

# **USER GUIDE FOR DENOVO GENOMICS PIPELINE**

## **Contents**

(1) Introduction

(2) Installation

    2.1: Directory structure of DeNoGAP package.

    2.2: System requirements.

    2.3: List of required Perl modules and external programs for DeNoGAP.

    2.4: Installation steps for DeNoGAP.

    2.5: Installation of R and required R-libraries.

    2.6: Installation and configuration of Apache webserver.

(3) Input Data Files

    3.1: Genome information file

    3.2: Configuration files

(4) SQLite database structure

(5) Running DeNoGAP pipeline

(6) Output directory structure.

(7) Exploring output using DeNoGAP graphical user interface.

## **1. Introduction**

Denovo Genomic Analysis Pipeline (DeNoGAP) is a software package for comparative analysis of multiple completed or draft genomes of prokaryotic species. The pipeline incorporates number of tools and databases for gene prediction, homolog prediction, ortholog prediction, functional annotation, phylogenetic profiling and core genome prediction.

## **2. Installation**

The package for DeNoGAP is available at <http://sourceforge.net/projects/denogap>.

### **2.1 Directory structure of DeNoGAP package:**

**bin:** contain main-pipeline execution script, installation script, and other analysis scripts.

**config:** contain configuration files for defining parameters for different analysis phases.

**lib:** contains DeNoGAP-specific Perl modules required for the analysis.

**exe:** directory to install external program by default.

**data:** directory to store input data files.

**doc:** contains manual for installing and using DeNoGAP.

**output:** directory to store output data / results.

**GUI:** contains cgi script and html file for graphical interface.

- **cgi-bin:** contains cgi scripts for connecting and accessing database using GUI.
- **html:** contains html page for graphical user interface to connect with the database.

### **2.2 SYSTEM REQUIREMENT**

**Operating system: Linux**

**RAM: minimum 2GB**

The package provides a script (install.pl) to install necessary programs and Perl modules for performing analysis with DeNoGAP on Linux platform. Currently DeNoGAP has been only tested on Ubuntu Linux system. Although not tested for other operating systems, we assume that DeNoGAP can also run under other Unix based OS after installation of necessary programs and prerequisites. Windows OS user can use ‘cygwin’ to run DeNoGAP.

## **2.3 List of required Perl modules / programs:**

### **Perl modules:**

FindBin, Env, Exporter, Getopt::Long, File::Basename, File::Copy, Tie::File  
Parallel::ForkManager, List::MoreUtils, List::Util, File::Path, Hash::Merge, DBI, CGI, English,  
File::Spec::Functions, FileHandle, IO::Scalar, IO::String, Mail::Send, Sys::Hostname,  
URI::Escape, XML::Parser, XML::Quote, Statistics::Basic, Statistics::R

### **BioPerl modules:**

Bio::Perl, Bio::SeqIO, Bio::Seq, Bio::SearchIO, Bio::Tools::Phylo::Phylip::ProtDist,  
Bio::AlignIO

### **R module:**

igraph

### **Programs:**

Muscle v3.8.31 or above (<http://www.drive5.com/muscle>)

Kalign2 (<http://msa.sbc.su.se/downloads/kalign>)

MCL (<http://micans.org/mcl>)

Hmmer version 3 or above (<http://selab.janelia.org/software/hmmer3>)

Phylip v3.6 or above (<http://evolution.gs.washington.edu/phylip>)

Glimmer (<http://ccb.jhu.edu/software/glimmer>)

Prodigal (<http://prodigal.googlecode.com>)

FragScan (<http://omics.informatics.indiana.edu/mg/get.php?software=FragGeneScan1.16.tar.gz>)

GeneMark (<http://opal.biology.gatech.edu>)

InterProScan5 (<https://code.google.com/p/interproscan>)

EMBOSS (<http://emboss.sourceforge.net>)

SQLite (<https://sqlite.org>)

R (<http://cran.r-project.org/mirrors.html>)

Apache (<http://www.apache.org>)

**Note:** DeNoGAP provides Perl script to install necessary program. However, users can skip this step if all required program are pre-installed on their system with PATH environmental variable set for each program to be accessed by DeNoGAP.

In some cases, DeNoGAP may fail to set the PATH environmental variable for the program due to restricted permissions. Under such circumstances, user may want to manually set the path for each program on their system.

Users can set path for each program by using putting following line in the “.bashrc” script.

[MUSCLE]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/muscle
```

[HMMER]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/ hmmer/bin
```

[MCL]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/ mcl/bin
```

[PHYLIP]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/ phylib/ phylib-3.695/exe
```

[GLIMMER]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/ glimmer/glimmer3.02/bin
```

```
export PATH=$PATH:~userpath/DeNoGAP/exe/ glimmer/glimmer3.02/scripts
```

[PRODIGAL]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/prodigal
```

[FragGeneScan]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/ frag_gene_scan/FragGeneScan1.16
```

[GeneMark]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/ genemark/genemark_suite_linux_64/gmsuite
```

[BLAST]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/ blast.ncbi-blast-2.2.29+/bin
```

[InterProScan]

```
export PATH=$PATH:~userpath/DeNoGAP/exe/ interproscan/Interproscan-5.7-48.0
```

## **2.4 STEP FOR INSTALLATION OF DeNoGAP:**

To install external programs using DeNoGAP install script, execute following command in the terminal:

```
cd DeNoGAP
```

```
cd bin
```

```
perl install.pl
```

Note: By default installation of most of the external programs will take place under “exe” directory of DeNoGAP package. However, some programs are installed under root directory by default, which requires administrative permissions from user for installation.

External Program installation script in DeNoGAP uses ftp URLs to download and install most up-to-date programs available till date. If users want to download any alternate version of any program, they can change the ftp URL for respective tool in the install script. Alternatively, they can manually install new version of the program.

## **2.5 Installation of R and its libraries**

As there are many different version of R available on different mirror sites, DeNoGAP do not automatically install R on user system. However user on Ubuntu Linux can use following steps to install R and other necessary libraries.

Method 1:

```
sudo apt-get install r-base
```

This command will install basic R libraries on Ubuntu.

Method 2:

Download R version best suited for your system from <http://cran.r-project.org/mirrors.html> and follow instruction in source for installation.

Once installed, open command line terminal and perform following steps to install required libraries.

- Type R on command line to open R command prompt.

- install.packages("igraph")
- q()

## 2.6 Installation and Configuration of Apache web server for DeNoGAP

DeNoGAP do not automatically install apache server. So please follow this steps to install apache on Ubuntu linux and configure it for GUI.

- sudo apt-get install apache2

This will set up apache server on user system. Open any Web Browser window and type <http://localhost> to test if apache is installed correctly or not. If installed correctly you should see a message on screen “It Works”.

- By default the web document root for apache server is “/var/www”. You can set your own document root by making changes in configuration files as shown below.
- Copy “GUI/cgi-bin” and “GUI/html” directories from DeNoGAP package to document root directory.

### Configuration of Apache Server for using GUI

Perform steps below on command line to configure apache server on linux/Unix system:

- cd /etc/apache2
- cd sites-available
- sudo gedit default

This will open “default” file using gedit text editor. “default” is a virtual host file. We will make changes in this file to access GUI for DeNoGAP

Inside “default” file make following changes to default configuration.

- (1) Change line “DocumentRoot /var/www” to “DocumentRoot /var/www/DeNoGAP/GUI”.
- (2) Change line “<Directory /var/www/>” to “<Directory /var/www/DeNoGAP/GUI>”
- (3) In the same <Directory> </Directory> block, change “Options Indexes FollowSymLinks MultiViews” to “Options Indexes FollowSymLinks MultiViews Includes ExecCGI”
- (4) Add a line after <Directory> </Directory> block,

- Script Alias “/var/www/DeNoGAP/GUI/cgi-bin” /usr/lib/cgi-bin
- (5) In the next <Directory> </Directory> block, Change <Directory /> to <Directory “/var/www/DeNoGAP/GUI/cgi-bin”>
- (6) Add line after the block,
- AddHandler cgi-script cgi pl
- (7) Save and Close “default” file.

Create symbolic link in “sites-enabled” directory to an virtual-host file in “sites-available” directory using following command:

- sudo a2ensite default

Restart Apache to read new configuration.

- sudo /etc/init.d/apache2 reload

Open Web Browser and type <http://localhost> to see web documents for DeNoGAP.

User should see two directories cg-bin and html. Instruction to use GUI is given in “Exploration of database section” below.

### 3. Input Data Files

DeNoGAP requires three input files to perform any analysis.

- (1) The tab-delimited file containing metadata information about the genomes used for the analysis.
- (2) The configuration file containing defined parameters for the analysis.
- (3) SQLite Database file to store output.

The format and description for each input file is given below.

#### 3.1 Genome Information File

The first line of the genome information file should start with “#” followed by tab-delimited column names. Remove any extra trailing spaces at the end of columns or line to avoid errors.

- **Mandatory Column names**
  - **genome\_name**: Full genome name.
  - **species**: Full name of the species.
  - **species\_type**: specify if genome belongs to bacteria or archaea. (Acceptable values: bacteria / archaea)
  - **abbreviation**: Short abbreviation for the genome. This will be used to identify and name all sequence files and output files. (Abbreviation should not have any dots or special characters except “\_”).
  - **genome\_type**: Indicate if genome is a reference or query. (Acceptable values: reference / query).
  - **outgroup**: Indicate if genome is an outgroup or not. (Acceptable values: Yes / No).
- **Optional Columns**
  - Users can create any number of new columns to add any additional information to the genome information table.
  - The name of additional columns should be in lower case without any special characters except “\_”.

The example table is given in the data directory of DeNoGAP package.

#### 3.2 Configuration Files

DeNoGAP uses separate configuration file for each analysis phase. All parameters, file paths, and directory paths required for performing the analysis should be defined in respective configuration file. Parameters are divided into different sections named within [] brackets. The

optional parameters can be disabled if not in use by prefixing parameter name as “#-parameter name” in the configuration file. The description of each configuration file and parameters included in it is given below:

- **PARSE\_GENBANK.config**

This configuration file defines parameters for extracting sequences and genomic information from the GenBank Files.

- **PARSE\_GENBANK:** Initiates parsing of genebank files (Default value: YES).
- **GENBANK\_DIR\_PATH:** Define directory path for genebank files.
- **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
- **GENOME\_DIR\_NAME:** Sub-directory name to store fasta formatted genome sequence files.
- **CDS\_DIR\_NAME:** Sub-directory name to store fasta formatted coding sequence files.
- **PROTEIN\_DIR\_NAME:** Sub-directory name to store fasta formatted protein sequence files.
- **FEATURE\_DIR\_NAME:** Sub-directory name to store tab-delimited genomic feature files.

- **PREDICT\_GENE.config**

This configuration file defines parameters to predict genes from the genome sequences using four gene prediction programs.

- **PREDICT\_GENE:** Initiate gene prediction analysis. (Default value: YES).
- **GENOME\_DIR\_PATH:** Define directory path for fasta-formatted genome sequence files.
- **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.

- **GILMMER\_RESULT\_DIR\_NAME:** Sub-directory name to store glimmer output files.
- **GENEMARK\_RESULT\_DIR\_NAME:** Sub-directory name to store GeneMark output files.
- **PRODIGAL\_RESULT\_DIR\_NAME:** Sub-directory name to store prodigal output files.
- **FRAGSCAN\_RESULT\_DIR\_NAME:** Sub-directory name to store fragscan output files.
- **CDS\_DIR\_NAME:** Sub-directory to store predicted coding sequence files.
- **PROTEIN\_DIR\_NAME:** Sub-directory to store translated protein sequence files.
- **FEATURE\_DIR\_NAME:** Sub-directory to store tab-delimited genomic feature files.
- **TRANSLATION\_CODE:** Genebank codon table for translating coding sequences into proteins.
- **OVERLAP\_BASE:** Number of overlapping bases allowed between adjacent genes.
- **PARALLEL\_CPU\_CORE:** Number of CPU core to be used for parallel processing.
- **GLIMMER3:** Define options for running glimmer3 program. Check available options from glimmer manual. All options should be defined within “ ”.
- **LONG\_ORF:** Define options for running long-orfs program. Check available options from glimmer manual. All options should be defined within “ ”.
- **MULTI\_EXTRACT:** Define options for running multi-extract program. Check available options from glimmer manual. All options should be defined within “ ”.
- **BUILD\_ICM:** Define options for running build-icm program. Check available options from glimmer manual. All options should be defined within “ ”.
- **GMSN:** Define options for running GeneMark program. Check available options from GeneMark manual. DeNoGAP automatically takes value for “-- name” and “-- species” options from the genome table. All other options should be defined here within “ ”.
- **PRODIGAL:** Define options for running Prodigal program. Check available options from prodigal help. DeNoGAP automatically takes value for -i , -t, -o, -a, -d , -s from the genome table. All other options should be defined here within “ ”.

- **FRAGSCAN:** Define options for running FragGeneScan program. Check available options from FragGeneScan help. DeNoGAP automatically takes value for “-genome” and “-out” options from the genome table. All other options should be defined here within “ ”.
- **GENE\_VERIFICATION.config**

This configuration file defines parameters to verify and annotate predicted protein sequences by comparing sequence match within Uniprot database.

  - **VERIFY\_SEQUENCE:** Initiate verification of predicted protein sequences using Uniprot database. (Default Value: YES).
  - **BLAST\_ALIGNMENT\_FILE:** Pairwise blast alignment result between protein sequences and UniPort database.
  - **FEATURE\_DIR:** Define full name of the directory including complete path containing genomic feature files.
  - **CDS\_DIR:** Define full name of the directory including complete path containing coding sequence files.
  - **PROTEIN\_DIR:** Define full name of the directory including complete path containing protein sequence files.
  - **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
  - **EVALUE\_THRESHOLD:** Define minimum e-value cut-off for significant hits.
  - **ALIGNMENT\_IDENTITY:** Define minimum sequence identity for significant hits.
  - **QUERY\_COVERAGE:** Define minimum query coverage for significant match.
  - **MIN\_PROTEIN\_LENGTH:** Define minimum protein sequence length cutoff to discard insignificant sequences.
- **LOAD\_DATA.config**

This configuration file defines parameters to load sequences and genomic feature information in to the SQLite database.

- **LOAD\_DATA:** Initiate module for loading sequences and genomic data. (Default: YES).
  - **FEATURE\_DIR:** Define full name of the directory including complete path containing genomic feature files.
  - **CDS\_DIR:** Define full name of the directory including complete path containing coding sequence files.
  - **PROTEIN\_DIR:** Define full name of the directory including complete path containing protein sequence files.
