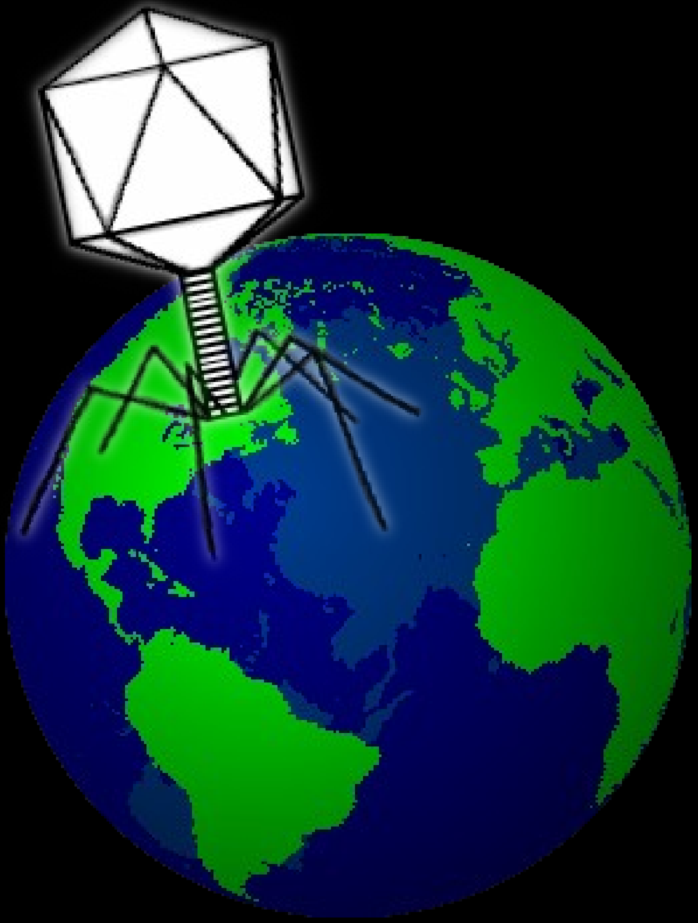


Characterization of environmental viral diversity using metagenomics

Florent Angly

RohwerLab

gateway to the future
CSRC
an interdisciplinary

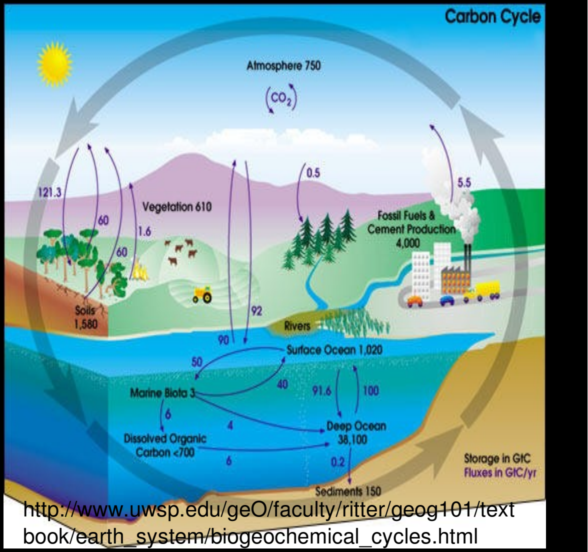
Ecological importance of viruses

Ubiquity

Large abundance

http://www.virusecology.org/MOVE/Method%206.html

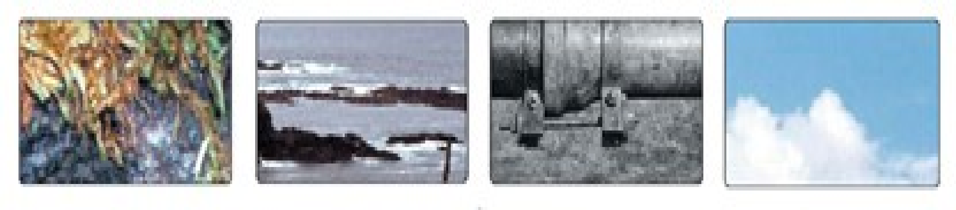
Global impact



http://www.uwsp.edu/geO/faculty/ritter/geog101/textbook/earth_system/biogeochemical_cycles.html

Culturable

Metagenomics



Environmental sample: uncultured phages

Environmental phage DNA

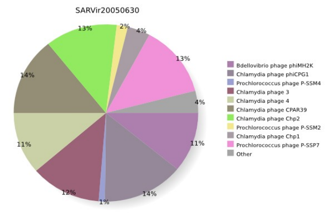
Who is there?

What do they do?

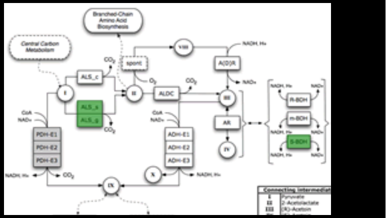
How many are they?

16S rDNA

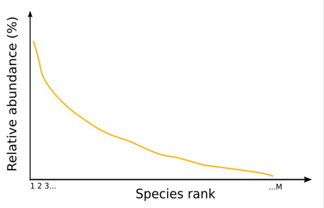
Taxonomy



Function

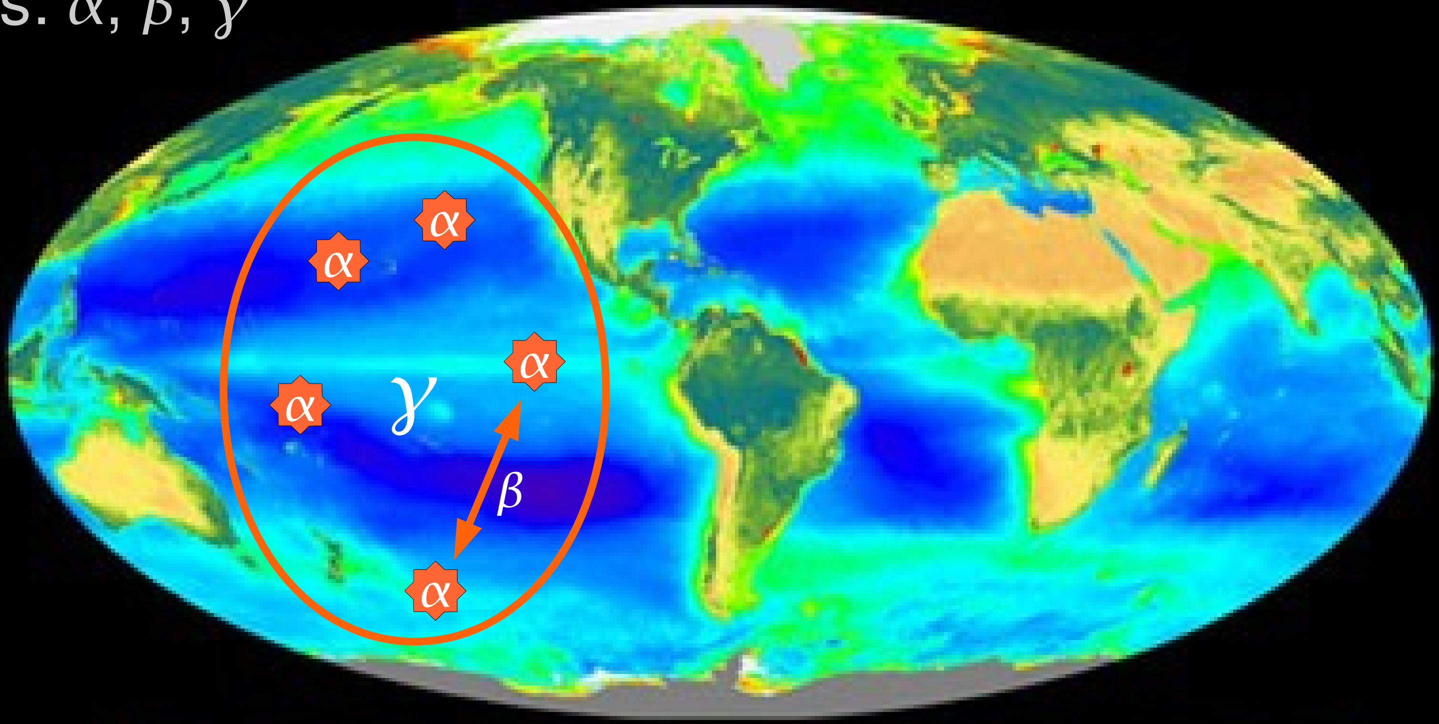


Diversity



Diversity

Diversity types: α, β, γ



http://www.gma.org/herring/biology/distribution/comparing_oceans.asp

α-diversity: richness, evenness, indices

Shannon-Wiener index:

$$H' = -\sum_{i=1}^{S} r_i \ln r_i$$

S: richness

r_i: relative abundance of the *i* th species

Hypothesis and plan

Is it possible to estimate viral diversity from metagenomes? If so, how?

