GenMAPP Gene Database for *Escherichia coli* K12

Ec-K12-Std_External_20060731x.gdb

ReadMe

Last revised: 3/2/07

This document contains the following:

1. Overview of the GenMAPP application and accessory programs
2. System Requirements and Compatibility
3. Installation Instructions
4. Gene Database Specifications
   a. Gene ID Systems
   b. Species
   c. Data Sources and Versions
   d. Database Report
5. Contact Information for support, bug reports, feature requests
6. Release notes
   a. Current version: Ec-K12-Std_External_20060731x.gdb
   b. Previous version: Ec-K12-Std_External_20060731.gdb
7. Database Schema Diagram

1. Overview of the GenMAPP application and accessory programs

GenMAPP (Gene Map Annotator and Pathway Profiler) is a free computer application for viewing and analyzing DNA microarray and other genomic and proteomic data on biological pathways. MAPPFinder is an accessory program that works with GenMAPP and Gene Ontology to identify global biological trends in gene expression data. The GenMAPP Gene Database (file with the extension *.gdb*) is used to relate gene IDs on MAPPs (*.mapp*, representations of pathways and other functional groupings of genes) to data in Expression Datasets (*.gex*, DNA microarray or other high-throughput data). GenMAPP is a stand-alone application that requires the Gene Database, MAPPs, and Expression Dataset files to be stored on the user's computer. GenMAPP and its accessory programs and files may be downloaded from <http://www.GenMAPP.org>. GenMAPP requires a separate Gene Database for each species. This ReadMe describes a Gene Database for *Escherichia coli* K12 that was built by the Loyola Marymount University (LMU) Bioinformatics Group using the program GenMAPP Builder, part of the open source XMLPipeDB project <http://xmlpipedb.cs.lmu.edu/>.

2. System Requirements and Compatibility

- This Gene Database is compatible with GenMAPP 2.0 and 2.1 and MAPPFinder 2.0. These programs can be downloaded from <http://www.genmapp.org>.
- System Requirements for GenMAPP 2.0/2.1 and MAPPFinder 2.0:
  Operating System: Windows 98 or higher, Windows NT 4.0 or higher (2000, XP, etc)
  Monitor Resolution: 800 X 600 screen or greater (SVGA)
  Internet Browser: Microsoft Internet Explorer 5.0 or later
  Minimum hardware configuration:
    Memory: 128 MB (512 MB or more recommended)
    Processor: Pentium III
    Disk space: 300 MB disk (more recommended if multiple databases will be used)

3. Installation Instructions

- Extract the zipped archive and place the file "Ec-K12-Std_External_20060731x.gdb" in the folder you use to store Gene Databases for GenMAPP. If you accept the default folder during the GenMAPP installation process, this folder will be C:\GenMAPP 2 Data\Gene Databases.

- To use the Gene Database, launch GenMAPP and go to the menu item *Data > Choose Gene Database*. Alternatively, you can launch MAPPFinder and go to the menu item *File > Choose Gene Database*.

4. Gene Database Specifications

a. Gene ID Systems

This *Escherichia coli* K12 Gene Database is “UniProt-centric” in that the main data source (primary ID system) for gene IDs and annotations is the UniProt complete proteome set for *Escherichia coli* K12, made available as an XML download by the Integr8 resource. In addition to UniProt IDs, this database provides the following proper gene ID systems that were cross-referenced by the UniProt data: Blattner, EchoBASE, and EcoGene. It also supplies UniProt-derived annotation links from the following systems: EMBL, InterPro, PDB, and Pfam. The Gene Ontology data has been acquired directly from the Gene Ontology Project. The GOA project was used to link Gene Ontology terms to UniProt IDs. Links to data sources are listed in the section below. This Gene Database also contains Affymetrix probe set identifiers for all Affymetrix *E. coli* microarrays (E_coli_2 Array, E. coli Genome Sense Array, E. coli Genome Antisense Array). Affymetrix identifiers were added to the Gene Database by GenMAPP.org.

