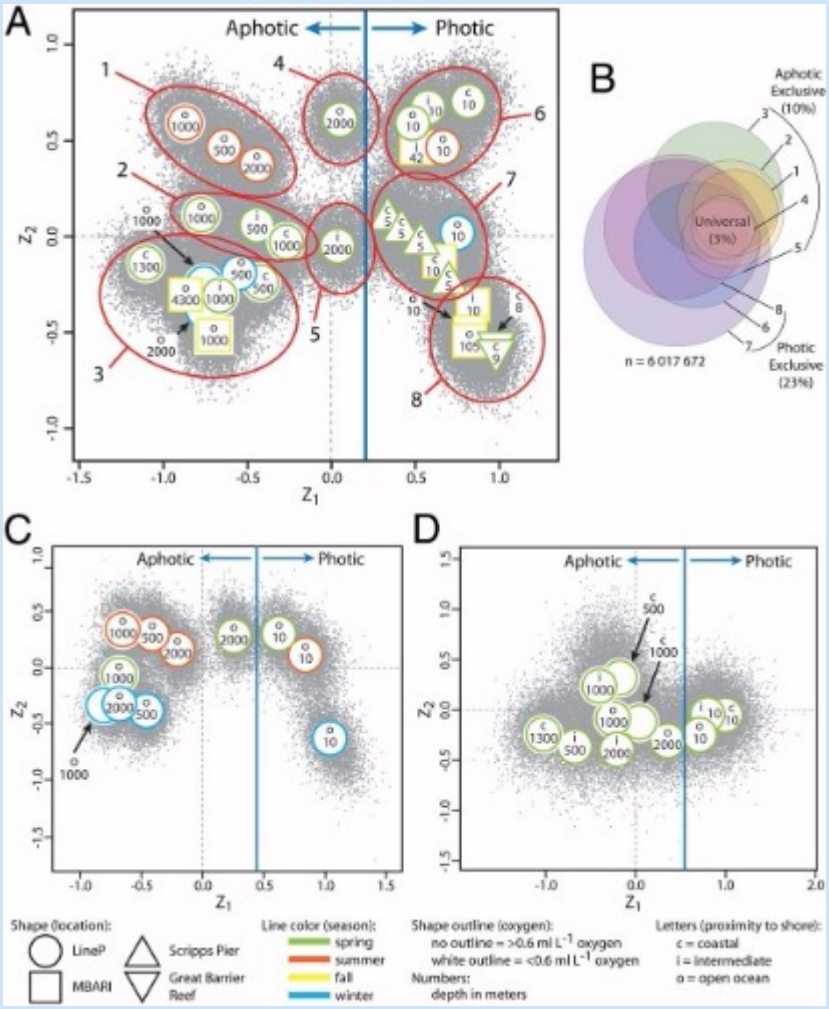
Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses

Bonnie Hurwitz, Ken Youens-Clark

Abstract

Long-standing questions in marine viral ecology are centered on understanding how viral assemblages change along gradients in space and time. However, investigating these fundamental ecological questions has been challenging due to incomplete representation of naturally occurring viral diversity in single gene- or morphology-based studies and an inability to identify up to 90% of reads in viral metagenomes (viromes). In this protocol, I describe how to use an annotation- and assembly-free strategy for comparative metagenomics that combines shared k-mer and social network analyses (regression modeling). This robust statistical framework enables visualization of complex sample networks and determination of ecological factors driving community structure. This tutorial describes a protocol to reproduce work from the Pacific Ocean virome comprised of 32 viromes from diverse sites in the Pacific Ocean.



"Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses" (July 7, 2014, doi: 10.1073/pnas.1319778111, PNAS July 22, 2014 vol. 111 no. 29 10714-10719)

Code is freely available at Github.

Citation: Bonnie Hurwitz, Ken Youens-Clark Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. **protocols.io**