1. α-diversity (PHACCS)
2. β-diversity (MaxiPhi)
3. Average genome length (GAAS)
4. Diversity workflow
5. Environmental viral diversity

- 1 -

α-diversity

PHACCS

α-diversity from metagenomic data

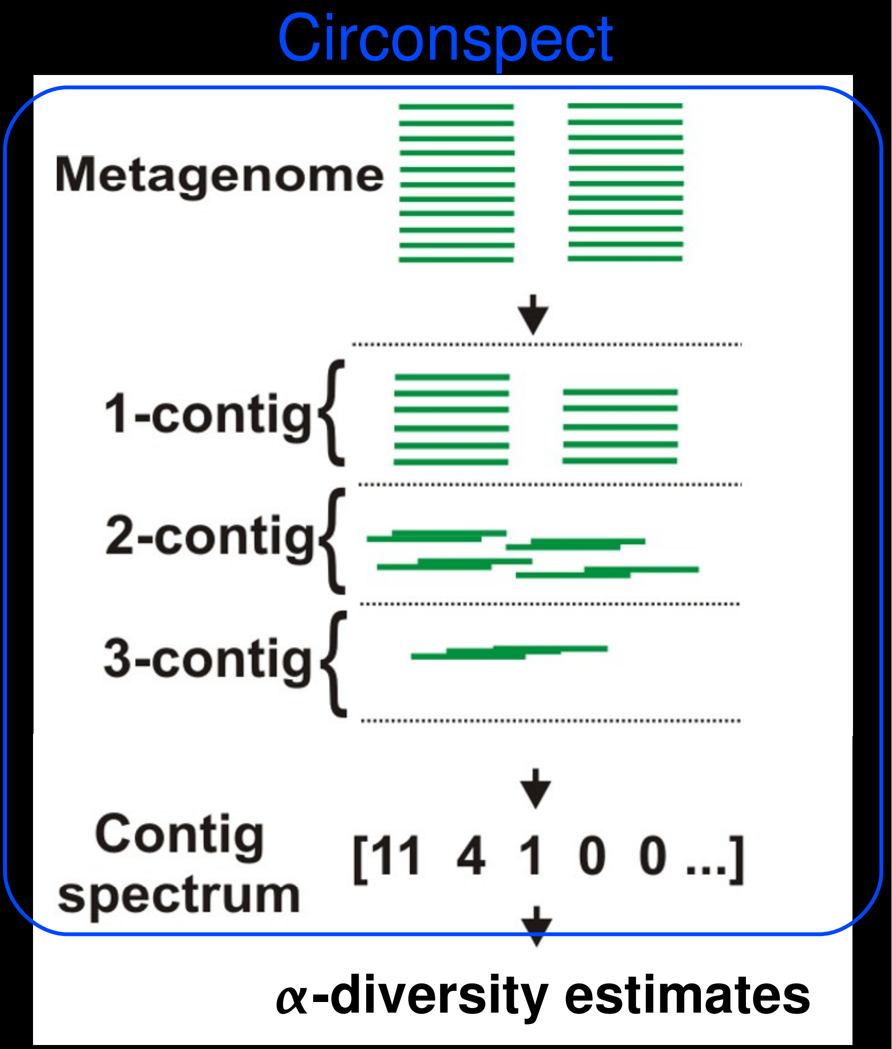
α-diversity: local diversity (one sample)

1. Assemble metagenomic sequences

2. Count the number of contigs

3. Assume that only sequences from the same species form contigs

4. Model diversity: the more abundant a species, the larger the number of its sequences forming contigs



Circonspect

Metagenome

1-contig

2-contig

3-contig

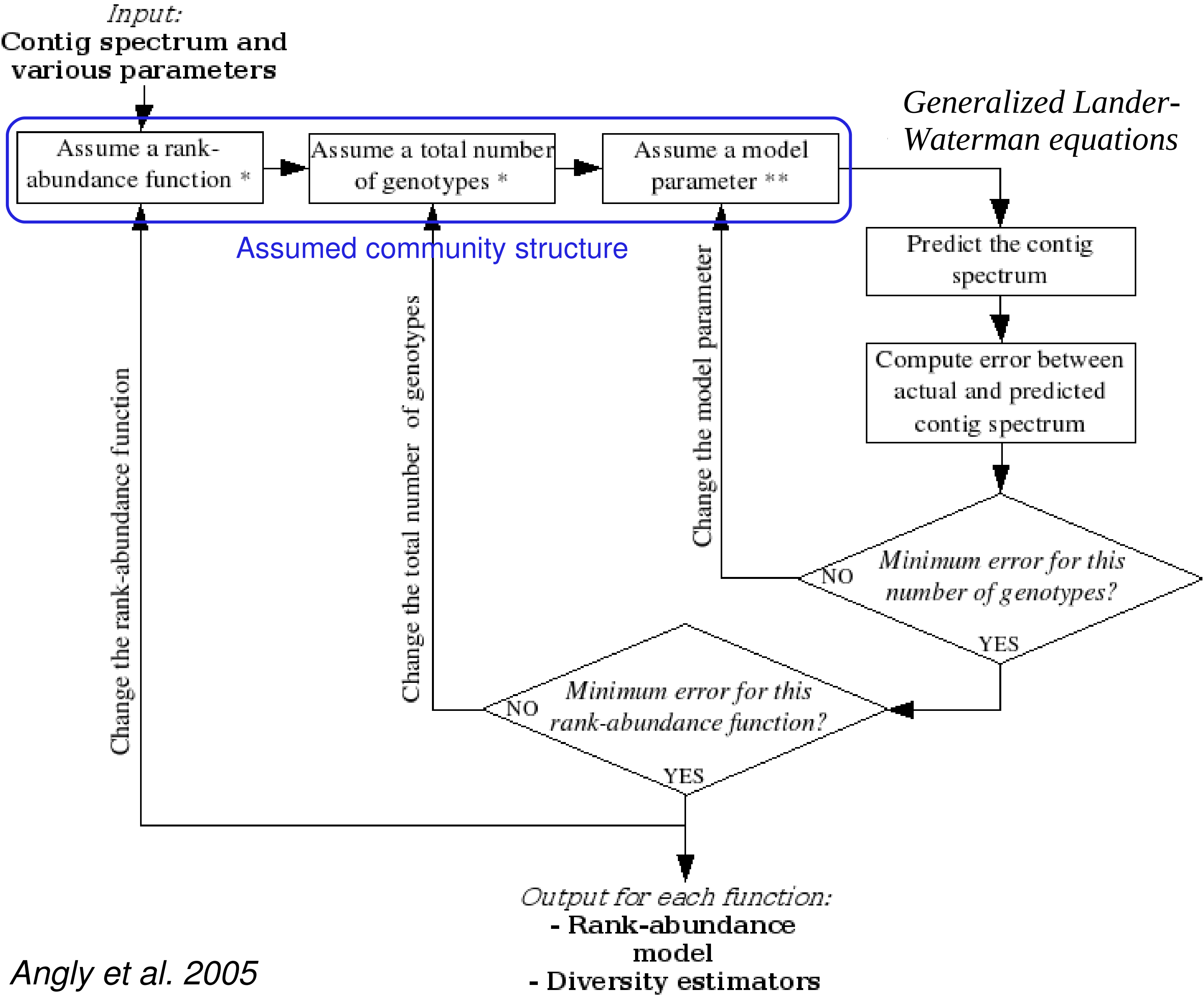
Contig spectrum

[11 4 1 0 0 ...]

α-diversity estimates

Characterization of environmental viral diversity using metagenomics

PHACCS
Phage communities from Contig Spectrum


Input:
Contig spectrum and various parameters
Assume a rank-abundance function *
Assume a total number of genotypes *
Assume a model parameter **
Assumed community structure
Generalized Lander-Waterman equations
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
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

Angly et al. 2005

$$c_q = \sum_{i=1}^{M} n_i w_{qi}$$

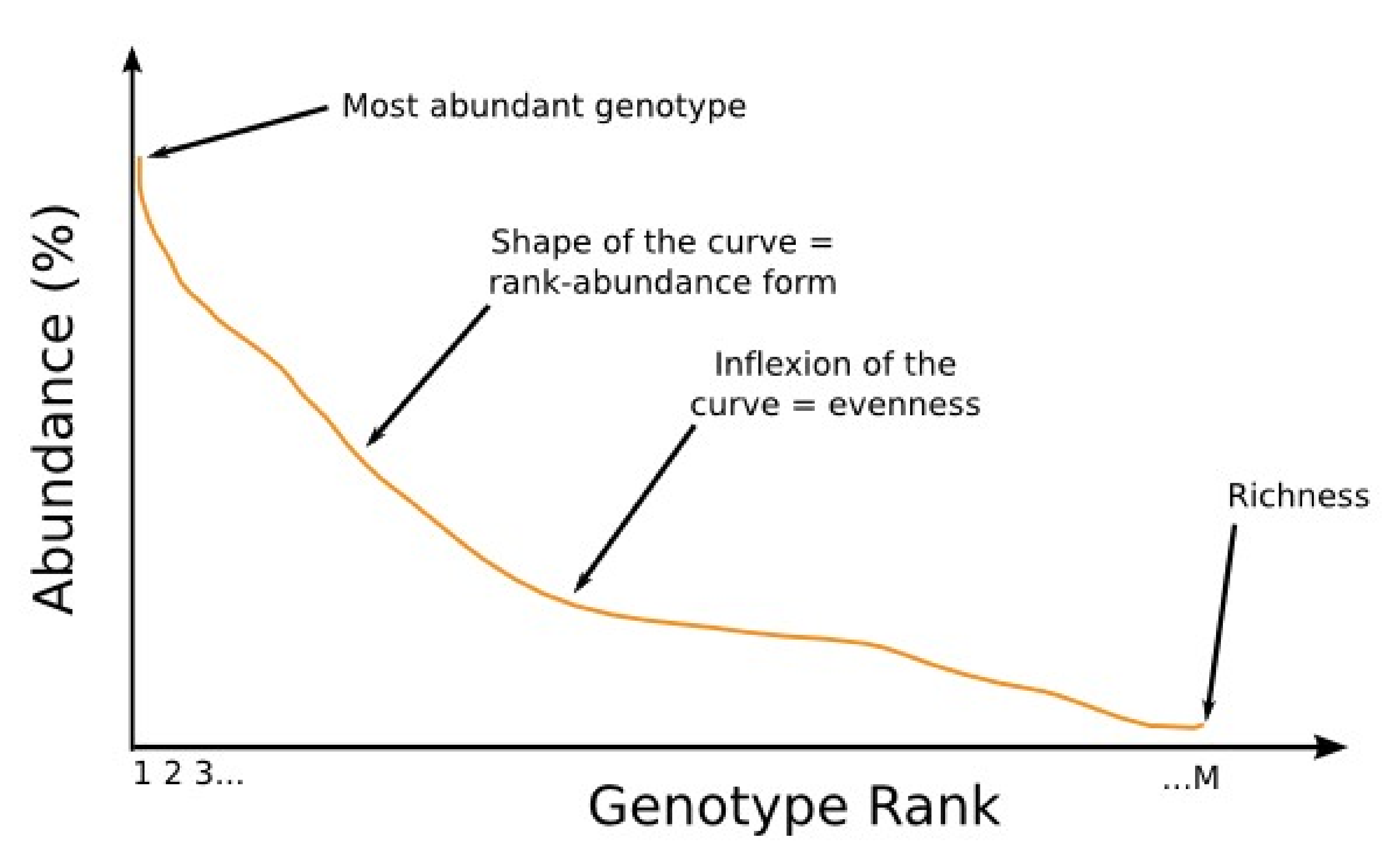
n_i : expected number of sequences for species *i*

w_{qi} : probability that a sequence of species *i* is in a contig of size *q*

Assumptions:

