

## User manual

### Gene Analysis

This code is comprised of two scripts. These are MAKE\_PANGENOME.R and PAN\_GWAS.R. make\_pangenome.r creates a pan-genome from a set of genomes or assemblies. PAN\_GWAS.R tests each gene in a specified pan-genome for association with a phenotype of interest.

#### MAKE\_PANGENOME.R

Constructs a pan-genome from a set of genomes or assembled contigs. Runs Prodigal to annotate genes and outputs a file containing all contigs annotated on all isolates. The script then runs CD-HIT to cluster annotated open reading frames. CD-HIT outputs a file containing the longest sequence in each cluster. The ids of these sequences are used as ids in the pan-genome. CD-HIT also outputs a file with the prefix ".clstr". This file contains more details of the sequences in each cluster.

Usage example:

```
Rscript /path/of/make_pangenome.r -contigFile contigs.txt -prodigal yes -  
similarity 0.7 -coverage 0.7 -prefix test -externalSoftwarePaths soft.txt
```

#### Inputs

##### **contigFile**

Tab-delimited file containing list of paths of genomes or assemblies in unzipped FASTA format. This file requires two columns, eg:

name	filePath
C00001200	/home/data/C00001200.velvet.fasta

**Note:** Genomes or assemblies need to be prefixed with unique, 9 character ids so that cluster names can be generated easily.

##### **prodigal**

Runs Prodigal (<https://github.com/hyattpd/Prodigal>) to annotate genes on assemblies (yes or no). The default value is yes.

##### **similarity**

Similarity threshold to use for clustering open reading frames with CD-HIT ([http://weizhong-lab.ucsd.edu/CD-HIT/wiki/doku.php?id=CD-HIT\\_user\\_guide](http://weizhong-lab.ucsd.edu/CD-HIT/wiki/doku.php?id=CD-HIT_user_guide)). Takes a value between 0.4 and 1.0.

##### **coverage**

Coverage threshold to use for clustering open reading frames with CD-HIT. Takes a value between 0.4 and 1.0. Defaults to 1.0 if not specified.

##### **prefix**

Prefix for output files.

##### **externalSoftwarePaths**

A tab delimited file containing the name and paths of the external software used in the analysis. This file should contain two columns with the headers *name* and *path*, which specify the name and path of each software package required. Users must install all dependencies prior to running the package.

For `make_pangenome.r`, the following dependencies are required, with the spellings below:

Prodigal  
CD-HIT  
makeblastdb  
blast

`make_pangeome.r` also requires the following R packages:  
Seqinr

### **scriptDir**

Directory in which the scripts called by `PAN_GWAS.R` are located. Default is the current working directory.

`make_pangenome.r` calls the following other R scripts:  
`parse_ncbi_protein.R`  
`annotate_clusters.R`

### **createDb**

Option to create blast databases and/or NCBI protein look-up table. This option requires either TRUE or FALSE to be specified; the default input is TRUE. If `createDb` is set to TRUE the user can input paths to create up to two BLAST databases and an NCBI protein look-up table.

### **db1**

Path to first BLAST database. This is the path to the output of `makeblastdb`, but excluding the extensions, eg, the directory containing the output of `makeblastdb` will include the following files: `db1.phr`, `db1.psq`, `db1.pin` and the input for `db1` will be `db1`. **db1** is required if **refseqFasta1** is not specified and **createDb** is set to FALSE.

### **db2**

Optional path to second BLAST database. This option is specified the same as **db1**.

### **refseqFasta1**

Path to file listing paths of FASTA files used to create first database. This is only needed if **createDb** is set to TRUE and **db1** is not specified.

### **refseqFasta2**

Path to file listing paths of FASTA files for optional creation of second BLAST database.

### **ncbiAnnot**

Path to NCBI protein look-up table. This option is needed if **ncbiSummary** is not specified and **createDb** is set to FALSE.

### **ncbiSummary**

Path to protein summary file downloaded from NCBI.

### **Outputs**

prefix.pangenome.fasta

FASTA file containing sequences of the longest sequence in each CD-HIT cluster, annotated with NCBI protein accessions.

Prefix.pangenome.varGenes

Tab-delimited file containing a matrix indicating the presence or absence of the gene clusters in each isolate. Column headers are genome names and row headers are gene cluster names.

prefix.clstr

CD-HIT output file listing all the sequences in each cluster.

prefix\_allprot.faa

File containing sequences of all the ORFs annotated by Prodigal, as annotated by Prodigal.

### **PAN\_GWAS.R**

Runs GWAS on a set of phenotypes using logistic regression and, optionally, corrects for population structure using Gemma. Creates Manhattan plots for each GWAS.

**Note:** For the Gemma analysis, a relatedness matrix based on the SNPs is required, so the SNP GWAS pipeline needs to be run first.

Usage example:

```
Rscript /path/of/PAN_GWAS.R -pangenome pangenome.varGenes -phenotype  
pheno.txt -gemma yes -relm relm.txt -externalSoftwarePaths soft.txt -script_dir  
/path/to/scripts
```

### **pangenome**

The path to a tab-delimited file containing pan-genome in the following format. Row names are gene cluster ids, column names are genome ids, cells denote presence (1) or absence (0) of each gene in each isolate. The prefix\_pangenome.varGenes file output by MAKE\_PANGENOME.R works as this input.

### **phenotype**

This specifies the path of a tab-delimited text file containing a table of phenotype data. This file requires at least two columns:

1. Unique ids for each isolate with header *name*
2. A column of phenotypes for each trait of interest. This column contains a binary phenotype for each trait in each isolate where "0"= indicates that

the isolate is a control and “1” indicates that the isolate is a case.  
Phenotype columns have the name of the trait as their header, eg:

name	trait1	trait2	trait3
genome1	0	1	1
genome2	1	0	1

### **gemma**

Option to run GEMMA (yes or no). Default=no.

### **relm**

Path to file containing the GEMMA relatedness matrix calculated on reference-based SNPs in the same dataset.

### **scriptDir**

Directory in which the scripts called by PAN\_GWAS.R are located. Default is the current working directory.

PAN\_GWAS.R calls the following scripts from the GWAS pipeline:  
do\_logreg\_chw.R

### **externalSoftwarePaths**

A tab delimited file containing the name and paths of the external software used in the analysis. This file should contain two columns with the headers *name* and *path*, which specify the name and path of each software package required. Users must install all dependencies prior to running the package.

For PAN\_GWAS.R the following dependencies are required, with the spellings below:

GEMMA

### **Outputs**

For each phenotype files with the following suffixes are output:

\_SNP\_logreg\_output.txt

Results from the uncorrected logistic regression GWAS

\_LMM\_corrected.txt

Results corrected for population structure using GEMMA (if GEMMA was run) and  $-\log_{10}$  p-values corrected.

\_logreg\_manhattan.png

Manhattan plot for uncorrected GWAS.

\_lmm\_manhattan.png

Manhattan plot for GEMMA-corrected GWAS.