GenMAPP Gene Database for *Mycobacterium smegmatis* (strain ATCC 700084 / mc(2)155)
Ms-Std_External_20110128.gdb
ReadMe

Last revised: 3/3/11

This document contains the following:

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 *Mycobacterium smegmatis* (strain ATCC 700084 / mc(2)155) that was built by the Loyola Marymount University (LMU) Bioinformatics Group using the program GenMAPP Builder 2.0, 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 "Ms-Std_External_20110128.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 *Mycobacterium smegmatis* Gene Database is UniProt-centric in that the main data source (primary ID System) for gene IDs and annotation is the UniProt complete proteome set for *Mycobacterium smegmatis*, 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: OrderedLocusNames, GeneId (NCBI), and RefSeq (protein IDs of the form YP_######). 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.

Proper ID System	SystemCode
UniProt	S
OrderedLocusNames	N
GeneId	L
RefSeq	Q

b. Species

This Gene Database is based on the UniProt proteome set for *Mycobacterium smegmatis* (strain ATCC 700084 / mc(2)155), taxon ID 246196.

c. Data Sources and Versions

- This *Mycobacterium smegmatis* Gene Database was built on January 28, 2011; this build date reflected in the filename Ms-Std_External_20110128.gdb. All date fields internal to the Gene Database (and not usually seen by regular GenMAPP users) have been filled with this build date.
- UniProt complete proteome set for *Mycobacterium smegmatis* (strain ATCC 700084 / mc(2)155), made available as an XML download by the Integr8 resource:
  <http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeId=25827>
  Filename: "25827.M_smegmatis.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 *Mycobacterium smegmatis* Gene Database was based on release 115 of Integr8 which was built from UniProt release 2010_11 and InterPro release 28.0 and was released on November 3, 2010.
- Gene Ontology gene associations are provided by the GOA project:
  <http://www.ebi.ac.uk/GOA/> as a tab-delimited text file. The *Mycobacterium smegmatis* GOA file was accessed from the Integr8 proteome set download page:
  <http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeId=25827>
  Filename: "25827.M_smegmatis.goa". The GOA file for this version of the *Mycobacterium smegmatis* Gene Database was based on GOA UniProt 89, released on November 17, 2010.
- Gene Ontology data is downloaded from
  <http://www.geneontology.org/GO.downloads.ontology.shtml>
  Data is released daily. For this version of the *Mycobacterium smegmatis* Gene Database we used the ontology version 1.1622 released on November 26, 2010.
  Filename: "go_daily-termdb.obo-xml.gz".

d. Database Report

- UniProt is the primary ID system for the *Mycobacterium smegmatis* Gene Database. The UniProt table contains all 6601 UniProt IDs contained in the UniProt proteome set for this species.
- The OrderedLocusNames ID system was derived from the cross-references in the UniProt proteome set. Each ID takes the form of MSMEG_####. We compared this table with the list of gene IDs in the JCVI Comprehensive Microbial Resource (CMR) at <http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gms>. There are 6880 protein-coding genes listed there. Of the protein-coding genes, 166 gene IDs do not appear in our Gene Database because they are not cross-listed in the UniProt XML. Four IDs appear in our Gene Database, but not in the CMR.
- The following table lists the numbers of gene IDs found in each gene ID system:

ID System	ID Count
EMBL	34
GeneId (NCBI)	6716
InterPro	3580
OrderedLocusNames	6718
PDB	87
Pfam	1761
RefSeq	6716
UniProt	6601

