Published online 28 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D613–D624***
doi:10.1093/nar/gks1235

EcoGene 3.0

Jindan Zhou and Kenneth E. Rudd*

Department of Biochemistry and Molecular Biology, The Miller School of Medicine at the University of Miami, Miami, FL 33143, USA

Received October 5, 2012; Revised October 31, 2012; Accepted November 1, 2012

ABSTRACT

EcoGene (http://ecogene.org) is a database and website devoted to continuously improving the structural and functional annotation of *Escherichia coli* K-12, one of the most well understood model organisms, represented by the MG1655(Seq) genome sequence and annotations. Major improvements to EcoGene in the past decade include (i) graphic presentations of genome map features; (ii) ability to design Boolean queries and Venn diagrams from EcoArray, EcoTopics or user-provided GeneSets; (iii) the genome-wide clone and deletion primer design tool, PrimerPairs; (iv) sequence searches using a customized EcoBLAST; (v) a Cross Reference table of synonymous gene and protein identifiers; (vi) proteome-wide indexing with GO terms; (vii) EcoTools access to >2000 complete bacterial genomes in EcoGene-RefSeq; (viii) establishment of a MySql relational database; and (ix) use of web content management systems. The biomedical literature is surveyed daily to provide citation and gene function updates. As of September 2012, the review of 37 397 abstracts and articles led to creation of 98 425 PubMed-Gene links and 5415 PubMed-Topic links. Annotation updates to Genbank U00096 are transmitted from EcoGene to NCBI. Experimental verifications include confirmation of a CTG start codon, pseudogene restoration and quality assurance of the Keio strain collection.

INTRODUCTION

The EcoGene database evolved from pre-genome era studies anticipating the genome era (1). EcoSeq was a growing collection of *Escherichia coli* K-12 genomic sequence contigs that were shotgun-assembled from the numerous gene-length DNA sequences in Genbank. EcoSeq was aligned, using the novel restriction map alignment software MapSearch (2), to EcoMap, a digitized version of the Kohara map, a complete genome restriction map (3). The gel-derived Kohara restriction map was replaced by DNA sequence restriction maps derived from sequenced regions, creating a hybrid map. Sequenced genes were aligned and oriented: EcoMap became the model for transitional genetic maps to localize and orient genes on the chromosome based on DNA sequences, restriction maps and genetic crosses (4). EcoSeq was one of the first large sets of DNA sequence data for which the genome orientation was known (>1.6 megabases) and was used to localize both known and novel DNA repeats, including the extreme DNA strand bias (skew) of the recombination hotspot Chi (5).

EcoGene was created by the development and application of new methods for accurately re-annotating the gene content of EcoSeq. Most re-annotations were done one protein coding sequence at a time, including careful consideration of DNA and protein sequence alignments, signal peptide predictions, and ribosome binding site predictions. The gene finding software GeneMark was used in a collaborative endeavor that assisted GeneMark's development (6,7). When the *E. coli* K-12 genome sequence was finished, these same methods were used to re-annotate the complete genome sequence for EcoGene.

The previous database publication described our basic annotation approaches including translation start site revisions, a compilation of the Verified Set of sequenced protein starts, as well as the identification and hypothetical reconstruction of deletion, frame-shifted and interrupted pseudogenes (8). The process of daily updating from the literature was described and currently 37 397 PubMed abstracts and full articles have been reviewed, creating 98 425 PubMed-to-Gene links. The EcoGene database and website previously described were rudimentary (see the *EcoGene MySql* section below). The map depictions at the original EcoGene.org website were static PDF maps produced by PrintMap (1).

The previous database article noted that the many updates in EcoGene differed substantially from the annotation in the *E. coli* K-12 MG1655(Seq) primary Genbank record U00096 (8,9). Genbank is the source database for gene names, product names and sequences used in many other databases, as well as source data for many bioinformatics studies. An annotation workshop held in

*To whom correspondence should be addressed. Tel: +1 350 243 6055; Fax: +1 350 243 3955; Email: krudd@med.miami.edu

2005 included participants from several *E. coli* databases, and was a conduit for the EcoGene revisions, particularly the extensive start site revisions, to migrate into Genbank U00096 (10). This was a follow-up to a previous annotation workshop in 2003 that established the multi-database annotation collaboration. The workshop report included a description of many of the annotation methods used at EcoGene during the previous decade. Although the report states that the start site revisions were performed at the workshop, all revisions had been reviewed and corrected during the previous decade at EcoGene. One of us (K.E.R.) demonstrated and explained the unpublished methodology used to correct start sites and frame-shifted pseudogenes, using over 200 examples shown to other curators at the workshop. Later, in April 2007, EcoGene.org became the source database for the Genbank U00096 and companion RefSeq record NC_000913 annotations. Updated MySql tables are transferred to NCBI from EcoGene.org to facilitate periodic annotation updates. A file comparing the update to the previous annotations is produced at NCBI and reviewed for comments and corrections by senior curators at UniProtKB (11), EcoCyc (12), NCBI (13), ASAP (14) and EcoGene. Updates are performed at EcoGene using a suite of in-house curation tools, some of which were developed with funds from a 1-year EcoliHub NIH sub-award in 2008–09.

Although Genbank U00096 is titled as the genome sequence for *E. coli* K-12 MG1655, a semi-motile strain represented by Coli Genetic Stock Center strain CGSC#6300, the genome sequence actually corresponds to MG1655 (Seq), a spontaneous hyper-motile variant (CGSC#7740) (15). The major difference between the two strains is the insertion of IS1H at *flhDC*, which causes hyper-motility. U00096.2 is a sequence-corrected version of U00096.1 (16). With the advent of next-generation sequencing, two groups have reported minor sequence errors in U00096.2 (17,18), and we have independently confirmed these few corrections (A. Richardson and K.E.R., unpublished). The corrections will appear in U00096.3.

