Published in final edited form as:
*Trends Microbiol*. 2009 July ; 17(7): 269–278. doi:10.1016/j.tim.2009.04.004.

# What we can learn about *Escherichia coli* through application of Gene Ontology

**James C. Hu**,
Dept. of Biochemistry and Biophysics, Texas A&M University and Texas Agrilife Research, College Station, TX 77843-2128, USA

**Peter D. Karp**,
Bioinformatics Research Group SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, USA

**Ingrid M. Keseler**,
Bioinformatics Research Group SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, USA

**Markus Krummenacker**, and
Bioinformatics Research Group SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, USA

**Deborah A. Siegele**
Department of Biology, Texas A&M University, College Station, TX 77843-3258, USA

James C. Hu: jimhu@tamu.edu; Peter D. Karp: pkarp@ai.sri.com; Ingrid M. Keseler: keseler@ai.sri.com; Markus Krummenacker: kr@ai.sri.com; Deborah A. Siegele: siegele@mail.bio.tamu.edu

## Abstract

How we classify the genes, products, and complexes that are present or absent in genomes, transcriptomes, proteomes, and other datasets helps us place biological objects into subsystems with common functions, see how molecular functions are used to implement biological processes, and compare the biology of different species and strains. Gene Ontology (GO) is one of the most successful systems for classifying biological function. Although GO is widely used for eukaryotic genomics, it has not yet been widely used for bacterial systems. The potential applications of GO are currently limited by the need to improve the annotation of bacterial genomes with GO and to improve how prokaryotic biology is represented in the ontology. In this review, we will discuss why GO should be adopted by microbiologists, and describe recent efforts to build and maintain high-quality GO annotation for *Escherichia coli* as a model system.

## Organizing the parts lists of life

High throughput DNA sequencing has progressed to the point where obtaining DNA sequence is no longer rate limiting for genome projects, especially for small bacterial genomes. Determining how the biology of an organism is controlled by its genome is now limited by understanding what functions are encoded in the mass of genomic DNA sequences coming from genomes and metagenomes. An important and currently limiting step in this process is annotation: categorizing the functions of the genes that comprise the parts list of an organism. Because core biological systems share common evolutionary origins, classification systems for categorizing gene function can be shared across all

Correspondence to: James C. Hu, jimhu@tamu.edu.

domains of life and functional insights from well-studied model organisms can be used to infer function in other systems. This review focuses on current efforts to apply the a universal categorization system, the Gene Ontology (GO) to *E. coli*, arguably one of the best-studied model organisms.

## Classification of E. coli gene function

Some of the earliest functional classification of gene products was applied to *E. coli* long before complete genomes were available. In the 1980s, Ingraham and colleagues[1] divided 892 mapped *E. coli* genes into a hierarchy with eleven top-level categories based on metabolic function. In 1993, Monica Riley extended these ideas in a revised clasification system, with six top-level groupings and 31 second-level groupings; online access to this categorization of *E. coli* genes was first provided by GenProtEc[2] (http://genprotec.mbl.edu/) and EcoCyc[3] (http://ecocyc.org). As complete bacterial genomes became available additional classification schemes were built on the foundation of the Riley's system. Fleischmann et al.[4] adapted the Riley classification scheme, but rearranged it into 102 functional categories in 14 higher level groups for the initial annotation of the *Haemophilus influenzae* Rd genome in 1995. This evolved into the 121 microbial TIGR role categories (http://cmr.jcvi.org/tigr-scripts/CMR/RoleIds.cgi)[5]. The Riley classification scheme was also reworked by Blattner et al. [6] into 22 high-level groupings for annotation of the completed *E. coli* K-12 genome. In 2000, Serres and Riley described another classification scheme, MultiFun[7], with the aim of supporting multiple functional annotations for each gene, rather than placing each gene in a single category. Although Multifun can be thought of as descended from the 1993 Riley classification scheme, it represents a major reworking, including a new system of numerical identifiers for terms. Multifun has been extensively used for annotation of bacterial genomes [8-10].

The Gene Ontology (GO) Consortium was begun in the late 1990s as a collaboration between three eukaryotic genome databases, FlyBase (the genome database for *Drosophila melanogaster*), Mouse Genome Informatics (MGI) and the *Saccharomyces* Genome Database (SGD) [11]. Since then, participants in the GO consortium have expanded dramatically, and now include several groups focused on bacterial systems: the Plant-Associated Microbe Gene Ontology Interest Group (PAMGO; http://pamgo.vbi.vt.edu/), the Institute for Genome Sciences at the University of Maryland (http://www.igs.umaryland.edu/), and, recently, EcoCyc (http://ecocyc.org) and EcoliWiki (http://ecoliwiki.net), the community annotation component of the EcoliHub (http://ecolihub.org) project. Toussaint and coworkers are developing GO for phage and plasmids as part of the ACLAME (a CLAssification of Mobile genetic Elements) project [12, 13].

A key feature of GO is its treatment of function as having three distinct senses, represented by three separate ontologies: Cellular Component, Molecular Function, and Biological Process. Cellular Component refers to not only subcellular localization to compartments such as the cytosol or periplasm, but also participation in multisubunit complexes. Molecular Function refers to the specific biochemical function of a gene product, while Biological Process covers multistep physiological processes, such as pathways. GO terms describe a property of a gene, gene product, or complex at varying levels of detail. GO terms are related to one another in a directed acyclic graph (or 'DAG'; Figure 1), where more detailed terms are described as children of more general terms. For example, the GO molecular function **galactokinase activity** (whose unique identifier within GO is GO: 0004335) is a child of two terms: **phosphotransferase activity, alcohol group as acceptor** (GO:0016773) and **carbohydrate kinase** (GO:0019200). These in turn have parent terms, as illustrated in Figure 1, tracing back to the ultimate ancestor, **molecular function** (GO: 0003674), the root of the molecular function ontology. In this case, the children are related

to the parents via an **is_a** relationship: the statements " **galactokinase activity** is_a **phosphotransferase activity, alcohol group as acceptor"** and " **galactokinase activity** is_a **carbohydrate kinase** " are both true, as are statements relating galactokinase activity and any other ancestor term. In GO, this relationship is called the true path rule, and it sets the rule for implicit annotation of gene products. When we annotate the *E. coli galK* product to **galactokinase activity** (that is, when we create an explicit association in a database between the gene product and the GO term), we are also asserting implicitly that it is a carbohydrate kinase, etc.

