QIIME: VIROME comparinator tool diversity metrics

Barbra Ferrell

Abstract

Viral Informatics Resource for Metagenome Exploration (VIROME) is a bioinformatics pipeline that classifies viral metagenome sequences after searching against annotated reference sequence databases (i.e., UniRef, SEED, ACLAME, COG, GO, KEGG, and PhageSEED) and a custom environmental database, Metagenomes Online (MgOl). VIROME's compare tool allows users to compare any number of metagenomes to one another by generating a Biological Observation Matrix (biom) file that contain counts of occurrences of custom observations (comparison metrics), and a mapping text file that provides per-sample metadata (library IDs and descriptions).

Quantitative Insights Into Microbial Ecology (QIIME) is an open-source software pipeline that allows users to interpret raw sequencing data and create graphical displays to interact with the data. QIIME is modular, providing users with the flexibility to select and integrate various applications.

While QIIME has many capabilities and online tutorials, this SOP is limited to the steps a user will take to use VIROME's compare tool output (biom and mapping files) to explore composition (what is in a sample), alpha-diversity (diversity within a sample), and beta-diversity (comparison of diversity between samples). This SOP generally follows guidance in online tutorials but with modifications unique to the VIROME Comparinator output analysis.

Citation: Barbra Ferrell QIIME: VIROME comparinator tool diversity metrics. **protocols.io**
dx.doi.org/10.17504/protocols.io.ec9baz6
Published: 11 Jan 2016

Before start

Make sure you have QIIME installed or are using the QIIME virtual machine.

Tutorials are available online at:

http://qiime.org/tutorials/index.html

This SOP was drafted based on tutorials available at the time of writing, including the 454 Overview Tutorial:

http://qiime.org/tutorials/tutorial.html and

Illumina Overview Tutorial:

http://nbviewer.ipython.org/github/biocore/qiime/blob/1.9.1/examples/ipynb/illumina_overview_tutorial

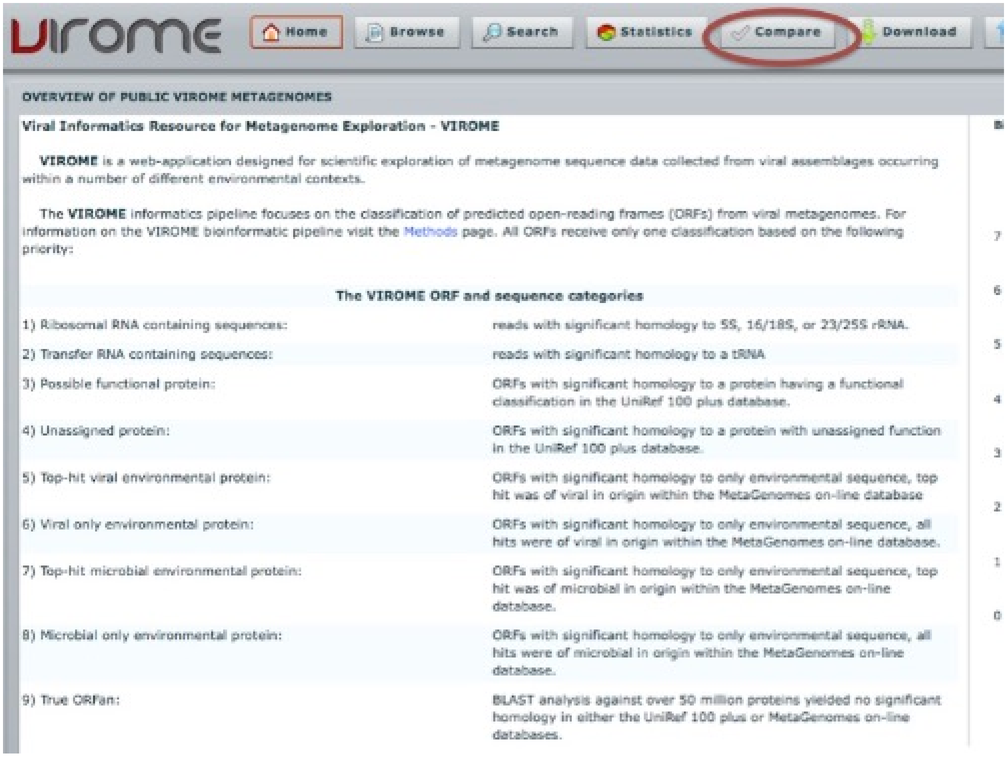
.ipynb

Protocol

Use VIROME's compare tool to generate biom and mapping files.

Step 1.

Visit the VIROME webapp site. Click the 'Compare' button on the top menu.



OVERVIEW OF PUBLIC VIROME METAGENOMES

Viral Informatics Resource for Metagenome Exploration - VIROME

VIROME is a web-application designed for scientific exploration of metagenome sequence data collected from viral assemblages occurring within a number of different environmental contexts.