Description

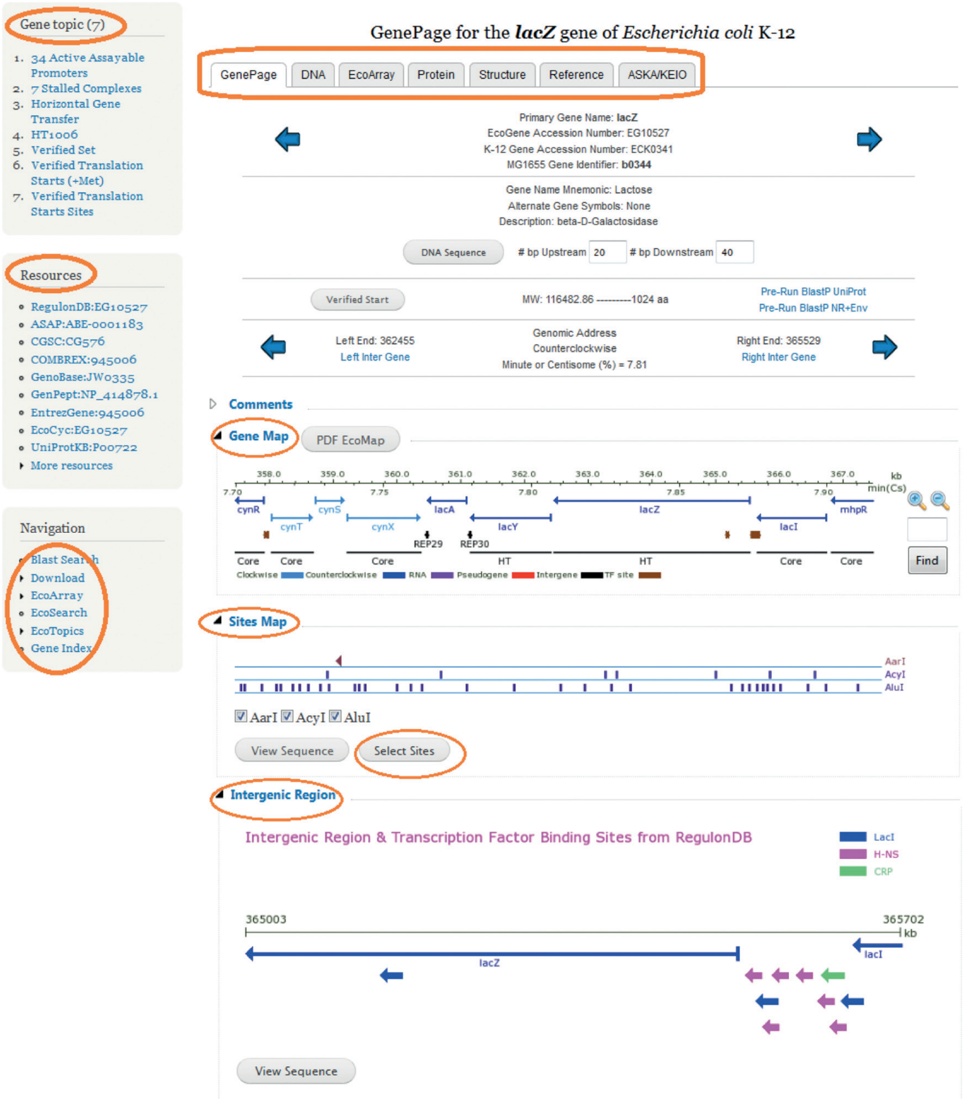
GenePages with graphic map depictions

EcoGene (http://ecogene.org) displays dynamically generated maps of genes, proteins and other features on GenePages. The tab-oriented approach taken for the EcoGene 3.0 GenePage design is depicted in Figure 1. Every protein and RNA-encoding gene has a GenePage with additional tabs displaying the DNA sequence, protein sequence, microarray results, PubMed-linked bibliographies, and PDB-linked structures. The GenePage tab also includes links to TopicPages related to the specific gene of interest. A resources menu to gene-oriented pages at external database resources and a navigation menu for easy access to other EcoGene functions are provided. Many GenePages also have comments noting recent discoveries, conflicts in the literature or curator interpretations. Comprehensive mini-reviews are not presented; however, links to EcoCyc provide the user with immediate access to their many well-written gene and protein function summaries. The GenePage bibliographies provide access to a comprehensive collection of daily-annotated PubMed abstracts, research articles and review articles that also contain summary material. Navigation arrows move to adjacent genes or move to genes in alphabetical order. Additional GenePage features display the length of the adjacent intergenic regions, link to pre-run BLAST output files, provide details on *N*-terminal protein sequencing (the Verified Set), display any number of upstream and downstream basepairs, explain the gene name mnemonic and provide alternate gene symbols (synonyms). A primary gene name is assigned considering historical precedence, accuracy, uniqueness and current usage. In addition to the EcoGene EG accession number for the gene, which refers to both the K-12 gene and alleles of the same gene in other *E. coli* strains, the ECK accession number, also maintained at EcoGene, refers specifically to the *E. coli* K-12 MG1655 allele.

Three regional maps, in Portable Network Graphic (PNG) format that allows for the images to be saved and used in publications and presentations, are found at the bottom of each GenePage. The first map is the Gene Map, containing gene depictions along with kb and centisome (%) scales. The Gene Map depicts a default 10-kb region of genomic DNA and can be zoomed in by users to define shorter regions in more detail. The Gene Map also illustrates the position of transcription factor binding sites (TFBSs) and intergenic repeats. Graphic tracks, found at the bottom of the Gene Map, show whether the gene is a Core gene, present in most *E. coli* strains, or a non-Core gene labeled as 'HT' to indicate they are predicted to be inherited through horizontal transmission, although gene loss could also account for the non-Core status in some cases (19). Objects within the Gene Map contain hyperlinks to other GenePages and TopicPages. Another link found within the Gene Map is PDF EcoMap, which allows the user to download a regional PDF map created using the PrintMap program (1). Full genome PDF maps in PrintMap format are available on the EcoTools page. The Site Map, the second of the three maps, depicts all the restriction sites for three default restriction enzymes (BamHI, EcoRI and HindIII), and can be customized by the user to show up to seven restriction enzyme sites. The list of restriction enzymes and their DNA recognition sequences were obtained from REBASE (20). Clicking on the restriction enzyme name next to the restriction maps links to additional information for that enzyme at http://rebase.neb.com. An additional feature of the Site Map is that it magnifies in concert with the Gene Map. The Intergenic Region maps at the bottom of the GenePages offer a more detailed depiction of TFBSs and intergenes. The transcription factors binding to the region are listed with hyperlinks to their GenePages, and intergenes are listed with hyperlinks to their TopicPages. TFBS arrows are linked to the RegulonDB source database (21).

GeneSets: EcoSearch, EcoTopics, EcoArray, Boolean queries and interactive Venn diagrams

A graphic presentation of Boolean query comparisons using two or three genesets can be displayed in an



GenePage for the ***lacZ*** gene of *Escherichia coli* K-12

Gene topic (7)

1. 34 Active Assayable Promoters
2. 7 Stalled Complexes
3. Horizontal Gene Transfer
4. HT1006
5. Verified Set
6. Verified Translation Starts (+Met)
7. Verified Translation Starts Sites

Resources

- RegulonDB:EG10527
- ASAP:ABE-0001183
- CGSC:CG576
- COMBREX:945006
- GenoBase:JW0335
- GenPept:NP_414878.1
- EntrezGene:945006
- EcoCyc:EG10527
- UniProtKB:P00722
- More resources

Navigation

- Blast Search
- Download
- EcoArray
- EcoSearch
- EcoTopics
- Gene Index

Primary Gene Name: **lacZ**
EcoGene Accession Number: EG10527
K-12 Gene Accession Number: ECK0341
MG1655 Gene Identifier: **b0344**

Gene Name Mnemonic: Lactose
Alternate Gene Symbols: None
Description: beta-D-Galactosidase

# bp Upstream 20 # bp Downstream 40

MW: 116482.86 ----------1024 aa

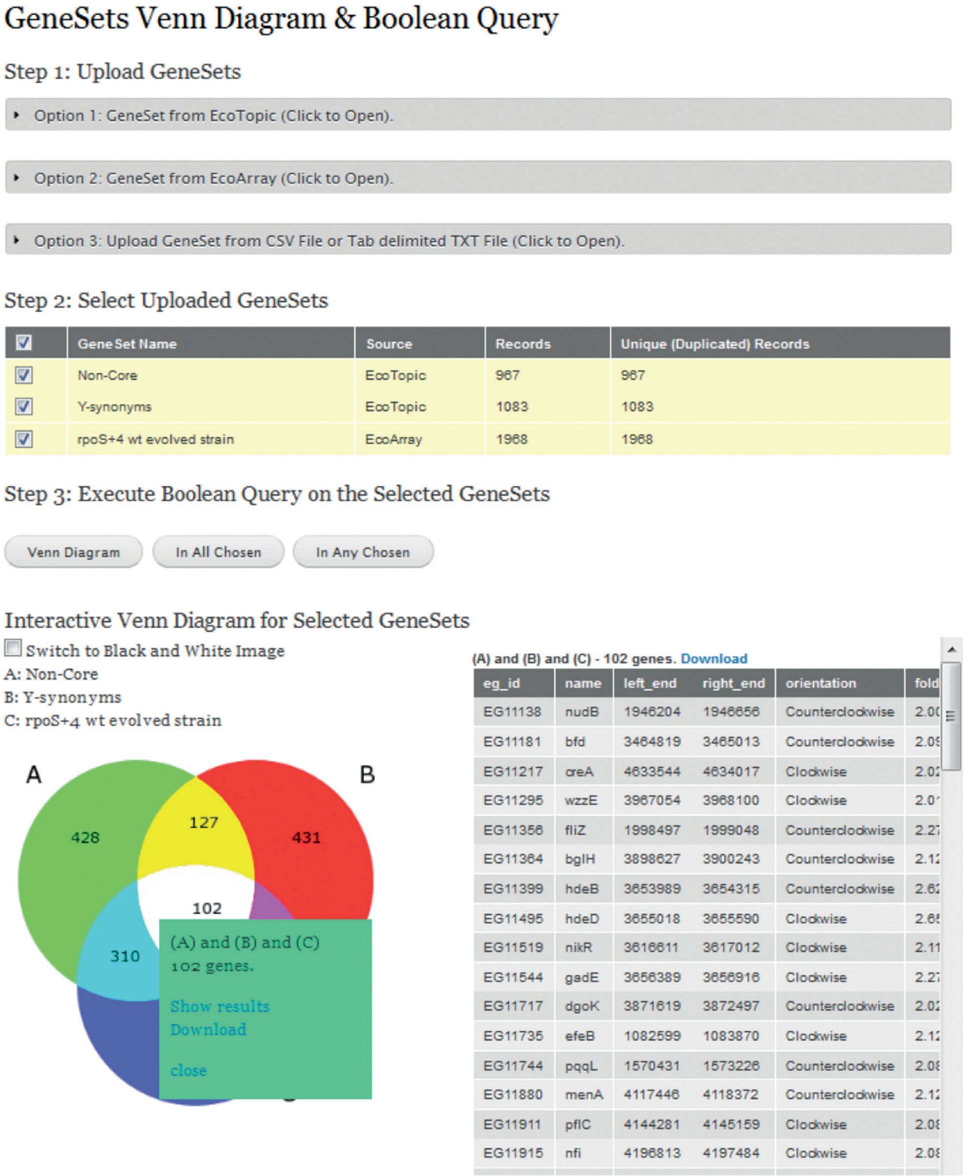
Left End: 362455
Genomic Address
Counterclockwise
Minute or Centisome (%) = 7.81
Right End: 365529