Each GO annotation for a gene product (Figure 2) should be associated with a literature reference and an evidence code to classify the basis on which a curator or a computational resource asserts the association between a gene product and a particular GO term. Evidence codes can be broadly grouped into those that are manually assigned, and those that are automatically assigned by a computer, as denoted by the IEA (Inferred from Electronic Annotation) evidence code. Note that the manually assigned annotations include both those with experimental evidence and those from computational analyses where a human curator has reviewed their validity.

Certain types of GO annotations also require a **with/from** component. For example, annotations based on sequence similarity evidence codes must denote what sequence the gene of interest is being compared to. Annotations based on interactions must include what biological entities the gene/product interacts with to provide evidence for the association to the GO term. Some annotations also include a **qualifier**. **Not** is used to record where experiments have falsified a hypothesized function. **Contributes_to** is used where a gene product is required for the function of a multisubunit complex, but does not contain the active site.

GO and GO annotations are human-readable, but are also meant to support machine inference systems, via a structure of relationships [14]. The ontology itself can be viewed as a set of declarative statements about classes of biological objects, e.g. **6-phosphofructokinase activity** (GO:0003872) is a kind of **phosphofructokinase activity** (GO:0008443), which includes other phosphofructokinase activities, such as **1-phosphofructokinase activity** (GO:0008662). Annotations are statements about instances of such abstract classes; annotating the *E. coli* phosphofructokinase encoded by *pfkA* with GO:0008662 asserts that the enzyme has the properties defined by that term.

## GO and E. coli

The GO consortium recently celebrated its tenth anniversary. Nevertheless a PubMed search of "gene ontology AND escherichia coli" or "gene ontology AND bacteria" yields primarily a mix of database descriptions and bioinformatic methods. There are more papers using GO to study the host responses to bacterial interactions than there are for using GO for studying bacteriology. This reflects how GO has been used more extensively for eukaryotic systems, and is the reason for the lack of *E. coli* examples below for how GO annotation has provided new insight into the biology of bacterial systems.

We argue below that GO has the potential to provide important biological insight for bacterial genomics and metagenomics. However, currently, exploiting the power of GO annotation is limited by the incompleteness of high quality GO annotation for bacterial gene products, and by the need for improvements in GO itself for prokaryotic biology. In the sections below, we will describe how GO and other structured classification systems could be used, how we are working to improve the quality of GO annotation for *E. coli* and how improvements to GO for the biology of prokaryotes are underway. Using *E. coli*, one of the

best-studied model organisms, as a focus for improving GO for prokaryotes will provide a high density of high-quality, experimentally-based functional annotations that will improve inference of function in other bacteria.

## Using GO: theory and practice

### Retrieving genes with related functions

The most basic uses of GO are to retrieve the GO annotations for a particular gene, and to find the set of genes that are annotated to a particular GO term. *E. coli* is unusual among model organisms in that a large number of independent databases contain *E. coli* GO annotations, including *E. coli*-centric databases, databases covering multiple microbes, and large resources such as UniProt, PDB, and Pfam. GO annotations for *E. coli* genes can be found on the online resources listed in Table 1, but in many cases the information available is incomplete or the update frequency is not clear. The complete current set of GO annotations for *E. coli* is being generated by EcoCyc and EcoliWiki (see below) and can be downloaded from the Gene Ontology Consortium website (http://www.geneontology.org/GO.current.annotations.shtml). Most users will find it easier to browse or search for *E. coli* GO annotations using web-based tools, such as AmiGO (http://amigo.geneontology.org), the GO consortium's browser, or *E. coli*-specific resources such as EcoCyc (http://ecocyc.org), or EcoliWiki (http://ecoliwiki.net).

For example, the EcoCyc web page for the essential cell division protein FtsK (http://biocyc.org/ECOLI/NEW-IMAGE?object=G6464-MONOMER), includes a list of GO annotations for FtsK, including GO:0051301 (cell division). To display other *E. coli* genes are involved in cell division, the user can click on the term to display the EcoCyc page for this GO term, which lists all EcoCyc gene products that are annotated with this GO term as “Term members”. Alternatively, this page can be reached by searching for “cell division” at EcoCyc.

At EcoliWiki, each gene product page includes a user-editable table for GO annotations, which also are listed as links at the bottom of the page. Clicking on one of these links takes the user to an EcoliWiki page that lists genes annotated to that GO term, while each GO ID in the table links to a page on the Gene Ontology Normal Usage Tracking System (GONUTS; http://gowiki.tamu.edu), a wiki-based GO browser (Box 1).

### Inferring function from homologs

By unifying annotation across species GO improves functional annotation of genes based on evolutionary relationships[15, 16]. GO annotations are sought for homologs identified by sequence similarity with your gene of interest, which lacks experimental evidence for function. Annotations found among the homologs are transferred by inference of common function to the gene of interest. With an explosion of sequences for both complete bacterial genomes and environmental metagenomes, annotation transfer based on homology will continue to grow in importance, and the GO consortium provides guidelines for practices used by participating databases for this kind of annotation transfer (http://geneontology.org/GO.evidence.shtml#computational). In particular, a homolog used to provide evidence for a GO annotation must be annotated based on experimental evidence. This policy is designed to avoid errors associated with transitive annotation [17, 18], where the function of gene A is inferred from homology with gene B, but B is annotated not based on experiments, but rather based on homology with gene C, which in turn may or may not be the subject of experimental tests. If A, B, and C encode multidomain proteins, experimental annotation of C may be based on a domain that is present in B, but not in A. The structure of GO

annotations facilitates good practice through the use of evidence codes and by recording the gene used to donate the annotation in the **with/from** field.

Finding the GO annotations for the best experimentally studied orthologs is straightforward. However, even for close orthologs, one cannot simply transfer all the annotations across species. This is perhaps safer for molecular function terms, but can be problematic for biological process and cellular component terms. Annotation transfer between must take into account the differences in basic cell architecture, not only between prokaryotes and eukaryotes, but also among prokaryotes. Shared biological processes may occur in different cellular compartments; automated methods sometimes assign organelle-specific GO terms to genes from *E. coli*. Gene products with the same molecular function do not always participate in all the same biological processes. For example, the MutS protein is associated with mismatch repair in all prokaryotes and eukaryotes that have been studied. However, methyl-directed mismatch repair is only found in a restricted subgroup of bacteria. MutS homologs are also associated with a variety of divergent processes in both prokaryotes and eukaryotes, ranging from limiting intergenomic recombination [19] to antibody class switching and somatic hypermutation of immunoglobulin genes [20, 21], processes which cannot be inferred just by homology.

