MG_HW10: Annotate anvio gene calls for function and taxonomy

Bonnie Hurwitz

Abstract

Citation: Bonnie Hurwitz MG_HW10: Annotate anvio gene calls for function and taxonomy. **protocols.io** dx.doi.org/10.17504/protocols.io.gczbsx6
Published: 08 Nov 2016

Protocol

Login
Step 1.

Login to the HPC

cmd COMMAND
```
ssh hpc
ice
```

Go to your user directory
Step 2.

Move into your user directory in bh_class

cmd COMMAND
```
cd /rsgrps/bh_class/user_id
```

Create new analysis directories
Step 3.

We are going to run analyses on the anvio gene calls to get function and taxonomy. Create these directories to store the analyses.

cmd COMMAND
```
mkdir taxonomy-genes
mkdir taxonomy-genes/std-err
mkdir taxonomy-genes/std-out
mkdir anvio-genes
mkdir function-genes
mkdir function-genes/std-err
mkdir function-genes/std-out
```

Get the nucleotide sequences for genes in Anvi'o
Step 4.

First we will need to download the anvio gene calls. Open Anvio on your computer. And type the following commands in the anvio terminal window.

cmd COMMAND

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

ANNOTATIONS

James Thornton Jr 29 Nov 2016

PC users

When you scp your files using Cygwin, move those files to a new folder in Documents. Then in docker quickstart terminal navigate to that folder and do pwd to get the full path. Then to launch Anvio:

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

Additional troubleshooting- if having issues do docker ps and see if there are existing sessions. If so do docker kill [session id]

Count the genes

Step 5.

How many genes did Anvi'o find on your contigs? Check by entering the following command in the Anvi'o terminal.

cmd COMMAND

```
egrep ">" nucleotides.faa | wc -l
```

Upload the genes to the HPC for further analysis

Step 6.

From your anvio terminal, you need to secure copy the anvio-genes.faa file back to the HPC.

cmd COMMAND

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

ANNOTATIONS

James Thornton Jr 10 Nov 2016

Windows users have to do this from Cygwin/Putty, not the Docker image for Anvi'o. For Cygwin you have to move/copy nucleotides.faa to C:/cygwin64/home/User/ then perform the above command in a terminal NOT logged into the HPC.

James Thornton Jr 10 Nov 2016

Include **netid**@sftp.hpc.arizona:/rsgrps/bh_class/username/anvio-genes

Go to the HPC

Step 7.

Go to the HPC and make sure your anvio-genes.faa file is there.

cmd COMMAND
```
cd /rsgrps/bh_class/username/anvio-genes
ls -l nucleotides.faa
```

Go to the taxonomy-genes directory

Step 8.

Go to the taxonomy-genes directory to perform the taxonomic analysis on the genes using centrifuge.

cmd COMMAND
```
cd /rsgrps/bh_class/bhurwitz/taxonomy-genes
```

Create the centrifuge_tax.sh script

Step 9.

Create the following script to run taxonomic analysis on the gene calls from Anvi'o.

cmd COMMAND
```
#!/bin/bash

#PBS -W group_list=bh_class
#PBS -q windfall
#PBS -l jobtype=cluster_only
#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.faa"
OUT_DIR="/rsgrps/bh_class/username/taxonomy-genes"
#---------------------------------------

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

centrifuge -x $CENT_DB -U $FASTA -S $OUT_DIR/centrifuge_hits.tsv --report-
file $OUT_DIR/centrifuge_report.tsv -f
```

Run the centrifuge_tax.sh script

Step 10.

Run the centrifuge_tax.sh script to get the taxonomic information for genes.

cmd COMMAND
```
chmod 755 centrifuge_tax.sh
qsub -o std-out/ -e std-err/ centrifuge_tax.sh
```

Go to the function-genes directory

Step 11.

Go to the function-genes directory. We are going to get the top pfam and kegg ids for each contig using the program uproc.

cmd COMMAND

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

Create the uproc_function.sh script

Step 12.

Create uproc_function.sh to run the functional analysis.

cmd COMMAND

```
#!/bin/bash

#PBS -W group_list=bh_class
#PBS -q windfall
#PBS -l jobtype=cluster_only
#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.faa"
OUT_DIR="/rsgrps/bh_class/username/function-genes"
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
```

ANNOTATIONS

Amy Hudson 09 Nov 2016

I am getting an Exit_status = 127 and my std-err says:

/rsgrps/bh_class/bin/uproc-dna: error while loading shared libraries: libuproc.so.2: cannot open shared object file: No such file or directory

/rsgrps/bh_class/bin/uproc-dna: error while loading shared libraries: libuproc.so.2: cannot open shared object file: No such file or directory

...

Di Ran 11 Nov 2016

I got the same errors.

Farideh Farahnak 13 Nov 2016

I have got the same errors

e s 15 Nov 2016

Same here

Bonnie Hurwitz 15 Nov 2016

Looks like the uproc binaries weren't installed in bh_class properly. We are fixing this.

Bonnie Hurwitz 15 Nov 2016

This has been fixed now. The due date for the assignment is now on Thursday.

Amy Hudson 16 Nov 2016

11/20 Ocelote doesn't work for qsub, Cluster/HTC/SMP seems to be working

_____

I'm now having errors for even submitting- anyone else?

When I log on to Ocelote and go to folder to run, this is the response:

qsub: Job rejected by all possible destinations

When I log on to sftp hpc

-bash: qsub: command not found

Run the uproc_function.sh script

Step 13.

Run the uproc_function.sh script

cmd COMMAND
```
chmod 755 uproc_function.sh
qsub -o std-out/ -e std-err/ uproc_function.sh
```

Convert the functional data into anvio format

Step 14.

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

cmd COMMAND
```
#! /uaopt/perl/5.14.2/bin/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";
}
```

Run format-anvio.pl

Step 15.

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
```

Download the functional data to your computer (where Anvi'o is running)

Step 16.

Download the functional data to your computer. Go to the directory where Anvi'o is running (on your computer) and type the following commands to download the files.

cmd COMMAND

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

Note the I use a "." here to mean the current directory in Anvi'o, on my computer this is: /Users/bhurwitz/skinmicrobiome

Download the taxonomic data to your computer (where Anvi'o is running)

Step 17.

Download the taxonomic data to your computer. Go to the directory where Anvi'o is running (on your computer) and type the following commands to download the files.

cmd COMMAND

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

Note the I use a "." here to mean the current directory in Anvi'o, on my computer this is: /Users/bhurwitz/skinmicrobiome

Upload the taxonomic data to the contigs.db

Step 18.

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
```

ANNOTATIONS

e s 17 Nov 2016

all I got is

File/Path Error: No such file: 'centrifuge_report.tsv' :/

Farideh Farahnak 17 Nov 2016

Hello,

You need to make a soft link from /usr/username/ on your MINGW environment to when you launch docker, here is the command:

$ docker run --rm -v ~:/my_data/ -it meren/anvio:latest

In here when I launch the Docker image on my desktop and type the above command on the prompt, I can see a soft link directory named "my_data" which gives me access to all the files I copied from HPC to my local machine, from that directory you can execute all the Anvio functionality.

Upload the functional data to the contigs.db

Step 19.

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
```

ANNOTATIONS

e s 17 Nov 2016

All I get is

File/Path Error: No such file: 'input_matrix.txt' :/

Document these steps in your google report

Step 20.

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?