Nov 13, 2017 Version 4

Assign taxonomy to gene calls using Centrifuge V.4

DOI
dx.doi.org/10.17504/protocols.io.ksrcwd6

James E Thornton Jr[1]
[1]Hurwitz Lab

Metafunc Course 2017

James E Thornton Jr
Hurwitz Lab

DOI: https://dx.doi.org/10.17504/protocols.io.ksrcwd6

Protocol Citation: James E Thornton Jr 2017. Assign taxonomy to gene calls using Centrifuge. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.ksrcwd6**

Protocol status: Working
We use this protocol and it's working

Created: November 13, 2017

Last Modified: March 01, 2018

Protocol Integer ID: 8753

Keywords: custom centrifuge pipeline, using centrifuge use, gene call, taxonomy, gene

Abstract

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

Troubleshooting

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

2 Extract gene calls from the contigs database.

Command

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

Note

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

Note

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

3 Log into the HPC

Command

```
$ ssh hpc
$ ocelote
```

4 Move into your class directory.

Command

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

5 Make an anvio-genes directory.

Command

```
$ mkdir anvio-genes
```

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

Command

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

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

Command

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

8 Move into the Centrifuge directory.

Command

```
$ cd Centrifuge
```

Dependencies

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

Command

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

10 Launch R.

Command

```
$ R
```

11 Get the "optparse" package.

Command

```
> </ProtocolCommand>
 <ProtocolNote id=
```

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

Command

```
> install.packages(
```

Note

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

13 Quit the R session. Do not save workspace image.

Command

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

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.

Command

```
$ nano config.sh
```

CENT_DB

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

FASTA_DIR

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

Note

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

TYPE

17 export TYPE="single"

FILE_EXT

18 export FILE_EXT='faa'

REPORT_DIR

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

Note

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

PLOT_OUT

20 export PLOT_OUT='/rsgrps/bh_class/username/anvio-genes/taxonomy/'

Note

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

PLOT_FILE and PLOT_TITLE

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'

Note

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

22 export FILE_TYPE="f"

Note

The nucleotides.fna file is in FASTA format.

EXCLUDE

23 The exclude parameter can be left blank.

export EXCLUDE=""

24 Save and quit config.sh

25 Move into the script directory.

Command

```
$ cd scripts
```

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

Command

```
$ nano centrifuge_single_tax.sh
```

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

Command

```
$ cd ..
```

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

Command

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

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

Command

```
$ stat -u username
```

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