Intergenic Region & Transcription Factor Binding Sites from RegulonDB

Figure 1. The main GenePage tab of the EcoGene 3.0 GenePage for *lacZ*. The main areas of interest are circled and described in the text.

interactive Venn diagram (Figure 2). Boolean queries can be executed using EcoGene genesets or user-specified genesets. EcoGene genesets are collections of genes clustered in EcoTopics or EcoArray. User-provided genesets are lists of EG accession numbers uploaded at a GeneSets Venn Diagram and Boolean Query page. EcoSearch can be used to obtain a list of user-specified EG numbers from a query or from an uploaded ID file containing gene names or any other unique gene identifiers; gene names that are synonyms are mapped to primary genes. The Cross Reference Mapping tool can also be used for this purpose (see *EcoTools* section). EcoSeach also allows searching of the extensive EcoRef bibliographies, including both GenePage and EcoTopic bibliographies. Both the EcoSearch results pages and the TopicPages display genesets on a small circular map linked to an interactive circular map design tool Circle Maps. For an example, see http://www.ecogene.org/genemap/map.php?search_topic = 560. Circle Maps can also be accessed directly from the EcoTools menu.

EcoTopics cover a wide range of topics and are used to cluster genes for any reason; however, typical EcoTopics have a hierarchal nature with many topics having multiple sub-topics and super-topics. In all, 559 EcoTopics exist, and are linked to 21 154 genes and 5415 PubMed abstracts. EcoTopics can vary in the number of associated genes. An EcoTopic containing information from shotgun proteomics can contain hundreds of genes, whereas other EcoTopics link to more specific topics, such as the 17 IS elements, or intergenes like the 5 REP elements and the 10 CRISPR repeats families. Some EcoTopics have no associated genesets and were created to have a place to attach bibliographies that might not associate with a gene, such as the metabolic modeling literature. However, others only serve as placeholders to aggregate PubMed citations; many of these do not currently contain a text description. EcoTopics are not intended to contain comprehensive curator-authored reviews. Reading review articles, which are culled into separate bibliographies as well as being in the main bibliographies, and reading



GeneSets Venn Diagram & Boolean Query

Step 1: Upload GeneSets

▸ Option 1: GeneSet from EcoTopic (Click to Open).

▸ Option 2: GeneSet from EcoArray (Click to Open).

▸ Option 3: Upload GeneSet from CSV File or Tab delimited TXT File (Click to Open).

Step 2: Select Uploaded GeneSets

☑	Gene Set Name	Source	Records	Unique (Duplicated) Records
☑	Non-Core	EcoTopic	967	967
☑	Y-synonyms	EcoTopic	1083	1083
☑	rpoS+4 wt evolved strain	EcoArray	1968	1968

Step 3: Execute Boolean Query on the Selected GeneSets

Interactive Venn Diagram for Selected GeneSets

□ Switch to Black and White Image

A: Non-Core

B: Y-synonyms

C: rpoS+4 wt evolved strain

(A) and (B) and (C) - 102 genes. Download

eg_id	name	left_end	right_end	orientation	fold
EG11138	nudB	1946204	1946656	Counterclockwise	2.00
EG11181	bfd	3464819	3465013	Counterclockwise	2.09
EG11217	creA	4633544	4634017	Clockwise	2.02
EG11295	wzzE	3967054	3968100	Clockwise	2.01
EG11356	fliZ	1998497	1999048	Counterclockwise	2.27
EG11364	bglH	3898627	3900243	Counterclockwise	2.12
EG11399	hdeB	3653989	3654315	Counterclockwise	2.62
EG11495	hdeD	3655018	3655590	Clockwise	2.65
EG11519	nikR	3616611	3617012	Clockwise	2.11
EG11544	gadE	3656389	3656916	Clockwise	2.27
EG11717	dgoK	3871619	3872497	Counterclockwise	2.02
EG11735	efeB	1082599	1083870	Clockwise	2.12
EG11744	pqqL	1570431	1573226	Counterclockwise	2.08
EG11880	menA	4117446	4118372	Counterclockwise	2.12
EG11911	pflC	4144281	4145159	Clockwise	2.08
EG11915	nfi	4196813	4197484	Clockwise	2.08

Figure 2. Interactive Venn diagrams and Boolean queries using genesets. Clicking on the number in any of the Venn diagram sectors produces a list of genes. The Venn diagrams can be saved as a PNG image for use in presentations or publications; a black-and-white version is available to save on printer and publishing costs.

introductions to articles is a good way to get expert reviews of many EcoTopics. EcoArray was originally designed in 2007 to capture *E. coli* microarray results that were not deposited in pubic databases and to allow access to expression data for a particular gene from the GenePages. Although many public websites and databases populated with *E. coli* transcription array data now exist, some datasets in EcoArray are still missing from public repositories and other databases. We are currently importing and comparing data from numerous sources as we revamp EcoArray, as will be reported separately.

PrimerPairs

PrimerPairs is an embedded design tool for obtaining genome-wide polymerase chain reaction (PCR) primer sequences, which can be used to generate desired deletion–insertions or to create an ordered clone library, using the current EcoGene annotations of gene intervals. The user has the option to set primer lengths, offset positions inside or outside of genes, add-on sequences for adding restriction sites or sequences to amplify kanamycin (kan) or chloramphenicol (cat) antibiotic resistance cassettes. PrimerPairs has the ability to detect and correct primer sequences that delete part of adjacent overlapping genes. The offending primers are repositioned automatically to avoid the double-deletion problem found in the Keio mutant collection (see below). PrimerPairs has previously been described in detail (22).

EcoBlast

The NCBI BLAST program (13) runs locally on the EcoGene server and is used to prepare pre-run BLAST search results for all protein sequences on a weekly basis. Links to these pre-run results are provided on each GenePage. A local BLAST interface, called EcoBlast, provides users with the ability to query input sequences against a set of in-house databases. These specialized databases include EcoProt and EcoGene [sequence libraries derived from K-12 MG1655(Seq) gene-only sequences], the current (M56) and original (M54) versions of Genbank U00096, the *Salmonella typhimurium* LT2 genome and proteome, sequence libraries of all complete *E. coli* or *Salmonella* genomes and proteomes and in-house compiled libraries of complete genomes and proteomes from Enterobacteriales species (excluding *E. coli*). Automated downloads of the comprehensive NCBI NT and NR non-redundant nucleotide and protein sequence libraries are performed weekly. The UniProtKB UniRef100 and Swiss-Prot/TrEMBL proteomes are also included in the EcoBlast Database menu. Finally, a comprehensive NR+EnvNR protein sequence database combining genomic and environmental sequences from NCBI is compiled in-house and gives comprehensive EcoBlastP results populated with intermediate homologs from EnvNR.