## Inferring function from GO term enrichment

Although molecular biology on a gene-by-gene basis continues to reveal functional information about uncharacterized genes, a variety of high-throughput ‘guilt by association’ approaches are now available. These include looking for patterns in gene expression from mRNA [22, 23] or proteomics profiles [24], protein-protein interactions from experiments [25], computational analyses [26] or combinations of these [27, 28]. In all cases, a list of genes that share experimental behavior with your favorite gene is returned. For expression studies, it would be a list of genes that are coexpressed with the gene of interest, usually identified by cluster analysis[29, 30]. Interaction studies might return a list of gene products that copurify in a pulldown experiment [31, 32], or a list of genes with synthetic phenotypes with your gene of interest. Phylogenetic profiling [33, 34] returns a list of genes that are present in the same set of genomes as your gene of interest.

The ‘guilt by association’ approach is based on the idea that the functions of genes on these lists, which are reflected in their GO annotations, will include functions relevant to the gene of interest. Although such lists often include genes that do not share functions with your gene, owing to indirect effects, nonspecific protein binding and so on, genes with shared functions will be enriched. Thus, functions of interest can be identified by finding GO terms that are statistically enriched among all terms found on the annotations of genes on your list. The details and problems with this approach have been reviewed elsewhere [35, 36]. There is clearly no single best way to do this kind of analysis; as of this writing, the GO consortium website lists 54 different third-party tools for analyzing expression data. Tools for analysis of expression data using GO have been reviewed extensively elsewhere [36]; term enrichment analysis should be applicable to the other kinds of data described above.

The statistical properties of the annotations themselves can be used for function prediction. Genes annotated to a particular GO term have a nonrandom distribution of annotations to other GO terms[37]; functions, processes and localization are intertwined. King et al. [37] modeled the statistical association of existing annotations within an organism. These models were used to predict missing GO annotations for genes from *Drosophila* or *Saccharomyces cerevisiae*. These predictions had an overall success rate of about 80% when 100 predictions were manually reviewed. Success was defined as human curators concluding that predicted annotations were either “known to be true” but not yet entered in the database (41 of 100) or plausible (42 of 100).

## Studying physiology, genetics and evolution

By aggregating the full set of GO term assignments within a given organism to a subset of higher-level (more abstract) GO terms, a snapshot of the functional repertoire of a genome can be generated, and the repertoires of different species or strains can be compared. Repertoires of GO terms have been used to compare the strategies used by the effectors secreted by different bacterial pathogens into host cells [38]. Correlation of GO process terms with bacteria occupying different ecological niches could provide interesting insights into microbial diversity and evolution.

GO term enrichment can be used to understand the physiological nature of a response to an environmental stress or change in growth conditions. One approach is to examine the GO terms used for genes found to be induced or repressed in an array experiment. Another approach is to examine ontology terms associated with genes with a similar mutant response to an environmenal condition. For example, Rooney and colleagues [39] recently examined ontology terms enriched in mutants from the systematic Keio knockout collection [40] that increased the sensitivity of *E. coli* to the alkylating agent methyl methanesulfonate. Analysis of enriched terms revealed not only the expected DNA damage response genes, but also genes suggestive of responses to RNA and protein damage. (Note that Rooney et al. used Multifun and not GO for this analysis.)

Transcription factors are often easily predicted from gene sequence, but identifying the physiological signals they respond to is not as straightforward. Computing GO term enrichment for the genes located near possible transcription factor binding sites could be used to generate hypotheses about the biological networks controlled by those factors. New methods like ChIP-chip [41] and ChIP-seq [42] have the potential for high-throughput identification of such sites.

## Limitations of GO

Although GO is structured to allow annotations to express the properties of gene products in a structured and controlled ontology, it should be noted that it has limitations with respect to the biological knowledge that can be expressed via GO annotations [5]. GO does not specify when or where a process or function occurs; consider how one might annotate the cellular component for the antisigma factor (a negative regulator of flagellin synthesis) encoded by *flgM*, which is cytosolic when it inhibits FliA-dependent transcription, but is secreted upon assembly of an intermediate structure formed during flagellar biosynthesis [43]. Although GO uses the **with/from** field to annotate interactions, it has very limited encoding of the relationships between nodes, such as the substrate–product relationships that connect consecutive steps in a biochemical pathway. By contrast, the ontology used within EcoCyc [44] and the related data model used by the Reactome project [45, 46] explicitly include inputs and outputs. These support richer sets of computational analyses than GO. Examples include computing the full complement of transportable substrates of an organism [47] and computing the full complement of known transcription-factor ligands for an organism [47].

Despite these limitations, we believe it is still worthwhile for *E. coli* researchers to embrace GO. GO is particularly well suited for classifying the similarities and differences in gene product function across all domains of life. Because *E. coli* is such an important model organism for basic molecular biology, adopting GO for annnotation of *E. coli* gene products is important for annotation transfer to all other organisms. Although other ontologies are designed to provide more expressive relationships, many of these are linked, or are being linked, to GO. Given the strong molecular genetics literature for *E. coli*, and the likelihood that high-throughput screens for phenotypic interactions will become common soon [48, 49],

connections between GO and phenotype ontologies [50, 51] will be useful for exploiting these data.

## Improving E. coli GO annotation

The ability of the *E. coli* biologist to use the kinds of applications described above is premised on a set of GO annotations for *E. coli* with good coverage and accuracy. Functional annotation of *E. coli* gene products has excellent coverage [47], with <15% of the genes assigned to the unknown function classes in Multifun (http://genprotec.mbl.edu/overview.html). Unfortunately, building the *E. coli* GO annotation is not as simple as just mapping of Multifun annotations onto GO owing to the different organization of the two systems and the different annotation practices used. For example, enzymes of the galactose catabolic operon *galEKT* and the GalR protein, which regulates transcription of this operon, are all assigned the Multifun term for carbon compound utilization. Translation from Multifun led to automated annotation of *galR* to the GO term for carbon utilization instead of to a term for regulation of carbon utilization. Similarly, translation of Multifun terms led to annotation of *galE*, which encodes the cytoplasmic UDP-galactose-4-epimerase, to GO component terms related to capsule and cell-surface antigens. Thus, while Multifun annotations can provide starting points for GO annotation, additional manual curation is required.