dx.doi.org/10.17504/protocols.io.efgbbjw

Published: 13 Jan 2016

Protocol

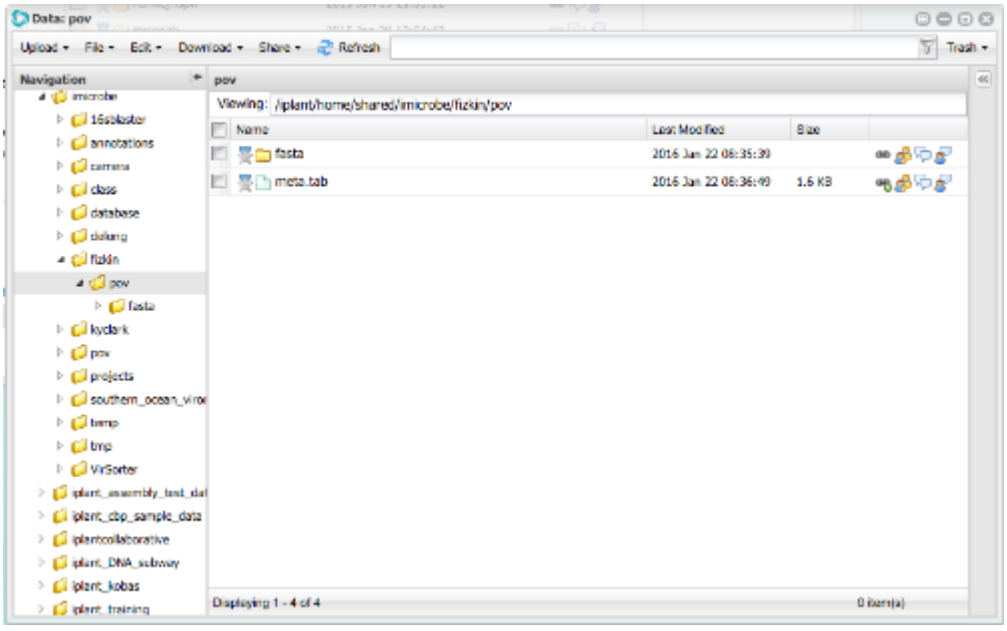
Step 1.

Log into to iPlant/CyVerse (http://www.cyverse.org/, http://de.iplantcollaborative.org) Discovery Environment.

Step 2.

Upload FASTA-formatted sequence files and a tab-delimited file of metadata. Example data can be found in the Data Store at "/iplant/home/shared/imicrobe/fizkin/pov." To view in the Discovery Environment:

- Click on the "Data" button in the DE
- Go to the "Community Data -> imicrobe -> fizkin -> pov -> fasta" directory



A sample metadata file is also included ("meta.tab"). The headers of the metadata file should include

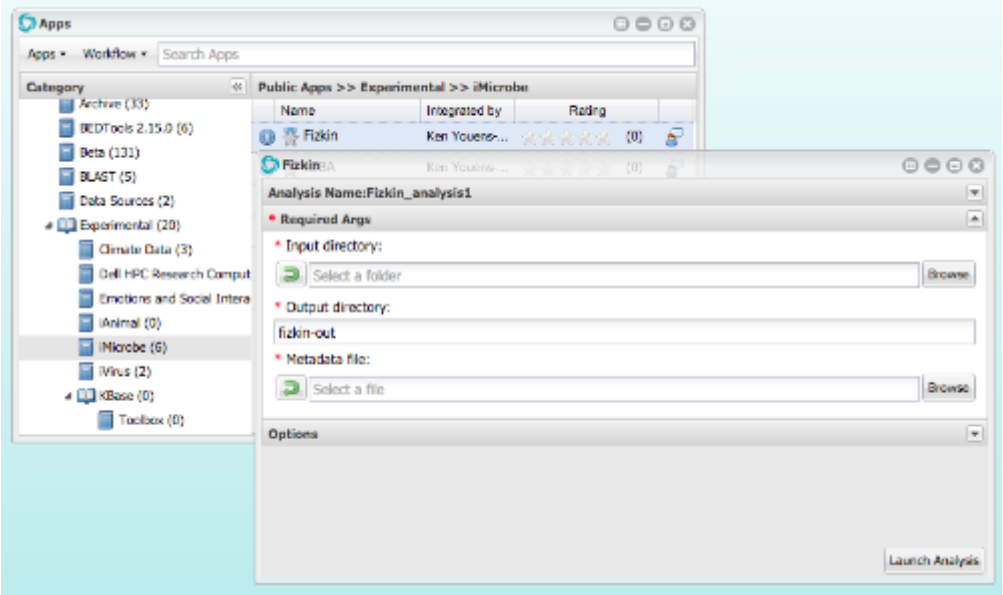
the 'name' of the file and fields ending in '.d' for 'discrete' value (e.g., 'Male' or 'Female'), '.c' for 'continuuous' data (e.g., numbers in a range), or '.ll' for 'latitude/longitude' data. Field names should not include underscores with the exception of 'lat_long.ll.'

