⚘ 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.

Protocol

login to the HPC
Step 1.

login to the HPC

cmd COMMAND
```
ssh hpc
ocelote
```

Create a new directory called function-summary
Step 2.

We already mapped our reads to functional categories using uproc. Now, we are going to group these into counts of higher level kegg categories. We will do this by mapping the kegg ids for specific genes, into broad kegg categories that they are linked to. Then, we will use these broad categories to create a bar chart in Anvi'o for each sample.

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

cmd COMMAND
```
mkdir /rsgrps/bh_class/username/function-summary
cd /rsgrps/bh_class/username/function-summary
```

Create files for uploading the annotation data to Anvi'o SAMPLES.db
Step 3.

In class, Ken wrote a script to convert the uproc output into a format we can upload into Anvi'o. The script he wrote is located [here](#).

This script takes output files from uproc and converts them into a file that Anvi'o can use. The

resulting tab-delimited output file contains the gene id, kegg or pfam id, and definitions for kegg or pfam ids.

Please add to this script to create a summary table for kegg ids at a higher level catagory. The kegg ids are useful to look at gene by gene in the viewer, but are also important to summarize into higher level categories to find patterns in the data. To do this, you will need to count how many reads have kegg ids that match to higher level categories. Use the file /rsgrps/bh_class/kegg_to_broadcat to convert the kegg ids from the uproc data to the broad categories. Note that there are 31 total "broad" categories, but this is way too many to display on our anvi'o bar 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).

Anvio works better if you have short descriptive names. And also give Anvio information about the sample. Choose one or two major categories to include (as in Table 1 below). Also remember that you will need to have a mapping file for converting the sample SRR id into a short sample name, and linking to the categories for the output.

An example output file is shown below (called samples-table). Note the sample name should match what you have used in the past, this may be an SRR number or maybe you gave it a more descriptive name.

📈 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 4.

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.

Step 5.

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
```

✪ NOTES
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 6.

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
```