**README:**

**Analysis Pipeline**

1. **Analysis pipeline for Illumina pair-end reads:**

Procedure to run Enrichment Analysis Pipeline:

Requirements:

bin folder = Contains all the scripts

src folder = Contains all the softwares

EnrichmentAnalysisPipeline.py

run\_QC.sh

Before you run the scripts export the folders

export PATH=pathtofolder/bin:$PATH

export PATH=pathtofolder/src:$PATH

***I) Run QC***

Have all the raw fastq files in a single folder and run the following script. The script uses fastqc and trim\_galore software to do the quality filter of 20 and minimum bp as 35. Make sure the software is imported before you run the script.

./run\_QC.sh 'Raw fastq files directory'

**Find Unknown species:**

./FindSpeciesInSample.py 'Fastq directory which are quality checked' 'NCBI nt database location'

Export the XML file in Megan software

***II) Find consensus genome, mapping and snps***

The "EnrichmentAnalysisPipeline.py" is a wrapper script which depending upon the parameters calls either

1) EnrichmentAnalysis\_NotSegmented.py

2) EnrichmentAnalysis\_Segmented.py

The sample information and the fasta file to be used as a reference is given in a csv file:

CD-Lib-0004\_S4\_L001\_R1\_001\_val\_1.fq,Rotv-CU938.fa

CD-Lib-0005\_S5\_L001\_R1\_001\_val\_1.fq,Rotv-YR062.fa

where, CD-Lib-0004\_S4\_L001\_R1\_001\_val\_1.fq is the forward read fastq file present in Fastq directory and Rotv-CU938.fa is the reference file with 11 segments in Reference directory. Please note that only the forward read fastq file should be given in the file, the code will find its pair by searching "R2" instead of R1.

For parallel run, for every fastq file in the csv file a ‘screen’ is created and the pipeline is run. You can check the jobs running by screen -ls command.

***Usage: EnrichmentPipeline.py [options]***

*Example: ./EnrichmentPipeline.py -r `pwd`/Reference -q `pwd`/Fastq -e umasang7@gmail.com -d /share/ncbi/nt -o `pwd`/Test -i `pwd`/run\_summary-tmp.csv -p F -s F -b `pwd` -t stringent*

*Options:*

*-h, --help show this help message and exit*

*-r REF\_DIR, --refdir=REF\_DIR*

*Directory of Reference fasta file*

*-q FQ\_DIR, --fastq\_dir=FQ\_DIR*

*Directory of Fastq files*

*-e EM, --email\_address=EM*

*Email address to access online NCBI*

*-d DB, --database=DB NCBI nt database*

*-o OUT\_DIR, --outdir=OUT\_DIR*

*Output directory*

*-i CSV\_FILE, --info=CSV\_FILE*

*csv file containing the sample and reference information*

*-p PAR, --parallel=PAR*

*Run each fastq file in the info file in Parallel or not. Takes T or F*

*-s SEG, --segment\_type=SEG*

*Virus in the info file in Segmented or not. Takes T or F*

*-b CODE\_DIR, --code\_dir=CODE\_DIR*

*The directory were the scripts are present*

*-t PARAM, --stringency=PARAM*

*Stringency for the consensus genome generation: Takes*

*value 'stringent' or 'lenient'. stringent = MIN\_COV 10*

*and Mismatches bwa defualt; lenient = minimum coverage*

*1 and 20bp mismatch*

Output:

The pipeline outputs,

i) Consensus genome generated: \*-hybrid.fa and \*-hybrid.annot

ii) SNPs found: \*.vcf

iii) Mapped bam file: \*-sort.bam

iv) Per position coverage: \*.cov

***III) Generate graphs***

\*\*\*\* This script is not packaged. Hence please edit the parameters and values within the script. For graph generation Circos software is used.

runGraphs.py is the wrapper script that calls generated\_circosPlot.py (to generate the graph) and bait\_effiency\_modules.py (For plotting the baits position)

Files required for Graph..

Sample related:

* Create a new directory and copy all the \*vcf , \*-hybrid.fa , \*.cov from the previous output.

Genome related:

* Rotv-CU938.coords (Gives the genome structure [represent as single chromosome or segmented])
* Rotv-CU938.genes (Gives the names of the segments or if the genome is split into genes then gene names)
* Rotv-CU938.baits: If it’s a capture and baits are present, this file should contain the baits in fasta format. The baits are then mapped to the \*hybrid.fa and positions are obtained. If it’s the Library, please make an empty file with name is Rotv-CU938.baits.

Software related:

* Copy all the \*.conf files from the ./bin folder to this directory. The “hist.conf” is the configuration file for the layers in the plot (e.g. Coverage layer, baits layer, snps layer etc..), change the display parameters in this if needed. ideogram.conf gives the config for the outer genome display.
* Copy the log\_axis.txt from the ./bin folder and edit it accordingly. By default the log axis is marked from 10^1 to 10^5. The changes for this is also in the hist.conf

\*\*\*\* Please make sure the consensus genome fasta file headers should be used as the genome or segment name in the graph related files (e.g. Rotv-CU938.coords).