User Guide (Version 1.0)

BayesDenovo: An accurate *de novo* transcriptome assembler using a Bayesian model

1. Introduction

The BayesDenovo package is developed and tested on Linux (Ubuntu) OS. The code is freely available under the MIT License (http://opensource.org/licenses/MIT).

Contact: (Jason) Jianhua Xuan at xuan@vt.edu

Version: 1.0

Version Date: Apr. 5, 2021

OS Platform: Linux (64 bit)

2. Quick Start with Demo Data

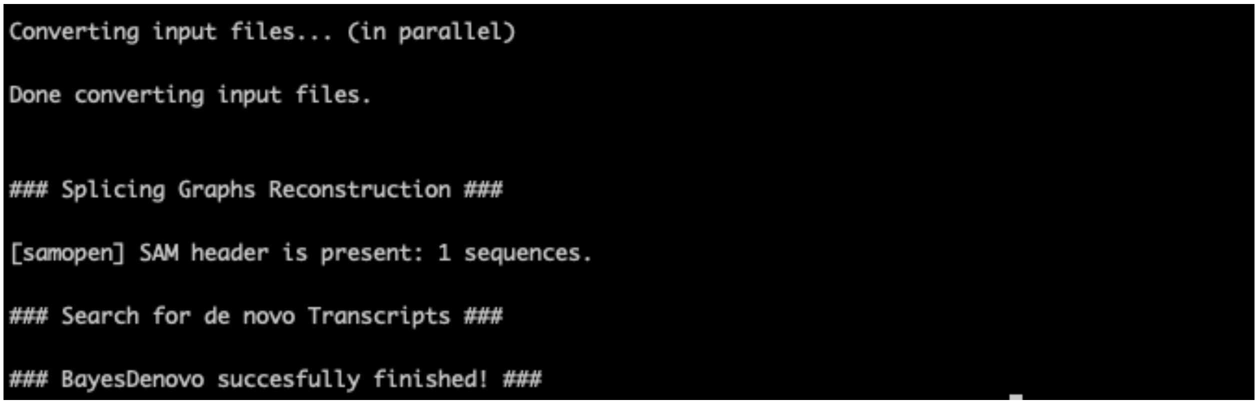
We prepared a demo paired-end fastq data in the folder "demo". There is also one script "run_BayesDenovo_demo.sh" in the package for users to test the demo data. In order to successfully complete the demo, the path of the package needs to be specified to the variable "PKG_PATH" in the script. The script can be run using

```
./run_BayesDenovo_demo.sh
```

If you see errors related to permission issues of this script, please run the following command to grant the permission:

```
chmod a+x run_BayesDenovo_demo.sh
```

If the package successfully finishes, you will see the following:



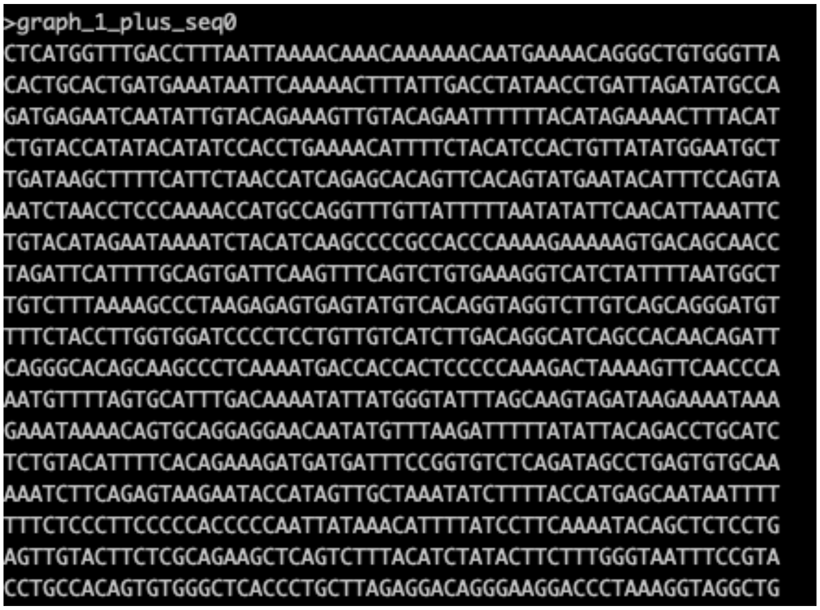
```
Converting input files... (in parallel)

Done converting input files.


### Splicing Graphs Reconstruction ###

[samopen] SAM header is present: 1 sequences.

### Search for de novo Transcripts ###

### BayesDenovo succesfully finished! ###
```