  - **ADJUST\_HEADER:** Default value: YES. Adjust sequence identifier and format it as “genome\_abbreviation|sequence\_identifier”.
- 
- **COMPARE\_REFERENCE.config**  
This configuration file defines parameter for pairwise sequence comparison between reference genomes using Phmmmer program.
    - **COMPARE\_REFERENCE:** (Default value: YES). Initiates pairwise sequence comparison between reference genomes defined by user in genome table.
    - **MODEL\_DB:** Define name for the database file to be created for Hidden Markov models of the protein families. (Default value: HMM\_MODEL\_DB).
    - **SEQ\_DB:** Define name for the database file to be created for Singleton protein family sequences. (Default value: HMM\_SEQ\_DB).
    - **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
    - **HMMER\_OPT:** Define options for running phmmmer program. Check available options for phmmmer from hmmer package manual. DeNoGAP automatically takes value for “-o” and “-domtblout” options from the genome table. All other options should be defined here within “ ”.
    - **MAX\_NUM\_DOMAIN:** Define maximum number of hmmer domains allowed between matched sequences. (Default value: 5).

- **ACCURACY\_THRESHOLD:** Define hmmer accuracy probability cutoff for significant match. The value range is between [0 - 1]. (Default value: 0.8)
  - **IDENTITY:** Define percentage identity cutoff for significant match. (Default value: 70).
  - **SIMILARITY:** Define percentage similarity cutoff for significant match. (Default value: 60).
  - **QUERY\_COVERAGE:** Define percentage cutoff for query sequence covered in a significant match. (Default value: 70).
  - **HMM\_COVERAGE:** Define percentage cutoff for hmm model sequence covered in a significant match. (Default value: 70).
  - **MIN\_CHIMERA\_IDENTITY:** Define percentage identity cutoff for predicting chimera-like match. (Default value: 70)
  - **MIN\_CHIMERA\_SIMILARITY:** Define percentage similarity cutoff for predicting chimera-like match. (Default value: 60)
  - **MIN\_CHIMERA\_QUERY\_COVERAGE:** Define percentage cutoff for query sequence covered in a chimera match. (Default value: 25).
  - **MIN\_CHIMERA\_HMM\_COVERAGE:** Define percentage cutoff for hmm model sequence covered in a chimera match. (Default value: 25).
  - **PARALLEL\_CPU\_CORE:** Define number of CPU cores to be used for the analysis. (Default value: 1).
  - **MCL\_INFLATION\_VALUE:** Define inflation vaule (I) for MCL clustering. (Default value: 1.5).
- 
- **PREDICT\_HMM\_FAMILY.config**  
This configuration files define parameters for iterative prediction of protein families in additional genomes.

- **PREDICT\_HMM:** Initiate iterative comparison of protein sequences from new genomes. (Default value: YES).
- **MODEL\_DB:** Define name for the database file to be created for Hidden Markov models of the protein families. (Default value: HMM\_MODEL\_DB).

- **SEQ\_DB:** Define name for the database file to be created for Singleton protein family sequences. (Default value: HMM\_SEQ\_DB).
- **HMM\_CLUSTER\_FILE:** Define complete path and file name of the seed family cluster file.
- **MODEL\_DB\_FILE (optional):** Define complete path and file name of the seed HMM model database file. If this option is disable, DeNoGAP first calculates hmm model for all families in the existing cluster file.
- **SINGLETON\_DB\_FILE (optional):** Define complete path and file name of the seed Singleton sequence database file. If this option is disable, DeNoGAP first calculates singleton sequence for all families in the existing cluster file.
- **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
- **HMMER\_OPT:** Define options for running hmmscan and phmmmer program. Check available options for hmmscan and phmmmer from hmmer package manual. DeNoGAP automatically takes value for “-o” and “-domtblout” options from the genome table. All other options should be defined here within “ ”.
- **MAX\_NUM\_DOMAIN:** Define maximum number of hmmer domains allowed between matched sequences. (Default value: 5).
- **ACCURACY\_THRESHOLD:** Define hmmer accuracy probability cutoff for significant match. The value range is between [0 - 1]. (Default value: 0.8)
- **IDENTITY:** Define percentage identity cutoff for significant match. (Default value: 70).
- **SIMILARITY:** Define percentage similarity cutoff for significant match. (Default value: 60).
- **QUERY\_COVERAGE:** Define percentage cutoff for query sequence covered in a significant match. (Default value: 70).
- **HMM\_COVERAGE:** Define percentage cutoff for hmm model sequence covered in a significant match. (Default value: 70).
- **MIN\_CHIMERA\_IDENTITY:** Define percentage identity cutoff for predicting chimera-like match. (Default value: 70)

- **MIN\_CHIMERA\_SIMILARITY:** Define percentage similarity cutoff for predicting chimera-like match. (Default value: 60)
  - **MIN\_CHIMERA\_QUERY\_COVERAGE:** Define percentage cutoff for query sequence covered in a chimera match. (Default value: 25).
  - **MIN\_CHIMERA\_HMM\_COVERAGE:** Define percentage cutoff for hmm model sequence covered in a chimera match. (Default value: 25).
  - **PARALLEL\_CPU\_CORE:** Define number of CPU cores to be used for the analysis. (Default value: 1).
  - **MCL\_INFLATION\_VALUE:** Define inflation vaule (I) for MCL clustering. (Default value: 1.5).
- 
- **PREDICT\_SUPER\_HOMOLOG.config**

This configuration file defines parameters for predicting super-homolog family and identifying links between partial gene families and full-length gene families.

    - **PREDICT\_SUPER\_HOMOLOG:** Initiate prediction of super-homolog families (Default value: YES).
    - **HMM\_CLUSTER\_FILE:** Define complete path and file name of the hmm-family cluster file.
    - **OLD\_HOMOLOG\_DATABASE\_FILE (optional):** Define complete path and file name of existing homolog database file. If value for this parameter is defined, DeNoGAP will use homolog cluster information from this database to append members from new genome to existing homolog families. If you don't want to use this parameter. Insert “#.” symbol before the parameter name to disable it.
    - **IDENTITY\_THRESHOLD:** Define percentage identity cutoff for significant partial match. (Default value: 70).
    - **SIMILARITY\_THRESHOLD:** Define percentage similarity cutoff for significant partial match. (Default value: 60).
    - **QUERY\_COVERAGE\_THRESHOLD:** Define percentage cutoff for query sequence covered in a significant partial match. (Default value: 70).

- **SUBJECT\_COVERAGE\_THRESHOLD:** Define percentage cutoff for subject sequence covered in a significant partial match. (Default value: 70).
  - **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
- 
- **PREDICT\_ORTHOLOG.config**  
This configuration file defines parameters for predicting ortholog and inparalog protein pairs and cluster ortholog families.
    - **PREDICT\_ORTHOLOG:** Initiate prediction of ortholog and in paralog pairs. (Default value: YES).
    - **CLUSTER\_ORTHOLOG:** Initiate clustering of ortholog and inparalog pairs. (Default value: YES).
    - **HMM\_CLUSTER\_FILE:** Define complete path and file name of the hmm-family cluster file.
    - **HOMOLOG\_CLUSTER\_FILE:** Define complete path and file name of super-homolog cluster file.
    - **OLD\_ORTHOLOG\_DATABASE\_FILE (optional):** Define complete path and file name of existing ortholog database. If value for this parameter is provided, DeNoGAP will only compute ortholog and paralog pair in new added genomes and append new information to the existing data.
    - **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
    - **ORTHOLOG\_DIVERGENCE\_THRESHOLD:** Define distance cut-off between [0 – 10] for predicting ortholog pairs in case out-group is absent. Less than 1 indicates similar sequences and greater than 1 indicates divergent sequences (Default value: 0.8).
    - **INPARALOG\_DIVERGENCE\_THRESHOLD:** Define distance cut-off between [0 – 10] for predicting inparalog pairs in case out-group is absent. Less than 1

indicates similar sequences and greater than 1 indicates divergent sequences (Default value: 0.5).

- **GLOBAL\_ALIGNMENT\_OVERLAP\_THRESHOLD:** Define sequence overlap cut-off between [0 – 1] for each pair of sequences in global alignment of homolog family. 0 indicates no overlap between sequences and 1 indicates complete overlap between sequences.
- **PAIRWISE\_SEQUENCE\_IDENTITY\_THRESHOLD:** Define identity cut-off between [0 – 100] for each pair of sequences in a homolog family. 0 indicates no identity at all between sequences and 100 indicates completely identical sequences.
- **PAIRWISE\_SEQUENCE\_SIMILARITY\_THRESHOLD:** Define similarity cut-off between [0 – 100] for each pair of sequences in a homolog family. 0 indicates no similarity at all between sequences and 100 indicates highly similar sequences.
- **MCL\_INFLATION\_VALUE:** Define inflation vaule (I) for MCL clustering. (Default value: 1.5).
- **PARALLEL\_CPU\_CORE:** Define number of CPU cores to be used for the analysis. (Default value: 1).

- **PHYLOGENETIC\_PROFILE.config**

This configuration file generates a phylogenetic profile matrix to represent presence or absence of protein families across genomes.

- **PHYLOGENETIC\_PROFILE:** Initiate analysis for making binary Phylogenetic profile. (Default value: YES).
- **ORTHOLOG\_CLUSTER\_FILE:** Define complete path and name of the protein family cluster file.
- **ORTHOLOG\_DATABASE\_FILE:** Define complete path and name of the ortholog database file for storing phylogenetic profile matrix.
- **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.

- **CORE\_GENOME.config**

This configuration file defines parameters for predicting core protein families and create concatenated core-genome alignment.