- No chimeric contigs
- Rank-abundance distribution
- Genomes have an average length

Community structure and diversity


Most abundant genotype
Shape of the curve = rank-abundance form
Inflexion of the curve = evenness
Richness
Abundance (%)
1 2 3...
...M
Genotype Rank

- 2 -

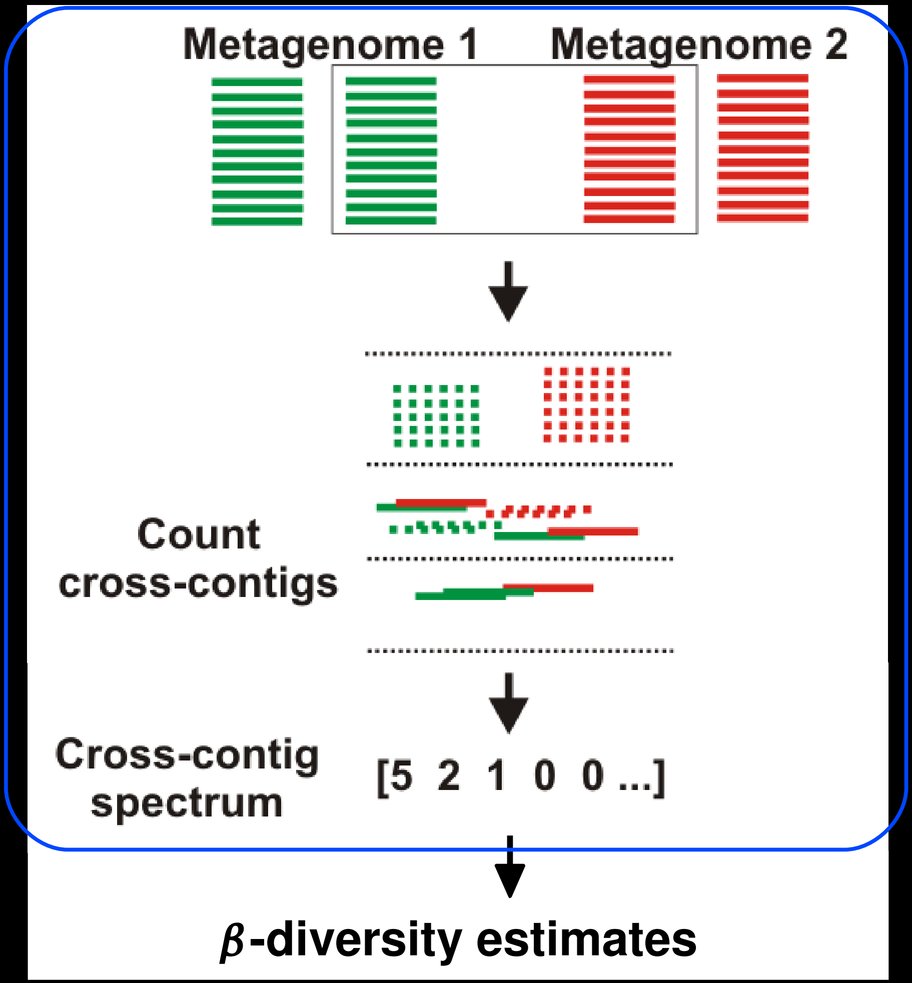
β-diversity

MaxiPhi

β-diversity from metagenomic data

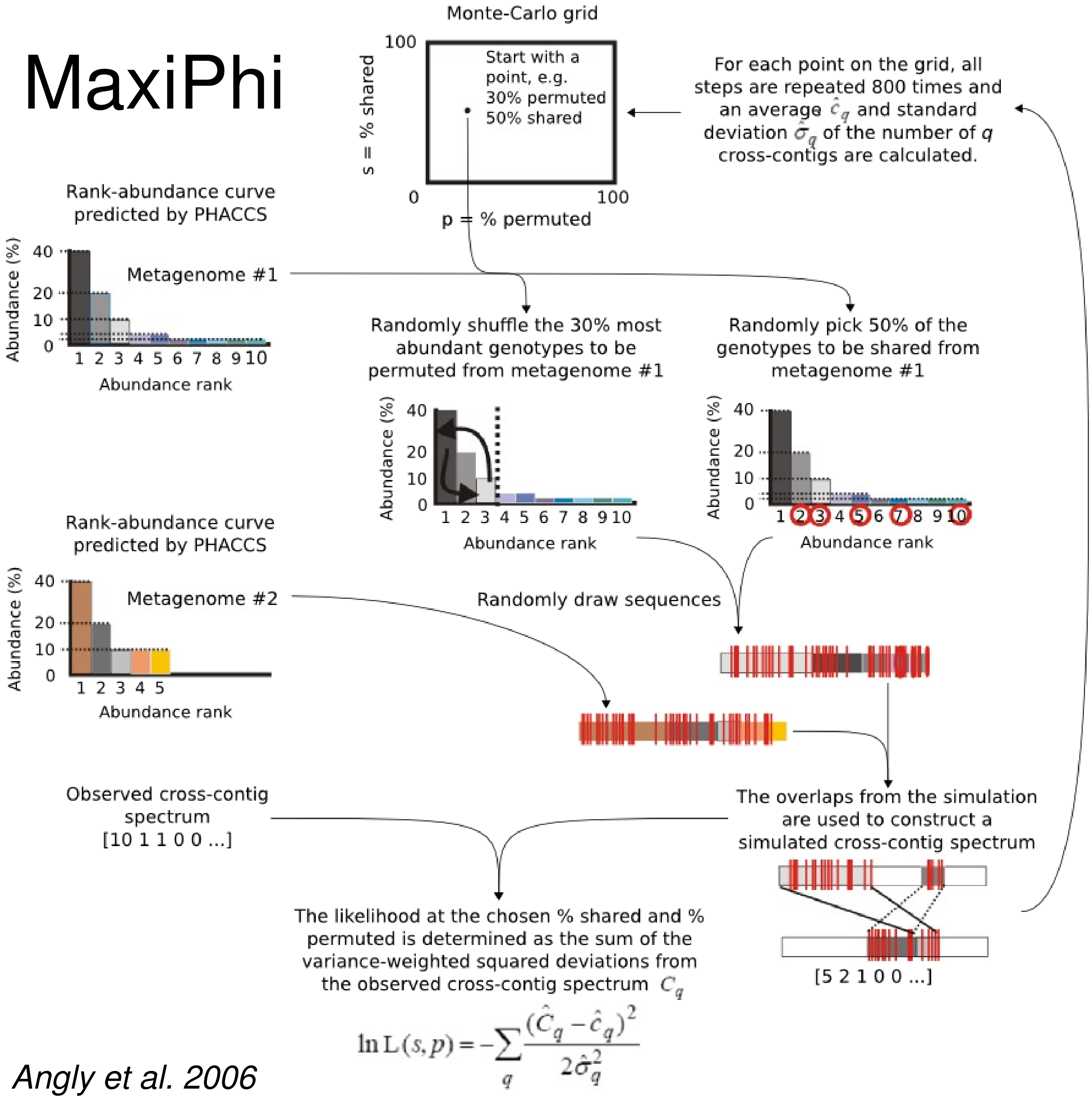
- β-diversity: difference in diversity between several samples

Cross-contig spectrum



Circonspect

MaxiPhi



Monte-Carlo grid

Start with a point, e.g. 30% permuted 50% shared

For each point on the grid, all steps are repeated 800 times and an average $\hat{c}_q$ and standard deviation $\hat{\sigma}_q$ of the number of q cross-contigs are calculated.

Rank-abundance curve predicted by PHACCS

Metagenome #1

Randomly shuffle the 30% most abundant genotypes to be permuted from metagenome #1

Randomly pick 50% of the genotypes to be shared from metagenome #1

Rank-abundance curve predicted by PHACCS

Metagenome #2

Randomly draw sequences

Observed cross-contig spectrum [10 1 1 0 0 ...]

The overlaps from the simulation are used to construct a simulated cross-contig spectrum

[5 2 1 0 0 ...]

