EIM User Manual

What is EIM?

EIM pipeline is a tool for improving the genomes reconstructed by reference-based assembly in terms of the number of mismatch and indel errors as well as the number of IUPAC codes. Currently, EIM supports three mappers: Bowtie2, BWA and Yara

How to use EIM?

Only Linux is supported. EIM pipeline has been tested on Ubuntu 15.10

Building EIM

First download the source package from http://bioinformatics.aut.ac.ir/EIM/.

Then extract EIM.tar.gz, 'cd' to the resulting directory and run 'make' command.

The package dependencies

EIM pipeline requires at least one of three mappers: Bowtie2, BWA and Yara. Thus a mapper has to be installed and added to the system path. These mappers can be downloaded from below addresses:

- Bowtie2 - 2.2.9 or newer

https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.2.9/

- BWA- 0.7.10 or newer

https://sourceforge.net/projects/bio-bwa/files/bwa-0.7.10.tar.bz2/download

- Yara sources are hosted on GitHub within the SeqAn library. Download the sources by executing:

```
$ git clone https://github.com/seqan/seqan.git
```

EIM applies SAMtools for making consensus sequences. Current version of EIM is compatible with version1.3 of SAMtools that is accessible from:

https://sourceforge.net/projects/samtools/files/samtools/1.3/

EIM uses Seqtk tool for processing sequences in the FASTA or FASTQ format. Download the sources by executing:

```
$ git clone https://github.com/lh3/seqtk.git
```

Running EIM

First create a folder with ‘data’ name inside ‘EIM-X.X’ directory and copy your read set and reference genome to it. Then ‘cd’ to ‘EIM-X.X’ directory and execute the script ‘runEIM.sh’ as follows:

```
$ ./runEIM.sh <mapper_name> <reads1> <reads2> <reference> <output_folder>
```

<mapper_name>

The name of read mapper by which the genome sequence is reconstructed and which is used in the pipeline for mapping. (Options: bowtie2, bwa and yara)

<reads1>

A fastq file containing first reads of a paired-end dataset.

<reads2>

A fastq file containing second reads of a paired-end dataset.

<reference>

A fasta file containing a reference genome sequence.

<output_folder>

A folder name for storing the temporary and final results of EIM pipeline.

Outputs

Usually upon completion of the successful running of EIM, the <output-folder> will contain some files including:

1. ‘contigs_mapping_X.fasta’
    - The contigs built by mapper X (bt: Bowtie2, bwa: BWA and yara: YARA)
2. ‘contigs_exact_sba.fasta’ or ‘contigs_exact_bt.fasta’
    - The contigs generated by the first step of EIM (Exact Mapping Step)

3. ‘contigs_EIM.fasta’
    - The final contigs generated by EIM pipeline.

Examples

First create a ‘data’ folder inside ‘EIM-X.X’ directory, download EcoliGenome.fa and one of the datasets such as ReadSet5 from http://bioinformatics.aut.ac.ir/EIM/, and extract them to ‘data’ folder. Then run one of following commands:

Run EIM with Bowtie2 mapper:

```
$ ./runEIM.sh bowtie2 readSet5.1.fastq readSet5.2.fastq EcoliGenome.fa  RS5_bt
```

Run EIM with BWA mapper:

```
$ ./runEIM.sh bwa readSet5.1.fastq readSet5.2.fastq EcoliGenome.fa RS5_bwa
```

Run EIM with Yara mapper:

```
$ ./runEIM.sh yara readSet5.1.fastq readSet5.2.fastq EcoliGenome.fa RS5_yara
```

Note: For large reference genome (> 20Mb), Bowtie2 mapper has to be applied for running EIM.