If you would like to run on other datasets, you can change the following variables in the scripts:

```
OUTPATH=     #the output folder, it will stop if the folder already exists
kmer=25 #kmer used in the assembly
seqType=fq # use fq for fastq file and fa for fasta file
LEFT=$PKG_PATH/demo/reads.left.fq #left of paired end reads
RIGHT=$PKG_PATH/demo/reads.right.fq #right of paired end reads
nCPU=5 #number of cores
```

3. Output

The assembled transcripts will be stored in file “assembly.fa” in the output folder. Here is a sample output from demo data:



```
>graph_1_plus_seq0
CTCATGGTTTGACCTTTAATTAAAACAAACAAAAAACAATGAAAACAGGGCTGTGGGTTA
CACTGCACTGATGAAATAATTCAAAAACTTTATTGACCTATAACCTGATTAGATATGCCA
GATGAGAATCAATATTGTACAGAAAGTTGTACAGAATTTTTTACATAGAAAACTTTACAT
CTGTACCATATACATATCCACCTGAAAACATTTTCTACATCCACTGTTATATGGAATGCT
TGATAAGCTTTTCATTCTAACCATCAGAGCACAGTTCACAGTATGAATACATTTCCAGTA
AATCTAACCTCCCAAAACCATGCCAGGTTTGTTATTTTTAATATATTCAACATTAAATTC
TGTACATAGAATAAAATCTACATCAAGCCCCGCCACCCAAAAGAAAAAGTGACAGCAACC
TAGATTCATTTTGCAGTGATTCAAGTTTCAGTCTGTGAAAGGTCATCTATTTTAATGGCT
TGTCTTTAAAAGCCCTAAGAGAGTGAGTATGTCACAGGTAGGTCTTGTCAGCAGGGATGT
TTTCTACCTTGGTGGATCCCCTCCTGTTGTCATCTTGACAGGCATCAGCCACAACAGATT
CAGGGCACAGCAAGCCCTCAAAATGACCACCACTCCCCCAAAGACTAAAAGTTCAACCCA
AATGTTTTAGTGCATTTGACAAAATATTATGGGTATTTAGCAAGTAGATAAGAAAATAAA
GAAATAAAACAGTGCAGGAGGAACAATATGTTTAAGATTTTTATATTACAGACCTGCATC
TCTGTACATTTTCACAGAAAGATGATGATTTCCGGTGTCTCAGATAGCCTGAGTGTGCAA
AAATCTTCAGAGTAAGAATACCATAGTTGCTAAATATCTTTTACCATGAGCAATAATTTT
TTTCTCCCTTCCCCCACCCCCAATTATAAACATTTTATCCTTCAAAATACAGCTCTCCTG
AGTTGTACTTCTCGCAGAAGCTCAGTCTTTACATCTATACTTCTTTGGGTAATTTCCGTA
CCTGCCACAGTGTGGGCTCACCCTGCTTAGAGGACAGGGAAGGACCCTAAAGGTAGGCTG
```

4. Options

The full options are shown below with explanation of each option:

```
# ** Required **
#  --seqType <string>  : type of reads: (fa, fq, cfa, cfq)
#  --bamfile <string>  : temporary pseudo alignment of reads
#  --fafile <string>   : temporary fasta file for pseudo alignment
#
#  If paired reads:
#     --left  <string>  : left reads
#     --right <string>  : right reads
#
#  Or, if unpaired reads:
#     --single <string>  : single reads
#
# ** Optional **
#  if strand-specific data, set:
#  --SS_lib_type <string> : Strand-specific RNA-Seq reads orientation. if paired: RF or FR,  if single: F or R.
#  --kmer_length/-k <int> : length of kmer, default: 25.
#  --output/-o <string> : name of directory for output, default: bayesdenovo_out_dir.
#  --CPU <int>  : number of CPUs, default: 2.
#  --pair_gap_length : gap length of paired reads, default: 200.
#  --min_seed_coverage <int> : minimum coverage of kmer as a seed, default: 2 .
#  --min_seed_entropy <float> : minimum entropy of kmer as a seed, default: 1.5.
#  --min_kmer_coverage <int> : minimum coverage of kmer used to extend, default: 1.
#  --min_kmer_entropy <float> : minimum entroy of kmer used to extend, default: 0.0.
#  --min_junction_coverage <int> : minimum of the coverage of a junction, default: 2.
#  --min_ratio_non_error <float> : min ratio for low/high alternative extension that is not an error,
default: 0.05.
#  --min_reads_span_junction<int>: minimum number of reads supporting a junction.
#  --min_cov_singletranscript : minimum coverage for single transcripts, default: 10.
#  --help/-h: show help information.
```