EcoTools: downloads and cross reference tool

In addition to the Boolean Query tool and PrimerPairs tool described above, the EcoTools page of EcoGene 3.0 has links to other data download pages, which are also listed in the EcoTools sidebar sub-menu. The EcoGene Database Table Download page allows the user to select fields from the current daily-updated database for export to a tab-delimited text file. This table includes information present on the GenePage and Protein tabs, including gene name, protein name, gene synonyms, molecular weight and map position. The table also includes several important accession numbers, such as EcoGene (EG) number, *E. coli* K-12 ECK number, UniProtKB/Swiss-Prot ID, GenBank GI. Recently, the Cross Reference Mapping and Download page was created for user access to many additional accession numbers and other gene identifiers such as gene name and synonyms. The *E. coli* K-12 gene accession numbers from >30 other databases were collected to enable the construction of hyperlinks to gene pages at other websites and to easily update tables that lack EG or ECK ids to the most recent gene names and functions. A Genome Sequence Download page enables whole genome DNA sequence or user-specified subsequence extraction. Some basic EcoGene navigation and retrieval functions, including gene-start or gene-stop anchored sub-sequence extractions, have been presented as detailed protocols (22).

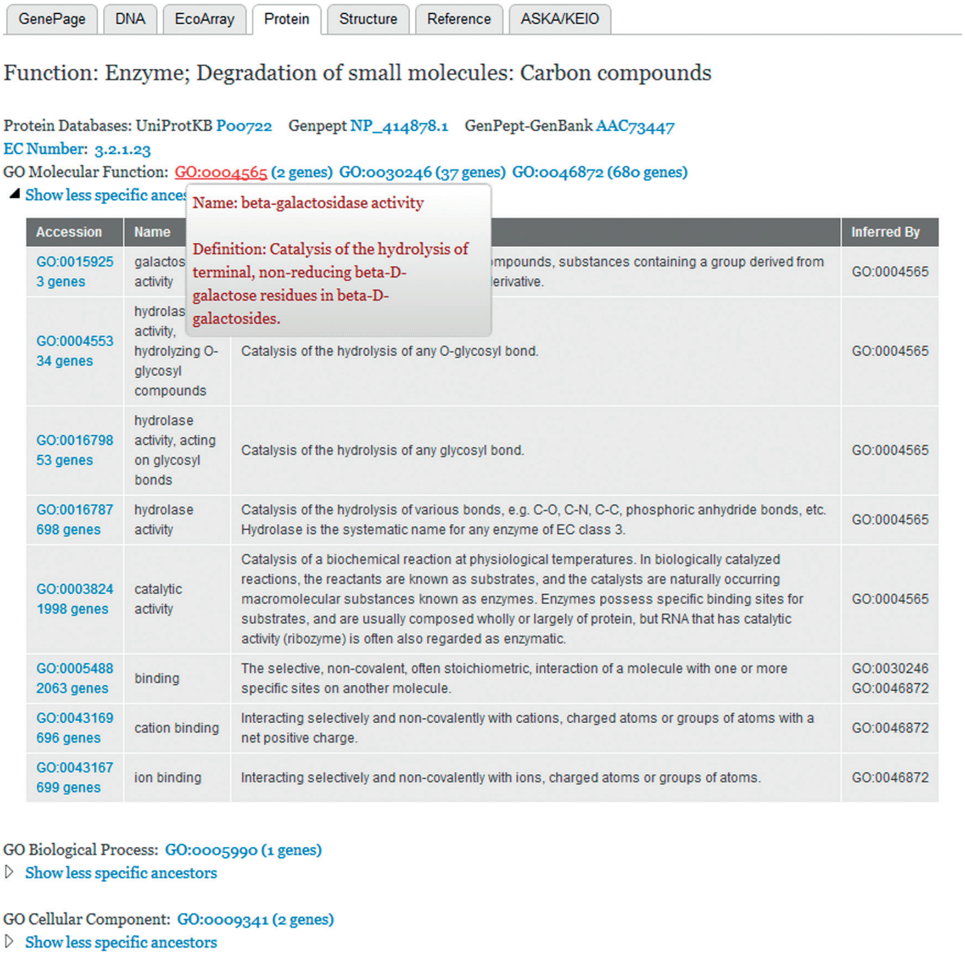
An Intergenic Region Download page enables the retrieval of the *E. coli* K-12 DNA sequences between genes. These intergenic sequences are named and oriented with respect to the flanking genes, and distances between genes are given. These intergenic sequences include many conserved features that lie between genes such as promoters, terminators, TFBSs and intergenic repeat families, e.g. CRISPR and REP elements (23). As these features are genetic determinants that lie outside of gene borders, they are termed intergenes. There are TopicPages for all the families of intergenic repeats. An option to exclude the intergenic repeats is provided causing the intergenic intervals to be broken up and named by the flanking genes and repeats. A link to the Intergenic Repeat Families super-TopicPage is provided on the EcoTools page. In addition to the DNA sequence files in FASTA format, a descriptive table of intergenic regions is provided for downloading. A download option to include adjacent gene overlaps as well as the intervals between genes is provided; the length of the overlap is provided as a negative interval length. Downloaded files have pseudogene gene names marked with apostrophes to indicate their pseudogene status. The pseudogene TopicPage explains the pseudogene annotation conventions used in EcoGene and can be accessed from the EcoTools page.

The EcoGene Tools page has files available for direct download, including EcoGene and EcoProt sequences files, with or without the partial and reconstructed pseudogene sequences, a PDF map of the entire chromosome in PrintMap format, and the Verified Set table (see below). Also provided is the EcoGene OpenSearch browser plug-in that enables EcoGene gene searches in browser toolbars.

GO term indexing

The Gene Ontology (GO) project is a collaborative system for gene products classification and description using three structured controlled vocabularies: cellular component, molecular function and biological process (http://www.geneontology.org/GO.doc.shtml) (24). All the GO terms currently associated with *E. coli* K-12 genes and the GO hierarchy of terms, names and definitions, including all ancestor and children relations, were obtained from the Amigo website (25). The *E. coli* GO term assignments at Amigo are compiled from several sources including EcoCyc (26), EcoliWiki (27), UniProtKB (11,28) and InterPro (29). We made two observations consistent with problems reported elsewhere (30). First, the GO annotations are incomplete and many proteins lack GO annotations consistent with their molecular function as currently understood. In some cases, a more specific child term can be applied to better describe the molecular function, whereas in other cases, the GO term is either missing or out-of-date. Second less specific ancestor terms are inconsistently and sporadically applied. Many new gene functions are first annotated in EcoGene and then propagated worldwide to other databases using the regularly updated Genbank U00096 record. We address the problem of incompleteness by monitoring and incorporating updates at Amigo and updating GO terms at EcoGene during the daily literature-based updating process. The EcoGene staff will periodically provide its updated GO term assignments to the GO consortium. To address the problem of sporadic ancestor term assignments, we remove all ancestors and retain only the most specific child terms for each function thread. We then derive all less-specific ancestor terms using the GO hierarchy.

The GO terms are displayed on the Protein tabs associated with GenePages (Figure 3). In addition to the GO term accession numbers, the number of genes



Function: Enzyme; Degradation of small molecules: Carbon compounds

Protein Databases: UniProtKB P00722 Genpept NP_414878.1 GenPept-GenBank AAC73447
EC Number: 3.2.1.23
GO Molecular Function: GO:0004565 (2 genes) GO:0030246 (37 genes) GO:0046872 (680 genes)
◢ Show less specific ances

Name: beta-galactosidase activity

Definition: Catalysis of the hydrolysis of terminal, non-reducing beta-D-galactose residues in beta-D-galactosides.

