GenMAPP Gene Database for *Shigella flexneri* 2a str. 301
Sf-Std_External_20151214.gdb
ReadMe

Last revised: 12/16/15

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 *Shigella flexneri* 2a str. 301 that was built by the Loyola Marymount University (LMU) Bioinformatics Group using the program GenMAPP Builder 3.0, 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 "Sf-Std_External_20151214.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 *Shigella flexneri* 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 *Shigella flexneri*, 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_###### and WP_#########). 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 *Shigella flexneri* 2a str. 301 (ATCC 29903), taxon ID 623.

c. Data Sources and Versions

- This *Shigella flexneri* Gene Database was built on December 14, 2015; this build date is reflected in the filename Sf-Std_External_20151214.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 *Shigella flexneri* 2a str. 301, downloaded from this page: <http://www.uniprot.org/uniprot/?query=proteome%3Aup000001006&sort=score>
  Filename: "uniprot-proteome%3AUP000001006.xml" (downloaded as a compressed .gz file and extracted)
  Version information for the proteome sets can be found at <http://www.uniprot.org/news/>
  The proteome set used for this version of the *Shigella flexneri* Gene Database was based on UniProt release 2015_11 released on November 11, 2015.
- Gene Ontology gene associations are provided by the GOA project: <http://www.ebi.ac.uk/GOA/> as a tab-delimited text file. The *Shigella flexneri* GOA file was accessed from the GOA proteomes FTP site: < ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/>
  Filename: "103.S_flexneri_301.goa". Version 11/19/2015 3:24:00 PM.
- Gene Ontology data is downloaded from <http://beta.geneontology.org/page/download-ontology>
  Data is released daily. For this version of the *Shigella flexneri* Gene Database we used the ontology version 2015-11-19 2:24:57 AM.
  Filename: "go_daily-termdb.obo-xml.gz".

d. Database Report

- UniProt is the primary ID system for the *Shigella flexneri* Gene Database. The UniProt table contains all 4103 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. The IDs are of the form SF####, S####, and CP####. Each ID appears once, however, the S#### IDs represent the same genes as SF#### which is why the count for the

OrderedLocusNames is almost twice as much as the Uniprot gene IDs in the table shown below. The CP#### format represents the plasmid IDs.

- Moreover, since we had to add ~92 IDs of the form SF#### from elsewhere in the XML file into our OrderedLocusNames tally, the rows in the TallyEngine results window has been duplicated. The OrderedLocusNames count is also twice as much (7567 * 2 = 15134) for the XML file, which is incorrect since we did not get a chance to only add the new counts to what we originally had, 7567 ordered locus names.
- Additionally, the PostgreSQL count for OrderedLocusNames is only 7660, compared to what we were expecting, 7661, from our past queries. The reason for the mismatch is still unknown at this point.
- According to the Ensembl Bacteria website, < http://bacteria.ensembl.org/Shigella_flexneri_2a_str_301/Info/Annotation/#assembly>, there is a total of 4446 coding genes, 379 non-coding genes, 259 pseudogenes, and 5086 gene transcripts for *Shigella flexneri* 2a str. 301. Of the protein coding genes, the 237 IDs do not appear in our Gene Database since they are not cross-listed in the UniProt XML file. ~100 plasmid IDs were also non-existent in the XML file.
- The following table lists the numbers of gene IDs found in each gene ID system:

ID System	ID Count Current version
EMBL	121
GeneID (NCBI)	3404
GeneOntology	6478
InterPro	5124
OrderedLocusNames	7661*
PDB	95
Pfam	2357
RefSeq	7501
UniProt	4103

* The 7661 count represents the total number of IDs of both the form SF#### and S####, which represent the same genes; when only the genes are counted and not the ordered locus IDs, the count would only be ~4209.

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

- The Gene Database for *Shigella flexneri* 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 *Shigella flexneri* 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: Sf-Std_External_20151214.gdb

- Trixie Roque, Jake Woodlee, Erich Yanoschik, Kristin Zebrowski, Kam D. Dahlquist and John David N. Dionisio contributed to this release.