b. Species

This Gene Database is based on the UniProt proteome set for *Escherichia coli* K12, taxon ID 83333. Two substrains of *E. coli* K12 have had their genome sequenced, MG1655 and W3110. This Gene Database contains data for the MG1655 strain (Blattner IDs). UniProt has a separate proteome set for the W3110 strain (taxon ID 316407). The W3110 strain uses IDs of the form “JWxxxx” and are not supported in this Gene Database. Note that although the taxon ID 83333 is the correct ID for *Escherichia coli* K12 and is listed on the Integr8 download site as such, the vast majority of entries in the UniProt proteome set actually refer to taxon ID 562, which belongs to *Escherichia coli* (with no strain designation).

c. Data Sources and Versions

- This *Escherichia coli* K12 Gene Database was built on July 31, 2006; this build date is reflected in the filename “Ec-K12-Std_External_20060731x.gdb”. All date fields internal to the Gene Database (and not usually seen by regular GenMAPP users) have been filled with the build date, except for the Affy table which shows the date 2/15/07. This version of the Gene Database containing Affymetrix probe set identifiers is distinguished from the previous version by the “x” in the filename. Versioning information for the individual data sources is given below.
- UniProt complete proteome set for *Escherichia coli* K12, made available as an XML download by the Integr8 resource:
  <http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeId=18>
  Filename: “18.E_coli_K12.xml” (downloaded as a compressed .gz file and extracted)
  Version information for the proteome sets can be found at
  <http://www.ebi.ac.uk/integr8/HelpAction.do?action=searchById&refId=5>
  The proteome set used for this version of the *Escherichia coli* K12 Gene Database was based on UniProt Knowledgebase release 8.3 on July 11, 2006.
- Gene Ontology gene associations are provided by the GOA project:
  <http://www.ebi.ac.uk/GOA/> as a tab-delimited text file. The *Escherichia coli* K12 GOA file was accessed from the Integr8 proteome set download page:
  <http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeId=18>
  Filename: “18.E_coli_K12.goa” (downloaded as a compressed .gz file and extracted)
  The GOA file for this version of the *Escherichia coli* K12 Gene Database was based on the July 2006 release of GOA (version information can be found as a date field within the GOA file itself).

- Gene Ontology data is downloaded from <http://www.godatabase.org/dev/database/> We use the monthly release available at the first of every month. For this version of the *Escherichia coli* K12 Gene Database we used the July 1, 2006 release. Filename: "go_200607-termdb.obo-xml.gz" (we extract the file and reverse the period and hyphen in the filename so it reads "go_200607-termdb-obo.xml".
- Affymetrix probe set identifiers and associations to UniProt and Blattner were collected from <http://www.affymetrix.com>, in the form of annotation files (.csv). Files were downloaded on February 14, 2007 and added to the Gene Database on February 15, 2007 by GenMAPP.org. Blattner associations were extracted from the "Transcript ID(Array Design)" field and UniProt associations were extracted from the "SwissProt" field.

d. Database Report

- UniProt is the primary ID system for the *Escherichia coli* K12 Gene Database. The UniProt table contains all 4329 UniProt IDs reported in the UniProt proteome set for this species.
- The Blattner IDs were derived from the cross-references in the UniProt XML for *Escherichia coli* K12. We compared our Blattner table with the table in the supplementary material from Riley *et al*. (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Research* 34: 1-9, "Supplementary_Table_1_Annotation_E._coli_Genes.xls". Our Blattner table contains 4466 identifiers. There are 219 Blattner IDs reported in the Riley et al. table that are not in our Gene Database. Of these:
  157 are RNA genes (tRNA, rRNA, or misc_RNA)
  1 is the origin of replication
  51 are protein coding sequences (CDS)
  10 do not have a feature designation
  Conversely, there are 200 Blattner IDs in our table that are not in the Riley table:
  104 are in the form of "ECOK12Fxxx" and correspond to proteins encoded by plasmid F
  51 contain a period in the ID, such as "bxxxx.y"
  45 are IDs that have been supplanted by a new ID, but are reported in UniProt as synonyms to other Blattner IDs.
- The *Escherichia coli* K12 Gene Database also contains 4156 EchoBASE IDs and 4224 EcoGene IDs that were cross-referenced by the UniProt XML.
- The *Escherichia coli* K12 Gene Database also contains 24692 Affymetrix probe set IDs.