Here is an example table:

name	lat_long.ll	biome.d	depth.c	season.d	transect.d
GD.Spr.C.8m.fa	-17.92522,146.14295	G	8	Spr	C
GF.Spr.C.9m.fa	-16.9207,145.9965833	G	9	Spr	C
L.Spr.C.1000m.fa	48.6495,-126.66434	L	1000	Spr	C
L.Spr.C.10m.fa	48.6495,-126.66434	L	10	Spr	C
L.Spr.C.1300m.fa	48.6495,-126.66434	L	1300	Spr	C
L.Spr.C.500m.fa	48.6495,-126.66434	L	500	Spr	C
L.Spr.I.1000m.fa	48.96917,-130.67033	L	1000	Spr	I
L.Spr.I.10m.fa	48.96917,-130.67033	L	10	Spr	I
L.Spr.I.2000m.fa	48.96917,-130.67033	L	2000	Spr	I
L.Spr.I.500m.fa	48.96917,-130.67033	L	500	Spr	I
L.Spr.O.1000m.fa	50.00167,-144.99899	L	1000	Spr	O
L.Spr.O.10m.fa	50.00167,-144.99899	L	10	Spr	O
L.Spr.O.2000m.fa	50.00167,-144.99899	L	2000	Spr	O
L.Sum.O.1000m.fa	50.00167,-144.99899	L	1000	Sum	O
L.Sum.O.10m.fa	50.00167,-144.99899	L	10	Sum	O
L.Sum.O.2000m.fa	50.00167,-144.99899	L	2000	Sum	O
L.Sum.O.500m.fa	50.00167,-144.99899	L	500	Sum	O
L.Win.O.1000m.fa	50.00167,-144.99899	L	1000	Win	O
L.Win.O.10m.fa	50.00167,-144.99899	L	10	Win	O
L.Win.O.2000m.fa	50.00167,-144.99899	L	2000	Win	O
L.Win.O.500m.fa	50.00167,-144.99899	L	500	Win	O
M.Fall.C.10m.fa	36.79683,-121.84667	M	10	Fall	C
M.Fall.I.10m.fa	36.12633,-123.4905	M	10	Fall	I
M.Fall.I.42m.fa	36.12633,-123.4905	M	42	Fall	I
M.Fall.O.1000m.fa	33.28683,-129.42833	M	1000	Fall	O
M.Fall.O.105m.fa	33.28683,-129.42833	M	105	Fall	O
M.Fall.O.10m.fa	33.28683,-129.42833	M	10	Fall	O
M.Fall.O.4300m.fa	33.28683,-129.42833	M	4300	Fall	O
SFC.Spr.C.5m.fa	32.86667,-117.25111	S	10	Spr	C
SFD.Spr.C.5m.fa	32.86667,-117.25111	S	10	Spr	C
SFS.Spr.C.5m.fa	32.86667,-117.25111	S	10	Spr	C
SMS.Spr.C.5m.fa	32.86667,-117.25111	S	10	Spr	C
STC.Spr.C.5m.fa	32.86667,-117.25111	S	10	Spr	C

Step 3.

Select the "Apps" button on the left, then look under "Public Apps -> Experimental -> iMicrobe -> Fizkin." Open the "Required Args" section and select your FASTA directory as the "Input directory." You can leave "Output directory" alone or change it if you wish. Use the file selector to find your "Metadata file" described in step 2.



Optional args:

- K-mer size: Default is 20. Values between 16 and 31 are best.
- Mode minimum: Default is 1. Increase to require more stringent matching.
- Max. num. sequences: Default is 300K. Use a lower value to reduce runtime. Use a higher value to get deeper coverage. Samples containing more than this parameter will be randomly sampled.
- Max. num. samples: Default is 15. Keep in mind that Fizkin runs a pair-wise analysis, so runtime is O(n^2). If your number of samples is greater than this argument, the samples will be randomly selected.
- Files list: The subset of files you wish to run, one file on each line

Step 4.
Press "Launch analysis" and wait for notification of the completion of your job.

Step 5.
Common failures include something like this from R (GBME):

Error in summary(fit1)$cov.unscaled[(2 * n):length(fit1$coef), (2 * n):length(fit1$coef)] :
subscript out of bounds
Calls: gbme -> gbme.glmstart -> as.matrix
Execution halted

This is usually due to the metadata being too homogenous or entirely heterogenous. Remove any offending metadata and try again.

Step 6.
The ultimate result should be a social network graph showing the grouping of samples similar to this: