USER GUIDE FOR DENOVO GENOMICS PIPELINE

Contents

Introduction

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

Installation

The package for DeNoGAP is available at the github site https://github.com/DSGlab/DeNoGAP.

Directory structure of DeNoGAP package:

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

cgi-bin: contains cgi scripts for connecting and accessing database using GUI.

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

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

exe: directory for installation of external programs.

html: contains html page for graphical user interface to connect with the database.

data: directory to store input data files.

output: directory to store output data / results.

doc: contains manual for setting up and using DeNoGAP.

SYSTEM REQUIREMENT

Operating system: Linux

RAM: at least 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 try using ‘cygwin’ to run DeNoGAP. We would release system-specific version in near future.

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

BioPerl modules:

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

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)

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

Users can either manually install each of the programs or run “install.pl” script under DeNoGAP for installation.

DeNoGAP do not install Apache. Users need to manually install apache server and configure it for running GUI for DeNoGAP database exploration.

Copy **cgi-bin** and **html** directories into the Apache document root directory.

To install required programs, execute following commands in terminal:

cd DeNoGAP

cd bin

perl install.pl <install_directory>

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 may need appropriate permissions from user for installation.

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.

(1) Genome Information File

The first line of the genome information file should start with “#” followed by tab-delimited column names.

- **Mandatory Column names**
  - **genome_name** : Full genome name.
  - **species**: Full name of the species**.**
  - **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.

(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 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 Phmmer 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 phmmer program. Check available options for phmmer 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).
- **CLUSTER_INDEX:** Define index value for naming the hmm family cluster file. (Default value: 1).

▪ **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:** Define complete path and file name of the seed HMM model database file.
- **SINGLETON_DB_FILE:** Define complete path and file name of the seed Singleton sequence database file.
- **HMM_FAMILY_OUTFILE:** Define complete path and file name for the final HMM model database output file.
- **SUPER_HOMOLOG_OUTFILE:** Define complete path and file name for the super-homolog cluster output 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 phmmer program. Check available options for hmmscan and phmmer 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).
- **CLUSTER_INDEX:** Define index value for naming the hmm family cluster file. (Default value: 2).

▪ **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.
- **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.
  - **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 for predicting ortholog pairs in case out-group is absent. (Default value: 0.8).

- **INPARALOG_DIVERGENCE_THRESHOLD:** Define distance cut-off for predicting inparalog pairs in case out-group is absent. (Default value: 0.5).
- **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).
  - **CLUSTER_FILE:** Define complete path and name of the protein family cluster file.
  - **GROUP_PROFILE:** Define complete path and name of the output profile file.
  - **GROUP_PROTEIN_TAB:** Define complete path and name of the tab-delimited list of protein sequence identifiers sorted by protein families.
  - **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).
  - **CLUSTER_FILE:** Define complete path and name of the protein family cluster 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).

(3) SQLite Database Schema

DeNoGAP uses SQLite database to store analyzed information. DeNoGAP creates 19 database tables to store results from various analysis phases. The description of each table and its column is given below:

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
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 -)
index_on_genome	INT	Order on the genome sequence
description	TEXT	Product description

Table: ProteinSequence		
Column	**Data type**	**Description**
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

Table: NucleotideSequence		
Column	**Data type**	**Description**
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

Table: Similarity		
Column	**Data type**	**Description**
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: LinkFamily		
Column	**Data type**	**Description**
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

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: 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		
Column	**Data type**	**Description**
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		
Column	**Data type**	**Description**
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		
Column	**Data type**	**Description**
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: PhylogeneticProfile		
Column	**Data type**	**Description**
id	TEXT	Ortholog family id
genome_name	TEXT	Genome abbreviation

Table: MapGeneIdtoGeneFamily		
Column	**Data type**	**Description**
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

RUNNING DeNoGAP

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** <sqlite database dir path> **-db_name** <name of the sqlite database> **-config** <path and name of configuration file> **-output_dir** <path 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.

- **Test parse genbank files**

  perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/PARSE_GENBANK.config –output_dir ../output/TEST_RUN

- **Test gene prediction**

  perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/PREDICT_GENE.config –output_dir ../output/TEST_RUN

- **Test gene verification**

  - Download and extract UniProt database. (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz)

  - Merge sequences predicted by multiple programs and single program for all genomes in a single sequence file using cat command.

- Perform protein Blast and store output in default format.
  blastp –query <All_predicted_sequence_file> -db <Uniprot_database_file> -evalue 1e-05 –out <blast_output_file> -max_target_seq 1
- Perform gene verification analysis using DeNoGAP. Change path and name of Blast alignment file in configuration file to match user-defined blast output file.

perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/GENE_VERIFICATION.config –output_dir ../output/TEST_RUN

- **Test Load Data**

perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/LOAD_DATA.config –output_dir ../output/TEST_RUN

- **Test Reference comparision**

perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/COMPARE_REFRENCE.config –output_dir ../output/TEST_RUN

- **Test Iterative HMM family prediction**

perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/PREDICT_HMM_FAMILY.config –output_dir ../output/TEST_RUN

- **Test Super-homolog family prediction**

perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/PREDICT_SUPER_HOMOLOG.config –output_dir ../output/TEST_RUN

- **Test Ortholog Prediction and Ortholog Clustering**

perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/PREDICT_ORTHOLOG.config –output_dir ../output/TEST_RUN

- **Test Phylogenetic profile**

  perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/PHYLOGENETIC_PROFILE.config –output_dir ../output/TEST_RUN

- **Test Core genome prediction**

  perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/CORE_GENOME.config -output_dir ../output/TEST_RUN

- **Test Function Annotation prediction**

  perl DeNoGap_v1.0.pl -genome_info ../data/test1/test_genome_table_1.txt –db_dir ../output/TEST_RUN –db_name test.sqlite –config ../config/ANNOTATION.config -output_dir ../output/TEST_RUN

OUTPUT

This section given as 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 SQLite database.

Analysis: Parse GenBank

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

- **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.

- **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 fasta formatted coding gene sequences (one file for each organism). Coding sequences predicted by multiple programs and single program for each genome are stored in separate sub-directories respectively.
- **Protein sequence directory**: This directory contains fasta formatted translated protein sequences (one file for each organism). Protein sequences predicted by multiple programs and single program for each genome are stored in separate sub-directories respectively.
- **Genomic feature directory**: This directory contains tab-delimited genomic feature files (one file for each organism). Genomic features predicted by multiple programs and single program for each genome are stored in separate sub-directories respectively.

Analysis: Gene prediction verification and annotation

The output directory for gene prediction verification stores output files for verified Coding sequences, protein sequences and their genomic features. It also stores genbank files for each genome with verified sequences.

- **VERIFY_CDS:** This sub-directory stores verified coding sequences for each genome (one file for each genome).
- **VERIFY_PROTEIN:** This sub-directory stores verified protein sequences for each genome (one file for each genome).
- **VERIFY_GENOME_FEATURE:** This sub-directory stores genomic features for verified sequences (one file for each genome).
- **VERIFY_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 un-parsed hmmscan and phmmer output files in alignment format and tabular format under HMM_FULL and HMM_DOM folders respectively (one file for each organism).
- **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 sub-directory stores files for HMM models and singleton sequences for protein families MODEL and SINGLETON sub-directories respectively (one file for each protein family).
- **HMM_DB:** This sub-directory stores HMM model database files and Singleton sequence database file.
- **SUPER_HOMOLOG:** This sub-directory stores output files from super-homolog family prediction.
- **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).
- **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.

- **PROJECT_DIR_NAME:** 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.