The **VIROME** informatics pipeline focuses on the classification of predicted open-reading frames (ORFs) from viral metagenomes. For information on the VIROME bioinformatic pipeline visit the Methods page. All ORFs receive only one classification based on the following priority:

The VIROME ORF and sequence categories

1) Ribosomal RNA containing sequences:	reads with significant homology to 5S, 16/18S, or 23/25S rRNA.
2) Transfer RNA containing sequences:	reads with significant homology to a tRNA
3) Possible functional protein:	ORFs with significant homology to a protein having a functional classification in the UniRef 100 plus database.
4) Unassigned protein:	ORFs with significant homology to a protein with unassigned function in the UniRef 100 plus database.
5) Top-hit viral environmental protein:	ORFs with significant homology to only environmental sequence, top hit was of viral in origin within the MetaGenomes on-line database
6) Viral only environmental protein:	ORFs with significant homology to only environmental sequence, all hits were of viral in origin within the MetaGenomes on-line database.
7) Top-hit microbial environmental protein:	ORFs with significant homology to only environmental sequence, top hit was of microbial in origin within the MetaGenomes on-line database.
8) Microbial only environmental protein:	ORFs with significant homology to only environmental sequence, all hits were of microbial in origin within the MetaGenomes on-line database.
9) True ORFan:	BLAST analysis against over 50 million proteins yielded no significant homology in either the UniRef 100 plus or MetaGenomes on-line databases.

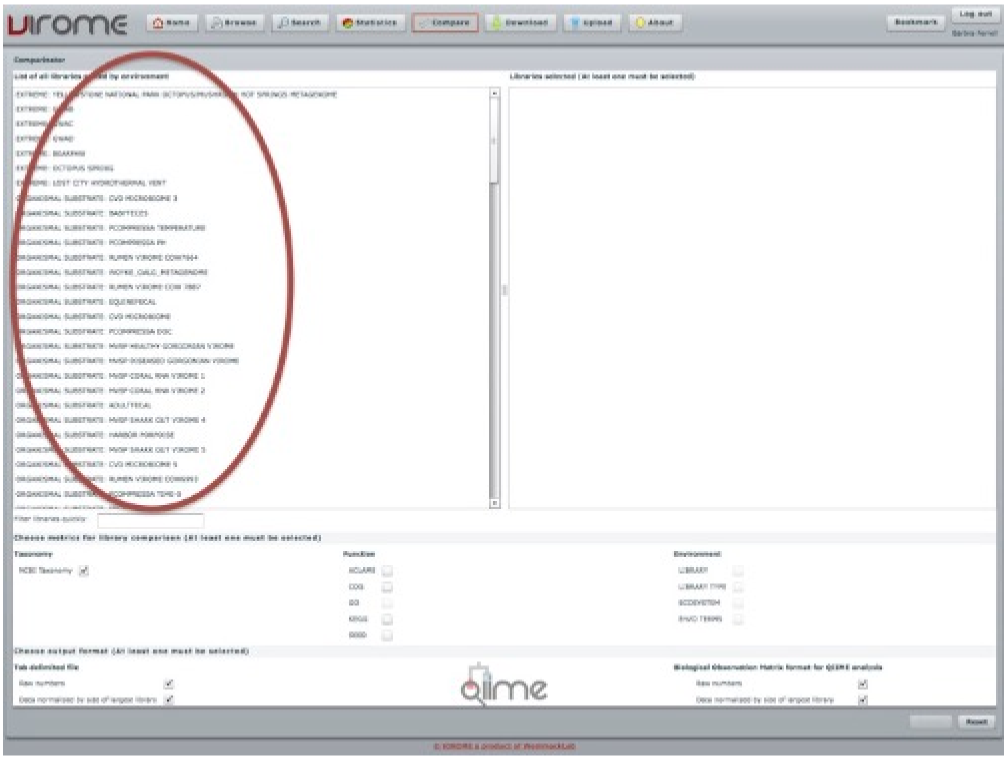
LINK:

http://virome.dbi.udel.edu

Use VIROME's compare tool to generate biom and mapping files.

Step 2.

Select libraries for comparison.



Comparinator

List of all libraries sorted by environment

EXTREME: YELLOWSTONE NATIONAL PARK OCTOPUS/MUSHROOM HOT SPRINGS METAGENOME
EXTREME: ESNB
EXTREME: DVAC
EXTREME: EWAD
EXTREME: BEAKMAN
EXTREME: OCTOPUS SPRING
EXTREME: LOST CITY HYDROTHERMAL VENT
ORGANISMAL SUBSTRATE: CVD MICROBIOME 3
ORGANISMAL SUBSTRATE: BABYTEETH
ORGANISMAL SUBSTRATE: PCOMPRESSA TEMPERATURE
ORGANISMAL SUBSTRATE: PCOMPRESSA PH
ORGANISMAL SUBSTRATE: RUMEN VIROME COW7664
ORGANISMAL SUBSTRATE: HOYKE_GALC_METAGENOME
ORGANISMAL SUBSTRATE: RUMEN VIROME COW 7887
ORGANISMAL SUBSTRATE: EQUINEFECAL
ORGANISMAL SUBSTRATE: CVD MICROBIOME
ORGANISMAL SUBSTRATE: PCOMPRESSA DOC
ORGANISMAL SUBSTRATE: MVSP HEALTHY GORGONIAN VIROME
ORGANISMAL SUBSTRATE: MVSP DISEASED GORGONIAN VIROME
ORGANISMAL SUBSTRATE: MVSP CORAL RNA VIROME 1
ORGANISMAL SUBSTRATE: MVSP CORAL RNA VIROME 2
ORGANISMAL SUBSTRATE: ADULTFECAL
ORGANISMAL SUBSTRATE: MVSP SHARK GUT VIROME 4
ORGANISMAL SUBSTRATE: HARBOR PORPOISE
ORGANISMAL SUBSTRATE: MVSP SHARK GUT VIROME 3
ORGANISMAL SUBSTRATE: CVD MICROBIOME 5
ORGANISMAL SUBSTRATE: RUMEN VIROME COW6993
ORGANISMAL SUBSTRATE: PCOMPRESSA TEMP-3