- **CORE\_GENOME:** Initiate analysis for predicting core-genome. (Default value: YES).
- **ORTHOLOG\_CLUSTER\_FILE:** Define complete path and name of the protein family cluster file.
- **ORTHOLOG\_DATABASE\_FILE:** Define complete path and name of the ortholog database file.
- **CORE\_ALIGNMENT\_FILE:** Define complete path and name of the concatenated core alignment file.
- **CORE\_THRESHOLD:** Define minimum percentage of required genome for predicting core-genome.
- **SEQUENCE\_TYPE:** Define sequence type (Options: nucleotide / protein). (Default value: protein).
- **INCLUDE\_OUTGROUP:** Include out-group sequences in core genome alignment. (Default value: NO).
- **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.

- **ANNOTATION.config**

This configuration files defines parameters for predicting functional annotation for protein sequences using InterProScan.

- **PREDICT\_ANNOTATION:** Initiates annotation of protein sequences. (Default value: YES).
- **INTERPRO\_SCAN\_PATH:** Define source directory path for interproscan databases and files.

- **INTERPRO\_SCAN\_OPTS:** Define options for running interproscan analysis. Check available options from InterProScan help. DeNoGAP automatically takes value for “-i”, “-f” and “-o” options from the genome table. All other options should be defined here within “ ”.
- **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
- **PARALLEL\_CPU\_CORE:** Define number of CPU cores to be used for the analysis. (Default value: 1).

## 4. SQLite Database Schema

DeNoGAP uses SQLite database to store analyzed information. DeNoGAP creates 3 databases: Master database, Homolog database and Ortholog database with total 20 tables to store results from various analysis phases. The description of each table and their columns is given below:

### (I) Master Database:

Table: OrganismInfo		
Column	Data type	Description
genome_name	TEXT	Full name of the genome
species	TEXT	Full name of the species
genome_abbreviation	TEXT	short name for the genome
genome_type	TEXT	Is Reference or Query
outgroup	TEXT	Is outgroup or not

Table: GeneFeature		
Column	Data type	Description
feature_id	TEXT	Unique sequence identifier for the gene
feature_type	TEXT	By default: CDS
protein_id	TEXT	Unique protein sequence identifier.
dbxref	TEXT	External database identifier for gene or protein.
genome_id	TEXT	Unique sequence identifier for the genome sequence
genome_type	TEXT	Chromosome / Plasmid / Contig
genome_name	TEXT	Full name of the genome
genome_length	INT	Length of the genome sequence
feature_start	INT	Start co-ordinate of the gene sequence
feature_end	INT	End co-ordinate of the gene sequence
nuc_length	INT	Length of the coding sequence
aa_length	INT	Length of the protein sequence

strand	INT	Genome strand on which gene is located (+ or -)
frame	INT	Coding frame for the coding sequence
index_on_genome	INT	Order on the genome sequence
description	TEXT	Product description
comment	TEXT	Comment by user

<b>Table: ProteinSequence</b>		
<b>Column</b>	<b>Data type</b>	<b>Description</b>
pseq_index_id	INT	Auto-incremented primary key index
protein_id	TEXT	Unique sequence identifier for protein
genome_abbreviation	TEXT	Short name for the genome
seq_type	TEXT	Protein
seq_length	INT	Length of amino acid sequence
aa_sequence	TEXT	Protein sequence

<b>Table: NucleotideSequence</b>		
<b>Column</b>	<b>Data type</b>	<b>Description</b>
nseq_index_id	INT	Auto-incremented primary key index
nucleotide_id	TEXT	Unique sequence identifier for CDS
genome_abbreviation	TEXT	Short name for the genome
seq_type	TEXT	Protein
seq_length	INT	Length of coding sequence sequence
nuc_sequence	TEXT	CDS sequence

<b>Table: Similarity</b>		
<b>Column</b>	<b>Data type</b>	<b>Description</b>
query_id	TEXT	Sequence identifier for the query protein
subject_id	TEXT	Identifier for the target sequence or target hmm

		group
query_length	INT	Length of query sequence
subject_length	INT	Length of target sequence or target hmm model
num_total_domain	INT	Total domains predicted in the query sequence
num_significant_domain	INT	Number of domains with significant match
query_start	INT	Start position of query sequence
query_end	INT	End position of query sequence
subject_start	INT	Start position of the target
subject_end	INT	End position of the target
evalue	REAL	Significance value
bit_score	INT	Bit score of the alignment
percent_identity	REAL	Percentage identity between sequences
percent_similarity	REAL	Percentage similarity between sequences
query_coverage	REAL	Percentage query sequence coverage
subject_coverage	REAL	Percentage target sequence coverage
pair_relation	TEXT	Match type (best / truncated / chimera / insignificant)

Table: DomainAnnotation		
Column	Data type	Description
protein_id	TEXT	Unique sequence identifier
genome_name	TEXT	Genome abbreviation
seq_len	INT	Length of query sequence
domain_id	TEXT	Unique identifier for the predicted domain
domain_name	TEXT	Name of the predicted domain
domain_start	INT	Start position of the domain
domain_end	INT	End position of the domain
significance_value	REAL	Significance of the domain match
description	TEXT	Domain description

**Table: InterProAnnotation**

<b>Column</b>	<b>Data type</b>	<b>Description</b>
protein_id	TEXT	Unique sequence identifier
genome_name	TEXT	Genome abbreviation
interpro_id	TEXT	Unique identifier for the interpro domain
interpro_name	TEXT	Name of the predicted interpro domain

**Table: GOAnnotation**

<b>Column</b>	<b>Data type</b>	<b>Description</b>
protein_id	TEXT	Unique sequence identifier
genome_name	TEXT	Genome abbreviation
go_id	TEXT	Unique identifier for the gene ontology term
go_category	TEXT	Classification category for go term
go_description	TEXT	Description of the go term

**Table: PathwayAnnotation**

<b>Column</b>	<b>Data type</b>	<b>Description</b>
protein_id	TEXT	Unique sequence identifier
genome_name	TEXT	Genome abbreviation
pathway_id	TEXT	Unique identifier for the predicted pathway
pathway_name	TEXT	Name of the predicted pathway

**Table: SignalPAnnotation**

<b>Column</b>	<b>Data type</b>	<b>Description</b>

protein_id	TEXT	Unique sequence identifier
genome_name	TEXT	Genome abbreviation
domain_start	REAL	Start position of signal peptide
domain_end	REAL	End position of signal peptide

<b>Table: TMHMMAnnotation</b>		
<b>Column</b>	<b>Data type</b>	<b>Description</b>
protein_id	TEXT	Unique sequence identifier
genome_name	TEXT	Genome abbreviation
domain_start	REAL	Start position of transmembrane
domain_end	REAL	End position of transmembrane

<b>Table: PhobiusAnnotation</b>		
<b>Column</b>	<b>Data type</b>	<b>Description</b>
protein_id	TEXT	Unique sequence identifier
genome_name	TEXT	Genome abbreviation
domain_name	TEXT	Name of domain
domain_description	TEXT	Description of domain
domain_start	REAL	Start position of domain
domain_end	REAL	End position of domain

## (II) Homolog Database:

<b>Table: LinkFamily</b>		
<b>Column</b>	<b>Data type</b>	<b>Description</b>
family_idA	TEXT	Group id of hmm family
family_idB	TEXT	Group id of hmm family
significance	REAL	Significance value between pair

**Table: GenetoSuperFamily**

Column	Data type	Description
gene_superfamily_index_id	INT	Auto increment primary key index
gene_id	TEXT	Unique sequence identifier
genome_name	TEXT	Genome abbreviation
hmm_family_id	TEXT	Hmm family identifier
super_family_id	TEXT	Super-homolog family identifier

**(III) Ortholog Database:****Table: MultipleAlignment**

Column	Data type	Description
seq_id	TEXT	Unique sequence identifier
genome_abbreviation	TEXT	Genome abbreviation
seq_type	TEXT	Genome abbreviation
alignment_id	TEXT	Super-homolog family identifier
alignment_length	INT	Length of super-homolog alignment
alignment_sequence	TEXT	Aligned sequence

**Table: DistancePair**

Column	Data type	Description
taxonA	TEXT	Genome abbreviation for SeqA
idA	TEXT	Unique sequence identifier SeqA
taxonB	TEXT	Genome abbreviation for SeqB
idB	TEXT	Unique sequence identifier SeqB
divergence	REAL	Pairwise distance between SeqA and SeqB
homolog_cluster_id	TEXT	Super-homolog family identifier

Table: OrthologPair		
Column	Data type	Description
taxonA	TEXT	Genome abbreviation for SeqA
idA	TEXT	Unique sequence identifier SeqA
taxonB	TEXT	Genome abbreviation for SeqB
idB	TEXT	Unique sequence identifier SeqB
divergence	REAL	Pairwise distance between SeqA and SeqB
homolog_cluster_id	TEXT	Super-homolog family identifier

Table: InParalogPair		
Column	Data type	Description
taxonA	TEXT	Genome abbreviation for SeqA
idA	TEXT	Unique sequence identifier SeqA
taxonB	TEXT	Genome abbreviation for SeqB
idB	TEXT	Unique sequence identifier SeqB
divergence	REAL	Pairwise distance between SeqA and SeqB
min_ortholog_divergence	REAL	Minimum pairwise distance between SeqA and ortholog from any other genome.
homolog_cluster_id	TEXT	Super-homolog family identifier

Table: PairwiseAlignment		
Column	Data type	Description
idA	TEXT	Unique sequence identifier SeqA
idB	TEXT	Unique sequence identifier SeqB
sequence_identity	REAL	Pairwise identity between idA and idB
sequence_similarity	REAL	Pairwise similarity between idA and idB
homolog_cluster_id	TEXT	Super-homolog family identifier

**Table: PhylogeneticProfile**

<b>Column</b>	<b>Data type</b>	<b>Description</b>
id	TEXT	Ortholog family id
genome_name	TEXT	Genome abbreviation

**Table: MapGeneIdtoGeneFamily**

<b>Column</b>	<b>Data type</b>	<b>Description</b>
familymap_index_id	INT	Auto-incremented primary key index
genefamily_id	TEXT	Unique identifier for predicted ortholog family
gene_id	TEXT	Unique identifier for the sequence
specie_abbreviation	TEXT	Genome abbreviation

## 5. Running DeNoGAP commands

In order to run analysis using DeNoGAP execute commands shown below in the command line terminal.

