PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information

Florent Angly, **Beltran Rodriguez-Brito**, **David Bangor**, **Pat McNairnie**, **Mya Breitbart**,

Peter Salamon, **Ben Felts**, **James Nulton**, **Joseph Mahaffy, Forest Rohwer**

by Yongan Zhao

Outline

- Background
- Goals
- Methods
- Results
- Conclusions

Background

- Importance of phages
    1. A phage is a virus infecting a prokaryote.
    2. Most abundant biological entities with an estimated 10^31 viral particles.
    3. Strong impact on microbial community biomass and structure because of their specific predation.
    4. Little is known about phage biodiversity.

- What's the biodiversity
    1. Richness: total number of different species.
    2. Evenness: the relative abundance of each species.

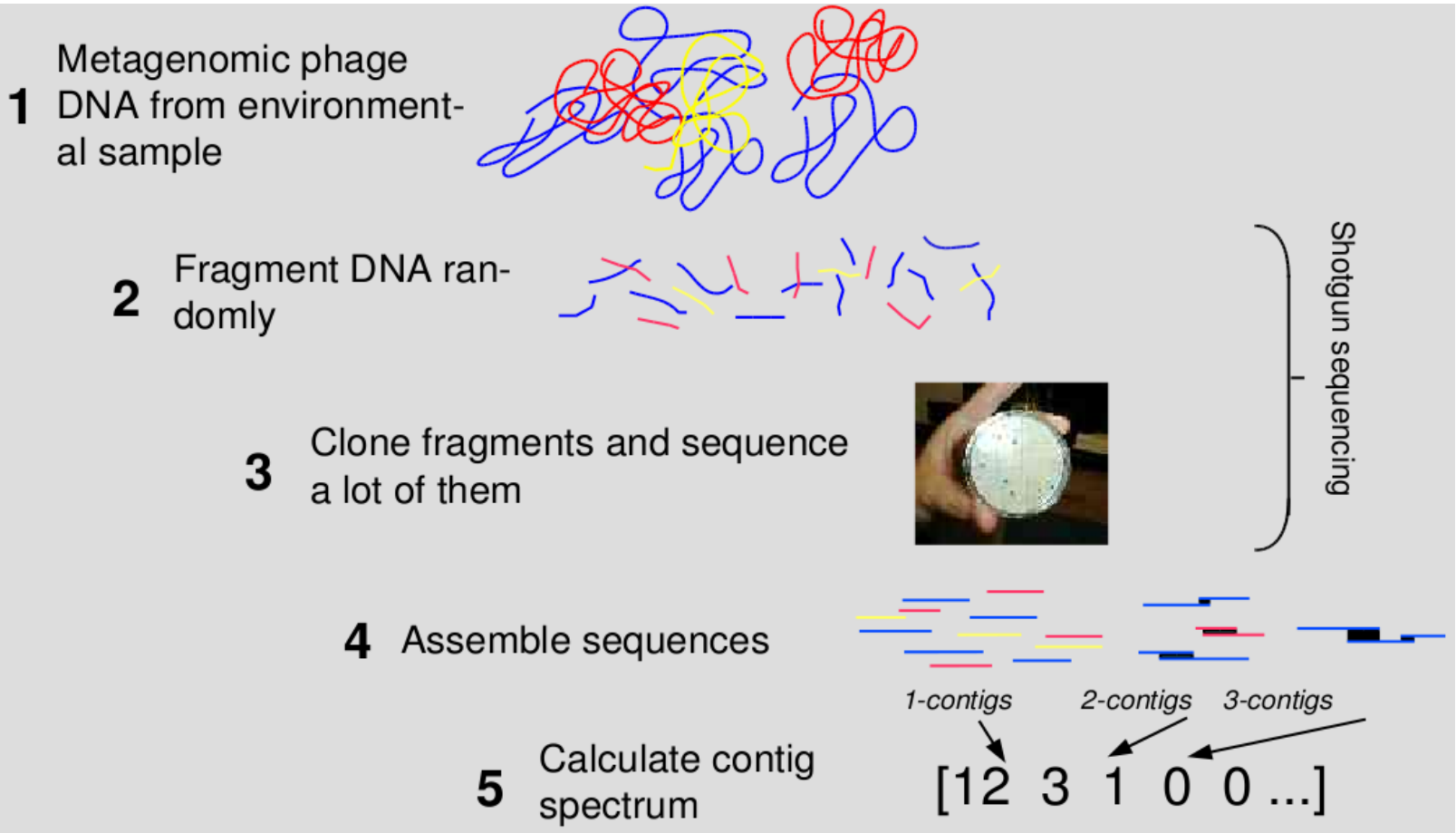
Goals

- PHACCS finds the most appropriate structure model by optimizing the model parameters until the predicted contig spectrum is as close as possible to the experimental one.

- Assess the biodiversity of uncultured viral communities
  1. Rank-abundance form
  2. Richness
  3. Evenness
  4. Shannon-Wiener index

What's the contig spectrum

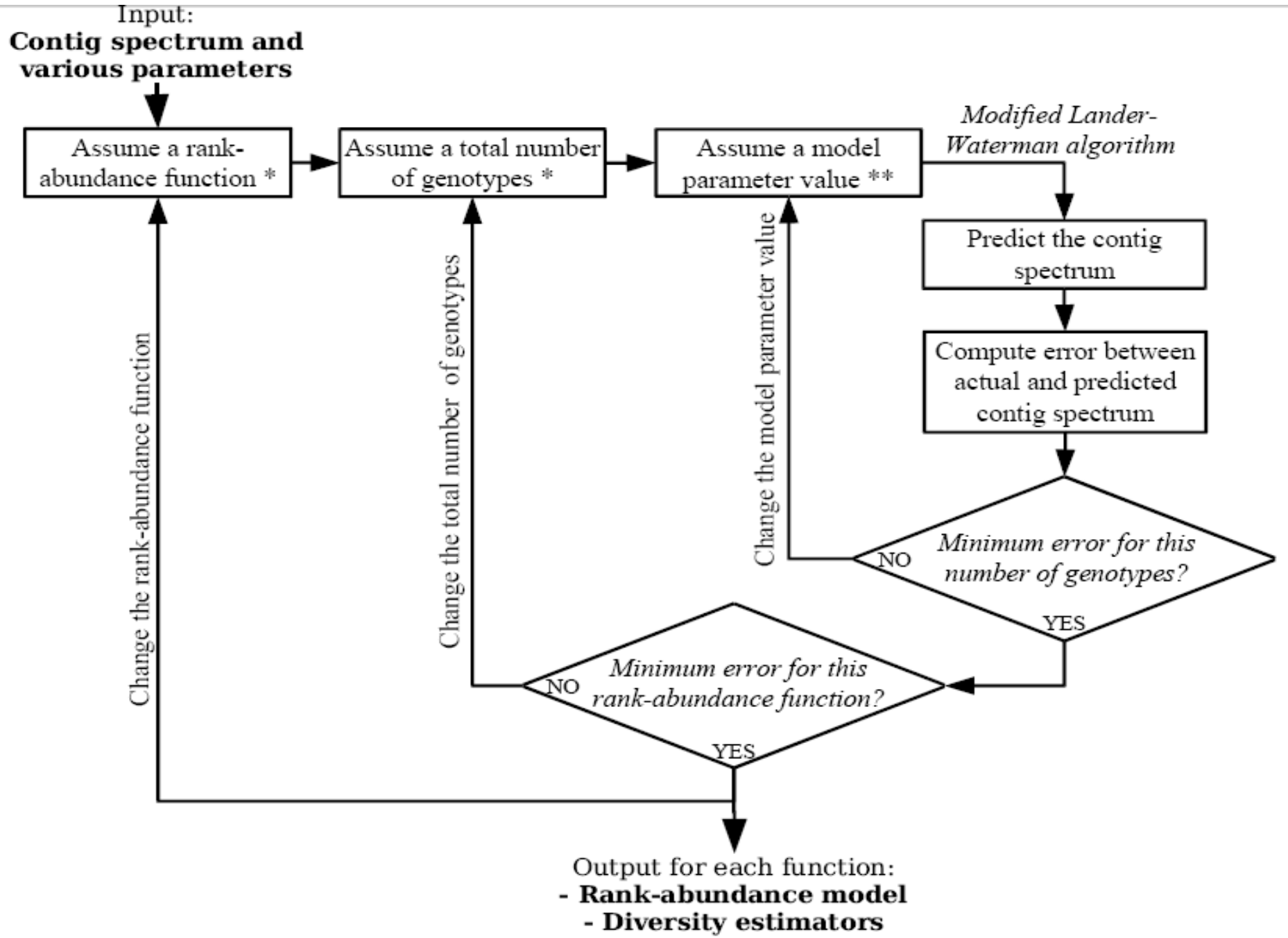
- The contig spectrum is a vector containing the number of contigs of size q .