Libraries selected (At least one must be selected)

Filter libraries quickly:

Choose metrics for library comparison (At least one must be selected)

Taxonomy
NCBI Taxonomy ☑

Function
ACLAME □
COG □
GO □
KEGG □
SEED □

Environment
LIBRARY □
LIBRARY TYPE □
ECOSYSTEM □
ENVO TERMS □

Choose output format (At least one must be selected)

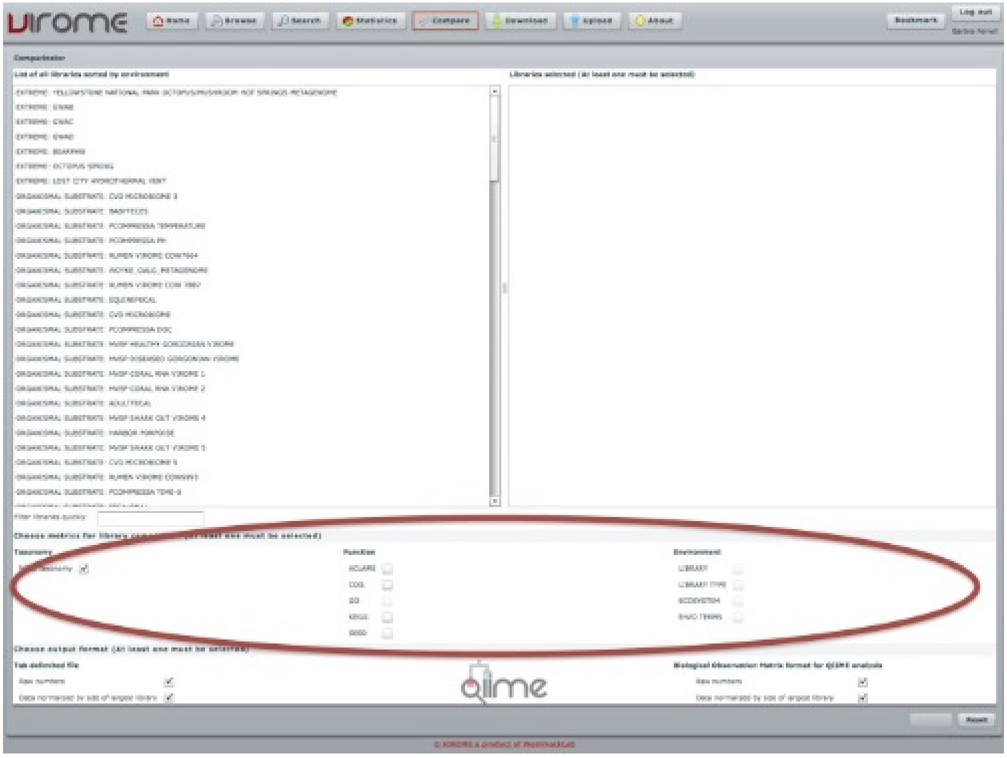
Tab delimited file
Raw numbers ☑
Data normalized by size of largest library ☑

Biological Observation Matrix format for QIIME analysis
Raw numbers ☑
Data normalized by size of largest library ☑

Use VIROME's compare tool to generate biom and mapping files.

Step 3.

Select metrics for comparison using the check boxes at the bottom of the page.



Comparinator

List of all libraries sorted by environment

EXTREME: YELLOWSTONE NATIONAL PARK OCTOPUS/MUSHROOM HOT SPRINGS METAGENOME
EXTREME: ESNB
EXTREME: DVAC
EXTREME: EWAD
EXTREME: BEAKMAN
EXTREME: OCTOPUS SPRING
EXTREME: LOST CITY HYDROTHERMAL VENT
ORGANISMAL SUBSTRATE: CVD MICROBIOME 3
ORGANISMAL SUBSTRATE: BABYTEETH
ORGANISMAL SUBSTRATE: PCOMPRESSA TEMPERATURE
ORGANISMAL SUBSTRATE: PCOMPRESSA PH
ORGANISMAL SUBSTRATE: RUMEN VIROME COW7664
ORGANISMAL SUBSTRATE: HOYKE_GALC_METAGENOME
ORGANISMAL SUBSTRATE: RUMEN VIROME COW 7887
ORGANISMAL SUBSTRATE: EQUINEFECAL
ORGANISMAL SUBSTRATE: CVD MICROBIOME
ORGANISMAL SUBSTRATE: PCOMPRESSA DOC
ORGANISMAL SUBSTRATE: MVSP HEALTHY GORGONIAN VIROME
ORGANISMAL SUBSTRATE: MVSP DISEASED GORGONIAN VIROME
ORGANISMAL SUBSTRATE: MVSP CORAL RNA VIROME 1
ORGANISMAL SUBSTRATE: MVSP CORAL RNA VIROME 2
ORGANISMAL SUBSTRATE: ADULTFECAL
ORGANISMAL SUBSTRATE: MVSP SHARK GUT VIROME 4
ORGANISMAL SUBSTRATE: HARBOR PORPOISE
ORGANISMAL SUBSTRATE: MVSP SHARK GUT VIROME 3
ORGANISMAL SUBSTRATE: CVD MICROBIOME 5
ORGANISMAL SUBSTRATE: RUMEN VIROME COW6993
ORGANISMAL SUBSTRATE: PCOMPRESSA TEMP-3

