Assign taxonomy to gene calls using Centrifuge

James Thornton Jr

Abstract

Uses a custom Centrifuge pipeline to assign taxonomy to gene calls.

Citation: James Thornton Jr Assign taxonomy to gene calls using Centrifuge. **protocols.io**
dx.doi.org/10.17504/protocols.io.kpfcvjn
Published: 07 Nov 2017

Protocol

Step 1.

Log into the HPC

cmd COMMAND

```
$ ssh hpc
$ ocelote
```

Step 2.

Move into your class directory.

cmd COMMAND

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

Step 3.

Clone the Centrifuge github repository.

cmd COMMAND

```
$ git clone git@github.com:jetjr/Centrifuge.git
```

Step 4.

Move into the Centrifuge directory.

cmd COMMAND

```
$ cd Centrifuge
```

Dependencies

Step 5.

This program uses R packages that must be installed prior to launching the job. Load the R module.

cmd COMMAND

```
$ module load unsupported
$ module load markb/R/3.1.1
```

Dependencies

Step 6.

Launch R.

cmd COMMAND

```
$ R
```

Dependencies

Step 7.

Get the "optparse" package.

cmd COMMAND

```
> install.packages("optparse", repos="http://R-Forge.R-project.org")
```

Dependencies

Step 8.

Get ggplot2 and plyr packages. You may be prompted to select a mirror. Any US server will work.

cmd COMMAND

```
> install.packages("ggplot2")
> install.packages("plyr")
```

NOTES

James Thornton Jr 07 Nov 2017

If you receive an error when installing the dependencies, continue with the protocol.

Dependencies

Step 9.

Quit the R session. Do not save workspace image.

cmd COMMAND

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

Step 10.

Edit the config.sh file to include the correct variable declarations. The following steps will detail how the config.sh file should be edited.

cmd COMMAND

```
$ nano config.sh
```

CENT_DB

Step 11.

export CENT_DB="/rsgrps/bh_class/b_compressed+h+v/b_compressed+h+v"

FASTA_DIR

Step 12.

export FASTA_DIR="/rsgrps/bh_class/username/prodigal"

NOTES

James Thornton Jr 07 Nov 2017

FASTA_DIR should point to the directory containing your nucleotides.fna file generated from step 2 and transfered to the anvio-genes directory.

TYPE

Step 13.

```
export TYPE="single"
```

FILE_EXT

Step 14.

```
export FILE_EXT="fna"
```

REPORT_DIR

Step 15.

```
export REPORT_DIR="/rsgrps/bh_class/username/taxonomy"
```

NOTES

James Thornton Jr 07 Nov 2017

The program will create this directory for you. Make sure to replace username.

PLOT_OUT

Step 16.

```
export PLOT_OUT='/rsgrps/bh_class/username/taxonomy/'
```

NOTES

James Thornton Jr 07 Nov 2017

Same as REPORT_DIR but make sure to include the trailing / as stated in the config.sh file.

PLOT_FILE and PLOT_TITLE

Step 17.

These should be named according to what sample your working with. For example, ocean data may name these:

```
export PLOT_FILE='ocean_depth'

export PLOT_TITLE='ocean_depth'
```

NOTES

James Thornton Jr 07 Nov 2017

PLOT FILE will be the file name of the bubble plot that is generated.

PLOT TITLE will be the title found on the actual plot.

FILE_TYPE

Step 18.

export FILE_TYPE="f"

NOTES

James Thornton Jr 07 Nov 2017

The nucleotides.fna file is in FASTA format.

EXCLUDE

Step 19.

The exclude parameter can be left blank.

export EXCLUDE=""

Step 20.

Save and quit config.sh

Step 21.

Submit the job using the submit script found in the Centrifuge directory.

cmd COMMAND

```
$ ./submit.sh
```

Step 22.

Status of the job can be determined by the following command:

cmd COMMAND

```
$ stat -u username
```