Published in final edited form as:

EcoSal Plus. 2014 May ; 6(1): . doi:10.1128/ecosalplus.ESP-0009-2013.

The EcoCyc Database

Peter D. Karp¹, Daniel Weaver¹, Suzanne Paley¹, Carol Fulcher¹, Aya Kubo¹, Anamika Kothari¹, Markus Krummenacker¹, Pallavi Subhraveti¹, Deepika Weerasinghe¹, Socorro Gama-Castro², Araceli M. Huerta², Luis Muñiz-Rascado², César Bonavides-Martinez², Verena Weiss², Martin Peralta-Gil², Alberto Santos-Zavaleta², Imke Schröder^{3,4}, Amanda Mackie⁵, Robert Gunsalus^{3,4}, Julio Collado-Vides², Ingrid M. Keseler¹, and Ian Paulsen⁵

¹Bioinformatics Research Group, SRI International, Menlo Park, USA

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

³Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, CA 90095, USA

⁴UCLA Institute of Genomics and Proteomics, University of California, Los Angeles, CA 90095, USA

⁵Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 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, 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 web site and downloadable software contain tools for analysis of high-throughput datasets. In addition, a steady-state metabolic flux model is generated from each new version of EcoCyc. The model can predict metabolic flux rates, nutrient uptake rates, and growth rates for different gene knockouts and nutrient conditions. This chapter provides a detailed description of the data content of EcoCyc, and of the procedures by which this content is generated.

1 EcoCyc Overview

EcoCyc¹ is a bioinformatics database that describes the genome and the biochemical machinery of *E. coli* K-12 MG1655. 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,

¹"EcoCyc" is pronounced "eeko-sike." It sounds like "ecology" and like "encyclopedia".

to facilitate a system-level understanding of *E. coli*. EcoCyc is an electronic reference source for *E. coli* biologists and for all researchers who work with *E. coli* and related microorganisms. In addition to the database, a steady-state metabolic flux model is available, generated from each new version of EcoCyc.

This chapter provides an overview of EcoCyc's data content and the procedures by which these data enter EcoCyc.

EcoCyc accelerates science. EcoCyc is designed for several different modes of interactive use via both the EcoCyc.org web site and in conjunction with the downloadable Pathway Tools [1] software (Section 13 lists the resources available to assist users in learning the web site and software)):

- 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.
- EcoCyc facilitates analysis of high-throughput data such as gene-expression and metabolomics data via tools for enrichment analysis, and for visualizing omics data on a metabolic map diagram, complete genome diagram, or regulatory network diagram.
- The EcoCyc metabolic flux model can predict growth or no-growth of wildtype and knock-out *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: [2, 3, 4, 5, 6].

Because the EcoCyc data are structured within a sophisticated ontology 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 use of EcoCyc to develop methods for studying path lengths within metabolic networks [7, 8, 9]; in studies relating protein structure to the metabolic network [10, 11]; and in analysis of the *E. coli* regulatory network [12, 13].

The development of many new bioinformatics methods requires high-quality, gold-standard datasets for the training and validation of those methods. EcoCyc has been used as a gold-standard dataset for the development of genome-context methods for predicting gene function [14, 15], operon-prediction methods [16, 17], prediction of promoters and transcription start sites [18, 19], regulatory network reconstruction [20], and the prediction of functional and direct protein-protein interactions [21, 22, 23]. The EcoCyc metabolic data have been used for studies concerning predicted metabolic networks and growth prediction [24, 25], and for model checking of a symbiotic bacteria's metabolic network [26].

Metabolic engineers alter microbes to produce biofuels, industrial chemicals, and pharmaceuticals; to de-grade toxic pollutants; and to sequester carbon [27, 28, 29]. 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 [30, 31, 32].

According to the Thomson Reuters Web of Knowledge citation index, as of August 2013, the 23 EcoCyc and RegulonDB papers authored since 1997 were cited by 2,395 publications from 1997–2013. According to Google Analytics, approximately 100,000 visitors query the EcoCyc website each year, generating 177,000 object page views per month on average in 2012.

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>).

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 organisms.

2 Overview of EcoCyc Data Content

EcoCyc covers a broad array of data types. Key to understanding the EcoCyc data and its 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., the linkage between a gene and gene product) 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.

No one-to-one mapping exists between EcoCyc classes and the data pages within the EcoCyc website, because one data page typically integrates information from multiple classes. For example, the pathway data page integrates information from objects in the classes Pathways, Reactions, Genes, Proteins, and Chemicals.