Libraries selected (At least one must be selected)

Filter libraries quickly:

Choose metrics for library comparison (At least one must be selected)

Taxonomy
NCBI Taxonomy ☑

Function
ACLAME □
COG □
GO □
KEGG □
SEED □

Environment
LIBRARY □
LIBRARY TYPE □
ECOSYSTEM □
ENVO TERMS □

Choose output format (At least one must be selected)

Tab delimited file
Raw numbers ☑
Data normalized by size of largest library ☑

Biological Observation Matrix format for QIIME analysis
Raw numbers ☑
Data normalized by size of largest library ☑

NOTES

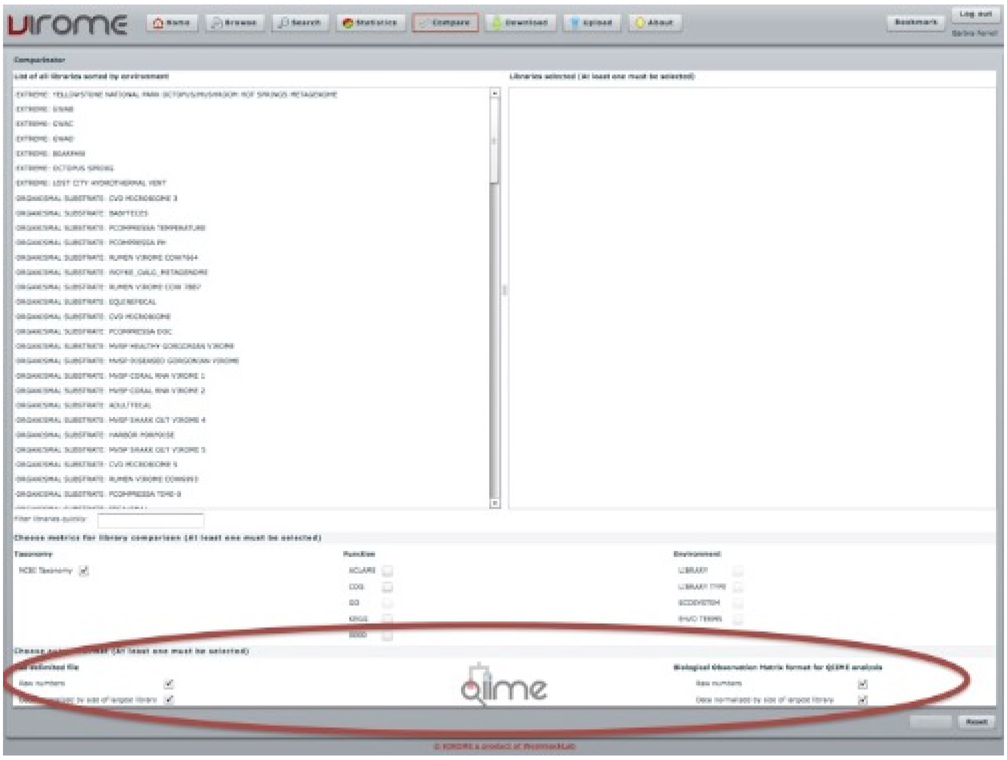
Barbra Ferrell 11 Jan 2016

The default metric is NCBI taxonomy. Multiple metrics can be selected at once.

Use VIROME's compare tool to generate biom and mapping files.

Step 4.

Select output file format.



