

ReadMe for GenMAPP Gene Database for *Vibrio cholerae* O1 biovar El Tor str. N16961, Vc-Std_External_20161009.gdb

Last revised: 12/21/16

This document contains the following:

1. Overview of GenMAPP application and accessory programs
2. System Requirements and Compatibility
3. Installation Instructions
4. Gene Database Specifications
 1. Gene ID Systems
 2. Species
 3. Data Sources and Versions
 4. Database Report
5. Contact Information for support, bug reports, feature requests
6. Release notes
 1. Current version: Vc-Std_External_20161009.gdb
 2. Previous version: Vc-Std_External_20101022.gdb
 3. Previous version: Vc-Std_External_20090622.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 *Vibrio cholerae* O1 biovar El Tor str. N16961 that was built by the Loyola Marymount University (LMU) Bioinformatics Group using the program GenMAPP Builder, part of the open source XMLPipeDB project <https://lmu-bioinformatics.github.io/xmlpipedb/>.

2. System Requirements and Compatibility: - This Gene Database is compatible with GenMAPP 2.0 and 2.1 ("GenMAPP Classic") and MAPPFinder 2.0. These programs can be downloaded from <https://github.com/GenMAPPCS/genmapp>. - System Requirements for GenMAPP 2.1 and MAPPFinder 2.0: - Operating System: Windows 98 or higher, Windows NT 4.0 or higher (2000, XP, etc), Windows 7 or higher - 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 "Vc-Std_External_20161009.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

i. Gene ID Systems

This *Vibrio cholerae* 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 *Vibrio cholerae*, 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, EnsemblBacteria, 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 |
| EnsemblBacteria | Eb |
| GeneID (NCBI) | L |
| RefSeq | Q |

ii. Species

- This Gene Database is based on the UniProt proteome set for *Vibrio cholerae* O1 biovar El Tor str. N16961 (ATCC 39315), taxon ID 243277.

iii. Data Sources and Versions

- This Gene Database was built on October 9, 2016; this build date is reflected in the filename Vc-Std_External_20161009.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 *Vibrio cholerae* O1 biovar El Tor str. N16961, downloaded from this page:
http://www.uniprot.org/uniprot/?query=organism%3A243277+keyword%3A1185&format=*&compress=yes Filename: "uniprot-organism%3A243277+keyword%3A1185.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 *Vibrio cholerae* Gene Database was based on UniProt release 2016_09 released on October 5, 2016.

- Gene Ontology gene associations are provided by the GOA project: <http://www.ebi.ac.uk/GOA/> as a tab-delimited text file. The *Vibrio cholerae* GOA file was accessed from the GOA proteomes FTP site: <ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/> Filename: "46.V_cholerae_ATCC_39315.goa". Version 10/4/16, 7:16:00 AM.
- Gene Ontology data is downloaded from <http://beta.geneontology.org/page/download-ontology> Data is released daily. For this version of the *Vibrio cholerae* Gene Database we used the ontology version 10/09/2016, 3:41 pm. Filename: "go_daily-termdb.obo-xml.gz".

iv. Database Report

- UniProt is the primary ID system for the *Vibrio cholerae* Gene Database. The UniProt table contains all 3784 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 VC_#### and once in the form of VC#####, (e.g., VC0001 and VC_0001) because IDs of both forms can be found in the literature.
- The following table lists the numbers of gene IDs found in each gene ID system:

| ID System | ID Count Current version | ID Count 20101022 version | ID Count 20090622 version |
|-------------------|--------------------------|---------------------------|---------------------------|
| EMBL | 176 | 293 | 199 |
| EnsemblBacteria | 3827 | N/A | N/A |
| GeneID (NCBI) | 3339 | 3827 | 3441 |
| GeneOntology | 6920 | 3829 | 3528 |
| InterPro | 4554 | 3942 | 2596 |
| OrderedLocusNames | 7664* | 7664* | 6890 |
| PDB | 312 | 157 | 67 |
| Pfam | 2200 | 1955 | 1449 |
| RefSeq | 6556 | 3827 | 3441 |
| UniProt | 3788 | 3784 | 3397 |

*There are 3832 unique genes/proteins in the current version of the Gene Database; the 7664 count represents the total number of IDs due to duplicate IDs of the form VC##### and VC_####.

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

- The Gene Database for *Vibrio cholerae* was built by the Loyola Marymount University (LMU) Bioinformatics Group using the program GenMAPP Builder, part of the open source XMLPipeDB project <https://lmu-bioinformatics.github.io/xmlpipedb/>.
- 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 Github repository at <https://github.com/lmu-bioinformatics/xmlpipedb>.
- For issues related to the *Vibrio cholerae* Gene Database, please contact: Kam D. Dahlquist, PhD., Department of Biology, Loyola Marymount University, 1 LMU Drive, MS 8888, Los Angeles, CA 90045-2659, kdahlquist at lmu dot 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

i. Current version: Vc-Std_External_20161009.gdb - Nicole Anguiano, Kam D. Dahlquist and John David N. Dionisio contributed to this release.

ii. Previous version: Vc-Std_External_20101022.gdb - Kam D. Dahlquist and John David N. Dionisio contributed to this release.

iii. Previous version: Vc-Std_External_20090622.gdb - Alexandra Alphonso, Derek Smith, John David N. Dionisio, and Kam D. Dahlquist contributed to the first release.

7. Database Schema Diagram



GenMAPP Gene Database Schema for *Vibrio cholerae* O1 biovar El Tor str. N16961 (20161009)