Genome

EcoCyc contains the complete genome sequence of *E. coli* and describes the nucleotide position and function of all known protein-coding and RNA-coding *E. coli* genes. Genome-related classes that are populated within EcoCyc include Genes, Pseudo-Genes, Promoters, DNA-Binding-Sites, and REP-Elements. Gene Ontology (GO) terms are assigned to genes both by EcoCyc curators and by import of GO terms from UniProt [33]. EcoCyc data on the essentiality of *E. coli* genes are described in Section 6.

Proteome

EcoCyc describes all known monomers and multimeric protein complexes of *E. coli*. EcoCyc contains extensive annotation of the features of *E. coli* proteins, such as

phosphorylation sites, metal-ion binding sites, and enzyme active sites, assigned by EcoCyc curators and imported from UniProt. Relevant classes within EcoCyc include Polypeptides and Protein-Complexes.

RNAome

EcoCyc describes all known RNAs and protein-RNA complexes of *E. coli*. Relevant classes within EcoCyc include RNAs, rRNAs, and Regulatory-RNAs. Note that EcoCyc does not explicitly represent messenger RNAs.

Regulation

EcoCyc contains the most complete description of the regulatory network of any organism. It covers *E. coli* operons, promoters, transcription factors, transcription-factor binding sites, attenuators, 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*. It describes each metabolic enzyme of *E. coli* including its cofactors, activators, inhibitors, and subunit structure.

Membrane Transporters

EcoCyc annotates *E. coli* transport proteins and the associated transport reactions that they mediate.

Growth Observations

EcoCyc integrates data on the growth of *E. coli* under many different growth conditions, as described in Section 5.

Database Links

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

3 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. EcoCyc is largely up to date with respect to its curation activities. As of October 2013, EcoCyc encodes information from more than 25,000 publications. A staff of four full-time curators updates the annotation of the *E. coli* genome on an ongoing basis.

The transcriptional regulatory information in EcoCyc and RegulonDB is curated by the group of Dr. Julio Collado-Vides at the UNAM; therefore, 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.

Curators collect gene, protein, pathway, and compound names and synonyms. They classify genes and gene products using the Gene Ontology [34] and MultiFun [35] ontologies, and they classify pathways within the Pathway Tools pathway ontology. Protein complex components and the stoichiometry of these subunits are captured; cellular localization of polypeptides and protein complexes is entered, as are experimentally determined protein molecular weights; enzyme activities and any enzyme prosthetic groups, cofactors, activators, or inhibitors are captured. Operon structure and gene regulation information are encoded.

Curators author textual summaries with extensive citations. Within the summaries for proteins, RNAs, pathways, and operons, curators capture 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.

4 Statistics on EcoCyc Content

Tables 1, 2, 3, and 4 present statistics on EcoCyc content. The listed numbers are current as of version 17.5, released in October 2013.

5 Conditions of *E. coli* Growth and Non-Growth

As of 2011, EcoCyc incorporates media that have been shown experimentally to support or not support growth of both wildtype and knock-out strains of *E. coli* K-12. This work has two goals. First is to assemble a comprehensive encyclopedia of *E. coli* growth conditions 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 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 20 media that are commonly 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), which measure cell respiration as a sensitive indicator of microbial growth [36]. The commercially available PM system for microorganisms provides a comprehensive set of phenotype tests 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 documents 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 *fur+* as well, according to Datsenko and Wanner, unpublished results.)
 This dataset 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 [37]
 Strain: *E. coli* K-12 MG1655 (American Type Culture Collection 700926)
 This dataset includes growth observations for the full complement of C, N, P, and S compounds under aerobic conditions. Bacteria were pre-grown on LB agar prior to the inoculation of Biolog plates and incubation at 37°C for 26 hours. The Omnilog instrument (a specialized incubator plus reader) was used for data collection and analysis.
- “The Evolution of Metabolic Networks of *E. coli*”
 Baumber *et al.* 2011 [38]
 Strain: *E. coli* K-12 MG1655
 This dataset consists of growth observations for 95 C compounds under aerobic and anaerobic conditions. Bacteria were pre-grown on Biolog Universal Growth Agar plus sheep blood (BUG-S), prior to inoculation of Biolog plates and incubation at 37°C. Growth was monitored by measuring optical density at 600nm with readings taken at 3, 6, 12, 24, and 48 hours (D. Baumber, personal communication).
- Mackie *et al.* 2013 [39]
 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 pre-grown on either LB

or R2A agar prior to inoculation of Biolog plates and incubation at 37°C for 48 hours. 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 [40]

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 prior to inoculation of Biolog plates and incubation at 37°C for 48 hours. 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.

