Script P5: Diversity

HANNIGAN GD, GRICE EA, ET AL.

Abstract

This protocol provides a method to predict phage community diversity using the algorithm PHACCS. Sequencing of viral communities often results in a high percent of unknown reads, largely due to our incomplete reference databases. To address this unknown factor of viromes, the algorithm PHACCS (PHAge Communities from Contig Spectra) was developed to predict virus community alpha diversity without the use of taxonomic references. Based on the methods from the following publication:

Hannigan, Geoffrey D., et al. "The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome." *mBio* 6.5 (2015): e01578-15.

Citation: HANNIGAN GD, GRICE EA, ET AL. Script P5: Diversity. **protocols.io**
dx.doi.org/10.17504/protocols.io.efrbbm6
Published: 10 Mar 2016

Guidelines

Required Software:

- PHACCS-1.1.3
- Circonspect-0.2.6
- GAAS-0.17
- Octave-3.8.1

Relevent Files

Output:

- Alpha_diversity/virus_and_phage_acc_numbers.txt
- Alpha_diversity/PHACCS_results_all.txt

R scripts: R5, R6, R10

Before start

Perl scripts and other supplemental information available at:

https://figshare.com/articles/The_Human_Skin_dsDNA_Virome_Topographical_and_Temporal_Diversity_Genetic_Enrichment_and_Dynamic_Associations_with_the_Host_Microbiome/1281248

Protocol

Virome Alpha Diversity

Step 1.

Make directory for the circonspect output files.

cmd COMMAND

```
mkdir ./negative_clean_subsample_150K_cironspect
```

NOTES

Geoffrey Hannigan 14 Jan 2016

Use randomly subsampled files here for speed purposes (150,000 subsampled). The majority of samples have more sequences than the subsampled depth, so it also provides normalization for the samples.

Virome Alpha Diversity

Step 2.

Run Circonspect to generate the contig spectra.

SOFTWARE PACKAGE (Unix)

Circonspect, 0.2.6

http://sourceforge.net/p/circonspect/code/ci/master/tree/

cmd COMMAND

```
ls ./neg_clean_subsample_150K | xargs -I {} --max-procs=16 Circonspect -o -
b ./negative_clean_subsample_150K_cironspect/{} -f ./neg_clean_subsample_150K/{}
```

Virome Alpha Diversity

Step 3.

In addition to the contig spectra, we need to use the program GAAS (suggested by PHACCS) to estimate the average virus/phage reference genome length for each sample.

Virome Alpha Diversity

Step 4.

First, download the list of all virus+phage accession numbers as a text file.

LINK:

http://www.ebi.ac.uk/genomes/virus.txt

Virome Alpha Diversity

Step 5.

Use the downloaded text file here to get all the fasta files.

LINK:

http://www.ebi.ac.uk/cgi-bin/sva/sva.pl?&do_batch=1

Virome Alpha Diversity

Step 6.

Upload the accession list and fasta file to the server. The final reference fasta is stored as ./reference/virus_and_phage_ref/virus_and_phage_ref.fasta

Virome Alpha Diversity

Step 7.

Run GAAS using the 150k sequence rarefied dataset.

SOFTWARE PACKAGE (Unix)

GAAS, 0.17

Florent Angly

http://sourceforge.net/p/gaas/code/ci/master/tree/

cmd COMMAND

```
mkdir ./GAAS_results_150k_subsample
mkdir ./tmp
ls ./neg_clean_subsample_150K | xargs -I {} --max-procs=4 mkdir ./tmp/{}
ls ./neg_clean_subsample_150K | xargs -I {} --max-
procs=4 mkdir ./GAAS_results_150k_subsample/{}
ls ./neg_clean_subsample_150K | xargs -I {} --max-procs=16 GAAS -gt 0 -v nucleic -e 1e-03 -
x ./tmp/{} -o ./GAAS_results_150k_subsample/{} -f ./neg_clean_subsample_150K/{} -
d ./reference/virus_and_phage_ref/virus_and_phage_ref.fasta
```

NOTES

Geoffrey Hannigan 14 Jan 2016

Need to make a set of unique tmp directories or else the procs all try to use the same directory with the same intermediate file names (this would be bad news!).

Geoffrey Hannigan 14 Jan 2016

Need to make a set of unique tmp directories or else the procs all try to use the same directory with the same intermediate file names (this would be bad news!).

Virome Alpha Diversity

Step 8.

Remove the tmp directory used for GAAS.

cmd COMMAND

```
rm ./tmp/
# WARNING: Always be careful when deleting directories, especially dirs with general names like
these.
```

Virome Alpha Diversity

Step 9.

Now that we have each sample's contig spectrum and average reference genome length, we can use PHACCS to estimate each sample's alpha diversity.

