A COMPUTATIONAL WORKFLOW FOR THE ESTIMATION OF ENVIRONMENTAL VIRAL DIVERSITY IN METAGENOMES

by

FLORENT E ANGLY

A Dissertation submitted to the Faculty of Claremont Graduate University and San Diego State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Faculty of Computational Science

Claremont and San Diego, California
2009

Approved by:

Forest Rohwer, Chair

We, the undersigned, certify that we have read this dissertation of Florent E Angly and approve it as adequate in scope and quality for the degree of Doctor of Philosophy.

Dissertation Committee:

Forest Rohwer, Chair

John Angus, Member

Rob Edwards, Member

Alpan Raval, Member

Peter Salamon, Member

ABSTRACT

A computational workflow for the estimation of environmental viral diversity in metagenomes

by

Florent E Angly

Claremont Graduate University and San Diego State University: 2009

Viruses and in particular phages, predators of Bacteria and Archaea, are numerically abundant in the environment and play important ecological roles. Yet, little is known about their diversity and distribution. The introduction of metagenomics has revolutionized the study of viral and microbial communities by bypassing the need to culture individual species, thus allowing access to their complete diversity. However, unlike for microorganisms, no standard technique exists to measure viral diversity from sequence data, and lab techniques are limiting.

In this thesis, computational methods were developed to quantify the diversity of viruses from metagenomic data. These methods use overlapping sequences (contigs) assumed to come from the same species. The modeling of the contigs characterizes viral community structure and α-diversity, or sample diversity. Assembling metagenomes pooled together produces contigs between

sequences from multiple samples (cross-contigs). Such contigs are indicative of common viruses and are the basis to estimate β-diversity (change in diversity between samples). Modeling the α-diversity and β-diversity of uncultured viral communities relies on knowing the average length of their genomes, which was calculated here from similarities of metagenomic reads to genomes of known length. The different programs necessary to the estimation of viral diversity were assembled into a workflow available online in order to offer the metagenomic community an easy way to assess metagenomic diversity.

The application of the viral diversity workflow suggests that there may be as many as 10^8 viral species on Earth, and that their distribution (e.g. diversity patterns) may be similar to that of microorganisms and macroorganisms. However, some biomes such as the air and deep subsurface remain unexplored. As additional metagenomes are produced and sampling resolution increases, this workflow for estimating diversity will prove invaluable to gain further insights into viral biogeography.

DEDICATION

To my family, friends and people that help giants becoming bigger.

ACKNOWLEDGMENTS

I want to thank the Gordon and Betty Moore Foundation and the National Science Foundation Biocomplexity Initiative for providing funding to support this research.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Viruses, biological entities incapable of reproducing without a host cell, are the most numerous biological entities on Earth, but their diversity is largely uncharacterized. Phages, viruses which infect Bacteria and Archaea, are especially diverse, with a number of extant phage species higher than that of other organisms. This thesis presents novel methods to characterize the diversity and distribution of viruses using metagenomic sequence data, a computational workflow incorporating these methods, and a case study of viral diversity in the world's oceans. The following is an introduction to the thesis and a review of the literature on viral metagenomics and diversity estimation.

The ecological importance of viruses

Viruses are ubiquitous and numerous in the environment, and are present in high abundances in terrestrial, aquatic and host-associated biomes [1-9], where their hosts are numerous. Many viruses also survive in extreme conditions such as high or low temperature, high pressure and salinity [10-17], and there is evidence that suggest their existence in the air column [18,19]. Observation of Virus-Like Particles (VLPs) with electronic and epifluorescence microscopy has revealed the presence of ~10 million VLPs per milliliter of seawater [20-23]. In the oceans, there are typically ~10 viral particles for each microbial cell [24]. The global number of viral particles was estimated to be ~10^{31} VLPs, based on the number of Bacteria and Archaea on Earth [25].

Not only are viruses abundant and ubiquitous, but they are also highly morphologically and genetically diverse. Circoviruses are the smallest known viruses, with an icosahedral capsid approximately 17 nm in diameter containing two genes on a circular single-stranded DNA molecule [26]. In contrast, the

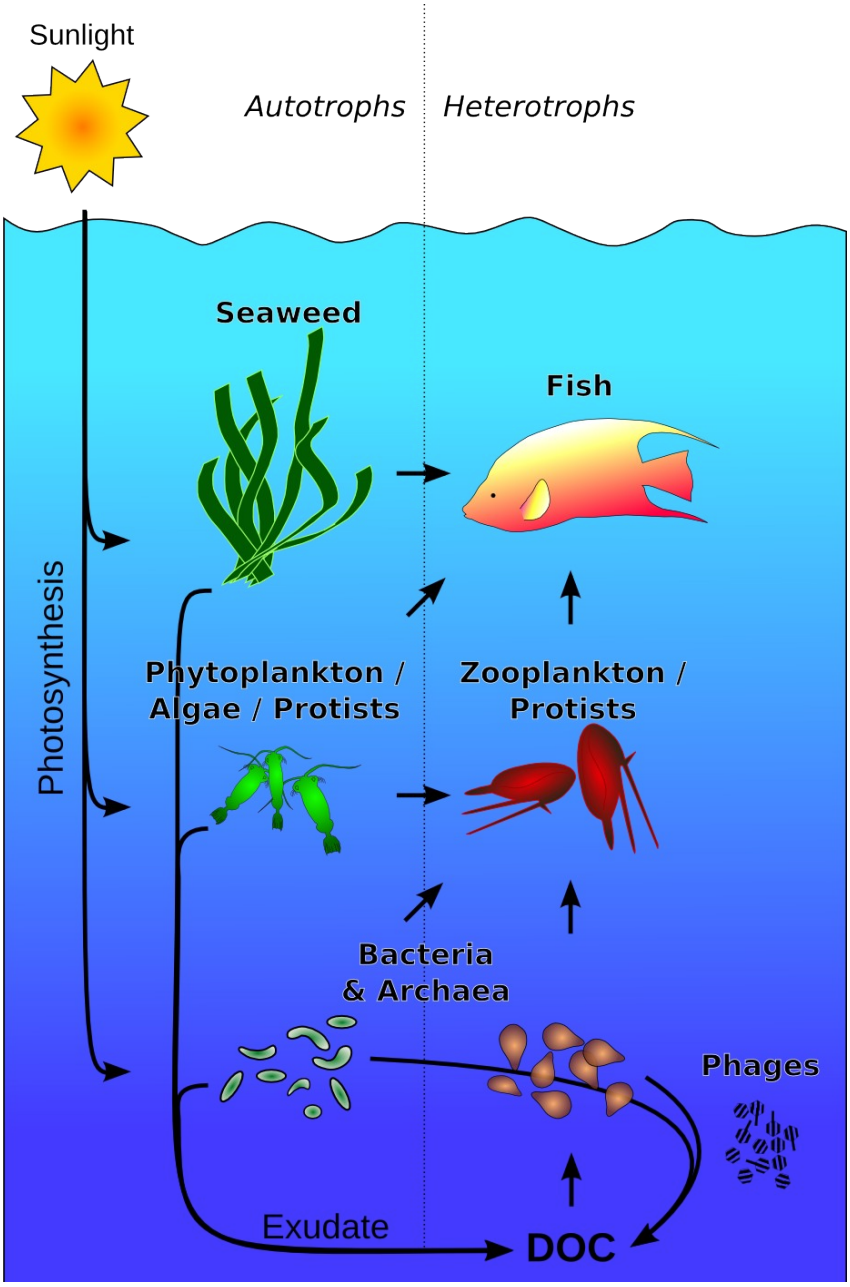


Figure 1.1: Role of phages in the marine food web.

Mamavirus has a 1.7 Mb double-stranded DNA genome, and its 750 nm large capsid is larger than some Bacteria [27].

Viruses, and in particular phages, play an important role in the marine food web. Bacteria incorporate Dissolved Organic Carbon (DOC) present in the water column for their growth [28,29]. The grazing of protists on Bacteria and of larger organisms on protists in turn, drives this carbon to higher levels of the food chain [28,29]. Instead of being sequestered in organisms of increasing size, the carbon contained in Bacteria can return to the DOC pool in the water column by the lytic action of phages [30,31] (Figure 1.1). This viral shunt directly affects important global biogeochemical processes such as the carbon cycle [32,33] and may have consequences that have to be integrated in global warming models [34].

Phages also impact microbial population dynamics, and their impact is as great as that of other predators of Bacteria, such as protists [35]. Predator-prey models such as “Kill the Winner” [36,37] have been advanced to explain the complex dynamics between phages and their hosts. Communities that follow Kill the Winner dynamics consist of a few highly abundant species and a large number of rare species. In Kill the Winner models, the most abundant bacterial hosts are more likely to be lysed due to increased contact with phage predators, and as the population size of these dominant bacteria is reduced, different bacterial species then become dominant [36,37]. The constant reciprocal pressure of phages on their hosts and of hosts on their phages [38] leads to co-evolutionary arms race called “Red Queen Effect” [39-41]. Only the species that

continually evolve to escape predation and outcompete other species maintain their fitness relative to the system and survive.

Viral metagenomics

Shotgun metagenomics was first developed to allow for the study of viral diversity without the limitations of culture-based and marker gene-directed approaches [42]. Metagenomics [43] combines genomics with ecology, and involves isolation of nucleic acids directly from environmental samples to obtain genomic sequences from the full cohort of organisms in an environment, as opposed to the genome of a single species [44,45]. Metagenomic approaches are ideal for studying viruses, since only a small fraction of the microorganisms are culturable [46] and phage species generally only have a very narrow number

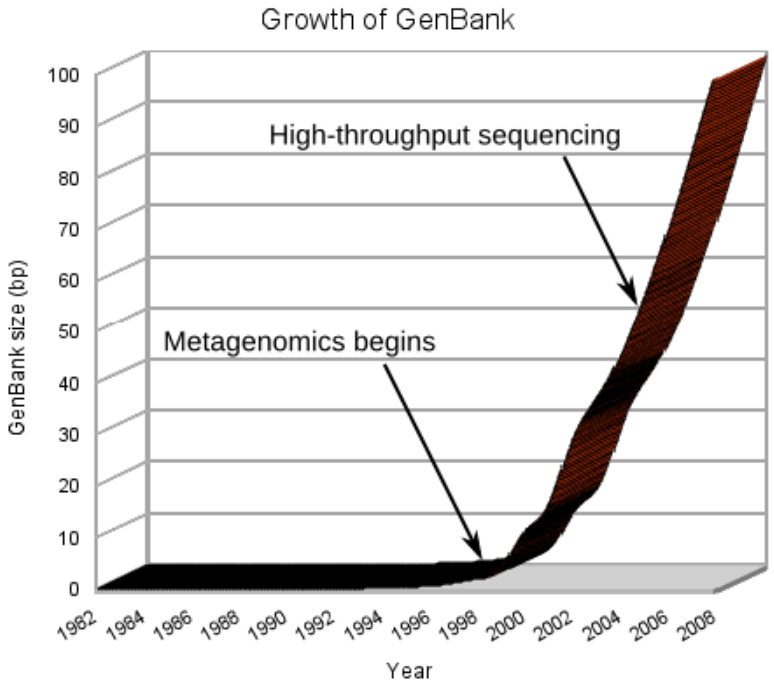


Figure 1.2: The evolution of the size of Genbank

of possible microbial hosts [47]. Metagenomics has been applied to viral communities in a variety of environments [1,48-61] and also to microbial communities [19,49,62-79]. In recent years, metagenomic methods combined with high-throughput sequencing [80-82] has generated unprecedented amounts of sequence data that are responsible for the exponential growth of public sequence databases such as GenBank [83] (Figure 1.2).

Metagenomic sequence data is used to find new enzymes [84-87], study evolutionary history [88], sequence novel organisms [50,62,63,89,90], and characterize the ecology of natural communities, groups of organisms living in the sample place at a given time [91]. In ecological studies, metagenomics is used to describe the structure and function of naturally-occurring communities by answering three central questions:

- Who is there? What species are present? (taxonomy)
- What are they doing? What genes do their genomes encode? (function)
- How many are there? How many different species/genes are present? (diversity)

To begin answering these questions, metagenomic sequences are usually compared to databases of annotated sequences (known species and known function) using local similarity search tools such as BLAST [92]. Public platforms for metagenome analysis such as MG-RAST [93], CAMERA [94] and IMG/M [95], or specialized software like MEGAN [96] and KARMA [97] extensively employ

similarity searches. Most of them annotate sequences using only the best similarity. However, the best similarity may not extend to the entirety of the query sequence, may not be from the most closely related organism, and metagenomic sequences may be highly similar to more than one sequence in the database [98]. Cutoff values for significant similarities are often determined arbitrarily and are based on BLAST expect values (E-values), which change depending on the size of the database used [99]. Additionally, in practice, many metagenomic sequences are be from novel organisms and thus have no similarities to sequences in existing databases. These sequences are categorized as unknown, and are often discarded in subsequent bioinformatic analyses [100].

Few methods can make use of all reads in a metagenomic dataset. The frequency of the oligomers in metagenomic sequences has been characterized previously. This similarity-independent method has shown that metagenomes from different biomes have distinct oligonucleotide signatures [101]. Assembly of metagenomic sequences, which plays an important role in this thesis to estimate diversity [48,50-53,57,102], also does not rely on the existence of similarities to sequences in databases. Sequence assembly is further an efficient method to reconstruct the genome sequence of unknown viruses [50,60,61,63].

Quantifying biodiversity

The estimation of diversity is more than an exercise in species enumeration. The loss of biodiversity has important socio-economical impacts [103,104].

Quantification of biodiversity is thus an important aspect of conservation efforts. Biodiversity in space is characterized in three ways [105,106]. α-diversity defines the diversity of a given location (or sample, or ecosystem), for example the number of bird species in a given wood. On a larger scale, γ-diversity captures the cumulative diversity of several locations, for example, the number of bird species in all the woods of a country. Finally, β-diversity measures the difference in diversity between several locations, for example how many species of bird are unique to each wood.

There are three components which comprise α-diversity: i) richness, or how many species there are (the more species, the more diverse the community), ii) evenness, or how evenly species are distributed in the community (if some species are numerically dominant, the community is considered less diverse), and iii) phylogenetic relatedness, or how closely related the species are (more phylogenetically distant species reflects a higher diversity) [107,108]. Many metrics capture one or several of these aspects of α-diversity into a single number. Let M be the number of species (richness) in a sample, R the total number of individuals in this sample, and f_i the relative abundance of the i^{th} species, then the following are defined:

- Margalef's richness [109]: A measure of richness normalized by sample size. $G = \frac{M - 1}{\ln R}$
- Shannon-Wiener index [110]: Adapted from information theory, it takes into

account species richness and relative abundance (on which evenness depends). $H' = -\sum_{i=1}^{M} f_i \ln f_i$

- Pielou's evenness [111]: $P = \frac{H'}{H'_{max}} = \frac{H'}{\ln M}$
- Simpson's index [112]: Measured as the probability that two individuals drawn at random from a community belong to different species.

$$D = 1 - \sum_{i=1}^{M} f_i^2$$

- Berger-Parker index [113]: This index is the abundance of the most abundant species. $B = \max_{1 \leqslant i \leqslant M} (f_i)$

The notion of species is difficult to define [114,115]. Taxons, Operational Taxonomic Units (OTUs), genotypes, or other taxonomy-related definitions are often used but biodiversity can also refer to more than the diversity of species. Species perform functions that are essential for the functioning of the ecosystem they live in. For example, corals on a reef provide shelter and breeding ground for a multitude of fish. In addition, corals are calcifying organisms that alter how much carbon dioxide is in the ocean. Functional diversity focuses on what the species do, not what they are. In fact, the diversity of functions performed in an ecosystem may be more important than the diversity of the species themselves for the proper functioning of this ecosystem [116]. The functional diversity of viruses and microorganisms can be accessed through their metabolism, i.e. their

gene content [64,117,118].

The species diversity of a community is reflected by its community structure, a representation of the arrangement of species inside their community (e.g., their relative abundance). Determining community structure may provide clues into the functioning and dynamics of its individuals. For example, the power law community structure often observed in viral communities [51,102,119] could be the result of a particular phage-host “Kill the Winner” dynamics [120]. Rank-abundance curves, or Whittaker plots [121], provide a visual representation of community structure. On these plots, the Y-axis represents the relative abundance of species, while on the X-axis, anonymous species are ranked by decreasing relative abundance, the species with rank 1 being the most abundant (Figure 1.3).

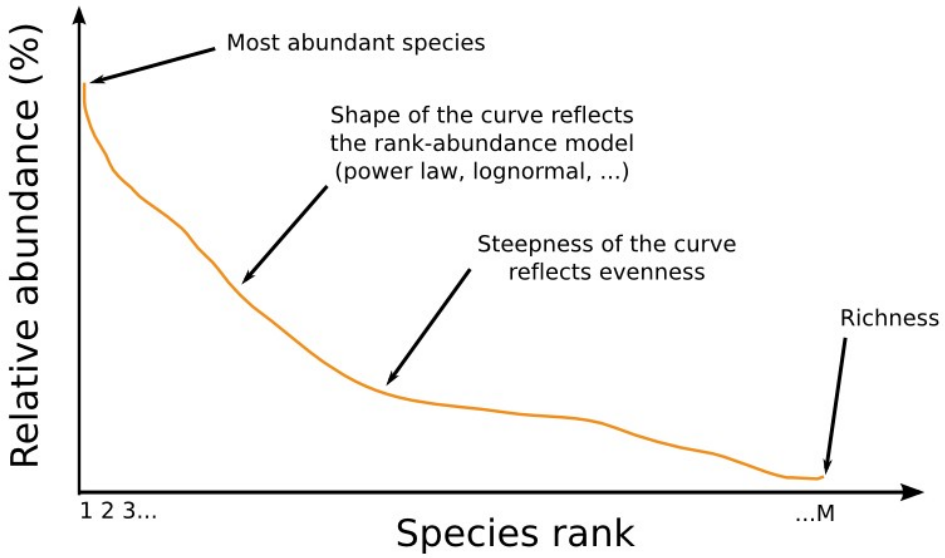


Figure 1.3: A rank-abundance curve depicts community structure as a list of anonymous species ranked by abundance.

Various rank-abundance models have been proposed to model community structure. Popular models have the common characteristic of exhibiting a large drop-off in the relative abundance of the first few species. In the following model equations, *M* represents the sample richness, f_i is the relative abundance of the i^{th} most abundant species, and *a* and *b* are parameters of the rank-abundance model to be determined:

- Power law: An empirical model that describes many natural phenomenons [122,123]. $f_i = a i^{-b}$ for $1 \leqslant i \leqslant M$
- Logarithmic: Another empirical model [123]. $f_i = a\left(\log(i+1)\right)^{-b}$ for $1 \leqslant i \leqslant M$
- Exponential: Empirical model [123]. $f_i = a\mathrm{e}^{-ib}$ for $1 \leqslant i \leqslant M$
- Broken-stick: An ecological model based on a partitioning of resources between species [124]. $f_i = \frac{R}{M}\sum_{h=i}^{M}\frac{1}{h}$ for $1 \leqslant i \leqslant M$, where *R* is the total number of individuals sampled.
- Niche preemption: Also based on resource partitioning [125]. $f_i = R\,a(1-a)^{i-1}$ and $f_M = R(1-a)^{M-1}$ for $1 \leqslant i \leqslant M-1$.
- Lognormal: A commonly used model with theoretical justifications [126].

$$f_i = \frac{e^{k_i \sigma}}{\sum_{h=1}^{M} e^{k_h \sigma}} \quad \text{with} \quad k_h = \frac{M}{\sqrt{2\pi}} \left(e^{-l_h^2/2} - e^{-l_{h+1}^2/2}\right) \quad , \quad l_1 = -\infty \quad ,$$

$$l_{h+1} = \sqrt{2}\, erf^{-1}\left(\frac{2}{M} + erf\left(\frac{l_h}{\sqrt{2}}\right)\right) \quad \text{and} \quad l_{M+1} = +\infty \quad \text{for} \quad 1 \leqslant i \leqslant M \quad \text{where } erf$$

is the error function and erf^{-1} its inverse.

- Unified neutral theory: Unlike other ecological models, the unified neutral theory assumes that the fitness of different species is the same [127,128]. The abundance of species in this model is caused by an equilibrium between speciation and extinction and can be solved numerically.

$$Pr(r_1, r_2, \ldots, r_M | \theta, R) = \frac{R!\,\theta^M}{1^{\phi_1} 2^{\phi_2} \ldots R^{\phi_R}\, \phi_1!\,\phi_2! \ldots \phi_R! \prod_{k=1}^{R} (\theta + k - 1)} \quad \text{where}$$

$\theta = 2\mathrm{R}\nu$. The symbol ν designates the speciation rate, ϕ_k the number of species with *k* individuals, and r_i the number of individuals belonging to species *i*.

Determining the rank-abundance model that best fits empirical species abundance observations is a non-trivial task that was originally done visually [129]. Visual fitting is inappropriate to distinguish between similar models and it is complicated by sampling biases that cause rare species to be undersampled, resulting in a lack of the tail of rank-abundance curves [130]. Tools were created recently to address these limitations [131,132]. Once the community structure is

known, it is straightforward to calculate a variety of diversity measures.

Patterns of diversity

Diversity in the environment has been reported to vary according to specific patterns potentially caused by global but poorly understood forces [133]. The latitudinal gradient of diversity has a long history and was first reported by von Humboldt [134]. He noted that as latitude increased, the variety of plants species decreased, i.e. their richness was higher at the equator than at the poles. Nowadays, it is recognized that species richness reaches a maximum at low latitude, not exactly 0° (Figure 1.4A). A similar pattern exists for elevation, the elevational gradient of diversity, in which richness is negatively correlated with altitude [135]. Modern impacts of humans on the environment [136] provide a good ground to study another gradient, the intermediate disturbance gradient, in which a disturbance that gradually increases in frequency or intensity causes diversity to progressively increase until it dramatically collapses [137-139]. A last pattern is the species-area relationship [140,141]; the number of species found in an area was found to correlate with the size of this area according to a power function: $M = d\,A^{e}$ where *M* is the species richness, *A* is the area and *d* and *e* are constants (Figure 1.4B).

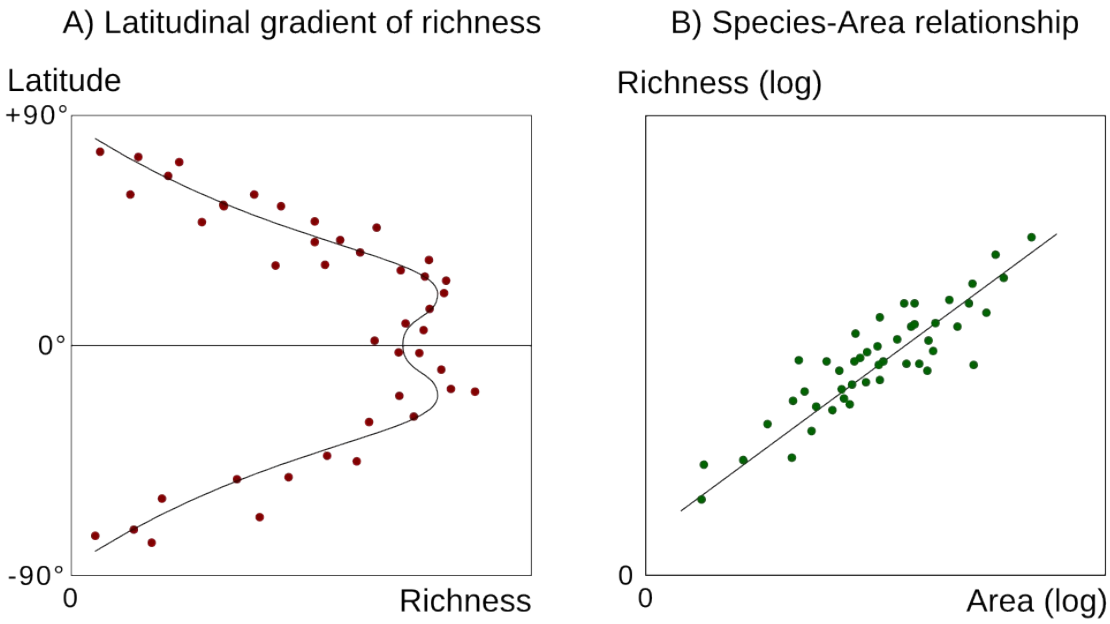


Figure 1.4: Theoretical data demonstrating different diversity patterns. A) The richness as a function of latitude follows a latitudinal gradient. B) The species-area relationship appears as a straight line on a log-log plot.

The latitudinal gradient of richness is the most well-known of the diversity patterns [142]. It is also very general and has been shown to range from aquatic to terrestrial biomes, for various organisms with a mass spanning over height orders of magnitude [143]. Despite this, it is unclear what causes it. Explanations for its existence have been advanced and are arranged in three categories. Historical reasons argue that the low species richness of the poles is due to the lack of time available for species to migrate and colonize these areas after historical events such as glaciations [144]. On the other hand, ecological factors have supporters that claim that increased richness in the tropics is reached because of larger speciation rates caused by stronger biotic interactions such as

predation, competition, and mutualism [145]. Finally, evolutionary hypotheses stipulate that a higher evolutionary rate in the tropics is responsible for higher speciation rates, and hence increased richness [146].

The diversity of microbial communities has been estimated using molecular methods such as the Polymerase Chain Reaction (PCR), Automated method of Ribosomal Intergenic Spacer Analysis (ARISA) [147], Terminal Restriction Fragment Polymorphism (TRFLP) [148], Pulse Field Gel Electrophoresis (PFGE) [149], and Denaturing or Temperature Gradient Gel Electrophoresis (DGGE and TGGE) [150,151]. Evidence from surveys using these techniques indicate that microorganisms may follow the same patterns of diversity as macroorganisms. For example, two studies suggest that marine Bacteria are subject to the latitudinal gradient of diversity [152,153].

Many of the tools used to investigate microbial diversity are not applicable to viruses because they lack common marker genes [154,155]. PFGE and other lab techniques that are used on viruses are often expensive, time-consuming and impractical for large scale studies. Therefore, it remains to be seen if viral communities follow the same patterns of diversity as microorganisms and macroorganisms, i.e. if they respond in the same way to the same global forces.

Characterizing viral biodiversity

Viruses have been referred to as the dark matter of the biosphere [156] because only a small fraction of their diverse species has been inventoried. In

this thesis, I show how to take advantage of the power of metagenomics by using all metagenomic sequences (including the unknowns) to investigate the diversity of uncultured viral communities. First, I detail a novel computational method to quantify the α-diversity of viral metagenomes in Chapter 2. Building on this method, Chapter 3 presents the first approach to evaluate metagenomic viral β-diversity. Then, Chapter 4 introduces an original program to estimate average genome length in microbial and viral metagenomes, which improves α and β-diversity estimations. Finally, I show in Chapter 5 how combining these various tools forms a comprehensive workflow for the characterization of viral diversity from natural communities.

CHAPTER 2: α-DIVERSITY

This chapter introduces PHAge Communities from Contig Spectrum (PHACCS), the first publicly available software designed to estimate viral α-diversity (diversity of a single sample). PHACCS uses contigs as the input to mathematical models of diversity, circumventing the limitations of similarity-based approaches. I developed this research tool and published it in BMC Bioinformatics [102]. The text of this article is attached in Appendix 1.

Hurdles to the estimation of viral α-diversity

α-diversity characterizes the diversity of a single community. Studies on microorganisms typically use the sequence of the 16S rDNA gene, which is a genetic marker shared by all Bacteria and Archaea, to estimate microbial phylogeny and α-diversity without cultivation [157-162].

There is no such common genetic marker for viruses that could be used to assess viral phylogeny and α-diversity [154,155]. Specific proteins of the phage capsid, tail or polymerase have been used to phylogenetically classify phages from specific taxa [163-167]. However, even though particular genes are conserved across one or several viral families, none is universal. Therefore, marker-based approaches are not appropriate to survey the viral communities.

Lab methods such as Denaturing Gradient Gel Electrophoresis (DGGE) [150], Temperature Gradient Gel Electrophoresis (TGGE) [151], and Pulsed-Field

Gel Electrophoresis (PFGE) [149], provide genetic fingerprints used to compare viral community diversity [168]. The number of bands obtained after running viral DNA on an electrophoretic gel is a proxy for species richness [169-173]. Though they may be useful to characterize and compare natural viral communities, these methods are limited in accuracy, reproducibility, and can have biases.

Defining viral species from sequence assembly

A computational method for the estimation of viral diversity from metagenomes (shotgun libraries) was originally developed in [51]. In this study, the investigators considered metagenomic reads which assembled with each other as belonging to the same species. Sequence assembly is typically used in a genomic context to join overlapping sequences into contigs for the establishment of the consensus sequence of a genome [174,175]. In the metagenomic context, by assuming that only sequences from the same species assemble together, the more contigs there are from a given species, the larger the relative abundance of that species in the community. This method is marker-independent and uses all metagenomic sequences for the estimation of diversity. Using mathematical modeling, it allows for a quantitative assessment of biodiversity, that is based not only on how many species are present, but also on how abundant they are.

No assembly software is specific for metagenomes, and chimeric contigs containing sequences from multiple species can be formed. The assembly-based

definition of a viral species is thus dependent on the stringency of the assembly parameters used. In [51], the best assembly parameters were determined by assembling 500 bp DNA fragments originating from 11 phage genomes using Sequencher [176]. The best parameter values determined heuristically were a minimum of 98% identity and 20 bp overlap between two reads. These parameters assembled only sequences from the same phage or very closely related phage species. Since there is a discrepancy between the assembly-based definition of a viral species and the actual viral taxonomy, the term genotype was introduced as a substitute for species.

The Community Lander-Waterman equations

The mathematical models used to estimate diversity from contigs were derived from the original Lander-Waterman equation [177] which expresses the expected number of sequences c_q that are part of a contig of size q as: $c_q = N w_q$ where N is the total number of sequences and w_q the probability that a sequence goes in a q-contig. The Community Lander-Waterman equations are generalized for a community of different species [51]. For a community with a given structure (rank-abundance equation) and richness M, the Community Lander-Waterman equation models the expected occurrence of contigs of different sizes (contig spectrum) (Figure 2.1) as: $c_q = \sum_{i=1}^{M} n_i w_{qi}$ where n_i indicates the number of reads of the i^{th} species.

Modeling viral community structure and diversity is an inverse problem; many community structures are empirically tested until the best-fitting one is found. In [51], the fit of different rank-abundance forms to a contig spectrum obtained from a marine viral community was quantified as the negative log-likelihood, i.e. the sum of the variance-weighted, squared deviations from the observed contig spectrum. Thus, the smaller the negative log-likelihood, the better the fit. Power law and exponential community structures were tested on two marine viral communities, resulting in a better fit of the power law model. The same diversity modeling technique was applied to uncultured viruses issued from human feces a year later [52], and the power law described the community the best.

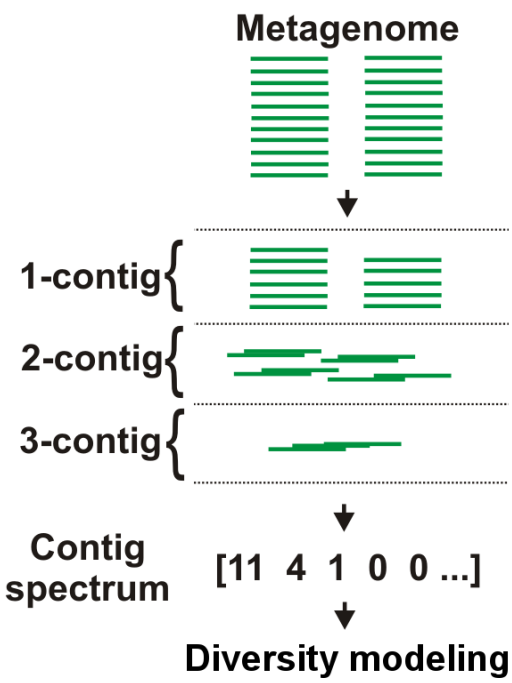


Figure 2.1: Metagenomic sequences are assembled into contigs. The number of contigs of each size is counted to determine the contig spectrum. Taken from Angly et al. (2006) PLoS Biol 4(11):e368 under the terms of the Creative Commons Attribution License.

Later [53], the diversity model was improved by representing the abundance

of phage species as a frequency (or relative abundance). Also, an alternative model appropriate for very even communities and a Monte-Carlo simulation were designed to compare to the original model. The application of the two new methods to newly generated near-shore and sediment viral metagenomes revealed no significant advantage over the original technique.

Modeling viral community structure and α-diversity

I developed PHACCS [102] to improve and extend the contig spectrum modeling approach and provide an easy-to-use web interface. In PHACCS, the Community Lander-Waterman equation was used and the error (opposite of the goodness of fit) between predicted and observed contig spectra was calculated as in the original model. In addition to the power law and exponential rank-abundance forms, PHACCS models communities using the logarithmic, broken stick, niche preemption and lognormal rank-abundance forms (see Chapter 1). To automatically determine the best-fitting model, I implemented an optimization algorithm that iteratively minimizes the error in PHACCS (Figure 2.2). PHACCS results present the community structure in both graphical and mathematical form, and the α-diversity estimates, including the richness, evenness, the Shannon-Wiener index and the Berger-Parker index (abundance of the most abundant genotype). Scientists can execute the PHACCS program online at http://biome.sdsu.edu/phaccs, or at http://portal.camera.calit2.net/ as part of the α-diversity workflow on CAMERA (see Chapter 5).

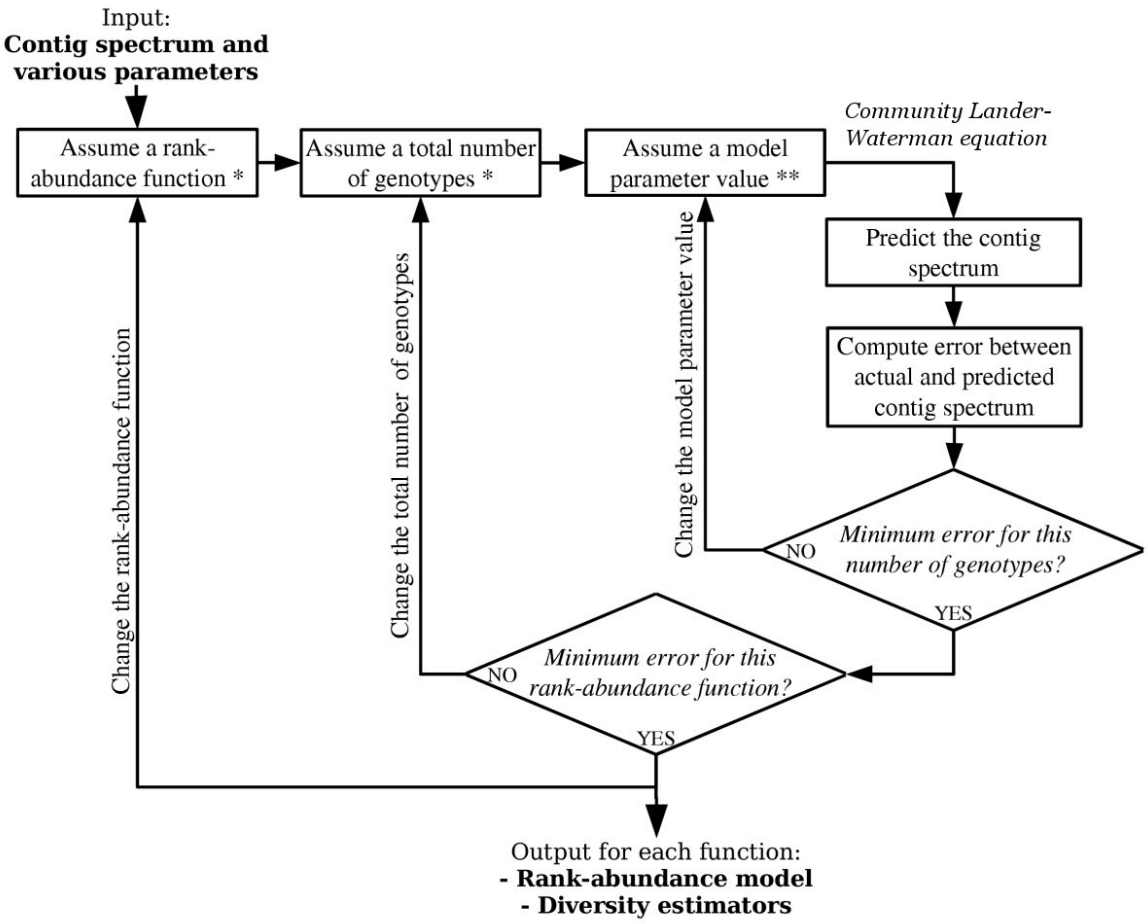


Figure 2.2: The PHACCS algorithm iteratively minimizes the error in fit of the rank-abundance model to the contig spectrum. Taken from Angly et al. (2005) BMC Bioinformatics 6:41 under the terms of the Creative Commons Attribution License.

The four viral metagenomes previously sequenced in [51-53] were analyzed with PHACCS [102] and compared. The power law was the best-fitting community structure in all cases. The viral communities were rich (between 2,390 and 7,340 genotypes), and exhibited different community structures (Figure 2.3). Viral and microbial communities have been reported to covary [178].

The viral diversity reported by PHACCS reflected the diversity of Bacteria in the sediments, water, and human digestive tract [179,180].

