C_HW10: Sample read count to functional categories for Anvi'o bar chart

Bonnie Hurwitz

Abstract

Create a script to add functional information about the samples into Anvi'o.

Citation: Bonnie Hurwitz C_HW10: Sample read count to functional categories for Anvi'o bar chart. **protocols.io** dx.doi.org/10.17504/protocols.io.gkibuue

Published: 29 Nov 2016

Protocol

login to the HPC

Step 1.

login to the HPC

cmd COMMAND

```
ssh hpc
ice
```

Create a new directory called function-reads

Step 2.

To get an idea of the broad functional categories in the samples we are going to use the reads to quantify hits to Kegg ids. Then we will group the hits by broad categories to include in a bar chart in Anvi'o for each sample. We are going to use a similar approach as we did before for annotating the function to genes (see previus protocol), except this time we will map the reads to kegg ids by using uproc.

First we need to create a diretory to run the analysis in:

cmd COMMAND

```
mkdir /rsgrps/bh_class/username/function-reads
cd /rsgrps/bh_class/username/function-reads
mkdir std-err
mkdir std-out
```

Create a script to run uproc on the reads

Step 3.

Now we need to create a script to run uproc on the reads for each of the samples. We will do this by running uproc on the reads that did not map to human (or the unmapped reads).

Create a script called: uproc_function.sh

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---------------
FASTQ_DIR="/rsgrps/bh_class/username/unmapped"
OUT_DIR="/rsgrps/bh_class/username/function-reads"
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 KEGG="$DATA/keggready"

cd $FASTQ_DIR
for file in `cat fastq-list`; do
   # filtered no human
   R1=$FASTQ_DIR/$file".paired.1.fastq"
   R2=$FASTQ_DIR/$file".paired.2.fastq"
   S=$FASTQ_DIR/$file".singletons.fastq"

   $UPROC --preds --short --threads 12 --
output $OUTPUT.$file.kegg $KEGG $UPROC_MODEL $R1 $R2 $S
done
```

Make sure your script is executable, and that no lines were split (the uproc command should be a single line).

Run the script on the HPC

Step 4.

Run the script above on the HPC to get read matches to kegg ids.

cmd COMMAND

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

Now write a script of your own to convert the output into a file for the Anvi'o SAMPLES.db

Step 5.

The output files will be called 'uproc-out.$file.kegg', where $file is one of your SRR file names. You will need to write a script that loops through each of these file names, converts the SRR file name into the body site 'short name' (e.g. Um for belly button), and then creates a read count for each of the broad kegg categories for each sample (use the file /rsgrps/bh_class/kegg_to_broadcat to convert the kegg ids in the uproc file to the broad categories). Also note that there are 31 categories, but this is way too many to display on our anvi'o graph. So the results should only display %reads that hit to the top five categories for all samples and group the rest into 'Other' (where Func1 is replaced by the 'broad category desc' in the output below).

Also remember that you will need to have a mapping file for converting the sample SRR id to the sample name, and linking to the categories for the output.

e.g.

SRR	Sample	Occlusion	MicroEnv
SRR1647143	Ra	Occluded	Sebaceous
SRR1647048	Sc	Exposed	Sebaceous
SRR1647047	Ax	Occluded	Moist
SRR1647142	Um	Occluded	Moist
SRR1647049	Fh	Exposed	Sebaceous
SRR1647046	Ac	IntOcculded	IntMoist
SRR1647045	Pa	Exposed	IntMoist
SRR1647141	Tw	Occluded	Moist

Note that this script should be similar to the script that Ken wrote in class to convert uproc output into input for Anvio. The main difference is that you need to create a summary table for the output.

EXPECTED RESULTS

Sample	Occlusion	MicroEnv	Func1	Func2	Func3	Func4	Func5	FuncOther
Ra	Occluded	Sebaceous	33	21	16	3	2	25
Sc	Exposed	Sebaceous	40	21	16	10	2	11
Ax	Occluded	Moist	30	18	5	10	1	36
Um	Occluded	Moist	28	13	6	10	25	18
Fh	Exposed	Sebaceous	27	21	12	10	2	28
Ac	IntOcculded	IntMoist	5	14	16	6	4	55
Pa	Exposed	IntMoist	6	21	16	10	2	45
Tw	Occluded	Moist	33	21	16	10	2	18

Download the output to your computer

Step 6.

Start Anvi'o as you have done in past protocols. Go into the directory with the anvi'o databases. Download the output (on the HPC) from the script above to your computer.

cmd COMMAND

```
scp sftp.hpc.arizona.edu:/rsgrps/bh_class/username/function-reads/samples-table .
```

where samples-table is the output of the script you wrote in the step above.

Upload the table into the SAMPLES.db in Anvi'o

Step 7.

Use the command below to create an anvi'o samples database.

cmd COMMAND

```
anvi-gen-samples-info-database --samples-information samples-table -o samples.db
```

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]

Start Anvi'o to see results

Step 8.

Start Anvi'o with the samples.db to visualize the results for the samples.

cmd COMMAND

```
anvi-interactive -p SAMPLES-MERGED/PROFILE.db -c contigs.db -s samples.db
```