Accession	Name		Inferred By
GO:0015925 3 genes	galactos activity	mpounds, substances containing a group derived from erivative.	GO:0004565
GO:0004553 34 genes	hydrolas activity, hydrolyzing O-glycosyl compounds	Catalysis of the hydrolysis of any O-glycosyl bond.	GO:0004565
GO:0016798 53 genes	hydrolase activity, acting on glycosyl bonds	Catalysis of the hydrolysis of any glycosyl bond.	GO:0004565
GO:0016787 698 genes	hydrolase activity	Catalysis of the hydrolysis of various bonds, e.g. C-O, C-N, C-C, phosphoric anhydride bonds, etc. Hydrolase is the systematic name for any enzyme of EC class 3.	GO:0004565
GO:0003824 1998 genes	catalytic activity	Catalysis of a biochemical reaction at physiological temperatures. In biologically catalyzed reactions, the reactants are known as substrates, and the catalysts are naturally occurring macromolecular substances known as enzymes. Enzymes possess specific binding sites for substrates, and are usually composed wholly or largely of protein, but RNA that has catalytic activity (ribozyme) is often also regarded as enzymatic.	GO:0004565
GO:0005488 2063 genes	binding	The selective, non-covalent, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule.	GO:0030246 GO:0046872
GO:0043169 696 genes	cation binding	Interacting selectively and non-covalently with cations, charged atoms or groups of atoms with a net positive charge.	GO:0046872
GO:0043167 699 genes	ion binding	Interacting selectively and non-covalently with ions, charged atoms or groups of atoms.	GO:0046872

GO Biological Process: GO:0005990 (1 genes)
▷ Show less specific ancestors

GO Cellular Component: GO:0009341 (2 genes)
▷ Show less specific ancestors

Figure 3. A GenePage Protein tab showing the GO terms with automatic expansion to a full list of less specific ancestor GO terms.

associated with specific GO terms are indicated in parentheses. Clicking the number of genes retrieves all those gene records. The link associated with the GO accession leads to the descriptive Amigo page for that GO term including a listing of all the ancestors and children for that GO term. Mousing over the GO term accession reveals its name and definition. All of the less-specific ancestors are hidden but can be revealed with a click. EcoGene is fully indexed with the expanded ancestor GO terms, and EcoSearch now has a dedicated search box for GO terms as an alternate way to retrieve genes without using the links on the Protein page. GO terms are being updated in-house by adding missing GO child terms, and by replacing child terms with more specific child terms if they exist.

EcoGene-RefSeq

To make the maps and tools developed for EcoGene available for other genomes, EcoGene-RefSeq was developed. Currently, 2074 complete bacterial genomes from the NCBI RefSeq project (31) are accessible through EcoGene-RefSeq. EcoGene-RefSeq is organized independently from EcoGene, with its own MySql database and web interface accessible from the EcoGene home page. Genome data is imported from NCBI in Generic Feature Format Version 3.0 format and stored in MySql tables. The EcoTools currently ported to EcoGene-RefSeq include PrimerPairs, Search and Download, GenePage navigation and Cross Reference mapping. An alphabetical gene index is also provided. A detailed bioinformatic description of EcoGene-RefSeq will be published separately (J.Z. and K.E.R., manuscript in preparation).

The verified set

The *E. coli* K-12 MG1655(Seq) genome sequence annotation is kept current with the biomedical literature by extensive manual curation at EcoGene.org. This includes updating the Verified Set of proteins whose starts have been determined by protein sequencing using Edman degradation. In 2000, the Verified Set contained 717 protein starts and citations to their published protein sequencing results (8). Currently, the Verified Set is composed of 923 verified protein starts identified in 818 publications. The verified proteins include 419 unprocessed proteins, 356 proteins whose *N*-terminal methionine is removed by methionine aminopeptidase, 142 exported proteins cleaved by signal peptidase I and six starts generated by various proteolytic cleavages. In addition to compiling published verifications, the EcoGene staff performs selected experimental verifications of predicted structural and functional annotations. The signal peptidase I cleavage sites and periplasmic locations were experimentally determined for HiuH and YhcN, as noted on their GenePages (N. Hus, R. Mitchell, M. Del Campo, and K.E.R., unpublished results). In most cases, the Verified Set also confirms a predicted translation start codon. The EcoTools page has a full table of the Verified Set, including PubMed identifiers as documentation. In addition, 82 proteins have been verified as lipoproteins in prior publications, and 38 proteins have been predicted to be lipoproteins in a comparative analysis of lipoprotein predictions in an EcoGene collaborative publication (32). These lipoproteins are clustered in EcoTopics.

CTG start codon verification

Start codon mutagenesis was used to confirm the second instance of CTG being used as a start codon in the *E. coli* K-12 genome. The *hda* gene was the first *E. coli* chromosomal gene shown to initiate translation with CTG (33). As part of an ongoing bioinformatics analysis of predicted initiation codons (8), two additional CTG starts were predicted based on sequence conservation and inspection of potential ribosome binding sites. Using start codon mutagenesis and protein fusions, the CTG start codon predicted for *yfjD* was confirmed as the start codon, while mutating the previously annotated TTG start at codon 9 had no effect on gene expression (D. Dague and K.E.R., unpublished results). The pseudogene *yghF'* encodes an *N*-terminal gene fragment that is also predicted to start with CTG at EcoGene.org, but this could not be confirmed because expression from YghF'-LacZ protein fusions was negligible even when additional *yghF'* upstream DNA was included in the plasmid construct.

Small protein and sRNA verifications

The prediction, annotation and functional characterization of small proteins and sRNAs remain a challenge (34,35). Two EcoGene collaborative projects combined bioinformatic and experimental approaches to identify novel small proteins and sRNAs in *E. coli*. Protein translations of 20 previously annotated and 18 newly predicted small proteins (16–50 amino acids), including nine new membrane proteins, were experimentally verified (36). Additionally, the novel IbsA-E toxins, encoded opposite the SibA-E sRNA antitoxins, are composed of only 18 or 19 amino acids and were experimentally characterized (37). The Sib sRNAs were originally predicted to be sRNAs encoded in the novel QUAD repeats, as reported in an earlier EcoGene bioinformatics study (23). The Sib sRNAs were experimentally verified independently in both the EcoGene and Storz laboratories and the QUAD repeats were renamed as SIB repeats (37).

The COMBREX-EcoGene collaboration

The COMBREX project aims to identify the functions of uncharacterized genes in *E. coli* and other bacteria (38,39). A 1-year sub-award from the COMBREX project funded a COMBREX–EcoGene collaboration to functionally characterize *E. coli* genes. The identification of YhiQ as RsmJ, the only unidentified 16S rRNA methyltransferase in *E. coli*, responsible for the m(2)G1516 modification, was a result of that collaboration (40). In addition, the *yaiX'* pseudogene was repaired and shown to facilitate clumping during vegetative growth. The previously uncharacterized YaiP, a predicted glucosyltransferase encoded in the *yaiX* operon was also demonstrated to be required for vegetative clumping, a result that could not be obtained until the *yaiX'* pseudogene was repaired (Dague *et al.*, abstract presented at The 2011 Madison Molecular Genetics of Bacteria and Phages Meeting, August 2–7, Madison, WI).

The E. coli *K-12 ignome*

The functional characterization of products from the y-genes (see below) *yihA, yeeX, yhgF, ybcJ, yhbY, yjgA, yhdE, yceF* and *yhc*N was initiated with the COMBREX funding. These are now ongoing EcoGene laboratory projects. These uncharacterized gene products are part of the *E. coli* K-12 *ignome*. We define an *ignome* as all the knowledge we lack about an organism, including structural, functional, regulatory, spatial and organization information about all cellular constituents. A practical *ignome* is all the knowable information about an organism, acknowledging that the acquisition of knowledge may be limited by costs (time, money, means) or by insolvable technical hurdles (Figure 4).

Keio collection and ASKA set verifications

The Keio collection contains deletions tagged with kanamycin resistance cassettes in the non-essential genes of *E. coli* K-12, created using high-throughput recombineering methodology (41,42). The Keio collection is a widely distributed and extensively used genetic resource that has a number of problems discovered during our extensive quality control assessment. Although the verifications are not completed, a preliminary summary of the problems we found has been published (22). One of the major problems relates to the lack of isogenicity for ~65% of the frozen cultures in the collection. These cultures contain variants that have acquired hyper-motility mutations mostly due to IS1 or IS5 insertions upregulating the *flhDC* locus. We have also observed that the semi-motile parent strain BW25113 is genetically unstable

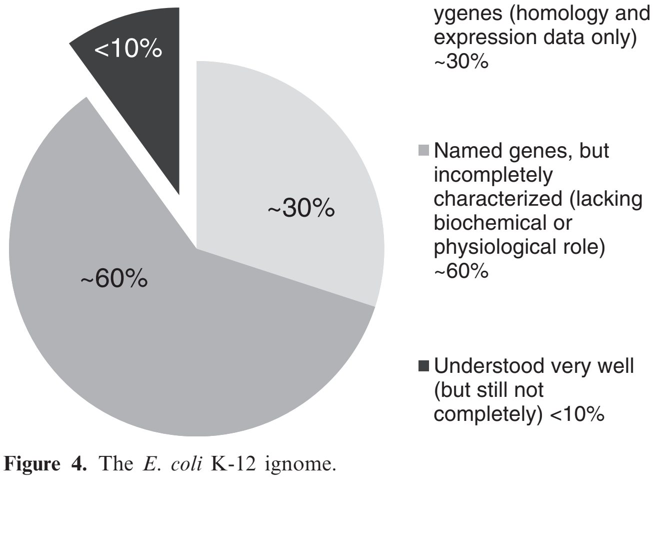


Figure 4. The *E. coli* K-12 ignome.

in the laboratory and can acquire hyper-motility mutations, as has been demonstrated for other semi-motile *E. coli* K-12 strains (15,43). This instability at *flhDC* is responsible for the difference between MG1655 and MG1655(Seq), as noted in the Introduction. To stabilize the Keio collection, we crossed the entire collection into the hyper-motile strain KRE10000 [MG1655(Seq) *rph*+] by P1 transduction and *re*-recombineering, as noted previously (22). This whole collection transduction outcross revealed that ~100 slow-growing strains had picked up unlinked growth defect suppressor mutations in the Keio collection. We also identified the same tandem duplications that were also found by the Keio group using their PCR primers (42).