The likelihood at the chosen % shared and % permuted is determined as the sum of the variance-weighted squared deviations from the observed cross-contig spectrum C_q

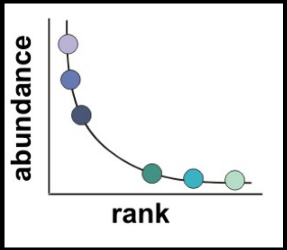
$$\ln \mathrm{L}(s,p) = -\sum_q \frac{(\hat{C}_q - \hat{c}_q)^2}{2\hat{\sigma}_q^2}$$

Angly et al. 2006

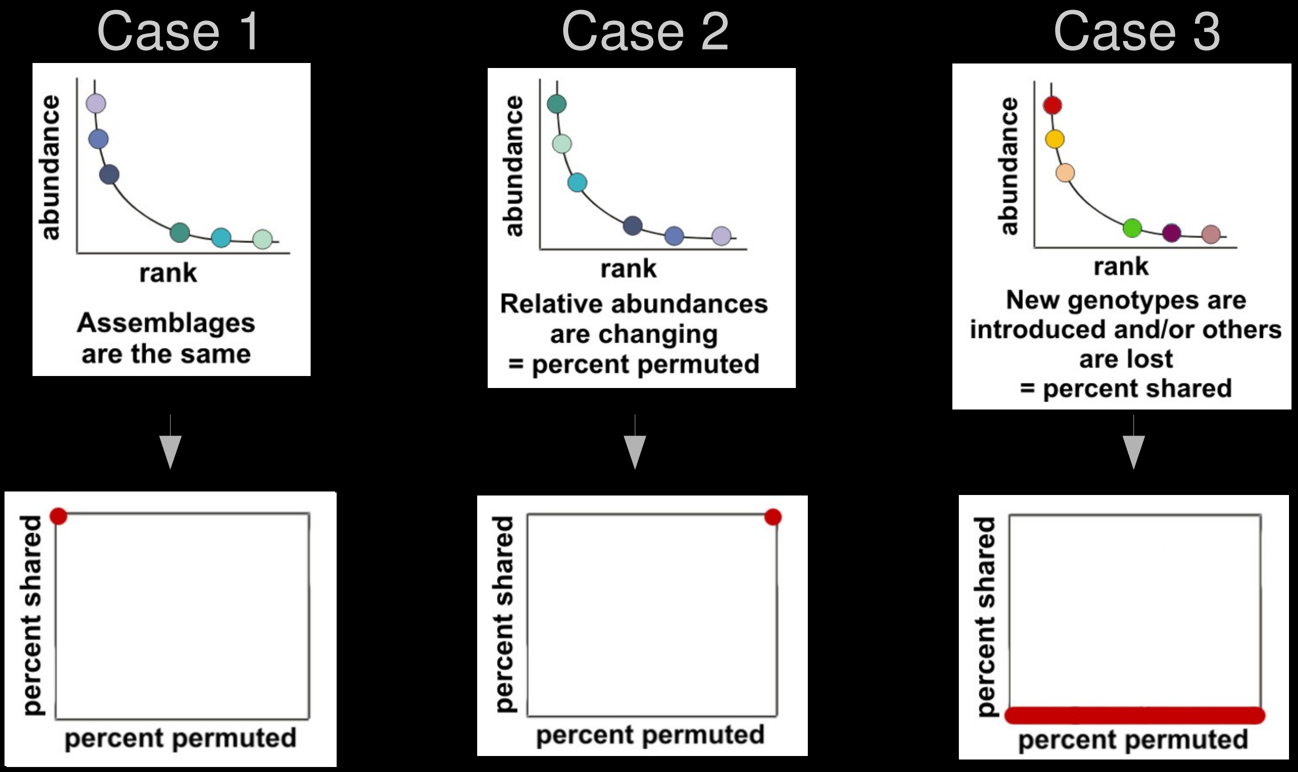
Modeling β-diversity

- Percent shared and percent permuted

Environment 1 (e.g., SAR)



Environment 2 (e.g., BBC)



Case 1

Assemblages are the same

Case 2

Relative abundances are changing = percent permuted

Case 3

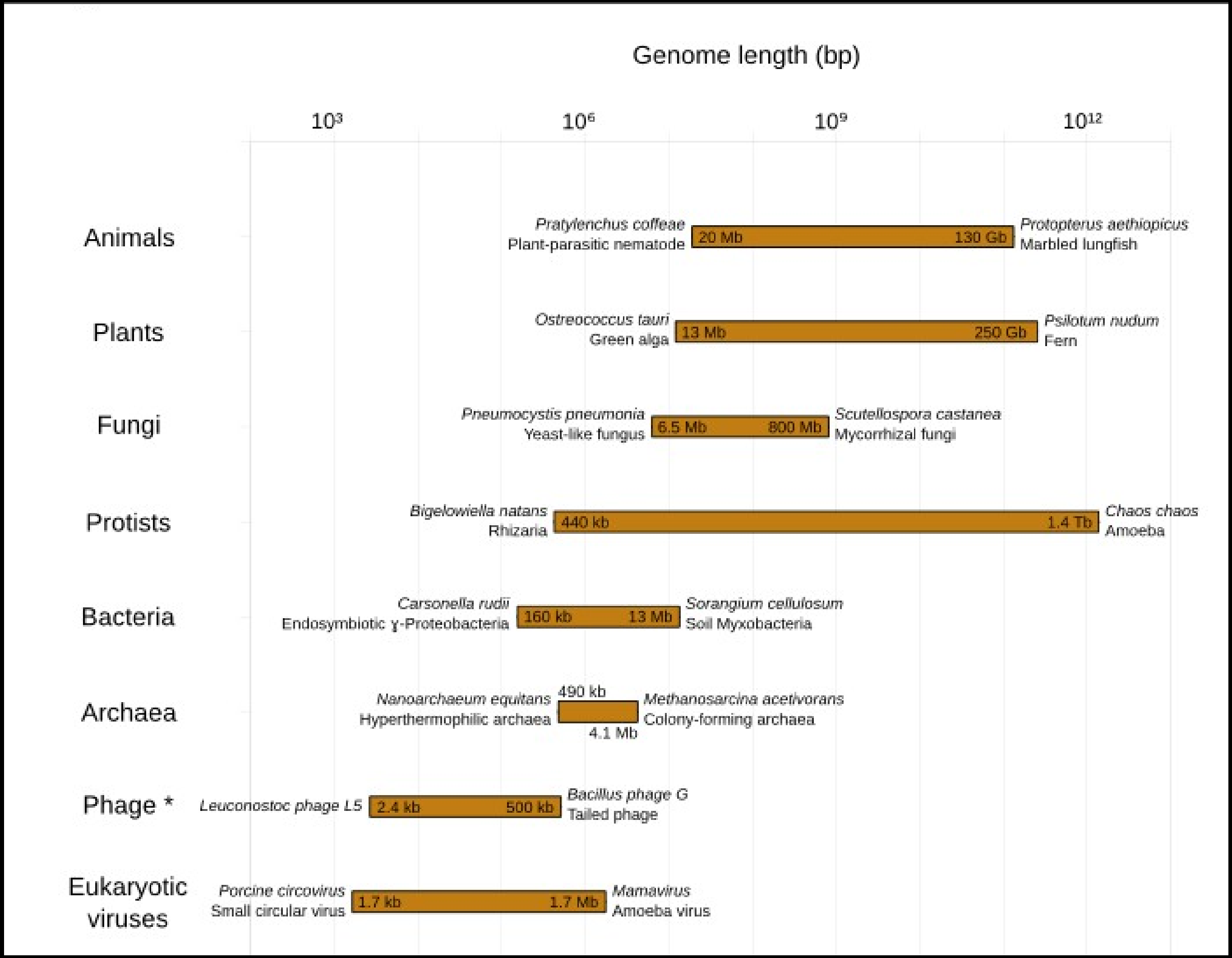
New genotypes are introduced and/or others are lost = percent shared

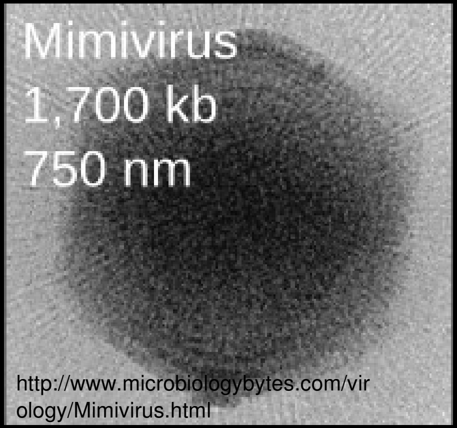
- 3 -

Average genome length

GAAS

Genome length of different organisms





Mimivirus
1,700 kb
750 nm

http://www.microbiologybytes.com/virology/Mimivirus.html

Circovirus
1.7 kb
17 nm

http://www.pcvd.org/

Microbes
and
viruses

Data compiled from the ICTVdb, NCBI RefSeq, Microbe Wiki, Fungal Genome Size Database, Plant DNA C-values Database and Animal Genome Size Database

GAAS

Metagenome

↓

BLAST ← Complete genome database

↓

Filter out weak similarities ★

↓

Keep & weight all similarities ★★

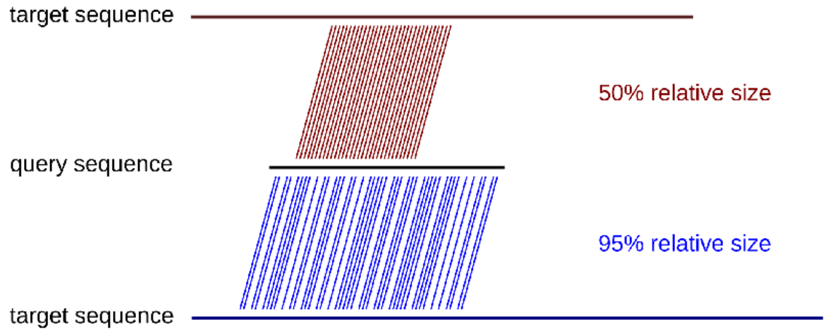
↓

Relative abundance ★★★

↓

Average genome length

★ Cutoff alignment E-value, identity percent, and relative length



★★ For each query sequence *i*, keep all the similarities to genomes *j* but give similarities a weight that has a statistical meaning:

