MG_HW7: Taxonomic Classification Using Centrifuge

James Thornton

Abstract

This protocol provides a procedure to generate taxonomic data from assembled contigs using centrifuge.

Citation: James Thornton MG_HW7: Taxonomic Classification Using Centrifuge. **protocols.io**
dx.doi.org/10.17504/protocols.io.f7mbrk6
Published: 25 Oct 2016

Guidelines

Centrifuge documentation

Protocol

Step 1.

Log in to the HPC cluster (ICE)

cmd COMMAND
```
$ ssh hpc
$ ice
```

NOTES
James Thornton Jr 25 Oct 2016

Option 3 for those with menu enabled.

Step 2.

Move into your class directory.

cmd COMMAND
```
$ cd /rsgrps/bh_class/username
```
Use YOUR username

Step 3.

Update your 'run-interactive.sh' script to include new memory allocations:

cmd COMMAND

```
#!/bin/bash

#PBS -W group_list=bh_class
#PBS -q windfall
#PBS -l jobtype=cluster_only
#PBS -l select=1:ncpus=12:mem=23gb
#PBS -l pvmem=22gb
#PBS -l walltime=24:00:00
#PBS -l cput=24:00:00
#PBS -M netid@email.arizona.edu
#PBS -m bea

#--------------EDIT THESE---------------
FASTA_DIR="/rsgrps/bh_class/username/fasta"
OUT_DIR="/rsgrps/bh_class/username/taxonomy"
BT2_OUT_DIR="/rsgrps/bh_class/username/unmapped"
#----------------------------------------

CENT_DB="/rsgrps/bh_class/b_compressed+h+v/b_compressed+h+v"
BT2_INDEX="/rsgrps/bh_class/bowtie2_index/human_index"

cd "$FASTA_DIR"
export FASTA_LIST="$FASTA_DIR/fasta-list"
ls *.fasta > $FASTA_LIST
echo "FASTA files to be processed:" $(cat $FASTA_LIST)

module load bowtie2/2.2.5
while read FASTA; do
  export FASTA="$FASTA"
  export FILE_NAME=`basename $FASTA | cut -d '.' -f 1`
  bowtie2 -x $BT2_INDEX -U $FASTA -f --very-sensitive-local -p 4 --
un $BT2_OUT_DIR/$FILE_NAME.unmapped

done < $FASTA_LIST

cd "$BT2_OUT_DIR"
export UNMAPPED_LIST="$BT2_OUT_DIR/unmapped-list"
ls *.unmapped > $UNMAPPED_LIST
echo "Running Centrifuge on the following files:" $(cat $UNMAPPED_LIST)

while read UNMAPPED; do
  export UNMAPPED="$UNMAPPED"
  export UNMAPPED_NAME=$(basename $UNMAPPED | cut -d '.' -f 1)
  centrifuge -x $CENT_DB -U $UNMAPPED -S $OUT_DIR/$UNMAPPED_NAME-classout --report-
file $OUT_DIR/$UNMAPPED_NAME-centrifuge_report.tsv -f
done < $UNMAPPED_LIST
```

As indicated in the script, edit the FASTA_DIR and OUT_DIR to include the path to YOUR Fasta files and the taxonomy directory created in the previous step. Remember to replace netid with YOUR netid to receive email notifications

NOTES

James Thornton Jr 25 Oct 2016

Note: run-interactive.sh should already exist from the last protocol but it can be created new if needed.

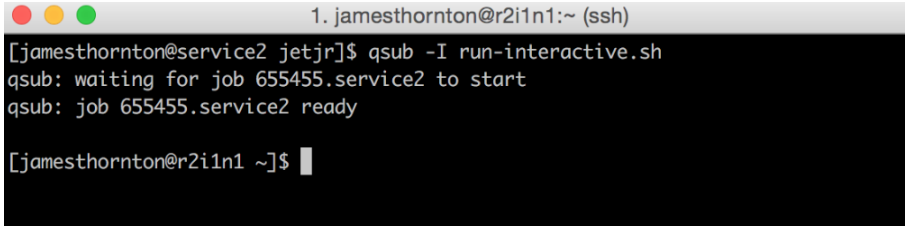
Step 4.