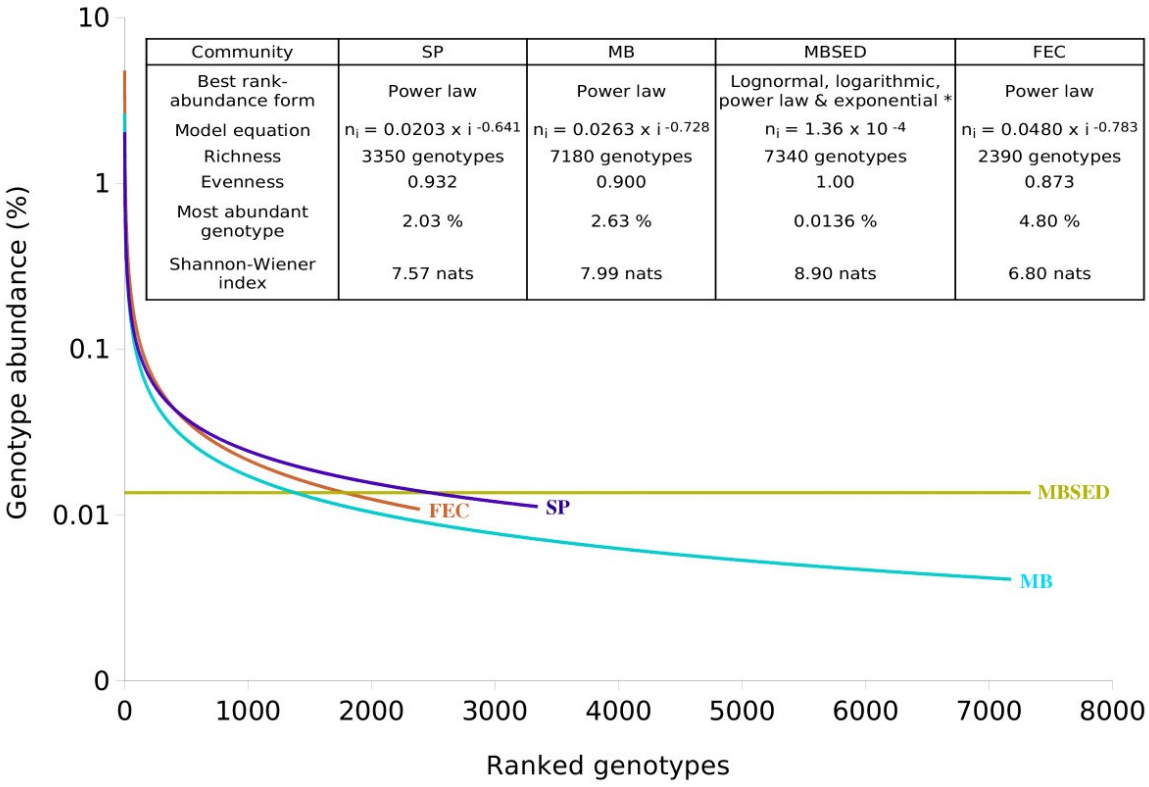


Figure 2.3: Rank-abundance form and α-diversity of four viral communities as determined by PHACCS. SP: Scripps Pier seawater, MB: Mission Bay seawater, MBSED: Mission Bay sediments, FEC: human feces. Taken from Angly et al. (2005) BMC Bioinformatics 6:41 under the terms of the Creative Commons Attribution License.

CHAPTER 3: β-DIVERSITY

This chapter reviews my study which contrasted the composition and distribution of viruses from four different oceanic provinces around North America. This work was published by PLoS Biology [50] and is attached in Appendix 2.

Measures of β-diversity

β-diversity is the difference in diversity between two samples and provides a quantification of differences in species composition between samples taken at different locations or times. Estimating the α-diversity of environmental viral metagenomes with PHACCS provided an opportunity to characterize the viral diversity patterns that exist in nature and determine what large-scale forces shape the evolution and distribution of viruses. However, α-diversity fails to reflect how viral communities with the same α-diversity differ from each other. This aspect is captured by measuring β- diversity.

There are many metrics for assessing β-diversity, both quantitative and qualitative. The simplest quantification of β-diversity is the total number of species unique to each sample *j*: $\beta = \sum_{j} (M_j - C)$, $0 \leqslant \beta \leqslant \sum_{j} M_j$, where M_j is the richness of the j^{th} sample and *C* is the number of species common to all samples. A higher β-diversity represents larger compositional differences between communities. In addition, indices of β-diversity have been developed

based on species presence/absence data and include:

- Whittaker's measure [105]: $W = T / \bar{M}$, where T is the combined richness of all communities, and $\bar{M}$ is their average richness.
- Sørensen similarity index [181]: $S = 2C / (M_1 + M_2)$, for two communities with richness M_1 and M_2. It ranges from 0 (no common species, largest β-diversity) to 1 (all species are in common, lowest β-diversity).

Some β-diversity metrics incorporate the relative abundance of the species in the calculation of diversity, including:

- Bray-Curtis index [182]: $BC = \frac{\sum_i |r_{1i} - r_{2i}|}{\sum_i (r_{1i} + r_{2i})}$ with r_{ji} the number of individuals belonging to species i in sample j.
- Morisita-Horn index [183]: This index is robust to variations in sample size and diversity. $MH = \frac{2 \sum_i r_{1i} r_{2i}}{(\lambda_1 + \lambda_2)\ R_1 R_2}$ where $\lambda_j = \frac{\sum_i r_{ji}^2}{R_j^2}$ and R_j the total number of individuals in sample j.

β-diversity is a fundamental attribute of biodiversity, but it is rarely studied across large spatial scales. A global survey compared the β-diversity of amphibians, birds, and mammals and showed that areas of high β-diversity coincide for these animal taxa, indicating that these regions are highly

susceptible to global climate change [184]. In addition to directing conservation efforts, these findings suggest that there are global processes which affect multiple taxa and lead to high levels of differentiation in natural communities. Viral β-diversity has yet to be characterized on a global scale, and it is unclear if viruses are under the same environmental pressures as macroorganisms.

Distribution of marine viruses

Genomic studies have found that phages represent the largest unexplored reservoir of sequence information in the biosphere [156,185,186]. In metagenomic surveys of viruses, the number of sequences from unidentified species was very high, as was the viral richness [48-58,119,185]. These data suggest that the composition of distinct marine viral communities is very different, i.e. that their β-diversity is large. However, phages are small and non-motile, and are passively transported by currents and winds [18,187-190]. Furthermore, the widespread presence of phage sequences indicates a possible global distribution for some phages [191,192]. Therefore, viruses in the marine environment could be cosmopolitan (have low β-diversity).

By comparing the community composition and β-diversity of four marine viral communities, I determined that phage communities are cosmopolitan, i.e. they exhibit low β-diversity [50]. Four viral metagenomes from distinct marine regions (Arctic Ocean, British Columbia Coast, Sargasso Sea and Gulf of Mexico) were sequenced, bioinformatically analyzed and then compared and contrasted to determine whether they contained mostly unique or mostly shared phage species. I used the Basic Local Alignment Search Tool (BLAST) [92] to identify phages with similarities to known phage genomes. The presence or absence of

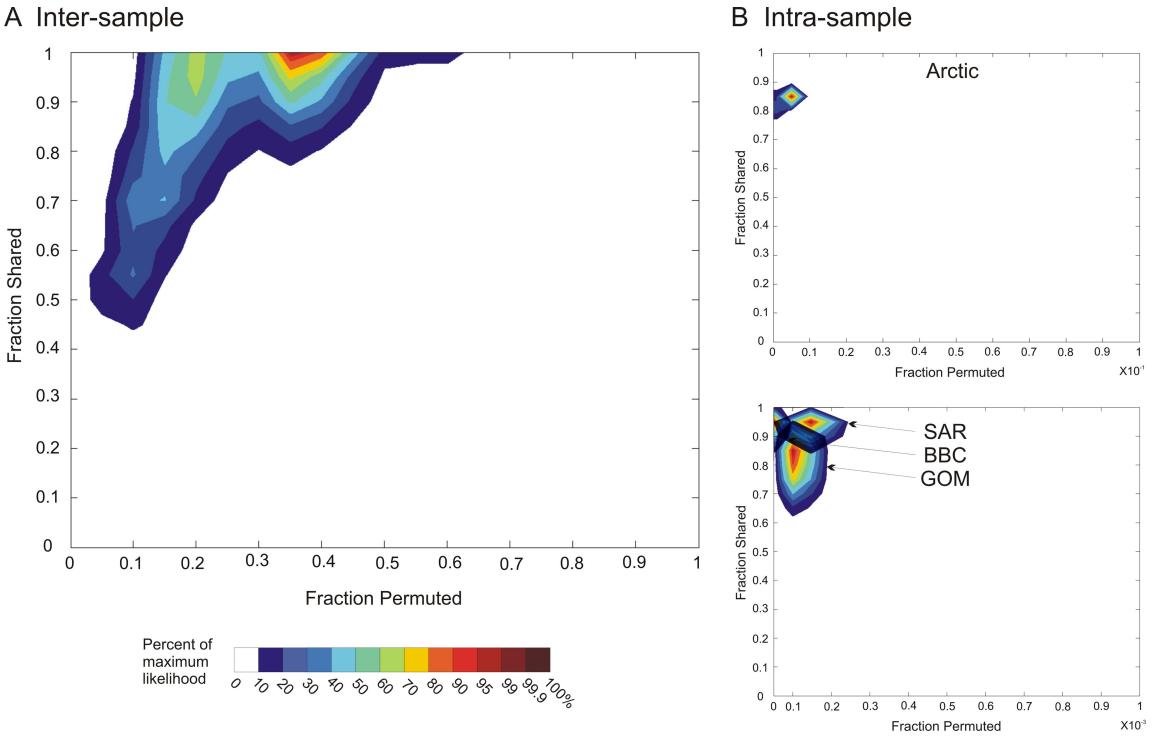


Figure 3.1: A) β-diversity contour plot for four marine viral metagenomes, B) Method controls. Arctic: Arctic Sea, SAR: Sargasso Sea, BBC: British Columbia coast, GOM: Gulf of Mexico. Taken from Angly et al. (2006) PLoS Biol 4(11):e368 under the terms of the Creative Commons Attribution License.

these known phages was plotted on the Phage Proteomic Tree [155], and UniFrac [193] was used to determine whether or not the communities were statistically different. The communities were region-specific (i.e. significantly different) despite sharing over a third of the identified phage species. This approach was limited by the large number of phages that are unsequenced and that were therefore overlooked in the analysis. To characterize the β-diversity of the four viral communities, I used a similarity-independent method called MAXIPHI (described below). All phage genotypes were shared, with a third of the most prevalent genotypes having a different abundance-rank (Figure 3.1A). Low β-diversity supports the notion that marine phages are cosmopolitan and that the unique nature of the viral communities from these marine regions is due to the same phages being present in different abundances.

Assembly of contigs and cross-contigs

The MAXIPHI method was central to the characterization of viral β-diversity and the conclusions of this study. I participated in the development of this novel tool and its validation using controls (Figure 3.1B). The method builds on the contig spectrum modeling approach detailed in Chapter 1. By using cross-contigs, contigs containing reads from multiple metagenomes (during the assembly of multiple metagenomes simultaneously) (Figure 3.2), the method extracts information about genotypes that are present in several viral communities.

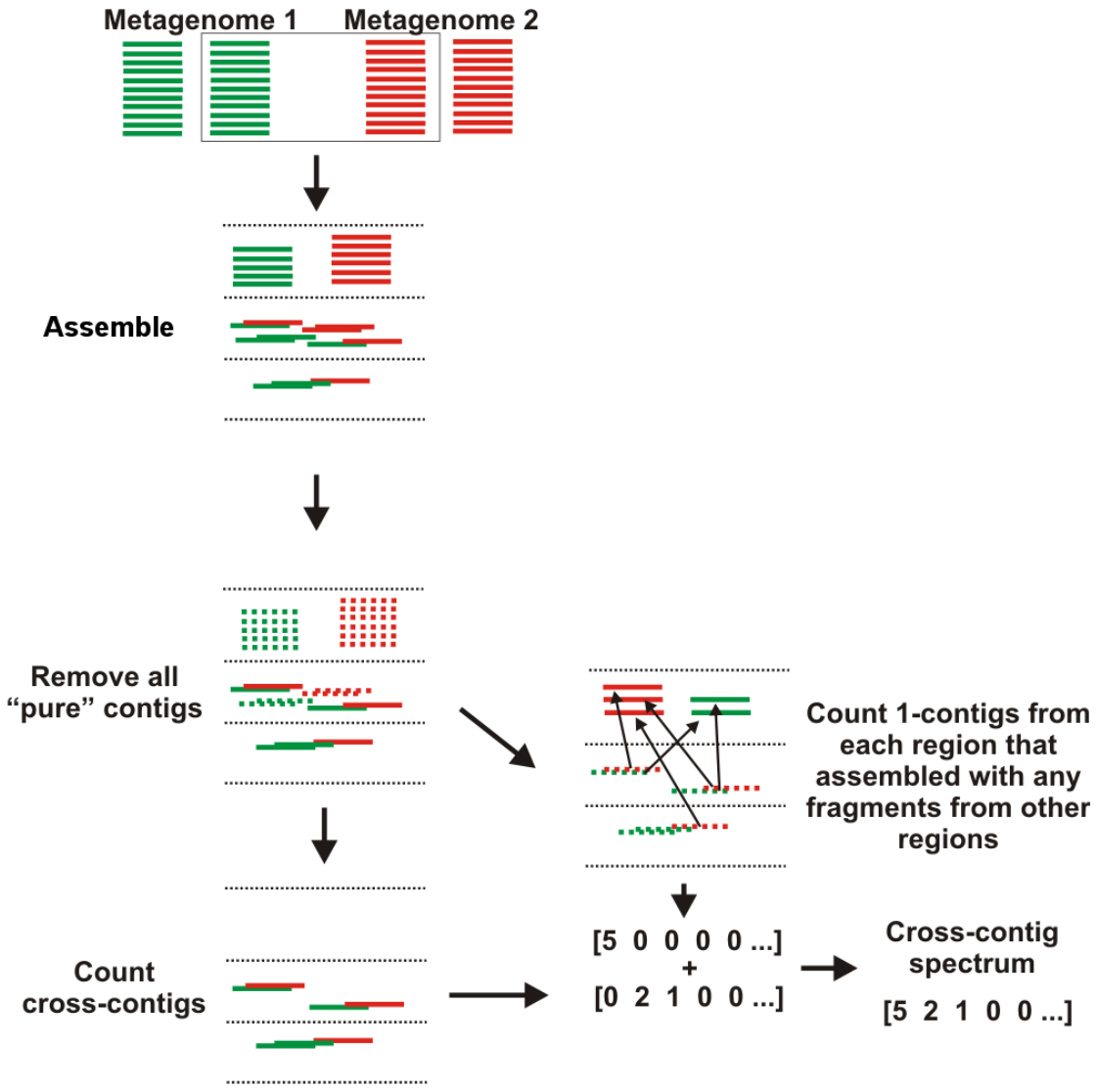


Figure 3.2: Forming a cross-contig spectrum requires assembling sequences from multiple metagenomes and removing contigs that contain sequences from only one metagenome. Adapted from Angly et al. (2006) PLoS Biol 4(11):e368 under the terms of the Creative Commons Attribution License.

To create contig spectra and cross-contig spectra in an automated manner, I designed and programmed Control In Research on CONtig spectra, CIRCONSPECT (http://sourceforge.net/projects/circonspect). The

CIRCONSPECT software assembles one or several metagenomes using TIGR Assembler [194] and calculates their contig or cross-contig spectrum. Since sequence assembly is a $O(n^2)$ problem (Figure 3.3) [195], it was more memory-

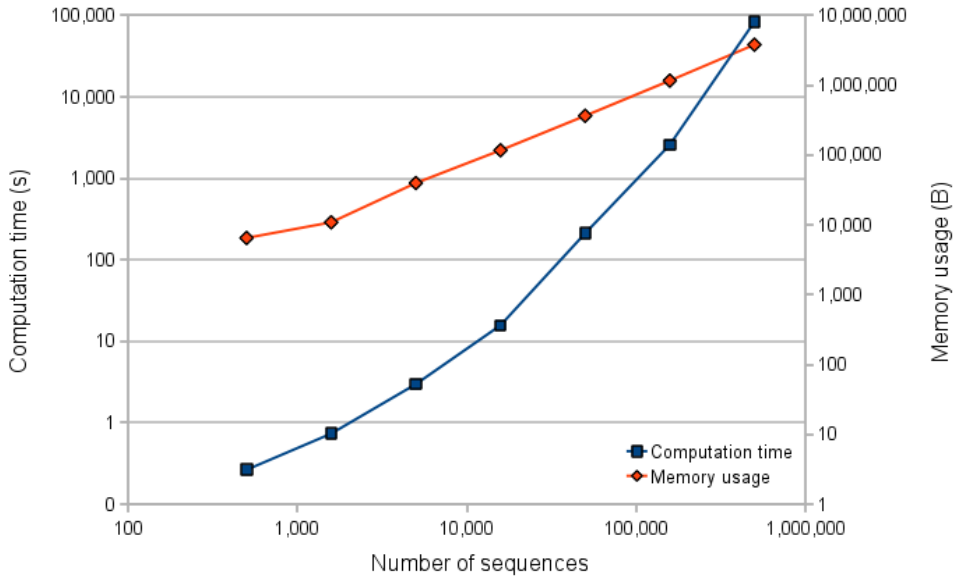


Figure 3.3: Effect of the number of sequences to assemble on the resource usage of TIGR Assembler.

efficient to implement a bootstrap procedure that repetitively assembles a random subset of the metagenomic sequences (e.g. 10,000 sequences) instead of all sequences (Figure 3.4). This partially alleviates the problem of large contigs broken into several smaller ones because of the assembler's inability to deal with the heterogeneous sequence information from multiple genomes. Further, provided a sufficiently large number of repetitions is performed, the bootstrap method covers the totality of the sequence data, from predominant genotypes to rare genotypes, and generates an accurate mean contig spectrum. When comparing different viral communities, using the same number of sequences in

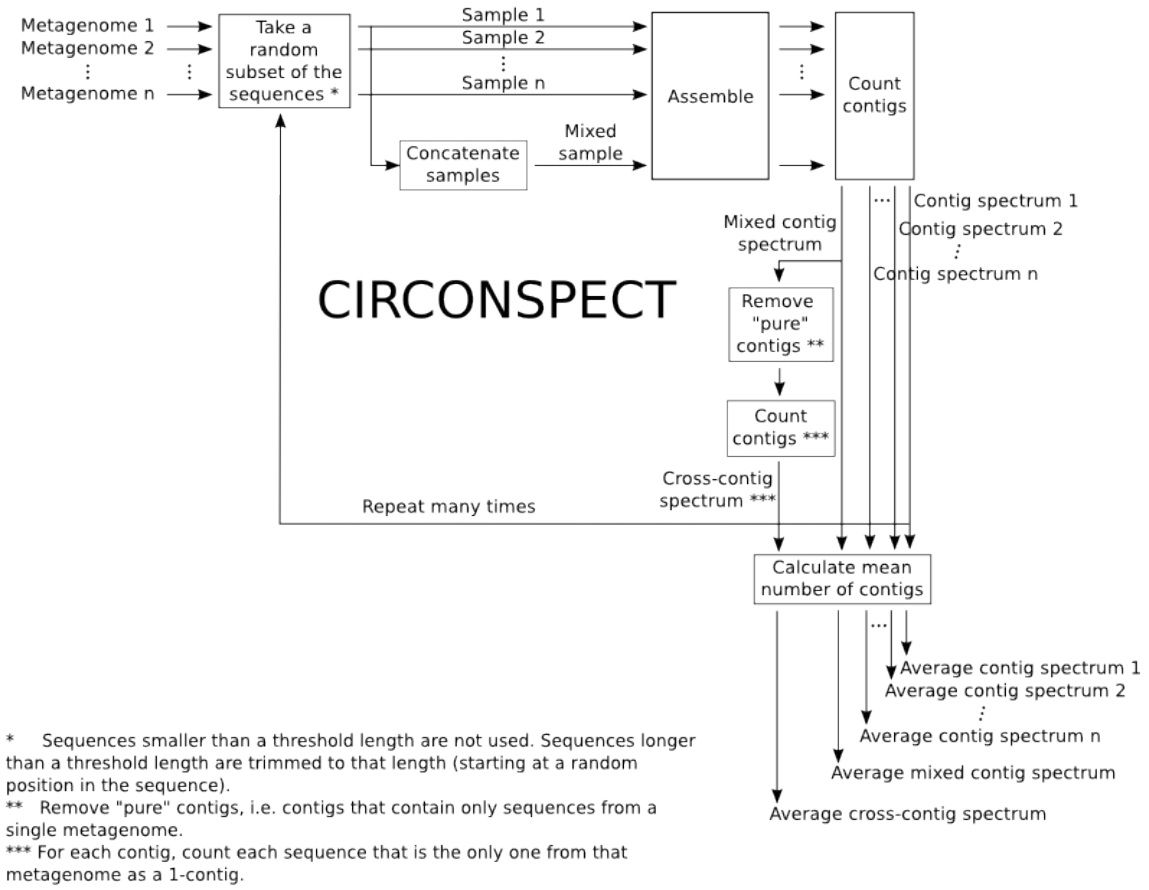


Figure 3.4: Flowchart of CIRCONSPECT, a program to automate the creation of contig spectra and cross-contig spectra in a controlled fashion.

the random subsets is also useful to compare metagenomes with a very different number of sequences. Another feature of CIRCONSPECT is the control of sequence length by trimming long sequences and discarding small ones. With this feature, one can force all the sequences to assemble to have the exact same length, e.g. 100 bp. This avoids the assumption that a distribution of sequences of different lengths is correctly represented by an average value in the average sequence length parameter used in PHACCS. Considering that distinct sequencing technologies used in metagenomics yield sequences of very different

lengths (e.g. ~100 bp for GS20 pyrosequencing, ~700 bp for Sanger sequencing), normalizing sequence length to the lowest common denominator in CIRCONSPECT allows to compare metagenomes without introducing bias.

In Sequencher, a minimum of 98% identity over 20 bp was used to assemble contig spectra [1,48,51-53,55,56,102]. TIGR Assembler implements a greedy overlap-layout-consensus algorithm [194] very different from the assembly algorithm of Sequencher. To accommodate for differences in the functioning the two programs, Circonspect's assembly parameters for TIGR Assembler were reevaluated and changed to 35 bp minimum overlap (and 98% minimum identity) [50,57,58,196].

Modeling the β-diversity of viral communities

MAXIPHI measures β-diversity in a quantitative way since it considers not only the species present but also what their abundances are. The method considers two types of differences in community structure that discriminate between different viral communities: the number of genotypes common to all communities (percent shared), and the number of the common genotypes with a different abundance-rank (percent permuted) (Figure 3.5).

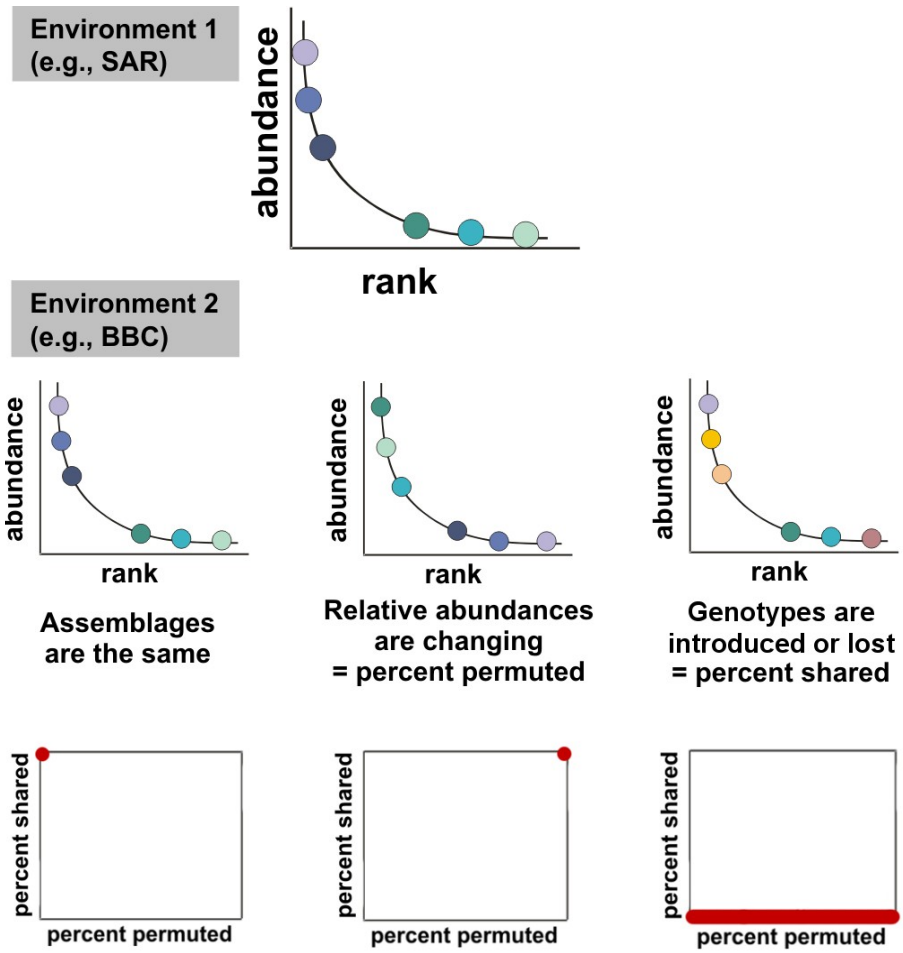


Figure 3.5: β-diversity in MAXIPHI is modeled using the number of genotypes in common and their abundance-rank. The three theoretical cases presented here are: two identical communities (same genotypes in the same abundance)(left), communities sharing the same genotypes but not in the same abundance (middle), and communities with no genotypes in common (right). Adapted from Angly et al. (2006) PLoS Biol 4(11):e368 under the terms of the Creative Commons Attribution License.

The β-diversity, or percent of species shared and percent of species permuted, was evaluated by performing Monte-Carlo simulations on the cross-

contig spectrum. Over the parameter space (*s*,*p*) representing the percent of shared species, *s*, and percent of species with a permuted abundance rank, *p*, many Monte-Carlo repetitions were performed in order to calculate a mean $\hat{c}_q$ and variance $\hat{\sigma}_q^2$ of the predicted cross-contig spectrum. A quasi- likelihood *L*(*s*,*p*) of matching the observed cross-contig spectrum $\hat{c}_q'$, used to generate a contour map of *L*, was obtained by $\ln L(s,p) = -\sum_q \frac{(\hat{c}_q' - \hat{c}_q)^2}{2\hat{\sigma}_q^2}$. The overall procedure is summarized as a flowchart (Figure 3.6).

The novel method to estimate the β-diversity of viruses from metagenomic data described above was essential to determine that the β-diversity of viruses in the oceans is low. The ability to estimate β-diversity of viral communities complements the α-diversity estimates to provide a more comprehensive view of the distribution of viral species in the environment.

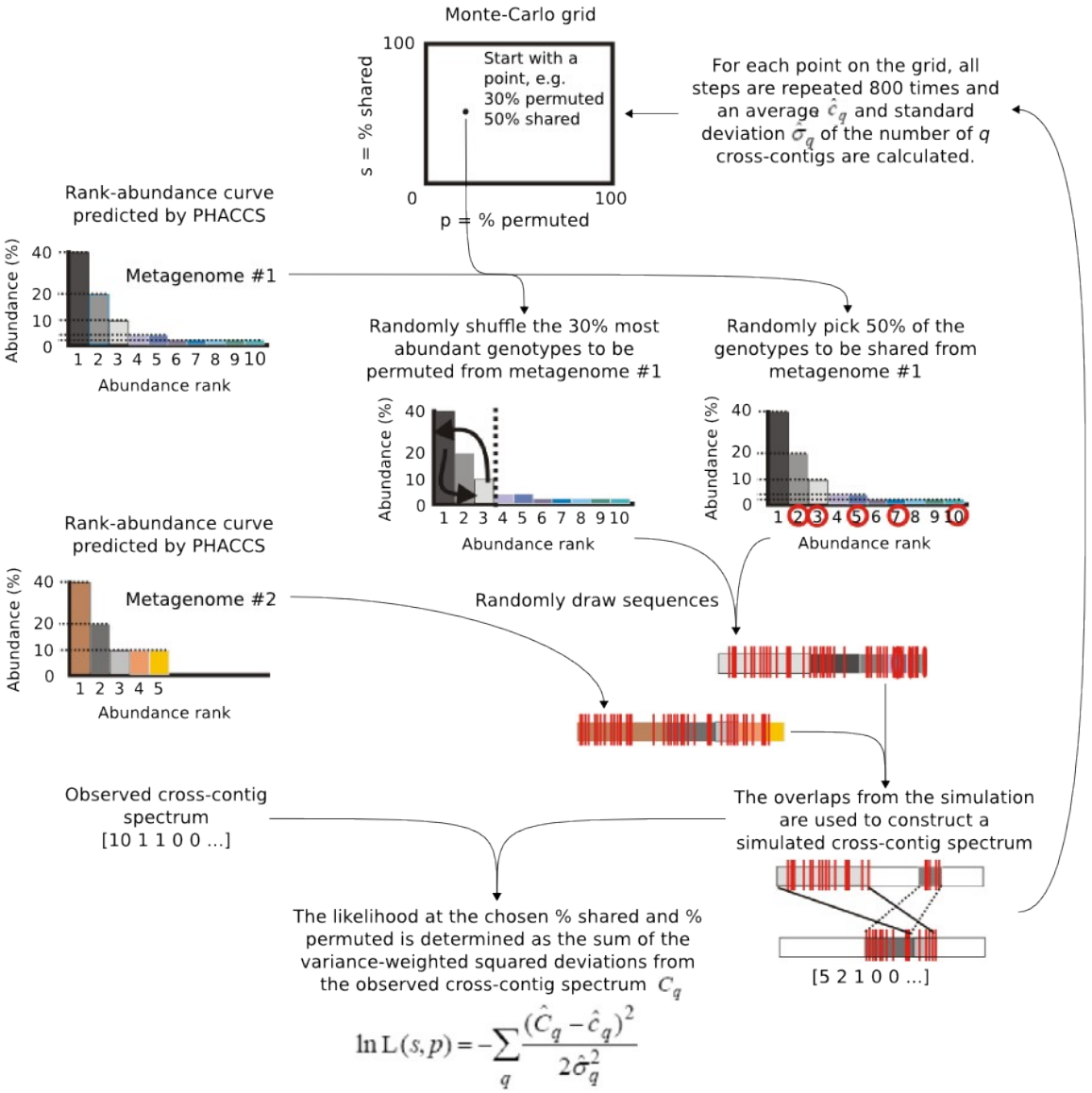


Figure 3.6: Overview of the Monte-Carlo procedure used in MAXIPHI. Adapted from Angly et al. (2006) PLoS Biol 4(11):e368 under the terms of the Creative Commons Attribution License.

CHAPTER 4: AVERAGE GENOME LENGTH

This chapter describes Genome relative Abundance and Average Size (GAAS), a novel metagenomic tool I developed to accurately estimate species relative abundance and average genome length for viral and microbial communities. This work was provisionally accepted for publication in PLoS Computational Biology in August 2009 and is attached as Appendix 3.

Influence of the average genome length on diversity estimates

Genome size refers to the amount of nucleic material in a genome, expressed as a weight or a number of base pairs. The models implemented in

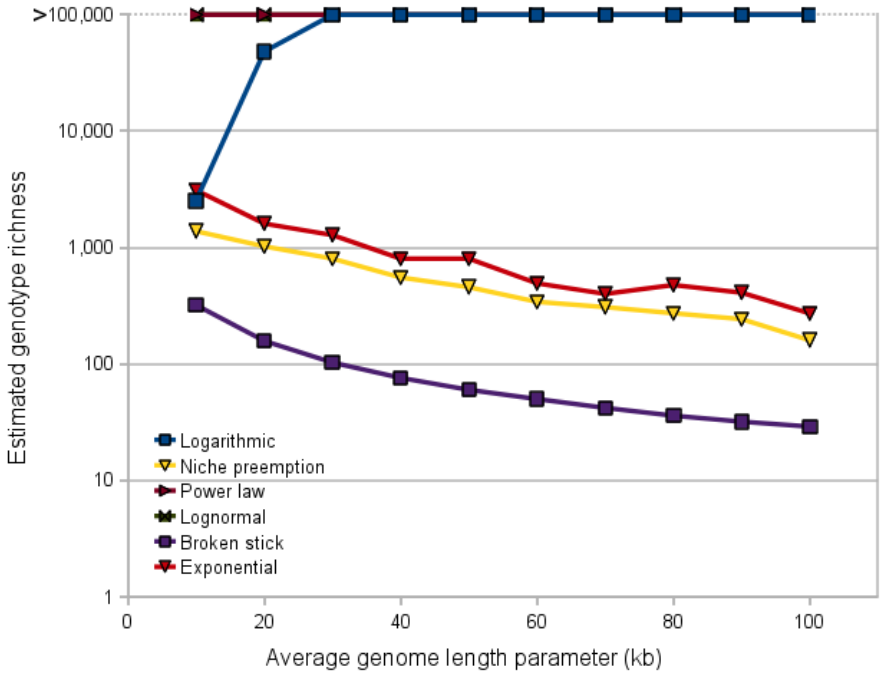


Figure 4.1: Effect of varying the average genome length on the richness estimates of PHACCS for the Sargasso Sea virome.

PHACCS to obtain diversity estimates use the average length of the genomes in a viral community as an input parameter. I tested varying the average genome length from 10 to 100 kb in the PHACCS analysis of the Sargasso Sea virome. Average genome length had a strong influence on the richness estimates. Different rank-abundance models responded differently to changes in average genome length, and the richness was changed by as much as ~40X in the case of the logarithmic model (Figure 4.1). An accurate average genome length is needed to maintain precision in the α-diversity computation.

Methods for estimating average genome length

The genome length of viruses spans three orders of magnitude (Figure 4.2), from the 1.7 kb circular single-stranded DNA genome of a Circovirus [197] to over 1.7 Mbp for the Mamavirus [27]. Pulsed Field Gel Electrophoresis (PFGE) has been used previously to characterize the genome size of viruses in natural communities. In various environments (e.g. rumen, freshwater, feces), PFGE determined the presence of phages with a genome length ranging from 10 kb to 850 kb [52,198-202]. In the oceans, viruses from 8 to 533 kb were detected using this method and the relative intensity of the bands on the PFGE gel allowed the estimation of an average of 50 kb [172,173,178,203,204]. Not having precise estimates of viral average genome length for more than the marine environment adds uncertainty to the exploration of viral α-diversity in new environments using PHACCS. In addition, PFGE's precision is dependent on the experimenter and is

time-consuming [205], making it impractical for large-scale studies. These limitations illustrate the need for a software solution to estimate average genome length in individual metagenomes.

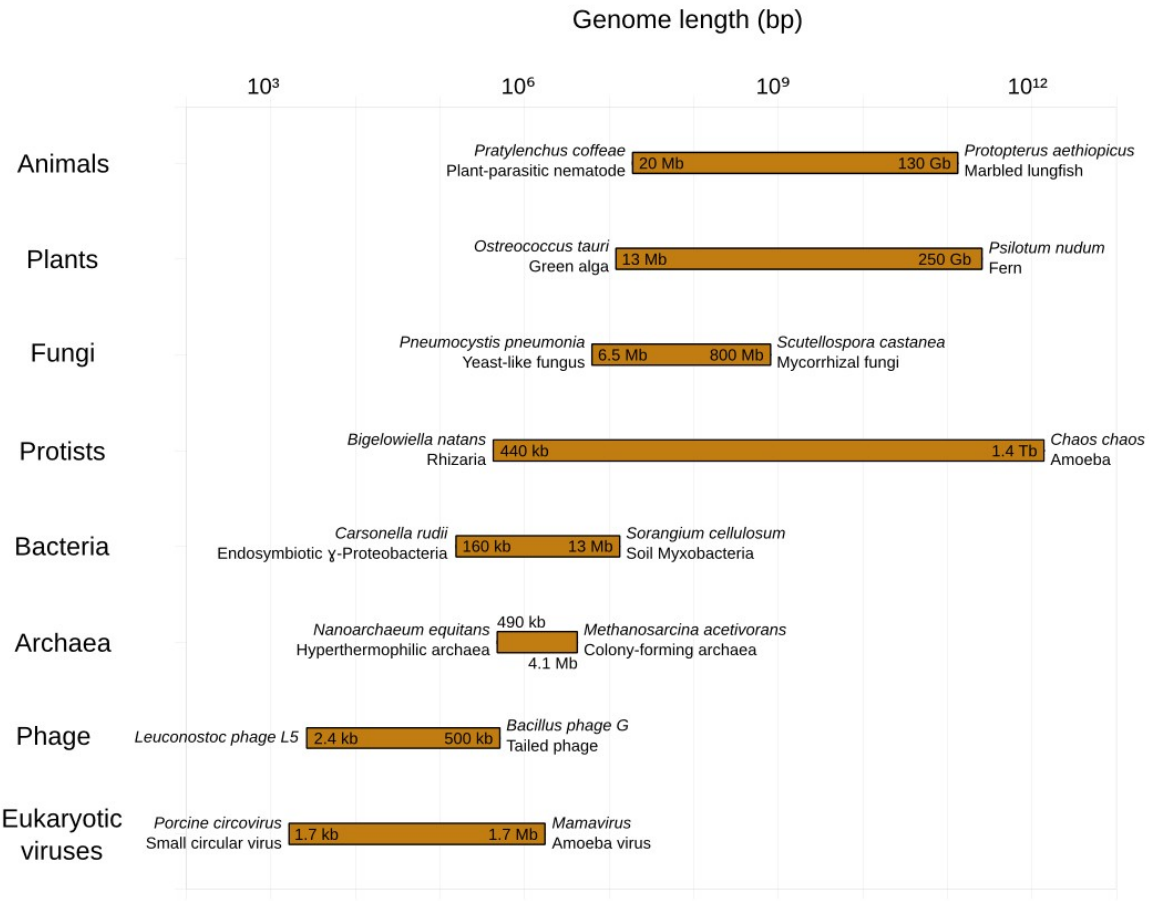


Figure 4.2: Upper and lower limits for the genome size of macroorganisms, microorganisms and viruses. I compiled the data on this graph from various sources: the NCBI RefSeq database, the ICTVdb, the Microbe Wiki, the Fungal Genome Size Database, the Plant DNA C-values Database and the Animal Genome Size Database.