NOTES

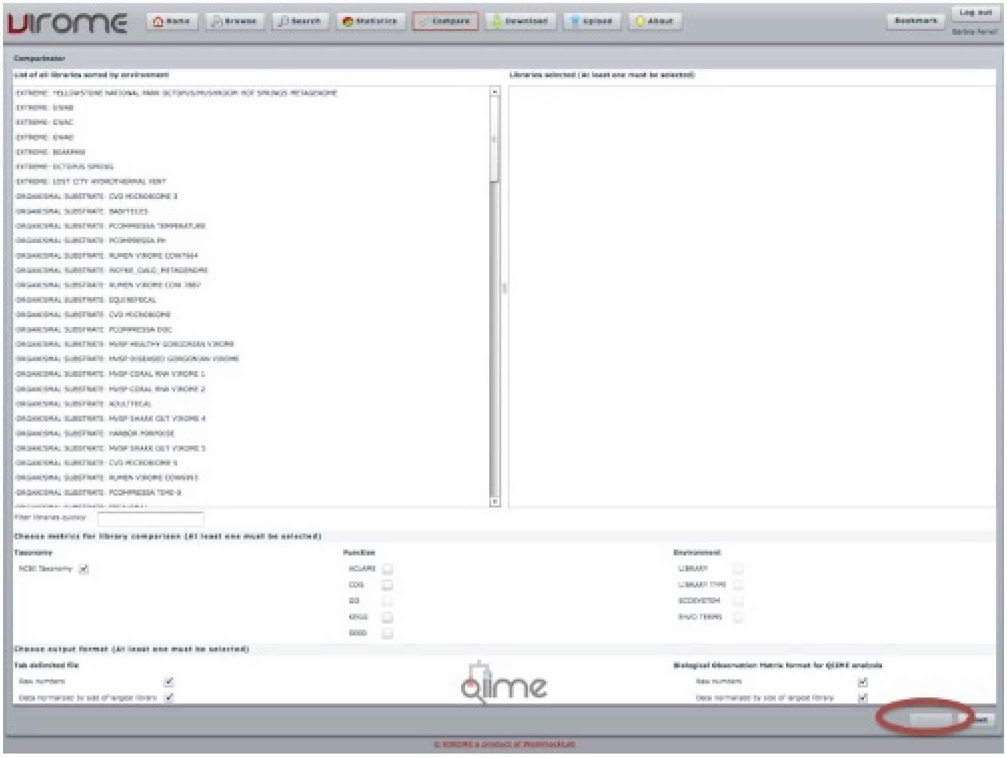
Barbra Ferrell 11 Jan 2016

The default file format is both tab-delimited and biom files for raw and normalized (individual ORF's count divided by the total number of predicted ORFs) counts. All tab-delimited and biom files for each metric are output to a single folder for that selected metric.

Use VIROME's compare tool to generate biom and mapping files.

Step 5.

Click Submit to compare.



NOTES

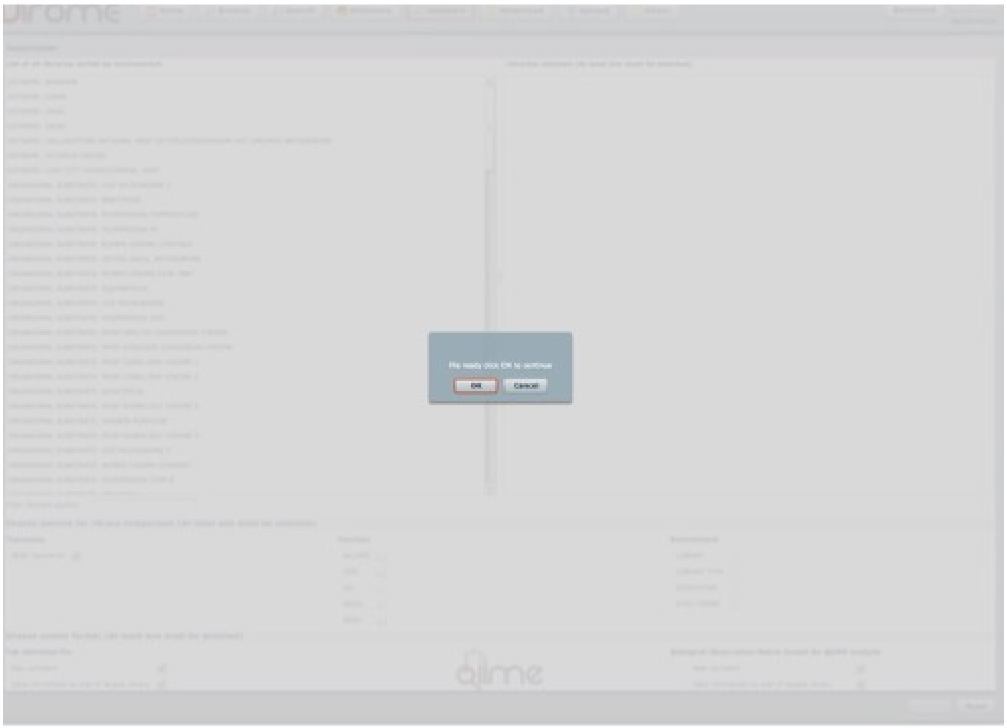
Barbra Ferrell 11 Jan 2016

File generation may take several minutes. When the files are ready, you will be asked to save them to your computer.