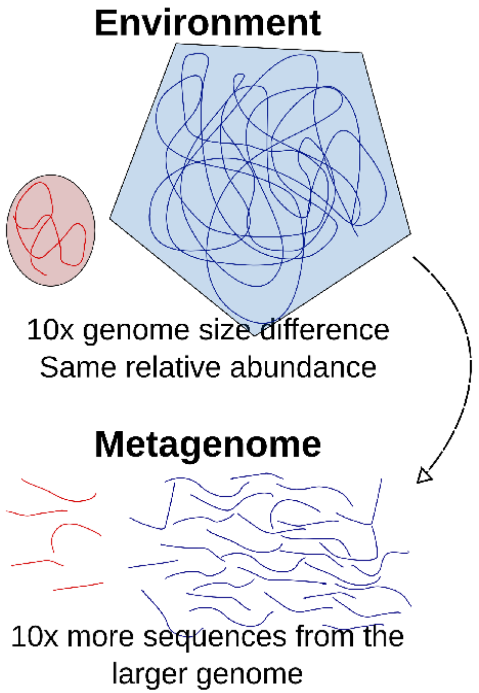
$$w_{ij} \propto k_i / E_{ij}$$

w_{ij} : weight

E_{ij} : E-value

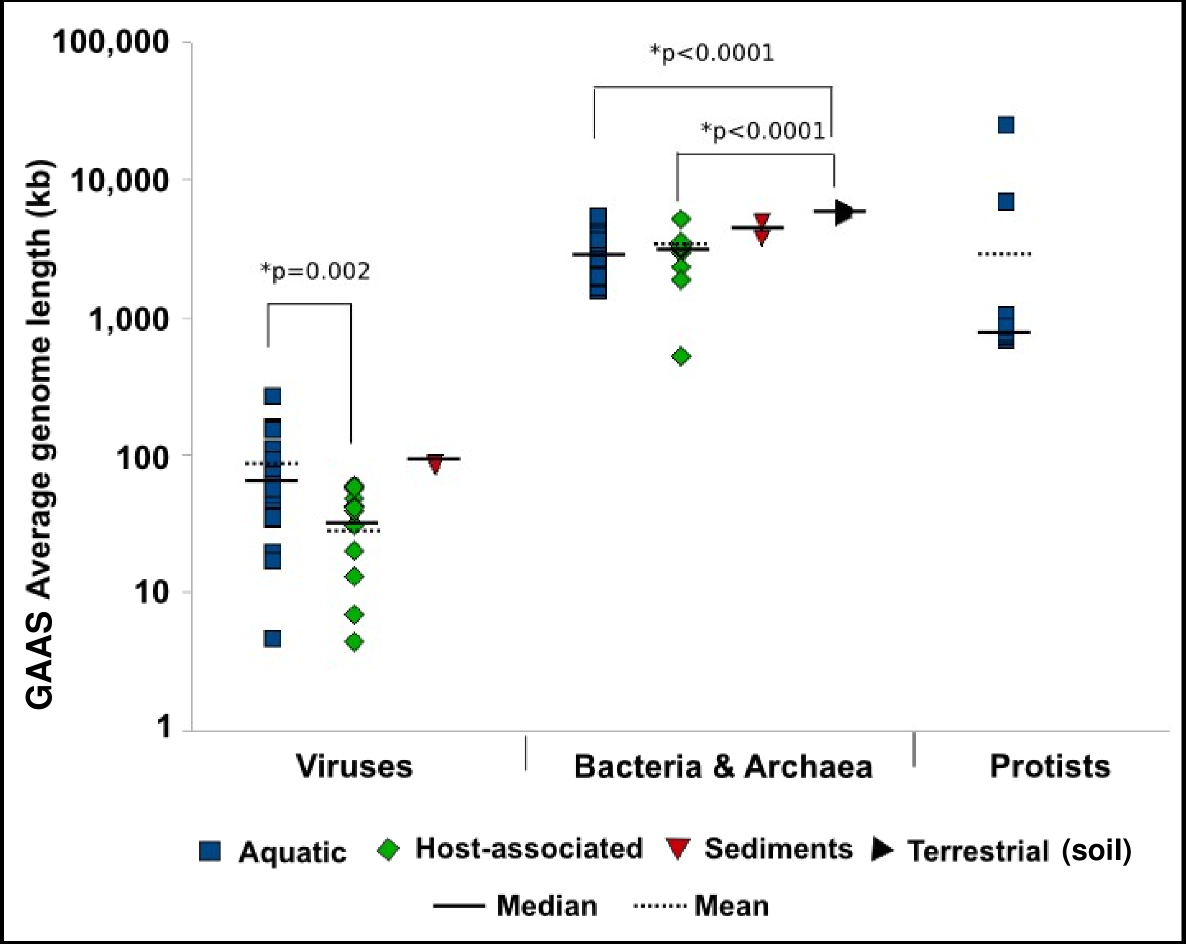
k_i : constant

★★★ Normalize by genome length

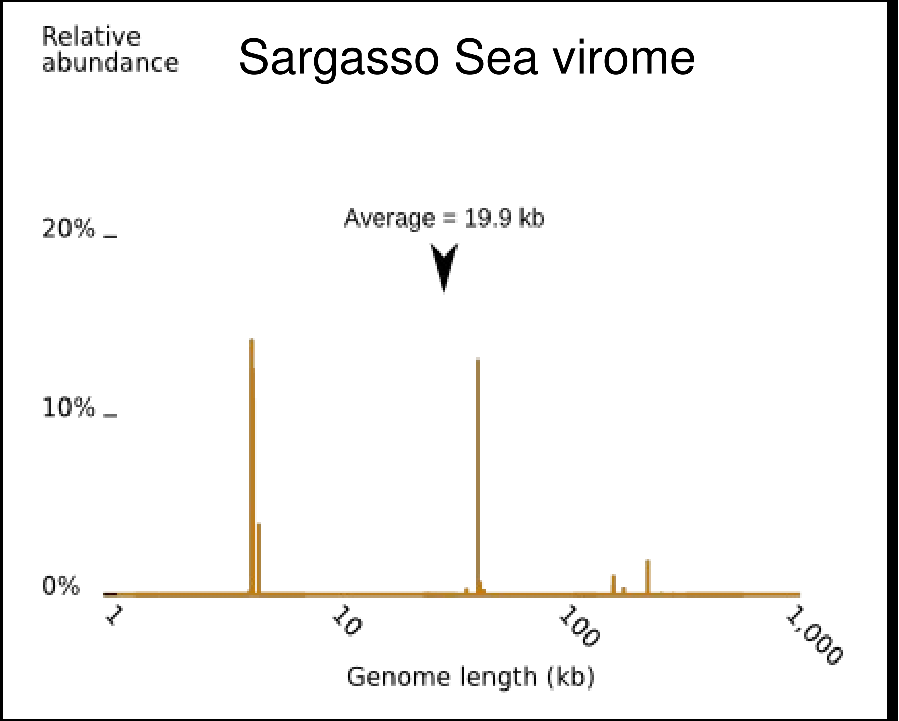

Environment
10x genome size difference
Same relative abundance
Metagenome
10x more sequences from the larger genome

Genome length in the environment

Meta-analysis of 174 metagenomes



- Variability between biomes
- Variability within biomes
- Variability within sample