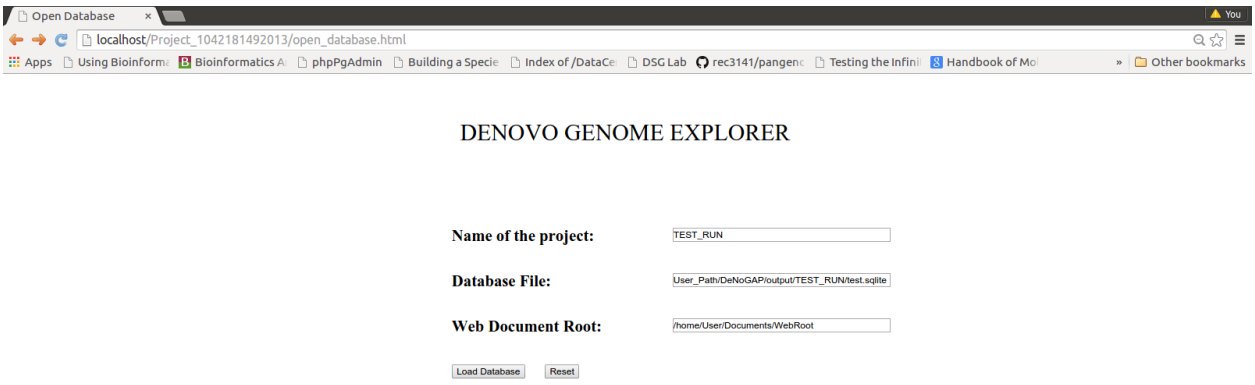
- **INTERPRO_SCAN_OUTPUT:** This directory stores all the output files and sub-directories for interproscan annotation.
- **XML_FILE:** This sub-directory stores un-parsed InterProScan output files in XML format (one file for each genome).
- **TAB_FILE:** This sub-directory stores un-parsed InterProScan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

