GenMAPP Gene Database for *Mycobacterium tuberculosis* str. H37Rv
Mt-Std_External_20101020.gdb
ReadMe

Last revised: 10/22/10

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 tuberculosis* str. H37Rv 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 "Mt-Std_External_20101020.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 tuberculosis* str. H37Rv 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 tuberculosis* str. H37Rv, 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 NP_###### and 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 tuberculosis* str. H37Rv (ATCC 25618), taxon ID 83332.

c. Data Sources and Versions

- This *Mycobacterium tuberculosis* str. H37Rv Gene Database was built on October 20, 2010; this build date reflected in the filename Mt-Std_External_20101020.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 tuberculosis* str. H37Rv, made available as an XML download by the Integr8 resource:
  <http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeId=30>
  Filename“30.M_tuberculosis_ATCC_25618.xml.gz” (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 tuberculosis* str. H37Rv Gene Database was based on the release 114 of Integr8 which was built from UniProt release 2010_10 and InterPro release 28.0 and was released on October 7, 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 tuberculosis* str. H37Rv GOA file was accessed from the Integr8 proteome set download page:
  <http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeId=30>
  Filename“30.M_tuberculosis_ATCC_25618.goa”. The GOA file for this version of the *Mycobacterium tuberculosis* str. H37Rv Gene Database was based on the GOA Proteome Sets 87.0 released on September 21, 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 tuberculosis* str. H37Rv Gene

Database we used the September 15, 2010 release.
Filename: “go_daily-termdb.obo-xml.gz”.

d. Database Report

- UniProt is the primary ID system for the *Mycobacterium tuberculosis* str. H37Rv Gene Database. The UniProt table contains all 3949 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 Rv####, Rv####.#, or Rv####c. We compared this table with the ID count in EnsemblBacteria, Genebuild May 2010 and Database version: 59.3a at <http://bacteria.ensembl.org/Mycobacterium/m_tuberculosis_h37rv/Info/StatsTable?db=core>. There are 3987 protein coding genes listed there, which is four less than the 3991 OrderedLocusNames IDs in the GenMAPP Gene Database.
- The following table lists the numbers of gene IDs found in each gene ID system:

ID System	ID Count
EMBL	218
GeneId (NCBI)	6912
InterPro	3156
OrderedLocusNames	3991
PDB	857
Pfam	1556
RefSeq	6912
UniProt	3949

Note that there are nearly twice as many GeneId and RefSeq IDs as there are UniProt IDs because the UniProt records cross reference both the *Mycobacterium tuberculosis* str. H37Rv and *Mycobacterium tuberculosis* str. CDC1551 (taxon ID 83331).

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

- The Gene Database for *Mycobacterium tuberculosis* str. H37Rv 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/>.
- For issues related to the *Mycobacterium tuberculosis* str. H37Rv Gene Database, please contact:

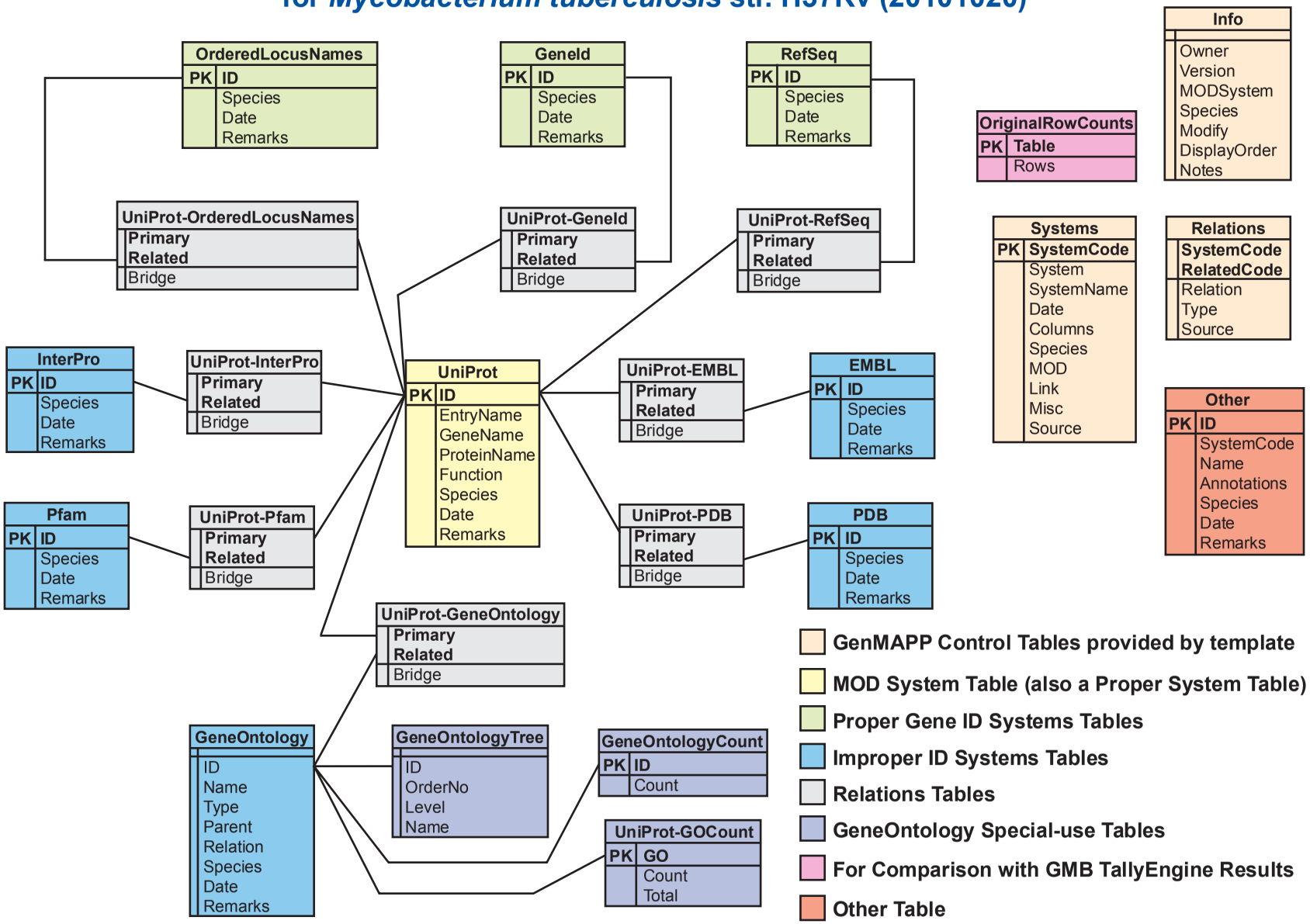
  Kam D. Dahlquist, PhD.
  Department of Biology
  Loyola Marymount University
  1 LMU Drive, MS 8220
  Los Angeles, CA 90045-2659
  kdahquist@lmu.edu
- For issues related to GenMAPP 2.0/2.1 or MAPPFinder 2.0 please contact GenMAPP support directly by e-mailing genmapp@gladstone.ucsf.edu or GenMAPP@googlegroups.com.

6. Release Notes

a. Current version: Mt-Std_External_20101020.gdb

- Kevin Paiz-Ramirez (lead), John David N. Dionisio, and Kam D. Dahlquist contributed to this release.

GenMAPP Gene Database Schema
for *Mycobacterium tuberculosis* str. H37Rv (20101020)



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).