For the past few years, EcoCyc [47] and EcoliWiki have been collaborating on improving and maintaining the GO annotations for *E. coli*, and, since the summer of 2008, we have been maintaining a gene association file for *E. coli* K-12 (gene_association.ecocyc) that is downloadable from the Gene Ontology Consortium website (http://geneontology.org/GO.current.annotations.shtml). Figure 3 shows the workflow for maintaining this annotation file. EcoCyc incorporates many electronic and experimental GO term annotations of *E. coli* gene products obtained from the “UniProt [multispecies] GO Annotations @ EBI” file downloaded from the Gene Ontology Consortium website (http://geneontology.org/GO.current.annotations.shtml). When this import was first performed in 2007, about 30,000 new IEA 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 [31, 32]. During the import procedure IEA-based annotations were removed if the annotated gene product also had more specific GO annotations based on experimental evidence codes. For example, if a gene product already contained an experimental annotation of the term **galactose kinase,** the software would not add an IEA-based annotation to the less-specific parent term **carbohydrate kinase**. This filtering led to the removal of about 1000 of these less specific and redundant annotations. UniProt annotations are reimported on a regular basis.

The annotations imported from the UniProt file are incorporated into the overall EcoCyc dataset, which includes manual annotation by EcoCyc curators. For example, EcoCyc curators are updating GO annotations for transcriptional regulators and enzymes within metabolic pathways. GO is also used for annotating newly identified functions for previously unannotated gene products. Whenever the curation for gene products is updated by EcoCyc curators, GO terms are updated as well.

In parallel, manual annotation of *E. coli* genes with GO is ongoing at EcoliWiki based on two complementary aims. First, EcoliWiki curators perform curation on genes targeted by the Reference Genomes project of the Gene Ontology Consortium [52]. This is a multi-organism effort to improve annotation consistency across databases, and to provide high-quality annotations as the basis for annotation transfer via homologs. Second, EcoliWiki is developing tools for doing GO annotation as a community curation activity [53]. As a wiki-

based system (Box 1), EcoliWiki allows registered users to edit any annotation on the website. An expert in *E. coli* biology who is unfamiliar with GO can begin an annotation that is refined into the required structure of a GO annotation by someone more familiar with annotation practice, but who is not an expert in the relevant subfields of *E. coli* biology. The rate of manual annotation at EcoliWiki is variable, but has reached 50–100 manual annotations per month.

To merge the annotations from these two efforts, EcoCyc generates a gene association file listing *E. coli* GO annotations from its quarterly releases, which is sent to the EcoliWiki team at Texas A&M. 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 in place with the hope that community members will eventually complete them. New versions of the file are deposited monthly at the GO consortium, and the merged and validated sets are returned to EcoCyc so that any new annotations can be incorporated into the next EcoCyc release. As of March 2009, gene_association.ecocyc includes 42482 GO annotations, including 6016 with non-IEA evidence codes. 1695 of 4472 *E. coli* gene products had one or more non-IEA GO annotations. Including IEA annotations, 3698 *E. coli* genes are annotated with GO. However, only 1813 genes are annnotated in all three aspects of GO.

## Improving GO

Development of the ontology is an ongoing activity of the GO consortium. One of the important properties of GO is that it is designed to accommodate changes [54], and the process of ontology development is meant to be transparent to the scientific community. The GO consortium maintains several email mailing lists for groups working with GO, and has a public Sourceforge bug tracker for requests related to changes in the ontology. These range from requests for new terms, to mark problematic terms as obsolete, and even to reorganization of whole branches of the ontology. For example, processes related to metabolism were rearranged into cellular and organismal processes [54] based on input from BioCyc [8].

Because GO was first developed for eukaryotic model organisms, terms for functions and processes found only in prokaryotes are often missing from the ontology. In addition, some terms that should apply to shared functions in prokaryotes and eukaryotes were defined in such a way that they excluded prokaryotes, either in how they were defined, or in their placement in the ontology. For example, terms related to photosynthetic electron tranport were recently reorganized to reflect the common origin and similar biochemistry of bacterial and chloroplast photosynthesis. In many cases, problematic definitions or placements led to the proliferation of *sensu* terms, e.g. sporulation *sensu* Bacteria, to the ontology. A major project of the GO Consortium over the past year was to eliminate unnecessary *sensu* terms, in order to make the distribution of common functions and processes across organisms more apparent.

Protein secretion illustrates some of the challenges for ontology development. All organisms use conserved components of a general protein translocation pathway to move unfolded preproteins across the membranes separating the cytosol from another compartment [55]. In eukaryotes, the translocon delivers proteins to the lumen of the endoplasmic reticulum, while in the bacteria the destination varies between Gram-negatives such as *E. coli*, where the protein goes to the periplasm and Gram-positives where the protein is secreted into the environment. The translocon that is used to import proteins into mitochondria is analogous

rather than homologous [56], but mitochondrial import shares other conserved components. GO terms describing the relevant functions and processes have been adjusted to capture both the shared biology and the distinct variations, but further modifications to the ontology in this area remain under discussion.

Although some issues in GO are found by browsing the ontology, many are uncovered only when GO terms are used to capture specific biological knowledge in the literature. This means that the quality control in areas of GO relevant to microbiologists is connected to how actively bacterial systems are being curated with GO. The PAMGO project has focused on the interaction of bacterial and fungal pathogens with their plant hosts, leading to the addition of hundreds of appropriate GO terms [57]. Annotation of *E. coli* is already stimulating reexamination of other aspects of GO.

## Concluding remarks and future perspectives

Gene Ontology (GO) is not the only categorization that can be used to represent biology in a form useable by both human curators and computer analyses, but its widespread adoption across diverse biological systems, the depth of what can be described through annotation, and the tools available or being developed for GO, have led the authors to participate in the ongoing effort to annotate *E. coli* gene products to GO. We hope this review will encourage other *E. coli* biologists to join this effort using the community annotation tool in EcoliWiki.

We envision future work building in three major areas (Box 2): improving the rate at which high-quality annotations are made, through either manual or automated methods, improving the structure of GO itself for prokaryotic biology, and applying GO to learn new information about bacterial biology. The first two are the focus of the work reviewed here; the third is likely to be dependent on the success of the other two.

However, it is important to recognize that adoption of GO for *E. coli* will cover only a tiny fraction of the diversity of gene function in prokaryotes. Inference of gene function and systems biology from other bacterial genomes and metagenomes requires the growing involvement of prokaryotic biologists in applying GO annotations to their favorite organism and in expansion of the Gene Ontology itself to include more terms for prokaryotic-specific functions and processes.