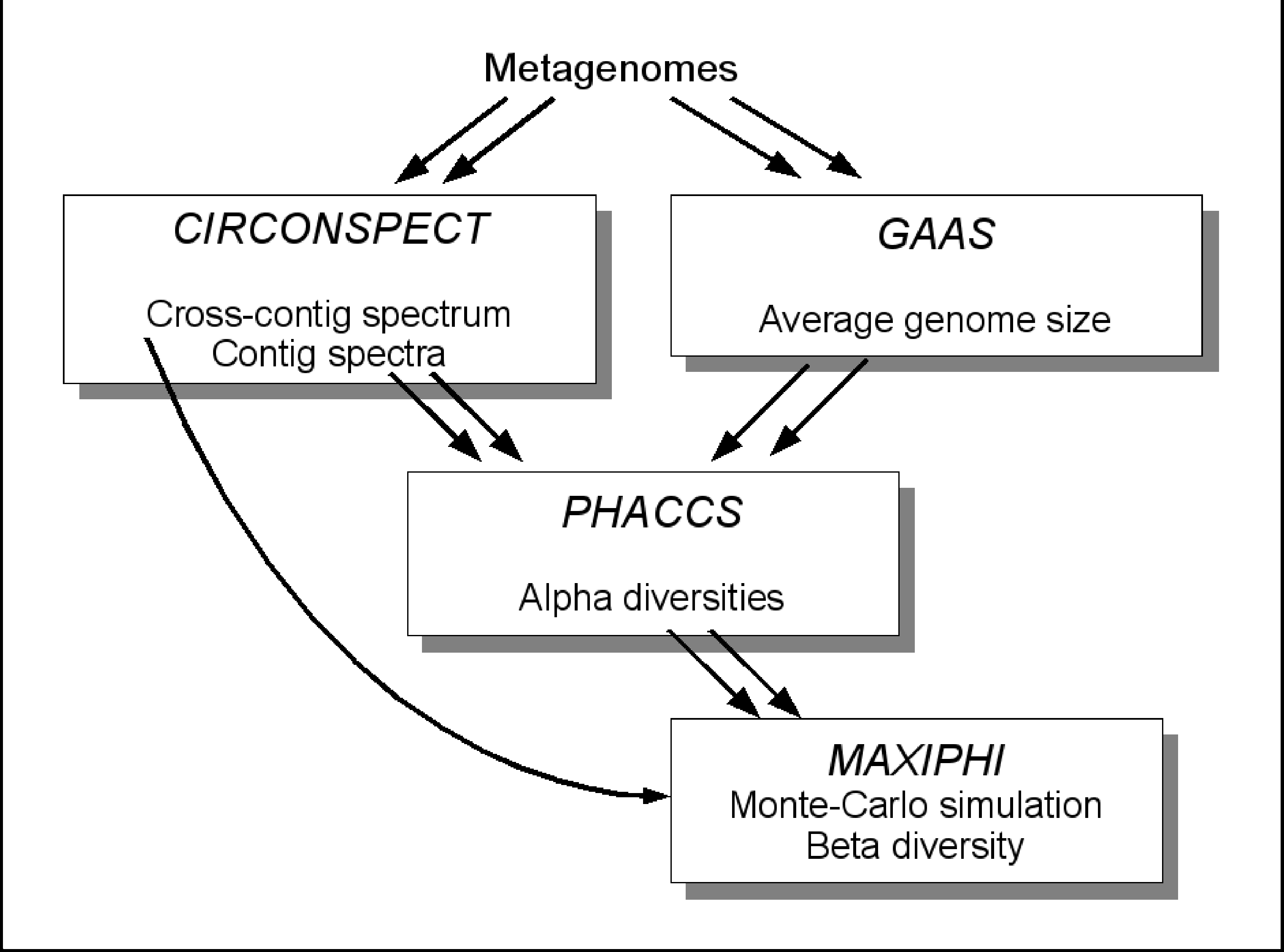


Angly et al. In revision

- 4 -

Diversity workflow

Diversity workflow



Metagenomes

CIRCONSPECT

Cross-contig spectrum
Contig spectra

GAAS

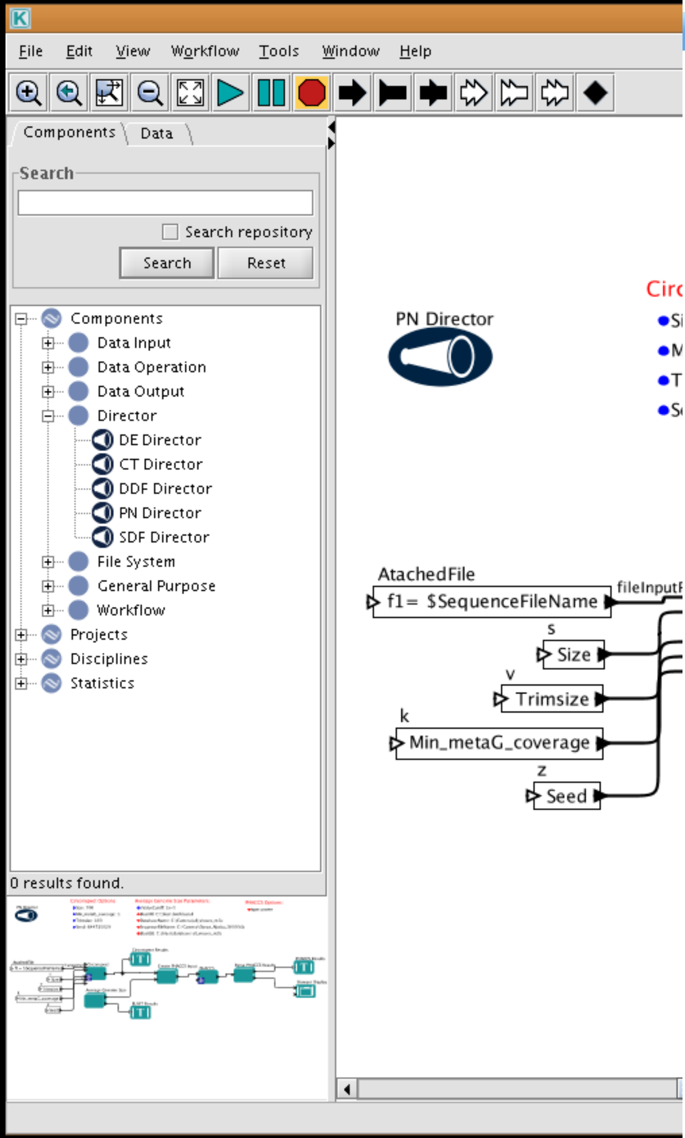
Average genome size

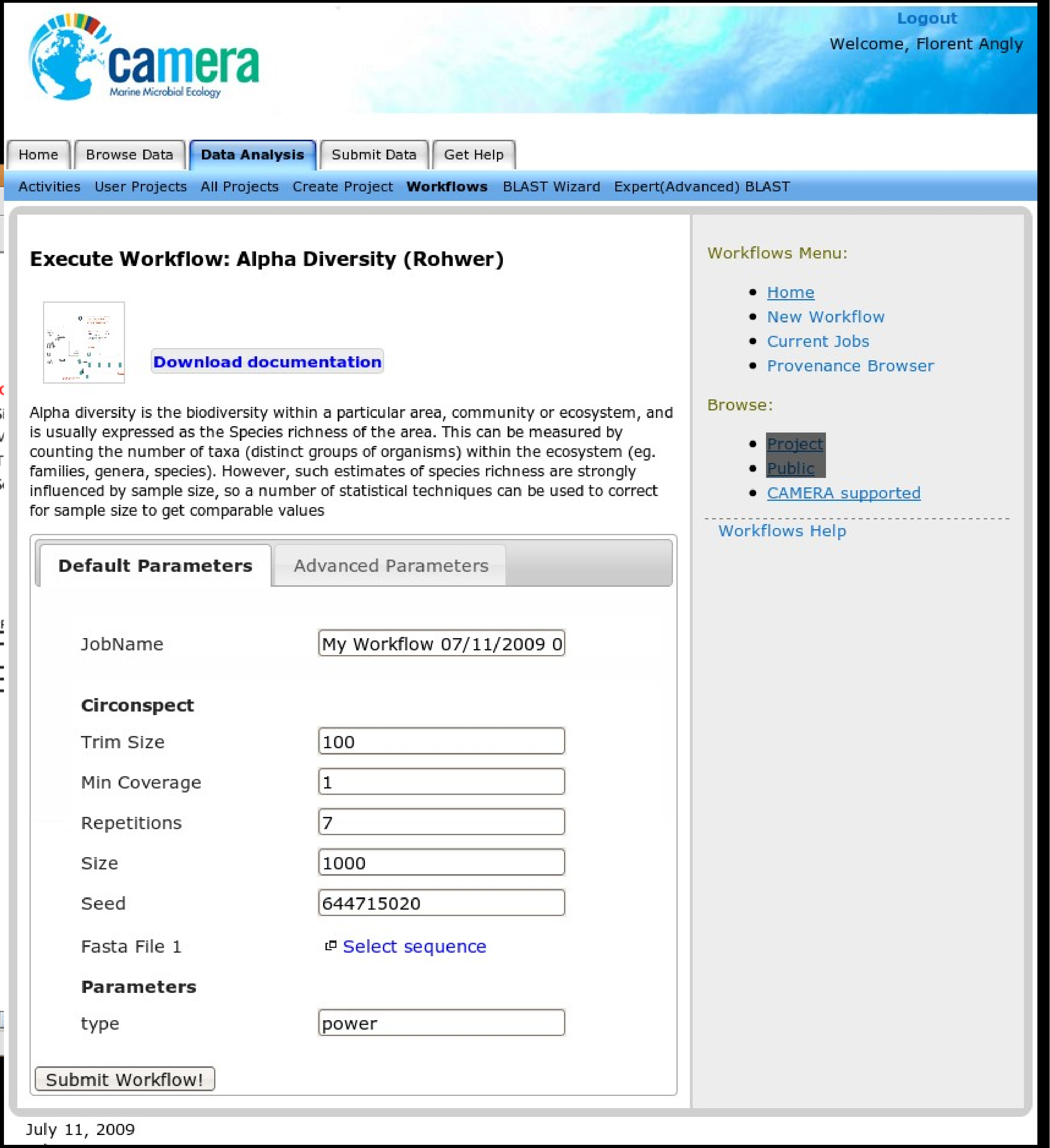
PHACCS

Alpha diversities

MAXIPHI
Monte-Carlo simulation
Beta diversity

α-diversity





Execute Workflow: Alpha Diversity (Rohwer)

Download documentation

Alpha diversity is the biodiversity within a particular area, community or ecosystem, and is usually expressed as the Species richness of the area. This can be measured by counting the number of taxa (distinct groups of organisms) within the ecosystem (eg. families, genera, species). However, such estimates of species richness are strongly influenced by sample size, so a number of statistical techniques can be used to correct for sample size to get comparable values

JobName	My Workflow 07/11/2009 0
Circonspect	
Trim Size	100
Min Coverage	1
Repetitions	7
Size	1000
Seed	644715020
Fasta File 1	Select sequence
Parameters	
type	power