A computational method, Effective Genome Size (EGS), was previously developed to calculate the average genome length in environmental samples using metagenomic data [206]. The EGS method relies on identifying selected marker genes that occur only once per genome, regardless of genome length, so that the total number of marker genes is inversely correlated with the average length of the genomes in the sample. From the density *D* of these genes in an environmental dataset, average genome length is calculated using the equation

$EGS = \frac{x + y K^{-z}}{D}$, with *K* the read length (bp), and *x*, *y* and z parameters that were calibrated using genomes of known size in public databases. The method performed well for the calculation of bacterial and archaeal average genome length. However, no set of markers is present in all viruses [154,155], and hence, the EGS method is not adapted to the study of phages communities.

Biological implications of average genome length

Average genome length is more than a parameter for the determination of viral diversity. For microorganisms, larger genome are characteristic of the copiotroph lifestyle [207] and is strongly correlated with a larger array of genes [208], used to process more resources [209]. The downside of a larger genome is a higher energetic maintenance cost and more complex regulation mechanisms [210]. Therefore bacterial species with larger genomes may be more adapted to environments with scarce but diverse resources, such as soil

[211]. In concordance with this hypothesis, the EGS method demonstrated that the average genome length of microorganisms was higher in soil samples than in other samples (Sargasso Sea, whale falls, acid mine drainage) [206]. Average genome length was also shown to be correlated with environmental complexity. It is not known whether the average genome length of viruses correlates with that of microorganisms and whether it is an indicator of environmental complexity.

Average genome length from sequence similarities

I designed the GAAS program (http://sourceforge.net/projects/gaas) to calculate the average genome length of uncultured viral and microbial communities, and also to provide more accurate estimates of community composition. I used GAAS to estimate average genome size and composition for metagenomes from diverse biomes and conducted a meta-analysis to determine if viral average genome length covaries with microbial average genome length. Complete details are given in Appendix 3. Briefly, GAAS is a novel tool that performs BLAST local similarity searches [92] between the metagenomic reads and a database of complete genomes to calculate average genome length. I assumed that the length of the genome from which a metagenomic sequence comes from is the same as that of the genome that it is similar to, because genome length tends to remain constant within taxa [212]. GAAS implements several methods described below that improve local similarity searches and correct for sampling biases (Figure 4.3).

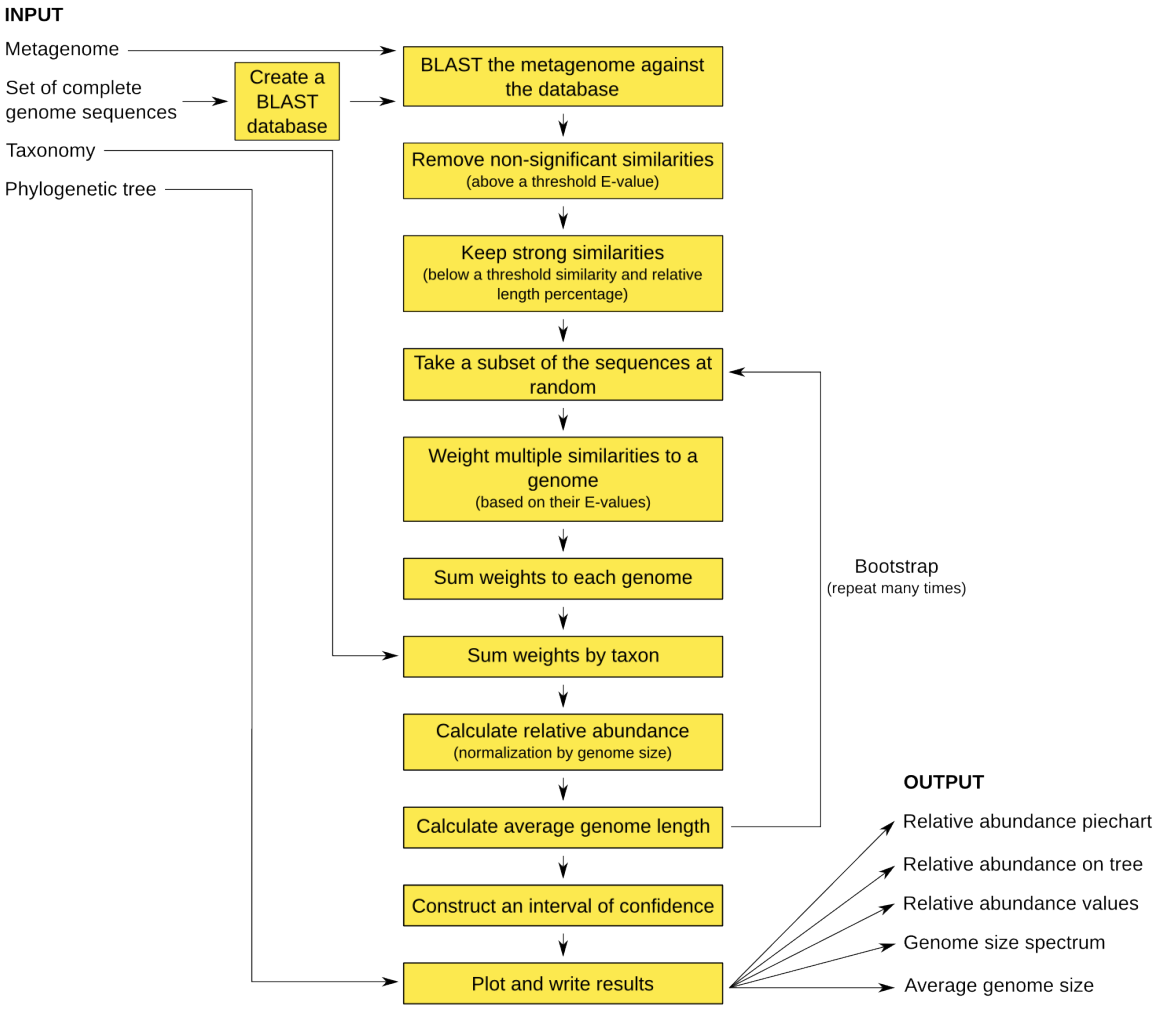


Figure 4.3: Flowchart illustrating how GAAS calculates community composition and average genome length.

E-values, or “expect values” [99], characterize how strong the similarity between two sequences is. Typical metagenomic studies that use BLAST simply use a cutoff E-value, that does not have an intuitive meaning, and that corresponds to a different threshold when using different databases. In GAAS, I used two criteria to select strong similarities likely to reflect sequence homology, a minimum alignment similarity and relative length (or hit coverage [213]). The

alignment relative length, or ratio of the alignment length over the query sequence length, is a way to remove short similarities, that are often similarities to protein domains present in a large number of unrelated taxa.

After the initial filtering, only the top similarity (the one with the lowest E-value) is usually kept for each metagenomic sequence. Instead, in GAAS, I kept all similarities that passed the filter and gave multiple similarities for a given metagenomic read different weights. Sequence similarity networks have used weighted E-values before [214] but the weights were calculated differently here, as inversely proportional to a per-genome expect value, i.e. an E-value normalized to the length of the target genome instead of to the length of the BLAST database used: $W_{uv} = g_v \frac{s'}{E_{uv}\, t_v'}$ where s' is the "effective length" [215] of the database (in number of residues), E_{uv} is the E-value between a metagenomic sequence u and a target genome v, t_v' is the effective length of the target genome v, and g_v is a constant such that $\sum_u W_{uv} = 1$.

Another correction originates from the observation that random shotgun libraries (e.g. metagenomes) are biased toward large genomes; the number of sequences from a given species is proportional not only to its relative abundance, but also to its genome length. While this is a well-known bias in proteomics [216-218], metagenomic studies typically ignore this effect. My correction consisted in normalizing the weights by the length of the genome to

obtain accurate genome relative abundance *f*: $f_v = o \frac{W_{uv}}{t_v}$ where *o* is a constant such that $\sum_j f_j = 1$, and t_v is the length of the target genome *v* (in bp).

To generate empirical confidence limits for length and relative abundance estimates, a bootstrapping procedure was implemented in GAAS. Empirical confidence intervals for genome relative abundance and average genome length were calculated by repeating the computation many times using a random subsample of the metagenome at each repetition. Confidence intervals were taken as the weighted percentiles of the observed estimates, e.g. 5th and 95th percentiles for a 90% confidence interval.

Method validation with simulated metagenomes

I validated the GAAS method using an extensive set of benchmarks (see Appendix 3 for details). The benchmarks consisted of ~10,000 simulated metagenomes, which were made with Grinder, a program I created and made available at http://sourceforge.net/projects/biogrinder. Grinder produces random shotgun libraries from complete genomes in a controlled fashion. The community structure of the genomes is a parameter (e.g. power law rank-abundance curve), and library parameters such as read length, coverage, sequencing error rate allow to produce realistic metagenomes. By creating simulated metagenomes of known composition, Grinder will help ground truth and improve metagenomic techniques. Running GAAS on simulated viral metagenomes showed that the

accuracy of GAAS estimates is higher than that obtained when using the standard BLAST parsing method (keeping the top similarity only, not normalizing by genome length). The benchmarks also demonstrated the applicability of GAAS to microbial metagenomes and to sequences ranging from 50 to 800 bp.

Average genome length in four biomes

To characterize average genome length in aquatic, terrestrial, sediment and host-associated biomes, I conducted a meta-analysis with GAAS using a large set of 175 viral and microbial metagenomes (Figure 4.4), presented in more details in Appendix 3. The average genome length changed significantly in different environments. However, the average genome length of different samples within a biome showed significant variations, suggesting that average genome lengths are not representative at the biome level. The comparison of average genome length of viruses and microorganisms sampled from the same environment at the same time showed that they were independent, likely reflecting how these organisms respond differently to environmental stresses. Also, this suggests that, as opposed to microorganisms, average viral genome length is not correlated with environmental complexity.

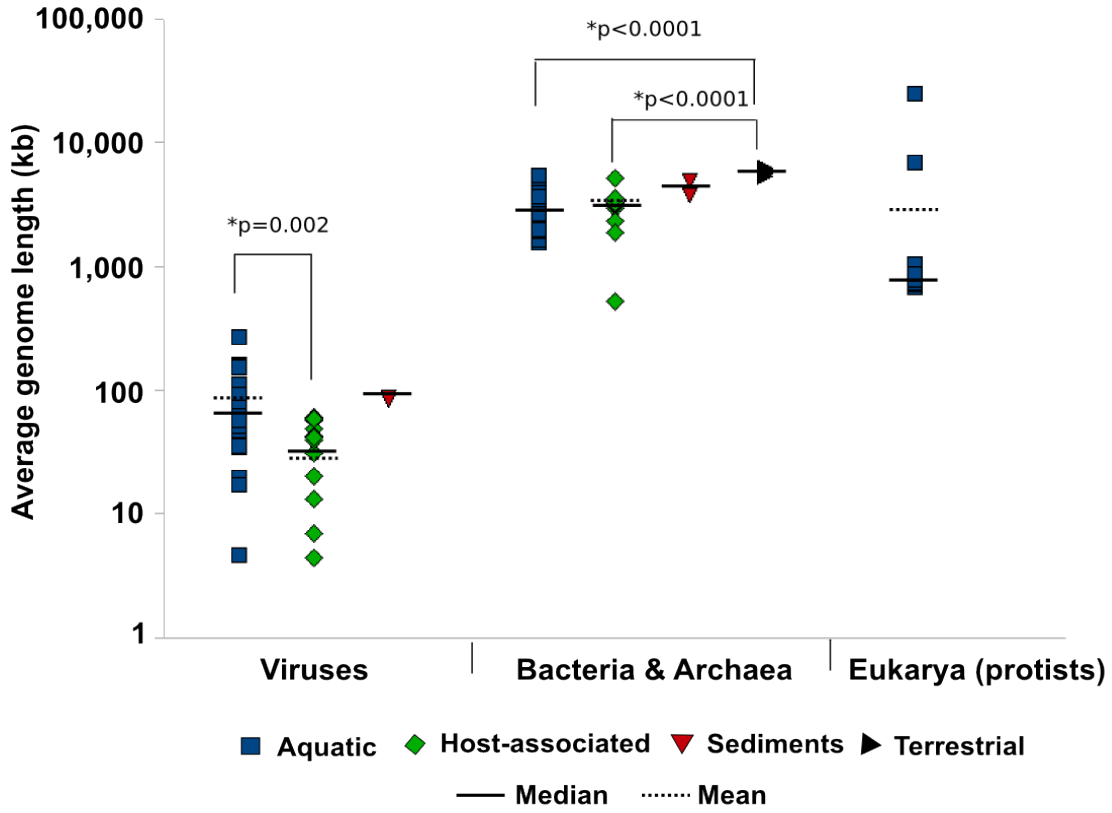


Figure 4.4: The average genome length of viruses, Archaea and Bacteria, and protists in different biomes as estimated by GAAS. Biomes were compared using non-parametric Wilcoxon tests (except for the sediments due to the small number of data points).

CHAPTER 5: A COMPUTATIONAL WORKFLOW FOR ESTIMATING VIRAL DIVERSITY

In the previous chapters, novel computational methods to estimate α-diversity, β-diversity and average genome length from viral metagenomes were discussed. The present chapter describes the synthesis of these different methodologies into a workflow that allows the automated estimation of viral diversity in metagenomes.

Biology and workflows

Biology has entered the age of information and relies heavily on computer programs for mining data and solving problems [219,220]. The computational aspect of biological research is referred to as bioinformatics, which largely consists of developing algorithms and performing *in silico* experiments. With the wealth of computational tools currently available, bioinformaticians can perform increasingly complex experiments which can be used to formulate new research hypotheses.

Bioinformatic experiments often represent a scientific workflow, a set of independent programs used in combination to perform advanced processing of data. A scientific workflow can be as simple as entering data into a website providing a specialized algorithm, copying the output and pasting it into another web-based program. With programming skills, one can use a more automated

approach, installing the software locally and writing a script that runs the programs and passes data between them. Dedicated programs such as Taverna [221] or Kepler [222] make it easier to compose scientific workflows to automate data analysis without programming knowledge.

Diversity workflow overview

The different independent computational elements necessary to estimate α- and β-diversity, CIRCONSPECT, GAAS, PHACCS and MAXIPHI can be integrated in a computational workflow to calculate the diversity of viral communities. To calculate α-diversity, average genome length must first be estimated using GAAS, which eliminates the need to rely on a hypothetical average to input into PHACCS. CIRCONSPECT is used to create contig spectra in an automated fashion from metagenomic sequences. Average genome length and contig spectra are then input to PHACCS, which finally estimates α-diversity (Figure 5.1).

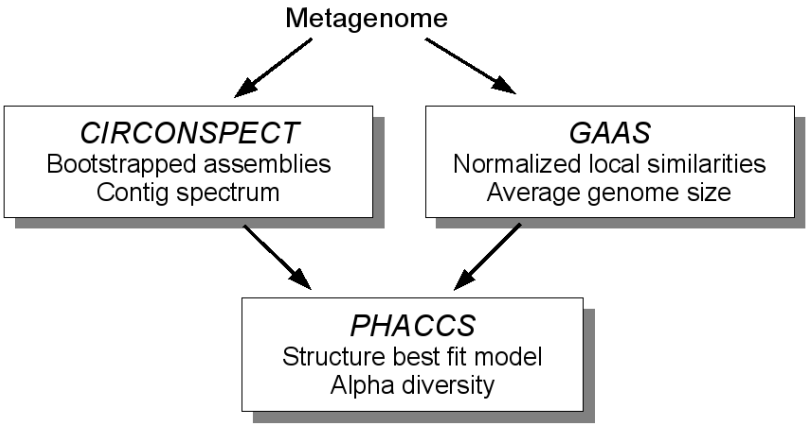


Figure 5.1: Conceptual overview of the α-diversity workflow

In the β-diversity workflow, the α-diversity of several metagenomes is computed. In addition, their cross-contig spectrum is determined by CIRCONSPECT. The community structure of all metagenomes predicted by PHACCS and their cross-contig spectrum are used in MAXIPHI to determine β-diversity (Figure 5.2).

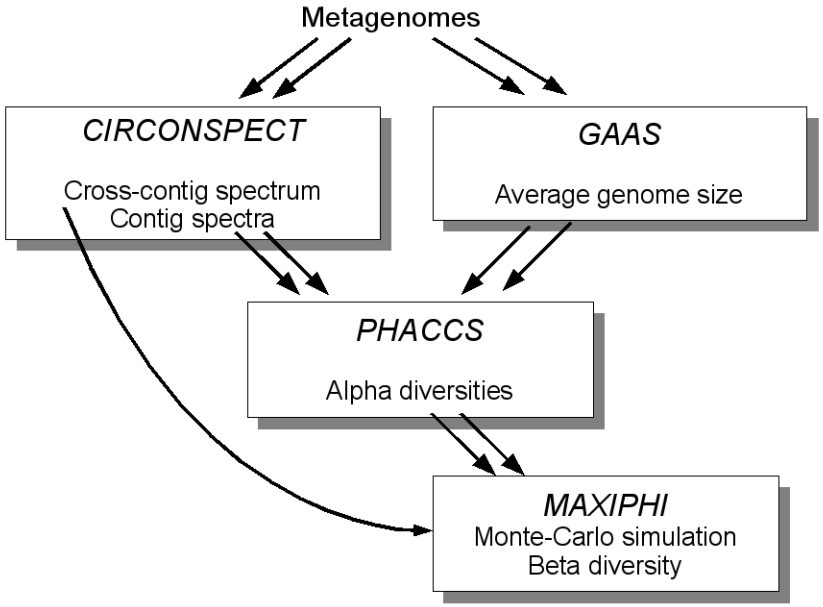


Figure 5.2: Conceptual overview of the β-diversity workflow

Implementation of the α-diversity workflow

CAMERA, the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis [94] is a platform that allows access to metagenomes and tools for their analysis through a web interface. This platform is supported by a 512-CPU cluster, 200 TB of storage, and is able to run BLAST analyses and generate recruitment plots. In version 2.0 (currently in public

preview phase at https://portal.camera.calit2.net/), the CAMERA software stack was reorganized around a Service Oriented Architecture [223,224]. This improvement makes CAMERA more flexible since the computing capacities are dissociated from the hosted software and data. Metagenomics means something different to different investigators and the new design of CAMERA better serves the various needs of the metagenomic community with its implementation of user-designed workflows.

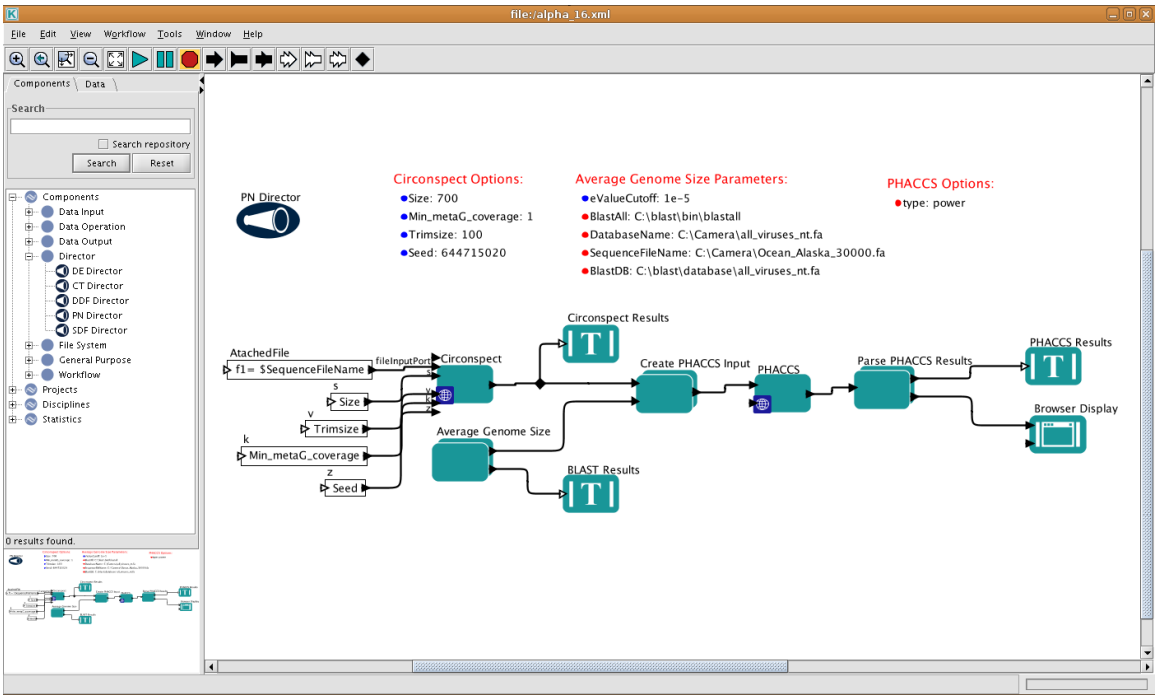


Figure 5.3: The α-diversity workflow implemented using REST web services in Kepler

In collaboration with the CAMERA staff, I composed the α-diversity workflow in Kepler (Figure 5.3), with the individual programs (GAAS, CIRCONSPECT and PHACCS) wrapped as Representational State Transfer (REST) web services

[225] hosted on the CAMERA servers. Integration of the α-diversity workflow in CAMERA now allows investigators to easily estimate the α-diversity of their metagenomes using a web interface (Figure 5.4).

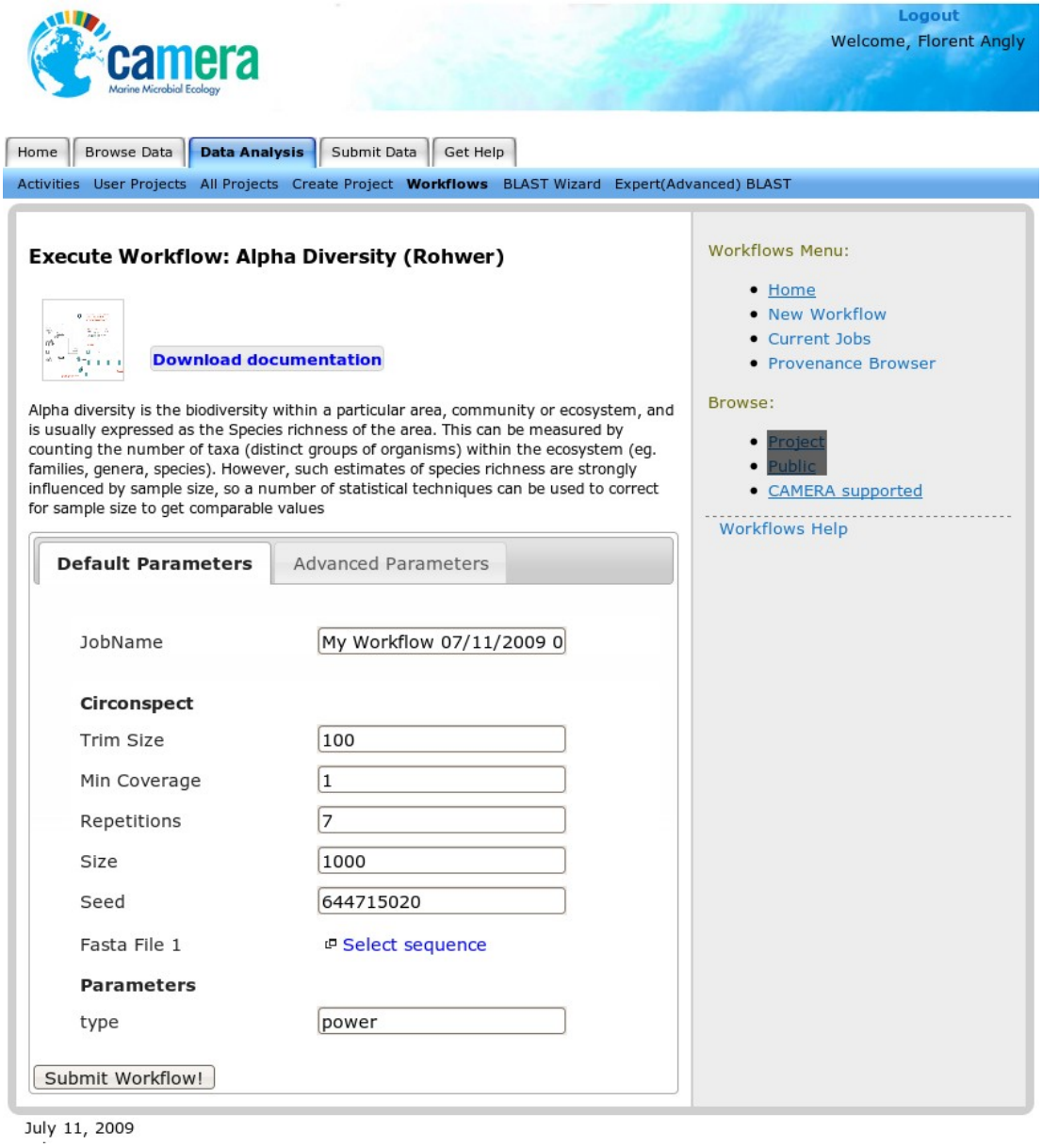


Figure 5.4: The web interface to the α-diversity workflow on CAMERA (https://portal.camera.calit2.net/)

The β-diversity workflow has not been composed yet since MAXIPHI's Monte-Carlo computer intensive methodology must be modified before it can be publicly released and executed on a large scale.

Revisiting previous diversity estimates

Using the diversity workflow, I re-estimated the α-diversity of eight viral metagenomes previously analyzed in Angly et al. 2005 and 2006 [50,102]: Scripps Pier (SP), Mission Bay (MB), Mission Bay Sediments (MBSED), Human Feces (FEC), Arctic Ocean (Arctic), British Columbia (BBC), Sargasso Sea (SAR), and Gulf of Mexico (GOM). My aim was to take advantage of the improvements made to the estimation of viral diversity since 2002 [51] to identify how the diversity estimates changed since their original publication.

These metagenomes had very different characteristics (Table 5.1), with metagenomes containing from 500 to over 700,000 sequences, and an average sequence length ranging from 100 to 700 base pairs. Due to these differences, computation parameters were selected to minimize bias as described below.

Table 5.1: Comparison of the characteristics of the eight viral metagenomes, sequenced by synthetic chain terminator chemistry (Sanger) or Roche 454 GS20 pyrosequencing (Pyro).

Viral metagenome	Biome	Sequencing method	Number of sequences	Mean sequence length (bp)	Total metagenome size (bp)
SP	Aquatic	Sanger	1,064	616.7	656,168
MB	Aquatic	Sanger	873	706.0	616,304
MBSED	Sediments	Sanger	1,156	635.4	734,497
FEC	Host-associated	Sanger	532	710.2	377,851
Arctic	Aquatic	Pyro	688,590	100.2	68,969,258
BBC	Aquatic	Pyro	416,456	103.2	42,976,291
SAR	Aquatic	Pyro	399,343	105.4	42,090,100
GOM	Aquatic	Pyro	263,908	102.6	27,086,439

The average genome length of the viromes was calculated with GAAS using tBLASTx against the NCBI RefSeq complete viral database with a minimum E-value of 10^{-3}. E-value based weights assigned to all significant similarities and genome length normalization were used to further refine BLAST results for average genome length calculation. The minimum relative alignment length was set to 40% and the alignment similarity to 40%, which allowed the recovery of a minimum of approximately 100 similarities for every metagenome. The estimated average genome length differed from the 50 kb originally assumed and ranged from 13.8 kb for the viruses in the Sargasso Sea to 71.8 kb for the viral communities of Scripps Pier (Table 5.2). These results are consistent with

previous reports of a large abundance of viruses with small genomes in the Sargasso Sea [50] and the significant fraction of Myoviridae with large genomes (>170 kb) detected in the Scripps Pier sample [51].

Table 5.2: Estimation of the average genome length of the viruses in the eight communities using GAAS.

Viral metagenome	Number of similarities	Number of similarities per sequence	Estimated average genome length (bp)
SP	300	0.282	71,786.2
MB	208	0.239	61,374.7
MBSED	652	0.564	58,863.4
FEC	91	0.171	28,875.7
Arctic	44,955	0.0653	67,035.0
BBC	35,741	0.0858	35,917.2
SAR	72,021	0.180	13,881.0
GOM	19,786	0.0750	51,994.6

Contig spectra were generated for all metagenomes using CIRCONSPECT. A sample size of 500 random sequences was chosen to accommodate the smallest metagenome analyzed (the fecal sample). Based on their length, sequences were either discarded or trimmed at a random position so that only sequences of 100 bp were assembled, a length slightly smaller than the average sequence length in the metagenomes with the shortest sequences. The assembly parameters for TIGR Assembler were a minimum overlap of 35 bp and minimum similarity of 98%, as in [50,57,58,196]. Random sampling was performed repeatedly until a coverage of 30x of the largest metagenomic library

was achieved, i.e. over 22,000 repetitions. The resulting average contig spectra are reported in Table 5.3. The contig spectrum obtained from the sediment sample had the smallest contig degree, i.e. the largest number of sequences in a contig was 2, as in [53]. Similarly to [50], the Gulf of Mexico sample had a contig degree much larger than the other samples, 49 sequences.

Table 5.3: Average contig spectra of the eight viromes calculated using CIRCONSPECT. All contig spectra were made from 500 sequences of 100 bp.

Viral metagenome	Average contig spectrum
SP	490.2277 4.7137 0.1008 0.0090 0.0013
MB	497.2539 1.3316 0.0269 0.0005
MBSED	499.8509 0.0746
FEC	493.3252 3.2448 0.0609 0.0006
Arctic	496.8570 1.5514 0.0131 0.0002
BBC	493.1876 2.3319 0.4799 0.1213 0.0307 0.0084 0.0022 0.0004 0.0001
SAR	487.9026 4.9393 0.5991 0.0888 0.0113 0.0013 0.0002
GOM	451.9391 4.0380 1.5288 0.9714 0.7159 0.5491 0.4218 0.3292 0.2553 0.2010 0.1566 0.1318 0.1057 0.0850 0.0720 0.0603 0.0521 0.0419 0.0356 0.0315 0.0261 0.0210 0.0189 0.0153 0.0125 0.0100 0.0104 0.0067 0.0064 0.0048 0.0046 0.0026 0.0028 0.0021 0.0019 0.0012 0.0010 0.0007 0.0007 0.0004 0.0004 0.0003 0.0004 0.0003 0.0001 0.0001 0.0002 0.0001 0.0001

The average genome lengths, average contig spectra and minimum contig overlap lengths were used in PHACCS to determine the α-diversity of the eight viral communities. All six rank-abundance models available in PHACCS were tested: power law, exponential, logarithmic, broken stick, niche preemption and

lognormal. The three best overall rank-abundance forms (with the smaller overall error) were, in order, the logarithmic, power law, and lognormal forms. The new estimates of richness, evenness and Shannon-Wiener index using the logarithmic model are presented in Table 5.4.

Table 5.4: Comparison of the α-diversity estimates of the eight viromes obtained using the original method, and the updated computational workflow. The estimates are derived from the logarithmic rank-abundance form, that fitted the different contig spectra overall the best. N/A: PHACCS could not estimate the diversity of these samples.

	Original estimates			New estimates		
Viral metagenome	*Richness*	*Evenness*	*Shannon-Wiener index*	*Richness*	*Evenness*	*Shannon-Wiener index*
SP	3,350	0.932	7.57	113	0.931	4.40
MB	7,180	0.900	7.99	994	0.943	6.51
MBSED	7,340	1.00	8.90	3,700	1.000	8.22
FEC	2,390	0.873	6.80	278	0.972	5.47
Arctic	532	0.964	6.05	257	0.971	5.39
BBC	129,000	0.918	10.8	>500,000	N/A	N/A
SAR	5,140	0.905	7.74	4,280	0.922	7.71
GOM	15,400	0.851	8.21	<1	N/A	N/A

The richest community (British Columbia) remained the richest after reanalysis, with several hundred thousands of genotypes. The previously least diverse community, from the feces sample, became the second least diverse, replaced by the Scripps Pier community (113 genotypes, 4.40 nats), for which an

order of magnitude richness change occurred during reanalysis. The reanalysis of another low diversity community, from the Arctic sample, did not change its status as a low-diversity samples, with an estimated 257 genotypes and 5.49 nats. As in the original analysis the diversity of viruses in sediments (MBSED) was found to be higher that in the water column above (MB).

In Figure 4.1, the logarithmic rank-abundance form was found to be the most sensitive to the average genome length parameter of PHACCS and increased average genome size in the logarithmic model produced increased richness estimates. The changes in diversity due to the reanalysis with the improved diversity workflow (Table 5.4) do not seem to be directly correlated with the changes in the estimated average genome size. Therefore, it is likely that changes in diversity estimates were driven by a combination of providing the average genome size estimated by GAAS, and of using the same random subset size and sequence length for all metagenomes in CIRCONSPECT. The community structure and α-diversity of two metagenomes, from the Sargasso Sea and Gulf of Mexico, could not be precisely determined, for reasons discussed below.

Improving the α-diversity workflow accuracy

Since the original contig spectrum modeling study [51], many methodological advances described in this thesis have been added to the viral diversity estimation methodology so that it is possible to compare metagenomes of a very

different nature. The impossibility of finding a precise community structure for some metagenomes (Table 5.4) suggests that there are still limitations. Possible ways to address these issues and improve the accuracy of viral diversity estimates are described below.

In the reanalysis of the Gulf of Mexico sample, no rank-abundance model could be determined. The contigs assembled from the Gulf of Mexico were larger than in the other samples, and further, the sequence dinucleotide entropy suggests that the Gulf of Mexico sequences had a different composition from the sequences in the other samples (Figure 5.5). I hypothesize that some contigs were formed between sequences containing a low nucleotide complexity, invalidating the assumption that only sequences from the same genotype assemble together (non-chimeric contigs). This issue could be resolved in

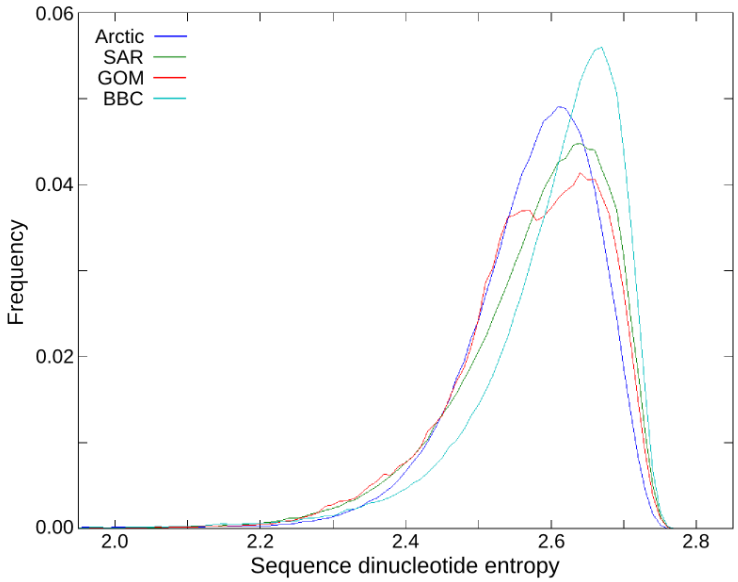


Figure 5.5: Entropy of the sequence dinucleotide frequencies for the four marine viromes.

CIRCONSPECT by filtering out sequences of low complexity either based on their dinucleotide entropy or as calculated by the DUST algorithm [226].

The accurate assembly of metagenomic sequences is critical for obtaining proper contig spectra. All assemblers work differently and no assembler has been specifically designed for metagenomic sequences. The features required for the inclusion of an assembler in CIRCONSPECT are the ability to take a user-specified minimum overlap length and minimum similarity percent. Recent investigations of TIGR Assembler show that it does not strictly respect the values entered by the user for these assembly parameters. Recently developed assemblers such as SR-ASM [195] and Minimus [227] are good candidates for the replacement of TIGR Assembler and inclusion in the diversity workflow.

Regardless of the assembler used, the proper assembly of sequences from single species into contigs depends on the stringency of the assembly parameters. Originally, a manual test on 11 phage genomes determined that a minimum of 98% similarity over at least 20 bp was sufficient to prevent the formation of most chimeric contigs [51] with Sequencher. Today, there are over 2,000 reference viral genomes available from the NCBI [228]. It would be valuable to perform a more systematic analysis of assembly parameters and how changing them affects the number of rightful and chimeric contigs formed. Grinder and CIRCONSPECT implement code that would help in that task. The optimal parameters to produce contig spectra would be the minimum overlap and identity values that minimize the number of chimeric contigs while forming a

sufficiently large number of contigs.