As part of our quality assessment of the Keio collection, we confirmed the auxotrophic phenotypes by feeding with specific nutrients (D. Dague and K.E.R., unpublished). A companion genome-wide set of clones, the ASKA library, has also been constructed and widely distributed (44). We tested the ASKA collection for the ability to complement Keio mutants and found that all the auxotrophs we tested were able to grow without supplements when they harbor un-induced cognate ASKA clones (D. Dague and K.E.R., unpublished). Bioinformatic validation of the deletion primers used to make the Keio collection found numerous adjacent-gene double deletions. Identification of these adjacent-gene double deletions inspired development of the PrimerPairs genome-wide primer design tool used to eliminate double deletions when designing PCR primers (22). We used ASKA clone complementation to show that the adjacent-gene 3′ double deletions in amino acid biosynthetic operons are innocuous, that is, complementing only the targeted gene restored prototrophy (D. Dague and K.E.R., unpublished). This demonstrates that most, if not all, of the small 3′ deletions in the genes adjacent to the genes targeted in 468 strains do not cause a mutant phenotype, and this is not unexpected, as proteins generally tolerate modifications at their C-termini, such as hexa-histidine tags. In contrast, the small adjacent-gene 5′ deletions completely eliminate adjacent-gene function in the few mutants that could be easily tested (D. Dague and K.E.R., unpublished). The 32 adjacent-gene 5′ deletion mutants have been reconstructed by *de novo* recombineering in KRE10000 (D. Dague and K.E.R., unpublished).

Additional information regarding the Keio strains and ASKA clones can be found through links on GenePages to the Genobase website (http://ecoli.naist.jp/GB8-dev/). GenePages also contain a Keio/ASKA tab with specific information on strains and clones, including their locations in microtiter trays, along with pictures of PCR-amplified fragments from the *flhDC* locus of each mutant in the Keio collection, indicating the presence or absence of IS1 or IS5 insertion sequences responsible for hyper-motility phenotypes. The PCR reactions were performed on the original frozen cells and identified many mixed wells with both semi-motile (IS−) and hyper-motile (IS+) cells (K.E.R. unpublished). When using the Keio mutants, care should be taken to follow standard bacterial genetics methodology of single

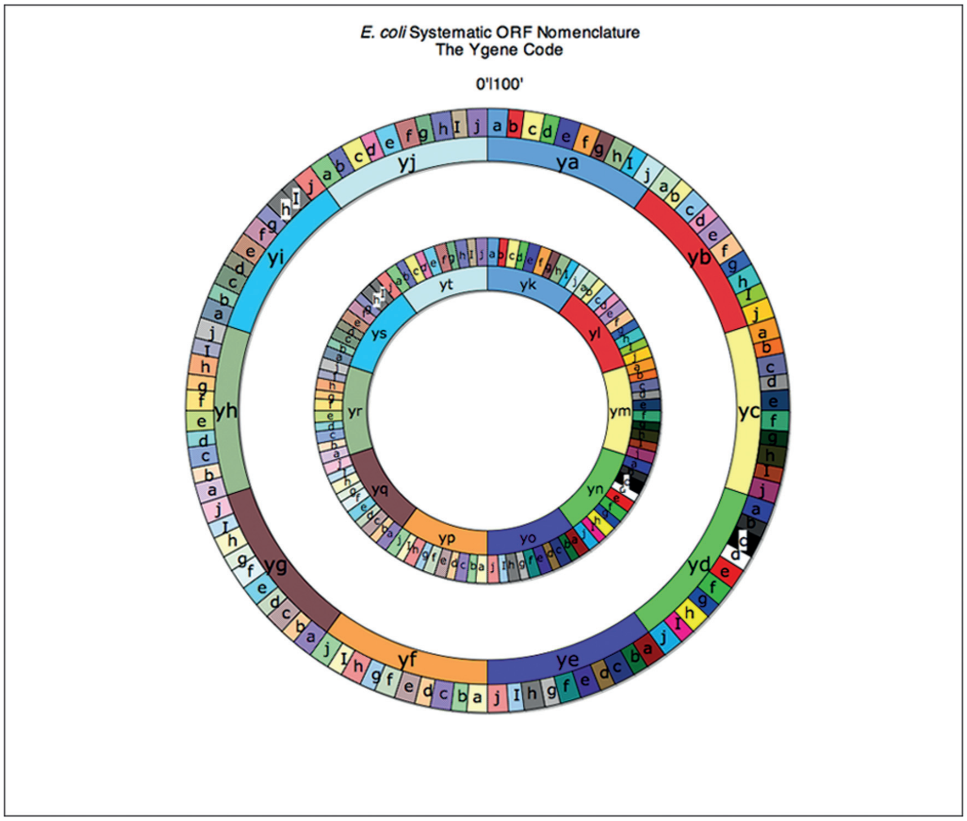


Figure 5. The y-gene code wheel depicting the systematic uncharacterized gene nomenclature rationale for *E. coli* K-12 is shown.

