285.

30. Muley VY, Ranjan A. 2012. Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. *PLoS One* **7:**e42057 http://dx.doi.org/10.1371/journal.pone.0042057.

31. Moreno-Hagelsieb G, Jokic P. 2012. The evolutionary dynamics of functional modules and the extraordinary plasticity of regulons: the *Escherichia coli* perspective. *Nucleic Acids Res* **40:**7104–7112 http://dx.doi.org/10.1093/nar/gks443.

32. Maheshwari S, Brylinski M. 2017. Across-proteome modeling of dimer structures for the bottom-up assembly of protein-protein interaction networks. *BMC Bioinformatics* **18:**257 http://dx.doi.org/10.1186/s12859-017-1675-z.

33. Kastenmüller G, Schenk ME, Gasteiger J, Mewes HW. 2009. Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. *Genome Biol* **10:**R28 http://dx.doi.org/10.1186/gb-2009-10-3-r28.

34. Kumar VS, Maranas CD. 2009. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLOS Comput Biol* **5:**e1000308 http://dx.doi.org/10.1371/journal.pcbi.1000308.

35. Thomas GH, Zucker J, Macdonald SJ, Sorokin A, Goryanin I, Douglas AE. 2009. A fragile metabolic network adapted for cooperation in the symbiotic bacterium *Buchnera aphidicola*. *BMC Syst Biol* **3:**24 http://dx.doi.org/10.1186/1752-0509-3-24.

36. Frazier ME, Johnson GM, Thomassen DG, Oliver CE, Patrinos A. 2003. Realizing the potential of the genome revolution: the genomes to life program. *Science* **300:**290–293 http://dx.doi.org/10.1126/science.1084566.

37. Bailey JE. 1991. Toward a science of metabolic engineering. *Science* **252:**1668–1675 http://dx.doi.org/10.1126/science.2047876.

38. Stephanopoulos G, Vallino JJ. 1991. Network rigidity and metabolic engineering in metabolite overproduction. *Science* **252:**1675–1681 http://dx.doi.org/10.1126/science.1904627.

39. Arense P, Bernal V, Charlier D, Iborra JL, Foulquié-Moreno MR, Cánovas M. 2013. Metabolic engineering for high yielding L(-)-carnitine production in *Escherichia coli*. *Microb Cell Fact* **12:**56 http://dx.doi.org/10.1186/1475-2859-12-56.

40. Jantama K, Zhang X, Moore JC, Shanmugam KT, Svoronos SA, Ingram LO. 2008. Eliminating side products and increasing succinate yields in engineered strains of *Escherichia coli* C. *Biotechnol Bioeng* **101:**881–893 http://dx.doi.org/10.1002/bit.22005.

41. Weber J, Hoffmann F, Rinas U. 2002. Metabolic adaptation of *Escherichia coli* during temperature-induced recombinant protein production: 2. Redirection of metabolic fluxes. *Biotechnol Bioeng* **80:**320–330 http://dx.doi.org/10.1002/bit.10380.

42. Averesch NJH, Martínez VS, Nielsen LK, Krömer JO. 2018. Toward synthetic biology strategies for adipic acid production: an in silico tool for combined thermodynamics and stoichiometric analysis of metabolic networks. *ACS Synth Biol* **7:**490–509 http://dx.doi.org/10.1021/acssynbio.7b00304.

43. Delépine B, Duigou T, Carbonell P, Faulon J-L. 2018. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metab Eng* **45:**158–170 http://dx.doi.org/10.1016/j.ymben.2017.12.002.

44. Ovsienko MV, Fedorova EN, Doroshenko VG. 2018. Vanillin resistance induced by BssS overexpression in *Escherichia coli*. *Appl Biochem Microbiol* **54:**21–25 http://dx.doi.org/10.1134/S0003683818010088.

45. UniProt Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* **41:**D43–D47.

46. Busch W, Saier MH Jr, International Union of Biochemistry and Molecular Biology (IUBMB). 2004. The IUBMB-endorsed transporter classification system. *Mol Biotechnol* **27:**253–262 http://dx.doi.org/10.1385/MB:27:3:253.

47. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**(D1)**:**D353–D361 http://dx.doi.org/10.1093/nar/gkw1092.

48. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D, Mao C, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJ, Yoo HS, Zhang C, Zhang Y, Sobral BW. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* **42**(D1)**:**D581–D591 http://dx.doi.org/10.1093/nar/gkt1099.

49. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nat Genet* **25:**25–29 http://dx.doi.org/10.1038/75556.

50. Serres MH, Riley M. 2000. MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb Comp Genomics* **5:**205–222 http://dx.doi.org/10.1089/mcg.2000.5.205.

51. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, Medina-Rivera A, Solano-Lira H, Bonavides-Martínez C, Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L, López-Fuentes A, Hernández-Koutoucheva A, Del Moral-Chávez V, Rinaldi F, Collado-Vides J. 2016. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res* **44**(D1)**:**D133–D143 http://dx.doi.org/10.1093/nar/gkv1156.