## Acknowledgments

We thank Michelle Gwinn-Giglio, Brenley McIntosh, and Anand Venkatraman for comments and help with the manuscript. We also thank the members of the GO consortium and the EcoliHub project, without whom our efforts on GO annotation for *E. coli* would not have been possible. Our work on GO is supported by NIH grant GM077678 to P.K. for EcoCyc and U24GM077905 to J.H. and D.S. as a subcontract to the EcoliHub project.

## Glossary

| | |
|---|---|
| **BioCyc** | BioCyc (http://biocyc.org) is a collection of Pathway/Genome Databases for hundreds of organisms built around the Pathway Tools software [8]. |
| **ChIP-chip and ChIP-seq** | ChIP is an abbreviation for Chromatin Immunoprecipitation, a method to recover DNA bound to a specific protein. In ChIP-chip [41] the DNA is identified by hybridization to a microarray, while in ChIP-seq [63], the DNA is identified by high-throughput sequencing. |

| | |
|---|---|
| **Directed Acyclic Graph (DAG)** | A directed acyclic graph is a set of nodes connected by unidirectional edges, and lacking cycles. Each node can have multiple descendants and/or multiple ancestors, but no ancestor is also a descendant. |
| **EcoCyc** | EcoCyc (http//ecocyc.org) is a literature-based, manually curated Pathway/Genome Database for *E. coli* K-12. EcoCyc is one of the BioCyc databases, and is a partner in the EcoliHub project. |
| **EcoliHub** | EcoliHub (http//ecolihub.org) is a project to provide cross-site unification of online resources related to *E. coli* K-12, its phages, plasmids, and mobile genetic elements. |
| **EcoliWiki** | EcoliWiki (http//ecoliwiki.net) is a wiki-based system for enlisting the scientific community in the ongoing annotation of *E. coli* K-12. EcoliWiki is an EcoliHub component (http://www.ecolicommunity.org). |
| **evidence code** | Evidence codes provide a controlled vocabulary of kinds of experimental or computational evidence used to support an annotation. GO evidence codes are available at the Gene Ontology Consortium website (http://geneontology.org). |
| **Phylogenetic profiling** | Phylogenetic profiling [64] uses patterns of co-occurrence of homologous genes to infer functional interactions. Phylogenetic profiling is based on the idea that all required activities for a biological process will be maintained together by selection for that process, and lost together when selection for the process is absent. |
| **Reactome** | Reactome (www.reactome.org/) is a curated database of metabolic pathways, with emphasis on those pathways found in humans. |
| **Sourceforge bug tracker** | Sourceforge (http//sourceforge.net/ is a free repository for open source software projects. Sourceforge provides bug tracker software to manage bug reports for these projects. |
| **Synthetic phenotypes** | Synthetic phenotypes are nonadditive phenotypes that are observed when multiple mutations are present in the same strain. Suppressors and synthetic lethals are examples of synthetic phenotypes. |
| **Term enrichment analysis** | Statistical analysis of whether an ontology term occurs more frequently in a list of gene annotations than would be expected by chance. See Rhee et al. [35] for additional discussion of approaches to term enrichment and pitfalls. |
| **Transitive annotation** | Annotation via two or more steps of annotation transfer by inference of shared function based on inferred homology. For example if A is homologous to B and B is homologous to C, and only A is annotated based on experimental evidence, transitive annotation would be tranfer of the annotation to C from A via B, instead of directly from A to C. This might occur if A and C are too distantly related to each other to score as homologs. However, it also can occur if A and C are each homologous to different parts of B, in which case the inference of common function is often invalid. |

## References

1. Ingraham, J., et al. Growth of the bacterial cell. Sinauer Associates, Inc; 1983.

2. Riley M, Space DB. Genes and proteins of *Escherichia coli* (GenProtEc). Nucleic Acids Res. 1996; 24:40. [PubMed: 8594596]
3. Karp PD, et al. EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. Nucleic Acids Res. 1996; 24:32–39. [PubMed: 8594595]
4. Fleischmann RD, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995; 269:496–512. [PubMed: 7542800]
5. Peterson JD, et al. The Comprehensive Microbial Resource. Nucleic Acids Res. 2001; 29:123–125. [PubMed: 11125067]
6. Blattner FR, et al. The complete genome sequence of *Escherichia coli* K-12. Science. 1997; 277:1453–1474. [PubMed: 9278503]
7. Serres MH, Riley M. MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. Microb Comp Genomics. 2000; 5:205–222. [PubMed: 11471834]
8. Karp PD, et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res. 2005; 33:6083–6089. [PubMed: 16246909]
9. Degnan PH, et al. Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. Genome Res. 2005; 15:1023–1033. [PubMed: 16077009]
10. Prickett MD, et al. BuchneraBASE: a post-genomic resource for *Buchnera* sp. APS. Bioinformatics. 2006; 22:641–642. [PubMed: 16397006]
11. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]
12. Toussaint A, et al. PhiGO, a phage ontology associated with the ACLAME database. Res Microbiol. 2007; 158:567–571. [PubMed: 17614261]
13. Leplae R, et al. ACLAME: a CLAssification of Mobile genetic Elements. Nucleic Acids Res. 2004; 32:D45–49. [PubMed: 14681355]
14. Smith B, et al. Relations in biomedical ontologies. Genome Biol. 2005; 6:R46. [PubMed: 15892874]
15. Martin DM, et al. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics. 2004; 5:178. [PubMed: 15550167]
16. Xie H, et al. Large-scale protein annotation through gene ontology. Genome Res. 2002; 12:785–794. [PubMed: 11997345]
17. Smith TF, Zhang X. The challenges of genome sequence annotation or “the devil is in the details”. Nat Biotechnol. 1997; 15:1222–1223. [PubMed: 9359093]
18. Iliopoulos I, et al. Evaluation of annotation strategies using an entire genome sequence. Bioinformatics. 2003; 19:717–726. [PubMed: 12691983]
19. Kang J, et al. Structural and functional divergence of MutS2 from bacterial MutS1 and eukaryotic MSH4-MSH5 homologs. J Bacteriol. 2005; 187:3528–3537. [PubMed: 15866941]
20. Wiesendanger M, et al. Somatic hypermutation in MutS homologue (MSH)3-, MSH6-, and MSH3/MSH6-deficient mice reveals a role for the MSH2-MSH6 heterodimer in modulating the base substitution pattern. J Exp Med. 2000; 191:579–584. [PubMed: 10662804]
21. Li Z, et al. The mismatch repair protein Msh6 influences the in vivo AID targeting to the Ig locus. Immunity. 2006; 24:393–403. [PubMed: 16618598]
22. Hughes TR, et al. Functional discovery via a compendium of expression profiles. Cell. 2000; 102:109–126. [PubMed: 10929718]
23. Doniger SW, et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol. 2003; 4:R7. [PubMed: 12540299]
24. VanBogelen RA, et al. Diagnosis of cellular states of microbial organisms using proteomics. Electrophoresis. 1999; 20:2149–2159. [PubMed: 10493120]
25. Vermeulen M, et al. High confidence determination of specific protein-protein interactions using quantitative mass spectrometry. Curr Opin Biotechnol. 2008; 19:331–337. [PubMed: 18590817]
26. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol. 2007; 3:e43. [PubMed: 17465672]

