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

PREREQUISITES:
BWA 0.6.1-r104
STAMPY 1.0.17
SAMTOOLS 0.1.18
BLAT
PRINSEQ 0.19.3
PERL 5.14.2
PERL Module: Parallel Manager (PERL)
PYTHON 2.7.3
BEDTOOLS v2.16.2
BLAST-2.2.17 (Only Fastacmd and
Formatdb are required)

INSTALLATION
1. Download copy of MetaGeniE. The root folder has wrapper metagenie.pl, four folders and iconfig.pm:
        a- bin folder consist of MetaGeniE perl programs
        b- scripts folder consist of bash scripts
        c- helper_scripts folder consist of perl programs for database formatting and other
utilities
        d- external folder consist of few dependencies required by MetaGeniE like blat, bedtools, prinseq and blast program (formatdb and fastacmd). Please download others for MetaGeniE to run successfully.
        e- iconfig.pm is configuration file required to set the paths and variables

2. Create a folder and Copy/link (ln -s sequence_file_name) the sequence read file (or in the metagenome sequencing file folder), copy iconfig.pm already bundled in MetaGeniE.

3. Set the variables and paths for executables, databases and dependencies in iconfig.pm. See "How To" for more information.

4. 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.

6. For analysis, see the respective temp folder created for example tmp_blat_bacteria will have *_SUMMARY file for the detected pathogens. 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
$run_bacdb="y"
and set the path for bacterial database folder. 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. Option 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

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 database can
be set up for querying
example (setting database for Release 57 of RefSeq Microbial dataset) :
for i in {1..97}; do wget
ftp://ftp.ncbi.nih.gov/refseq/release/microbial/microbial.$i.1.genomic.
fna.gz; done;
2. for i in {1..97}; do perl reformatHeader.pl
microbial.$i.1.genomic.fna microbial.$i.1_rf; done; # scripts are available in helper_scripts folder in MetaGeniE downloadable bundle. Do
not use file extension for new files created/existing file. See #5 below.
3. for i in {1..97}; do perl cntFastaSeq.pl microbial.$i.1_rf microbial.$i.genSize; done;
4. cat *genSize > GenomeDesc; (Note: The final file name should be GenomeDesc)
5. Index database file example
bwa index -a bwtsw microbial.94.1_rf (for larger number of files, set pbs
script or use: for i in {1..97}; do bwa index -a bwtsw microbial_file; done;).
IMPORTANT: Please do not use any file extension (*.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.nih.gov/refseq/release/microbial/
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 <rawat.arun_atr_gmail_dot_com>.