12 means 12 contigs containing one reads;
3 means 3 contigs containing 2 reads;
1 means 1 contig containing 3 reads.

biome.sdsu.edu/phaccs/data/**PHACCS_presentation.pdf**

Workflow


Input:
Contig spectrum and various parameters
Assume a rank-abundance function *
Assume a total number of genotypes *
Assume a model parameter value **
Modified Lander-Waterman algorithm
Predict the contig spectrum
Compute error between actual and predicted contig spectrum
Minimum error for this number of genotypes?
NO
YES
Change the model parameter value
Minimum error for this rank-abundance function?
NO
YES
Change the total number of genotypes
Change the rank-abundance function
Output for each function:
- Rank-abundance model
- Diversity estimators

Rank-abundance models

Six basic functional forms of relative rank-abundance

- Empirical models: power law, logarithmic, exponential
- Ecological models: broken stick, niche preemption
- Lognormal distributions

They model the relative frequency of each genotype.

- For example, power law: $n_i = a\, i^{-b}$,

  n_i is the relative frequency of the ith genotype, a is relative abundance of the most abundant genotype, i is the rank index from 1 to total number of genotypes, and b is the evenness parameter

Breitbart M, Salamon P, Andresen P, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F: **Genomic analysis of uncultured marine viral communities.** *Proc Natl Acad Sci USA* 2002, **99:**14250-5.

Lander-Waterman algorithm for one genotype

background: the probability that two starting points will be within a distance x of each other is 1 - exp(-ax), where 1/a is the average distance between starting points.

- P(*two starting point will be within a distance s-o*):1-exp(-N*(s-o)/L)
- P(*a random selected fragment is part of a q-contig*):$w_q = q\, p^{q-1} (1-p)^2$
- The expected number of q-contig member: $c_q = N w_q$

N is the number of fregments, L is the length of the genome, s is the average length of fregments, o is the length of the minimum overlap

Modified Lander-Waterman algorithm for several genotype

- Each viral genotype will make its contribution to observed q-contigs ---> community contig spectrum can be calculated as the sum of individual contig spectra

$$c_q = \sum_{i=1}^{M} n_i w_{qi} \quad \text{with} \quad \sum_{i=1}^{M} n_i = N \quad \text{for} \quad 1 \le i \le M$$

E-M algorithm

E-M algorithm is method for finding maximum likelihood estimates of parameters in statistical models. EM is an iterative method which alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the latent variables, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

$$\text{Error: } F = \sum_{q=1}^{L} \frac{(c_q{}' - c_q)^2}{V_q} \text{ with } V_q = \sum_{i=1}^{M} n_i w_{qi}(1 - w_{qi})$$

http://en.wikipedia.org/wiki/Expectation_maximization

Assessments

- The best model is used to assess diversity
- The richness is estimated as equal to the number of different genotypes found in the community structure model
- A measure for diversity:

- Shannon-Wiener index H' (in nats): $H' = -\sum_{i=1}^{S} r_i \ln r_i$

- The evenness:

Evenness E: $E = H'/H_{max} = H'/\ln S$

Limitations

- It is sensitive to the contig spectrum quality
  1. The same clone should be sequenced only once
  2. The sequences should be trimmed to remove ambiguities
  3. The assembly parameters should be sufficiently stringent

- The stringent parameters may omit some true-contigs

- Assume all DNA fragments and all the genotypes have the same size

Results

Four viral metagenomes belonging to different ecosystems were tested. Two of these were phage community samples of near-shore surface seawater from Scripps Pier (SP) and Mission Bay (MB), San Diego, California, USA. The two other samples are sediments from Mission Bay (MBSED) and human feces (FEC).

Table 2: Best descriptive rank-abundance form for the viral communities as determined by PHACCS. The error represents the variance weighted sum squared deviation between the experimental and the predicted contig spectra. For each community, the best descriptive function is the one that minimizes the error. The best fit obtained for each rank-abundance form was ranked according to the error in ascending order.

SP			MB			MBSED			FEC		
Rank	*Model*	*Error*	*Rank*	*Model*	*Error*	*Rank*	*Model*	*Error*	*Rank*	*Model*	*Error*
1	**Power law**	**1.84**	1	**Power law**	**2.15**	1	**Power law**	**0.0104**	1	**Power law**	**9.79**
2	Lognormal	1.93	2	Lognormal	2.36	1	**Lognormal**	**0.0104**	2	Lognormal	10.2
3	Logarithmic	2.57	3	Logarithmic	2.88	1	**Logarithmic**	**0.0104**	3	Logarithmic	10.3
4	Broken stick	10.7	4	Broken stick	14.6	1	**Exponential**	**0.0104**	4	Broken stick	52.2
5	Exponential	12.0	5	Exponential	16.2	5	Niche preemption	0.0139	5	Exponential	60.0
5	Niche preemption	12.0	5	Niche preemption	16.2	6	Broken stick	0.0157	5	Niche preemption	60.0

Results

Community	SP	MB	MBSED	FEC
Best rank-abundance form	Power law	Power law	Lognormal, logarithmic, power law & exponential *	Power law
Model equation	$n_i = 0.0203 \times i^{-0.641}$	$n_i = 0.0263 \times i^{-0.728}$	$n_i = 1.36 \times 10^{-4}$	$n_i = 0.0480 \times i^{-0.783}$
Richness	3350 genotypes	7180 genotypes	7340 genotypes	2390 genotypes
Evenness	0.932	0.900	1.00	0.873
Most abundant genotype	2.03 %	2.63 %	0.0136 %	4.80 %
Shannon-Wiener index	7.57 nats	7.99 nats	8.90 nats	6.80 nats