Another limitation of the current implementation of the diversity workflow is that the Community Lander-Waterman model cannot account for contig length. For example, a contig with high coverage (a large number of sequences in a short contig) indicates a genotype with large relative abundance. However, using only a contig spectrum, a high coverage contig is not distinguishable from a longer contig with an identical number of sequences. Improvements in diversity estimation should consider modeling contig coverage [229] and length [230] to complement the contig spectrum model.

The modifications suggested above could improve the accuracy of diversity estimates. Quantifying the gains could be done by assuming viral communities of given community structure and genome sequence (from NCBI RefSeq), and simulating random shotgun libraries (metagenomes). Grinder, the tool I created to benchmark the GAAS method, could be used to groundtruth the α-diversity methodology. Similarly, simulated metagenomes with a varying number of sequences in common could be produced to quantify how accurate the β-diversity estimates are.

Chapter 6: CONCLUSIONS

Many methods to analyze metagenomes are limited by their reliance on similarities to existing sequences. In this thesis, similarity-independent techniques based on metagenomic read assembly were developed to characterize the α and β-diversity of uncultured viral communities.

Innovative methods for characterizing viral diversity

PHACCS is the first software that uses the Community Lander-Waterman equation to model the expected abundance of contigs (contig spectra) based on metagenomic data. PHACCS characterizes the rank-abundance distribution and α-diversity (richness, evenness, Shannon-Wiener index) of uncultured environmental viral communities. In order to process the data easily, it was necessary to automate the creation of contig spectra, a process which was implemented in CIRCONSPECT. CIRCONSPECT was expanded and further developed to produce cross-contig spectra, contigs made of sequences from different metagenomes and used in MAXIPHI. The MAXIPHI method addressed β-diversity, a generally unexplored area of viral metagenomics. β-diversity was characterized by modeling the percentage of shared species and species shifted in abundance between viral communities, as estimated by Monte-Carlo simulations. Finally, to avoid assumptions about average genome length used in the modeling of community structure, I created GAAS, which provides estimates of genome length spectrum and average, based on finding local similarities. The

robustness of the GAAS estimates relies on the accurate determination of the relative abundance of genomes by normalizing for statistical biases such as genome length. By creating a workflow for the calculation of viral diversity on the CAMERA community platform, investigators without bioinformatic expertise can obtain diversity estimates.

Viral diversity modeling started when Sanger sequencing was the standard sequencing technology. High-throughput sequencing appeared in the last few years, bringing larger datasets with different characteristics such as shorter reads and different types of sequencing errors. The computational methodologies developed in this thesis have been updated through several generations of sequencing platforms and viral diversity was calculated from very different metagenomes while avoiding the introduction of potential biases.

Insights into the ecology of viruses

The methods discussed in this thesis have been applied to numerous environmental metagenomes to gain insights into viral diversity and ecology. This work corroborates previous evidence that viruses are the most diverse biological entities on Earth; a richness from 10^2 to 10^5 viral genotypes was reported in the marine habitat [49,50,102]. The β-diversity analysis of marine viruses suggests that oceanic viruses are cosmopolitan, even though they form location-specific assemblages [50]. The calculation of average genome length in different biomes supports these results, by demonstrating large variability in the community

structure of viruses from different marine locations (Appendix 3). This variability between sites may be shaped in part by the existence of a latitudinal gradient of richness for marine viruses [50], suggesting that viruses obey some of the large scale laws that apply to microorganisms and macroorganisms. The soil is an environment reported to harbor a large diversity of micro-organisms [231]. Viral communities in the soil were even richer than in any other biome analyzed, ranging from 10^3 to 10^7 species [48]. From these numbers, the projected world diversity of viruses could be as high as 10^8 species. Insights were obtained into the evolution of viral genomes; the independence between the average length of viral and microbial genomes indicates that identical environmental pressures have different consequences on the evolution and genome length of these organisms (Appendix 3).

Future computational and biological prospects

Estimates of diversity are sensitive to the average genome length parameter. PFGE and GAAS results indicate that the distribution of genome length in a viral community is broad and multimodal. Using an average length to represent a distribution of genome lengths may lead to a loss of precision or cases where the best community structure cannot be estimated. Future efforts to redesign contig spectrum modeling might avoid assuming an average genome length by using a Markov Chain Monte Carlo (MCMC) approach that allows every viral genome to have its own length. Such an MCMC approach is feasible [232], and while it

would be more computationally intensive, it would likely produce more robust estimates of community structure and diversity. Tests of the diversity workflow on microbial metagenomes have been inconclusive so far. This is perhaps because microbial genomes are composed of several replicons, e.g. their main genome and the different plasmids they carry. These microbial replicons differ in length by orders of magnitude and their sequences assemble independently. A MCMC modeling approach that uses a dynamic length for the genome or replicon length may predict the diversity of microorganisms in metagenomes in the same way as for viruses. Taking this MCMC approach further, it would be possible to rescind the assumption that viral communities follow an empirical rank-abundance form and let each replicon take an arbitrary relative abundance. This would resolve the controversial issue of determining what rank-abundance model is best to model environmental phage communities.

The α and β-diversity methods only characterize taxonomic diversity, or the diversity of species. Recently, tools to detect open reading frames in short metagenomic sequences were introduced [233-235]. These tools could extract metagenomic sequences that code for genes and, used as the input for the diversity workflow, these sequences could thus allow calculating the functional diversity, or diversity of genes in metagenomes. Studies suggest that ecosystem stability is correlated more directly with functional diversity than with taxonomic diversity [116,236-238], and the simultaneous estimation of taxonomic and functional diversity will allow to test if it applies to viruses and microorganisms in

addition to macroorganisms. Potentially, metagenomic taxonomic and functional diversity could also help determining the “health” of a particular ecosystem [239,240].

The application of diversity measures to environmental viral samples supports the existence of patterns of diversity for viruses. The richness of viruses in the oceans seems to obey a latitudinal gradient [50] and preliminary estimates of viral diversity in the Line Islands (not shown) are in agreement with the intermediate disturbance theory. Metagenomics has been growing fast and the number of locations from which a viral metagenome is available is steadily increasing (Figure 6.1). This provides the opportunity for applying the workflow of diversity to a larger number of metagenomes and add statistical confidence to our observations of diversity patterns. While many viral metagenomes are from the marine environments, other biomes have been characterized, including host-associated and terrestrial systems. Viral communities from subterranean environments and ambient air have yet to be studied using metagenomics. Sampling these two remaining major biomes, should be a priority to get an accurate picture of the diversity of viruses on Earth and calculate their global richness.

Figure 6.1: Google Maps plot of the location of the viral metagenomes collected so far.

REFERENCES

1. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, et al. (2008) Viral diversity and dynamics in an infant gut. Res Microbiol 159: 367-373.

2. Bergh O, Børsheim KY, Bratbak G, Heldal M (1989) High abundance of viruses found in aquatic environments. Nature 340: 467-468.

3. Hennes KP, Suttle CA (1995) Direct counts of viruses in natural waters and laboratory cultures by epifluorescence microscopy. Limnol Oceanogr 40: 1050-1055.

4. Lemke MJ, Wickstrom CE, Leff LG (1997) A preliminary study on the distribution of viruses and bacteria in lotic habitats. Arch Hydrobiol 141: 67-74.

5. Ashelford KE, Day MJ, Fry JC (2003) Elevated abundance of bacteriophage infecting bacteria in soil. Appl Environ Microbiol 69: 285-289.

6. Cornax R, Moriñigo MA, Gonzalez-Jaen F, Alonso MC, Borrego JJ (1994) Bacteriophages presence in human faeces of healthy subjects and patients with gastrointestinal disturbances. Zentralbl Bakteriol 281: 214-224.

7. Hennes K, Simon M (1995) Significance of bacteriophages for controlling Bacterioplankton growth in a mesotrophic lake. Appl Environ Microbiol 61: 333-340.

8. Ashelford KE, Day MJ, Bailey MJ, Lilley AK, Fry JC (1999) In situ population dynamics of bacterial viruses in a terrestrial environment. Appl Environ Microbiol 65: 169-174.

9. Torrella F, Morita RY (1979) Evidence for a high incidence of bacteriophage particles in the waters of Yaquina Bay, Oregon: ecological and taxonomical implications. Appl Environ Microbiol 37: 774-778.

10. Danovaro R, Manini E, Dell'Anno A (2002) Higher abundance of bacteria than of viruses in deep mediterranean sediments. Appl Environ Microbiol 68: 1468-1472.

11. Rachel R, Bettstetter M, Hedlund BP, Häring M, Kessler A, et al. (2002) Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. Arch Virol 147: 2419-2429.

12. Breitbart M, Wegley L, Leeds S, Schoenfeld T, Rohwer F (2004) Phage community dynamics in hot springs. Appl Environ Microbiol 70: 1633-1640.

13. Kyle JE, Eydal HSC, Ferris FG, Pedersen K (2008) Viruses in granitic groundwater from 69 to 450m depth of the Aspo hard rock laboratory, Sweden. ISME J 2: 571-574.

14. Säwström C, Lisle J, Anesio AM, Priscu JC, Laybourn-Parry J (2008) Bacteriophage in polar inland waters. Extremophiles 12: 167-175.

15. Danovaro R, Serresi M (2000) Viral density and virus-to-bacterium ratio in deep-sea sediments of the Eastern Mediterranean. Appl Environ Microbiol 66: 1857-1861.

16. Kepner RL, Wharton RA, Suttle CA (1998) Viruses in Antarctic lakes. Limnol Oceanogr 43: 1754-1761.

17. Wilson WH, Lane D, Pearce DA, Ellis-Evans JC (2000) Transmission electron microscope analysis of virus-like particles in the freshwater lakes of Signy Island, Antarctica. Polar Biol 23: 657-660.

18. Baylor E, Baylor M, Blanchard D, Syzdek L, Appel C (1977) Virus transfer from surf to wind. Science 198: 575-580.

19. Tringe SG, Zhang T, Liu X, Yu Y, Lee WH, et al. (2008) The airborne metagenome in an indoor urban environment. PLoS ONE 3: e1862.

20. Maranger R, Bird DF (1995) Viral abundance in aquatic systems: a comparison between marine and fresh waters. Mar Ecol Prog Ser 121: 217-226.

21. Noble RT, Fuhrman JA (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquat Microb Ecol 14: 113-118.

22. Marie D, Brussaard CPD, Thyrhaug R, Bratbak G, Vaulot D (1999) Enumeration of marine viruses in culture and natural samples by flow cytometry. Appl Environ Microbiol 65: 45-52.

23. Paul JH, Jiang SC, Rose JB (1991) Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration. Appl Environ Microbiol 57: 2197-2204.

24. Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic

ecosystems. Microbiol Mol Biol Rev 64: 69-114.

25. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. Proc Nat Acad Sci USA 95: 6578-6583.

26. Niagro FD, Forsthoefel AN, Lawther RP, Kamalanathan L, Ritchie BW, et al. (1998) Beak and feather disease virus and porcine circovirus genomes: intermediates between the geminiviruses and plant circoviruses. Arch Virol 143: 1723-1744.

27. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, et al. (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455: 100-104.

28. Azam F, Fenchel T, Field JG, Gray JS, Meyer-Reil LA, et al. (1983) The ecological role of water-column microbes in the sea. Mar Ecol Prog Ser 10: 257-263.

29. Pomeroy LR (1974) The ocean's food web, a changing paradigm. BioScience 24: 499-504.

30. Bratbak G, Thingstad F, Heldal M (1994) Viruses and the microbial loop. Microb Ecol 28: 209-221.

31. Murray AG, Eldridge PM (1994) Marine viral ecology: incorporation of bacteriophage into the microbial planktonic food web paradigm. J Plankton Res 16: 627-641.

32. Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399: 541-8.

33. Kirchman DL (2000) Microbial ecology of the oceans. 1st ed. Wiley-Liss.

34. Zepp RG, Sonntag C (1995) The role of nonliving organic matter in the earth's carbon cycle. John Wiley and Sons.

35. Pernthaler J (2005) Predation on prokaryotes in the water column and its ecological implications. Nat Rev Micro 3: 537-546.

36. Thingstad TF, Lignell R (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquat Microb Ecol 13: 19-27.

37. Fuhrman JA, Schwalbach M (2003) Viral influence on aquatic bacterial communities. Biol Bull 204: 192-195.

38. Weitz JS, Hartman H, Levin SA (2005) Coevolutionary arms races between bacteria and bacteriophage. Proc Nat Acad Sci USA 102: 9535-9540.

39. Van Valen L (1973) A new evolutionary law. Evol Theor 1: 1-30.

40. Lively CM (1996) Host-parasite coevolution and sex. BioScience 46: 107-114.

41. Dybdahl MF, Storfer A (2003) Parasite local adaptation: Red Queen versus Suicide King. Trend Ecol Evol 18: 523-530.

42. Rohwer F, Seguritan V, Choi DH, Segall AM, Azam F (2001) Production of shotgun libraries using random amplification. Biotechniques 31: 108-12,114-6,118.

43. Handelsman J, Rondon MR, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245-R249.

44. Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 38: 525-552.

45. Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet 6: 805-814.

46. Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39: 321-46.

47. Ackermann H, DuBow M (1987) Volume I: General properties of bacteriophages. Viruses of prokaryotes. CRC Press, Inc, Vol. 1. p. 202.

48. Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, et al. (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of Bacteria, Archaea, Fungi, and viruses in soil. Appl Environ Microbiol 73: 7059-7066.

49. Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, et al. (2008) Microbial ecology of four coral atolls in the northern Line Islands. PLoS ONE 3: e1584.

50. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. PLoS Biology 4: e368.

51. Breitbart M, Salamon P, Andresen B, Mahaffy J, Segall A, et al. (2002) Genomic analysis of uncultured marine viral communities. Proc Nat Acad Sci USA 99: 14250-14255.

52. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185: 6220-6223.

53. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, et al. (2004) Diversity and population structure of a near-shore marine-sediment viral community. Proc Biol Sci 271: 565–574.

54. Vega Thurber RL, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, et al. (2008) Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. Proc Nat Acad Sci USA 105: 18413-8.

55. Culley AI, Lang AS, Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. Science 312: 1795-1798.

56. Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, et al. (2007) Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol 73: 7629-7641.

57. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, et al. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452: 340-343.

58. Marhaver KL, Edwards RA, Rohwer F (2008) Viral communities associated with healthy and bleaching corals. Environ Microbiol 10: 2277-2286.

59. Leroy M, Prigent M, Dutertre M, Confalonieri F, Dubow M (2008) Bacteriophage morphotype and genome diversity in Seine River sediment. Freshwater Biol 53: 1176-1185.

60. Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, et al. (2008) Assembly of viral metagenomes from Yellowstone hot springs. Appl Environ Microbiol 74: 4164-4174.

61. Djikeng A, Kuzmickas R, Anderson NG, Spiro DJ (2009) Metagenomic analysis of RNA viruses in a fresh water lake. PLoS ONE 4: e7264.

62. Woyke T, Xie G, Copeland A, González JM, Han C, et al. (2009) Assembling

the marine metagenome, one cell at a time. PLoS ONE. 4: e5299.

63. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.

64. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554-557.

65. Sogin ML, Elwood HJ, Gunderson JH (1986) Evolutionary diversity of eukaryotic small-subunit rRNA genes. Proc Nat Acad Sci USA 83: 1383-1387.

66. Cottrell MT, Waidner LA, Yu L, Kirchman DL (2005) Bacterial diversity of metagenomic and PCR libraries from the Delaware River. Environmental Microbiology 7: 1883-1895.

67. Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, et al. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7: 57.

68. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.

69. Wegley L, Edwards R, Rodriguez-Brito B, Liu H, Rohwer F (2007) Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. Environmental Microbiology 9: 2707-2719.

70. Martín HG, Ivanova N, Kunin V, Warnecke F, Barry KW, et al. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotech 24: 1263-1269.

71. Legault B, Lopez-Lopez A, Alba-Casado J, Doolittle WF, Bolhuis H, et al. (2006) Environmental genomics of *Haloquadratum walsbyi* in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. BMC Genomics 7: 171.

72. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, et al. (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 55: 205–211.

73. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006)

An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-131.

74. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, et al. (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443: 950-955.

75. Yokouchi H, Fukuoka Y, Mukoyama D, Calugay R, Takeyama H, et al. (2006) Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using Φ29 polymerase. Environmental Microbiology 8: 1155-1163.

76. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5: e77.

77. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, et al. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450: 560-565.

78. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480-484.

79. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.

80. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

81. Bennett S (2004) Solexa Ltd. Pharmacogenomics 5: 433-438.

82. Fu Y, Peckham HE, McLaughlin SF, Ni JN, Rhodes MD, et al. (2008) SOLiD system sequencing and 2 base encoding. Cold Spring Harbor Laboratory, NY.

83. Benson D, Boguski M, Lipman D, Ostell J, Ouellette B, et al. (1999) GenBank. Nucleic Acids Research 27: 12-17.

84. Ferrer M, Martínez-Abarca F, Golyshin PN (2005) Mining genomes and metagenomes for novel catalysts. Curr Opin Biotech 16: 588-593.

85. Lorenz P, Eck J (2005) Metagenomics and industrial applications. Nat Rev

Micro 3: 510-516.

86. Henne A, Daniel R, Schmitz RA, Gottschalk G (1999) Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4- hydroxybutyrate. Appl Environ Microbiol 65: 3901-3907.

87. Knietsch A, Bowien S, Whited G, Gottschalk G, Daniel R (2003) Identification and characterization of coenzyme B12-dependent glycerol dehydratased and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures. Appl Environ Microbiol 69: 3048-3060.

88. Eisen JA, Fraser CM (2003) Phylogenomics: intersection of evolution and genomics. Science 300: 1706-1707.

89. Quaiser A, Ochsenreiter T, Klenk H, Kletzin A, Treusch AH, et al. (2002) First insight into the genome of an uncultivated crenarchaeote from soil. Environ Microbiol 4: 613-611.

90. Hallam SJ, Konstantinidis KT, Putnam N, Schleper C, Watanabe Y, et al. (2006) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. Proc Nat Acad Sci USA 103: 18296-18301.

91. Fauth JE, Bernardo J, Camara M, Resetarits WJ, Van Buskirk J, et al. (1996) Simplifying the jargon of community ecology: a conceptual approach. Am Nat 147: 282-286.

92. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-10.

93. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

94. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. PLoS Biol 5: e75.

95. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, et al. (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36: 534-538.

96. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of

metagenomic data. Genome Res 17: 377-86.

97. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, et al. (2008) Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res 36: 2230-9.

98. Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52: 540-542.

99. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Nat Acad Sci USA 87: 2264-2268.

100. Hugenholtz P, Tyson GW (2008) Microbiology: metagenomics. Nature 455: 481-483.

101. Willner D, Thurber RV, Rohwer F (2009) Metagenomic signatures of 86 microbial and viral metagenomes. Environ Microbiol 11: 1752-1766.

102. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6: 41.

103. Costanza R, Folke C (1997) Valuing ecosystem services with efficiency, fairness, and sustainability as goals. Nature's services. p. 392.

104. Chapin III FS, Zavaleta ES, Eviner VT, Naylor RL, Vitousek PM, et al. (2000) Consequences of changing biodiversity. Nature 405: 234-242.

105. Whittaker RH (1960) Vegetation of the Siskiyou Mountains, Oregon and California. Ecol Monogr 30: 279-338.

106. Whittaker RH (1972) Evolution and measurement of species diversity. Taxon 21: 213-251.

107. Vane-Wright RI, Humphries CJ, Williams PH (1991) What to protect? Systematics and the agony of choice. Biol Conserv 55: 235-254.

108. Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61: 1-10.

109. Margalef R (1957) The theory of information in ecology. General Systems Bulletin 3: 36-71.

110. Shannon CE, Weaver W (1963) The mathematical theory of communication. Urbana: Univ of Illinois Press.

111. Pielou EC (1996) The measurement of diversity in different types of biological collections. J Theor Biol 13: 131-144.

112. Simpson EH (1949) Measurement of diversity. Nature 163: 688.

113. Berger WH, Parker FL (1970) Diversity of planktonic foraminifera in deep-sea sediments. Science 168: 1345–134.

114. Cohan FM (2002) What are bacterial species? Annu Rev Microbiol 56: 457-487.

115. Hey J (2001) The mind of the species problem. Trends Ecol Evol 16: 326-329.

116. Tilman D, Knops J, Wedin D, Reich P, Ritchie M, et al. (1997) The influence of functional diversity and composition on ecosystem processes. Science 277: 1300-1302.

117. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33: 5691-5702.

118. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629-632.

119. Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Micro 3: 504-510.

120. Hoffman KH, Rodriguez-Brito B, Breitbart M, Bangor D, Angly F, et al. (2005) The structure of marine phage populations. Proceedings of ECOS 2005.

121. Whittaker RH (1965) Dominance and diversity in land plant communities: Numerical relations of species express the importance of competition in community function and evolution. Science 147: 250-260.

122. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509-512.

123. Ulrich W (2001) Models of relative abundance distributions I: model fitting by stochastic models. Pol J Ecol 49: 145-157.

124. Pielou EC (1975) Ecological diversity. New York: Wiley.

125. MacArthur RH (1957) On the relative abundance of bird species. Proc Nat Acad Sci USA 43: 293-295.

126. Sugihara G (1980) Minimal community structure: an explanation of species abundance patterns. Am Nat 116: 770-787.

127. Hubbell SP (2001) The unified neutral theory of biodiversity and biogeography. Princeton University Press.

128. Leigh EG (2007) Neutral theory: a historical perspective. J Evolution Biol 20: 2075-2091.

129. Tokeshi M (1993) Species abundance patterns and assemblage structure. Adv Ecol Res 24: 111-186.

130. Preston FW (1948) The commonness and rarity of species. Ecol 29: 254-283.

131. Drozd P, Novotny V (n.d.) PowerNiche: Niche division models for community analysis. Ceske Budejovice, Czech Republic.

132. Etienne RS, Olff H (2005) Confronting different models of community structure to species-abundance data: a Bayesian model comparison. Ecol Lett 8: 493-504.

133. Shmida A, Wilson MV (1985) Biological determinants of species diversity. J Biogeogr 12: 1-20.

134. Humboldt AV (1860) Ansichten der Natur. J. G. Cotta.

135. Darwin C (1839) Journal of researches into the natural history and geology of the countries visited during the voyage of H. M. S. *Beagle* round the world, under the command of Capt. Fitz Roy, R. N. London: Henry Colburn.

136. Halpern BS, Walbridge S, Selkoe KA, Kappel CV, Micheli F, et al. (2008) A global map of human impact on marine ecosystems. Science 319: 948.

137. Connell JH (1978) Diversity in tropical rain forests and coral reefs. Science 199: 1302-1310.

138. Grigg RW, Maragos JE (1974) Recolonization of hermatypic corals on

submerged lava flows in Hawaii. Ecology 55: 387.

139. Grime JP (1973) Competitive exclusion in herbaceous vegetation. Nature 242: 344-347.

140. Connor EF, McCoy ED (1979) The statistics and biology of the species-area relationship. Am Nat 113: 791.

141. Arrhenius O (1920) Distribution of the species over the area. Medd fr K Vet Akad Nobelinstitut 4.

142. Hawkins BA (2001) Ecology's oldest pattern? Trends Ecol Evol 16: 470.

143. Hillebrand H (2004) On the generality of the latitudinal diversity gradient. Am Nat 163: 192-211.

144. Gaston KJ, Blackburn TM (2000) Pattern and processes in macroecology. Oxford: Blackwell Scientific.

145. Pianka ER (1966) Latitudinal gradients in species diversity: A review of concepts. Am Nat 100: 33.

146. Cardillo M, Orme CDL, Owens IPF (2005) Testing for latitudinal bias in diversification rates: an example using New World birds. Ecology 86: 2278-2287.

147. Fisher MM, Triplett EW (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. Appl Environ Microbiol 65: 4630-6.

148. Liu WT, Marsh TL, Cheng H, Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl Environ Microbiol 63: 4516-4522.

149. Schwartz DC, Cantor CR (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell 37: 67-75.

150. Fischer SG, Lerman LS (1979) Length-independent separation of DNA restriction fragments in two-dimensional gel electrophoresis. Cell 16: 191-200.

151. Thatcher DR, Hodson B (1981) Denaturation of proteins and nucleic acids by thermal-gradient electrophoresis. Biochem J 197: 105–109.

152. Pommier T, Canback B, Riemann L, Boström KH, Simu K, et al. (2007) Global patterns of diversity and community structure in marine bacterioplankton. Mol Ecol 16: 867-880.

153. Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, et al. (2008) A latitudinal diversity gradient in planktonic marine bacteria. Proc Nat Acad Sci USA 105: 7774-7778.

154. Hendrix RW (1999) Evolution: the long evolutionary reach of viruses. Curr Biol 9: R914-R917.

155. Rohwer F, Edwards RA (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol 184: 4529-35.

156. Filée J, Tétart F, Suttle CA, Krisch HM (2005) Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. Proc Nat Acad Sci USA 102: 12471-12476.

157. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, et al. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Nat Acad Sci USA 82: 6955-6959.

158. Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221–271.

159. Olsen GJ, Larsen N, Woese CR (1991) The ribosomal RNA Database project. Nucleic Acids Res 19: 2017–2021.

160. Amann R, Ludwig W, Schleifer K (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59: 143-169.

161. Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, et al. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet 4: e1000255.

162. Fox G, Stackebrandt E, Hespell R, Gibson J, Maniloff J, et al. (1980) The phylogeny of prokaryotes. Science 209: 457-463.

163. Le Marrec C, van Sinderen D, Walsh L, Stanley E, Vlegels E, et al. (1997) Two groups of bacteriophages infecting *Streptococcus thermophilus* can be distinguished on the basis of mode of packaging and genetic determinants for major structural proteins. Appl Environ Microbiol 63: 3246-3253.

164. Fuller NJ, Wilson WH, Joint IR, Mann NH (1998) Occurrence of a sequence

in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. Appl Environ Microbiol 64: 2051-2060.

165. Tétart F, Desplats C, Kutateladze M, Monod C, Ackermann H, et al. (2001) Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. J Bacteriol 183: 358-366.

166. Proux C, van Sinderen D, Suarez J, García P, Ladero V, et al. (2002) The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. J Bacteriol 184: 6026-6036.

167. Filée J, Forterre P, Sen-Lin T, Laurent J (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. J Mol Evol 54: 763-773.

168. Weinbauer MG (2004) Ecology of prokaryotic viruses. FEMS Microbiol Rev 28: 127-181.

169. Short SM, Suttle CA (1999) Use of the polymerase chain reaction and denaturing gradient gel electrophoresis to study diversity in natural virus communities. Hydrobiologia 401: 19-33.

170. Short SM, Suttle CA (2002) Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. Appl Environ Microbiol 68: 1290-1296.

171. Steward GF, Montiel JL, Azam F (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. Limnol Oceanogr 45: 1697-1706.

172. Sandaa R, Larsen A (2006) Seasonal variations in virus-host populations in Norwegian coastal waters: focusing on the cyanophage community infecting marine *Synechococcus* spp. Appl Environ Microbiol 72: 4610-4618.

173. Wommack KE, Ravel J, Hill RT, Chun J, Colwell RR (1999) Population dynamics of Chesapeake bay virioplankton: total-community analysis by pulsed-field gel electrophoresis. Appl Environ Microbiol 65: 231-40.

174. Staden R (1982) An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. Nucleic Acids Res 10: 2951-2961.

175. Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, et al. (1984) DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature

310: 207-211.

176. Gene Codes (n.d.) Sequencher. Ann Arbour, MI.

177. Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2: 231-9.

178. Hewson I, Winget DM, Williamson KE, Fuhrman JA, Wommack KE (2006) Viral and bacterial assemblage covariance in oligotrophic waters of the West Florida shelf (Gulf of Mexico). J Mar Biol Assoc UK 86: 591-603.

179. Hewson I, Vargo GA, Fuhrman JA (2003) Bacterial diversity in shallow oligotrophic marine benthos and overlying waters: effects of virus infection, containment, and nutrient enrichment. Microb Ecol 46: 322-336.

180. Mai V, Morris JGJ (2004) Colonic bacterial flora: changing understandings in the molecular age. J Nutr 134: 459-464.

181. Sørensen TA (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on Danish commons. K Dan Vidensk Selsk Biol Skr 5: 1-34.

182. Bray JR, Curtis JT (1957) An ordination of the upland forest communities of Southern Wisconsin. Ecol Monogr 27: 326-349.

183. Horn HS (1966) Measurement of "overlap" in comparative ecological studies. Am Nat 100: 419-424.

184. McKnight MW, White PS, McDonald RI, Lamoreux JF, Sechrest W, et al. (2007) Putting beta-diversity on the map: broad-scale congruence and coincidence in the extremes. PLoS Biol 5: e272.

185. Rohwer F (2003) Global phage diversity. Cell 113: 141.

186. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, et al. (2003) Origins of highly mosaic mycobacteriophage genomes. Cell 113: 171-182.

187. Baylor ER, Baylor MB (1980) Surf-to-wind transfer of viruses. Ann NY Acad Sci 353: 201-208.

188. Martin C (1988) The application of bacteriophage tracer techniques in South West Water. Water Environ J 2: 638-642.

189. Epstein HT (2000) The properties of bacteriophages. Advances in virus research. Academic Press.

190. Moebus K (1983) Lytic and inhibition responses to bacteriophages among marine bacteria, with special reference to the origin of phage-host systems. Helgol Mar Res 36: 375-391.

191. Breitbart M, Miyake JH, Rohwer F (2004) Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol Lett 236: 249-256.

192. Casas V, Jon Miyake, Heather Balsley, Julie Roark, Serena Telles, et al. (2006) Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California. FEMS Microbiol Lett 261: 141-149.

193. Lozupone C, Hamady M, Knight R (2006) UniFrac - An online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics 7: 371.

194. Sutton GG, White O, Adams MD, Kervalage AR (1995) TIGR Assembler: a new tool for assembling large shotgun sequencing projects. Genome Science & Technology 1: 9-19.

195. Blazewicz J, Bryja M, Figlerowicz M, Gawron P, Kasprzak M, et al. (2009) Whole genome assembly from 454 sequencing output via modified DNA graph concept. Comput Biol Chem 33: 224-230.

196. McDaniel L, Breitbart M, Mobberley J, Long A, Haynes M, et al. (2008) Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. PLoS ONE 3: e3263.

197. Studdert MJ (1993) Circoviridae: new viruses of pigs, parrots and chickens. Aust Vet J 70: 121-122.

198. Klieve AV, Swain RA (1993) Estimation of ruminal bacteriophage numbers by pulsed-field gel electrophoresis and laser densitometry. Appl Environ Microbiol 59: 2299–2303.

199. Sandaa R, Foss Skjoldal E, Bratbak G (2003) Virioplankton community structure along a salinity gradient in a solar saltern. Extremophiles 7: 347-351.

200. Filippini M, Middelboe M (2007) Viral abundance and genome size

distribution in the sediment and water column of marine and freshwater ecosystems. FEMS Microbiol Ecol 60: 397-410.

201. Otawa K, Lee S, Yamazoe A, Onuki M, Satoh H, et al. (2007) Abundance, diversity, and dynamics of viruses on microorganisms in activated sludge processes. Microb Ecol 53: 143-152.

202. Wu Q, Liu W (2009) Determination of virus abundance, diversity and distribution in a municipal wastewater treatment plant. Water Res 43: 1101-1109.

203. Fuhrman JA, Griffith JF, Schwalbach MS (2002) Prokaryotic and viral diversity patterns in marine plankton. Ecol Res 17: 183-194.

204. Holmfeldt K, Middelboe M, Nybroe O, Riemann L (2007) Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their Flavobacterium hosts. Appl Environ Microbiol 73: 6730-6739.

205. Graves LM, Swaminathan B (2001) PulseNet standardized protocol for subtyping *Listeria monocytogenes* by macrorestriction and pulsed-field gel electrophoresis. Int J Food Microbiol 65: 55-62.

206. Raes J, Korbel J, Lercher M, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. Genome Biol 8: R10.

207. Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, et al. (2009) The genomic basis of trophic strategy in marine bacteria. Proc Nat Acad Sci USA 106: 15527-15533.

208. Gregory TR, DeSalle R (2005) Comparative genomics in prokaryotes. The Evolution of the Genome. San Diego: Elsevier. pp. 585-675.

209. Ranea JAG, Buchan DWA, Thornton JM, Orengo CA (2004) Evolution of protein superfamilies and bacterial genome size. J Mol Biol 336: 871-87.

210. Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. Annu Rev Genet 38: 771-92.

211. Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Nat Acad Sci USA 101: 3160-3165.

212. Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of

change and exchange. J Mol Evol 44: 383-397.

213. Sadreyev R, Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 326: 317-336.

214. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS ONE 4: e4345.

215. Altschul SF, Gish W (1996) Local alignment statistics. Methods Enzymol 266: 460-80.

216. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, et al. (2006) Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J Proteome Res 5: 2339-2347.

217. Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, et al. (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. Methods 40: 303-311.

218. Paoletti AC, Parmely TJ, Tomomori-Sato C, Sato S, Zhu D, et al. (2006) Quantitative proteomic analysis of distinct mammalian mediator complexes using normalized spectral abundance factors. Proc Nat Acad Sci USA 103: 18928-18933.

219. Kanehisa M, Bork P (2003) Bioinformatics in the post-sequence era. Nat Genet 33 Suppl: 305-310.

220. Nakai K, Vert J (2002) Genome informatics for data-driven biology. Genome Biol 3: reports4010.1-reports4010.3.

221. Oinn T, Addis M, Ferris J, Marvin D, Senger M, et al. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20: 3045-3054.

222. Altintas I, Berkley C, Jaeger E, Jones M, Ludäscher B, et al. (2004) Kepler: an extensible system for design and execution of scientific workflows. Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM'04). IEEE Computer Society.

223. MacKenzie CM, Laskey K, McCabe F, Brown P, Metz R (2006) Reference model for Service Oriented Architecture. OASIS.

224. Brown A, Johnston S, Kelly K (2002) Using Service-Oriented Architecture and component-based development to build web service applications. Rational Software Corporation.

225. Fielding RT (2000) Architectural styles and the design of network-based software architectures University of California, Irvine.

226. Morgulis A, Gertz EM, Schaffer AA, Agarwala R (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol 13: 1028-40.

227. Sommer D, Delcher A, Salzberg S, Pop M (2007) Minimus: a fast, lightweight genome assembler. BMC Bioinformatics 8: 64.

228. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61-65.

229. Cohan F, Krizanc D, Lu Y (2007) Estimating bacterial diversity from environmental DNA: a maximum likelihood approach. The 2007 International Symposium on Bioinformatics Research and Applications (ISBRA 2007). Atlanta, GA, USA: Springer. p. 653.

230. Koerbitz P (2006) Modeling of phage communities. San Diego State University.

231. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, et al. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 66: 2541-2547.

232. Domes JW, Druken BK, Salamon P (2007) PHACCS III: a tool for predicting environmental characteristics of bacteriophage. San Diego State University Mathematics Research Experience for Undergraduates and Teachers.

233. Krause L, Diaz NN, Bartels D, Edwards RA, Puhler A, et al. (2006) Finding novel genes in bacterial communities isolated from the environment. Bioinformatics 22: e281-289.

234. Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15: 387-396.

235. Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: predicting genes in metagenomic sequencing reads. Nucl Acids Res 37: W101–W105.

236. Wardle DA, Bonner KI, Barker GM, Yeates GW, Nicholson KS, et al. (1999) Plant removals in perennial grassland: vegetation dynamics, decomposers, soil biodiversity and ecosystem properties. Ecol Monogr 69: 535-568.

237. Huston MA (1997) Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. Oecologia 110: 449-460.

238. Hooper DU, Vitousek PM (1997) The effects of plant composition and diversity on ecosystem processes. Science 277: 1302-1305.

239. Visser S, Parkinson D (1992) Soil biological criteria as indicators of soil quality: soil microorganisms. Am J Alternative Agr 7: 33-37.

240. Xu F, Tao S, Dawson RW, Li P, Cao J (2001) Lake ecosystem health assessment: indicators and methods. Water Res 35: 3157-3167.

APPENDICES

Appendix 1: PHACCS

Appendix 2: MAXIPHI

Appendix 3: GAAS

Appendix 1: PHACCS

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, and Forest Rohwer.

BMC Bioinformatics 6, no. 41 (March 2). 2005.

BMC Bioinformatics

BioMed Central

Software

Open Access

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

Florent Angly[1,2], Beltran Rodriguez-Brito[2,4], David Bangor[2,3], Pat McNairnie[2], Mya Breitbart[2], Peter Salamon[3], Ben Felts[3], James Nulton[3], Joseph Mahaffy[3] and Forest Rohwer*[2,5]

Address: [1]Ecole Supérieure de Biotechnologie de Strasbourg, Boulevard Sébastien Brandt, 67413 Illkirch, France, [2]Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA, [3]Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA, [4]Computational Science Research Center, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA and [5]Center For Microbial Sciences, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA

Email: Florent Angly - fangly@projects.sdsu.edu; Beltran Rodriguez-Brito - brodrigu@rohan.sdsu.edu; David Bangor - heimdalle@yahoo.com; Pat McNairnie - arcum@cox.net; Mya Breitbart - mya@sunstroke.sdsu.edu; Peter Salamon - salamon@math.sdsu.edu; Ben Felts - ben.felts@conexant.com; James Nulton - jnulton@mail.sdsu.edu; Joseph Mahaffy - mahaffy@math.sdsu.edu; Forest Rohwer* - forest@sunstroke.sdsu.edu

* Corresponding author

Published: 02 March 2005

BMC Bioinformatics 2005, **6**:41 doi:10.1186/1471-2105-6-41

Received: 18 December 2004
Accepted: 02 March 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/41

Abstract

Background: Phages, viruses that infect prokaryotes, are the most abundant microbes in the world. A major limitation to studying these viruses is the difficulty of cultivating the appropriate prokaryotic hosts. One way around this limitation is to directly clone and sequence shotgun libraries of uncultured viral communities (i.e., metagenomic analyses). PHACCS http://phage.sdsu.edu/phaccs, Phage Communities from Contig Spectrum, is an online bioinformatic tool to assess the biodiversity of uncultured viral communities. PHACCS uses the contig spectrum from shotgun DNA sequence assemblies to mathematically model the structure of viral communities and make predictions about diversity.