```
cd DeNoGap_v1.0
```

```
cd bin
```

```
perl Denogap_v1.0.pl -genome_info <genome information file> -db_dir <full path sqlite database dir path> -db_name <name of the master sqlite database> -config <full path and name of configuration file> -output_dir <full path and name to the output directory>
```

### Running DeNoGAP with test dataset

The package contains 2 dataset that are under **data/test1** and **data/test2** directories.

The data/test1 contains input files to initiate new analysis with new set of genomes.

The data/test2 contains input files to add new genomes to the existing analysis database.

**Note 1:** Sometimes copying command directly from documentation file and pasting in terminal window may cause errors. If so, please type the command in terminal as shown below.

**Note 2:** Replace ~userpath with path to DeNoGAP directory as per the user's directory path.

- **Test parse genbank files**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENOGENAP/data/test1/test_genome_table_1.txt -db_dir  
/home/~/userpath/DENOGENAP/output/TEST_RUN -db_name test.sqlite -config  
/home/~/userpath/DENOGENAP/config/PARSE_GENBANK.config -output_dir  
/home/~/userpath/DENOGENAP/output/TEST_RUN
```

- **Test gene prediction**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENOGENAP/data/test1/test_genome_table_1.txt -db_dir  
/home/~/userpath/DENOGENAP/output/TEST_RUN -db_name test.sqlite -config  
/home/~/userpath/DENOGENAP/config/PREDICT_GENE.config -output_dir  
/home/~/userpath/DENOGENAP/output/TEST_RUN
```

- **Test gene verification and annotation**

- Download and extract UniProt database.

([ftp://ftp.uniprot.org/pub/databases/uniprot/current\\_release/knowledgebase/complete/uniprot\\_sprot.fasta.gz](ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz))

- Perform protein Blast with all predicted sequences against Uniprot database and store output in default format (-outfmt 0).