5. Contact Information for support, bug reports, feature requests

- The Gene Database for *Escherichia coli* K12 was built by the Loyola Marymount University (LMU) Bioinformatics Group using the program GenMAPP Builder, part of the open source XMLPipeDB project <http://www.cs.lmu.edu/~xmlpipedb>.
- For support, bug reports, or feature requests relating to XMLPipeDB or GenMAPP Builder, please consult the XMLPipeDB Manual found at <http://www.cs.lmu.edu/~xmlpipedb/documentation.shtml> or go to our SourceForge site <http://www.cs.lmu.edu/~xmlpipedb>.
- For issues related to the *Escherichia coli* K12 Gene Database, please contact:

  Kam D. Dahlquist, Ph.D.
  Department of Biology
  Loyola Marymount University
  1 LMU Drive, MS 8220
  Los Angeles, CA 90045-2659
  kdahlquist@lmu.edu

- For issues related to GenMAPP 2.0/2.1 or MAPPFinder 2.0, please contact GenMAPP support directly by e-mailing genmapp@gladstone.ucsf.edu.

6. Release Notes

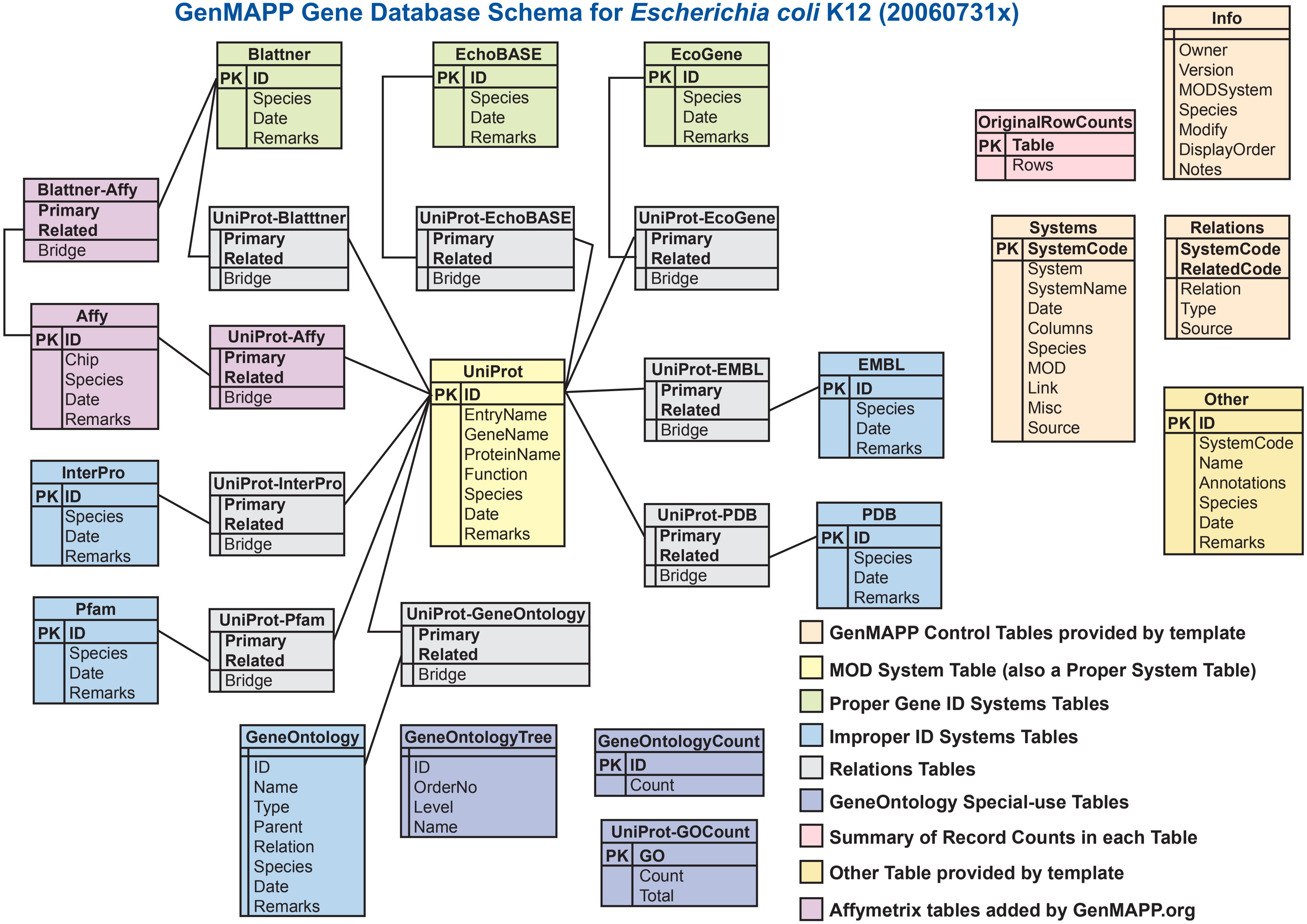
a. Current version: Ec-K12-Std_External_20060731x.gdb

