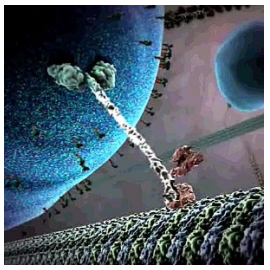
Nov 18, 2017

Annotate gene function with Uproc



DOI
dx.doi.org/10.17504/protocols.io.kt8cwrw

James E Thornton Jr[1]

[1]Hurwitz Lab

Metafunc Course 2017

James E Thornton Jr
Hurwitz Lab

DOI: dx.doi.org/10.17504/protocols.io.kt8cwrw

Protocol Citation: James E Thornton Jr 2017. Annotate gene function with Uproc. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.kt8cwrw**

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

Created: November 17, 2017

Last Modified: March 15, 2018

Protocol Integer ID: 8800

Abstract

This protocol details the steps to annotate Anvi'o gene calls for function using Uproc.

1 Log into the HPC.

Command

```
$ ssh hpc
$ ocelote
```

2 Move into your anvio-genes directory.

Command

```
$ cd /rsgrps/bh_class/username/anvio-genes
```

3 Make a "function" directory.

Command

```
$ mkdir function
```

4 Move into the function directory.

Command

```
$ cd function
```

5 Create uproc_function.sh to run the functional analysis.

Note

Your gene calls should be located in /rsgrps/bh_class/username/anvio-genes. It maybe called nucleotides.fna or nucleotides.faa. Make sure the FASTA variable uses the correct name.

Command

Make sure to replace netid AND the

```
#!/bin/bash
#PBS -W group_list=bh_class
#PBS -q windfall
#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=
```

6 Make a standard out and standard error directory.

Command

```
$ mkdir std-err std-out
```

7 Run the uproc_function.sh script.

Command

```
$ qsub -e std-err -o std-out uproc_function.sh
```

8 Check the status of your job. Continue to the next step upon sucessful job completion.

Command

```
$ qstat -u username
```

9 Move into your function directory.

Command

```
$ cd /rsgrps/bh_class/username/anvio-genes/function
```

10 Create a perl script called format-anvio.pl to convert the functional data into anvio format.

Command

```
#! /usr/bin/env perl
use strict;
if (@ARGV != 4) { die
```

11 Run format-anvio.pl to convert the functional data into anvio format.

Command

```
chmod 755 format-anvio.pl
./format-anvio.pl uproc-out.kegg /rsgrps/bh_class/kegg_to_desc uproc-
kegg-anvio kegg
./format-anvio.pl uproc-out.pfam /rsgrps/bh_class/pfam_to_domain
uproc-pfam-anvio pfam
cat uproc-kegg-anvio > input_matrix.txt
egrep -v
```

12 Download the functional data to your computer. Go to the directory where you are storing your Anvi'o data (contigs.db).

Command

```
scp netid@sftp.hpc.arizona.edu:/rsgrps/bh_class/username/anvio-
genes/function/input_matrix.txt .
```

Note

This step is done on a local terminal (not the HPC).

13 Download the taxonomy data to your computer. Go to the directory where you are storing your Anvi'o data (contigs.db).

Command

```
scp netid@sftp.hpc.arizona.edu:/rsgrps/bh_class/username/anvio-
genes/taxonomy/*centrifuge_report.tsv .
scp netid@sftp.hpc.arizona.edu:/rsgrps/bh_class/username/anvio-
genes/taxonomy/*centrifuge_hits.tsv .
```

14 Open Anvio.

For those using the Docker image only execute the following command:

Command

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

15 From the Anvi'o terminal, type the following command to upload the taxonomic data for the gene calls.

Command

```
anvi-import-taxonomy -c contigs.db -i centrifuge_report.tsv
centrifuge_hits.tsv -p centrifuge
```

Note

You should be in the directory that contains both your contig database and the centrifuge files.

16 From the Anvi'o terminal, type the following command to upload the functional data for the gene calls.

Command

```
anvi-import-functions -c contigs.db -i input_matrix.txt
```

17 Please document these steps in the methods section of your report. Note what programs you used for each step and what the parameters were. How many genes did Anvi'o find? Is this different from your analyses? Why? Anvi'o uses the "-p meta" parameter for metagenomics datasets, which is stricter than what you first ran. How many of your genes from Anvio had a match to a known bacteria? How many matched known proteins for kegg or pfam?