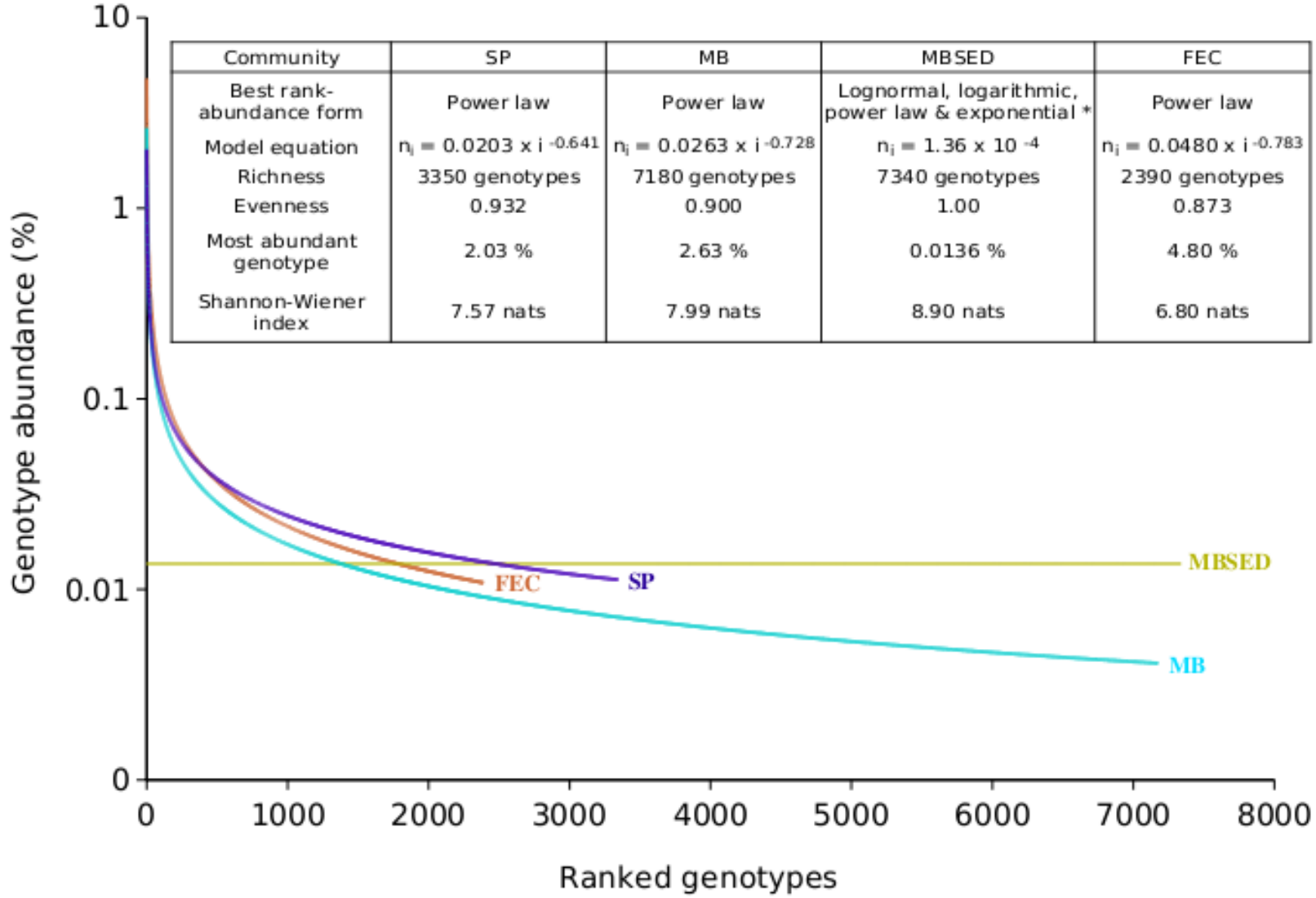


Figure 2
Comparison of the structure and diversity of the different viral communities using PHACCS. The graphics represent rank-abundance curves, where the abundance of each genotype is plotted versus its abundance rank, the genotype of rank one being the most abundant. The curves were obtained by plotting the PHACCS rank-abundance values of the different communities on the same axis. *The predicted community structure for MBSED was the same for the lognormal, logarithmic, power and exponential rank-abundance forms. As a consequence, the diversity predictions were also the same.

http://biome.sdsu.edu/phaccs/

Parameters:

Contig spectrum:	[1021 17 3 0 0 0]
Avg. genome size:	50000 bp
Avg. fragment length:	650 bp
Min. overlap length:	35 bp
Genotype range:	between 1 and 100000
Precision:	2

Rank-abundance form: power

> Structure model

• Error:	3.7
• Comment:	-
• Model parameter 1:	0.73
• Model parameter 2:	0.026
• Model equation:	$n_i = 0.026\ i^{-0.73}$

> Diversity estimates

• Richness:	8100 genotypes
• Evenness:	0.9
• Most abundant genotype:	2.6 % of the community
• Shannon-Wiener index:	8.1 nats