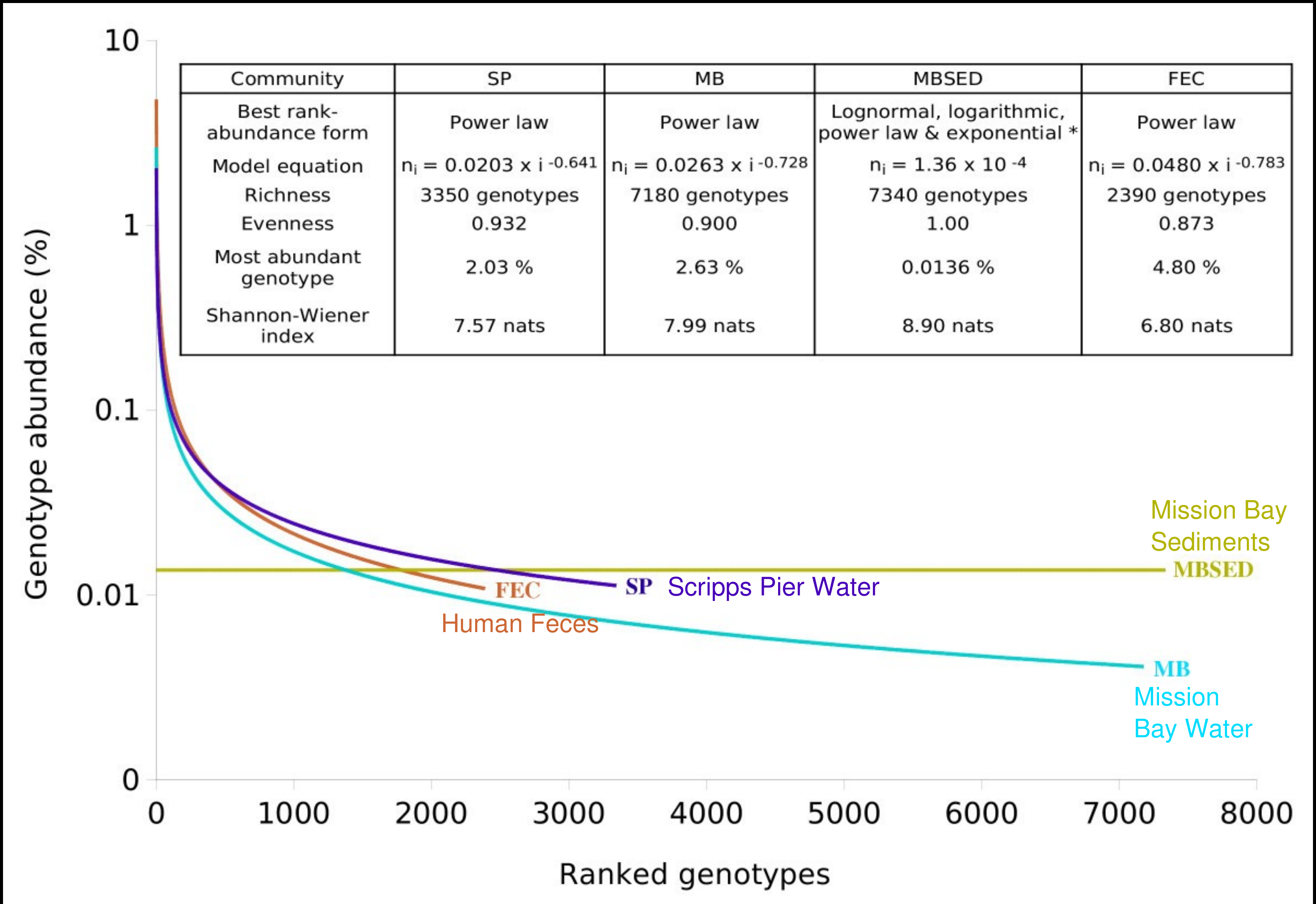
July 11, 2009

- 5 -

Environmental viral diversity

α-diversity of four viromes

- PHACCS estimate of community structure and diversity



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

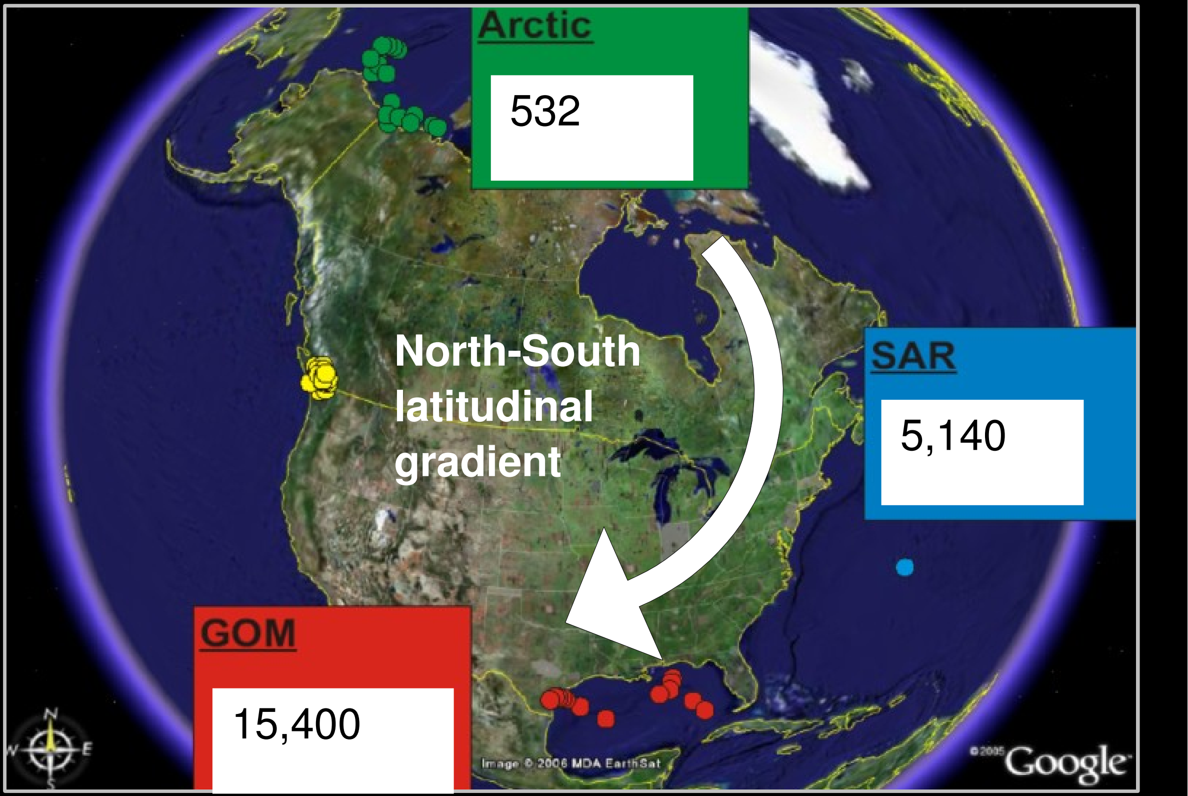
Angly et al. 2005

Marine latitudinal gradient of diversity

- The latitudinal richness gradient is the most documented pattern:
  Higher diversity close to the tropics

- Affects macroorganisms (*Hillebrand et al. 2004*), microorganisms (*Pommier et al. 2007*, *Fuhrman et al. 2008*)
- Affects viruses?
- How universal are the rules that govern the distribution of life on Earth?

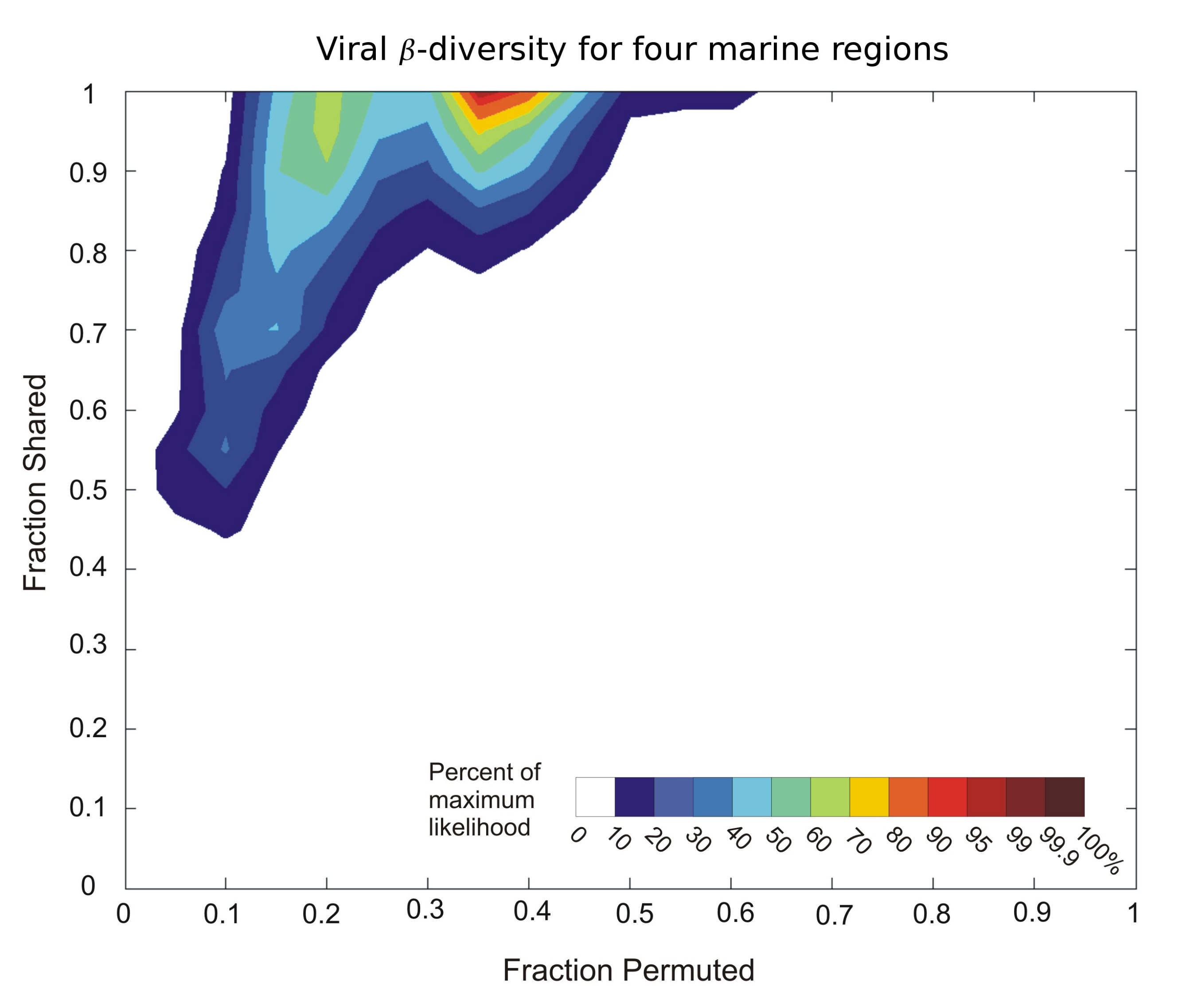
Viral richness in the oceans



Angly et al. 2006

Marine β-diversity

- Viruses are dispersed world-wide
- "Everything is everywhere"
- Marine viruses are cosmopolitan but the environment selects!