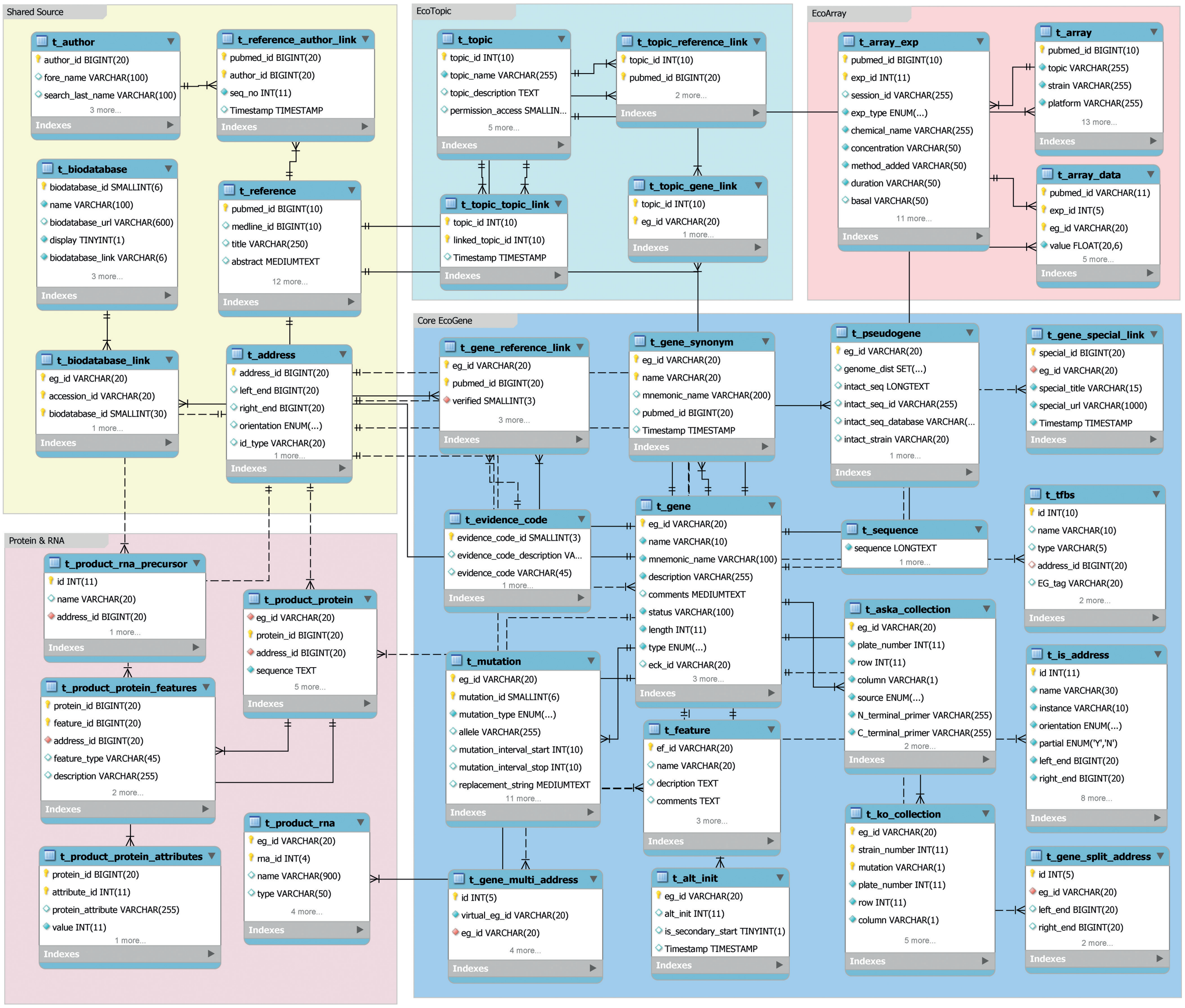


Figure 6. EcoGene 3.0 MySql database schema. Some minor fields have been deleted for clarity. The five modules are color-coded.

colony purification, phenotypic (motility) testing and P1 transduction as necessary to restore isogenicity. Published high-throughput results that used Keio mutants directly from trays without single colony purification need to be carefully re-examined.

Y-gene nomenclature

Genes without names or known functions were previously assigned y-gene names using a systematic nomenclature based on map location (4). Although initially assigned based on map position, e.g. *yaaA-Z* are genes in the first centisome (map minute, %) interval, the y-gene names are not intended to denote map position and have been used to denote the orthologous gene in *S. typhimurium* despite having a different map location. If any particular centisome interval has more than 26 unnamed genes, a second set of y-names can be used, e.g. *ykaA-Z* for the first centisome interval (Figure 5). Many y-genes have now been assigned functions, but authors often choose not to assign new meaningful mnemonic gene names as previously suggested (4) because they may not fully understand the biochemical activity or physiological role of the newly characterized gene. However, *E. coli* genes have been traditionally assigned names at the first, not final, characterization effort; as long as the initial mnemonic does not turn out to be completely incorrect, which rarely happens, it should be retained. This means that mnemonic names describing the functions or phenotypes can and should be assigned after their initial characterization. On the other hand, some authors assign new genes names to y-genes without first checking at EcoGene.org to see if the gene name has been previously used. Mnemonic gene names that do not refer to gene function or mnemonics that are common words, which are poor terms for computer searches, are discouraged. In addition, some authors assign new genes even when the primary gene name mnemonic is still valid just because they have discovered more detailed functional information. This re-naming can go on indefinitely, engenders confusion and is to be discouraged. These superfluous, albeit informative, re-namings are entered in EcoGene as gene synonyms; the primary gene names and synonyms for *E. coli* K-12 are maintained at EcoGene.org. Authors can consult the EcoGene curator (K.E.R.) on the choice of new gene names. A TopicPage devoted to the y-genes with additional information is available at EcoGene.org (http://www.ecogene.org/topic/364).

The EcoGene MySql relational database

The first edition of EcoGene used static HTML pages, essentially publishing rows from a spreadsheet into early GenePages as periodic updates (8). To provide continuously updated information to EcoGene.org, a live relational database is required. In 2005, a relational schema for EcoGene was created to build cross-references for the breadth of information being published describing the biology of the *E. coli* K-12 laboratory workhorse strain and to accommodate a detailed level of structural annotation, including intergenes and pseudogenes (Figure 6). The schema is modular, which assists the database to be robust, efficient and flexible in our experience. Five data-specific modules are displayed: Core EcoGene, Shared Sources, Protein & RNA, EcoTopic and EcoArray. Additional modules include the MySql tables for the Drupal content management system (http://drupal.org/) and recently added tables for GO term descriptions and an ancestor-expanded gene-GO linking table.

The Core EcoGene module holds the genome sequence, the gene-pubmed cross-references and genomic address for genes and intergenes. It includes information for genes, synonyms, IS elements, repeat DNA, e.g. REPs and CRISPRs, and TFBSs (the latter imported from RegulonDB). Pseudogenes interrupted by IS insertion have split addresses analogous to exons. The Genbank format 'join statements' describing hypothetical reconstructions of the split genes that are displayed on their GenePages are derived from the t_split_gene_address table. IS elements present at multiple locations are represented in the t_gene_multi_address table, allowing for direct GenePage navigation among the various IS alleles and locations. Information related to the widely used Keio mutant and ASKA clone collections are also stored here.

The Shared Sources module holds tables for genomic gene and feature basepair positions, data derived from PubMed records and curated links to external databases. The Protein & RNA module contains product information derived from bacterial RefSeq records; for EcoGene-curated *E. coli* K-12 MG1655, this is where the GenBank record product descriptions are updated and stored between U00096 updates. The EcoTopic and EcoArray modules were added to support the ongoing development of two ambitious EcoGene projects.

EcoGene.org content management platform

EcoGene 2.0 was the default EcoGene.org website from 2006 until EcoGene 3.0 was officially launched in September 2012. During a 1-year beta deployment period, >80% of EcoGene.org usage switched to EcoGene 3.0. EcoGene 2.0 will remain available for users who are accustomed to its interface; it accesses the same daily updated MySql database as EcoGene 3.0, but all new development projects will be restricted to EcoGene 3.0. For example, the interactive Venn diagram tool is only available in EcoGene 3.0. EcoGene.org, launched in 2006, is powered by the open source content management software PHP-Nuke (http://phpnuke.org) and supported by a MySql database (http://www.mysql.com). However, in 2010 we desired improved capabilities and support, as well as a new look-and-feel for the EcoGene.org interface, so we evaluated popular open source content management platforms and chose Drupal to build EcoGene 3.0. Most of the EcoGene.org 2.0 PHP (http://www.php.net) software was rewritten to be compatible with Drupal (http://drupal.org). The EcoGene.org website operational algorithms, scripts and routines were also renovated and streamlined to facilitate future improvement and expansion.

ACKNOWLEDGEMENTS

We thank Fred Blattner and Guy Plunkett for facilitating the transfer of annotation update responsibility for Genbank U00096 to K.E.R. We acknowledge the participation of William Klimke (NCBI), Ingrid Keseler (EcoCyc), Guy Plunkett (ASAP), Eric Neeno-Eckwall (ASAP) and Andrea Auchincloss (UniProtKB) for reviewing the periodic annotation updates sent from EcoGene to NCBI. We acknowledge computer programming on EcoGene 2.0 by Dian Fan and Feng Liu. We would like to thank Julio Collado-Vides of the University Nationale de Mexico for permission to use data from RegulonDB in EcoGene. We thank Andrew Richardson for technical assistance, use of unpublished results, and a critical reading of the manuscript. We acknowledge Andrew Richardson, Yusuf Kaya, Kristi Jones, Darryl Dague, Joseph Reed, and Christopher Crowley for work on the Keio collection validation and various laboratory projects associated with EcoGene and COMBREX-EcoGene. We also acknowledge technical assistance from Nir Hus, Mark Del Campo and Rick Mitchell for unpublished results from verification studies. We acknowledge Kristi Jones for her help in designing and curating EcoArray.

FUNDING

National Institutes of Health [1-R01-GM58560, 3-U24-GM077905-3S1 subaward #4102-25541, RC2-GM92602 subaward]. Funding for open access charge: Waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Rudd,K.E., Miller,W., Werner,C., Ostell,J., Tolstoshev,C. and Satterfield,S.G. (1991) Mapping sequenced *E. coli* genes by computer: software, strategies and examples. *Nucleic Acids Res.*, **19**, 637–647.
2. Miller,W., Ostell,J. and Rudd,K.E. (1990) An algorithm for searching restriction maps. *Comput. Appl. Biosci.*, **6**, 247–252.