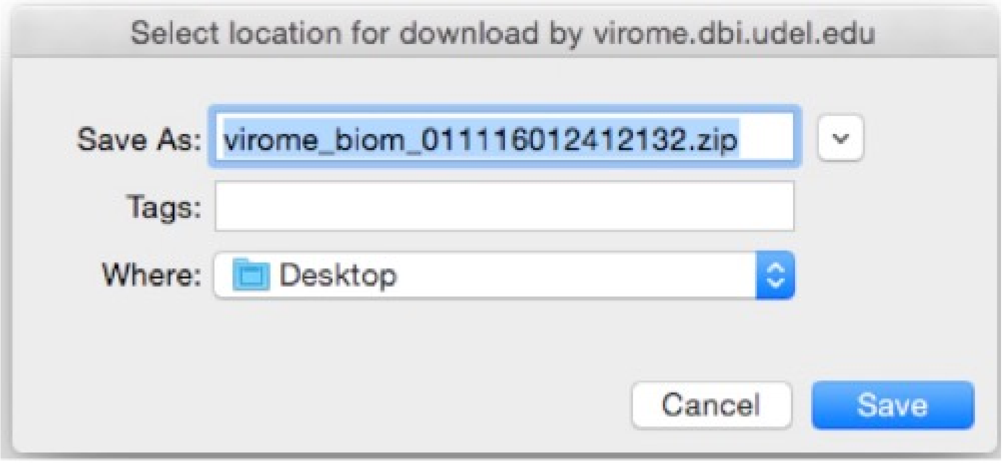
Use VIROME's compare tool to generate biom and mapping files.

Step 6.

Save files(s) to desktop as a compressed .zip folder.



File ready click OK to continue



Select location for download by virome.dbi.udel.edu

Save As: virome_biom_011116012412132.zip

Tags:

Where: Desktop

Prepare files for use in QIIME pipeline.

Step 7.

In a Terminal window, unzip the VIROME output folder.

cmd COMMAND

```
unzip VIROME_BIOM_123456789123456.zip
```

unzip the VIROME output folder and its contents

NOTES

Barbra Ferrell 11 Jan 2016

Note: you can also unzip the folder outside of the terminal window by double clicking on the compressed folder on your desktop. However, unzipping in this manner generates files that the computer sees as executable files, rather than readable text, tab de-limited, or biom files.

Prepare files for use in QIIME pipeline.

Step 8.

Make sure all files in the unzipped VIROME output are readable text or biom files. Each metric selected for comparison will have its own subfolder, and each subfolder will contain the raw and/or normalized text and/or biom files, and an additional VIROME-generated mapping text file requried for the QIIME analysis.

File permissions should read:
-rw-r--r--@

NOTES

Barbra Ferrell 11 Jan 2016

The file type can be easily viewed in your document Finder by navigating to the appropriate folder and checking the “Kind” of file listed in the document details (should be “text” instead of “Unix Executable File”). Alternatively, in Terminal, change to the appropriate VIROME output directory. Change to one of the subfolders, e.g. taxonomy:
cd taxonomy

Use the following command to view the contents of the VIROME output folder with folder permissions:
ls -l

File permissions should read:
-rw-r--r--@

If the contents are instead executable files, permission will read drwxr-xr-x@. To change permissions, first navigate back up to the VIROME output directory, then enter:
chmod -R 644 *

Enter QIIME working environment.

Step 9.

In Terminal, enter the QIIME working environment.

LINK:

http://www.wernerlab.org/software/macqiime

NOTES

Barbra Ferrell 11 Jan 2016

This assumes MacQIIME is already installed on your computer. MacQIIME can be downloaded and installed from http://www.wernerlab.org/software/macqiime

Alternatively, you can use QIIME loaded onto a server.

Barbra Ferrell 11 Jan 2016

In Terminal, enter

macqiime (if running on your local computer), OR
qiime (when running on a server)

Note that this should change your terminal prompt.

Enter QIIME working environment.

Step 10.

Change directories to the appropriate VIROME output subfolder, e.g. taxonomy.

Summarize libraries by composition.

Step 11.

Summarize libraries by composition.

cmd COMMAND

```
summarize_taxa_through_plots.py -i taxonomy.raw.biom -o taxonomy_raw_summary -m mapping.txt
```

Run the QIIME workflow to evaluate library composition based on VIROME calls. Modify this command based on the appropriate input file. User provides the name for the output folder, here shown corresponding to the input file.

NOTES

Barbra Ferrell 11 Jan 2016

This script will group OTUs by different classification levels, generating new tables at each level. Using a VIROME output, you will evaluate compositions/classifications based on VIROME calls, rather than from *de novo* OTU picking that you may perform using QIIME pipelines with other types of input (e.g. raw sequence data).

Barbra Ferrell 11 Jan 2016

You will need to pass the script the following information:

script command		summarize_taxa_through_plots.py
input file	-i	*input_biom_file*.biom
output folder	-o	*output_folder*
mapping file	-m	mapping.txt

For example:
summarize_taxa_through_plots.py -i taxonomy.raw.biom -o taxonomy_summary -m mapping.txt

Summarize libraries by composition.

Step 12.

Review output of composition summary. For visual evaluation of the composition area or bar charts, open the area_charts.html and bar_charts.html files located in the taxa_summary_plots folder.

cmd COMMAND

```
open taxonomy_raw_summary/taxa_summary_plots/area_charts.html

open taxonomy_raw_summary/taxa_summary_plots/bar_charts.html
```

Open the library composition area or bar plots chart, shown here when still in the original metric subfolder. Note that the taxa_summary_plots folder and charts are automatically named from within the QIIME script, so that portion of these commands should not need modification.