Results: PHACCS builds models of possible community structure using a modified Lander-Waterman algorithm to predict the underlying contig spectrum. 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. This model is the basis for making estimates of uncultured viral community richness, evenness, diversity index and abundance of the most abundant genotype.

Conclusion: PHACCS analysis of four different environmental phage communities suggests that the power law is an important rank-abundance form to describe uncultured viral community structure. The estimates support the fact that the four phage communities were extremely diverse and that phage community biodiversity and structure may be correlated with that of their hosts.

Background

Most environmental viruses are phages (a.k.a., bacteriophages) that infect prokaryotic cells, both Bacteria and Archaea. On average there are about ten phage particles per host cell [1]. Extrapolations from the number of prokaryotes [2] make phages the most abundant biological entities in the biosphere with an estimated 10^{31} viral particles. By killing prokaryotes, phages can strongly impact microbial community biomass [3] and structure [4]. Despite their importance, very little is known about phage biodiversity.

Traditionally, the study of environmental phage diversity, dynamics, and ecology requires growing prokaryotes on microbiology plates and infecting them with phages. However this standard technique is limited by the fact that only a small fraction of environmental microbes are readily cultured [5] and that each phage species generally only has a very narrow number of possible microbial hosts [6]. In addition, even if it is possible to observe phages with an electron microscope, pictures are not sufficient to identify species because of the low taxonomic resolution of viral morphology. Cultivating and observing phages do not permit to assess environmental phage diversity.

Biodiversity is composed of richness, or total number of different species [7], and evenness, expressing the relative abundance of each species [8]. The Shannon-Wiener index quantifies diversity as a single term combining richness and evenness [9]. A high richness and high evenness together represent a high level of diversity.

A new approach to accessing natural microbial diversity is through the creation of shotgun sequence libraries from environmental metagenomes (sum of all genomes) [10-14], so that the genetic information of each genotype of the community is recorded, qualitatively (sequence) and quantitatively (abundance of each sequence). The community is analyzed by sequencing a part of the library. The metagenomic data used here is the contig spectrum, determined by assembly of environmental random shotgun DNA fragments. The contig spectrum is a vector containing the number of contigs (groups of overlapping sequences) of size q (number of sequences in the group) [10]. The stringency of the assembly parameters can be varied so that only sequences belonging to the same genotype overlap. Thus, for one genotype, the bigger the contigs in the contig spectrum, the higher the number of copies and the more abundant this genotype. Based on this, the contig spectrum provides important information about the abundance and diversity of genotypes within a community.

In this work, we present PHACCS (PHAge Communities from Contig Spectrum), an online computational tool to assess the diversity and structure of environmental viral communities from the contig spectrum of shotgun sequence data. The PHACCS program and its predictions are first described and then used to analyze four environmental viral communities.

Implementation

Platform and software

The standalone core mathematics for PHACCS consists of Matlab (MathWorks Inc., Natick, MA.) scripts that are partly based on the previous works [10-12]. A CGI (Common Gateway Interface) script written in PERL (Practical Extraction and Report Language) is used to input and output data from and to an HTML (Hyper Text Markup Language) interface. PHACCS was developed and tested on a Linux-based (2.6.6 kernel) personal computer running PERL 5.8.3 (with CGI module), Matlab 6.5.0, and Apache 2.0.50 web server.

Obtaining a contig spectrum

The input for PHACCS is the contig spectrum, a vector containing the number of q-contigs (groups of q overlapping sequences) from the *in silico* assembly of random shotgun DNA fragments. Detailed information about the way to get viral metagenomes and their contig spectrum can be found in [10-12]. Briefly, viral communities were isolated via tangential flow filtration and cesium chloride centrifugation, and their DNA was extracted. The DNA was randomly fragmented, used to create a linker amplified shotgun library [15] and clones were sequenced (between 500 and 1200 for studies [10-12]). The sequence assembly program Sequencher (Gene Codes Corp., Ann Arbor, MI.) was used to assemble phage sequences having at least 98% identity on at least 20 bp [10]. The stringency of the assembly parameters was experimentally determined so that only fragments belonging to the same genotype assemble together. Closely related phage genomes (e.g., coliphages T3 and T7) can be discriminated using these parameters [10]. The number of contigs of each size was then recorded to generate the contig spectrum. The number of sequences in the largest contig defines the contig spectrum degree.

Modified Lander-Waterman algorithm

PHACCS uses a modified version of the Lander-Waterman algorithm [16] to predict a contig spectrum from assumed population parameters. The original Lander-Waterman algorithm is a way of predicting the contig spectrum of a randomly fragmented genome (e.g., a single viral species) given: i) the length L of the genome, ii) the number N of DNA fragments studied, iii) the average size s of these fragments, and iv) the minimum overlap length o for the sequence assembly [16]. Given this data, the predicted values of the following quantities are calculated:

• Probability p of an overlap: $p = 1 - e^{-Nx/L}$ with $x = s - o$

• Probability w_q for a fragment to be part of a q-contig (overlap of q fragments):

$w_q = qp^{q-1}(1-p)^2$

• Expected number of fragments c_q that are part of a q-contig: $c_q = Nw_q$

• Contig spectrum: $[c_1 \frac{c_2}{2} \frac{c_3}{2} ...]$

The modified Lander-Waterman algorithm is a generalization of the original algorithm to a group of M different genotypes (e.g., a whole viral community) [10]. The predicted contig spectrum can be calculated as the sum of the contig spectra for each individual genotype i.

• Expected number of fragments c_q part of a q-contig:

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

w_{qi} is the probability for a fragment to be part of a q-contig for the genotype i and n_i is the expected number of fragments for the genotype i.

In this modified algorithm, since there are several genotypes, an assumption about their underlying distribution within the community in terms of abundance has to be made.

Relative rank-abundance forms

PHACCS offers six basic functional forms of relative rank-abundance for biological populations: the power law, logarithmic, exponential, broken stick, niche preemption, and lognormal distributions.

The first three functional forms are empirical models that were designed to describe an asymptotic drop-off in the abundance [17]:

• Power: $n_i = ai^{-b}$ for $1 \le i \le M$

• Logarithmic: $n_i = a(\log(i+1))^{-b}$ for $1 \le i \le M$

• Exponential: $n_i = ae^{-ib}$ for $1 \le i \le M$

The parameter a represents the abundance of the most abundant genotype, b is a parameter related to the evenness, and M is the number of different genotypes in the community.

Two ecological models are based on a partitioning of resources between species [18,19]:

• Broken stick: $n_i = \frac{N}{M} \sum_{x=i}^{M} \frac{1}{x}$ for $1 \le i \le M$

• Niche preemption: $n_i = Nk(1-k)^{i-1}$ and $n_M = N(1-k)^{M-1}$ for $1 \le i \le M - 1$

The broken stick function has only one parameter, M, and assumes a random distribution of resources, whereas in the niche preemption function, each species takes only a fraction k of the remaining resources in the environment.

The sixth functional form is the lognormal distribution. It is the most commonly used species distribution, with numerous theoretical justifications in the literature [20,21]. The relationship is specified as species density versus abundance and needs to be transformed to give a rank-abundance relationship. Our rank-abundance form was obtained by dividing the area under the normal distribution with standard deviation σ into M equal area slices and associating an abundance n_i with the i-th slice by calculating an average value for the abundance within the slice. The result is:

• Lognormal: $n_i = \frac{e^{m_i \sigma}}{\sum_{j=1}^{M} e^{m_j \sigma}}$ with $m_i = \frac{M}{\sqrt{2\pi}}(e^{-t_i^2/2} - e^{-t_{i+1}^2/2})$,

$t_1 = -\infty$, $t_{i+1} = \sqrt{2} erf^{-1}\left(\frac{2}{M} + erf(\frac{t_i}{\sqrt{2}})\right)$ and $t_{M+1} = +\infty$ for $1 \le i \le M$

where erf is the error function and erf^{-1} its inverse.

Modeling the viral community structure

The PHACCS algorithm is represented in Figure 1. The experimentally determined contig spectrum of a sample and the other parameters needed for the modified Lander-Waterman algorithm are the input. For a given rank-abundance function, assumed values of the function parameters (number of different genotypes, as well as b for the power law, logarithmic, exponential and lognormal distributions and k for niche preemption) are used to predict a contig spectrum using the modified Lander-Waterman algorithm. To determine the model fitness, the error between the actual and the predicted contig spectrum is calculated as the variance-weighted sum of squared deviations, L being the contig spectrum vector length and c_q' the experimental number of fragments that belong to a q-contig:

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

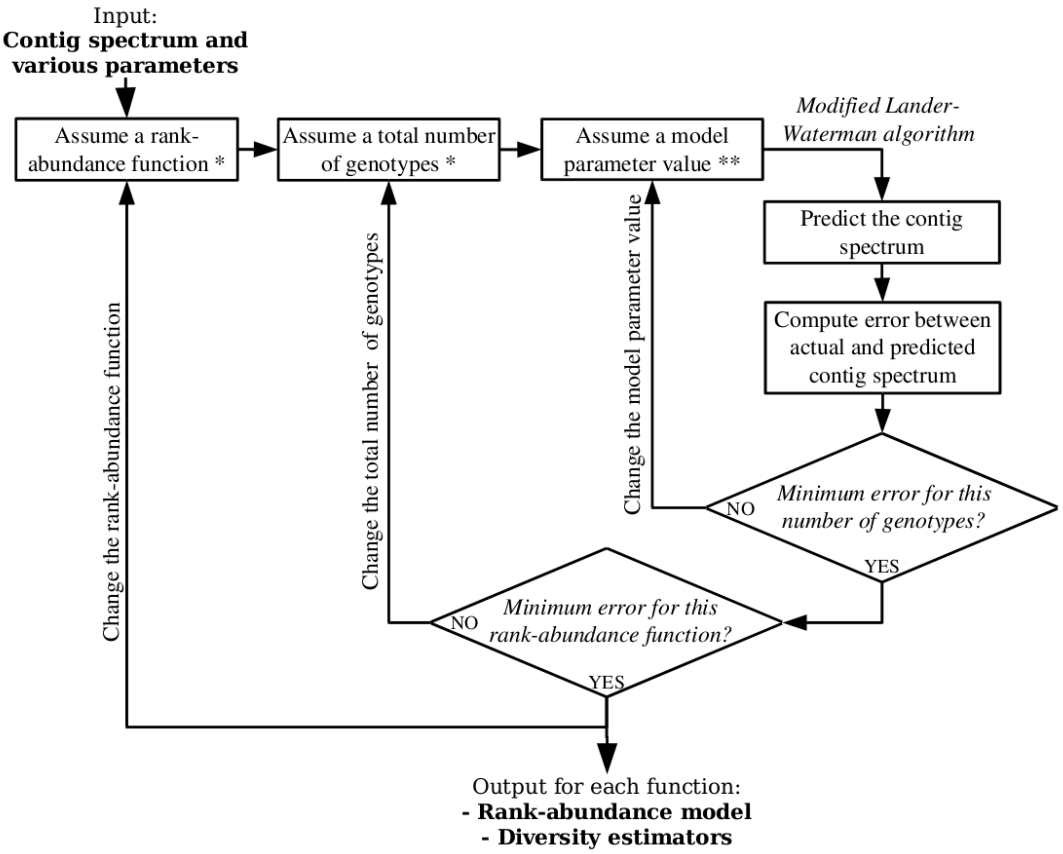


Figure 1
Flowchart of PHACCS. *The rank-abundance functions and the range of genotypes to use can be defined by the user. **This parameter represents b for the power law, logarithmic, lognormal and exponential distributions and k for the niche preemption. This parameter is not applicable to the broken stick.

The best descriptive model for a community structure is defined as the one with the smallest error. For each rank-abundance function tested, the global minimum for the error is found by optimizing the value of the function parameters.

The values of the error can be roughly interpreted as logarithms of odds ratios of the observed contigs being seen from community distributions of the specified forms. Thus a value of 0.1 for the difference in errors between two models corresponds to an odds ratio of $e^{0.1}$ which is about 11:10 between the two models. This means that the model with the smallest error is about 10% more likely to give rise to the observed data.

Predicting the viral community diversity

For each rank-abundance form, the best model is used by PHACCS to assess diversity. The richness *S* is estimated as equal to the number of different genotypes *M* found in the community structure model. The abundance of the most abundant genotype is also directly determined from the model as the highest rank-abundance value. The

Table 1: Test data used for the study of the phage communities with PHACCS [10-12]. The average fragment sequence length was determined using Sequencher after sequence trimming (maximum one ambiguity on 99 bp at each extremity). A 98% identity for a minimal overlap length of 20 bp was used for sequence assembly with Sequencher to obtain the contig spectra. The average genome length was determined by pulse field gel electrophoresis [12, 22].

Community	SP (Scripps Pier)	MB (Mission Bay)	MBSED (Mission Bay Sediments)	FEC (Fecal)
Contig spectrum *	1021 17 2 0 ...	841 13 2 0 ...	1152 2 0 ...	482 18 2 2 0 ...
Avg. community genome size	50 kb	50 kb	50 kb	30 kb
Avg. shotgun fragment length	663 bp	663 bp	570 bp	699 bp

* The number of trailing zeros was set to 10 for each contig spectrum.

Shannon-Wiener index, which is a measure for diversity, is calculated using the relative rank-abundance values $r_i = n_i/N$ of all individual genotypes i [9]:

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

The evenness is derived from H' [18]:

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

Comparison of four phage communities

As a case study, four viral metagenomes obtained from previous studies and 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 [10]. The two other samples are sediments from Mission Bay (MBSED) [11] and human feces (FEC) [12]. A compilation of the data for these samples is presented in Table 1. These four datasets were analyzed with PHACCS using all six rank-abundance models.

Results

Best abundance forms

The errors obtained from the contig spectrum analysis of the different samples are presented in Table 2. For each sample the best descriptive model of the community structure is the one with the smallest error. The SP community was best described by using the power law (error of 1.84), closely followed by the lognormal (error of 1.93) and logarithmic (error of 2.57) distributions. The exponential and niche preemption distributions had poor fits, with errors of 12.0. The MB community modeling gave qualitatively the same results. Power law was the best fit with an error of 2.15 and exponential and niche preemption were last with an error of 16.2. The FEC community also had the same sequence of best fitting rank-abundance forms. The best model was given by using the power law form (error 9.79). Exponential and niche preemption did a poor job of explaining the data, coming in last with an error of 60.0. For the MBSED community, the power law, lognormal, logarithmic and exponential distributions all tied for the best fit (with an error of 0.0104), whereas broken stick gave the worst fit (error of 0.0157).

Phage community diversity and structure

The different diversity indicators and the rank-abundance curves obtained by using the best descriptive model for each sample are summarized in Figure 2. The MBSED community was the richest with an estimated 7340 different phage genotypes. MB had ~7180 different genotypes, SP ~3350, and FEC was the least rich sample with ~2390 different genotypes. MBSED was the most even community with the maximum possible evenness of 1.00 (flat rank-abundance curve), followed by SP (evenness of 0.932), MB (evenness of 0.900), and FEC (evenness of 0.873). The most abundant genotype represented 4.80% of the total community for FEC, 2.63% for MB, 2.03% for SP and around 0.01% for MBSED. Based on the Shannon-Wiener diversity index, MBSED was overall the most diverse community with 8.90 nats, then MB (7.99 nats), SP (7.57 nats), and finally FEC (6.80 nats), the least diverse community.

Discussion

Using PHACCS

PHACCS is publicly accessible at http://phage.sdsu.edu/phaccs and the source code is freely available [see Additional file 1]. The biological information PHACCS needs as an input is the viral community's contig spectrum, average genome size, average shotgun DNA sequence length, and the minimum overlap length used for the assembly. PHACCS has two HTML interfaces. The basic interface assumes default values for marine phage communities (average genome size of 50 kb, average fragment length of 650 bp and minimum overlap of 20 bp). All rank-abundance forms (power law, expoential, logarithmic, lognormal, broken stick and niche preemption distributions) are tested for up to 100,000 genotypes. In the advances inter-

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

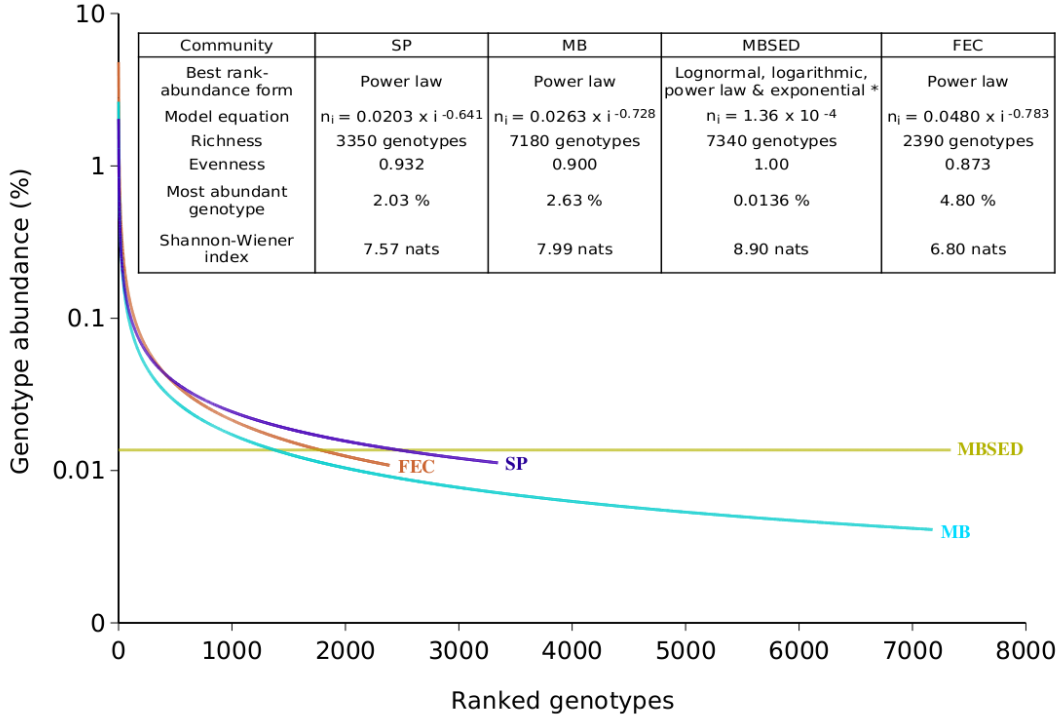


Figure 2
Comparison of the structure and diversity of the different viral communities using PHACCS. The graphics represent rank-abundance curves, where the abundance of each genotype is plotted versus its abundance rank, the genotype of rank one being the most abundant. The curves were obtained by plotting 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.

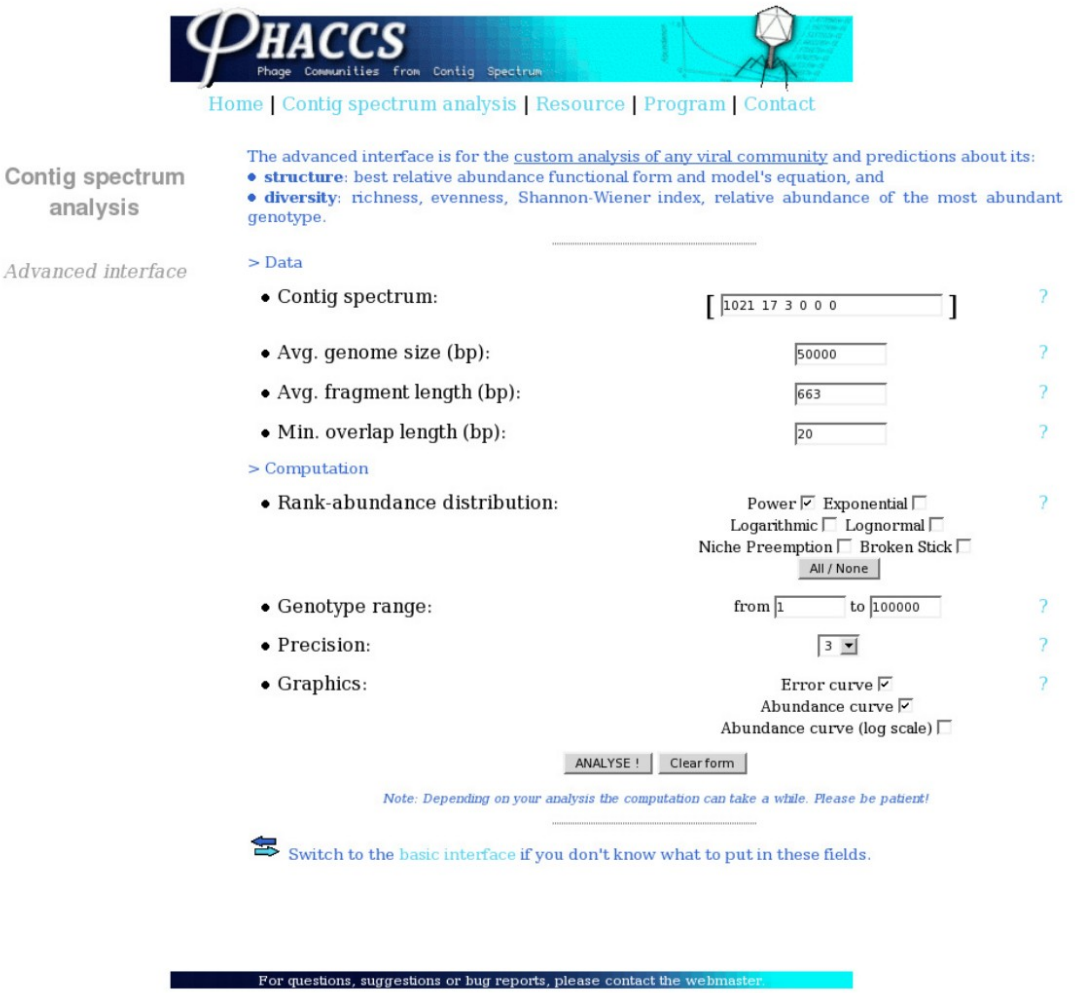


Figure 3
Screenshot of PHACCS' advanced web interface.

face (Figure 3) the user can change all biological and computational parameters.

PHACCS analyses are computer intensive. On a dual-Opteron™ server, the computation for the SP sample takes ~5 minutes. The broken stick and lognormal rank-abundance forms account for most of the computation time (data not shown). Increasing the range of genotypes to search dramatically increases the time needed to complete the analysis (data not shown).

PHACCS estimations about the virus community are: i) structure – best descriptive rank-abundance form, model equation and error, and ii) diversity – richness, evenness,

• **Project home page:** http://phage.sdsu.edu/phaccs

• **Operating system(s):** Unix based system for PHACCS and its web interface. Platform independent for PHACCS core.

• **Programming language:** Matlab (for the core scripts) and Perl

• **Other requirements:** For the interface: CGI.pm Perl module, ppmtogif, webserver program (to use PHACCS as a web service)

• **License:** GNU GPL

Authors' contributions

FA developed the PHACCS main program and its interface. BRB helped with the programming. BRB, DB, PMN, PS, BF, JN and JM developed the modified Lander-Waterman algorithm and implemented it with Matlab. FR and MB helped write the manuscript and provided the test datasets. All authors read and approved the final manuscript.

Additional material

Additional File 1

This file contains the script files part of PHACCS. These files are either standard text or picture files.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-6-41-S1.zip]

Acknowledgements

This work was supported by NSF 0316518 (FR) and an EPA STAR fellowship (MB). We thank Ines Thiele, Steve Rayhawk and Cynthia Steiner for useful suggestions and comments.

References

1. Wommack KE, Colwell RR: **Virioplankton: Viruses in aquatic ecosystems.** *Microbiol Mol Biol Rev* 2000, **64:**69-114.
2. Whitman WB, Coleman DC, Wiebe WJ: **Prokaryotes: The unseen majority.** *Proc Natl Acad Sci USA* 1998, **95:**6578-6583.
3. Wilcox RM, Fuhrman JA: **Bacterial viruses in coastal seawater: lytic rather than lysogenic production.** *Mar Ecol Prog Ser* 1994, **114:**35-45.
4. Thingstad TF: **Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems.** *Limnol Oceanogr* 2000, **45:**1320-1328.
5. Staley JT, Konopka A: **Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats.** *Annu Rev Microbiol* 1985, **39:**346.
6. Ackermann H-W, DuBow MS: **Viruses of prokaryotes; Volume I: General properties of bacteriophages.** CRC Press, Inc; 1987:202.
7. McIntosh RI: **An index of diversity and the relation of certain concepts to diversity.** *Ecology* 1967, **48:**392-404.
8. Hill MO: **Diversity and evenness: a unifying notation and its consequences.** *Ecology* 1973, **54:**427-431.
9. Shannon CE, Weaver W: **The mathematical theory of communication.** Univ of Illinois Press, Urbana; 1963.
10. 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.
11. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, Rohwer F: **Diversity and population structure of a near-shore marine sediment community.** *Proc R Soc Lond* 2004, **271:**565-574.
12. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F: **Metagenomic analysis of an uncultured viral community from human feces.** *J Bacteriol* 2003, **185:**6620-6223.
13. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Bram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428:**37-43.
14. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.
15. Rohwer F: **Construction and Analyses of Linker-Amplified Shotgun Libraries (LASLs).** [http://www.sci.sdsu.edu/PHAGE/LASL/index.htm].
16. Lander ES, Waterman MS: **Genomic mapping by fingerprinting random clones: a mathematical analysis.** *Genomics* 1988, **2:**231-239.
17. Ulrich W: **Models of relative abundance distributions I: Model fitting by stochastic models.** *Pol J Ecol* 2001, **49:**145-157.
18. Pielou EC: *Ecological Diversity* New York: John Wiley & Sons, Inc; 1975.
19. McArthur RH: **On the relative abundance of bird species.** *Proc Nat Acad Sci USA* 1957, **43:**293-295.
20. Sugihara G: **Minimal community structure: an explanation of species abundance patterns.** *Am Nat* 1980, **116:**770-187.
21. McGill BJ: **A test of the unified neutral theory of biodiversity.** *Nature* 2003, **422:**881-885.
22. Steward GF, Montiel JL, Azam F: **Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments.** *Limnol Oceanogr* 2000, **45:**1697-1706.
23. Hoffmann KH, Rodriguez-Brito B, Breitbart M, Bangor D, Angly F, Felts B, Nulton J, Rohwer F, Salamon P: **Power law rank-abundance relationships in marine phage populations.** in press.
24. Hewson I, Vargo GA, Fuhrman JA: **Bacterial diversity in shallow oligotrophic marine benthos and overlying waters: effects of virus infection, containment, and nutrient enrichment.** *Microb Ecol* 2003, **46:**322-36.
25. Mai V, Morris JG Jr: **Colonic Bacterial flora: changing understandings in the molecular age.** *J Nutr* 2004, **134:**459-464.

Appendix 2: MAXIPHI

The Marine Viromes of Four Oceanic Regions.

Florent E Angly, Ben Felts, Mya Breitbart, Peter Salamon, Robert A. Edwards, Craig Carlson, Amy M. Chan, Matthew Haynes, Scott Kelley, Hong Liu, Joseph M. Mahaffy, Jennifer E. Mueller, Jim Nulton, Robert Olson, Rachel Parsons, Steve Rayhawk, Curtis A. Suttle, and Forest Rohwer.

PLoS Biology 4, no. 11 (November 1). 2006.

The Marine Viromes of Four Oceanic Regions

Florent E. Angly[1,2], Ben Felts[2,3], Mya Breitbart[1¤], Peter Salamon[2,3], Robert A. Edwards[1,2,4,5], Craig Carlson[6], Amy M. Chan[7], Matthew Haynes[1], Scott Kelley[1,4], Hong Liu[1], Joseph M. Mahaffy[2,3], Jennifer E. Mueller[1], Jim Nulton[2,3], Robert Olson[8], Rachel Parsons[9], Steve Rayhawk[1,2], Curtis A. Suttle[7,10,11], Forest Rohwer[1,4*]

1 Department of Biology, San Diego State University, San Diego, California, United States of America, **2** Computational Science Research Center, San Diego State University, San Diego, Califomia, United States of America, **3** Department of Mathematics, San Diego State University, San Diego, California, United States of America, **4** Center for Microbial Sciences, San Diego State University, San Diego, California, United States of America, **5** Fellowship for Interpretation of Genomes, Burr Ridge, Illinois, United States of America, **6** University of California Santa Barbara, Santa Barbara, California, United States of America, **7** Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, British Columbia, Canada, **8** Argonne National Laboratory, Argonne, Illinois, United States of America, **9** Bermuda Biological Station for Research, St. George's, Bermuda, **10** Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada, **11** Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

Viruses are the most common biological entities in the marine environment. There has not been a global survey of these viruses, and consequently, it is not known what types of viruses are in Earth's oceans or how they are distributed. Metagenomic analyses of 184 viral assemblages collected over a decade and representing 68 sites in four major oceanic regions showed that most of the viral sequences were not similar to those in the current databases. There was a distinct "marine-ness" quality to the viral assemblages. Global diversity was very high, presumably several hundred thousand of species, and regional richness varied on a North-South latitudinal gradient. The marine regions had different assemblages of viruses. Cyanophages and a newly discovered clade of single-stranded DNA phages dominated the Sargasso Sea sample, whereas prophage-like sequences were most common in the Arctic. However most viral species were found to be widespread. With a majority of shared species between oceanic regions, most of the differences between viral assemblages seemed to be explained by variation in the occurrence of the most common viral species and not by exclusion of different viral genomes. These results support the idea that viruses are widely dispersed and that local environmental conditions enrich for certain viral types through selective pressure.

Citation: Angly F, Felts B, Breitbart M, Salamon P, Edwards R, et al. (2006) The marine viromes of four oceanic regions. PLoS Biol 4(11): e368. DOI: 10.1371/journal.pbio.0040368

Introduction

Most marine viruses are phages (bacteriophages) that kill the heterotrophic and autotrophic microbes (both Bacteria and presumably Archaea) that dominate the world's oceans [1]. Phages and the other major microbial predator guild, nanoflagellates, control the numbers of marine microbes to a concentration of about $\sim 5 \times 10^5$ cells per ml of surface seawater [2,3].

Phages affect microbial evolution by inserting themselves into genomes as prophages. Prophages often account for most of the difference between strains of the same microbial species [4], and they can dramatically change the phenotype of the hosts via lysogenic conversion. For example, many nonpathogens and pathogens only differ by prophages that encode exotoxin genes [5]. Phages also affect microbial evolution by moving genes from host to host. It has been hypothesized that most of the orphan open reading frames (ORFans) in microbial genomes are actually of phage origin [6]. Phages may also affect microbial evolution by killing specific microbes. Various Lotka-Volterra models, called "kill-the-winner," predict that as one microbial strain becomes dominant, its viral predator kills it and leaves open a niche that can be used by a related strain that is resistant to the phage [7,8]. This model may explain the enormous microdiversity observed in microbial communities [9].

The advent of whole-community genome sequencing (i.e., metagenomics) is rapidly changing the way viral and microbial diversity are assayed. Using this approach, it is possible to rapidly characterize the metabolic diversity and community structure of any microbial ecosystem [10–19]. We studied the marine viral metagenome (virome) of four oceanic regions. The viromes were obtained by pyrosequencing uncultured viral assemblages that were integrated over 4,600 km in distance, 3,000 m in depth, and over a decade in time in order to characterize them and identify patterns of viral distribution and diversity.

Materials and Methods

Samples and Sequencing

Samples were collected from four oceanic regions (Figure 1). Briefly, the viral samples were concentrated on tangential flow filters (30–100-kD cutoff), distributed into 50-ml tubes and stored at 4 °C in the dark. A single sample was collected from the Sargasso Sea (labeled SAR) on 30 June 2005.

Academic Editor: Nancy A. Moran, University of Arizona, United States of America

Received April 8, 2006; **Accepted** September 5, 2006; **Published** November 7, 2006

DOI: 10.1371/journal.pbio.0040368

Abbreviations: bp, base pair; PTP, permutation tail probability; ssDNA, single-stranded DNA

* To whom correspondence should be addressed. E-mail: forest@sunstroke.sdsu.edu

¤ Current address: University of South Florida, St. Petersburg, Florida, United States of America

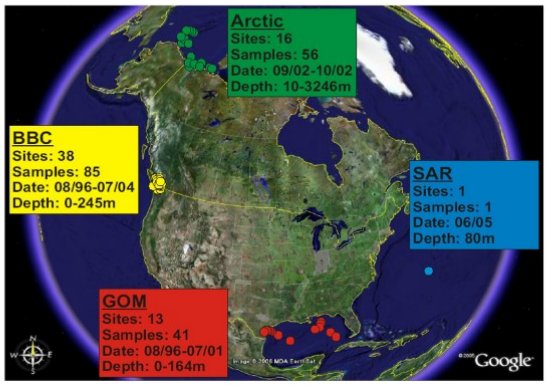


Figure 1. Sampling Sites

The circles represent the sampling locations in the Sargasso Sea (SAR), Gulf of Mexico (GOM), British Columbia (BBC), and the Arctic Ocean. The number of samples taken at each location and combined for sequencing, as well as the date and depth range, are shown in the boxes.

DOI: 10.1371/journal.pbio.0040368.g001

Chloroform was added to this sample to stop microbial growth. Integrative samples, representing multiple sites and times, were assembled from the Gulf of Mexico (labeled GOM; 13 sites; 42 individual samples), the British Columbia coastal waters (labeled BBC; 38 sites; 85 individual samples), and the Arctic Ocean (labeled Arctic; 16 sites; 56 individual samples). These samples represent the combined viral assemblages of four oceanic regions over approximately one decade (sample details are described in Protocol S1).

Viral particles were purified using a combination of filtration and density-dependent centrifugation ([10]; http://scums.sdsu.edu/isolation.html, accessed 15 September 2006). The cesium chloride gradient was designed to recover virions with densities from 1.35 g ml^{-1} to 1.5 g ml^{-1}. Viral DNA was isolated by a formamide/CTAB extraction [20], and the resulting DNA was amplified with Genomiphi and sequenced using pyrophosphate sequencing (454 Life Sciences, Branford, Connecticut, United States) [21] (see Protocol S1 for details on the technology). Each Genomiphi reaction started with 100–150 ng of DNA, above the 10 ng recommended by the manufacturer. A total of 181,044,179 base pairs (bp) of DNA sequence data was generated from the four libraries (SAR, 42 Mbp; GOM, 27 Mbp; BBC, 43 Mbp; and Arctic, 69 Mbp). The difference in library size was due to differences in number of successful reads during the pyrosequencing. The 1,768,297 sequences had an average length of 102 bp. The GOM, BBC, Arctic, and SAR metagenomes are deposited on the SDSU Center for Universal Microbe Sequencing website at (http://scums.sdsu.edu/phage/Oceans, accessed 15 September 2006).

Bioinformatics

The metagenome sequences from each of the libraries were compared to the SEED nonredundant database and environmental database using BLASTX [22]. The SEED includes the GenBank database supplemented with other complete and draft genome sequences. The environmental database consists of the microbial assemblages from the Iron Mountain acid mine drainage [16], Sargasso Sea [17], whale fall [18], and Minnesota farm soil [18]. All large-scale computational analyses were performed on the Terraport and National Microbial Pathogen Data Resource cluster at Argonne National Laboratory. Individual analyses were performed on a 12-node Orion desktop cluster (Orion, Santa Clara, California, United States).

These comparisons were supplemented with more extensive TBLASTN and with TBLASTX comparisons [22] of either selected portions of the data against the complete nonredundant database or the whole library compared to boutique databases. The same cutoff E value was always used for the same database and BLAST search method. In addition, the sequences were compared to the phage and prophage sequences from 510 genomes of the phage genome database (RA Edwards, unpublished data). A FASTA file of these genomes is at http://scums.sdsu.edu/phage/Oceans.

Taxonomic Composition of the Metagenomes

In an approach similar to previous work [10–12], the best similarity for each metagenomic sequence was automatically parsed and assigned as "known" if there was a significant similarity ($E \leq 10^{-5}$) to a sequence from the nonredundant nucleotide database, else "environmental" for a significant similarity to any environmental database sequence, and else "unknown" (if there was no significant similarity to any database). The number of similarities in each group was then counted (Figure 2A). These numbers were also averaged for the four samples. In a second step, the sequences from the "known" group were classified as viral, bacterial, archaeal, or eukaryotic based on their highest similarity (Figure 2B). To assess the contribution of the prophages (often similar to bacterial sequences), TBLASTX was used to compare the sequences against the complete phage genome sequences. Any significant similarity in the previous four taxonomic groups that was also similar to a prophage sequence was assigned to the prophage group instead. The prophage sequences for these analyses were extracted from complete microbial genomes. A complete list is available at the supporting website (http://scums.sdsu.edu/phage/Oceans). The average of these numbers for the four samples was also calculated.