6 Essential Gene Information

As of 2011, EcoCyc incorporates several large-scale datasets 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 to 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.* [41]

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

This study used a genetic footprinting technique with a Tn5 based transposome system and reported unambiguous assessment of approximately 87% of *E. coli* ORFs for essentiality. 626 genes were identified as essential for aerobic growth in rich media while 3126 genes were dispensable. Note that the inability to obtain an insertion mutant using this system may in some cases be a reflection of the non-

targeted 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.* [42] and corrections [43]

Strain: *E. coli* K-12 BW25113 (*rpoS*(Am) *rph-1* λ - *rrnB3* *lacZ*4787 *hsdR*514 (*araBAD*)567 (*rhaBAD*)568 *rph-1*)

This study created 3985 in-frame, single gene deletion mutants using the lambda RED recombinase system. 303 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 [43]. This study also profiled the growth of the mutants in minimal glucose MOPS 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.* [44]

Strain: *E. coli* K-12 BW25113 (*rpoS*(Am) *rph-1* λ - *rrnB3* *lacZ*4787 *hsdR*514 (*araBAD*)567 (*rhaBAD*)568 *rph-1*) (the same as in [42])

This study used the Keio collection of single gene knockout mutants and profiled them for growth on glycerol supplemented minimal medium. 119 genes were identified as essential for growth on glycerol. They also combined these observations with those made by Baba *et al* [42] 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.* [45]

This publication used the experimental data regarding conditional gene essentiality from Joyce *et al.* [44] and from Baba *et al.* [42] and compared it with the computationally predicted essential genes in their genome scale metabolic reconstruction of *E. coli*. This dataset is included in EcoCyc to facilitate benchmarking of computational predictions of essentiality from the EcoCyc model with computations from the model of Feist *et al.* Multicopy suppression underpins metabolic evolvability.

- “Multicopy suppression Underpins Metabolic Evolvability”

Patrick *et al.* 2007 [46]

Strain: *E. coli* BW25113 (*rpoS*(Am) *rph-1* λ - *rrnB3* *lacZ4787* *hsdR514* (*araBAD*)567 (*rhaBAD*)568 *rph-1*)

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

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

7 EcoCyc Metabolic Flux Model

A quantitative steady-state metabolic flux model has been derived from EcoCyc using flux balance analysis (FBA) [47, 48]. By running this model with different parameters, scientists can model the growth of *E. coli* under different nutrient conditions and for different gene knock-outs. Every time the model is executed, it is freshly generated from EcoCyc, meaning that as the reactions in EcoCyc are updated due to curation, the model automatically reflects those changes.

The EcoCyc FBA model is distinct from the *E. coli* FBA models derived by the Palsson group [49, 45, 50], but these models have much in common because EcoCyc and the iAF1260 model were partially unified in 2007 [45], and both groups consult the other's work when updating their models.

The Supplementary Information, provided as the accompanying paper-suppl-info.xlsx file, details the *E. coli* biomass metabolite set used to model biomass production metabolite requirements in EcoCyc FBA. This metabolite set is derived from the iJO1366 model WT biomass reaction of Orth *et al.* [50]. The Supplementary Information also contains a description of the nutrient and secretion metabolite sets that supply inputs and outputs to the FBA model, as well as a description of differences between the EcoCyc FBA biomass metabolite set and the iJO1366 WT biomass reaction.

To run the EcoCyc FBA model, download and install a Pathway Tools software configuration that includes EcoCyc, and invoke the MetaFlux modeling component of Pathway Tools (see Chapter 8 of the Pathway Tools User's Guide).

EcoCyc provides several example files describing invocations of the FBA model under different nutrient conditions. Those files are found within the installed Pathway Tools directory tree at pathway-tools/aic-export/pgdbs/biocyc/ecocyc/VERSION/data/fba/. Output files produced as a result of successful FBA runs on the supplied. fba files are also included. The supplied input files are:

1. GlucoseAer.fba: 10 mmol/gCDW/hr glucose uptake, minimal media, aerobic conditions
2. GlucoseAnaer.fba: 10 mmol/gCDW/hr glucose uptake, minimal media, anaerobic conditions

3. GlycerolAer.fba: 10 mmol/gCDW/hr glycerol uptake, minimal media, aerobic conditions
4. GlycerolAnaer.fba: 10 mmol/gCDW/hr glycerol uptake, minimal media, anaerobic conditions

7.1 External Flux Predictions

MetaFlux metabolic flux predictions from EcoCyc version 17.5 for aerobic growth on glucose and glycerol are given in Tables 5 and 6. Model predictions for anaerobic growth on glucose and glycerol are given in Tables 7 and 8. In all cases, the uptake rate of the carbon source is set to an upper bound reflecting experimental uptake rates in mmol/gCDW/hr. O₂ uptake rates are set to an upper bound of 0.00 mmol/gCDW/hr under anaerobic conditions. All other nutrient sources are left free.