```
blastp -query <All_predicted_sequence_file> -db <Uniprot_database_file> -eval 1e-05 -out <blast_output_file> -max_target_seq 1
```

- Perform gene verification and annotation analysis using DeNoGAP.

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENOGap/data/test1/test_genome_table_1.txt -db_dir  
/home/~/userpath/DENOGap/output/TEST_RUN -db_name test.sqlite -config  
/home/~/userpath/DENOGap/config/GENE_VERIFICATION.config -output_dir  
/home/~/userpath/DENOGap/output/TEST_RUN
```

- **Test Load Data**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENOGap/data/test1/test_genome_table_1.txt -db_dir  
/home/~/userpath/DENOGap/output/TEST_RUN -db_name test.sqlite -config  
/home/~/userpath/DENOGap/config/LOAD_DATA.config -output_dir  
/home/~/userpath/DENOGap/output/TEST_RUN
```

- **Test Reference comparision**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENOGap/data/test1/test_genome_table_1.txt -db_dir  
/home/~/userpath/DENOGap/output/TEST_RUN -db_name test.sqlite -config  
/home/~/userpath/DENOGap/config/COMPARE_REFERENCE.config -output_dir  
/home/~/userpath/DENOGap/output/TEST_RUN
```

- **Test Iterative HMM family prediction**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENOGap/data/test1/test_genome_table_1.txt -db_dir  
/home/~/userpath/DENOGap/output/TEST_RUN -db_name test.sqlite -config
```

```
/home/~/userpath/DENoGAP/config/PREDICT_HMM_FAMILY.config --output_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN
```

- **Test Super-homolog family prediction**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENoGAP/data/test1/test_genome_table_1.txt --db_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN --db_name test.sqlite --config  
/home/~/userpath/DENoGAP/config/PREDICT_SUPER_HOMOLOG.config --output_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN
```

- **Test Ortholog Prediction and Ortholog Clustering**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENoGAP/data/test1/test_genome_table_1.txt --db_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN --db_name test.sqlite --config  
/home/~/userpath/DENoGAP/config/PREDICT_ORTHOLOG.config --output_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN
```

- **Test Phylogenetic profile**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENoGAP/data/test1/test_genome_table_1.txt --db_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN --db_name test.sqlite --config  
/home/~/userpath/DENoGAP/config/PHYLOGENETIC_PROFILE.config --output_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN
```

- **Test Core genome prediction**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENoGAP/data/test1/test_genome_table_1.txt --db_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN --db_name test.sqlite --config  
/home/~/userpath/DENoGAP/config/CORE_GENOME.config --output_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN
```

- **Test Function Annotation prediction**

```
perl DeNoGap_v1.0.pl -genome_info  
/home/~/userpath/DENoGAP/data/test1/test_genome_table_1.txt --db_dir
```

