GenMAPP Gene Database for *Streptococcus pneumoniae* str. TIGR4

Streptococcus_pneumoniae_TIGR4_20131125.gdb

ReadMe

Last revised: 03 Dec 2013

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 *Streptococcus pneumoniae* str. TIGR4 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 "Streptococcus_pneumoniae_TIGR4_20131125.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 *Streptococcus pneumoniae* 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 *Streptococcus pneumoniae*, made available as an XML download. 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_######). 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 (NCBI)	L
RefSeq	Q

b. Species

This Gene Database is based on the UniProt proteome set for *Streptococcus pneumoniae* serotype 4 (strain ATCC BAA-334 / TIGR4), taxon ID 170187.

c. Data Sources and Versions

- This *Streptococcus pneumoniae* Gene Database was built on 25 November 2013; this build date is reflected in the filename Streptococcus_pneumoniae_TIGR4_20131125.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 *Streptococcus pneumoniae* serotype 4 (strain ATCC BAA-334 / TIGR4), downloaded from this page: <http://www.uniprot.org/uniprot/?query=organism%3a170187+keyword%3a181&format=*>
  Filename: "uniprot-organism%3A170187+keyword%3A181.xml"
  Version information for the proteome sets can be found at <http://www.uniprot.org/news/>
  The proteome set used for this version of the *S. pneumoniae* Gene Database was based on UniProt release 2013_11 released on November 13, 2013.
- Gene Ontology gene associations are provided by the GOA project: <http://www.ebi.ac.uk/GOA/> as a tab-delimited text file. The *S. pneumonia*e str. TIGR4 GOA file was accessed from the GOA proteomes FTP site: < ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/>
  Filename: "57.S_pneumoniae_TIGR4.goa". Version 11/12/2013 2:49:00 PM.
- Gene Ontology data is downloaded from <http://beta.geneontology.org/page/download-ontology>
  Data is released daily. For this version of the *Streptococcus pneumoniae* Gene Database we used the ontology version 2013-11-20.
  Filename: "go_daily-termdb.obo-xml.gz".

d. Database Report

- UniProt is the primary ID system for the *Streptococcus pneumoniae* Gene Database. The UniProt table contains all 2126 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 appears twice, once in the form of SP_#### and once in the form of SP####, (e.g., SP0001 and SP_0001) because IDs of both forms can be found in the literature. We compared this table with the list of gene IDs in the EnsemblBacteria Gene Annotation at http://bacteria.ensembl.org/streptococcus_pneumoniae_tigr4/Info/Annotation/#genebuild. There are 2125 protein coding genes listed and 58 non-coding genes. Of the protein coding genes, 58 gene IDs do not appear in our Gene Database because they are not cross-listed in the UniProt XML file. Additionally, one gene in the UniProt XML file did not appear in the Ensembl database.
- The following table lists the numbers of gene IDs found in each gene ID system:

ID System	ID Count Current version
EMBL	201
GeneOntology	3648
InterPro	2641
OrderedLocusNames	4252*
PDB	225
Pfam	1277
RefSeq	2105
UniProt	2109

*There are 2126 unique genes/proteins in the current version of the Gene Database; the 4252 count represents the total number of IDs due to duplicate IDs of the form SP#### and SP_####.

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

- The Gene Database for *Streptococcus pneumoniae* 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 *Streptococcus pneumonaie* 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. Current version: Streptococcus_pneumoniae_TIGR4_20131125.gdb

- Tauras Vilgalys, Alina Vreeland, Kevin Meilak, Kam D. Dahlquist, and John David N. Dionisio contributed to this release.