NOTES

Barbra Ferrell 11 Jan 2016

An output folder is created in the working directory. The output includes:

Log file

Plots folder (taxa_summary_plots)
Note that the name of the plots folder is assigned as part of the QIIME workflow, and is not a variable passed by the user through the command line. Regardless of the VIROME comparison metric (e.g., taxonomy, ACLAME, SEED, or ecosystem), the plots folder is called taxa_summary_plots.
Along with several subfolders, the plots folder contains two key files:
area_charts.html
bar_charts.html
Use the open command in Terminal to view these charts in a web browser.

Biom and txt files at each classification level.
These files can be opened in Excel for running additional statistics.

Alpha-diversity

Step 13.

Create a parameters file for alpha diversity and save in the top level VIROME output folder that contains:
alpha_diversity:metrics shannon,chao1,observed_otus

NOTES

Barbra Ferrell 11 Jan 2016

The alpha_rarefaction.py workflow includes an alpha_diversity.py script, which determines alpha diversity based on several parameters. The default parameters are PD_whole_tree, chao1, and observed otus. To confirm default parameters, on the command line enter:
alpha_diversity.py -h
and view the list of metrics. However, the PD_whole_tree (phylogenetic diversity whole tree) metric requires a phylogenetic tree .tre file, which is not part of the VIROME output (generation of a .tre file is part of the QIIME pipeline to generate OTUs, which is skipped since we are moving ahead with OTUs called by VIROME). Therefore, with a VIROME output, you must create a separate parameters .txt file in order to modify the default alpha diversity metrics.

Barbra Ferrell 11 Jan 2016

Create a plain text file (i.e., alpha_params.txt) in the high level VIROME output folder that contains:
alpha_diversity:metrics shannon,chao1,observed_otus
A list of other metrics is available by entering on the command line
alpha_diversity.py -s
but are not included as part of this SOP and may required additional input files (e.g. the phylogenetic tree file).

NOTE: the alpha_params.txt file is not necessarily unique to each VIROME comparison metric - you may choose to run shannon, chao1, and observed_otus for each comparison metric (i.e., each VIROME subfolder). Therefore, you can generate a single generic alpha_params.txt file to place in the top level VIROME output folder, and refer to it for each alpha diversity analysis you choose to perform.

Alpha-diversity

Step 14.

Compute alpha-diversity, or within-sample diversity, and generate alpha rarefaction plots.

cmd COMMAND

```
alpha_rarefaction.py -i taxonomy_raw_biom -o taxonomy_raw_alpha -m mapping.txt -
p ,./alpha_params.txt
```

Run the QIIME workflow to evaluate alpha diversity based on VIROME calls. Modify this command based on the appropriate input file. User provides the name for the output folder, here shown corresponding to the input file. User must modify the default parameters for alpha diversity, pointing to a parameters file that here is shown in the master VIROME output folder.

NOTES

Barbra Ferrell 11 Jan 2016

You will need to pass the script the following information:

script command		alpha_rarefaction.py
input file	-i	*input_biom_file*.biom
output folder	-o	*output_folder*
mapping file	-m	mapping.txt
parameter file	-p	*parameter_file*.txt

For a single generic alpha_params.txt file in your VIROME output foler:

```
alpha_rarefaction.py -i taxonomy.raw.biom \
-o taxonomy_alpha -m mapping.txt -p ../alpha_params.txt
```

For individual parameters file within each VIROME output subfolder:

```
alpha_rarefaction.py -i taxonomy.raw.biom \
-o taxonomy_alpha -m mapping.txt -p taxonomy_alpha_params.txt
```

Barbra Ferrell 11 Jan 2016

Other considerations:

There are several other parameters of interest in the alpha_rarefaction.py workflow, primarily num_steps, min_rare_depth, and max_rare_depth. The multiple_rarefactions.py script within the workflow rarefies (subsamples) the OTU table, creating a series of subsampled OTU tables by random sampling without replacement. The alpha_rarefaction.py workflow uses the following defaults:

min_rare_depth, default: 10
max_rare_depth, default: median sequence/sample count
num_steps, default: 10

While these parameters may be changed on the command line by entering -n, --min_rare_depth, or -e, respectively, the cases in which you may want to modify those defaults are not individually discussed in this SOP.

Alpha-diversity

Step 15.

Review output of alpha diversity analysis. For visual evaluation of the alpha rarefaction plots, open the rarefaction_plots.html file located in the alpha_rarefaction_plots folder.

cmd COMMAND

```
open taxonomy_raw_alpha/alpha_rarefaction_plots/rarefaction_plots.html
```

Open the alpha rarefaction plots, shown here when still in the original metric subfolder. Note that the alpha_rarefaction_plots folder and charts are automatically named from within the QIIME script, so that portion of this command should not need modification.

NOTES

Barbra Ferrell 11 Jan 2016

An output folder is created in the working directory based on the output folder name you passed in the command line (e.g., taxonomy_alpha). Many intermediate files generated through the alpha_rarefaction.py workflow are not retained. The output includes:

Log file

Folder (alpha_div_collated) with collated alpha diversity measurement files
There will be one .txt file for each alpha diversity metric used. Note that the QIIME workflow automatically assigns the name of each file based on the corresponding metric.

