==

Mixture model for strains reconstruction

Author: XIN LI

Date: 04/12/2020

==

######Working OS#################

Our model is working on Linux Operating System.

###### Virtual Environment #########

Most Easy way to set up environment is using conda. A mixture.yml file was provided in the source directory.

(1) Install Conda:

https://docs.conda.io/projects/conda/en/latest/user-guide/install/

(2) Open a terminal in mixture model directory, create virtual environment by command:

```
conda env create -f ./mixture.yml
```

(3) Before running model, go into the virtual environment by command:

```
conda activate mixture
```

###### Common problem with Virtual Environment #########

After you install some package, if our model are running with error. It is mostly caused by the rpy2 package, you probably manually install R, and then install rpy2.

(1) Inside the environment created above, Conda install R

```
conda install -c r r
```

(2) Inside the environment created above, install rpy2 with specific version

```
pip install rpy2
```

###### Manual Install Without Virtual Environment #######

If you do not like conda or can not use it by conda for some reason, then in your own environment, the following prerequisite software with required version should be installed:

(1) python=3.7.6

If there is no python 3.7.6 installed, you can download and install python from (http://www.python.org/download/). You can use "python -V" command to check whether python is installed and the version of Python.

(2) Install R
https://www.r-project.org/ or install it by conda(https://anaconda.org/r/r)

(3) pysam=0.15.3

Detail to install pysam in the link: https://pysam.readthedocs.io/en/latest/installation.html. Or use three commands below one by one:

```
conda config --add channels r

conda config --add channels bioconda

conda install pysam
```

(4) rpy2==2.9.4

```
conda install rpy2
```

(5) jenkspy=0.1.5

```
pip install jenkspy
```

(6) scipy=1.4.1

```
conda install scipy=1.4.1
```

(7) Bowtie2 2.3 or newer (optional if you are only using our preprocessing script)

```
Conda install bowtie2
```

(8) Samtools 1.9 (optional if you are using our preprocessing script)

```
Conda install samtools=1.9 or https://anaconda.org/bioconda/samtools
```

===

###### Preprocessing

Since MixtureS require bam format input, you had better input the sorted Bam file as input. However, if you only have FastQ format data, you can get a hint from our common preprocessing pipeline. But please check your experimental specification before applying our preprocessing step, because our preprocessing may be not suitable for your specific experiments. So you may need to adjust our preprocessing script based on your own experiments or trim read protocol, like single-end read, trim customized bar code or others.

(1) Use conda virtual environment in **Virtual Environment:**

(2) Manually prerequisite

If you would like to use our preprocessing steps, there are two more additional prerequisite, see (7) and (8) in **Manual Install Without Virtual Environment**.

(3) Sample command:

```
python preprocessing.py --sample_name test --pair1 ./example_test_data/test.read1.fastq --
pair2 ./example_test_data/test.read2.fastq --process 6 --
genome_name ./example_test_data/GCF_000016525.1_ASM1652v1_genomic.fna --res_dir ./test_res_data
```

--sample_name: sample name

--pair1: forward read for FastQ

--pair2: reverse read for FastQ

--process: # of processor to run

--genome_name: reference genome

--res_dir : result directory

==

###### Parameters for Run Mixture model #######

(1) Go into virtual environment if you deploy by virtual environment

```
conda activate mixture
```

(2) Run our mixture model:

```
python ./mixture_model.py
```

--sample_name sample_name: Given a unique sample name for running on samples

--genome_len genomeLength: the corresponding genome length

--genome_name genomeName: the genome name(name inside FASTA file)

--genome_file_loc genomeLoc: genome FASTA file location. Absolute file location is better

--bam_file bam_file: the sorted bam file which have been mapped to the reference genome.

--res_dir directory_name: the result directory name

(3) sample input

The test data has been saved in the directory of example_test_data, so you can just navigate into directory where mixture_model.py located. And use below command to run example data. The result will save into directory test_result.

```
python mixture_model.py --sample_name test --genome_len 1853160 --genome_name NC_009515.1 --
genome_file_loc ./example_test_data/GCF_000016525.1_ASM1652v1_genomic.fna --
bam_file ./example_test_data/test_sorted.bam --res_dir ./test_result
```

==

###### Interpret results #######

1.sampleName_strains file: This is the final prediction of strains after combining k-mer results. It includes coverage with corresponding polymorphic sites. In the example below, there are two strains predicted. Each strain starts with '>' and its corresponding coverage. From next row, they are the

polymorphic sites of above strain. Each row is represented by allele and its position. In the example below, two strains with coverage 10.402205435195318 and 23.599084994628157. The first strain has six polymorphic sites, and the second strain has 5 polymorphic sites.

output example:

```
>10.402205435195318
C,807324
A,260641
G,869404
T,143273
G,902916
A,754450
>23.599084994628157
G,576282
G,1500940
C,162805
T,924389
C,1722675
```

2.Polymorphic sites:

First column is location, and followed by frequency of A, C, G ,T. For example, in the first row below, the location is 6696 with frequency of C 6 and frequency of G 175.

```
6696    0       9       175     0
7148    161     0       6       0
7589    0       0       4       203
10403   0       203     4       0
```

==

Contact Information

Please do not hesitate to reach out to me if you have questions.

Xin Li (xli042@knights.ucf.edu)

Haiyan Nancy Hu (haihu@cs.ucf.edu)

Xiaoman Shawn Li (xiaoman@mail.ucf.edu)