27. von Mering C, et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 2002; 417:399–403. [PubMed: 12000970]
28. Sprinzak E, et al. Characterization and prediction of protein-protein interactions within and between complexes. Proc Natl Acad Sci U S A. 2006; 103:14718–14723. [PubMed: 17003128]
29. Slonim DK. From patterns to pathways: gene expression data analysis comes of age. Nat Genet. 2002; 32 Suppl:502–508. [PubMed: 12454645]
30. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet. 2001; 2:418–427. [PubMed: 11389458]
31. Butland G, et al. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. Nature. 2005; 433:531–537. [PubMed: 15690043]
32. Arifuzzaman M, et al. Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. Genome Res. 2006; 16:686–691. [PubMed: 16606699]
33. Wu J, et al. Gene annotation and network inference by phylogenetic profiling. BMC Bioinformatics. 2006; 7:80. [PubMed: 16503966]
34. Pellegrini M, et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA. 1999; 96:4285–4288. [PubMed: 10200254]
35. Rhee SY, et al. Use and misuse of the gene ontology annotations. Nat Rev Genet. 2008; 9:509–515. [PubMed: 18475267]
36. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics. 2005; 21:3587–3595. [PubMed: 15994189]
37. King OD, et al. Predicting gene function from patterns of annotation. Genome Res. 2003; 13:896–904. [PubMed: 12695322]
38. Lindeberg M, et al. Gene Ontology annotation highlights shared and divergent pathogenic strategies of type III effector proteins deployed by the plant pathogen Pseudomonas syringae pv tomato DC3000 and animal pathogenic Escherichia coli strains. BMC Microbiology. in press.
39. Rooney JP, et al. Systems based mapping demonstrates that recovery from alkylation damage requires DNA repair, RNA processing, and translation associated networks. Genomics. 2008
40. Baba T, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006; 2:20060008.
41. Buck MJ, Lieb JD. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics. 2004; 83:349–360. [PubMed: 14986705]
42. Valouev A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 2008
43. Chevance FF, Hughes KT. Coordinating assembly of a bacterial macromolecular machine. Nat Rev Microbiol. 2008; 6:455–465. [PubMed: 18483484]
44. Karp PD. An ontology for biological function based on molecular interactions. Bioinformatics. 2000; 16:269–285. [PubMed: 10869020]
45. Stein LD. Using the Reactome database. Curr Protoc Bioinformatics. 2004; Chapter 8, unit 8 7
46. Matthews L, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2008
47. Karp PD, et al. Multidimensional annotation of the *Escherichia coli* K-12 genome. Nucleic Acids Res. 2007; 35:7577–7590. [PubMed: 17940092]
48. Typas A, et al. High-throughput, quantitative analyses of genetic interactions in E. coli. Nat Methods. 2008
49. Butland G, et al. eSGA: E. coli synthetic genetic array analysis. Nat Methods. 2008
50. Sprague J, et al. The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. Nucleic Acids Res. 2008; 36:D768–772. [PubMed: 17991680]
51. Mabee PM, et al. Phenotype ontologies: the bridge between genomics and evolution. Trends Ecol Evol. 2007; 22:345–350. [PubMed: 17416439]
52. GO Consortium. The Gene Ontology project in 2008. Nucleic Acids Res. 2008; 36:D440–444. [PubMed: 17984083]

53. Hu JC, et al. The emerging world of wikis. Science. 2008; 320:1289–1290. [PubMed: 18535227]
54. Consortium, G.O. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 2006; 34:D322–326. [PubMed: 16381878]
55. Cao TB, Saier MH Jr. The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. Biochim Biophys Acta. 2003; 1609:115–125. [PubMed: 12507766]
56. Dolezal P, et al. Evolution of the molecular machines for protein import into mitochondria. Science. 2006; 313:314–318. [PubMed: 16857931]
57. Torto-Alalibo T, et al. The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: Community development of new Gene Ontology terms describing biological processes involved in microbe-host interactions. BMC Microbiology. in press.
58. Hodis E, et al. Proteopedia - a scientific 'wiki' bridging the rift between three-dimensional structure and function of biomacromolecules. Genome Biol. 2008; 9:R121. [PubMed: 18673581]
59. Arshinoff BI, et al. Xanthusbase: adapting wikipedia principles to a model organism database. Nucleic Acids Res. 2007; 35:D422–426. [PubMed: 17090585]
60. Huss JW 3rd, et al. A gene wiki for community annotation of gene function. PLoS Biol. 2008; 6:e175. [PubMed: 18613750]
61. Mons B, et al. Calling on a million minds for community annotation in WikiProteins. Genome Biol. 2008; 9:R89. [PubMed: 18507872]
62. Waldrop M, et al. Big data: Wikiomics. Nature. 2008; 455:22–25. [PubMed: 18769412]
63. Mardis ER. ChIP-seq: welcome to the new frontier. Nat Methods. 2007; 4:613–614. [PubMed: 17664943]
64. Pellegrini M, et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999; 96:4285–4288. [PubMed: 10200254]

**Wiki-based community annotation**

Wikis, named for the Hawaiin word for “quickly”, are systems for collaborative generation of publicly accessible web pages. The best known is Wikipedia, an online encyclopedia generated by volunteers. Recently, wikis have been adopted for a variety of biological information resources [58-62]. Some distinguishing features of wikis are that:

- Registered users can edit any content, including material written by others.
- Edits appear immediately on the website without the need for approval from a gatekeeper.
- All revisions are kept, so that users can view every past version of a page and see all the revisions.