```
/home/~/userpath/DENoGAP/output/TEST_RUN -db_name test.sqlite -config  
/home/~/userpath/DENoGAP/config/ANNOTATION.config -output_dir  
/home/~/userpath/DENoGAP/output/TEST_RUN
```

## 6. Output Directory Structure

This section gives an overview of the output directory structure and output files created for each analysis phase.

**Database directory:** User while running DeNoGAP should define the name and path of database directory.

**Database name:** User while running DeNoGAP should define the name of the master SQLite database file.

### Analysis: Parse GenBank

The output directory for parsed genbank files contains four sub-directories defined in the configuration file.

- **GENEBANK\_DATA**

- **Genome sequence directory:** This directory contains fasta formatted genome sequence files (one file for each organism).
- **Coding sequence directory:** This directory contains fasta formatted coding gene sequences (one file for each organism).
- **Protein sequence directory:** This directory contains fasta formatted protein sequences (one file for each organism).
- **Genomic feature directory:** This directory contains tab-delimited genomic feature files (one file for each organism).

### Analysis: Gene Prediction

The output directory for gene prediction contains seven sub-directories defined in the configuration file.

- **GENE\_PREDICTION:**

- **Glimmer directory:** This directory stores output from Glimmer software (one directory for each genome).
- **GeneMark directory:** This directory stores output from GeneMark software (one directory for each genome).
- **Prodigal directory:** This directory stores output from Prodigal software (one directory for each genome).
- **FragGeneScan directory:** This directory stores output from FragGeneScan software (one directory for each genome).
- **Coding sequence directory:** This directory contains two sub-directories “MULTIPLE” and “SINGLE”.

- **MULTIPLE**: This sub-directory stores coding sequences in fasta format predicted by more than one gene prediction program (one file for each genome).
  - **SINGLE**: This sub-directory stores coding sequences in fasta format predicted by only one gene prediction program (one file for each genome).
- **Protein sequence directory**: This directory contains two sub-directories “MULTIPLE” and “SINGLE”.
  - **MULTIPLE**: This sub-directory stores translated protein sequences in fasta format for genes predicted by more than one gene prediction program (one file for each genome).
  - **SINGLE**: This sub-directory stores translated protein sequences in fasta format for genes predicted by only one gene prediction program (one file for each genome).
- **Genomic feature directory**: This directory contains two sub-directories “MULTIPLE” and “SINGLE”.
  - **MULTIPLE**: This sub-directory stores genomic features in tabular format for genes predicted by more than one gene prediction program (one file for each genome).
  - **SINGLE**: This sub-directory stores genomic features in tabular format for genes predicted by only one gene prediction program (one file for each genome).

## **Analysis: Gene prediction verification and annotation**

The output directory for sequence verification stores output files for verified Coding sequences, protein sequences and their genomic features along with Uniprot annotation. It also stores genbank files for each genome with verified sequences.

### ▪ **SEQUENCE\_VERIFICATION**

- **CDS\_SEQUENCE**: This sub-directory stores verified coding sequences for each genome (one file for each genome).
- **PROTEIN\_SEQUENCE**: This sub-directory stores verified protein sequences for each genome (one file for each genome).
- **GENOMIC\_FEATURE**: This sub-directory stores genomic features and Uniprot annotation for verified sequences (one file for each genome).
- **GENBANK\_FILE**: This sub-directory stores genebank file for each genome (one file for each genome).

## Analysis: Homolog prediction

The output directory for homolog prediction stores output files for reference genome comparison, iterative hmm-family prediction and ortholog prediction.

- **HOMOLOG\_SCAN:** This directory stores all the output files and sub-directories for homolog prediction analysis.
  - **HMMER\_OUT:** This directory stores sub-directories to store un-parsed hmmscan and phmmmer output files in alignment format and tabular format.
    - **HMM\_DOM:** This sub-directory stores hmmer results in domain table format.
    - **HMM\_FULL:** This sub-directory stores hmmer results in default alignment format.
  - **BEST\_PAIR:** This sub-directory stores similarity information for highly similar sequences (one file for each organism).
  - **CHIMERA\_PAIR:** This sub-directory stores similarity information for chimera-like protein sequences (one file for each organism).
  - **ALL\_PAIR:** This sub-directory stores similarity information for all kind of pairwise matches including highly significant hit, significant partial hits, chimera-like hits and insignificant sequence match.
  - **MCL:** This sub-directory stores output from MCL clustering.
  - **HMM:** This directory stores sub-directories for HMM models and singleton sequences.
    - **MODEL:** This sub-directory stores HMM models for protein families.
    - **SINGLETON:** This sub directory stores singleton protein families.
  - **HMM\_DB:** This sub-directory stores HMM model database files and Singleton sequence database file.
  - **HOMOLOG:** This sub-directory stores tab-delimited file showing linked-relationship between partial and full-length HMM families.
  - **ORTHOLOG:** This directory consists of other sub-directories for storing output files from ortholog prediction analysis.
    - **PAIR\_DISTANCE:** This sub-directory stores pairwise genetic distance between each pair of proteins in a super-homolog family (one file for each super-homolog family).
    - **PAIR\_ORTHOLOG:** This sub-directory stores pairwise genetic distance between each pair of ortholog proteins (one file for each super-homolog family).
    - **PAIR\_INPARALOG:** This sub-directory stores pairwise genetic distance between each pair of inparalog proteins (one file for each super-homolog family).

- **ALIGNMENT:** This sub-directory stores multiple alignment of each super-homolog family (one file for each super-homolog family).
  - **PAIRWISE\_ALIGNMENT:** This sub-directory stores pairwise local alignment result for each pair of sequence in a homolog family (one file for each super-homolog family).
  - **ORTHO\_CLUSTER:** This sub-directory stores ortholog protein family information (one file for each super-homolog family).
- **RESULT:** This sub-directory stores clustering results for latest protein hmm models, super-homolog family and ortholog family. It also stores latest HMM model database and Singleton sequence database.

### **Analysis: Phylogenetic profile**

The output directory to store results from phylogenetic profile analysis.

- **PROFILE:** The name for this output directory is taken from the PROJECT\_DIR\_NAME parameter in the configuration file. It stores presence and absence information of the protein families across compared genomes as a binary matrix and tabular list.

### **Analysis: Core Genome Prediction**

The output directory for core genome prediction stores amino acid sequences and alignments for core gene families.

- **CORE\_SEQ:** This sub-directory stores amino acid sequences for the core gene families in fasta format (one file for each core family).
- **CORE\_ALN:** This sub-directory stores amino acid sequence alignment for the core gene families in fasta format (one file for each core family).
- **CORE\_ALIGNMENT\_FILE:** The name and location of concatenated core alignment file is taken from the CORE\_ALIGNMENT\_FILE parameter in the configuration file.

### **Analysis: InterProScan Annotation**

The output directory for InterProScan annotation stores output files for domain prediction, pathway prediction, gene ontology, and signal peptide prediction.

- **INTERPRO\_ANNOTATION:** This directory stores all the output files and sub-directories for interproscan annotation.
  - **XML\_FILE:** This sub-directory stores un-parsed InterProScan output files in XML format (one file for each genome).

- **TAB\_FILE:** This sub-directory stores un-parsed InterProScan output files in Tab-delimited format (one file for each genome).
- **DOMAIN:** This sub-directory stores parsed domain information (one file for each genome).
- **INTERPRO:** This sub-directory stores parsed interpro domain information (one file for each genome).
- **GENE\_ONTOLOGY:** This sub-directory stores parsed gene ontology information (one file for each genome).
- **PATHWAY:** This sub-directory stores parsed metabolic pathway information (one file for each genome).
- **SIGNALP:** This sub-directory stores parsed signal peptide information (one file for each genome).
- **TMHMM:** This sub-directory stores parsed transmembrane information (one file for each genome).
- **PHOBIUS:** This sub-directory stores parsed results from phobius (one file for each genome).

## 7. Graphical Interface for DeNoGAP database exploration

Steps for using GUI to explore DeNoGAP database are as follow:

- Open [http://localhost/html/open\\_database.html](http://localhost/html/open_database.html) in the Web Browser

### DENOVO GENOME EXPLORER

<b>Name of the project:</b>	<input type="text" value="Test_Set"/>
<b>Central Database File:</b>	<input type="text" value="/home/ds glab/Documents/RunAnalysis/MicrobialGenom"/>
<b>Homolog Database File:</b>	<input type="text" value="/home/ds glab/Documents/RunAnalysis/MicrobialGenom"/>
<b>Ortholog Database File:</b>	<input type="text" value="/home/ds glab/Documents/RunAnalysis/MicrobialGenom"/>
<b>Web Document Root:</b>	<input type="text" value="/home/ds glab/Documents/DATABASE"/>
