==========

BHap

Author: XIN LI

Date: August, 6, 2018

==========

Prerequisite Software

(1)Python 2.7 should be installed on your system.

First, you need to install python (recommend: python 2.7.x-64bit), which is already installed by default for most of current Linux systems. If there is no python 2.7 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)Python required package: Numpy 1.14.1 or higher Installation instructions for numpy (http://docs.scipy.org/doc/numpy/user/install.html)

---

python -m pip install --user numpy

---

(3)Python required package: networkx 2.1 or higher Installation instructions for networkx (https://networkx.github.io/documentation/stable/install.html)

---

pip install networkx

---

(4)Python required package: scipy 1.0.0(https://www.scipy.org/install.html)

---

python -m pip install --user scipy

==========

Parameters

python ./run_BHap.py

-d resName: result name, BHap will automatically generate a directory with this name, and stores the predicted results.

-r ReferenceGenome: the path to reference genome in FASTA format

-t ReadsType: read type, fastq is preferred

-1 ForwardSeq: path to forward seq

-2 ReverseSeq: path to reverse seq

-l ReadLen： Input data read length

-c Coverage: Input data coverage

-g GenomeLength: the length of the reference genome

==========
==========

Example

The test data has been saved in the directory of test_data, so you can just navigate into directory where run_BHap.py located. And use below command to run example data.

---

python ./run_BHap.py -d ./tmpName -r ./test_data/GCF_000016525.1_ASM1652v1_genomic.fna -t fastq -1 ./test_data/forward.fastq -2 ./test_data/reverse.fastq -l 100 -c 100 -g 1853160

==========
==========

Interpret results

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

polymorphic sites of above haplotype. Each row is represented by allele and its position. In the example below, two haplotypes with coverage 10.402205435195318 and 23.599084994628157. The first haplotype has six polymorphic sites, and the second haplotype 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.total_time: total running time for the whole procedure

==========

####### 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)