Assembly and Verification of Single-Stranded DNA, the chp1-Like Microphage from the Sargasso Sea

The single-stranded DNA (ssDNA) chp1-like microphage was partially assembled from all of the sequences that had significant TBLASTX similarities ($E \leq 10^{-5}$). The assembly parameters were a minimal match percentage of 85% and a 20-bp minimum overlap using Sequencher 4.0 (Gene Codes, Ann Arbor, Michigan, United States). These sequences alone did not result in the assembly of a complete genome due to areas with low similarity to known chp1-like microphage. To complete the assembly, batches of sequences from the Sargasso Sea sample were added to these assemblies until complete coverage was obtained (the consensus sequence is in Protocol S1). The PCR primers SARssDNAF (5′ TGC GGA GAA TAT GGT GAT GA 3′), SARssDNAR1 (5′ CGG TTA TTA CGC CTG TCG TT 3′), and SARssDNAR2 (5′ CCA TGG TAG GGC AGA GGT AA 3′) were designed based on the consensus sequence. A PCR was run against the original Sargasso Sea sample DNA. The reaction mixture (50 μl total volume) contained target DNA, 1 mM of each primer, and 1X FideliTaq master mix (USB, Cleveland,

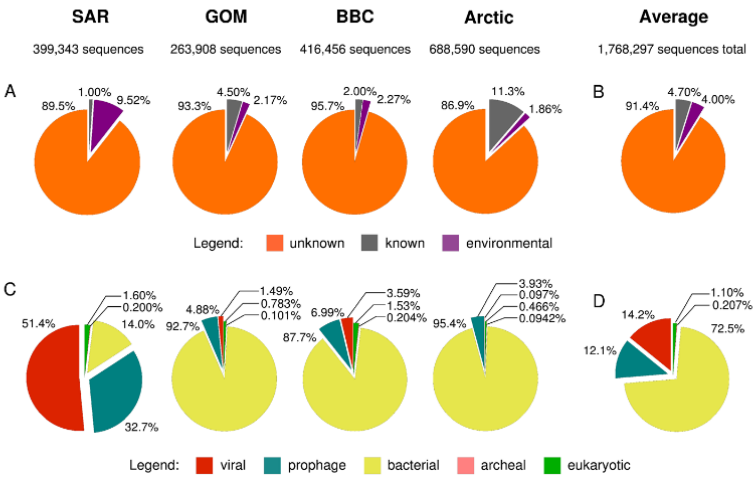


Figure 2. Composition of the Assemblage Genome Sequences as Determined by Similarity to Known DNA and Protein Sequences

(A) The percent of "known" sequences compared to the SEED and environmental databases. A sequence was considered "known" if it had a significant similarity ($E < 10^{-5}$) to the SEED, else "environmental" if it had a similarity to any environmental database, and else "unknown".

(B) Breakdown of the "known" sequences into viral (both eukaryotic and bacteriophages), prophage, Bacteria, Archaea, or Eukarya.

DOI: 10.1371/journal.pbio.0040368.g002

Ohio, United States). The thermocycler conditions were: 5 min at 94 °C; 30 cycles of 1 min at 94 °C, 1 min at 65 °C – 0.5 °C per cycle, 3 min at 72°C; and 10 min 72 °C. Positive PCR products were sequenced for verification of sequence length, and identity was confirmed using TBLASTX.

Permutation Tail Probability Tests of Phylogenic Similarity between Phage Assemblages

Because the sequences did not originate from a single genetic locus, the evolutionary relationships could not be determined by using standard alignment-based phylogenetic analyses. To determine phylogeny, the sequences were first mapped to the Phage Proteomic Tree based on their best TBLASTX similarity. The version of the Phage Proteomic Tree used here contained 510 complete phage genomes (http://scums.sdsu.edu/phage/Oceans) and was constructed as described previously [23]. Permutation tail probability (PTP) was then used to infer phylogenetic similarity among the phage assemblages. The PTP test uses phylogenetic parsimony to determine whether a given characteristic correlates with phylogeny [24]. Briefly, if a sequence had a best similarity to a phage genome on the Phage Proteomic Tree, it was scored on a tree using Phylogenetic Analysis Using Parsimony software (PAUP) [25]. The number of steps that would be required to produce a tree from one sample to another was then determined. To assign significance, this value was compared to a distribution produced by randomizing the input tree 10,000 times.

Genetic Isolation by Distance of the Phage Assemblages

Isolation by Distance Web Service (IBDWS) ([26]; http://biome.sdsu.edu/ibdws) was used to test for a correlation between the geographic distance between two samples and the genetic divergence between viral assemblages. This online software uses Mantel tests to determine whether marine phages in closer physical proximity have greater genetic similarity (as measured by Φ_{ST}) than those separated by large geographic distances. For these tests, the current datasets were combined with data from the California coast [10]. The Arlequin program [27] was used to calculate Φ_{ST}. The Φ_{ST} statistic compares the phylogenetic diversity within each assemblage to the total phylogenetic diversity of the combined assemblages using the equation:

$$\Phi_{ST} = (\theta_T - \theta_W)/\theta_T \quad (1)$$

where θ_T is the total phylogenetic diversity of two assemblages and θ_W is the phylogenetic diversity within each assemblage or population. A Φ_{ST} value close to zero means there is complete overlap in the phylogenetic diversity, whereas values greater than zero indicate increasing levels of phylogenetic differentiation up to a value of 1, indicating complete differentiation.

Assembly and Mathematical Modeling of Viral Assemblage Diversity

To estimate viral diversity, sets of 10,000 random sequences from each oceanic region were assembled using TIGR Assembler [28] with a minimum overlap length of 35 bp, a minimal match percentage of 98% and no alignment error in 32 bp to identify overlapping sequences (contigs) [10]. The Perl script used to automate this task is available at http://scums.sdsu.edu/phage/Oceans. Average contig spectra were calculated (Figure S3) over ten repetitions, and the maximum likelihood assemblage structure of the marine viral assemblages was determined using mathematical rank-abundance models in PHAge Communities from Contig Spectra (PHACCS) ([29]; http://biome.sdsu.edu/phaccs). Random subsamples of the metagenomes were used instead of the totality of the whole metagenomes, because PHACCS analyses are more robust at low coverage [10,11,29]. The diversity estimates for the best-fitting assemblage model were used for each oceanic region. Detailed graphical explanations of these procedures are given in Protocol S1.

Table 1. Number of Similarities to Phage Genomes and Groups of Interest in the Four Metagenomes

Group of Interest	Phage Species	Marine Region			
		SAR	GOM	BBC	Arctic
Cyanophage	*Prochloro. marinus* ϕ P-SSM2	4661[a]	589[a]	1190[a]	148[a]
	Prochloro. marinus ϕ P-SSP7	4493[a]	81[a]	86	16
	Prochloro. marinus ϕ P-SSM4	1759[a]	263[a]	587[a]	51
	Synechococcus ϕ S-PM2	1107[a]	196[a]	474[a]	54
Prophage	*Br. melitensis* 16M ϕ Bruc1 pro-ϕ	12	115[a]	92	700[a]
	Yersinia pestis ϕ Yers2 pro-ϕ	12	60[a]	34	386[a]
	Escherichia coli ϕ CP4–6 pro-ϕ	4	52	24	364[a]
	Agro. tumefaciens ϕ Tum2 pro-ϕ	14	43	55	281[a]
	Escherichia coli ϕ CP037–7 pro-ϕ	6	40	11	240[a]
	Xy. fastidiosa ϕ Xpd5 pro-ϕ	34	36	23	187[a]
	Escherichia coli ϕ CP037–4 pro-ϕ	3	29	11	146[a]
	Mesorhizobium loti ϕ Meso1 pro-ϕ	32	35	176[a]	56
	Pseudo. putida ϕ PP03 pro-ϕ	397	57[a]	96	1
chp1-like microphage (ssDNA)	*Bd. bacteriovorus* ϕ MH2K	1835[a]	20	115	0
	Chlamydia ϕ 4	1757[a]	5	119	0
	Chlamydia ϕ 3	1572[a]	9	119	0
	Chlamydia psittaci ϕ 2	568[a]	2	29	0
	Chlamydia psittaci ϕ chp1	519[a]	16	60	0
	Chlamydia ϕ CPAR39 pro-ϕ	1548[a]	14	112	0
Miscellaneous	*Salmonella* ϕ epsilon15	56	41	172[a]	116[a]
	Burkholderia thailandensis ϕ E125	7	29	29	111[a]
	Roseobacteria SIO67 ϕ SIO1	360	409[a]	465[a]	36
	Rhodothermus marinus ϕ RM 378	301	93[a]	206[a]	20
	α-proteobacteria ϕ JL001	333	45	197[a]	55
	Bordetella ϕ BIP-1	128	62[a]	167[a]	63
	Pseudo. aeruginosa ϕ PaP3	123	55	161[a]	10

[a]The ten most abundant similarities are noted for each sample.
Prochloro., *Prochlorococcus*; Br., Brucella; Agro., Agrobacterium; Pseudo., Pseudomonas; Bd., Bdellovibrio.
DOI: 10.1371/journal.pbio.0040368.t001

To analyze the degree of similarity between the viral assemblages, the amount of overlap between the assemblages was determined by assembling a mixed sample of 10,000 fragments obtained by pooling 2,500 fragments from each region. The fact that fragments from one region assembled with fragments from another region indicates overlap between the metagenomes of the two regions, and the extent of this overlap quantifies the similarity. The contig spectrum obtained from the mixed sample was modified in two respects to give what is called the cross-contig spectrum (Figure S4). First, any contig that contained fragments exclusively from a single region was removed (i.e., only contigs that included fragments from more than one region were counted). Thus for the contigs of size $q > 1$, $\hat{C}_q$, the number of q-contigs from the pooled sample that included fragments from more than one region, was calculated. Second, the number of 1-contigs from each region that assembled with any fragments from other regions was used as the number of 1-cross-contigs, $\hat{C}_1$. The resulting cross-contig spectrum $[\hat{C}_1, \hat{C}_2, \hat{C}_3, ...]$ was then compared to the mean cross-contig spectrum from simulated mixtures of the four assemblages. To simulate such mixtures requires a model of which genomes with a certain rank and abundance in one assemblage correspond to which genomes in another.

There are many ways to envision morphing one assemblage of genotypes (species defined on the genomic level by assembly of sequences) into another. For these analyses, two morphing modes were considered (Figure S5): (i) varying the proportion of genotypes that were shared between assemblages and (ii) varying the proportion of the genotypes whose abundance ranks were shuffled (i.e., subjected to a random permutation). Using these two degrees of freedom, s (percent shared) and p (percent permuted), Monte Carlo analyses were performed to estimate the degree of morphing as measured by these two parameters to find maximum-likelihood values for s and p based on the closeness of the match to the cross-contig spectrum found for the pooled sample.

The Monte Carlo simulations were all performed using the best-fit models for each region. The cross-contig spectrum based on the mixed sample was used to perform these simulations (Figure S6). Each simulation included 861 pairs of s and p values spanning a 21×41 grid between 0% and 100% for each parameter. Each simulation randomly permuted the abundance rank of p of the most abundant genotypes, randomly assigned s of the genotypes to be shared, and determined the resulting predicted cross-contig spectrum. This was repeated 100 times for each combination of s and p values. The entire simulation, including the selection of the 2,500 fragments from each region, was repeated eight times resulting in 800 predicted cross-contig spectra for each combination of parameter values. The mean $\hat{c}_q$ and variance $\hat{\sigma}_q^2$ of these 800 values were then used to construct a quasi-likelihood $\mathcal{L}(s,p)$

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

of matching the observed cross-contig spectrum, thereby generating a contour map of L as in [11]. This log likelihood

would be expected if each cross-contig value were normally distributed. The contour map of the quasi-likelihood landscape was produced from this grid of 861 quasi-likelihood values. As a control, the whole procedure was repeated for all regions with nonoverlapping subsets of sequences all taken from the same geographical region (rather than from four different regions).

Results/Discussion

"Community" is commonly defined several ways, including "the species that occur together in space and time" [30] and "an association of interacting populations" [31]. Assemblage is probably the most proper term to describe viral groups, and most instances of "community" in the literature, both by ourselves and others, is not correct. See [32] for a disambiguation of some important ecological terms.

General Characteristics of the Marine Viral Metagenomes

On average, >91% of the sequences were not significantly similar to those in the extant databases (Figure 2A). A partial explanation for the high percentage of unknowns is almost certainly due to the shorter sequences (~100 bp on average) that are generated by pyrosequencing at 454 Life Sciences. Previous viral metagenomic studies that used Sanger sequencing (~650 bp fragments) found that >60% of the sequences were unknowns [33]. The Arctic Ocean sample had the highest percentage of known similarities (11%) to the SEED database, mostly because of the large number of prophage-like sequences (Table 1). Comparison of the marine viral sequences to the environmental database did not yield a significant number of new similarities compared to the SEED database (~2% to the environmental database), with the notable exception of the Sargasso Sea sample, where >9% of the similarities were to the environmental database, presumably because the major sources of sequences for the environmental database were the Sargasso Sea microbial metagenomes, originally collected in 2003 [17]. The overlap between the viral metagenome and the microbial metagenomes raises several important points. First, a significant number of viral sequences are retained on the larger-pore filters, either as free viruses, proviruses, or in cells undergoing a burst. The latter explanation was hypothesized by Delong et al. [19], who observed a large number of viral similarities at one depth at the Hawaii Oceanic Time-series (HOT) station. Second, the microbial assemblages in the Sargasso Sea appear to be relatively stable over prolonged periods (~2 y). Finally, the small amount of sampling and sequencing represented by these two studies (~10^{12} bp) is already constricting the unknown sequence space of the Sargasso Sea. With the continual decline in Sanger sequencing costs and introduction of large-scale pyrosequencing, metagenomic approaches should be able to characterize global sequence diversity in a relatively short period of time.

Among the fraction of sequences with similarity to the SEED database, most of the "knowns" were similarities to bacterial sequences in the Arctic, British Columbia, and Gulf of Mexico samples (Figure 2B). This can be accounted for by the following: (i) the larger number of microbial rather than viral genomes in the database, (ii) unidentified prophages within microbial genomes, (iii) the large amount of horizontal gene transfer between phages and their hosts, (iv) phages

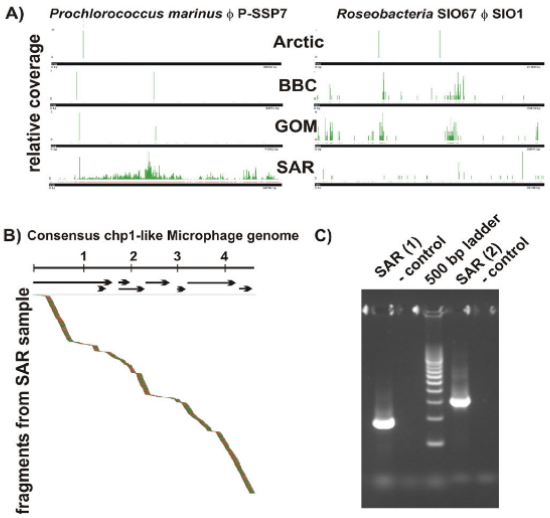


Figure 3. Distribution of Similarities and Assembly Controls

(A) Distribution of similarities between the four metagenome samples to the *P. marinus* ϕ P-SSP7 and *Roseobacteria* SIO67 ϕ SIO1 genomes (as determined by BLASTN analysis). The green bars represent the average number of sequences averaged over 100 bp windows.

(B) Comparison of fragments from the Sargasso Sea metagenome against the consensus ssDNA chp1-like microphage genome. The consensus from this assembly is in the Protocol S1.

(C) PCR verification of chp1-like microphages in original SAR sample. PCR primers were designed based on a consensus sequence from the assembly shown in (B). SAR1 is a ~900-bp fragment and SAR2 is a ~1,500-bp fragment.

DOI: 10.1371/journal.pbio.0040368.g003

carrying full genes from their host, as observed in sequenced phage genomes [34,35], and (v) the overall larger size of bacterial genes relative to viral genes, statistically increasing the probability of sequencing and hitting them.

The sample from the Sargasso Sea was exceptional in that the majority of "known" sequences were most similar to three *Prochlorococcus* phage genomes (Table 1) originally isolated from the same area of the ocean [34]. This finding suggests that just a few phage genomes from novel environments will greatly increase our understanding of viral diversity in these environments. The distribution of BLASTN similarities along the *Prochlorococcus marinus* ϕ P-SSP7 genome [34] is shown in Figure 3A. There is almost complete coverage of the genome within the Sargasso Sea sample. In contrast, the similarly sized *Roseobacteria* SIO67 ϕ SIO1 genome [36], which was isolated from near-shore waters in California, is only sparsely covered in the Sargasso Sea sample, but has higher coverage in the Gulf of Mexico and British Columbia samples. This supports the idea that certain phage groups are more prevalent in certain biogeographic regions. This general pattern was reinforced by the observation of a number of phage genomes and groups prevalent in different oceanic regions (Table 1).

The five most abundant putative viral-encoded enzymes (Table 2) appear to be involved in scavenging host nucleotides (e.g., riboreductases) and supporting host metabolism through the infection cycle (e.g., carboxylyases and transferases). The viral fraction also contained *psbA* genes, which encode the D1 protein of photosystem II in the cyanobac-

Table 2. The Most Abundant Enzyme-Coding Genes in the Four Oceanic Viral Metagenomes

Marine Region	Enzyme Name	EC number	Gene Occurrences
Sargasso Sea (SAR)	Ribonucleotide reductase of class Ia (aerobic), alpha subunit	1.17.4.1	89
	Ribonucleoside-diphosphate reductase	1.17.4.1	75
	Ribonucleotide reductase of class II (coenzyme B12-dependent)	1.17.4.1	50
	GTP cyclohydrolase I, type 2	3.5.4.16	37
	Adenine-specific methyltransferase	2.1.1.72	22
Gulf of Mexico (GOM)	Formate dehydrogenase-O, major subunit	1.2.1.2	27
	Carbamoyl-phosphate synthase large chain	6.3.5.5	25
	Cytochrome c oxidase polypeptide I	1.9.3.1	24
	Ribonucleotide reductase of class II (coenzyme B12-dependent)	1.17.4.1	23
	DNA polymerase III alpha subunit	2.7.7.7	23
British Columbia coast (BBC)	Ribonucleotide reductase of class II (coenzyme B12-dependent)	1.17.4.1	34
	DNA polymerase III alpha subunit	2.7.7.7	22
	3-polyprenyl-4-hydroxybenzoate carboxylyase	4.1.1.-	18
	Cytochrome c oxidase polypeptide I	1.9.3.1	18
	Ribonucleotide reductase of class Ia (aerobic), alpha subunit	1.17.4.1	18
Arctic Ocean	3-polyprenyl-4-hydroxybenzoate carboxylyase	4.1.1.-	205
	DNA polymerase III alpha subunit	2.7.7.7	185
	Cytochrome c oxidase polypeptide I	1.9.3.1	175
	Isoleucyl-tRNA synthetase	6.1.1.5	157
	Methylcrotonyl-CoA carboxylase carboxyl transferase subunit	6.4.1.4	155

EC number, Enzyme Commission number.
DOI: 10.1371/journal.pbio.0040368.t002

teria. The majority of sequenced cyanophages carry this gene, and evidence is mounting that the cyanophages need the D1 protein for successful infection and replication [34,37,38]. The occurrence of *psbA* was lowest in the Arctic sample, probably reflecting a decrease in the host and cyanophage numbers in the colder environments.

Discovery of an Abundant Marine ssDNA Phage Group

The Sargasso Sea sample had a large number of sequences (6% of the total; Table 1) with significant similarities to chp1-like Chlamydiamicrovirus (Microviridae family). These viruses are small ssDNA phages. Assemblies from these sequences resulted in the near-complete genomes of several marine Microviridae phages from the Sargasso Sea sequences (Figure 3B). To our knowledge, this is the first report describing the presence of this phage group in the marine environment, which was previously overlooked because the amplification and cloning methods excluded ssDNA viruses. The only other report of ssDNA viruses in the marine environment was a Circovirus that infected diatoms [39]. However, the marine sequences in this study did not show any similarity to that virus. Sequences with significant similarity to the chp1-like phages were observed less frequently in the British Columbia (~10-fold less common than in SAR) and Gulf of Mexico samples (~100-fold less common than in SAR). No sequences from this group were found in the Arctic sample (Table 1 and Figure 4). Primers were designed against these genomes and appropriately sized DNA fragments were amplified from the Sargasso Sea sample (Figure 3C). No amplicons were detected in the Gulf of Mexico or British Columbia samples, suggesting that they were present at numbers below the level of detection in this PCR or had a divergent sequence. A geographical constraint that limits the distribution of these viruses would be most consistent with these results. However concerns about sample amplification and storage bias make it impossible to accurately access the relative abundances of these viruses at this point.

Every Phage Everywhere?

The distribution of similarities to the chp1-like Microphage, *P. marinus* ϕ P-SSP7, *Roseobacteria* SIO67 ϕ SIO1, and others in the viral-fraction suggests that viruses have restricted geographical distributions similar to those observed in micro- and macro-organisms [40,41]. This is in contrast to studies that have shown that identical phage genes are distributed throughout the biosphere and that phages from soils and sediments can replicate in marine microbial populations [3,42,43]. To determine whether all marine phages are spread everywhere or if there is a strong regionalization, three different approaches were used.

A new version of the Phage Proteomic Tree was constructed, and similarities from the samples were mapped onto this tree (Figure 4). Eighty-four phage species were specific to one marine region, whereas 45 were common to all four. From the remaining phage species, 102 were found in several oceanic regions. The phylogenetic parsimony of phages from each sample was compared to the Phage Proteomic Tree using the PTP tests, because viruses do not have a single genetic locus conserved across all genomes. The PTP test showed that the distribution of phages in the marine samples is not random. First, marine phages are phylogenetically distinct from the available genomes, suggesting a "marineness" to the group as a whole ($p < 0.0001$; 10,000 randomizations). Second, there was a significant difference between phages from the different oceanic regions ($p < 0.0001$; 10,000 randomizations), supporting a geographical specificity for viruses despite the wide prevalence of some phage species.

An Isolation By Distance (IBD) approach demonstrated that there was a significant positive correlation between geographic distance (km) and genetic distance (as measured by Φ_{ST}) (Mantel test; $Z = -78.9$; $r = 0.585$; $p < 0.017$) (Figure 5),

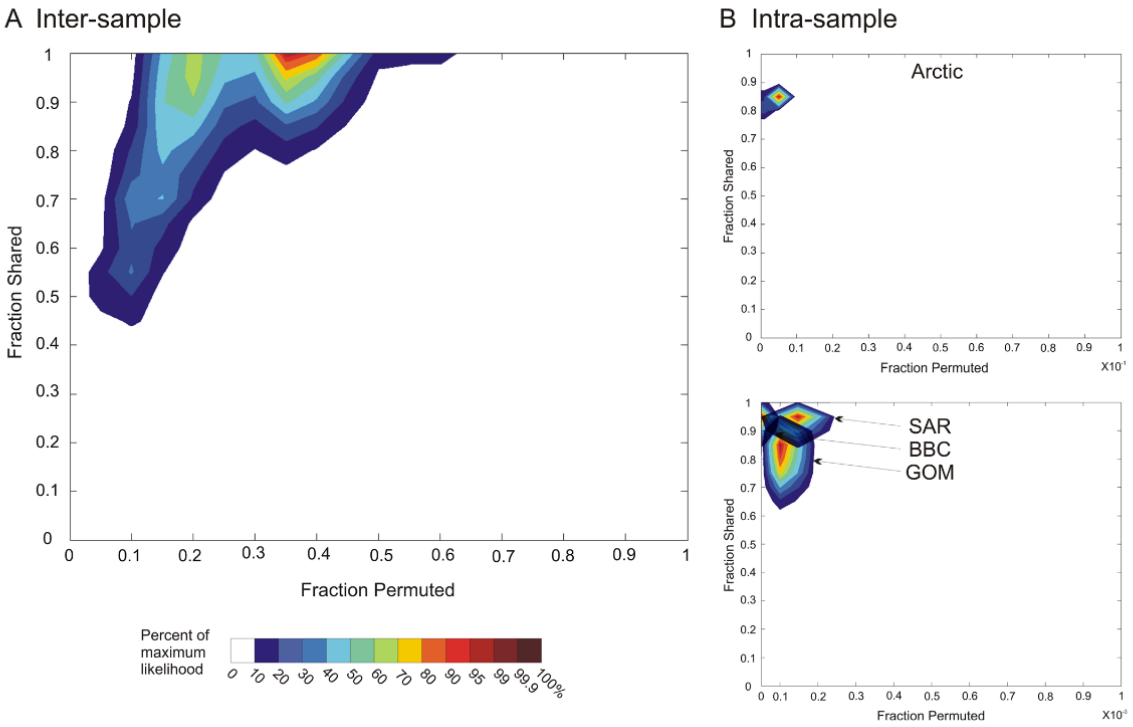


Figure 6. Monte Carlo Simulation of Cross-Contigs between Metagenomic Samples

(A) For the intersample analysis, the maximum likelihood occurred at 35% fraction permuted and 100% fraction shared. (B) The maximum likelihood was between 0% and 0.5% fraction permuted and 85% and 95 % fraction shared for the intrasample controls.

DOI: 10.1371/journal.pbio.0040368.g006

metagenome was the least genotype-rich (532 predicted genotypes) and diverse (H′ of 6.05 nats).

Being located on the west coast of the North American continent, the coast of British Columbia is in an upwelling area. It is also enclosed and fed by many rivers. These conditions might importantly increase the diversity of microbial communities and thus provide an explanation for the very high viral assemblage diversity estimated in this oceanic region. Omitting the BBC, the viral diversity of the other regions (the Gulf of Mexico, Sargasso Sea, and Arctic Ocean) correlate with the well-established North-South latitudinal diversity gradient [44], with a larger diversity at lower latitudes. Planktonic diversity patterns of near-shore versus off-shore (more diverse plankton assemblages off-shore) [45] were not observed here; the large spatial scale of the sampling probably masked this effect if present.

Assemblies of the mixed sample were used to predict global viral diversity using PHACCS. A total of 57,600 different viral genotypes in all four regions (H′ of 9.8 nats) was estimated. This number is smaller than the number of genotypes predicted in the BBC sample, which may indicate an undersampling for the mixed metagenome or be due to some of the assumptions of the model. Taken together, these data indicate that the global marine viral richness could be as high as a few hundred thousand species, with a regional richness sometimes almost as high, likely because of migration processes.

Integrative Versus Single Samples

It was expected that the integrated samples would be more even because it is assumed the viruses that were most abundant at one spatial-temporal time point would be rarer at another ("kill-the-winner" hypothesis). As summarized in Table 3, the evenness of the single time point sample (SAR 0.905) fell in between that of the three integrated samples (Arctic 0.964; BBC 0.918; GOM 0.851). Similarly, the predicted richness (5140 genotypes) and diversity (H′ 7.74 nats) at the single point represented by the Sargasso Sea sample fell in between that of the integrated samples (richness 532–129,000; H′ 6.05–10.8 nats). Because of factors with a supposedly greater impact, like latitude, it is not clear that integrating individual samples gave a greater depth of coverage.

Without a doubt, many interesting trends based on depth and a wide variety of other spatial, biological, and temporal parameters were missed by the integrative sampling used here. However, this sampling does provide a useful overview of the marine virome on a global and regional scale. Currently, there are no real criteria as to what constitutes a useful size or time scale for sampling natural viral assemblages, so there is no particular advantage or disadvantage to keeping samples separate or analyzing them as a metadataset.

Rather the sampling scheme should be driven by the question being addressed. Viral assemblages are interesting in their own right, not just in context of their host communities. However, future studies should also start cross-correlating the viruses with their hosts. Of particular interest will be determining if the "islands" and ORFans observed in microbial genomes are represented in the virome [6,46].

Potential Sampling and Processing Biases

Sampling bias in the current datasets was primarily due to loss of large viruses during filtering. Currently, there is no experimental method to avoid this problem. The cesium chloride gradients used here recover all known phage groups, and essentially all the viral-like particles in the starting samples migrate to the proper density in these preparations (as observed by epifluorescence microscopy; unpublished data). Unfortunately, the cloning methods used here will not recover RNA viruses. Suttle et al. [47,48] have shown that RNA viruses are present in the marine environment. Whereas most electron microscopy [49,50] and nucleic acid–based studies [51] have not found RNA viruses in large numbers, RNA viruses are still believed to be important components of the marine virome that need additional study.

Another potential source of bias is the different times that the samples were stored before processing. Phage particles are very stable and often stored for decades at 4 °C. This is a commonly known lab phenomenon and is supported by the observation that the oldest viral concentrates (~12 y old) in this study had very high concentration of viruses (>10^9 viral-like particles per ml). Different phages, however, may have different decay rates under these conditions. This does not seem to be especially problematic, because there is no correlation between the types of viruses observed and the storage time. For example, the Arctic and SAR samples are the most recently harvested samples, yet they have the biggest differences in terms of types of phages (Table 1). Nonetheless, there may be effects of storage on the composition of the viral assemblages. For this reason, analyses based on absolute abundances of one specific virus to another were avoided in this study. Instead, the presence of a sequence in the metagenome was simply assumed to mean that the virus was in the original sample (i.e., an occurrence).

Whole-genome amplification techniques introduce biases in the relative concentrations of different genomes. Tests of Genomiphi by the manufacturer and others [52,53] have not reported a significant bias in the amplification of circular double-stranded DNA (dsDNA), with the exception of very small dsDNA targets (<1 kb), which are much smaller than the vast majority of marine viruses, and of ssDNA, which will probably be a preferred target for the DNA polymerase. Although not bias-free, Genomiphi is the most accurate amplification method available [54]. Interesting trends associated with viral assemblage structure may have been missed because of our choice of using presence/absence data for the analyses presented here, but by being conservative there should not be any effects of storage, amplification, and sampling biases on our interpretations.

Conclusion

The metagenomic analysis of viral assemblages from the Arctic Ocean, the coast of British Columbia, the Gulf of Mexico, and the Sargasso Sea presented here has changed our perception on the composition of viral assemblages in the sea. First, there is clear evidence that the composition of viral assemblages varies in different geographic regions probably reflecting selective pressure. Previously overlooked viral groups, such as ssDNA viruses and prophages, can be major constituents of marine viral assemblages (Sargasso Sea and Arctic Ocean, respectively). Second, global viral diversity is high (possibly a few hundred thousand viral species), but regional diversity can be almost as high due to viral migration. This migration provides opportunities for global exchange of DNA among viral genomes, as predicted by the mosaic model [55]. Viral diversity also varied according to latitude, with a higher richness at low latitudes. Finally, it seems that although some viral species are endemic and others are ubiquitous, the vast majority are widespread and shared between several oceanic regions. Invasion and replacement by new phages does not appear to be an important structuring factor for these viral assemblages. What sets different assemblages apart is likely the change in abundance of its most abundant members, supporting to some extent the old tenet "everything is everywhere, but, the environment selects" [56] for marine viruses.

Supporting Information

Figure S1. Frequency of Homopolymeric Tracts in the Four Marine Viromes, the Complete Phage Genomes, and Twenty, Randomly Chosen Microbial Genomes

The tracts from 3 nucleotides (nt) to 15 nt were counted and normalized to the number of bases in each sequence. One 3-nt tract is

Table 3. Viral Assemblage Structure Predicted from Assembly of Metagenomic Sequences

Sample	Richness	Evenness	Most Abundant Genotype (%)	Shannon-Wiener Index
Arctic	532 genotypes	0.964	2.27	6.05 nats
BBC	129,000 genotypes	0.918	7.28	10.8 nats
GOM	15,400 genotypes	0.851	13.3	8.21 nats
SAR	5140 genotypes	0.905	8.45	7.74 nats
Mixed	57,600 genotypes	0.895	9.34	9.81 nats

Ten separate assemblies of 10,000 sequences chosen at random from each library were performed for each sample. For the mixed sample, 2,500 randomly chosen fragments were used from each library. The average contig spectrum was used to predict assemblage structure using PHACCS.

DOI: 10.1371/journal.pbio.0040368.t003

found approximately every 30 bp, whereas one 15-nt tract is found approximately every 10 million bp. The 510 complete phage genomes totaled 18,909,173 bp in length, and the microbial genomes totaled 22,110,123 bp in length. The lengths of the pyrosequenced libraries are given in the text.

Found at DOI: 10.1371/journal.pbio.0040368.sg001 (1.3 MB TIF)

Figure S2. Relative Abundance of Phages in the Four Metagenomes

Because of way the samples were stored and the long storage time, the distribution shown may not accurately reflect the reality.

Found at DOI: 10.1371/journal.pbio.0040368.sg002 (104 KB TIF)

Figure S3. Determining a Normal Contig Spectrum

Found at DOI: 10.1371/journal.pbio.0040368.sg003 (135 KB TIF)

Figure S4. Getting a Cross-Contig Spectrum.

Found at DOI: 10.1371/journal.pbio.0040368.sg004 (3.1 MB TIF)

Figure S5. The Possible Scenarios Considered in the Monte Carlo Simulation to Explain the Observed Cross-Contigs

Found at DOI: 10.1371/journal.pbio.0040368.sg005 (143 KB TIF)

Figure S6. Analyzing a Cross-Contig Spectrum

Found at DOI: 10.1371/journal.pbio.0040368.sg006 (589 KB TIF)

Protocol S1. Details on Materials and Methods.

Found at DOI: 10.1371/journal.pbio.0040368.sd001 (39 KB PDF)

Accession Numbers

The Genome Projects Database (http://www.ncbi.nlm.nih.gov/Genomes) accession numbers for the sequences are 17765 (GOM), 17767 (BBC), 17769 (Arctic), and 17771 (SAR); the Genome Catalogue (http://gensc.sf.net) accession numbers are 000002__GCAT (GOM), 000003__GCAT (BBC), 000004__GCAT (Arctic), and 000005__GCAT (SAR); and the GOLD database (http://www.genomesonline.org) GOLDstamps are GM00060 (GOM), GM00061 (BBC), GM00062 (Arctic), and GM00063 (SAR).

Acknowledgments

The GOM, BBC, and Arctic samples were collected with the generous help of the crew and scientists aboard the research vessels *Longhorn*, *Mirai*, *Raddisson*, *Walton Smith*, and *Vector*. We are grateful to A.M. Comeau, A.C. Ortmann, C.M. Short, S.M. Short, M.G. Weinbauer, and S.W. Wilhelm for sample collection and processing; as well as K. Shimada, E.C. Carmack, and J. Paul for providing the opportunity to participate in the *Mirai* and *Walton Smith* expeditions. The Sargasso Sea samples were collected with the assistance of R. Morris and the Captain and Crew or the R.V. *Weatherbird II*.

Author contributions. FR conceived and designed the experiments. CC, AC, MH, and RP performed the experiments. FA, BF, MB, PS, RE, SK, HL, JMM, JEM, JN, SR, CS, and FR analyzed the data. RO contributed reagents/materials/analysis tools. FA, BF, MB, RE, JEM, CS, and FR wrote the paper.

Funding. These collections were supported by NSERC grants to CAS (Discovery, Shiptime and Research Network [CASES]) and by ONR and NSF grants to CAS. Ship time and collection was supported by an NSF microbial observatory grant to CAC. The Marine Microbial Initiative by the Gordon and Betty Moore Foundation (FR) sponsored the sequencing, bioinformatics, and mathematical analyses.

Competing interests. The authors have declared that no competing interests exist.

References

1. Weinbauer MG (2004) Ecology of prokaryotic viruses. FEMS Microbiol Rev 28: 127–181.
2. Wilcox RM, Fuhrman JA (1994) Bacterial viruses in coastal seawater: Lytic rather than lysogenic production. Mar Ecol Prog Ser 114: 35–45.
3. Sano E, Carlson S, Wegley L, Rohwer F (2004) Movement of viruses between biomes. Appl Environ Microbiol 70: 5842–5846.
4. Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brussow H (2003) Phage as agents of lateral gene transfer. Curr Opin Microbiol 6: 417–424.
5. Davis BM, Waldor MK (2002) Mobile genetic elements and bacteria pathogenesis. In: Craig NL, Gragie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington (DC): ASM Press. pp. 1040–1055.
6. Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: The genealogy of ORFans in *E. coli*. Genome Res 14: 1036–1042.
7. Thingstad TF, Lignell R (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquat Microb Ecol 13: 19–27.
8. Hoffmann K, Rodriguez-Brito B, Breitbart M, Bangor D, Angly FE, et al. (2005) The structure of marine phage populations. In: Kjelstrup S, Hustad J, Gundersen T, Røsjorde A, Tsatsaronis G, editors. Trondheim (Norway): Tapir Academic Press. 5 p.
9. Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, et al. (2005) Genotypic diversity within a natural coastal bacterioplankton population. Science 307: 1311–1313.
10. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99: 14250–14255.
11. Breitbart M, Felts B, Kelley S, Mahaffy J, Nulton J, et al. (2004) Diversity and population structure of a nearshore marine sediment viral community. Proc R Soc Lond Ser B Biol Sci 271: 565–574.
12. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 85: 6220–6223.
13. Breitbart M, Rohwer F (2005) Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. Biotechniques 39: 729–736.
14. Edwards R, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, et al. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7.
15. Zhang T, Breitbart M, Lee W, Run J-Q, Wei C, et al. (2006) RNA viral community in human feces: Prevalence of plant pathogenic viruses. PLoS Biol 4: e3. DOI: 10.1371/journal.pbio.0040003.
16. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37–43.
17. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74.
18. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554–557.
19. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496–503.
20. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: A laboratory manual. New York: Cold Spring Harbor Laboratory Press. 1659 p.
21. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–480.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
23. Rohwer F, Edwards R (2002) The phage proteomic tree: A genome based taxonomy for phage. J Bacteriol 184: 4529–4535.
24. Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. Appl Environ Microbiol 68: 3673–3682.
25. Swofford DL (2000) PAUP*. Phylogenetic analysis using parsimony (*and other methods). 4.0 ed. Sunderland (Massachusetts): Sinauer Associates.
26. Jensen JL, Bohonak AJ, Kelley ST (2005) Isolation by distance, web service. BMC Genetics 6: 13.
27. Schneider S, Kueffer JM, Roessli D, Excoffier L (1997) Arlequin: A software for population genetic data analysis. 1.1 ed. Geneva: Genetics and Biometry Lab, Department of Anthropology.
28. Sutton GG, White O, Adams MD, Kerlavage AR (1995) TIGR assembler: A new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1: 9–19.
29. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6: 41.
30. Begon M, Harper JL, Townsend CR (1990) Ecology: Individuals, populations, and communities. Boston: Blackwell Scientific Publications. 945 p.
31. Ricklefs RE (1990) Ecology. New York: WH Freeman. 896 p.
32. Fauth JE, Bernardo J, Camara M, Resetarits WJ, Buskirk JV, et al. (1996) Simplifying the jargon of community ecology: A conceptual approach. Am Nat 147: 5.
33. Edwards R, Rohwer F (2005) Viral metagenomics. Nature Rev Microbiol 3: 504–510.
34. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. PLoS Biol 3: e144. DOI: 10.1371/journal.pbio.0030144.
35. Seguritan V, Feng IW, Rohwer F, Swift M, Segall AM (2003) Genome sequences of two closely related Vibrio parahaemolyticus phages, VP16T and VP16C, that coexist in the same host. J Bacteriol 185: 6434–6447.