7.2 Improvement of the Metabolic Model

With each EcoCyc release we plan to include an improved version of the EcoCyc metabolic flux model that reflects recent improvements to our knowledge of the *E. coli* metabolic network.

Model predictions can differ from experimental measurements due 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.

8 Update Frequency

The EcoCyc.org and BioCyc.org websites and downloadable files are updated 3–4 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.

9 Data Sources Incorporated into EcoCyc

EcoCyc includes data imported from the following bioinformatics databases. In most cases, the data are re-imported once or twice per year. We note that many literature references within EcoCyc were obtained from PubMed.

9.1 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. We import all protein features with experimental or non-experimental evidence qualifiers except for 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. We import citations associated with UniProt protein features if they include an associated PubMed ID.

The import of protein features into EcoCyc is done via the UniProt Feature Importer tool within the Pathway Tools software.

9.2 Gene Ontology

For several years, EcoCyc and EcoliWiki/PortEco have been collaborating on improving and maintaining the GO annotations for *E. coli*. GO and its applications are described in more detail in [56]. Since the summer of 2008, we have been periodically generating a file containing all *E. coli* K-12 GO term annotations, called `gene_association.ecocyc`, that may be obtained from the Gene Ontology Consortium.

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 one 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. During the import of GO terms from UniProt into EcoCyc, a filtering operation is applied to prune GO term annotations based solely on computational (IEA) evidence if the EcoCyc gene product already has more specific GO annotations (in other words, GO terms that are children of the GO term being imported) that have experimental evidence available. For example, if a gene product already contained an experimental annotation of the term "galactose kinase," the software would not add the computational annotation "carbohydrate kinase." This filtering leads to the removal of approximately 1,000 of these less specific and redundant annotations.

A gene association file is generated from the quarterly EcoCyc releases. This file is sent to the EcoliWiki team at Texas A&M for further processing. At EcoliWiki, annotations made in the wiki-based community annotation system since the last EcoCyc update are added to the file, along with annotations containing qualifiers (mainly `contributes_to`) not yet 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. EcoliWiki runs the GO consortium validation

scripts and deposits the file with the GO consortium via their Concurrent Versioning System.

9.3 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 [57] was incorporated into EcoCyc in version 10.0 (March 2006).

9.4 RefSeq Collaboration

EcoCyc is involved in a collaboration to update the genome annotation of the GenBank (U00096.2) and RefSeq (NC 000913.2) entries for *E. coli* K-12 MG1655 on an ongoing basis. The primary collaborators include EcoCyc, EcoGene, UniProtKB/Swiss-Prot, 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.

9.5 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.

10 EcoCyc Accession Numbers

10.1 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.

- **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 id 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, then 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 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 [57]. ECK numbers are stored in the EcoCyc slot Accession-2. For only the first 18 or so genes in the *E. coli* K-12 MG1655 genome are the b-number and ECK number the same number; for subsequent genes the numbers have diverged.

11 Other *E. coli* and *Shigella* PGDBs in BioCyc

EcoCyc is part of the larger BioCyc collection of Pathway/Genome Databases (PGDBs). BioCyc version 17.5 (2013) includes 160 *E. coli* and *Shigella* PGDBs. Most of these PGDBs were generated computationally and lack the extensive manual literature-based curation of the EcoCyc K-12 database. The *E. coli* genomes in BioCyc are focused on complete genomes, and do not include draft genomes.

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). This procedure 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. This curation is supported by the PortEco project.

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 webpages.

12 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 using the “Report Errors or Provide Feedback” link at the bottom of every data page. Please email suggestions or questions to biocyc-support at ai dot sri dot com.

During every EcoCyc release we email a summary of new developments to our biocyc-users mailing list. To subscribe to this mailing list, please see <http://biocyc.org/subscribe.shtml>.

13 How to Learn More

- How to use a Pathway Tools website such as EcoCyc
- Downloadable instructional videos on how to use EcoCyc
- Pathway/Genome Database Concepts Guide
- Publications on EcoCyc: [58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70]
- BioCyc User's Guide
- MetaCyc User's Guide
- Guide to the Pathway Tools Schema
- How to download Pathway Tools and organism flat-file databases

14 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.*, *Nuc Acids Res*, 41:D605–12 2013.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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 was a valued advisor 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. 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, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee T, et al. Pathway Tools version 13.0: Integrated Software for Pathway/Genome Informatics and Systems Biology. *Brief Bioinform.* 2010; 11:40–79. Available from: <http://bib.oxfordjournals.org/cgi/content/abstract/bbp043>. 10.1093/bib/bbp043 [PubMed: 19955237]
2. Kim KS, Lee S, Ryu CM. Interspecific bacterial sensing through airborne signals modulates locomotion and drug resistance. *Nat Commun.* 2013; 4:1809. [PubMed: 23651997]
3. Bower JM, Gordon-Raagas HB, Mulvey MA. Conditioning of uropathogenic *Escherichia coli* for enhanced colonization of host. *Infect Immun.* 2009 May; 77(5):2104–2112. [PubMed: 19255192]
4. Rhodius V, Dyk TKV, Gross C, LaRossa RA. Impact of genomic technologies on studies of bacterial gene expression. *Annu Rev Microbiol.* 2002; 56:599–624. [PubMed: 12142487]