SOFTWARE PACKAGE (Unix)

PHACCS, 1.1.3

Florent Angly & Forest Rohwer

cmd COMMAND

```
run.phaccs.with.avg.genome.length () {
    # Set the variable number for the average genome length
    # Use echo to STDOUT as an easy way to make sure the vairables are set as expected
    echo Variable = ${1}
    export SHORT_FILENAME=$(echo ${1} | sed 's/\.fa.csp//')
    echo SHORT_FILENAME = ${SHORT_FILENAME}
    export GAAS_NUM=$(awk 'FNR == 2 {print $1}' ./GAAS_results_150k_subsample/${SHORT_FILEN
AME}.fa/gaas_${SHORT_FILENAME}_average.txt)
    echo GAAS_NUM = ${GAAS_NUM}
    octave --silent --path /appl/PHACCS-113/ --
eval "results = phaccs('./negative_clean_subsample_150K_cironspect/${1}', [], [], [], [], [
'$GAAS_NUM'], {'power'}, 1, 1000000, [], [], [], [1]); richness = results{1}.richness; even
ness = results{1}.evenness; most_abund = results{1}.most_ab; sw_index = results{1}.sw_index
; save './negative_clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/${1}_ric
hness.txt' richness; save './negative_clean_subsample_150K_phaccs_with_genome_lengths_highe
r_richness/${1}_evenness.txt' evenness; save './negative_clean_subsample_150K_phaccs_with_g
enome_lengths_higher_richness/${1}_most_abund.txt' most_abund; save './negative_clean_subsa
mple_150K_phaccs_with_genome_lengths_higher_richness/${1}_sw_index.txt' sw_index; exit"
}
export -f run.phaccs.with.avg.genome.length
```

NOTES

Geoffrey Hannigan 14 Jan 2016

Because PHACCS was written in MatLab, we used the open source MatLab alternative Octave to run the script. We also pulled out the diversity information from the PHACCS output and put it together into one easy-to-use file. This output was used in R analysis.

Geoffrey Hannigan 14 Jan 2016

Because PHACCS was written in MatLab, we used the open source MatLab alternative Octave to run the script. We also pulled out the diversity information from the PHACCS output and put it together into one easy-to-use file. This output was used in R analysis.

Virome Alpha Diversity

Step 10.

Make directory for the output.

cmd COMMAND

```
mkdir ./negative_clean_subsample_150K_phaccs_with_genome_lengths_higher_richness
ls ./negative_clean_subsample_150K_cironspect/*.csp | sed 's/.*\///g' | xargs -I {} --max-
procs=16 sh -c "run.phaccs.with.avg.genome.length {}"
```

Virome Alpha Diversity

Step 11.

Make a directory for the alpha diversity output.

cmd COMMAND

```
mkdir ./results_negative_clean_subsample_150K_phaccs_with_genome_lengths_higher_richness
```

Virome Alpha Diversity

Step 12.

Generate table with sample name in first column and diversity metric in the second column.

cmd COMMAND

```
head ./negative_clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/*_richness.
txt | grep -
v \# | sed '/^$/d' | sed s'/^.*\///g' | sed s'/_R1.*//' | awk 'NR % 2 {printf $0"\t"} !(NR
% 2) {print $0}' | sed s'/a/NA/g' | sed '1 s/^/SampleID\tRichness\n/' > ./results_negative_
clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/PHACCS_richness.txt
head ./negative_clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/*_evenness.
txt | grep -
v \# | sed '/^$/d' | sed s'/^.*\///g' | sed s'/_R1.*//' | awk 'NR % 2 {printf $0"\t"} !(NR
% 2) {print $0}' | sed s'/a/NA/g' | sed '1 s/^/SampleID\tEvenness\n/' > ./results_negative_
clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/PHACCS_evenness.txt
head ./negative_clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/*_most_abun
d.txt | grep -
v \# | sed '/^$/d' | sed s'/^.*\///g' | sed s'/_R1.*//' | awk 'NR % 2 {printf $0"\t"} !(NR
% 2) {print $0}' | sed s'/a/NA/g' | sed '1 s/^/SampleID\tMost_Abund\n/' > ./results_negativ
e_clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/PHACCS_most_abund.txt
head ./negative_clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/*_sw_index.
txt | grep -
v \# | sed '/^$/d' | sed s'/^.*\///g' | sed s'/_R1.*//' | awk 'NR % 2 {printf $0"\t"} !(NR
% 2) {print $0}' | sed s'/a/NA/g' | sed '1 s/^/SampleID\tSW_Index\n/' > ./results_negative_
clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/PHACCS_sw_index.txt
```

Virome Alpha Diversity

Step 13.

Paste together the resulting files.

cmd COMMAND

```
paste ./results_negative_clean_subsample_150K_phaccs_with_genome_lengths_higher_richness/PH
```

```
ACCS_richness.txt ./results_negative_clean_subsample_150K_phaccs_with_genome_lengths_higher
_richness/PHACCS_evenness.txt ./results_negative_clean_subsample_150K_phaccs_with_genome_le
ngths_higher_richness/PHACCS_most_abund.txt ./results_negative_clean_subsample_150K_phaccs_
with_genome_lengths_higher_richness/PHACCS_sw_index.txt | cut -
f 1,2,4,6,8 > ./results_negative_clean_subsample_150K_phaccs_with_genome_lengths_higher_ric
hness/PHACCS_results_all.txt
```

NOTES

Geoffrey Hannigan 14 Jan 2016

The resulting PHACCS diversity data can be found in PHACCS_results_all.txt. This data was analyzed according to Script R2.

Whole Metagenome

Step 14.

Whole metagenome alpha diversity is calculated in R using metaphlan taxonomy output.

Beta Diversity

Step 15.

Virome and whole metagenome beta diversity is calculated in R and uses the "OTU Table" format contig relative abundance table.