Graphical Interface for DeNoGAP database exploration

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

- Open http://localhost/html/open_database.html in the Web Browser



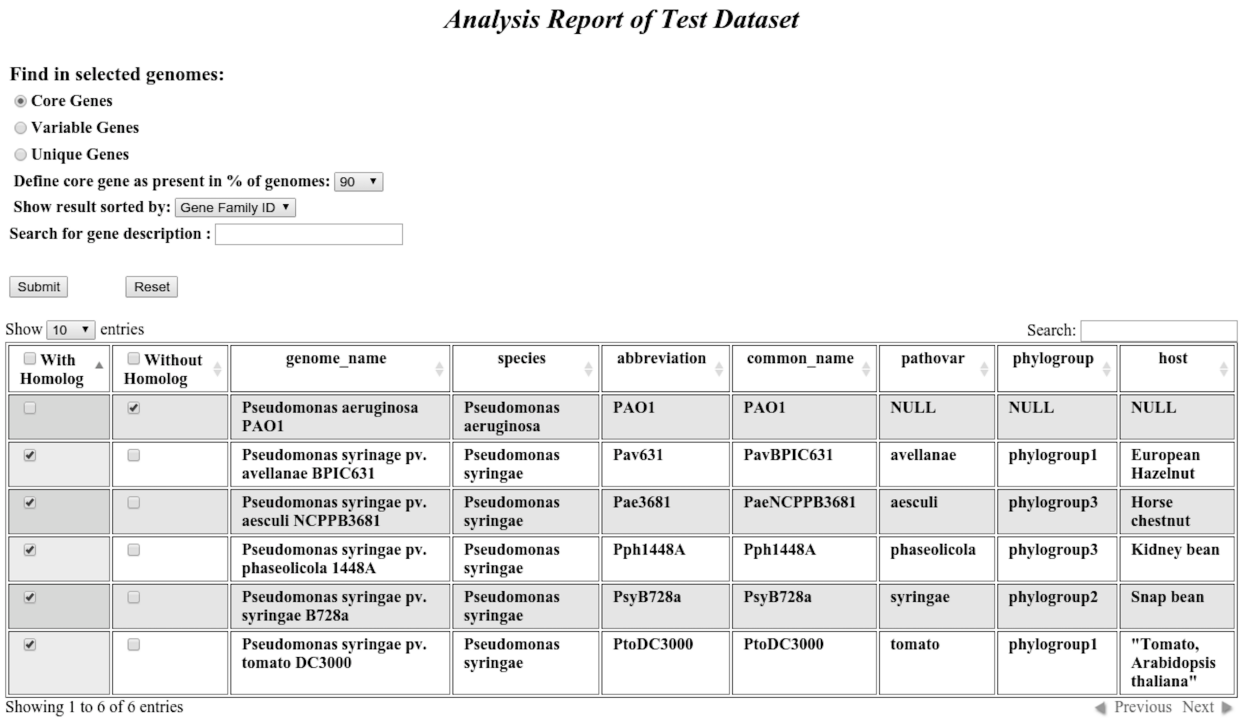
- Enter name of the Project:
  TEST_RUN

- Enter full path and name of the SQLite database file:
  /home/~User_path/DeNoGAP/output/TEST_RUN/test.sqlite

- 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 Dataset

Find in selected genomes:

- (•) **Core Genes**
- ( ) **Variable Genes**
- ( ) **Unique Genes**

Define core gene as present in % of genomes: 90

Show result sorted by: Gene Family ID

Search for gene description :

Submit Reset

Show 10 entries Search:

With Homolog	Without Homolog	genome_name	species	abbreviation	common_name	pathovar	phylogroup	host
☐	☑	Pseudomonas aeruginosa PAO1	Pseudomonas aeruginosa	PAO1	PAO1	NULL	NULL	NULL
☑	☐	Pseudomonas syrinage pv. avellanae BPIC631	Pseudomonas syringae	Pav631	PavBPIC631	avellanae	phylogroup1	European Hazelnut
☑	☐	Pseudomonas syringae pv. aesculi NCPPB3681	Pseudomonas syringae	Pae3681	PaeNCPPB3681	aesculi	phylogroup3	Horse chestnut
☑	☐	Pseudomonas syringae pv. phaseolicola 1448A	Pseudomonas syringae	Pph1448A	Pph1448A	phaseolicola	phylogroup3	Kidney bean
☑	☐	Pseudomonas syringae pv. syringae B728a	Pseudomonas syringae	PsyB728a	PsyB728a	syringae	phylogroup2	Snap bean
☑	☐	Pseudomonas syringae pv. tomato DC3000	Pseudomonas syringae	PtoDC3000	PtoDC3000	tomato	phylogroup1	"Tomato, Arabidopsis thaliana"

Showing 1 to 6 of 6 entries Previous 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

Result of Analysis

2371 shown in result

Filter row with term:

Genome Name: Pseudomonas syringae pv. tomato DC3000

Show Gene Information

Group ID	Gene ID	Gene Description	Genes per Genome	Details
orthoCluster_102.1	PSPTO_3556	Glycolate oxidase, subunit GlcE	2	
orthoCluster_103.1	PSPTO_5125	Uncharacterized protein	2	
orthoCluster_104.1	PSPTO_4243	Urea amidolyase-related protein	2	
orthoCluster_105.1	PSPTO_4936	Methyl-accepting chemotaxis protein	2	
orthoCluster_106.1	PSPTO_2481	TonB protein	2	
orthoCluster_107.1	PSPTO_2980	Oxidoreductase, 2OG-Fe(II) oxygenase family	2	
orthoCluster_108.1	PSPTO_3813	Uncharacterized protein	1	
orthoCluster_109.1	PSPTO_3460	D-isomer specific 2-hydroxyacid dehydrogenase family protein	2	
orthoCluster_110.1	PSPTO_2442	CheW domain protein	1	
orthoCluster_111.1	PSPTO_0114	GGDEF domain/EAL domain protein	2	
orthoCluster_112.1	PSPTO_3682	Uncharacterized protein	2	
orthoCluster_113.1	PSPTO_2615	GAF domain protein	1	
orthoCluster_114.1	PSPTO_5166	Membrane protein, putative	1	
orthoCluster_115.1	PSPTO_2685	Shikimate 5-dehydrogenase, putative	1	
orthoCluster_116.1	PSPTO_2177	2-dehydro-3-deoxygalactonate kinase	1	
orthoCluster_117.1	PSPTO_1790	Acyl-CoA dehydrogenase family protein	2	
orthoCluster_118.1	PSPTO_5257	Uncharacterized protein	1	
orthoCluster_119.1	PSPTO_5408	Uncharacterized protein	2	
orthoCluster_120.1	PSPTO_3841	Ribonuclease E	2	
orthoCluster_121.1	PSPTO_4289	Uncharacterized protein	1	
orthoCluster_122.1	PSPTO_4925	Uncharacterized protein	1	
orthoCluster_123.1	PSPTO_2480	Methyl-accepting chemotaxis protein	1	
orthoCluster_124.1	PSPTO_4373	Sensor histidine kinase ColS	2	
orthoCluster_125.1	PSPTO_3597	Uncharacterized protein	2	
orthoCluster_126.1	PSPTO_0941	Tellurium resistance protein TerA	2	
orthoCluster_127.1	PSPTO_4847	Penicillin-binding protein	1	
orthoCluster_128.1	PSPTO_3494	Inositol 2-dehydrogenase	2	
orthoCluster_129.1	PSPTO_4536	Peptide ABC transporter, ATP-binding protein	2	
orthoCluster_13.1	PSPTO_2148	Pyoverdine sidechain peptide synthetase IV, D-Asp-L-Ser component	5	
orthoCluster_13.2	PSPTO_2149	Pyoverdine sidechain peptide synthetase III, L-Thr-L-Ser component	2	
orthoCluster_130.1	PSPTO_1639	tRNA 5-methylaminomethyl-2-thiouridine biosynthesis bifunctional protein MnmC	1	

- **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 gene family Id
- **Gene ID combo box:** List Inparalog gene Ids for each gene family. Select gene Id from the list to view detailed information.

- **Gene Description:** Show product description for selected gene Id.
- **Detail:** Select gene Id to display detailed information about the gene.

- Click “Show Information” to display detailed information for selected gene Id.

- Gene Detail Information

GENE : PSPTO_3556

Genomic Feature:	
Locus Tag:	PSPTO_3556
Species / Strain Name:	Pseudomonas syringae pv. tomato DC3000
Species Abbreviation:	PtoDC3000
Genome ID:	NC_004578
Genome Type:	chromosome
Genome Length:	6397126
Index on Genome:	3586
Genomic Location:	4013245 : 4014303 (+)
Gene Name:	Not available
Feature Type:	CDS
Protein Length:	353
Nucleotide Length:	1059
Product Description:	Glycolate oxidase, subunit GlcE
Comparative Genomics Information:	
Homolog Group:	Cluster_102
Ortholog Group:	orthoCluster_102.1
HMM Model Group:	Group4271
Annotation:	

GO Annotation:

GO ID	GO Category	GO Description
GO:0008762	MOLECULAR_FUNCTION	UDP-N-acetylmuramate dehydrogenase activity
GO:0050660	MOLECULAR_FUNCTION	flavin adenine dinucleotide binding
GO:0055114	BIOLOGICAL_PROCESS	oxidation-reduction process
GO:0016491	MOLECULAR_FUNCTION	oxidoreductase activity

PFam:

PFam ID	Pfam Name	Description	Start	End
PF01565	FAD_binding_4	FAD binding domain	16	136

InterPro:

InterPro ID	Description
IPR006094	FAD linked oxidase, N-terminal