3. Kohara,Y., Akiyama,K. and Isono,K. (1987) The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell*, **50**, 495–508.
4. Rudd,K.E. (1998) Linkage map of *Escherichia coli* K-12, edition 10: the physical map. *Microbiol. Mol. Biol. Rev.*, **62**, 985–1019.
5. Blaisdell,B.E., Rudd,K.E., Matin,A. and Karlin,S. (1993) Significant dispersed recurrent DNA sequences in the *Escherichia coli* genome. Several new groups. *J. Mol. Biol.*, **229**, 833–848.
6. Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res.*, **22**, 4756–4767.
7. Borodovsky,M., McIninch,J.D., Koonin,E.V., Rudd,K.E., Medigue,C. and Danchin,A. (1995) Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res.*, **23**, 3554–3562.
8. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
9. Blattner,F.R., Plunkett,G. 3rd, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
10. Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T. *et al.* (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res.*, **34**, 1–9.
11. Apweiler,R., Jesus Martin,M., O'onovan,C., Magrane,M., Alam-Faruque,Y., Antunes,R., Barrera Casanova,E., Bely,B., Bingley,M., Bower,L. *et al.* (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
12. Keseler,I.M., Collado-Vides,J., Santos-Zavaleta,A., Peralta-Gil,M., Gama-Castro,S., Muniz-Rascado,L., Bonavides-Martinez,C., Paley,S., Krummenacker,M., Altman,T. *et al.* (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.*, **39**, D583–D590.
13. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.
14. Glasner,J.D., Rusch,M., Liss,P., Plunkett,G. 3rd, Cabot,E.L., Darling,A., Anderson,B.D., Infield-Harm,P., Gilson,M.C. and Perna,N.T. (2006) ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. *Nucleic Acids Res.*, **34**, D41–D45.
15. Barker,C.S., Pruss,B.M. and Matsumura,P. (2004) Increased motility of *Escherichia coli* by insertion sequence element integration into the regulatory region of the *flhD* operon. *J. Bacteriol.*, **186**, 7529–7537.
16. Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S., Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. *et al.* (2006) Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol. Syst. Biol.*, **2**, 2006 0007.
17. Freddolino,P.L., Amini,S. and Tavazoie,S. (2012) Newly identified genetic variations in common *Escherichia coli* MG1655 stock cultures. *J. Bacteriol.*, **194**, 303–306.
18. Fabich,A.J., Leatham,M.P., Grissom,J.E., Wiley,G., Lai,H., Najar,F., Roe,B.A., Cohen,P.S. and Conway,T. (2011) Genotype and phenotypes of an intestine-adapted *Escherichia coli* K-12 mutant selected by animal passage for superior colonization. *Infect. Immun.*, **79**, 2430–2439.
19. Davids,W. and Zhang,Z. (2008) The impact of horizontal gene transfer in shaping operons and protein interaction networks—direct evidence of preferential attachment. *BMC Evol. Biol.*, **8**, 23.
20. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2010) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.
21. Gama-Castro,S., Salgado,H., Peralta-Gil,M., Santos-Zavaleta,A., Muniz-Rascado,L., Solano-Lira,H., Jimenez-Jacinto,V., Weiss,V., Garcia-Sotelo,J.S., Lopez-Fuentes,A. *et al.* (2011) RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.
22. Zhou,J. and Rudd,K.E. (2011) Bacterial genome reengineering. *Methods Mol. Biol.*, **765**, 3–25.
23. Rudd,K.E. (1999) Novel intergenic repeats of *Escherichia coli* K-12. *Res. Microbiol.*, **150**, 653–664.
24. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
25. Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B. and Lewis,S. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.
26. Hu,J.C., Karp,P.D., Keseler,I.M., Krummenacker,M. and Siegele,D.A. (2009) What we can learn about *Escherichia coli* through application of Gene Ontology. *Trends Microbiol.*, **17**, 269–278.
27. McIntosh,B.K., Renfro,D.P., Knapp,G.S., Lairikyengbam,C.R., Liles,N.M., Niu,L., Supak,A.M., Venkatraman,A., Zweifel,A.E., Siegele,D.A. *et al.* (2012) EcoliWiki: a wiki-based community resource for *Escherichia coli*. *Nucleic Acids Res.*, **40**, D1270–D1277.
28. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Mun Chan,W., Eberhardt,R. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.
29. Burge,S., Kelly,E., Lonsdale,D., Mutowo-Muellenet,P., McAnulla,C., Mitchell,A., Sangrador-Vegas,A., Yong,S.Y., Mulder,N. and Hunter,S. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database (Oxford)*, **2012**, bar068.
30. Faria,D., Schlicker,A., Pesquita,C., Bastos,H., Ferreira,A.E., Albrecht,M. and Falcao,A.O. (2012) Mining GO annotations for improving annotation consistency. *PLoS One*, **7**, e40519.
31. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
32. Gonnet,P., Rudd,K.E. and Lisacek,F. (2004) Fine-tuning the prediction of sequences cleaved by signal peptidase II: a curated set of proven and predicted lipoproteins of *Escherichia coli* K-12. *Proteomics*, **4**, 1597–1613.
33. Su'etsugu,M., Nakamura,K., Keyamura,K., Kudo,Y. and Katayama,T. (2008) Hda monomerization by ADP binding promotes replicase clamp-mediated DnaA-ATP hydrolysis. *J. Biol. Chem.*, **283**, 36118–36131.
34. Rudd,K.E., Humphery-Smith,I., Wasinger,V.C. and Bairoch,A. (1998) Low molecular weight proteins: a challenge for post-genomic research. *Electrophoresis*, **19**, 536–544.
35. Hemm,M.R., Paul,B.J., Miranda-Rios,J., Zhang,A., Soltanzad,N. and Storz,G. (2010) Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. *J. Bacteriol.*, **192**, 46–58.
36. Hemm,M.R., Paul,B.J., Schneider,T.D., Storz,G. and Rudd,K.E. (2008) Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol. Microbiol.*, **70**, 1487–1501.
37. Fozo,E.M., Kawano,M., Fontaine,F., Kaya,Y., Mendieta,K.S., Jones,K.L., Ocampo,A., Rudd,K.E. and Storz,G. (2008) Repression of small toxic protein synthesis by the Sib and OhsC small RNAs. *Mol. Microbiol.*, **70**, 1076–1093.
38. Roberts,R.J. (2011) COMBREX: COMputational BRidge to EXperiments. *Biochem. Soc. Trans.*, **39**, 581–583.
39. Roberts,R.J., Chang,Y.C., Hu,Z., Rachlin,J.N., Anton,B.P., Pokrzywa,R.M., Choi,H.P., Faller,L.L., Guleria,J., Housman,G. *et al.* (2011) COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. *Nucleic Acids Res.*, **39**, D11–D14.
40. Basturea,G.N., Dague,D.R., Deutscher,M.P. and Rudd,K.E. (2012) YhiQ is RsmJ, the methyltransferase responsible for methylation of G1516 in 16S rRNA of *E. coli*. *J. Mol. Biol.*, **415**, 16–21.
41. Baba,T., Ara,T., Hasegawa,M., Takai,Y., Okumura,Y., Baba,M., Datsenko,K.A., Tomita,M., Wanner,B.L. and Mori,H. (2006)

Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.*, **2**, 2006.0008.

42. Yamamoto,N., Nakahigashi,K., Nakamichi,T., Yoshino,M., Takai,Y., Touda,Y., Furubayashi,A., Kinjyo,S., Dose,H., Hasegawa,M. *et al.* (2009) Update on the Keio collection of *Escherichia coli* single-gene deletion mutants. *Mol. Syst. Biol.*, **5**, 335.
43. Wang,X. and Wood,T.K. (2011) IS5 inserts upstream of the master motility operon *flhDC* in a quasi-Lamarckian way. *ISME J.*, **5**, 1517–1525.
44. Kitagawa,M., Ara,T., Arifuzzaman,M., Ioka-Nakamichi,T., Inamoto,E., Toyonaga,H. and Mori,H. (2005) Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 ORF archive): unique resources for biological research. *DNA Res.*, **12**, 291–299.