5. Gonzalez R, Tao H, Purvis JE, York SW, Shanmugam KT, Ingram LO. 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.* 2003 Mar; 19(2):612–623. [PubMed: 12675606]
6. Taoka M, Yamauchi Y, Shinkawa T, Kaji H, Motohashi W, Nakayama H, et al. Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. *Mol Cell Proteomics.* 2004 Aug; 3(8):780–787. [PubMed: 15115803]
7. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. *Science.* 2002 Aug; 297(5586):1551–1555. [PubMed: 12202830]
8. Simeonidis E, Rison SC, Thornton JM, Bogle ID, Papageorgiou LG. Analysis of metabolic networks using a pathway distance metric through linear programming. *Metab Eng.* 2003 Jul; 5(3): 211–219. [PubMed: 12948755]
9. Arita M. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA.* 2004 Feb; 101(6):1543–1547. [PubMed: 14757824]
10. Jardine O, Gough J, Chothia C, Teichmann SA. Comparison of the small molecule metabolic enzymes of *Escherichia coli* and *Saccharomyces cerevisiae*. *Genome Res.* 2002 Jun; 12(6):916–929. [PubMed: 12045145]
11. Rison SC, Thornton JM. Pathway evolution, structurally speaking. *Curr Opin Struct Biol.* 2002 Jun; 12(3):374–382. [PubMed: 12127458]
12. Ma HW, Kumar B, Dittes U, Gunzer F, Buer J, Zeng AP. An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nuc Acids Res.* 2004; 32(22):6643–6649.
13. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet.* 2002 May; 31(1):64–68. [PubMed: 11967538]
14. Karimpour-Fard A, Leach SM, Gill RT, Hunter LE. Predicting protein linkages in bacteria: Which method is best depends on task. *BMC Bioinformatics.* 2008; 9:397. [PubMed: 18816389]
15. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. Prolinks: A database of protein functional linkages derived from coevolution. *Genome Biol.* 2004; 5(5):R35. [PubMed: 15128449]
16. Price MN, Huang KH, Alm EJ, Arkin AP. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nuc Acids Res.* 2005; 33(3):880–892.
17. Steinhauser D, Junker BH, Luedemann A, Selbig J, Kopka J. Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics.* 2004 Aug; 20(12):1928–1939. [PubMed: 15044239]
18. Burden S, Lin YX, Zhang R. Improving promoter prediction for the NNPP2.2 algorithm: A case study using *Escherichia coli* DNA sequences. *Bioinformatics.* 2005 Mar; 21(5):601–607. [PubMed: 15454410]
19. Gordon L, Chervonenkis AY, Gammelman AJ, Shahmuradov IA, Solovyev VV. Sequence alignment kernel for recognition of promoter regions. *Bioinformatics.* 2003 Oct; 19(15):1964–1971. [PubMed: 14555630]
20. Fu Y, Jarboe LR, Dickerson JA. Reconstructing genome-wide regulatory network of *E. coli* using transcriptome data and predicted transcription factor activities. *BMC Bioinformatics.* 2011; 12:233. [PubMed: 21668997]
21. Watanabe RL, Morett E, Vallejo EE. Inferring modules of functionally interacting proteins using the Bond Energy Algorithm. *BMC Bioinformatics.* 2008; 9:285. [PubMed: 18559112]
22. Muley VY, Ranjan A. Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. *PLoS One.* 2012; 7(7):e42057. [PubMed: 22844541]
23. Moreno-Hagelsieb G, Jokic P. The evolutionary dynamics of functional modules and the extraordinary plasticity of regulons: the *Escherichia coli* perspective. *Nucleic Acids Res.* 2012; 40(15):7104–12. [PubMed: 22618875]
24. Kastenmuller G, Schenk ME, Gasteiger J, Mewes HW. Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. *Genome Biol.* 2009; 10(3):R28. [PubMed: 19284550]