**Future directions**

Using GO to provide biological insight for biology requires further work in three major areas. Although these apply to application of GO to both prokaryotes and eukaryotes, special issues arise for both *E. coli* in particular and prokaryotes in general.

### Improving the annotation process

Currently, manually curated annotations based on experimental evidence codes are the standard for high-quality annotations. GO annotations with IEA (Inferred from Electronic Annotation) are filtered out for many applications, including displays on AmiGO (http://amigo.geneontology.org), the GO consortium browser. However, the throughput of manual curation is limited by the number of curators who are doing annotation. Combinations of two approaches are likely to help: 1) improvements in computational methods such as natural language processing and text mining can, at minimum, make manual curation more efficient by identifying candidate annotations for further human curation. 2) increasing the contribution of community curation. Whether the broader community can or will contribute significantly to high-quality GO annotation is an open question being tested by community annotation systems, including wikis (Box 1).

Improvements in annotation transfer via homology are also needed. For prokaryotes, and especially for bacteriophage genomes, transfer over larger evolutionary distances are often required compared to transfer between eukaryotic orthologs.

### Improving the ontology

The Gene Ontology is a work in progress, and is especially sparse in its coverage of the vast diversity of functions found in prokaryotes. Revision of GO tends to be driven by areas where active annotation is ongoing. Just as early microbial genome projects tended to focus on pathogens, this tendency biases the structure of GO in certain directions. This can be a problem when using the structure of the ontology itself to analyze annotations. For example, the distance terms to the ontology root is sometimes used to measure the information content of a set of annotations.

### Applications of GO and other ontologies for biological inference

Some of the current uses of GO to infer gene function are described in the text, but tools like term enrichment analysis are arguably crude examples of the kind of computational inference invoked to justify the formal structures of ontologies like GO. Future analyses are likely to incorporate annotations, the structures of GO and other ontologies, phylogenetic analyses, and systems analysis to provide deeper biological insight.

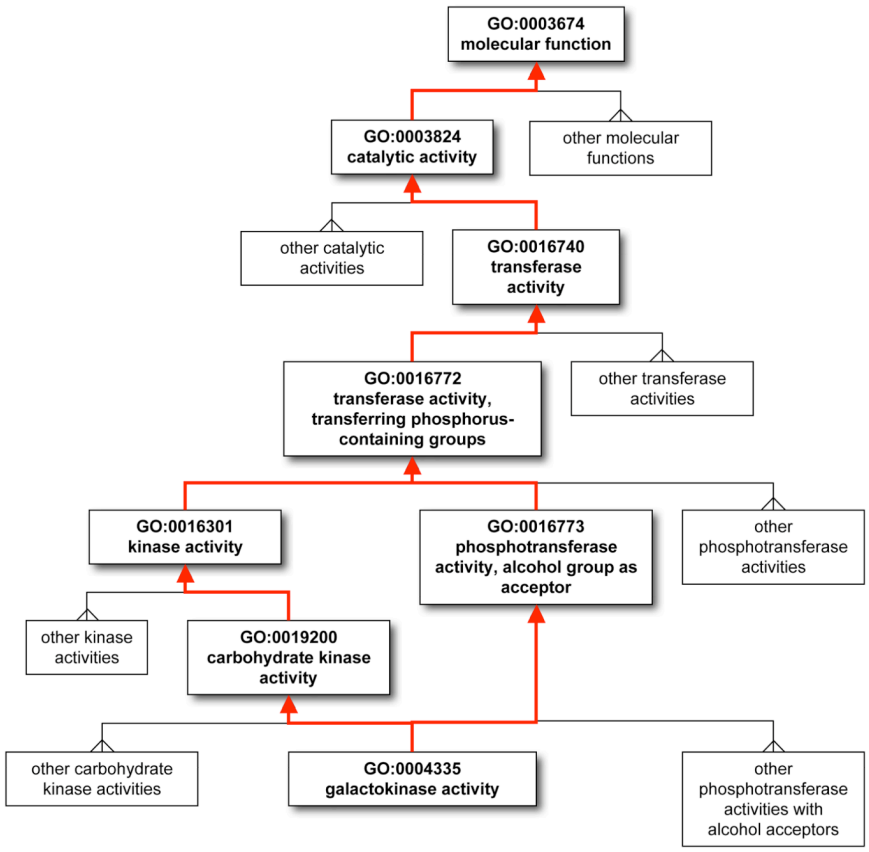
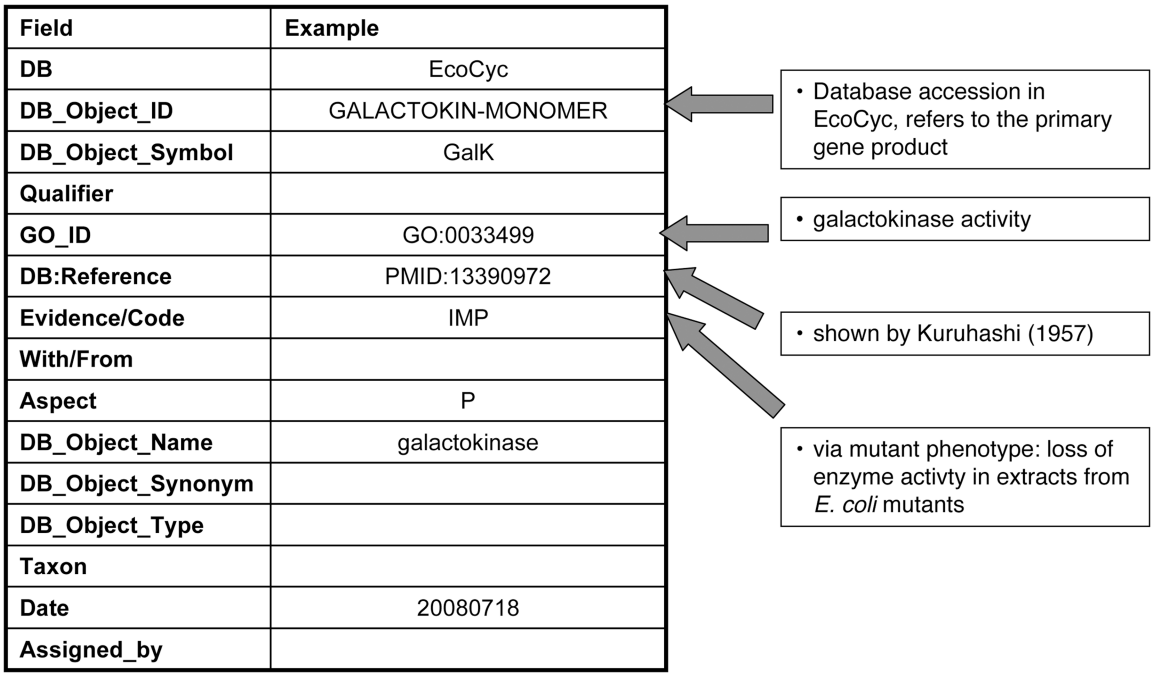



