Assign taxonomy to gene calls using Centrifuge Version 4

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.ksrcwd6
Published: 13 Nov 2017

Protocol

Step 1.

Navigate to the directory on your local machine that contains the contigs.db generated during the Anvi'o protocol.

Step 2.

Extract gene calls from the contigs database.

cmd COMMAND

```
$ anvi-get-dna-sequences-for-gene-calls -c CONTIGS.db -o nucleotides.faa
```

NOTES

James Thornton Jr 07 Nov 2017

Important: nucleotides.fna was generated in the prodigal protocol. HOWEVER, we will be using this version from Anvi'o for taxonomy assignment.

James Thornton Jr 07 Nov 2017

Remember windows users you must launch Anvio using docker.

docker run --rm -v /path/to/files:/my_data -p 8080:8080 -it meren/anvio:latest

ANNOTATIONS

Eldridge Wisely 15 Nov 2017

Mine says:

WARNING

==========

You did not provide any gene caller ids. As a result, anvi'o will give you back

sequences for every 65293 gene call stored in the contigs database. Brace

yourself.

Does this mean that my Anvio step didn't work correctly?

Step 3.

Log into the HPC

```
cmd COMMAND
$ ssh hpc
$ ocelote
```

Step 4.

Move into your class directory.

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

Step 5.

Make an anvio-genes directory.

```
cmd COMMAND
$ mkdir anvio-genes
```

Step 6.

On your local machine, scp the nucleotides.fna file generated from step 2 into the newly created anvio-genes directory.

```
cmd COMMAND
$ scp nucleotides.fna username@sftp.hpc.arizona.edu:/rsgrps/bh_class/username/anvio-genes
```

Step 7.

Clone the Centrifuge github repository into your class directory on the HPC.

```
cmd COMMAND
$ pwd
/rsgrps/bh_class/username
$ git clone git@github.com:jetjr/Centrifuge.git
```

Step 8.

Move into the Centrifuge directory.

```
cmd COMMAND
$ cd Centrifuge
```

Dependencies

Step 9.

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 10.

Launch R.

cmd COMMAND

```
$ R
```

Dependencies

Step 11.

Get the "optparse" package.

cmd COMMAND

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

NOTES

James Thornton Jr 08 Nov 2017

Choose yes if promted to use a personal library.

Dependencies

Step 12.

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 13.

Quit the R session. Do not save workspace image.

cmd COMMAND

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

Step 14.

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 15.

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

FASTA_DIR

Step 16.

export FASTA_DIR='/rsgrps/bh_class/username/anvio-genes'

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 17.

export TYPE="single"

FILE_EXT

Step 18.

export FILE_EXT='faa'

REPORT_DIR

Step 19.

export REPORT_DIR='/rsgrps/bh_class/username/anvio-genes/taxonomy/'

NOTES

James Thornton Jr 07 Nov 2017

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

PLOT_OUT

Step 20.

export PLOT_OUT='/rsgrps/bh_class/username/anvio-genes/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 21.

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 22.

export FILE_TYPE="f"

NOTES
James Thornton Jr 07 Nov 2017

The nucleotides.fna file is in FASTA format.

EXCLUDE
Step 23.

The exclude parameter can be left blank.

export EXCLUDE=""

Step 24.

Save and quit config.sh

Step 25.

Move into the script directory.

cmd COMMAND
```
$ cd scripts
```
Step 26.

Edit the PBS variables in centrifuge_single_tax.sh to include the bh_class group and your email.

#PBS -W group_list=bh_class

#PBS -M netid@email.arizona.edu

cmd COMMAND
```
$ nano centrifuge_single_tax.sh
```
Step 27.

Save and quite centrifuge_single_tax.sh. Then move back into the main Centrifuge directory.

cmd COMMAND
```
$ cd ..
```
Step 28.

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

cmd COMMAND
```
$ ./submit.sh
```
Step 29.

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

cmd COMMAND
```
$ stat -u username
```
Step 30.

A successful job will generate a centrifuge_report.tsv file in anvio-genes/taxonomy.