25. Kumar VS, Maranas CD. GrowMatch: An automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput Biol*. 2009 Mar.5(3):e1000308. [PubMed: 19282964]
26. Thomas GH, Zucker J, Macdonald SJ, Sorokin A, Goryanin I, Douglas AE. A fragile metabolic network adapted for cooperation in the symbiotic bacterium *Buchnera aphidicola*. *BMC Syst Biol*. 2009; 3:24. [PubMed: 19232131]
27. Frazier ME, Johnson GM, Thomassen DG, Oliver CE, Patrinos A. Realizing the potential of the genome revolution: The Genomes to Life program. *Science*. 2003 Apr; 300(5617):290–293. [PubMed: 12690188]
28. Bailey JE. Toward a Science of Metabolic Engineering. *Science*. 1991 Jun.252:1668–1675. [PubMed: 2047876]
29. Stephanopoulos G, Vallino JJ. Network rigidity and metabolic engineering in metabolite overproduction. *Science*. 1991 Jun.252:1675–1681. [PubMed: 1904627]
30. Arense P, Bernal V, Charlier D, Iborra JL, Foulquie-Moreno MR, Canovas M. Metabolic engineering for high yielding L(–)-carnitine production in *Escherichia coli*. *Microb Cell Fact*. 2013; 12(1):56. (ENG). [PubMed: 23718679]
31. Jantama K, Zhang X, Moore JC, Shanmugam KT, Svoronos SA, Ingram LO. Eliminating side products and increasing succinate yields in engineered strains of *Escherichia coli* C. *Biotechnol Bioeng*. 2008 Dec; 101(5):881–893. [PubMed: 18781696]
32. Weber J, Hoffmann F, Rinas U. Metabolic adaptation of *Escherichia coli* during temperature-induced recombinant protein production: 2. Redirection of metabolic fluxes. *Biotechnol Bioeng*. 2002 Nov; 80(3):320–330. [PubMed: 12226865]
33. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nuc Acids Res*. 2013; 41(Database issue):D43–7.
34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for the unification of biology. *Nature Genetics*. 2000; 25:25–29. [PubMed: 10802651]
35. Serres MH, Riley M. MultiFun, a multifunctional classification scheme for *Escherichia coli* K–12 gene products. *Genome Biol*. 2000; 5(4):205–222.
36. Bochner BR, Gadzinski P, Panomitros E. Phenotype Microarrays for High-Throughput Phenotypic Testing and Assay of Gene Function. *Genome Res*. 2001 Jul; 11(7):1246–1255. [PubMed: 11435407]
37. Oun MA, Suthers PF, Jones GI, Carter BR, Saunders MP, Maranas CD, et al. Genome scale reconstruction of a *Salmonella* metabolic model: comparison of similarity and differences with a commensal *Escherichia coli* strain. *J Biol Chem*. 2009; 284(43):29480–8. [PubMed: 19690172]
38. Baumler DJ, Peplinski RG, Reed JL, Glasner JD, Perna NT. The evolution of metabolic networks of *E. coli*. *BMC Syst Biol*. 2011; 5:182. [PubMed: 22044664]
39. Mackie A, Paley S, Keseler IM, Shearer A, Paulsen IT, Karp PD. Addition of *Escherichia coli* K–12 Growth-Observation and Gene Essentiality Data to the EcoCyc database. *J Bacteriol*. 2013
40. Yoon SH, Han MJ, Jeong H, Lee CH, Xia XX, Lee DH, et al. Comparative multi-omics systems analysis of *Escherichia coli* strains B and K–12. *Genome Biol*. 2012; 13(5):R37. [PubMed: 22632713]
41. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, et al. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol*. 2003 Oct; 185(19):5673–5684. [PubMed: 13129938]
42. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of *Escherichia coli* K–12 in-frame, single-gene knockout mutants: The Keio collection. *Mol Syst Biol*. 2006; 2:2006.0008.
43. Yamamoto N, Nakahigashi K, Nakamichi T, Yoshino M, Takai Y, Touda Y, et al. Update on the collection of *Escherichia coli* single-gene deletion mutants. *Mol Syst Biol*. 2009; 5:335. [PubMed: 20029369]
44. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, et al. Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *J Bacteriol*. 2006 Dec; 188(23):8259–8271. [PubMed: 17012394]
45. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for *Escherichia coli* K–12 MG1655 that accounts for 1260 ORFs and