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

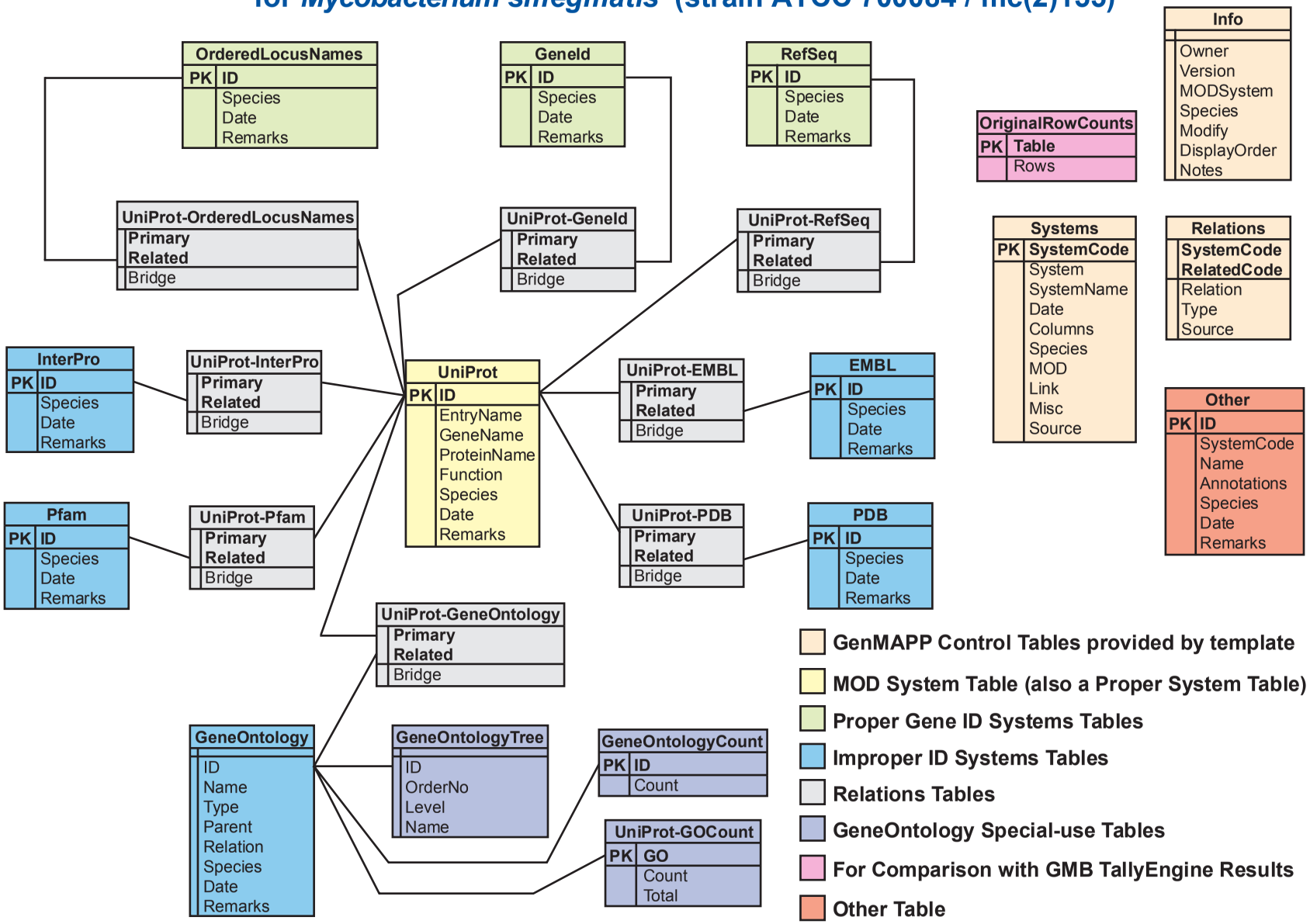
- The Gene Database for *Mycobacterium smegmatis* was built by the Loyola Marymount University (LMU) Bioinformatics Group using the program GenMAPP Builder, part of the open source XMLPipeDB project <http://xmlpipedb.cs.lmu.edu/>.
- For support, bug reports, or feature requests relating to XMLPipeDB or GenMAPP Builder, please consult the XMLPipeDB Manual found at <http://xmlpipedb.cs.lmu.edu/documentation.shtml> or go to our SourceForge site <http://sourceforge.net/projects/xmlpipedb/>.
  E-mail: xmlpipedb-developer@lists.sourceforge.net
- For issues related to the *Mycobacterium smegmatis* Gene Database, please contact:
  Kam D. Dahlquist, PhD.
  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 or GenMAPP@googlegroups.com.

6. Release Notes

a. First release: Ms-Std_External_20110128.gdb

- Richard Brous, Margie Doyle, Andrew Herman, John David N. Dionisio, and Kam D. Dahlquist contributed to the first release. This Gene Database was generated as a final project in the BIOL/CMSI 367/HNRS 398-05: Biological Databases course at Loyola Marymount University in Fall 2010.

GenMAPP Gene Database Schema
for *Mycobacterium smegmatis* (strain ATCC 700084 / mc(2)155)



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).