52. Bochner BR, Gadzinski P, Panomitros E. 2001. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res* **11:**1246–1255 http://dx.doi.org/10.1101/gr.186501.

53. AbuOun M, Suthers PF, Jones GI, Carter BR, Saunders MP, Maranas CD, Woodward MJ, Anjum MF. 2009. Genome scale reconstruction of a *Salmonella* metabolic model: comparison of similarity and differences with a commensal *Escherichia coli* strain. *J Biol Chem* **284:**29480–29488 http://dx.doi.org/10.1074/jbc.M109.005868.

54. Baumler DJ, Peplinski RG, Reed JL, Glasner JD, Perna NT. 2011. The evolution of metabolic networks of *E. coli*. *BMC Syst Biol* **5:**182 http://dx.doi.org/10.1186/1752-0509-5-182.

55. Mackie A, Paley S, Keseler IM, Shearer A, Paulsen IT, Karp PD. 2014. Addition of *Escherichia coli* K–12 growth-observation and gene essentiality data to the EcoCyc database. *J Bacteriol* **196:**982–988. doi:10.1128/JB.01209-13.

56. Yoon SH, Han MJ, Jeong H, Lee CH, Xia XX, Lee DH, Shim JH, Lee SY, Oh TK, Kim JF. 2012. Comparative multi-omics systems analysis of *Escherichia coli* strains B and K-12. *Genome Biol* **13:**R37 http://dx.doi.org/10.1186/gb-2012-13-5-r37.

57. Gerdes SY, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabási AL, Oltvai ZN, Osterman AL. 2003. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol* **185:**5673–5684 http://dx.doi.org/10.1128/JB.185.19.5673-5684.2003.

58. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2:**2006.0008. doi:10.1038/msb4100050.

59. Yamamoto N, Nakahigashi K, Nakamichi T, Yoshino M, Takai Y, Touda Y, Furubayashi A, Kinjyo S, Dose H, Hasegawa M, Datsenko KA, Nakayashiki T, Tomita M, Wanner BL, Mori H. 2009. Update on the Keio collection of Escherichia coli single-gene deletion mutants. *Mol Syst Biol* **5:**335 http://dx.doi.org/10.1038/msb.2009.92.

60. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BØ, Agarwalla S. 2006. Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *J Bacteriol* **188:**8259–8271 http://dx.doi.org/10.1128/JB.00740-06.

61. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3:**121 http://dx.doi.org/10.1038/msb4100155.

62. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. 2007. Multicopy suppression underpins metabolic evolvability. *Mol Biol Evol* **24:**2716–2722 http://dx.doi.org/10.1093/molbev/msm204.

63. Orth JD, Thiele I, Palsson BØ. 2010. What is flux balance analysis? *Nat Biotechnol* **28:**245–248 http://dx.doi.org/10.1038/nbt.1614.

64. McCloskey D, Palsson BO, Feist AM. 2013. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol* **9:**661 http://dx.doi.org/10.1038/msb.2013.18.

65. Reed JL, Vo TD, Schilling CH, Palsson BO. 2003. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* **4:**R54 http://dx.doi.org/10.1186/gb-2003-4-9-r54.

66. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ. 2011. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. *Mol Syst Biol* **7:**535 http://dx.doi.org/10.1038/msb.2011.65.

67. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, Feist AM, Palsson BO. 2017. iML1515, a knowledgebase that computes Escherichia coli traits. *Nat Biotechnol* **35:**904–908 http://dx.doi.org/10.1038/nbt.3956.

68. Oursel D, Loutelier-Bourhis C, Orange N, Chevalier S, Norris V, Lange CM. 2007. Lipid composition of membranes of *Escherichia coli* by liquid chromatography/tandem mass spectrometry using negative electrospray ionization. *Rapid Commun Mass Spectrom* **21:**1721–1728 http://dx.doi.org/10.1002/rcm.3013.

69. Latendresse M, Krummenacker M, Trupp M, Karp PD. 2012. Construction and completion of flux balance models from pathway databases. *Bioinformatics* **28:**388–396 http://dx.doi.org/10.1093/bioinformatics/btr681.

70. Latendresse M. 2014. Efficiently gap-filling reaction networks. *BMC Bioinformatics* **15:**225 http://dx.doi.org/10.1186/1471-2105-15-225.

71. Paley SM, Karp PD. 2006. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res* **34:**3771–3778 http://dx.doi.org/10.1093/nar/gkl334.

72. Paley S, Parker K, Spaulding A, Tomb JF, O'Maille P, Karp PD. 2017. The Omics Dashboard for interactive exploration of gene-expression data. *Nucleic Acids Res* **45:**12113–12124 http://dx.doi.org/10.1093/nar/gkx910.

