MG_HW7: Taxonomic Classification Using Centrifuge Version 2

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.f7sbrne
Published: 26 Oct 2016

Guidelines

Centrifuge documentation

Protocol

Step 1.

Log in to the HPC cluster (ICE)

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

NOTES
James Thornton Jr 26 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.

Make a new directory called "taxonomy"

cmd COMMAND
```
$ mkdir taxonomy
```

```
$ mkdir unmapped
```

Step 4.

Copy the following into a new script named centrifuge_tax.sh:

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 26 Oct 2016

Important: For this to work you Fasta files must end with the extension .fasta

Step 5.

Submit centrifuge_tax.sh using qsub:

cmd COMMAND

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

Step 6.

Once the job is running it will loop through all of your Fasta files and run centrifuge to generate taxonomic data. This will take about 1 hour to generate reports for all 6 of your fasta files. You can use qstat to check the status of your job.

cmd COMMAND

```
$ qstat -u username
```

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

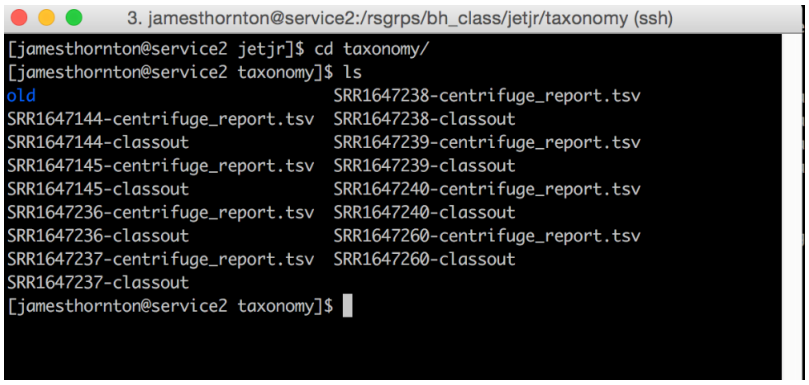
Step 7.

Once the job is complete move into your taxonomy directory and ensure all output files are there. If the job was successful there should be a total of 6 "classout" files and 6 "centrifuge_report.tsv" files.

cmd COMMAND

```
$ cd taxonomy
$ ls
```

EXPECTED RESULTS



Step 8.

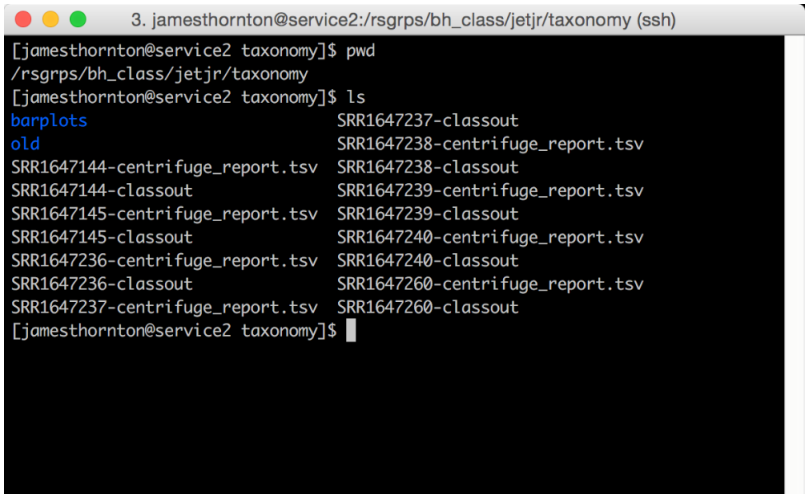
In your taxonomy directory make a new directory called barplots

cmd COMMAND

```
$ mkdir barplots
```

Make sure you are in /rsgrps/bh_class/username/taxonomy for this to work correctly

EXPECTED RESULTS



Step 9.

Copy + Paste the following into a script called cent_barplots.R

Important: Edit cent.dir and out.dir to include the correct paths

- Edit cent.dir to include the path to your taxonomy directory (/rsgrps/bh_class/username/taxonomy/)
- Edit out.dir to include the path to your barplots diretory (/rsgrps/bh_class/username/taxonomy/barplots/)

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

NOTES

James Thornton Jr 26 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.

