MetaGeniE Documentation

MetaGenome Explorer (MetaGeniE) is a distributed infrastructure to handle high number of metagenome sequences and accurately identify infections even to species/strain from clinical/metagenome samples.

PREREQUISITES:

BWA 0.7.10
STAMPY 1.0.17
SAMTOOLS 1.1
BLAT
PRINSEQ 0.19.3
PERL 5.14.2
PERL Module: Parallel Manager (PERL)
PYTHON 2.7.3
BEDTOOLS v2.25
BLAST-2.2.17 (Only Fastacmd and Formatdb are required)

INSTALLATION

1. Download copy of MetaGeniE.
The root folder has following:
    a. metagenie.pl: wrapper
    b. iconfig.pm: is configuration file required to set the paths and variables
    c. bin folder consists of MetaGeniE perl programs
    d. scripts folder consists of bash scripts
    e. helper_scripts folder consists of perl programs for database formatting and other utilities
    f. external folder consists of few dependencies required by MetaGeniE like Blat, Prinseq. Please download others for MetaGeniE to run successfully.
    g. Test folder: Assist in pre-installation
2. Create a folder and Copy/link the sequence read file (ln -s sequence_file_name). Also copy iconfig.pm already bundled in MetaGeniE in the same folder. Cat the input sequence file (example zcat).
3. Set the variables and paths for executables, databases and dependencies in iconfig.pm. See "How To" for more information.
Note: The iconfig.pm should be copied/present in the same folder as metagenome sequence files (recommended) or should be in the PATH.
5. CD into the metagenome sequence folder and run metagenie.pl with the required parameter (LSF/PBS scripts available).
6. For analysis, see the respective temp folder created for example: tmp_bwa_bacteria/tmp_blat_bacteria will have *_SUMMARY file for the detected pathogens. You can sort it preferably by genome coverage or other options like %genome coverage etc.
The unmapped reads remaining after running each module will be named as such example: *_readReduct.fasta and *_pathoDetect.fasta.
The logs are generated in log folder example log_stats file has the breakdown of the filtration and alignment statistics.

HOW TO:

1. Running different module of MetaGeniE:

The option to run Read-Reduct and Patho-Detect can be set up from instructions: metagenie.pl -man (metagenie.pl -help).

Example if Read-Reduct module is turn off, there is no need to set any database/executable/option related to this module (see iconfig.pm for details).

2. Setting the variables and paths for executables and databases:

MetaGeniE uses Variables and Options. PLEASE DO NOT DELETE ANY VARIABLES/OPTIONS. This might have adverse effect on the MetaGeniE. Variables/database setup can be turned off/on with corresponding option.

Example: To turn on search against Bacterial database, set the path for bacterial database folder and set $run_bacdb="y" in the iconfig.pm.

If you do not want to use Human database, set $run_ref_hg19="n". If you turn off any option, then you do not need to set path for this option. Options can only use following values: (y/Y or n/N)

3. Setting the database is described in section "How to set database from public resources"

4. Indexing the database. See section "Indexing the database"

HOW TO SET DATABASE FROM PUBLIC RESOURCES

A. Set the human database (See DATABASE SOURCES).

Human Database is used only in Read-Reduct Module:

1. Download the human database
2. Cat/merge all the chromosomes to single fasta file
3. Index database file example

bwa index -a bwtsw -p hs_ref_GRCh37_p5 hs_ref_GRCh37_p5.fa

4. Only Human database requires STAMPY indexing example

stampy.py --species=human --assembly=hs_ref_GRCh37_p5 -G hs_ref_GRCh37_p5 hs_ref_GRCh37_p5.fa

stampy.py -g hs_ref_GRCh37_p5 -H hs_ref_GRCh37_p5

Note: Stampy usage not recommended for faster analysis and should be used if you want to remove higher % of the divergent human DNA.

5. Repeat Database is already bundled with MetaGeniE or latest database can be downloaded from Repbase (Repeat database does not require indexing).

B. Set Microbial database (bacteria, viral and fungal) (See DATABASE SOURCES).

This database is used only in Patho-Detect Module:

1. Multiple fragments of reference database can be set up for example setting database for Complete Bacterial genomes from NCBI:

wget ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz

The downloaded file needs to be split into multiple partition (~1GB to 3GB depending on available memory) for memory management (see unix split command).

2. Format headers of the split reference database fasta files:

for i in {1..x}; do perl reformatHeader.pl microbial_file_$i.fna microbial_file_$i_rf; done;

scripts are available in helper_scripts folder in MetaGeniE downloadable bundle. Do not use file extension for new files created/existing file. See #5 below.

3. Generate genome file for final summary report:

for i in {1..x}; do perl cntFastaSeq.pl microbial_file_$i_rf microbial_file_$i.genSize; done;

4. Join the file:
*cat *genSize > GenomeDesc;*
(Note: The final file name should be GenomeDesc)
5. Index reference database as formatted above:
bwa index -a bwtsw microbial_file_$i_rf
(for larger number of files, set pbs script.
IMPORTANT: Please do not use any file extension (like *.fna/*.fasta) for the microbial database file. This allows BLAT to automatically pick multiple fragments of the database from BWA indexes.

DATABASE SOURCES
Human Hg19
ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/
Human Korean Genome
ftp://ftp.kobic.kr/pub/KOBIC-KoreanGenome/
Human Chinese Genome ftp://public.genomics.org.cn/BGI/yanhuang/fa/
Bacterial Genome ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz
Viral Genome ftp://ftp.ncbi.nih.gov/refseq/release/viral/
Fungi Genome ftp://ftp.ncbi.nih.gov/refseq/release/fungi/

TESTING
Test database and Instructions to run are available in the MetaGeniE downloadable bundle.
BUG REPORTS ----------
Please report your bug/requirements at the email
<metagenie.dev_atr_gmail_dot_com>.