73. Kayser A, Weber J, Hecht V, Rinas U. 2005. Metabolic flux analysis of *Escherichia coli* in glucose-limited continuous culture. I. Growth-rate-dependent metabolic efficiency at steady state. *Microbiology* **151:**693–706 http://dx.doi.org/10.1099/mic.0.27481-0.

74. Belaich A, Belaich JP. 1976. Microcalorimetric study of the anaerobic growth of *Escherichia coli*: growth thermograms in a synthetic medium. *J Bacteriol* **125:**14–18.

75. Varma A, Boesch BW, Palsson BO. 1993. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl Environ Microbiol* **59:**2465–2473.

76. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. 2013. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* **7:**74 http://dx.doi.org/10.1186/1752-0509-7-74.

77. Weaver DS, Keseler IM, Mackie A, Paulsen IT, Karp PD. 2014. A genome-scale metabolic flux model of Escherichia coli K-12 derived from the EcoCyc database. *BMC Syst Biol* **8:**79 http://dx.doi.org/10.1186/1752-0509-8-79.

78. Hu JC, Karp PD, Keseler IM, Krummenacker M, Siegele DA. 2009. What we can learn about *Escherichia coli* through application of Gene Ontology. *Trends Microbiol* **17:**269–278 http://dx.doi.org/10.1016/j.tim.2009.04.004.

79. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett G III, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL. 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res* **34:**1–9 http://dx.doi.org/10.1093/nar/gkj405.

80. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muñiz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, Kaipa P, Spaulding A, Pacheco J, Latendresse M, Fulcher C, Sarker M, Shearer AG, Mackie A, Paulsen I, Gunsalus RP, Karp PD. 2011. EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res* **39**(Database)**:** D583–D590 http://dx.doi.org/10.1093/nar/gkq1143.

81. Keseler IM, Bonavides-Martínez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, Peralta-Gil M, Santos-Zavaleta A, Shearer AG, Karp PD. 2009. EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res* **37**(Database)**:**D464–D470 http://dx.doi.org/10.1093/nar/gkn751.

82. Karp PD, Keseler IM, Shearer A, Latendresse M, Krummenacker M, Paley SM, Paulsen I, Collado-Vides J, Gama-Castro S, Peralta-Gil M, Santos-Zavaleta A, Peñaloza-Spínola MI, Bonavides-Martinez C, Ingraham J. 2007. Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res* **35:**7577–7590 http://dx.doi.org/10.1093/nar/gkm740.

83. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD. 2005. EcoCyc: a compre-

hensive database resource for *Escherichia coli*. *Nucleic Acids Res* **33:** D334–D337 http://dx.doi.org/10.1093/nar/gki108.

84. Karp PD, Arnaud M, Collado-Vides J, Ingraham J, Paulsen IT, Saier MHJ. 2004. The *E. coli* EcoCyc database: no longer just a metabolic pathway database. *ASM News* **70:**25–30.

85. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S. 2002. The EcoCyc Database. *Nucleic Acids Res* **30:**56–58 http://dx.doi.org/10.1093/nar/30.1.56.

86. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. 2000. The EcoCyc and MetaCyc databases. *Nucleic Acids Res* **28:**56–59 http://dx.doi.org/10.1093/nar/28.1.56.

87. Karp PD. 1999. Using the EcoCyc Database, p 269–280. *In* Bishop M (ed), *Nucleic Acid and Protein Databases and How To Use Them*. Academic Press, London, UK.

88. Karp PD, Riley M. 1999. EcoCyc: the resource and the lessons learned, p 47–62. *In* Letovsky S (ed), *Bioinformatics Databases and Systems*. Kluwer Academic Publishers, Norwell, MA.

89. Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M. 1999. Eco Cyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res* **27:**55–58 http://dx.doi.org/10.1093/nar/27.1.55.

90. Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M. 1998. EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res* **26:**50–53 http://dx.doi.org/10.1093/nar/26.1.50.

91. Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M. 1997. EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res* **25:**43–51 http://dx.doi.org/10.1093/nar/25.1.43.

92. Karp PD, Riley M, Paley SM, Pelligrini-Toole A. 1996. EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res* **24:**32–39 http://dx.doi.org/10.1093/nar/24.1.32.

93. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, Latendresse M, Muñiz-Rascado L, Ong Q, Paley S, Peralta-Gil M, Subhraveti P, Velázquez-Ramírez DA, Weaver D, Collado-Vides J, Paulsen I, Karp PD. 2017. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res* **45**(D1)**:**D543–D550 http://dx.doi.org/10.1093/nar/gkw1003.

94. Keseler IM, Skrzypek M, Weerasinghe D, Chen AY, Fulcher C, Li GW, Lemmer KC, Mladinich KM, Chow ED, Sherlock G, Karp PD. 2014. Curation accuracy of model organism databases. *Database (Oxford)* **2014:**1–6 http://dx.doi.org/10.1093/database/bau058.