Folder (alpha_rarefaction_plots) with rarefaction plots
The make_rarefaction_plots.py scripts creates plots of alpha diversity vs. simulated sequence effort, known as rarefaction plots. Each curve represents a sample/library, and can be grouped by the sample metadata supplied in the mapping file (in this case, only by the library ID or description as provided by the VIROME mapping file).
The subfolder **average_tables** contains .png files of the average plots for each metric and category.
The **rarefaction_plots.html** file can be opened with a web browser by
open rarefaction_plots.html
Once the browser window is open, various metrics and categories can be selected to view plots.

Beta-diversity

Step 16.

Create a parameters file for beta diversity and save in the top level VIROME output folder that contains:
beta_diversity:metrics bray_curtis

NOTES

Barbra Ferrell 11 Jan 2016
Create a plain text file (i.e., beta_params.txt) in the high level VIROME output folder that contains:
beta_diversity:metrics bray_curtis
A list of other metrics is available by entering on the command line
beta_diversity.py -s
but are not included as part of this SOP and may required additional input files (e.g. the phylogenetic tree file).

NOTE: the beta_params.txt file is not necessarily unique to each VIROME comparison metric - you may choose to run bray_curtis for each comparison metric (i.e., each VIROME subfolder). Therefore, you can generate a single generic beta_params.txt file to place in the top level VIROME output folder, and refer to it for each beta diversity analysis you choose to perform.

Beta-diversity

Step 17.

Compute beta diversity, or between-sample diversity, and generate ordination (Principal Component Analysis, PCoA) plots.

COMMAND

```
beta_diversity_through_plots.py -i taxonomy.raw.bio -o taxonomy_raw_beta -m mapping.txt -
p ../beta_params.txt
```

Run the QIIME workflow to evaluate beta diversity based on VIROME calls. Modify this command based on the appropriate input file. User provides the name for the output folder, here shown corresponding to the input file. User must modify the default parameters for beta diversity, pointing to a parameters file that here is shown in the master VIROME output folder.

NOTES

Barbra Ferrell 11 Jan 2016

You will need to pass the script the following information.

script command		beta_diversity_through_plots.py
input file	-i	*input_biom_file*.biom
output folder	-o	*output_folder*
mapping file	-m	mapping.txt
parameter file	-p	*parameter_file*.txt

For a single generic beta_params.txt file in your VIROME output folder:

```
beta_diversity_through_plots.py -i taxonomy.raw.biom \
-o taxonomy_beta -m mapping.txt -p ../beta_params.txt
```

For individual parameters files within each VIROME output subfolder:

```
beta_diveristy_through_plots.py -i taxonomy.raw biom \
-o taxonomy_beta -m mapping.txt -p taxonomy_beta_params.txt
```

Beta-diversity

Step 18.

Review output of beta diversity analysis. For visual evaluation of the beta diversity PCoA plots, open the index.html file located in the *bray_curtis*_emperor_pcoa_plot folder. Note that the name of this output subfolder is dependent on the metrics identified in the beta parameters text file.

COMMAND

```
open taxonomy_raw_beta/bray_curtis_emperor_pcoa_plot/index.html
```

Open the beta diversity Principal Coordinate Analysis PCoA plots, shown here when still in the original metric subfolder. Note that the *_emperor_pcoa_plot folder is automatically named but is based on each beta diversity metric identified in the beta parameters text file. This portion of the command may need to be modified based on selected metrics.

NOTES

Barbra Ferrell 11 Jan 2016

Note that you may get an error related to a "VisibleDeprecationWarning". Many QIIME attributes or metrics are periodically updated - this warning is related to one of those updates and does not indicate an error in the beta_diversity_through_plots.py execution.

An output folder is created in the working directory based on the output folder name you passed in the command line (e.g., taxonomy_beta). The output includes:

Log file

Distance matrix file for each metric (ending in _dm.txt, e.g., bray_curtis_dm.txt)
There will be one *_dm.txt file for each beta diversity metric used. Note that the QIIME workflow automatically assigns the name of each file based on the corresponding metric.

Text file for each metric (ending in _pc.txt, e.g., bray_curtist_pc.txt)
There will be one *_pc.txt file for each beta diversity metric used, which contains the results for the Principal Coordinates Analysis (PCoA). Note that the QIIME workflow automatically assigns the name of each file based on the corresponding metric.

Folder for each metric with PCoA plots (ending in _emperor_pcoa_plot, e.g., bray_curtis_emperor_pcoa_plot)

Principal Coordinates Analysis (PCoA) helps extract and visualize a few components of variation from complex, multidimensional (multi-component) data. PCoA plots map the samples/libraries present to axes such that the first principal coordinate (PC1) explains the maximum amount of variation between samples/libraries, the second (PC2) contains the second largest amount of variation, etc. Emperor, a tool for visualizing high throughput community data, allows for the visualization of PCoA plots in three dimensions.

There will be one *_emperor_pcoa_plot folder for each beta diversity metric used, which contains a subfolder with information required for the Emperor visualization and an **index.html** file. Use the command

```
open index.html
```

within the plots subfolder to view the PCoA plots.

When viewing the PCoA plots, each SampleID in the mapping.txt file (each library from VIROME) will get a different color. Since the VIROME output does not break each library down further, the PCoA plots will have only one dot per library.