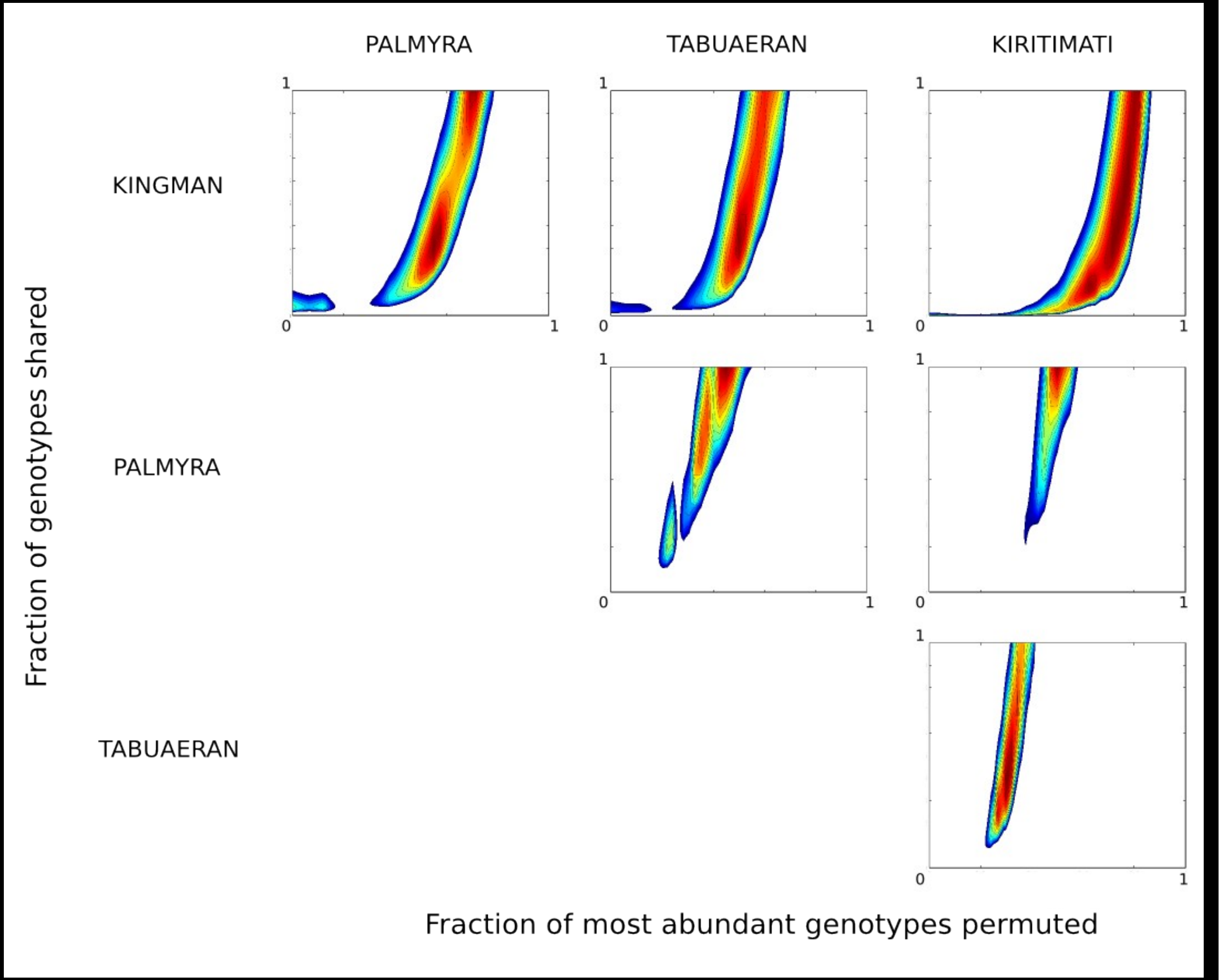


Angly et al. 2006

Diversity in the Line islands coral reefs

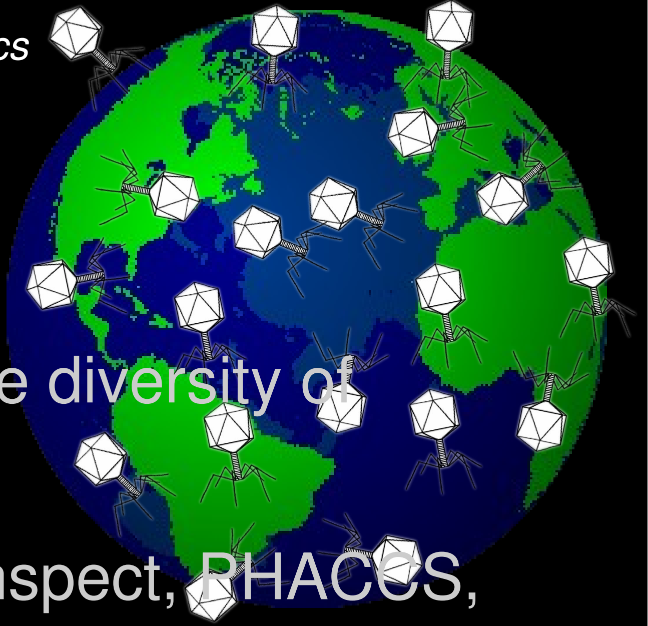
Atoll	***Human population***	***Viral richness***
Kingman	0	8,380
Palmyra	20	17,100
Tabuaeran	1,000	24,800
Kiritimati	5,100	102,000

Viral β-diversity



Sandin et al 2008

Plot by Steve Rayhawk



Conclusions

Metagenomics is well suited to investigate the diversity of uncultured viral communities

Novel computational methods: GAAS, Circonspect, PHACCS, MaxiPhi

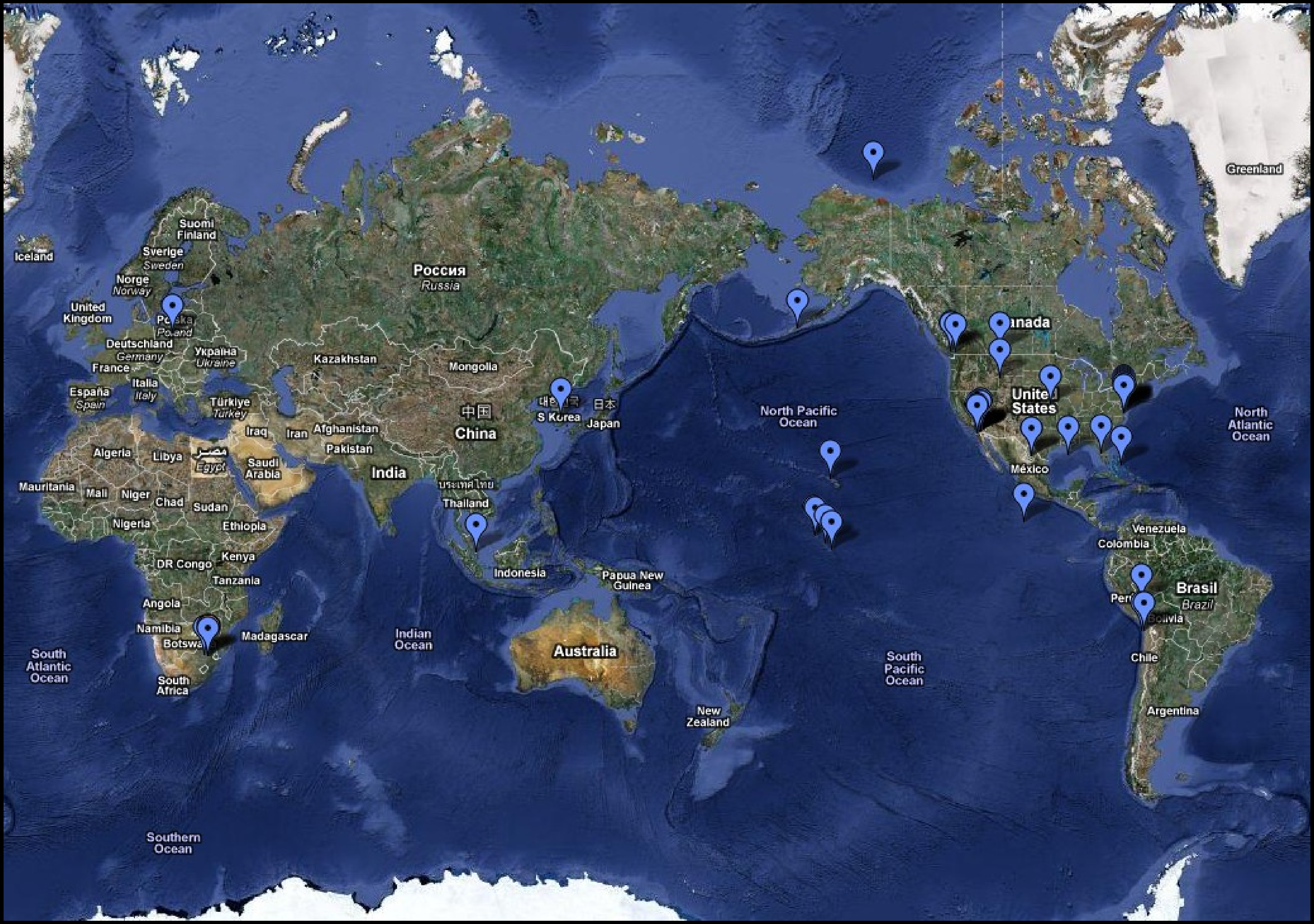
Publicly available tools and integration into easy-to-use software workflow

Diversity methodology does not require similarities to databases

Viral diversity may follow the same patterns of diversity as microorganisms and macroorganisms

As more and more viral metagenomes are sequenced, the metagenomic diversity workflow will be used to analyze the global virome and estimate Earth's total viral richness

Viral metagenome locations



Acknowledgments

SAN DIEGO STATE UNIVERSITY

Rob Edwards
Peter Salamon
Ben Felts
Jim Nulton
Joseph Mahaffy
Robert Schmieder
Liz Dinsdale
Alejandra Prieto-Davo

NSF

NSF Biocomplexity

RohwerLab

Forest
Bahador
Beltran
Dana
John
Katie
Linda
Liz
Mark
Matt
Mike
Yan Wei

Anna
Becky
Bethany
Betty
Christelle
Cynthia
David
Danielle
Elysa
Emiko
Fairoz
Karin
Jennifer
John
Marina
Megan
Morrigan
Mya
Neilan
Olga
Priscila
Selina
Steve
Veronica
Yanan
...

Gordon and Betty
MOORE
FOUNDATION

Characterization of environmental viral diversity using metagenomics

Characterization of environmental viral diversity using metagenomics

Characterization of environmental viral diversity using metagenomics