36. Rohwer F, Segall AM, Steward G, Seguritan V, Breitbart M, et al. (2000) The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with non-marine phages. Limnol Oceanogr 42: 408–418.
37. Zeidner G, Bielawski JP, Shmoish M, Scanlan DJ, Sabehi G, et al. (2005) Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. Environmental Microbiology 7: 1505–1513.
38. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, et al. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol 4: e234. DOI: 10.1371/journal.pbio.0040234.
39. Nagasaki K, Tomaru Y, Takao Y, Nishida K, Shirai Y, et al. (2005) Previously unknown virus infects marine diatom. Appl Environ Microbiol 71: 3528–3535.
40. Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, et al. (2006) Microbial biogeography: Putting microorganisms on the map. Nat Rev Microbiol 4: 102–112.
41. Dolan JR (2005) An introduction to the biogeography of aquatic microbes. Aquat Microb Ecol 41: 39–48.
42. Breitbart M, Rohwer F (2004) Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol Lett 236: 245–252.
43. Short CM, Suttle CA (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in marine and freshwater environments. Appl Environ Microbiol 71: 480–486.
44. Hillebrand H (2004) Strength, slope and variability of marine latitudinal gradients. Marine Ecol Prog Ser 273: 251–267.
45. Hutchinson GE (1961) The paradox of the plankton. Am Nat 95: 137–145.
46. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311: 1768–1770.
47. Culley AI, Lang AS, Suttle CA (2003) High diversity of unknown picorna-like viruses in the sea. Nature 424: 1054–1057.
48. Culley AI, Lang AS, Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. Science 312: 1795–1798.
49. Nomizu T, Mizuike A (1986) Electron microscopy of submicron particles in natural waters: Specimen preparation by centrifugation. Mikrochim Acta 1: 65–72.
50. Moebus K (1991) Preliminary observations on the concentration of marine bacteriophages in the water around Helgoland. Helgo Meeresunters 45: 411–422.
51. Steward GF, Montiel JL, Azam F (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. Limnol Oceanogr 45: 1697–1706.
52. Hutchison IIICA., Smith HO, Pfannkoch C, Venter JC (2005) Cell-free cloning using phi29 DNA polymerase. Proc Natl Acad Sci U S A 102: 17332–17336.
53. Yokouchi H, Fukuoka Y, Mukoyama D, Calugay R, Takeyama H, et al. (2006) Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using phi-29 polymerase. Environ Microbiol 8: 1155–1163.
54. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, et al. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics: 7: 216.
55. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing evolution of viruses. Trends Microbiol 8: 499–500.
56. de Wit R, Bouvier T (2006) "Everything is everywhere, but, the environment selects"; what did Baas Becking and Beijerinck really say? Environ Microb 8: 755–758.

Supplementary Material for "Marine Viromes of Four Oceanic Regions" by Angly et al.

Sample collection and preparation

Field collection from the Sargasso Sea (SAR): Samples were collected on June 30, 2005 from Hydrostation S (32° 10' N, 64° 30' W) located in the northwestern Sargasso Sea approximately 26 km southeast of the island of Bermuda, essentially the same spot the Surge Samples were collected. Water from 80 m was retrieved via the *R.V. Weatherbird II's* CTD rosette equipped with 12 liter Niskin bottles. This depth was targeted because summer subsurface maximum in viral like particles (VLP) is historically located between 80 and 100 m. Upon recovery of CTD, seawater was transferred to a pre-cleaned 150 L polypropylene carboy via acid washed silicone tubing. Viruses were concentrated using a Pelicon II tangential flow filtration system (Millipore Corporation; Bedford, MA) equipped with a 30 kd Biomax cassette filter (0.5 m^2; modified polyethersulfone). Prior to collection the filter was cleaned with 0.1 N H_3P0_4 recirculated for 45 minutes followed by 0.1N NaOH recirculated for 45 minutes. One-hundred and fifty liters of seawater was recirculated over the 30 kD filter until the retentate volume was ~150 ml. This concentration step took ~3 hours to complete. The final concentrate was distributed into four 45 ml aliquots in sterile Falcon tubes. Five ml of chloroform was added to each tube, stored and shipped at 4° C until arrival at SDSU. The final viral concentration was approximately 1.4 X 10^9 VLP ml^{-1}.

Field collection from the Gulf of Mexico (GOM), the Bay of British Colombia (BBC), and the Arctic: Samples were collected from the Gulf of Mexico, the Arctic Ocean, the Strait of Georgia, British Columbia and adjacent inlets, and Barclay Sound on the west coast of Vancouver Island. Further details on the samples and locations are listed in Supplemental Table S1. Seawater samples (10 to 200 liters) were prefiltered through 142 mm-diameter glass fiber filters (1.2 µm nominal pore-size: type GC50, Advantec MFS, Dublin, CA, or 0.7 µm nominal pore-size: type GF/F, Whatman, Clifton, NJ) followed by either 0.45-µm-pore-size (type GVWP, Millipore, Bedford, MA) or 0.2 µm-pore-size (Gelman, East Hills, NY) membrane filters. Virus-sized particles in the filtrate were concentrated ca. 50- to 700-fold by ultrafiltration using a 10 or 30 kDa-cutoff Amicon/Millipore (S1Y10/S1Y30/S10Y30) spiral cartridge and then stored at 4°C in the

dark. One milliliter of each virus concentrate (VC) was combined into one of the following mixes based on its geographical origin. The Gulf of Mexico mix consisted of 41 different virus communities (3.65 x 10^{10} vlps per ml), the Arctic mix consisted of 56 different virus communities (1.11 x 10^{10} vlps per ml), while the British Columbia mix consisted of 85 different communities (3.4 x 10^{10} vlps per ml). Virus abundances in each mixture were counted using SYBR Green I (Invitrogen, Carlsbad, CA) and epifluorescence microscopy.

Table S1. Temporal and spatial sampling of 4 marine provinces.

Location	Number of Stations	Number of Samples	Sampling Dates	Salinity (psu)	Temp (ºC)	Depth (m)
SAR						
Hydrostation S	1	1	6/05	36.7	19.8	80
GOM						
Western GOM	2	6	6/95 and 7/96	36.2 to 36.4	20.2 to 30.5	surface to 164
Texas Coast	5	13	6/94 to 7/96	26.1 to 40.6	14.7 to 30.5	surface to 5
Northeast GOM	6	14	7/01	31.5 to 36.6	18.7 to 30.0	1 to 120
Eastern GOM	2	8	7/01	33.5 to 36.6	21.6 to 30.0	3 to 90
Arctic						
Chukchi Sea	7	14	9/02	26.8 to 35.0	-1.4 to 5.4	10 to 3246
Canadian Arctic [Beaufort Sea, MacKenzie Shelf, Amundsen Gulf]	23	42	9/02 to 10/02	20.3 to 34.9	-1.5 to 1.5	2 to 968
BBC						
Strait of Georgia and adjacent inlets, Barclay Sound	38	85	8/96 to 7/04	14.0 to 31.3	7.0 to 22.6	surface to 245

Preparation of Marine Phage DNA samples for pyrosequencing at SDSU: Viral concentrates were filtered with 0.22 µm Sterivex cartridge filters to remove any microbial contaminants. Viral particles were treated with DNase, RNase, and purified by cesium chloride (CsCl) gradient centrifugation. Approximately 8.5 ml of viral concentrate, with CsCl added to create a density of 1.15 g ml^{-1}, was layered onto a step gradient comprised

of CsCl solutions at 1.7 g ml^{-1}, 1.5 g ml^{-1}, and 1.25 g ml^{-1}. CsCl solutions were made up in filtered and autoclaved seawater, obtained separately from the samples. The gradients were centrifuged at 22,000 rpm in an SW41 swinging bucket rotor at 4° C for 2 hours. After centrifugation, the 1.5 ml corresponding to the 1.5 g ml^{-1} gradient step plus the interfaces above and below, were withdrawn from the tubes with a syringe and a 18 gauge hypodermic needle.

DNA was extracted by addition of buffers to yield final concentrations of 0.2 M Tris, pH 8, and 5mM EDTA, followed by addition of 1 volume of deionized formamide. These samples were then incubated at room temperature for 30 min, after which 2 volumes of 100% ethanol were added to precipitate the DNA. The DNA was pelleted by centrifugation at 12,000 rpm in a fixed-angle rotor at 4° C for 20 min. The pellets were washed with 70% ethanol, and then resuspended in 567µl TE buffer (10 mM Tris, pH 8.0, 1mM EDTA) and then 30 µl of 10% SDS (sodium dodecyl sulfate) and 3 µl of proteinase K (20 mg ml^{-1}) were added, and the samples were incubated at 37° C for 1 hour. Subsequently 100 µl of 5 M NaCl and 80 µl of CTAB NaCl solution (0.7 M NaCl, 10% w/v cetyl trimethyl ammonium bromide) were added, followed by incubation for 10 min. at 65° C. The samples were then extracted with 1 volume of chloroform, then with 1 volume of phenol:chloroform:isoamyl alcohol (1:1:24), and finally with 1 volume of chloroform, after which 0.7 volumes of isopropyl alcohol (2-propanol) were added to precipitate the DNA. After storage overnight at -20° C, sample tubes were centrifuged at 12,000 rpm in a microcentrifuge at 4° C. The pellets were washed with 70% ethanol, dried and resuspended in 50 µl H_2O.

The DNA was amplified with Genomiphi kits (Amersham; Ø-29 DNA polymerase) for 18 hours in a thermal cycler, using multiple 20 µl reactions containing 50-100 ng of the isolated DNA as template. After amplification, the resulting DNA was purified with silica columns (Qiagen) to remove the enzyme, dNTPs, and primers, then ethanol precipitated and resuspended in H_2O to yield a DNA concentration of ~0.3 mg ml^{-1}. DNA samples (~10 µg each) were sequenced using pyrophospate sequencing technology (454 Life Sciences, Inc, Branford, CT).

454 pyrosequencing

Potential errors associated with 454 pyrosequencing: There are two main concerns associated with pyrosequencing: 1) random errors, where an incorrect base is substituted for a correct base, and 2) systematic errors, due to homopolymeric runs (i.e., runs of the same base). Since none of the marine virome sequences were known, the random error rate can not be directly determined from the data. It is assumed that the error rate is approximately the same as other investigators are reporting from very deep coverage of known sequences (e.g. primers included in the sequence). In these cases the error rate seems to be much less than 1 incorrect base in 1,000 reads (Edwards, personal communication).

454 Life Sciences, Inc, assert that their sequencing technology is accurate up to at least 8 homopolymeric nucleotides. To test this assertion, and to estimate the effect of these errors on the sequence analysis performed here, the frequency of homopolymeric runs from 3 nt to 15 nt were calculated for each of the four marine viromes, a database of 510 complete phage genomes, and 20 complete microbial genomes (Figure S1).

In general the marine virome contained very similar numbers of homopolymeric tracts as the microbial genomes. For unknown reasons there appear to be less 9-mers through 13-mers in the completed phage genomes than in either the microbial genomes or the viral libraries sequenced here. No 14 nt homopolymeric tracts were found in any of the 510 complete phage genomes. Presumably the higher packing density of genes in phage genomes, and the decreased information contained in long homopolymeric tracts is selected against in these genomes. In contrast to the rumored problems with homopolymeric tracts, this analysis seems to demonstrate that there are about as many tracts in 454 pyrosequenced databases as in complete bacterial genomes, and in fact the Sargasso sample sequenced here appears to contain a few more of these tracts than other databases.

In total, 15,543 sequences containing homopolymeric tracts between 9 and 15 nt were found in the four libraries (Table S2). Therefore, less than 1% of the sequences contain a homopolymeric tract that would be susceptible to the compression error of concern with pyrosequencing. We therefore conclude that the errors associated with compression of consecutive nucleotides is negligible in comparison to the number of

sequences we have generated, and other researchers are demonstrating that the random error of pyrosequencing is significantly less than any other sequencing technology.

Figure S1. Frequency of homopolymeric tracts in the four marine viromes, the complete phage genomes, and twenty, randomly chosen microbial genomes. The tracts from 3 nt to 15 nt were counted and normalized to the number of bases in each sequence. One 3 nt tract is found approximately every 30 bp, while one 15 nt tract is found approximately every 10 million bp. The 510 complete phage genomes totaled 18,909,173 bp in length, and the microbial genomes totaled 22,110,123 bp in length. The lengths of the pyrosequenced libraries are given in the text.

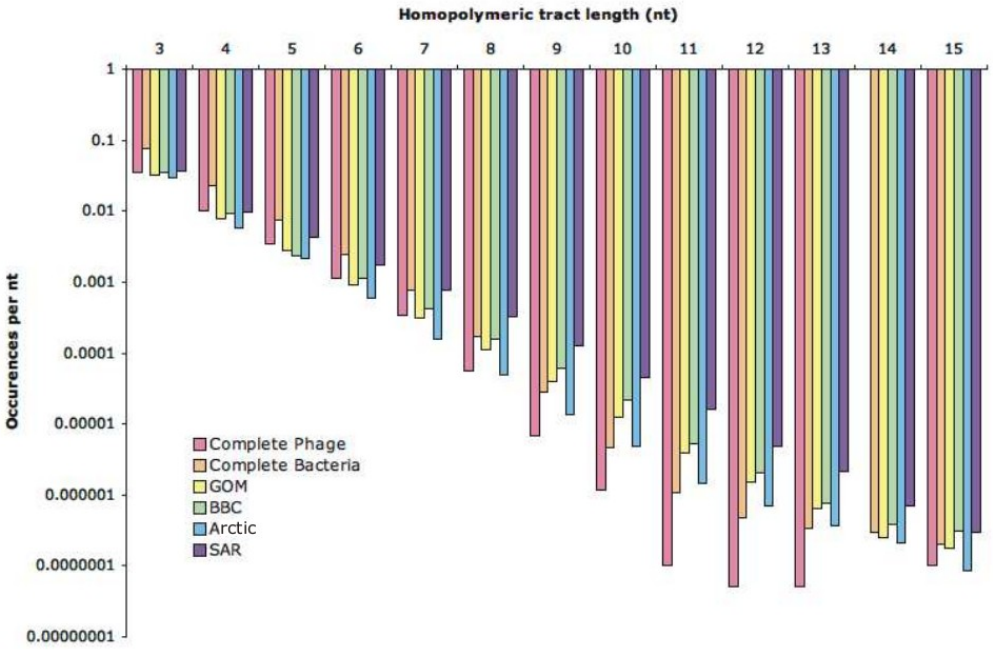


Table S2. Number of homopolymeric tracts, and number of sequences containing them for each of the four marine virome libraries.

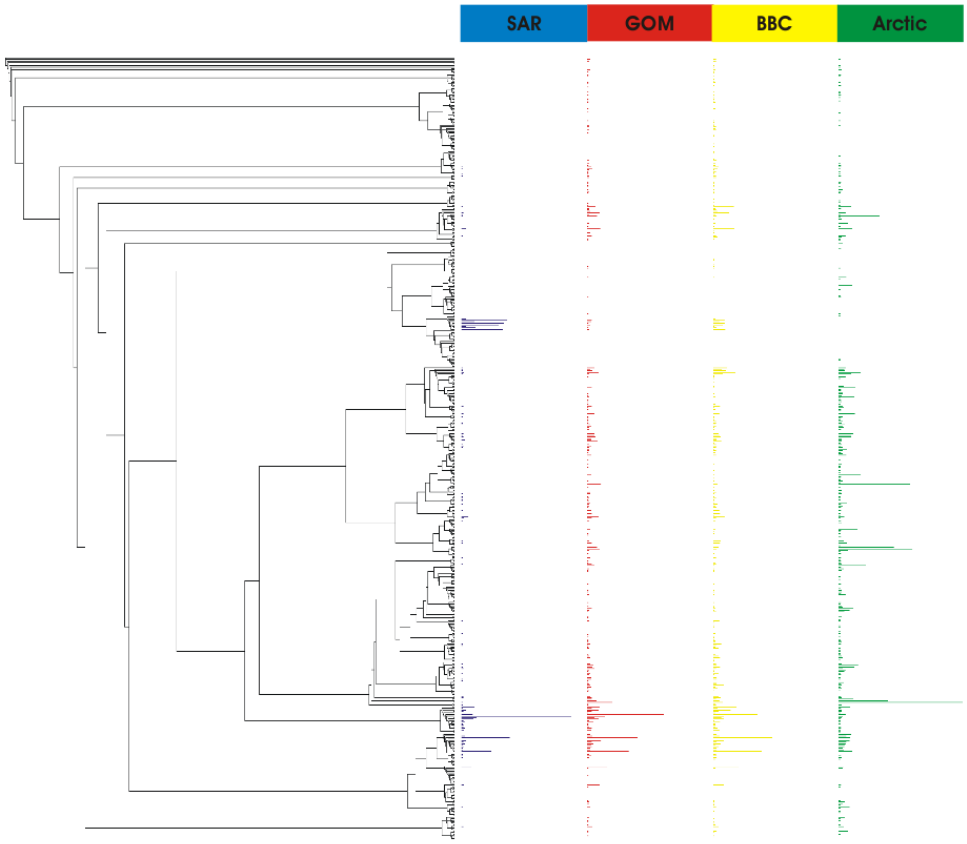
Library	*Number of tracts*	*Sequences with >1 tract*
GOM	1,659	1,650
BBC	4,053	3,832
Arctic	1,499	1,498
SAR	8,590	8,563
Total	15,801	15,543

Distribution of the marine viral sequences on the Phage Proteomic Tree

A new version of the Phage Proteomic Tree containing 510 phage genomes was constructed as described previously. All sequences were analyzed by TBLASTX against a database containing all of the completely sequenced phage genomes. Similarities with an E-value < 10^{-6} against this database (approximately equivalent to an E-value of 0.001 against the SEED nr database) were considered significant. For each sequence with a significant similarity to the phage genome database, the top TBLASTX similarity was recorded. To determine which phage phylogenetic groups were seen in each of the marine samples, each genome with at least one top significant similarity in a particular sample was marked with a solid line in the corresponding column next to its position on the Phage Proteomic Tree. A version with the relative abundance of the phages was also constructed (Figure S2). In that case, the length of the bars is proportional to the relative abundance of the phage species in the community.

Figure S2. Relative abundance of phages in the 4 metagenomes.

Note that due to the way the samples were stored and the long storage time, the distribution shown may not accurately reflect the reality.



Analyses of the chp1-like Microphage

Please note: The SAR chp1-like Microphage is a consensus sequence constructed from the SAR metagenome. The input sequences actually represent a group of closely related viruses.

>SAR chp1-like Microphage

CCCAGCGTGCTGGGTTTAATCTCATGTCGTAGTGTAGTAAAGGTATCGTACCTTCCAAGCCACCGTTAGGAATAACATT
CTGTTCCATCATACCAAGGTATTCTGGACGTTGTAAACGTGCGTCTGGTGAGGTCACTCCGAAATGTGATTGTAATATTT
CGGTGTATCTTGTACCGCCTCGCGCGTCTTTCTCATAAAGACGTTGAATTTGAAACGCTTCGCGGAGTTCGTTAATTGTT
GCAGCTGCTGCATCTGTAAGATCTGCTGTAAAGTTGAAATCTTGCGCTGTTGCAAGGTTAGCTGCGCCTCCACCTGGGA
AAGTATTGGCGTAATAAAAAATCTTGGTTGTTAGATGCATCGTTATACATTGCGATGTATTTTCCATCGCCGGTAATTGA
TGACCCGCCAACTGCTGAGTAGGTAATTGGTGCTTCCGTTCCCAGTGGTAGTGTTACTGCATCGCCCTTTTGAGGGAAT
GGTAAACTGCTGGTAAAATAATCGTGGCGTTTACCACGCTTTTGTAGTGTGTAAGTTGTGTAAGTGTCCGGGCCATCGC
CTTTGTCTACTGTTAAGCTATCTTGTAGGTTTTTCGTCACGAAACCACTCATTCCATATAAGGTTATAAGCTCGTCCGTG
TAAGTTATTAAAATCTATACCAGCAATTTGCGTGGGAAGTCCCATGTAATCAAAAAGGCTATTTTCTGCCACCGTAGCT
CCGGTAATTTGAGGTACTAGAAAGTCTGTGCTATCGCCTGGATCGTCTTGTGCCCCGTTAAACTTTTCCCAATTGTCCCA
GATCAATCTATTGGGGACAAAGAAAAAGGAATGTCTCAACATACATATTATCCATAATTGGATATATCGGCGTTGCTAA
TCGACCAAAACCCTGTTGCGTTCATTTGAAAAGTATCGCCTGGTAGAACTTCGTCTACATAGATAGGGACCAGGTACCC
TGAATCGAATGTTGTCTTAAGGCCGTGCACACGGTTAAAGGTACTACGTTGAATATCCGCTTGTGGTACTCTGCTAAAT
TCGTGTGTAAGTGTAGTGGGTAGGCTACCCATTGGTCCGCCGAGCATATTAGTCTCCTAATGTTTCCATCTCKATAATTT
TCTTTGGTTTTTCCTGACCGGTAATTATTCCGGTCGTTTCGTCAAAGCTACCCAATCTGTGAAGCGAAAAATCGCTAGGG
TGCTTTGCGAATGCGTGATCCTTATTGTTGATCACTATGTCTTGAACCGCTCTTACTGCGGTTCCATCTTTAATCTCTAGA
AAGGGTTGTGAGTACATCTCAGCCTTTCTGTCATATACTGCGTAATAAACTTTCTTCATYTTTTCCTCCCGTGGAATATT
GTTACGGGAGAATCTTACGCATAATATACAATAGACGTCAATAGTTTATGTAACTGTTTGTTTGGACTCTTGTTACCCTG
TAAATGATTCATTTTATGACATTTTACAGGTTTCGAACTAATCGTTCGAGTTTTTTTATTTTTATTTCTTCTGACACCCAG
AGGTCATCCATCGCCTTTATTATATTCAATTATAGTCTCTGGCGCCTGTTCTTTTCGCTTCTCTTTCAGCTGCTCAAAGTA
TTCGGGATCGTGTTTCTGTAGTTCCTTATCGTAATACCTAGGAACTTTCATTTTTATACCATCGTGAACGATGTAATCGT
GCAAATGTGCATCTGTCCATCCGTATTTCCAATACCATTGATTTCCGATCCCGTTTTGAGGTTTTTKGTTTATTTCCACGC
GACATTGTCGCGTATTGGTTATCGAGATCGTATTCGATCTGACCTGTTTCGGGGTTTATATATTGCTCAGGGGGGCCCTC
CCCTTTCGCTCTTTTCATTACATACCGTGCCACATAATGGGCACTTTCGTATGTACACGCCCCAATTCTGTGGTAGCCGT
GGGGCCACAGTTCTTCTAATTCGGGTGATATATATAATTCGTTACCTAGTTTTTTTTCCCATAATTGTTTGTCTGGAAAAT
CATACCCGAATATTATTGCATGATAGTGGGGGCGTTTGTTTTCATCACCATATTCTCCGCAGTGAAAGAACTTAATGTCT
TTTCCTTTTTTTTTGCGGAGCCGTTTCAAAAATCTCTGAAACTCGGTGATGTCCAGAGACCAAGGGCGAGGGCGCTGTTC
AAGGGTCTCTGGGTTTATTGTTAAGGTTATGAAGCAATTGTGTTCGTGCATCTGGGCTTCATGCATACATCTGATAGCCC
ATTCACGACTGTGTTGCAGTCGCAACCCCAGCATTGACCACATGGAAGAATTAAAAGCCCTTTGCATATGCAAAGGGCT
TTAATCTTCCCTGTGGTCAGTGCTGGGGTTGCAGACTGCAACACAGTAGAGAATGGGCTATCAGATGTATGCATGAAGC
CCAGATGCACGAACACAATTGCTTCATAACCTTAACAATAAACCCAGAGACCCTTGAACAGCGCCCTCGCCCTTGGTCT
CTGGACATCACCGAGTTTCAGAGATTTTTTGAAACGGCTCCGCAAAAAAAACAGAAAAGGACATTAAGTTCTTTCATTG
CGGAGAATATGGTGATGAAAACAAACGCCCCCACTATCATGCAATAATATTCGGGTATGATTTCCAGACAAACAATTA
TGGGAAAAAAAAACTAGGTAACGAATTATATATATCCCCCGAATTAGAAGAACTGTGGCCCCATGGCTACCACAGAAT
TGGGGCGTGTACATACGAAAGTGCCCATTATGTGGCACGATATGTTATGAAAAGAGCGAAAGGGGAGGGGCCCCCCTG
AGCAATATATAAACCCCGAAACAGGTCAGATCGAATACGATCTCGATAACCAATACGCGACAATGTCGCGTGGAAATA
AACAAAAAAACCTCAAAACGGGATCGGAAATCAATGGTATTGGAAATACGGATGGACAGATGCACATTTGCACGATTAC
ATCGTTCACGATGGTATAAAAATGAAAGTTCCTAGGTATTACGATAAGGAACTGGAAAAATACGATCCTGAATACTTTC
AGGAATTGAAAGCGAAGCGGAAAGAACAGTCACCAGAGACTATAATAGAATATAATAAGGCGATGGATGACCTCTGG
GTGTCAGAAGAAATAAAAATAAAAAAACTCGAACGATTAGTTCGAAACTTGTAAAATGTCATAAAATGAATCATTTAC
AGGGTAACAAGAGTCCAAACAAACAGTTACATAAACTATTGACGTCTATTGTATATTATGCGTAAGATTCTCCCGTAAC
AATATTCCACGGGAGGAAAAGATGAAGAAAGTATATTACGCAGTGTATGACAGAAAAGCAGAGATGTATTCACAGCCT
TTTCTAGAGATAAAAGACGGTACAGCAATAAGGGCTGTTCAGGACATAGTAATCAACAGTAAAGACCATGCGTTCGCA
AAACATCCCAGAGATTTCACATTATTCAGACTGGGTGAATTTGACGAAACGACAGGCGTAATAACCGGACAGGATAAA
CCGAAACAGATCATAGAGATTGAAACACTTGGAGAGTTAAAAAATGCTAGGCGGACCAATGGGCACCCTGCCCACCAC
ATTATCACACGAATTCTCACGCGTACCTCAAGCAGATATTCAACGTAGTACCTTTAACCGTGTACACGGGCTTAAAACA
ACATTCGATAGTGGATACTTGGTTCCGATATTCGTCGACGAAGTTCTCCCCGGCGATACGTTTCAATGTAGCGCGACGG
GCCTTTGGTCGCCTTTCAACTCCTCTCTACCCAGTAATGGATAACATGTATGTAGAAACATTCTTTTTCTACGTCCCAAA
TCGTATTATCTGGGACAACTGGGAGAAACTCAACGGTGCACAGGATGATCCGAACGACAGTACAGATTTTCTGGTTCCC
CAAATACAATCGGCAACAATAGCTCAGGATACTCTTTTCGATTATATGGGACTTCCCACCAAGACAGCAGGTTTGAACT
TTAACAACCTGCACGGTAGAGCATACAACCTCATCTGGAACGAATGGTTCCGAGATGAAAATTTACAGGATTCCCTAGT
AGTAGATAAGGACGATGGCCCTGACACTTTAACAGATTATACACTACAAAAAACGTGGTAAAAGACACGATTATTTTA
CCTCTGCCCTACCATGGCCTCAGAAAGGCGATGCAGTAAACCTACCACTCGGAACATCTGCTCCAGTAGCAACGGATTC
CGCAGATGGTGAAAACATAGCAGTATATTCAACAGGATTAGGCGGCTATACCAATATGGCGGCGAATGGAACCTTTGT
GGAAAACCCGTTCGGCGGTGGAACCGAAGACCGCTCACTATATGCCGACCTAACAGATGCAACAGCAGCAACAATCAA
CGAATACGCGAAGCGTTTCAAATCCAGAGACTTCTGGAGCGTGACGCTAGGGGCGGCACAAGATATACCGAGATTTTA
CAATCCCATTTTGGAGTAACCTCACCAGACGCCGCTTACAGCGTCCGAGTATCTCGGCGGCTCAAAAACAGAAATAAA
CATGCAGCCAATTCCACAGACTGGTTCAACAGACAGTACATCTCCTCAAGGTAACCTAGCAGCAATAGGTACAGCATC
ATCCAGAGGCGGATTTTAATAAGTCTTTTGTAGAACATGGTGTAATTATCGGAATGGCATGCGTATTTGCAGACTTAAC

TTATCAACAAGGGTATGAACCGTATGTGGTCACGTCGTGACCGCTGGGACTTTTATTGGCCAGCTCTCGCCCATTTAGG
CGAACAAGCGGTTCTAAACCAAGAAATCTATTATCAAAACACTTCAGCGGATTCCCAGACCTTTGGCTATCAGGAACGC
TGGGCAGAATATAGATATAAACCAAGCCAGATCACTGGCAAAATGCGTTCGAACGCAACAGGCACCTTTAGACGTATG
GCACTTGGCACAGGATTTCTCCTCGCCTGCCGGCACTCAACTCTTCATTCATCGAAGAAAACCCACCCATCGATCGGGT
TATCGCAGTAACCGACGAACCACAATTCATCTGGGACTGGTACTTCGATCTTAAATGTACAAGACCAATGCCTGTTTAT
TCAGTACCAGGCTTAATCGATCACTTCTAGGTGCAATATGAATGGATTCAAGTGGACCATTATTCTTGGCGTTCTTCGCA
AGTATGC

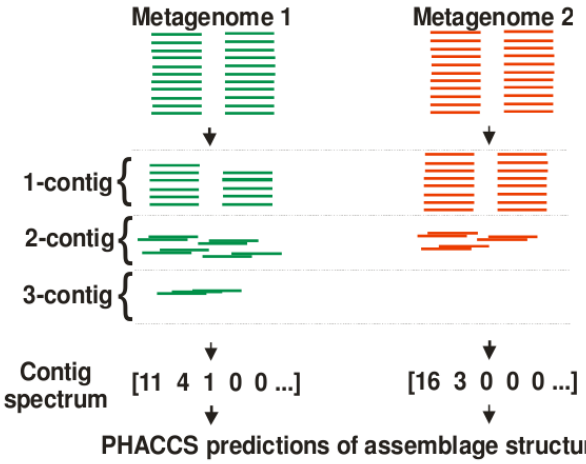
>chlamyd4 (Chlamydia phage 4; AY769964.1)
ATGGTTAGGAATCGGCGTTTGCCTTCAGTTATGAGTCATTCTTTCGCGCAAGTCCCATCAGCGCGAATTCAGAGAAGTT
CTTTTGATAGATCTTGTGGTTTAAAAACTACATTCGACGCCGGTTACCTAATCCCTATCTTTTGTGATGAAGTTCTCCCT
GGAGATACTTTCTCCTTGAAAGAGGCGTTTTTAGCACGTATGGCAACGCCTATCTTTCCTCTTATGGATAATTTGCGTTT
AGATACGCAGTATTTCTTTGTTCCTCTTCGACTATTATGGTCGAATTTTCAAAAGTTCTGTGGAGAACAAGATGATCCTG
GAGATTCTACAGATTTTCTTACCCCAATTTTGACCGCTCCTCAGAATGGTGGTTTTGCTGAAGGATCGATCCATGATTAT
CTTGGTCTACCTACTAAAGTTGCAGGAGTTCAATGTGTTGCGTTTTGGCACAGAGCTTACAATTTGATTTGGAACCAGTA
CTATCGTGATGAAAATATTCAGGATTCAGTTGAAGTGCAAATGGGAGATACCACTGCAGATGAAGTGAACAATTATAA
GCTTCTTAAGCGCGGGAAGCGTTATGATTATTTCACTTCATGTCTCCCTTGGCCACAAAAAGGTCCTGCAGTGACAATC
GGAGTTGGAGGTATTGTTCCTGTTCAAGGTTTAGGAATTCAATGGGGCGGTTCTACAGGTCCAAATCCTATAACTGCTT
CTGATTGGAGAGATTCCGTTAATCCTACATATGTAAATTCTGCAACGCAGACGCCTACAGGAACGAATAAGATTTTGAG
TTATGGTCAGGCGTATTATATTAAGAAGCCTGGAGAACCAGCTACAGATCCTGCACCTAGGGCTTATGTAGATTTAGGT
TCGACTTCTCCTGTGACGATTAATTCTCTTCGTGAAGCTTTCCAATTGCAAAAGCTTTATGAGAGAGATGCCCGTGGTGG
AACAAGGTACATTGAGATTATTCGTTCCCATTTCAATGTGCAGTCTCCAGATGCAAGGTTGCAACGTGCAGAGTATCTT
GGAGGTTCTTCAACTCCTGTGAATATTTCTCCGATTCCACAGACTTCCTCAACAGACTCCACATCTCCTCAAGGAAATCT
TGCTGCTTATGGTACAGCGATTGGATCGAAGCGAGTCTTCACAAAGTCCTTCACAGAACATGGTGTAATCCTTGGATTA
GCCTCTGTACGCGCCGATCTCAACTATCAGCAAGGTTTGGATAGGATGTGGTCACGAAGAACGCGCTGGGACTTTTACT
GGCCTGCTCTTAGCCATTTAGGTGAGCAAGCTGTGCTCAATAAAGAGATCTATTGCCAAGGTCCTGCAGTTAAGGATGC
TCAGAATGGCAATGTTGTTGTGGATGAGCAAGTCTTTGGATATCAGGAGAGATTTGCGGAGTATCGCTATAAGACTTCG
AAAATTACTGGCAAGTTCCGATCAAATGCTACAAGTTCTTTAGATTCATGGCATTTAGCTCAGGAATTTGAGAATCTTC
CAACACTTTCTCCGGAGTTTATCGAAGAAAATCCTCCTATGGATCGTGTTCTTGCTGTAAATACTGAGCCAGATTTTCTT
TTAGATGGCTGGTTTTCATTGCGTTGTGCAAGACCAATGCCTGTCTACTCTGTTCCAGGCCTCATTGATCATTTCTAATTT
CTACTCAGTTTTCCGATTTGATAAAGCAAACTCACGTTCGTAGATAAGTGAGTACGGTGAAGACCAAAACGGAAAGCT
GAGGCGTAAAAATGTGGAGAATTTATGAATCCCGAACAACTTACGAACACTCTCGGTTCAGCAGTTTCTGGAGTTGCGC
AAGGATTATCCTTTCTCCCTGGAATAGCTTCCGGAGTTTTAGGATATCTTGGTGCACAAAAGCAAAATGCCACTGCGAA
GCAAATTGCTAGAGAGCAAATGGCTTTTCAGGAGCGCATGTCTAACACGGCATACCAACGTGCCATGGAAGACATGAA
GAAAGCTGGCCTTAACCCTATGTTAGCTTTTTCTAAAGGCGGTGCTTCTTCTCCTGCAGGAGCGTCATGGTCTCCGAATA
ATCCTGTAGAAAATGCGATGAATTCTGGCCTTGCCGTGCAAAGACTTACTTACGAACGTAAGAAAATGCAGGCAGAGC
TTCAGAATCTTCGTGAGCAGAACCGTTTGATTAGAAATCAAGCAATACGTGAAGGCTATCTCGCAGAACGAGATAAAT
ATATGCGTGTTGCTGGAGTTCCTGTGGCCACTGAGATGTTAGATAAGACTTCTGGTCTTATCTCATCTTCAGCTAAGGCA
TTTAAGAATCTTTTTTCAAGAAAAGGAAGGTAGATGTTTAAGTCGGCATATTCCGAAAAAAAATCTGTAAAGATGAAGT
TCACACAGAAATCTTTGACGCAGCAACACAACAAAGATGAGTGTGATATTAACAACATCGTCGCAAAACTCAACGCTA
CAGGCGTTTTAGAGCACGTAGAGCGACGATCTCCACGTTATATGGACTGTATGGACCCTATGGAGTATTCCGAGGCTCT
AAACGTCGTTATTGAGGCTCAGGAGCAATTTGACTCTTTACCAGCCAAAATTCGTGAACGTTTTGGAAATGATCCAGAA
GCGATGCTCGATTTCTTGAGCCGTGAAGAAAATTATGAAGAAGCAAAGGCGTTAGGTTTTGTTTATGAAGATGGAACTT
CTGGAGCACCTCAAACATTTTTTGAAGCTGATCCTAAAGATGATCAAAATGTGGCAAACCAAGAACCTGGATTAGCCC
AAAAATGAGCAAATTTTGTGCAAAAAAGTGTGCAAAAAAATGTGCAAAAAATGGGCCAAAAATTGCCCCCAAAATCG
GAGCATTTTACGAGAGAAAAACACCAGCGTGTAACAGTCTTACTTGATCTGTTACACGCCTGGTGGTCGGAATTGTAAG
GAAATTTTTTAAAACTAAGCCCTATTTAGGGCCCAAAATTTAAGCTTAAAATGAGGTTAAAAAATGGCACGAAGATAC
AGACTTTCGCGACGCAGAAGTCGACGACTTTTTTCAAGAACTGCATTAAGAATGCATCGAAGAAATAGACTTCGAAGA
ATTATGCGTGGCGGCATTAGGTTTTAGTTTTGGATGTTAAGGAAATCTTTAAGGTTATGCTAAATTAGCTGCTATGTATA
ATTTGGCTCGTGACGAATGTATGTCATATTCGCACCGTTTACAATTACACAGCAGTTGAAGGCTTAGACGTTGATTTTTA
ATGTCTTAGCCTTCATTTTTGGTTTAGTGTGATTGCAAATGAGGTGCTCATGACGTGCATTTCTCCTTTTGTATGTTTTAT
AGATCCTTGTAACCAGCTCTGGTTTCCCAAAGGTGAGAAGTCTTCTAAACCTTGGGATAAAGTCCGTGAATTAAATGCT
TTTGAGCAAACGCAACCTGAAGAGTATCGAAAACGTTGGATTTTGATGCCTTGCCGTAGGTGCAAGTTTTGTAGAGTGC
AGAATGCAAAGATTTGGTCGTATCGTTGCATGCACGAAGCGTCTTTATATTCTCAGAATTGCTTTTTAACTTTGACTTAT
GAGGATCAGCATCTTCCAGAGAATGGTTCTCTGGTAAGAAATCATCCGACTTTGTTTCTTAGGCGATTGAGAGAGCACA
TTTCTCCTCATAAGATTCGTTATTTTGGATGTGGTGAATATGGATCGAAATTACAAAGGCCTCATTATCATCTTCTTATTT
ATAATTACGATTTTCCTGATAAAAAGCTCTTGAGTAAAAAGCGTGGCAATCCTCTCTTTGTTTCTGAGAAGTTAATGCA
GCTTTGGCCGTATGGATTCTCTACAGTGGGATCTGTAACGCGGCAAAGTGCAGGTTATGTAGCGCGCTATTCTTTGAAG
AAAGTGAGTAGAGATATTTCTCAAGATCATTATGGTCAAAGACTTCCGGAGTTTCTTATGTGTTCTCTTAAACCAGGAA
TAGGAGCGGATTGGTATGAGAAATATAAACGCGATGTCTATCCTCAGGATTATCTTGTTGTGCAAGATAAAGGGAAGT
CTTTTACGACGCGTCCTCCACGTTACTATGATAGCTACATTCTCGGTTTGATCCGGAAGAGATGGACGAGGTCAAACAA
AAACGTGTAGAGAAAGTCATGGCTTTGCCTGAGCTATCTCAGGATAAGGCTGAGGTGAAGCAATATATTTTCAATGACC
GTACGAAGAGACTCTTTAGAGACTATGAGGAGGAGAGTTACTAAACTTTTTTAAAAAATAGGAGCTTTTTTCAATGAAA
GTTTTTACAGTGTTTGATATTAAGACGGAAATTTATCAGCAGCCTTTTTTTATGCAGGCTACGGGAGCGGCAATCAGAG
CGTTTTCCGATATGGTAAATGAGGATCCTACAAAGAATCAATTTGCCGCGCATCCTGAAGATTACATTCTCTATGAGAT
TGGATCTTACGATGACTCTACTGGAACTTTCATTCCCTTAGATGTGCCTAAAGCCTTAGGAACAGGCTTGGATTTTAAGC
ACAAACAGTAGGGAAGAT

Global sample diversity and cross-contig simulation

For these analyses, 2500 random sequences were taken from each of the 4 metagenomes, totaling 10,000 sequences. This is the **mixed sample**. The sequences were assembled with TIGR_Assembler using a minimum of 98% identity over at least 20 bp and no sequence alignment error in 32 bp ("-g 1" argument).