- thermodynamic information. *Mol Syst Biol.* 2007; 3:121–38. Available from: <http://www.nature.com/doifinder/10.1038/msb4100155>. [PubMed: 17593909]
46. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy suppression underpins metabolic evolvability. *Mol Biol Evol.* 2007; 24(12):2716–22. [PubMed: 17884825]
 47. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? *Nat Biotechnol.* 2010; 28(3):245–8. [PubMed: 20212490]
 48. Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc.* 2010; 5(1):93–121. [PubMed: 20057383]
 49. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* 2003; 4(9):R54. [PubMed: 12952533]
 50. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism — 2011. *Mol Syst Biol.* 2011; 7:535. [PubMed: 21988831]
 51. Kayser A, Weber J, Hecht V, Rinas U. Metabolic flux analysis of *Escherichia coli* in glucose-limited continuous culture. I. Growth-rate-dependent metabolic efficiency at steady state. *Microbiology.* 2005; 151:693–706. [PubMed: 15758216]
 52. Ibarra RU, Edwards JS, Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature.* 2002; 420:186–189. [PubMed: 12432395]
 53. Belaich A, Belaich JP. Microcalorimetric study of the anaerobic growth of *Escherichia coli*: growth thermograms in a synthetic medium. *J Bacteriol.* 1976; 72:497–9.
 54. Varma A, Boesch BW, Palsson BO. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Applied and Environmental Microbiology.* 1993; 59(8):2465–73. [PubMed: 8368835]
 55. Dharmadi Y, Murarka A, Gonzalez R. Anaerobic fermentation of glycerol by *Escherichia coli*: anew platform for metabolic engineering. *Biotechnology and Bioengineering.* 2006; 94(5):821–9. [PubMed: 16715533]
 56. Hu JC, Karp PD, Keseler IM, Krummenacker M, Siegele DA. What we can learn about *Escherichia coli* through application of Gene Ontology. *Trends Microbiol.* 2009 Jul; 17(7):269–278. [PubMed: 19576778]
 57. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, et al. *Escherichia coli* K-12: A cooperatively developed annotation snapshot—2005. *Nuc Acids Res.* 2006; 34(1):1–9.
 58. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muniz-Rascado L, et al. EcoCyc: A Comprehensive Database of *Escherichia coli* Biology. *Nuc Acids Res.* 2011; 39:D583–90.
 59. Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, et al. EcoCyc: A comprehensive view of *E. coli* biology. *Nuc Acids Res.* 2009; 37:D464–70. Available from: <http://nar.oxfordjournals.org/cgi/reprint/gkn751?ijkey=7epgizfnGFYQHCE&keytype=ref>.
 60. Karp PD, Keseler IM, Shearer A, Latendresse M, Krummenacker M, Paley SM, et al. Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nuc Acids Res.* 2007; 35:7577–90. Available from: <http://nar.oxfordjournals.org/cgi/content/full/35/22/7577>.
 61. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, et al. EcoCyc: A comprehensive database resource for *E. coli*. *Nuc Acids Res.* 2005; 33:D334–7. Available from: http://nar.oupjournals.org/cgi/content/full/33/suppl_1/D334?ijkey=80p4BbGpEFjLQ&keytype=ref.
 62. Karp PD, Arnaud M, Collado-Vides J, Ingraham J, Paulsen IT, Saier MHJ. The *E. coli* EcoCyc Database: No Longer Just a Metabolic Pathway Database. *ASM News.* 2004; 70(1):25–30.
 63. Karp PD, Riley M, Saier M, Paulsen IT, Paley S, Pellegrini-Toole A. The EcoCyc Database. *Nuc Acids Res.* 2002; 30(1):56–8.
 64. Karp PD, Riley M, Saier M, Paulsen IT, Paley S, Pellegrini-Toole A. The EcoCyc and MetaCyc Databases. *Nuc Acids Res.* 2000; 28(1):56–59.
 65. Karp, PD. Nucleic Acid and Protein Databases and How To Use Them. London: Academic Press; 1999. Using the EcoCyc Database; p. 269–280.

66. Karp, PD.; Riley, M. Bioinformatics Databases and Systems. Norwell, MA: Kluwer Academic Publishers; 1999. EcoCyc: The resource and the lessons learned; p. 47-62.
67. Karp P, Riley M, Paley S, Pellegrini-Toole A, Krummenacker M. EcoCyc: Electronic Encyclopedia of *E. coli* Genes and Metabolism. Nuc Acids Res. 1999; 27(1):55–58.
68. Karp P, Riley M, Paley S, Pellegrini-Toole A, Krummenacker M. EcoCyc: Electronic Encyclopedia of *E. coli* Genes and Metabolism. Nuc Acids Res. 1998; 26(1):50–53.
69. Karp P, Riley M, Paley S, Pellegrini-Toole A, Krummenacker M. EcoCyc: Electronic Encyclopedia of *E. coli* Genes and Metabolism. Nuc Acids Res. 1997; 25(1):43–50.
70. Karp P, Riley M, Paley S, Pellegrini-Toole A. EcoCyc: Electronic Encyclopedia of *E. coli* Genes and Metabolism. Nuc Acids Res. 1996; 24(1):32–40.

Table 1

Genes and gene products in EcoCyc. 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 IS elements are included in the count of Genes but are not included in the sub-categories of genes.