**Figure 1.**

Relationships between GO terms in a Directed Acyclic Graph (DAG). The figure illustrates a subset of the molecular function DAG for galactokinase_activity (GO:0004335). Arrows indicate relationships of the **is_a** type. The ancestors of GO:0004335 are highlighted back to the root of the molecular function ontology via arrows highlighted in red. The other boxes indicate alternative branches from each ancestor; these contain many GO terms (not shown).



| Field | Example |
|---|---|
| **DB** | EcoCyc |
| **DB_Object_ID** | GALACTOKIN-MONOMER |
| **DB_Object_Symbol** | GalK |
| **Qualifier** | |
| **GO_ID** | GO:0033499 |
| **DB:Reference** | PMID:13390972 |
| **Evidence/Code** | IMP |
| **With/From** | |
| **Aspect** | P |
| **DB_Object_Name** | galactokinase |
| **DB_Object_Synonym** | |
| **DB_Object_Type** | |
| **Taxon** | |
| **Date** | 20080718 |
| **Assigned_by** | |



**Figure 2.**

A typical *E. coli* GO annotation. The structure of an annotation for the *E. coli* galK gene product is shown. Note that the text of the database accessions from EcoCyc should not be used to infer anything about the protein.

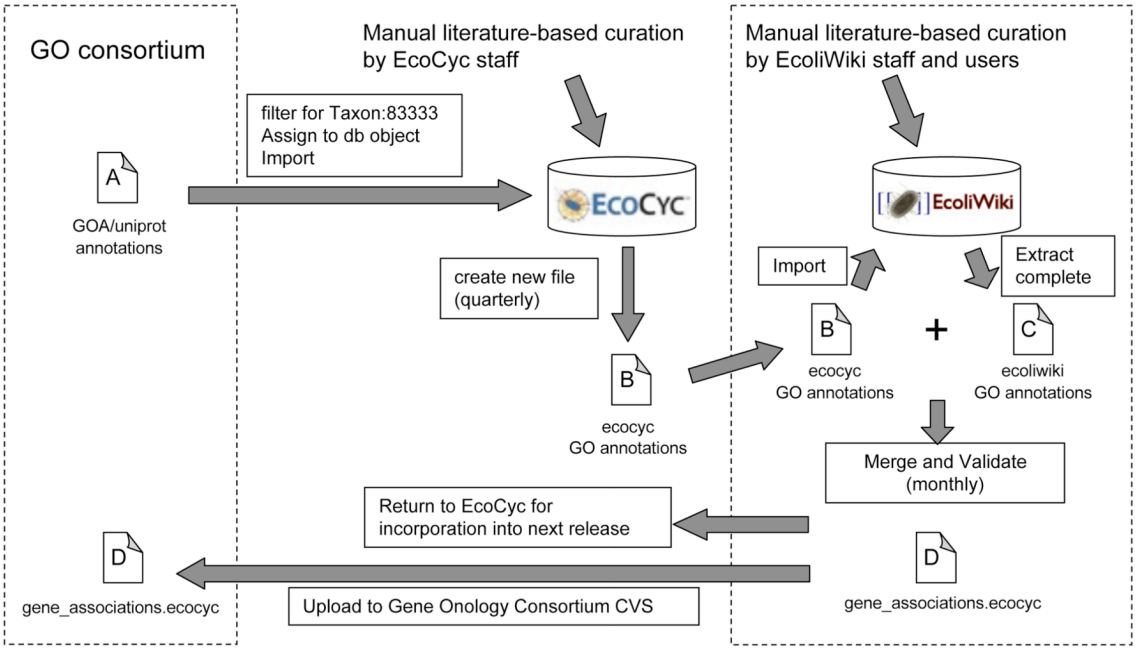



**Figure 3.**

Workflow for updating gene_association.ecocyc, the list of GO annotations maintained by EcoCyc/EcoliWiki. GO annotations from UniProt (A) are downloaded from the GO Consortium and imported to EcoCyc after filtering and processing as described in the text. A file merging these annotations with manual annotations made by EcoCyc curators (B) is sent to EcoliWiki. This is merged with manual annotations extracted from EcoliWiki (C) to generate a final gene association file (D), which is submitted back to the GO consortium.

## Table 1
### Online resources displaying GO terms for *E. coli* genes

| Database | URL | Notes |
|---|---|---|
| EcoCyc | http://ecocyc.org/ | GO terms, evidence (icons) and references, updated quarterly |
| EcoliWiki | http://ecoliwiki.net/ | GO terms, evidence codes, references and qualifiers, updated continuously and revision history viewable via wiki recent changes |
| ASAP | https://asap.ahabs.wisc.edu/asap/home.php | GO terms and evidence codes, updates given timestamps |
| PEC | http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp | GO terms and references |
| Genobase | http://www.shigen.nig.ac.jp/ecoli/strain/top/top.jsp | GO terms and references |
| GenExpDB | http://chase.ou.edu/oubcf/tools/annot.php | GO terms |
| Comprehensive Microbial Resource (CMR) | http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi | GO terms and evidence codes |
| coliBASE | http://xbase.bham.ac.uk/colibase/ | GO terms and referencesGO terms without GO ID numbers, |
| CyberCell | http://redpoll.pharmacy.ualberta.ca/CCDB/ | Blattner Ontology terms, update log available |
| MicrobesOnline | http://www.microbesonline.org/ | GO terms and non-standardized terms, updated every 6-12 months |
| RegulonDB | http://regulondb.ccg.unam.mx/ | GO terms |
| TransportDB | http://www.membranetransport.org/ | GO terms, non-standardized evidence codes and references |
| NCBI | http://www.ncbi.nlm.nih.gov/sites/gquery | GO terms without GO ID numbers |
| RCSB PDB | http://www.rcsb.org/pdb/home/home.do | GO terms |
| Pfam | http://pfam.sanger.ac.uk/ | GO terms |
| UniProt | http://www.ebi.ac.uk/uniprot/index.html | GO terms, evidence codes and references |