Assembly with Megahit

James Thornton Jr

Abstract

Co-assembly using Megahit.

Citation: James Thornton Jr Assembly with Megahit. **protocols.io**
dx.doi.org/10.17504/protocols.io.j8tcrwn
Published: 10 Oct 2017

Protocol

Step 1.

Log into the HPC.

cmd COMMAND

```
$ ssh hpc
$ ocelote
```

Step 2.

From your home directory, open .bashrc file for editing.

cmd COMMAND

```
$ nano .bashrc
```

NOTES

James Thornton Jr 09 Oct 2017

Remember, you are already in your home directory after logging into ocelote.

Step 3.

Input the following line into your .bashrc file:

cmd COMMAND

```
export PATH=/rsgrps/bh_class/bin:$PATH
```

This will allow us to execute tools found in /rsgrps/bh_class/bin without specifying the path name.

Step 4.

Save and close the .bashrc file

Step 5.

Move into your project directory.

cmd COMMAND

```
$ cd /rsgrps/bh_class/username
```

Step 6.

Create a directory for assembly output. Then move into that directory.

cmd COMMAND

```
$ mkdir assembly
$ cd !$
```

Step 7.

Make directories for standard out and standard error.

cmd COMMAND

```
mkdir std-out std-err
```

Step 8.

Before we continue, determine if you have single end or paired end files. If you have two files per SRR number, you have paired end reads. Otherwise, you have single end reads.

1. If you have single end reads proceed to step 5.

2. If you have paired end reads, skip to step 6.

Step 9.

Assembly script for SINGLE END FILES

Create a script called run-assembly.sh

cmd COMMAND

```
#!/bin/bash

#PBS -W group_list=bh_class
#PBS -q windfall
#PBS -l select=1:ncpus=20:mem=40gb
#PBS -l pvmem=38gb
#PBS -l walltime=24:00:00
#PBS -l cput=48:00:00
#PBS -M netid@email.arizona.edu
#PBS -m bea

FASTQ_DIR='/rsgrps/bh_class/username/fastq'
ASSEM_DIR='/rsgrps/bh_class/username/assembly'
MIN_CONTIG_LEN=500
OUT_DIR='/rsgrps/bh_class/username/assembly/megahit-out'

cd $ASSEM_DIR

SINGLEs=`ls $FASTQ_DIR/*.fastq | python -
```

```
c 'import sys; print ",".join([x.strip() for x in sys.stdin.readlines()])'`

megahit -r $SINGLEs --preset meta-sensitive --min-contig-len $MIN_CONTIG_LEN -o $OUT_DIR -
t 12
```

NOTES

James Thornton Jr 09 Oct 2017

OUT_DIR does NOT need to be created prior to running this script. Megahit will make the directory on its own.

Step 10.

Assembly script for PAIRED END FILES

Create a script called run-assembly.sh

cmd COMMAND

```
#!/bin/bash

#PBS -W group_list=bh_class
#PBS -q windfall
#PBS -l select=1:ncpus=20:mem=40gb
#PBS -l pvmem=38gb
#PBS -l walltime=24:00:00
#PBS -l cput=48:00:00
#PBS -M netid@email.arizona.edu
#PBS -m bea

FASTQ_DIR='/rsgrps/bh_class/username/fastq'
ASSEM_DIR='/rsgrps/bh_class/username/assembly'
MIN_CONTIG_LEN=500
OUT_DIR='/rsgrps/bh_class/username/assembly/megahit-out'

cd $ASSEM_DIR

R1s=`ls $FASTQ_DIR/*_1.fastq | python -
c 'import sys; print ",".join([x.strip() for x in sys.stdin.readlines()])'`
R2s=`ls $FASTQ_DIR/*_2.fastq | python -
c 'import sys; print ",".join([x.strip() for x in sys.stdin.readlines()])'`

megahit -1 $R1s -2 $R2s --preset meta-sensitive --min-contig-len $MIN_CONTIG_LEN -
o $OUT_DIR -t 12
```

NOTES

James Thornton Jr 09 Oct 2017

OUT_DIR does NOT need to be created prior to running this script. Megahit will make the directory on its own.

Step 11.

Run the assembly:

cmd COMMAND

```
$ chmod +x run-assembly.sh
$ qsub -e std-err/ -o std-out/ run-assembly.sh
```

Step 12.

You can check the status of your job with the following command:

cmd COMMAND

```
$ qstat -u username
```

NOTES

James Thornton Jr 09 Oct 2017

Job runtime will vary depending on the size of your dataset.

Step 13.

Upon job completion, inspect the log file in the megahit-out folder to determine the quality of the assembly including: N50, number of total contigs, maximum/minimum lengths.