- This release is a modification of the first release of a standard *Escherichia coli* K12 Gene Database. The data in the Gene Database remains the same with the following additions.
- This Gene Database also contains Affymetrix probe set identifiers for all Affymetrix *E. coli* microarrays (E_coli_2 Array, E. coli Genome Sense Array, E. coli Genome Antisense Array). Affymetrix identifiers were added to the Gene Database by GenMAPP.org. These Affymetrix probe set identifiers were related to both UniProt and Blattner identifiers.
- This ReadMe document was updated on March 2, 2007.

b. Previous version: Ec-K12-Std_External_20060731.gdb

- This was the first release of a standard *Escherichia coli* K12 Gene Database.
- Unlike the official Gene Databases from GenMAPP.org, in the *Escherichia coli* K12 Gene Database, PDB has been designated as an “improper” gene ID system and cannot be used as an ID for gene objects on MAPPs. This action was taken because PDB IDs can refer to structures containing two or more different polypeptides and thus do not refer to a unique protein molecule.
- This ReadMe document was updated on September 24, 2006. The Gene Database itself has not changed. However, changes were made to this document to reflect the following:
  - The filename of the database has been changed from “Ec-Std_20060731.gdb” to “Ec-K12-Std_External_20060731.gdb” to be consistent with the name of the database that can be downloaded via GenMAPP.org. “K12” was added to reduce ambiguity between *Escherichia coli* K12 and other strains of *E. coli*. “External” was added to indicate that the Gene Database was created by a group external to GenMAPP.org.
  - The URL for the XMLPipeDB Project has changed from <http://www.cs.lmu.edu/~xmlpipedb/> to <http://xmlpipedb.cs.lmu.edu/>.
  - A clarification to the taxon IDs used in the UniProt source data was made.

GenMAPP Gene Database Schema for *Escherichia coli* K12 (20060731x)



NOTE: Some Relations tables are not shown. All possible pairwise Relations tables exist between Proper ID systems and between Proper and Improper ID systems, but not between Improper ID systems (i.e., Proper-Proper, Proper-Improper, but NOT Improper-Improper).