<input type="button" value="Load Database"/> <input type="button" value="Reset"/>	

- Enter name of the Project:  
TEST\_RUN
- Enter full path and name of the Master SQLite database file:  
/home/~/User\_path/DeNoGAP/output/TEST\_RUN/test.sqlite
- Enter full path and name of the Homolog SQLite database file:  
/home/~/User\_path/DeNoGAP/output/TEST\_RUN/HomologDB\_<time\_stamp>
- Enter full path and name of the Ortholog SQLite database file:  
/home/~/User\_path/DeNoGAP/output/TEST\_RUN/OrthologDB\_<time\_stamp>
- Enter full path and name of the Apache Web Document root:  
/home/~/User\_path/Documents/WebRoot

- Click “Load Database”.
- **Analysis Query Interface**

*Analysis Report of Test\_Set*

**Find in selected genomes:**

\* Core Genes  
○ Variable Genes  
○ Unique Genes  
Define core gene as present in % of genomes: 100 ▾  
Show result sorted by: Gene Family ID ▾  
Search for gene description :

Show 10 ▾ entries

		genome_name	species	species_type	abbreviation
<input type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. tomato T1	Pseudomonas syringae	bacteria	PtoT1
<input type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. tomato Max13	Pseudomonas syringae	bacteria	PtoMax13
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. tomato DC3000	Pseudomonas syringae	bacteria	PtoDC3000
<input type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. tabaci str. ATCC 11528	Pseudomonas syringae	bacteria	Pta11528
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. syringae B728a	Pseudomonas syringae	bacteria	PsyB728a
<input type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. pisi 1704B	Pseudomonas syringae	bacteria	Ppi1704B
<input type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. phaseolicola 1448A	Pseudomonas syringae	bacteria	Pph1448A
<input type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. oryzae 36_1	Pseudomonas syringae	bacteria	Por36_1
<input type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. morsprunorum str. M302280	Pseudomonas syringae	bacteria	Pmo302280PT
<input type="checkbox"/>	<input type="checkbox"/>	Pseudomonas syringae pv. mori str. 301020	Pseudomonas syringae	bacteria	Pmo301020

Showing 1 to 10 of 24 entries (filtered from 142 total entries)

Search: Pseudomonas syringae

Previous 1 2 3 Next

- **Query Options:**
  - **Core Genes:** Show core gene families present in selected genomes.
  - **Variable Genes:** Show accessory gene families present in selected genomes.
  - **Unique Genes:** Show genome-specific families present in selected genomes.
  - **Define core gene threshold:** Define percentage cut-off of genomes for predicting core genes.
  - **Show result sorted by:** Gene families / Gene ID. (By default: Core genes are sorted by gene families and Variable and Unique genes are sorted by gene id).
  - **Search description:** Show results for genes having specific functional description only.
- **Genome Table:**
  - **Search:** Filter and display rows containing specific text in any column only.
  - **With Homolog:** Select genomes for which homolog should be present.
  - **Without Homolog:** Select genomes for which homolog should be absent.
- Click “Submit”.
- Result Table

## Search Result

Genome Name: [Pseudomonas syringae pv. syringae B728a](#)

[Show Gene Information](#)

Show [10](#) entries

Search:  [Copy](#) [CSV](#) [Excel](#)

Group ID	Homolog Group ID	HMM Group ID	Gene ID	Genome Name	Gene Description	Details
Cluster_10212.1	Cluster_10212	Group2510	Q4ZL22	Pseudomonas syringae pv. syringae B728a	ATP synthase subunit delta	<a href="#">...</a>
Cluster_10520.1	Cluster_10520	Group5462	Q4ZMP2	Pseudomonas syringae pv. syringae B728a	Elongation factor Tu	<a href="#">...</a>
Cluster_10870.1	Cluster_10870	Group4888	Q4ZMN2	Pseudomonas syringae pv. syringae B728a	Transcription termination/antitermination protein NusG	<a href="#">...</a>
Cluster_11008.1	Cluster_11008	Group3165	Q4ZRH5	Pseudomonas syringae pv. syringae B728a	6-carboxy-5,6,7,8-tetrahydropterin synthase	<a href="#">...</a>
Cluster_11077.1	Cluster_11077	Group5954	Q4ZZW4	Pseudomonas syringae pv. syringae B728a	D-amino acid dehydrogenase	<a href="#">...</a>
Cluster_11238.1	Cluster_11238	Group4397	Q4ZRK4	Pseudomonas syringae pv. syringae B728a	Isocitrate dehydrogenase NADP-dependent, monomeric type	<a href="#">...</a>
Cluster_11536.1	Cluster_11536	Group3052	Q4ZM01	Pseudomonas syringae pv. syringae B728a	S-adenosylmethionine synthase	<a href="#">...</a>
Cluster_1198.1	Cluster_1198	Group4047	Q4ZZ83	Pseudomonas syringae pv. syringae B728a	Dihydroxy-acid dehydratase	<a href="#">...</a>
Cluster_1254.1	Cluster_1254	Group1221	Q4ZQB3	Pseudomonas syringae pv. syringae B728a	Unannotated protein sequence	<a href="#">...</a>
Cluster_12972.1	Cluster_12972	Group2410	Q4ZUP4	Pseudomonas syringae pv. syringae B728a	Unannotated protein sequence	<a href="#">...</a>

Showing 1 to 10 of 70 entries

Previous [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) Next

- **Genome Name:** Select genome name to view result for specific genome.
- **Filter row with terms:** Filter and display rows having specific description term only.
- **Group Id:** Display ortholog gene family id
- **Homolog Group Id:** Display homolog family id.
- **HMM Group Id:** Display HMM family id.
- **Gene ID:** Display Locus\_tag id for selected genome in each ortholog family.
- **Gene Description:** Show product description for selected Locus tag Id.
- **Detail:** Select Locus tag Id to display detailed information about the gene.
  
- Click “Show Information” to display detailed information for selected gene Id.

- Gene Detail Information

Edit		GENE : Q4ZL22					
Genomic Feature:							
Locus Tag:	Q4ZL22						
Protein Id:	Pyr_5123						
External database Id:	YP_238188.1						
Species / Strain Name:	Psedomonas syringae pv. syringae B728a						
Species Type:	bacteria						
Species Abbreviation:	PsB728a						
Genome ID:	NC_007005.1						
Genome Type:	Chromosome						
Genome Length:	6093698						
Index on Genome:	5249						
Genomic Location:	60081466 : 6082002 (-1)						
Gene Name:	Not available						
Feature Type:	CDS						
Protein Length:	179						
Nucleotide Length:	537						
Product Description:	ATP synthase subunit delta						
Comments:	Produces ATP from ADP in the presence of a proton gradient across the membrane. The alpha chain is a regulatory subunit.						
Comparative Genomics Information:							
Homolog Group:	Cluster_10212						
Ortholog Group:	Cluster_10212.1						
HMM Model Group:	Group2510						
Annotation:							
GO Annotation:		GO ID	GO Category	GO Description			
		GO:0005886	Cellular Component	The membrane surrounding a cell that separates the cell from its external environment. It consists of a phospholipid bilayer and associated proteins.			
PFam:		PFam ID	Pfam Name	Description	Start	End	Significance
		PF00006	ATP-synt_ab	ATP synthase alpha/beta family, nucleotide-binding domain	150	376	1e-10
InterPro:		InterPro ID		Description			
		IPR020003		ATPase, alpha/beta subunit, nucleotide-binding domain, active site			

- Click edit button to make changes to the annotation stored in the database for individual gene.

- Annotation Edit form for selected gene

**EDIT GENE ANNOTATION**

<b>Locus Tag :</b>	<input type="text" value="Q4ZL22"/>		
<b>Protein Id :</b>	<input type="text" value="Psy_5123"/>		
<b>External database Id :</b>	<input type="text" value="YP_238188.1"/>		
<b>Species / Strain Name :</b>	<input type="text" value="Pseudomonas syringae pv. syringae B728a"/>		
<b>Species Abbreviation :</b>	<input type="text" value="PsyB728a"/>		
<b>Genome ID :</b>	<input type="text" value="NC_007005.1"/>		
<b>Genome Type :</b>	<input type="text" value="Chromosome"/>		
<b>Genome Length :</b>	<input type="text" value="6093698"/>		
<b>Index on Genome :</b>	<input type="text" value="5249"/>		
<b>Feature Start :</b>	<input type="text" value="6081466"/>		
<b>Feature End :</b>	<input type="text" value="6082002"/>		
<b>Strand :</b>	<input type="button" value="+ ▲"/> <input type="button" value="▼ -"/>		
<b>Frame :</b>	<input type="text" value="1"/>		
<b>Gene Name :</b>	<input type="text" value="Not available"/>		
<b>Feature Type :</b>	<input type="text" value="CDS"/>		
<b>Protein Length :</b>	<input type="text" value="179"/>		
<b>Nucleotide Length :</b>	<input type="text" value="537"/>		
<b>Product Description :</b>			
<input type="text" value="ATP synthase subunit delta"/> <text style="font-size: small; margin-top: 5px;">//</text>			
<b>Homolog Group :</b>	<input type="text" value="Cluster_10212"/>		
<b>Ortholog Group :</b>	<input type="text" value="Cluster_10212.1"/>		
<b>HMM Group :</b>	<input type="text" value="Group2510"/>		
<b>GO Id :</b>	<input type="text" value="GO:0005886"/> <b>GO Category :</b> <input type="text" value="Cellular Component"/> <b>GO Description :</b> <input type="text" value="The membrane surrounding a cell that separates the cell from its external environment"/>		
<b>Pfam Id :</b>	<input type="text" value="PF00006"/> <b>Pfam Name :</b> <input type="text" value="ATP-synt_ab"/> <b>Pfam Description :</b> <input type="text" value="ATP synthase alpha/beta family, nucleotide-binding domain"/>		
<b>Pfam Start :</b>	<input type="text" value="150"/>	<b>Pfam End :</b> <input type="text" value="376"/>	<b>Significance :</b> <input type="text" value="1e-10"/>
<b>InterPro Id :</b>	<input type="text" value="IPR020003"/>	<b>InterPro Description :</b> <input type="text" value="ATPase, alpha/beta subunit, nucleotide-binding domain, active site"/>	
<b>Pathway Id :</b>	<input type="text" value="ko00190"/>	<b>Pathway Description :</b> <input type="text" value="Oxidative phosphorylation"/>	

- Click Submit button at the bottom of edit page to store edited information in the database.
- Refresh gene detail information page to display edited information.