Genes/Proteins	
Genes	4501
Protein-Coding Genes	4284
tRNA Genes	86
rRNA Genes	22
Regulatory RNA Genes	41
Other RNA Genes	56
Pseudogenes	133
Polypeptides	4393
Protein Complexes	995
Heteromultimeric Protein Complexes	290
Homomultimeric Protein Complexes	705
Protein Features	23,114
Enzymes (excluding transporters)	1245
Transporters	267

Table 2

Gene annotation status in EcoCyc. Genes of known molecular function have experimental evidence for their assigned function, whereas genes of predicted molecular function have had their function predicted computationally.

Gene Annotation Status	
Genes of known or predicted molecular function	3127
Genes of known molecular function	2710
Genes of predicted molecular function	417
Genes of unknown molecular function	1374

Table 3

Reactions, compounds and pathways in EcoCyc. Superpathways are connected sets of base metabolic pathways (connected via shared substrates.)

Reactions/Metabolites/Pathways	
Metabolic Reactions	1443
Transport (including Electron Transfer) Reactions	379
Pathways	401
Small-molecule metabolism base pathways	291
Signaling pathways	29
Superpathways	81
Metabolites	2466
Metabolites that are substrates of enzyme-catalyzed reactions	1331
Metabolites that are physiological enzyme regulators	121
Metabolites that are cofactors or prosthetic groups	56
Transported metabolites	274

Table 4

Regulation-related objects and interactions in EcoCyc. Each member of “Instances of Regulation of Transcription Initiation” describes a single regulatory interaction between a transcription factor and its binding site.

Transcriptional/Translational Regulation	
Transcription Units	4510
Promoters	3777
Terminators	259
Transcription Factors	194
Transcription Factor Binding Sites	2773
Instances of Regulation of Transcription Initiation	3293
Instances of Regulation by Transcriptional Attenuation	20
Instances of Regulation of Translation	146

Table 5

Comparison of experimental aerobic glucose-limited chemostat growth data with EcoCyc and iJO1366 FBA model predictions (389 reactions active in EcoCyc). Metabolite uptake and production rates in units of mmol/gCDW/hr. Growth in units of hr^{-1} . Experimental data from Kayser *et al.* [51].

	Experimental	EcoCyc
Glucose uptake	3.008	3.008
Growth rate	0.300	0.276
O ₂ uptake	7.413	4.472
NH ₄ uptake	2.367	3.026
Sulfate uptake		0.068
Phosphate uptake		0.288
CO ₂ production	7.38	5.480
H ₂ O production		13.026
H ⁺ production		2.582

Table 6

EcoCyc FBA model performance for aerobic glycerol-limited growth (385 reactions active in Eco-Cyc).

Uptake and production rates in units of mmol/gCDW/hr. Growth in units of hr⁻¹. Experimental OUR/G1-UR estimated graphically from Ibarra *et al.* [52].

	Experimental	EcoCyc
Glycerol uptake	10	10.00
Growth rate		0.53
O ₂ uptake	11	8.96
NH ₄ uptake		5.80
Sulfate uptake		0.13
Phosphate uptake		0.55
CO ₂ production		5.74
H ₂ O production		30.36
H ⁺ production		4.95

Table 7

EcoCyc FBA model performance for anaerobic glucose-limited growth (383 reactions active in EcoCyc). Uptake and production rates in units of mmol/gCDW/hr. Growth in units of hr^{-1} . Experimental data from Belaich and Belaich [53] via Varma *et al.* [54]. FHL set inactive for purposes of comparison.

	Experimental	EcoCyc
Glucose uptake	10.0	10.00
Growth rate	0.30	0.25
O ₂ uptake	0.00	0.00
NH ₄ uptake		2.76
Sulfate uptake		0.06
Phosphate uptake		0.26
CO ₂ production		0.27
H ₂ O production		0.57
H ⁺ production		27.14
Acetate production	7.5	8.03
Formate production	11.3	16.76
Succinate production	1.2	0.00
Ethanol production	8.7	7.73

Table 8

EcoCyc FBA model performance for anaerobic glycerol-limited growth (374 reactions active). Uptake and production rates in units of mmol/gCDW/hr. Growth in units of hr⁻¹. Quantitative experimental rates not currently available; for a qualitative description of anaerobic glycerol fermentation, see Dharmadi *et al.* [55].

	EcoCyc
Glycerol uptake	10.00
Growth rate	0.08
O ₂ uptake	0.00
NH ₄ uptake	0.88
Sulfate uptake	0.02
Phosphate uptake	0.84
CO ₂ production	0.17
H ₂ O production	3.13
H ⁺ production	9.47
Acetate production	0.00
Formate production	8.72
Succinate production	0.00
Ethanol production	8.90