Submit run-interactive using qsub:

cmd COMMAND

```
$ qsub -e std-err/ -o std-out/ centrifuge_tax.sh
```

EXPECTED RESULTS



Step 5.

Once the job is ready move back into your class directory.

cmd COMMAND

```
$ qstat -u username
```

Use YOUR username Under S (Status) 'Q' means queued, 'R' means running

Step 6.

Make a directory named 'taxonomy'.

cmd COMMAND

```
$ cd taxonomy
$ ls
```

Step 7.

Run Centrifuge on your fixed-contigs.fa:

cmd COMMAND

```
$ module load R
```

EXPECTED RESULTS

For my 3000+ contigs it took ~2 minutes to run. The time it takes will vary depending on the amount of contigs you have.

NOTES

James Thornton Jr 25 Oct 2016

Important: This step is written under the assumption you are executing it while in /rsgrps/bh_class/username

If you are somewhere else while trying to execute this command it will NOT work.

Step 8.

Once the job is complete, move into the taxonomy directory and ensure the correct output is there.

cmd COMMAND

```
$ ./cent_barplots.R
```

EXPECTED RESULTS

```
2. jamesthornton@service2:/rsgrps/bh_class/jetjr/taxonomy (ssh)
[jamesthornton@service2 jetjr]$ cd taxonomy/
[jamesthornton@service2 taxonomy]$ ls
centrifuge_report.tsv  classification_out
[jamesthornton@service2 taxonomy]$
```

Step 9.

Sort centrifuge_report.tsv by the 5th column descending and put the output into a temp_sorted.tsv file:

cmd COMMAND

```
#!/usr/bin/env Rscript

#----------EDIT HERE----------
cent.dir <- "/rsgrps/bh_class/username/taxonomy/"
out.dir <- "/rsgrps/bh_class/username/taxonomy/barplots/"
#--------------------------------

file.names <- dir(cent.dir, pattern="-centrifuge_report.tsv")

gen_barplot <- function (data) {
  data_title <- gsub("-centrifuge_report.tsv", "", data)
  data <- read.delim(paste0(i, data))
  total_reads <- sum(data$numReads)
  proportion_classified <- data$numReads / total_reads
  data["proportion_classified"] <- proportion_classified
  read_subset <-
 subset(data, proportion_classified > 0.005, select = c("name", "numReads", "proportion_cla
ssified"))
  read_subset$numReads <- as.numeric(read_subset$numReads)
  png(filename=paste0(out.dir,data_title,"_taxonomy.png"), width = 600, height = 600)
  op <- par(mar=c(15, 8, 4, 2) + 0.1, mgp = c(10, 1, 0))
  p1 <-
 barplot(read_subset$proportion_classified, main=paste0("Read Proportional Classification:
",data_title), names.arg = read_subset$name, las=2, cex.names = 1, cex.axis = 1, ylab="Prop
ortion Classified", ylim = c(0, 0.90))
  grid(nx=NA, ny=NULL)
  print(p1)
  dev.off()
}

for (i in cent.dir) {
  lapply(file.names, gen_barplot)
}
```

Make sure to edit username in cent.dir and out.dir to include YOUR path. Also ensure that both

cent.dir and out.dir end with the slash

Step 10.

During the sort the column headers are placed at the bottom of the file. Fix this with the following command:

cmd COMMAND

```
$ echo $(tail -1 temp_sorted.tsv) | cat -
 temp_sorted.tsv | sed '$d' > centrifuge_sorted.tsv
```

The last line of temp_sorted is piped into cat of temp_sorted.tsv, placing it at the top of the file. sed '$d' deletes the last line which is the column headers which is now at the top

Step 11.

Once you have centrifuge_sorted.tsv, remove the temp_sorted.tsv file.

cmd COMMAND

```
> q()
Save workspace image? [y/n/c]: n
```

Step 12.

Take a look at your now sorted centrifuge_sorted.tsv file. Report on some of the most abundant organisms with their read counts. Make sure to state the methods used to obtain these results.