Annotate gene function with Uproc

James Thornton Jr

Abstract

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

Citation: James Thornton Jr Annotate gene function with Uproc. **protocols.io**
dx.doi.org/10.17504/protocols.io.kt8cwrw
Published: 17 Nov 2017

Protocol

Step 1.

Log into the HPC.

cmd COMMAND
```
$ ssh hpc
$ ocelote
```
Step 2.

Move into your anvio-genes directory.

cmd COMMAND
```
$ cd /rsgrps/bh_class/username/anvio-genes
```
Step 3.

Make a "function" directory.

cmd COMMAND
```
$ mkdir function
```
Step 4.

Move into the function directory.

cmd COMMAND
```
$ cd function
```
Step 5.

Create uproc_function.sh to run the functional analysis.

cmd COMMAND
```
#!/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="/rsgrps/bh_class/username/anvio-genes/nucleotides.fna"
OUT_DIR="/rsgrps/bh_class/username/anvio-genes/function"
OUTPUT="$OUT_DIR/uproc-out"
#---------------------------------------

export UPROC="/rsgrps/bh_class/bin/uproc-dna"
export DATA="/rsgrps/bh_class/data/uproc"
export UPROC_MODEL="$DATA/model"
export UPROC_OUT_DIR="$OUT_DIR/uproc-out"
export PFAM="$DATA/pfam27ready"
export KEGG="$DATA/keggready"


$UPROC --preds --short --threads 12 --output $OUTPUT.pfam $PFAM $UPROC_MODEL $FASTA

$UPROC --preds --short --threads 12 --output $OUTPUT.kegg $KEGG $UPROC_MODEL $FASTA
```

Make sure to replace netid AND the "EDIT THESE" section

NOTES

James Thornton Jr 16 Nov 2017

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.

Step 6.

Make a standard out and standard error directory.

cmd COMMAND

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

Step 7.

Run the uproc_function.sh script.

cmd COMMAND

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

Step 8.

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

cmd COMMAND

```
$ qstat -u username
```

Step 9.

Move into your function directory.

cmd COMMAND

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

Step 10.

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

cmd COMMAND

```
#! /usr/bin/env perl
use strict;

if (@ARGV != 4) { die "Usage: format-anvio.pl uproc-file function-to-desc out source\n"; }

my $uproc = shift @ARGV;
my $function = shift @ARGV;
my $out = shift @ARGV;
my $source = shift @ARGV;

open (F, $function) || die "I need a kegg or pfam desc file\n";
open (U, $uproc) || die "I need the uproc file\n";
open (OUT, ">$out") || die "Cannot open out\n";

my %id_to_desc;
while (<F>) {
   chomp $_;
   my ($id, $desc) = split (/\t/, $_);
   $id_to_desc{$id} = $desc;
}
print OUT "gene_callers_id\tsource\taccession\tfunction\te_value\n";
while (<U>) {
   chomp $_;
   my @fields = split (/,/, $_);
   my $gene = $fields[1];
   $gene =~ s/\|.*//;
   my $id = $fields[6];
   my $score = $fields[7];
   my $desc = "NONE";
   if (exists $id_to_desc{$id}) {
      $desc = $id_to_desc{$id};
   }
   print OUT "$gene\t$source\t$id\t$desc\t$score\n";
}
```

Step 11.

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

cmd 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 "gene_callers_id" uproc-pfam-anvio >> input_matrix.txt
```

Step 12.

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

cmd COMMAND

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

NOTES

James Thornton Jr 16 Nov 2017

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

Step 13.

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

cmd 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 .
```

Step 14.

Open Anvio.

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

cmd COMMAND

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

Step 15.

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

cmd COMMAND

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

NOTES

James Thornton Jr 16 Nov 2017

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

ANNOTATIONS

James Thornton Jr 21 Nov 2017

If your centrifuge report and hits files are not named as shown in the step, rename them.

Step 16.

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

cmd COMMAND

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

Step 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?