James Thornton Jr 26 Oct 2016

This R script will calculate the total number of reads and then divide the classified reads by the total for each hit generating a proportion classified statistic. Only hits with a proportion of 0.5% of reads classified will be plotted.

Step 10.

Once you have edited cent.dir and out.dir save and close the file. Cat cent_barplots.R and copy the entire script.

cmd COMMAND

```
$ chmod +x cent_barplots.R
```

NOTES

James Thornton Jr 26 Oct 2016

Make sure you copy the ENTIRE script.

Step 11.

Load the module R:

cmd COMMAND

```
$ module load R
```

Step 12.

Load R. You should see a prompt once executed:

cmd COMMAND

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

EXPECTED RESULTS

```
3. jamesthornton@service2:/rsgrps/bh_class/jetjr/taxonomy (ssh)
[jamesthornton@service2 taxonomy]$ module load R
[jamesthornton@service2 taxonomy]$ R

R version 2.15.2 (2012-10-26) -- "Trick or Treat"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-unknown-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' for quit R.

>
```

Step 13.

Paste what you copied from step 10 into the R prompt.

EXPECTED RESULTS

```
3. jamesthornton@service2:/rsgrps/bh_class/jetjr/taxonomy (ssh)

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> #----------EDIT HERE----------
> cent.dir <- "/rsgrps/bh_class/jetjr/taxonomy/"
> out.dir <- "/rsgrps/bh_class/jetjr/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_classifi
ed"))
+   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="Proportion 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)
+ }
```

Step 14.

Press enter to execute the R script. You should see something similar to what is shown below.

EXPECTED RESULTS

```
3. jamesthornton@service2:/rsgrps/bh_class/jetjr/taxonomy (ssh)
[10,] 11.5
[11,] 12.7
[12,] 13.9
[13,] 15.1
[14,] 16.3
[15,] 17.5
[16,] 18.7
[17,] 19.9
[18,] 21.1
[19,] 22.3
[20,] 23.5
[21,] 24.7
[22,] 25.9
[23,] 27.1
     [,1]
[1,]  0.7
[2,]  1.9
[3,]  3.1
[4,]  4.3
[5,]  5.5
[6,]  6.7
     [,1]
[1,]  0.7
[2,]  1.9
[3,]  3.1
     [,1]
[1,]  0.7
[2,]  1.9
     [,1]
[1,]  0.7
[2,]  1.9
[3,]  3.1
[4,]  4.3
>
```

Step 15.

Type q() and press enter to quit R. Press n + enter when asked to save.

cmd COMMAND

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

Step 16.

Move into your barplots directory and make sure you have 6 .png images.

cmd COMMAND

```
$ cd /rsgrps/bh_class/username/taxonomy/barplots
$ ls
```

EXPECTED RESULTS

```
3. jamesthornton@service2:/rsgrps/bh_class/jetjr/taxonomy/barplots (ssh)
[jamesthornton@service2 taxonomy]$ cd barplots/
[jamesthornton@service2 barplots]$ ls
SRR1647144_taxonomy.png  SRR1647236_taxonomy.png  SRR1647238_taxonomy.png  SRR1647240_taxonomy.png
SRR1647145_taxonomy.png  SRR1647237_taxonomy.png  SRR1647239_taxonomy.png  SRR1647260_taxonomy.png
[jamesthornton@service2 barplots]$
```

Step 17.

To view the images you must scp to your local machine. Open a new terminal (don't log into hpc). Determine where you want to store the files on your local machine and move into that directory.

NOTES

James Thornton Jr 26 Oct 2016

Windows users using Cygwin, your file will be stored in C:/cygwin64/home/USER. Just open a new

terminal window and proceed to next step (you can't move to a specific local directory).

Step 18.

Execute the following command to scp the .png files to your local machine:

cmd COMMAND

```
$ scp netid@hpc.arizona.edu:/rsgrps/bh_class/username/taxonomy/barplots/*.png .
```

Replace netid and username. (They may be different).

Step 19.

You can now open the images on your local machine. Reminder that windows users will have their images in C:/cygwin64/home/USER.

Step 20.

Report on what you've found for each sample. Make sure to state the method used to obtain these results.