A **contig spectrum** was determined by counting the number of *q*-contigs, where *q* is the number of fragments in any particular contig (Figure S3).

Figure S3. Determining a contig spectrum.



For a particular set of sequences, the average fragment length was 102 bp. All of the contigs of 5 or more sequences (i.e., q>4) only contained sequences from one library and the contig spectrum from the mixed sample was:

Mixed contig spectrum: [9474 130 26 13 7 5 2 0 2 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0]

q	#	Samples that each contig came from
15	1	1 GOM – i.e., all 15 sequences that assembled were from GOM
14	1	1 GOM
10	1	1 GOM
9	2	1 BBC, 1 SAR
7	2	1 GOM, 1 BBC
6	5	2 GOM, 1 BBC, 2 SAR
5	7	4 GOM, 3 SAR

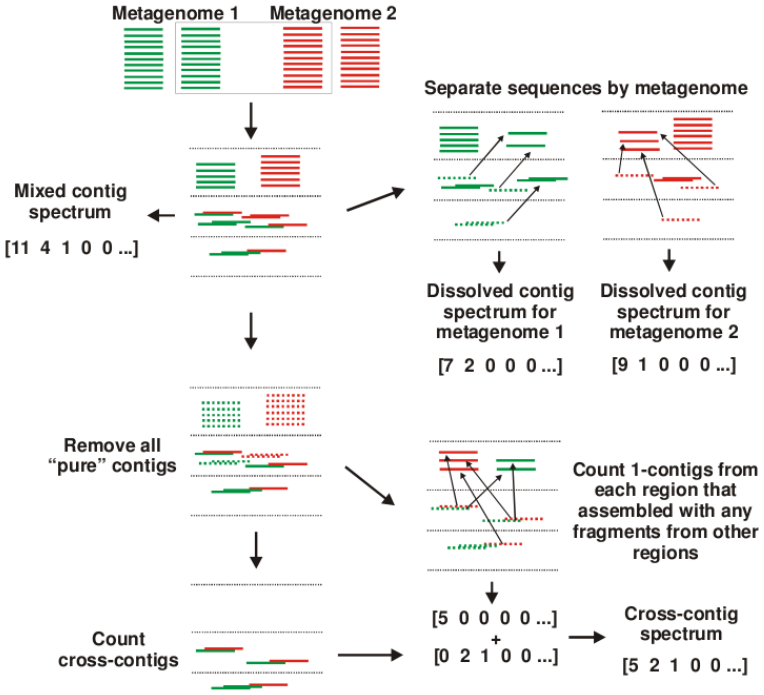
4	13	5 GOM, 3 BBC, 4 SAR, 1 cross
3	26	11 GOM, 1 Arctic, 1 cross
2	130	36 GOM, 10 BBC, 24 Arctic, 41 SAR, 19 cross
1	9474	singletons: 2297 GOM, 2446 BBC, 449 Arctic, 2324 SAR

q = number of fragments in each contig (size of the contig)

= number of contigs

To determine a **cross contig spectrum**, only sequences that assembled with sequences from other regions were kept. The number of *q*-cross-contigs was then counted as the number of remaining contigs of *q* sequences. The total number of singletons (1-contigs) from each region that assembled with any fragments from other regions was the number of 1-cross-contigs. The method to determine this cross-contig spectrum is represented in Figure S4. In the example above, the cross-contig spectrum was: [42 19 1 1 0]

Figure S4. Getting a cross contig spectrum.



Dissolved contig spectra were calculated for each separate metagenome by determining how many of the contigs came from only one metagenome. The dissolved contig spectra were not used in the manuscript except as a check on the methodology (i.e., they should be similar to contig spectra obtained when assembling individual metagenomes).

After repeating the process 10 times to get a better coverage of the metagenomes, the resulting contig spectra were averaged yielding the following:
Average mixed contig spectrum: [8870.1 227.5 49.9 23.4 11.8 7.4 4.3 3.2 2.8 1.3 1.7 1.2 0.7 0.5 0.9 0.4 1 0.4 0.6 0.4 0.3 0.2 0.1 0.2 0.1 0.1 0.2 0.2 0.1 0.2 0.2 0.3 0.2 0 0.3 0.1 0 0.1 0.3 0.2 0 0 0.1 0 0 0 0 0.2 0.1 0 0 0 0 0 0 0 0 0 0]
Average cross-contig spectrum: [48.9 23.5 1.4 0.2]

To estimate community structure and diversity, the averaged mixed contig spectrum was analyzed using PHACCS (http://biome.sdsu.edu/phaccs) using the following parameters for the example above: 102 bp for the average fragment size, an average genome length of 50 kb, and looking for up to 100,000 genotypes. The results are presented in Table S3.

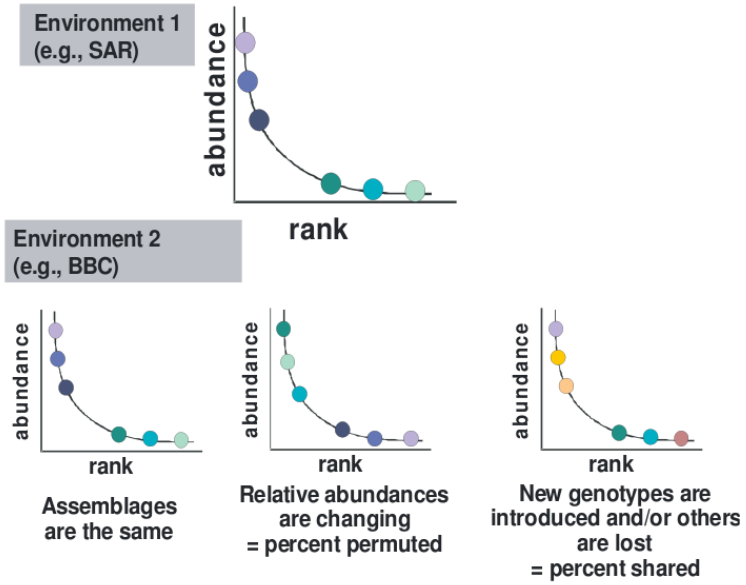
Table S3. Example of PHACCS output using the average mixed contig spectrum mentioned above. The best fit (lowest error) in this example was for a logarithmic distribution of the genotypes.

	Error	Richness	Evenness	% most abundant	Shannon (nats)
Power law	4560.1	100,000+ *			
Exponential	26,208	10,001	NaN	8.3849	NaN
Logarithmic	2324.8	57,572	0.89481	9.3394	9.8078
Lognormal	3906.3	100,000+ *			
Niche premption	26,208	10,001	NaN	8.3849	NaN
Broken stick	20,095	53	0.89884	8.5979	3.5687

* 100,000+ means that the best parameters for the tested distribution were not found by PHACCS using the specified input parameters.

The **Monte Carlo simulation** was used to determine whether differences between observed viruses within a community are due to changes in their relative rank (i.e., the abundance they make in the community) or because they are fundamentally different viruses (illustrated in Figure S5).

Figure S5. The possible scenarios considered in the Monte Carlo simulation to explain the observed cross contigs.



The average cross-contig spectra were then compared with simulated average cross contig spectra from simulated mixtures of the four communities (Figure S6).

Appendix 3: GAAS

The GAAS Metagenomic Tool and its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes.

Florent E. Angly, Dana Willner, Alejandra Prieto-Davó, Robert A. Edwards, Robert Schmieder, Rebecca Vega-Thurber, Dionysios A. Antonopoulos, Katie Barott, Matthew T. Cottrell, Christelle Desnues, Elizabeth A. Dinsdale, Mike Furlan, Matthew Haynes, Matthew R Henn, Yongfei Hu, David L. Kirchman, Tracey McDole, John D. McPherson, Folker Meyer, R. Michael Miller, Egbert Mundt, Robert K. Naviaux, Beltran Rodriguez-Mueller, Rick Stevens, Linda Wegley, Lixin Zhang, Baoli Zhu, Forest Rohwer

PLoS Computational Biology. 2009. In press.

The GAAS Metagenomic Tool and its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes

Florent E. Angly [1 2] *, Dana Willner [1], Alejandra Prieto-Davó [1], Robert A. Edwards [1 3 4], Robert Schmieder [2 3], Rebecca Vega-Thurber [6], Dionysios A. Antonopoulos [5], Katie Barott [1], Matthew T. Cottrell [7], Christelle Desnues [8], Elizabeth A. Dinsdale [1], Mike Furlan [1], Matthew Haynes [1], Matthew R Henn [9], Yongfei Hu [10], David L. Kirchman [7], Tracey McDole [1], John D. McPherson [11], Folker Meyer [4], R. Michael Miller [5], Egbert Mundt [12], Robert K. Naviaux[13], Beltran Rodriguez-Mueller [1 2], Rick Stevens [4], Linda Wegley [1], Lixin Zhang [10], Baoli Zhu [10], Forest Rohwer [1]

[1] Biology Department, San Diego State University, San Diego, CA, USA
[2] Computational Science Research Center, San Diego State University, San Diego, CA, USA
[3] Computer Science Department, San Diego State University, San Diego, CA, USA
[4] Mathematics and Computer Science Division, Argonne National Lab, Argonne, IL, USA
[5] Biosciences Division, Argonne National Laboratory, Argonne, IL, USA
[6] Biology Department, Florida International University, Miami, FL, USA
[7] School of Marine Science and Policy, University of Delaware, Lewes, DE, USA
[8] URMITE, Centre National de la Recherche Scientifique UMR IRD 6236, Université de la Méditerranée, Marseille, France
[9] The Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA
[10] CAS Key laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
[11] Ontario Institute for Cancer Research, MaRS Centre, Toronto, ON, Canada
[12] Poultry Diagnostic and Research Center, College of Veterinary Medicine, The University of Georgia, Athens, GA, USA
[13] School of Medicine, University of California of San Diego, San Diego, USA

* Corresponding author

Running title: Genome relative Abundance and Average Size

Author summary:

Metagenomics uses DNA or RNA sequences isolated directly from the environment to determine what viruses or microorganisms exist in natural communities and what metabolic activities they encode. Typically, metagenomic sequences are compared to annotated sequences in public databases using the BLAST search tool. Our methods, implemented in the Genome relative Abundance and Average Size (GAAS) software, improve the way BLAST searches are processed to estimate the taxonomic composition of communities and their average genome length. GAAS provides a more accurate picture of community composition by correcting for a systematic sampling bias towards larger genomes, and is useful in situations where organisms with small genomes are abundant, such as disease outbreaks caused by small RNA viruses.

Microbial average genome length relates to environmental complexity and the distribution of genome lengths describes community diversity. A study of the average genome length of viruses and microorganisms in four different biomes using GAAS on 169 metagenomes showed significantly different average genome sizes between biomes, and large variability within biomes as well. This also revealed that microbial and viral average genome sizes in the same environment are independent of each other, which reflects the different ways that microorganisms and viruses respond to stress and environmental conditions.

Abstract

Metagenomic studies characterize both the composition and diversity of uncultured viral and microbial communities. BLAST-based comparisons have typically been used for such analyses; however, sampling biases, high percentages of unknown sequences, and the use of arbitrary thresholds to find significant similarities can decrease the accuracy and validity of estimates. Here, we present Genome relative Abundance and Average Size (GAAS), a complete software package that provides improved estimates of community composition and average genome length for metagenomes in both textual and graphical formats. GAAS implements a novel methodology to control for sampling bias via length normalization, to adjust for multiple BLAST similarities by similarity weighting, and to select significant similarities using relative alignment lengths. In benchmark tests, the GAAS method was robust both to high percentages of unknown sequences and to variations in metagenomic sequence read lengths. Re-analysis of the Sargasso Sea virome using GAAS indicated that standard methodologies for metagenomic analysis may dramatically underestimate the abundance and importance of organisms with small genomes in environmental systems. Using GAAS, we conducted a meta-analysis of microbial and viral average genome lengths in over 150 metagenomes from four biomes to determine whether genome lengths vary consistently between and within biomes, and between microbial and viral communities from the same environment. Significant differences between biomes and within aquatic sub-biomes (oceans, hypersaline systems, freshwater, and microbialites) suggested that average genome length is a fundamental property of environments driven by factors at the sub-biome level. The behavior of paired viral and microbial metagenomes from the same environment indicated that microbial and viral average genome sizes are independent of each other, but indicative of community responses to stressors and environmental conditions.

Introduction

Metagenomic approaches to the study of microbial and viral communities have revealed previously undiscovered diversity on a tremendous scale [1,2]. Metagenomic sequences are typically compared to sequences from known genomes using BLAST to estimate the taxonomic and functional composition of the original environmental community [3]. Many software tools designed to estimate community composition (e.g. MEGAN) annotate sequences using only the best similarity [4]. However, the best similarity is often not from the most closely related organism [5]. In addition, most metagenomes contain a large percentage of sequences from novel organisms which cannot be identified by BLAST similarities, further complicating analysis [1,6,7].

Mathematical methods based on contig assembly have been developed to estimate viral diversity and community structure from metagenomic sequences regardless of whether they are similar to known sequences [8]. These similarity-independent methods require the input of the average genome length of viruses from a given sample [8]. Having an accurate value of this average is important because it takes a potentially large range spanning 3 orders of magnitude, and has a large influence on the diversity estimates. Average genome length for an environmental community can be determined using Pulsed Field Gel Electrophoresis (PFGE) [9,10]. PFGE gives a spectrum of genome lengths in a microbial or viral consortium, indicated by electrophoretic bands on an agarose gel, which can be used to calculate an average genome length. Due to the large variability of dsDNA virus genome length, PFGE can discriminate and identify dominant viral populations [11]. However, PFGE is limited because the bands are not independent and a single band can contain different DNA sequences [12,13].

Average genome length in environmental samples has also been used as a metric to describe

community diversity and complexity [9,14-17]. In PFGE, both a larger size range and a greater number of bands indicate a wider variety of genomes and hence, a more diverse community [9,14,16,17]. The average genome length of a microbial community has been shown to serve as a proxy for the complexity of an ecosystem [15]. Longer average genome lengths indicate higher complexity [15], since larger bacterial genomes can encode more genes and access more resources [18].

Here we introduce Genome relative Abundance and Average Size (GAAS), the first bioinformatic software package that simultaneously estimates both genome relative abundance and average genome length from metagenomic sequences. GAAS is implemented in Perl and is freely available at http://sourceforge.net/projects/gaas/. Unlike methods that rely on microbial marker genes to estimate genome length, the GAAS method can be applied to viruses, which lack a universally common genetic element [19]. GAAS determines community composition and average genome length using a novel BLAST-based approach that maintains all similarities with significant relative alignment lengths, assigns them statistical weights, and normalizes by target genome length to calculate accurate relative abundances. Using GAAS, the community composition and average genome length for over 150 viral and microbial metagenomes was derived from four different biomes, including the Sargasso Sea virome previously described in Angly et al. [1]. The average genome lengths were used in a meta-analysis to determine how genome length varies at three levels: between biomes (e.g. terrestrial versus aquatic), between related sub-biomes (e.g. ocean versus freshwater), and between microbial and viral communities sampled from the same environment.

Results and Discussion

Accuracy of GAAS estimates

GAAS provided more accurate estimates of average genome length and community composition than standard BLAST searches (i.e. no length normalization, no relative alignment length filtering, top BLAST similarity only) (Figure 1). The accuracy of GAAS estimates was benchmarked using artificial viral metagenomes. To simulate environmental metagenomes, 80% of species were treated as unknowns and viral communities were created with either power law or uniform rank-abundance structures. The error for power law metagenomes was consistently higher than for the uniform case (data not shown). Significance of BLAST similarities was determined using relative alignment length and percentage of similarity in addition to an E-value cutoff. The accuracy of GAAS was dramatically increased by normalizing for genome length; average errors decreased significantly for community composition ($p<0.001$, Mann-Whitney U test), as well as genome length ($p<0.001$, Mann-Whitney U test) (Figure 1 A, B). Metagenomes consist of sequence fragments derived from the available genomes in an environment [20]. Even if two genomes are present in equal abundances, a larger genome has a higher probability of being sampled because it will produce more fragments of a given size per genome (Figure S1). Length normalization in GAAS corrected for this sampling bias inherent to the construction of random shotgun libraries such as metagenomes. Using all similarities weighted proportionally to their E-values further reduced errors in composition. This reduction was significant in comparison to average error when only the top BLAST similarity was used ($p<0.001$, Mann-Whitney U test) (Figure 1 C). When no species were treated as unknown, the error on the GAAS estimates decreased dramatically (Figure S2). GAAS performed well in benchmarks using artificial microbial metagenomes obtained from JGI (Figure S3). Figure S4 shows that it is harder to distinguish between

closely related strains than unrelated species using local similarities: the error on the relative abundance estimates is higher than for more distantly related microorganisms (Figure S3). However, GAAS improves both estimates of relative abundance and average genome length, from ~2% relative error for the average genome size when keeping only the top similarity to ~0.2% using all similarities and weighting them (Figure S4).

Read length does not matter for GAAS

Variations in metagenomic read lengths did not affect the accuracy of GAAS relative genome length estimates (Figure 2, Figure S5, Figure S6). GAAS was benchmarked on simulated viral metagenomes containing 50, 100, 200, 400, or 800 base pair sequences. Read length had no effect on the accuracy of average genome length estimates ($p=0.408$, Kruskal-Wallis test). Average errors in composition increased significantly ($p<0.001$, Kruskal-Wallis test) with increasing read length, but there was only a very weak positive correlation between increased errors and longer reads ($tau=0.07$, $p<0.001$). The accuracy of GAAS estimates was thus not very susceptible to changes in read length on average. This contrasts with a report on the inappropriateness of short reads for characterizing environmental communities, mainly on the basis that they miss more distant homologies than longer sequences [21]. In addition, the longest reads tested here (800 bp) achieved both the lowest and highest error on the relative abundance estimates (Figure S5). This indicates that the choice of appropriate filtering parameters is more important for longer sequences than for short sequences. In summary, GAAS can be used to accurately and effectively estimate both composition and average genome length for sequences from a variety of available technologies: very short (~50 bp) sequences obtained by reversible chain termination sequencing (e.g. Solexa), mid-size sequences produced by Roche 454 pyrosequencing (~100-400 bp), and long 700+ bp reads sequenced by synthetic chain-terminator

chemistry (Sanger).

Re-analysis of the Sargasso Sea Virome

Re-analysis of the Sargasso Sea virome using GAAS revealed that small ssDNA phages were more important than previously assessed, representing ~80% of the viral community (Figure 3). Community composition and average genome size for the Sargasso Sea virome were calculated using both the GAAS method and the standard method (no length normalization, top similarities only) for comparison. Both the pie charts and length spectra in Figure 3 were generated directly by GAAS. Using the standard method, the Sargasso Sea viral community was dominated by *Prochlorococcus* phages (64%), with lesser abundances of *Chlamydia* phages (15%), *Synechococcus* phages (12%), *Bdellovibrio* phages (3%) and *Acanthocystis chlorella* viruses (2%). In contrast, using GAAS, *Chlamydia* phages were the most abundant organism (79%), whereas *Prochlorococcus* phages only comprised 16% of the community. The presence of *Chlamydia* phages in the Sargasso Sea was previously verified experimentally using molecular methods [1]. In contrast to the standard method, the GAAS method also indicated very low relative abundances (<1%) of *Synechococcus* phages and *Chlorella* viruses, which have larger genomes.

Most of the variations in community composition estimates were explained by differences in viral genome lengths (Figure 3, right panel). The corrected relative abundance estimates provided by GAAS indicated that species with larger genomes were less abundant than previously thought, and that normalizing by genome length was essential for accurate estimation of community composition (as shown in benchmark tests, Figure 1). A lack of normalization could lead to poor and possibly misleading community composition estimates, as our results have shown, since relative abundance does not equal percentage of similarities.

Phages with small genomes (20-40 kb) are believed to be the most abundant oceanic viruses [11]. In the re-analysis of the Sargasso Sea metagenome, GAAS estimated that 80% of the viral particles were *Microviridae* (mainly *Chlamydia* phages), viruses with a genome size smaller than 10 kb. Multiple Displacement Amplification (MDA) was used during the preparation of the Sargasso Sea virome and could have led to over-representation of this viral family. Despite this potential bias, the *Chlamydia* phage content of this virome was still higher than in all viromes prepared with MDA (except for the stromatolite viromes [6]) (data not shown). In addition, diverse marine circovirus-like genomes, with a length of less than 3 kb, have also been reported in the Sargasso Sea [22], suggesting that small single-stranded viruses play important roles in this marine habitat.

Average genome length varies significantly between and within biomes

Both microbial and viral average genome lengths calculated by GAAS were significantly different between marine, terrestrial, and host-associated biomes (Figure 4A, Table S1, Table S2). Of the 169 metagenomes analyzed, 146 had a sufficient number of similarities for estimation of average genome length. The average for genome length across all aquatic viral metagenomes was consistent with the previous estimate of 50 kb for marine systems using PFGE by Steward et al. [9]. Host-associated and aquatic viromes had average genome lengths spanning a wide range, from 4.4 to 51.2 kb and from 4.6 to 267.9 kb respectively. Viral average genome lengths were significantly smaller in host-associated metagenomes than in aquatic systems ($p=0.002$, Mann-Whitney U test). Estimates of microbial average genome length for aquatic and terrestrial biomes were similar to those predicted using the Effective Genome Size (EGS) method [15], a computational technique based on finding conserved bacterial and archaeal markers in metagenomic sequences. Aquatic microbiomes also showed large variation in average genome sizes, ranging from 1.5 to 5.5 Mb for Bacteria and Archaea

and from 0.7 to 25.7 Mb for protists. Microbial average genome lengths in the terrestrial biome were significantly higher than in the host-associated and aquatic biomes ($p<0.0001$, Mann-Whitney U test). Genome lengths of Bacteria and Archaea from soil environments have previously been shown to be larger than those observed in other biomes [15]. A larger genome is characteristic of the copiotroph lifestyle [23] as it provides microbes a selective advantage in the complex soil environment where scarce but diverse resources are available [24].

Microbial and viral average genome lengths were also significantly different between aquatic sub-biomes. Aquatic metagenomes were grouped into five categories (ocean, freshwater, hypersaline, microbialites, and hot springs) to determine if the variation in average genome lengths could be accounted for by the influence of distinct sub-biomes (Figure 4B, Table S1, Table S2). Other biomes did not include enough metagenomes from different sub-biomes to allow for meaningful classification and analysis. While average genome lengths still varied over a range of values in sub-biomes, the variability was much lower than in the aquatic biome as a whole (Table S1). The average genome sizes in oceanic viromes varied from 20 to 163 kb, well within the range described in [17]. In hypersaline metagenomes, the average genome length varied from 51 to 263 kb, which is comparable to viral genome sizes detected in ponds of similar salinities [16]. A number of average genome lengths were significantly different between sub-biomes for both viruses and microbes (Figure 4B). The stromatolite metagenomes had an average genome length which was significantly different from the oceanic and hypersaline sub-biomes ($p < 0.05$, Mann-Whitney U test), but not from freshwater systems. Oceanic and hypersaline environments were not significantly different. In comparison with the biome level (Figure 4A), the range of average genome lengths at the sub-biome level was reduced (Figure 4B). This suggests that differences in average genome lengths may be driven by environmental factors at a more

specific level (e.g. the sub-biome) than what can be encompassed by general biome classifications. Previous work has demonstrated that both metabolic profiles and dinucleotide composition vary at the sub-biome level, and significant differences between both composition and metabolic functions have been reported for marine (ocean), hypersaline, microbialite, and freshwater environments [7,25].

Microbial and viral average genome lengths are independent

Microbial and viral average genome lengths varied independently of each other across biomes and aquatic sub-biomes, and reflected differences in the way microbial and viral consortia react to stressors and environmental conditions (Figure 5). Using GAAS estimates for average genome lengths, we compared 25 pairs of viral and microbial metagenomes sampled from the same environment at the same time point. Viral and microbial community compositions have been shown previously to co-vary [26], however, there was no consistent trend between microbial and viral average genome length across all biomes (Kendall's tau=-0.21, p=0.10).

Most viromes in this analysis were obtained by the collection of viral particles small enough to pass through 0.22 μm pore size filters. The four viral metagenomes collected using 0.45 μm filters [27] had a larger viral average genome length (in light blue in Figure 5). These data show that large viruses may be omitted when sampling with 0.22 μm filters and the capsid size of DNA viruses is likely positively correlated with their genome length. Sampling biases, however, do not account for the independence of viral and microbial length reported here.

Paired metagenomes from oceanic and hypersaline aquatic sub-biomes were characterized by small fluctuations in viral genome lengths coupled with large variations in microbial genome lengths. The four paired ocean metagenomes (Figure 5, light blue squares) were taken from waters surrounding coral atolls in the Northern Line Islands [27]. Microbial communities changed dramatically along a

gradient of human disturbance, with populations of pathogens and heterotrophic microbes increasing with human activity [27], which could have resulted in large differences in average microbial genome lengths between atolls. Across all four atolls, viral communities were dynamic but dominated in general by *Synechococcus* and *Prochlorococcus* phage, according to both the original [27] and the GAAS analysis (not shown). The large genome of these widespread phages resulted in a less variable viral average genome length. In hypersaline metagenomes (Figure 5, blue diamonds), a similar trend of low variation in viral genome lengths coupled with larger ranges of microbial genome lengths was observed. This corresponded to known differences in the ranges of genome lengths of dominant halophilic viruses and microbes. The most abundant viruses in hypersaline systems have genome lengths between 32 and 63 kb, while predominant Halobacteria have genome lengths varying across a larger range, from 2.6 to 4.3 Mb [28,29].

The relationship between viral and microbial average genome lengths in manipulated coral metagenomes reflected differences in how viral and microbial consortia reacted to stress (Figure 5, yellow triangles). Five of the six manipulated metagenome pairs used in this analysis were metagenomes from *Porites compressa* corals subjected to a variety of stressors [30,31]. Nutrient, DOC, temperature, and pH stress all resulted in an increased abundance of large herpes-like viruses over the control, which could lead to increased average viral genome lengths overall [30]. However, shifts in the microbial consortia (consisting of Bacteria, Archaea, and eukaryotes) were more variable depending on which stressor was applied [31]. For example, temperature stressed corals showed a dramatic increase in fungal taxa, which could be driving the larger average microbial genome length seen here.

Conclusions

The GAAS software package implements a novel methodology to accurately estimate community composition and average genome length from metagenomes with statistical confidence. GAAS provides the user with both textual and graphical outputs, including genome length spectra, relative abundance pie charts, and relative abundances mapped to phylogenetic trees. GAAS can easily be applied to any database of complete sequences to perform taxonomic or functional annotations, and provides filtering by relative alignment length as a standard for selecting significant similarities regardless of which database is used. Since GAAS controls for sampling bias towards larger genomes and considers all significant BLAST similarities, it has the potential to identify key players in ecosystems that may be ignored by other analyses. For example, the re-analysis of the Sargasso Sea virome indicated that small ssDNA phage were very abundant and may play a previously overlooked role in the oceanic ecosystem. GAAS could also be applied in metagenomic studies of disease outbreaks and epidemics. Many emerging and highly virulent human pathogens are ssRNA viruses with small genomes, which could be missed by standard analysis methods, which do not normalize for genome length. Meta-analysis using GAAS provided insight into how environmental factors may affect average genome lengths in microbial and viral communities and the relationships between them. The lack of covariance between microbial and viral average genome lengths indicates that natural and applied stressors have different effects on microbes and viruses from the same environment.

Materials and methods

GAAS: Genome relative Abundance and Average Size in random shotgun libraries

GAAS software package

GAAS was implemented as a standalone software package in Perl and is freely available at http://sourceforge.net/projects/gaas/. It accepts and produces files in standard formats (FASTA sequences, Newick trees, tabular BLAST results, SVG graphics). The GAAS methodology is described in detail below and is outlined in Figure 6.

Similarity filtering

BLAST analyses (NCBI BLAST 2.2.1) were conducted through GAAS in order to determine significant similarities between metagenomic sequences and completely sequenced genomes. Similarities were filtered based on a combination of maximum E-value, minimum similarity percentage and minimum relative alignment length. E-value filtering removed non-significant similarities, and the alignment similarity percentage and relative length were used to select for strong similarities likely to reflect the taxonomy of the metagenomic sequences. E-values depend on the size of the database and the absolute length of alignments between query and target sequences, and thus may not be comparable between analyses [32,33]. Relative alignment length, also called alignment coverage [34], is the ratio of the length of the alignment to the length of the query sequence (Figure S7). It is independent of the database size and sequence length, and provides an intuitive and consistent threshold to select significant similarities. Since the ends of sequenced reads can be of lower quality, similarities were kept only if the length of the alignment represented the majority of the length of the query sequence. Sequences with no similarity satisfying the filtering criteria were ignored in the rest of the analysis.

Similarity weighting

In order to avoid the loss of relevant similarities by reliance upon smallest E-values alone [5], all significant similarities for each query sequence (as defined by our criteria above) were kept and assigned weights as follows.

Based on the Karlin-Altschul equation, the expect value E_{ij} between a metagenomic query sequence i and a target genome sequence j is given by: $E_{ij} = m_i' n' e^{-S'_{ij}}$ where m'_i is the effective query sequence length, n' is the effective database size (in number of residues) and S'_{ij} is the high-scoring pair (HSP) bitscore [32]. Using the effective length corrects for the "edge effect" of local alignment and is significant for sequences smaller than 200 bp such as sequences produced by the high throughput Roche-454 GS20 platform. Assuming that a query sequence is more likely to have local similarities to longer target genomes, each of the E-values can be reformulated into an expect value F_{ij} of a similarity in a given target genome by: $F_{ij} = m_i' t_j' e^{-S'_{ij}} = E_{ij} t_j' / n'$ where t_j' is the effective length [35] of the target genome j. Using the length of the target genome in the F-value produces an expect value relative to the target genome, not to the totality of the genome database (as is the case of the E-value).

From F_{ij}, a weight w_{ij} can be calculated as $w_{ij} = z_i / F_{ij}$ with z_i being a constant such that for a given metagenomic query sequence i, $\sum_j w_{ij} = 1$. This weight carries the statistical meaning of the expect value of the similarity relative to the given genome in such a way that the larger the expect value, the lower the weight. Therefore, for a given query sequence i, the weight was calculated as

$$w_{ij} = \frac{z_i}{E_{ij} t_j'} .$$

Genome relative abundance using genome length normalization

The relative abundance of sequences in a random shotgun library is proportional not only to the relative abundance of the genomes in the library but also to their length. Similarly to the normalization used in proteomics [36-38], normalization by genome length is needed to obtain correct relative abundance of the species in a metagenome. For each target genome *j*, the weights w_{ij} to that genome were added to obtain W_j. The weighted similarities W_j to each genome were then normalized by the actual length t_j of the genome (including chromosomes, organelles, plasmids and other replicons) to obtain accurate relative abundance estimates: $W_j = x / t_j$ where *x* is a constant such that $\sum_j W_j = 1$.

Average genome length calculation

GAAS relies on the relatively stable genome size found within taxa [39] to calculate average genome length. The average genome length was calculated as a weighted average of individual genome lengths. The length of the genome for each individual organism identified in the metagenome was weighted by the relative abundance of that organism as calculated by GAAS. Thus, the mean genome length *L* was calculated as: $L = \sum_k r_k l_k$ where r_j was the relative abundance of organism *k*, and l_j its individual genome length.

Confidence intervals for relative abundance and average genome length estimates

A bootstrap procedure was implemented in GAAS to provide empirical confidence intervals for relative abundance and average genome length estimates. The estimation of community composition and average genome length was repeated many times using a random subsample of 10,000 sequences for each repetition. Confidence intervals were determined based on the percentiles of the observed

estimates, e.g. 5th and 95th percentiles for a 90% confidence interval.

Reference databases for viral, microbial and eukaryotic metagenomes

NCBI RefSeq (ftp://ftp.ncbi.nih.gov/refseq/release) (Release 32, August 31, 2008) was used as the target database for the estimation of taxonomic composition and average genome size. Three databases containing exclusively complete genomic sequences were created from the viral, microbial, and eukaryotic RefSeq files. All incomplete sequences were identified as having descriptions containing words such as "shotgun", "contig", "partial", "end" and "part", and were removed from the database.

A taxonomy file containing only the taxonomic ID of the sequences in these three databases was produced using the NCBI Taxonomy classification. Sequences with a description matching the following words were excluded from that file unless the chromosomal sequences were also available for the same organism: "plasmid", "transposon", "chloroplast", "plastid", "mitochondrion", "apicoplast", "macronuclear", "cyanelle" and "kinetoplast". The complete viral, microbial, and eukaryal sequence files with accompanying taxonomic IDs are available at http://biome.sdsu.edu/gaas/data/.

Mapping to phylogenetic trees

Similarly to the Interactive Tree Of Life (ITOL) [40] and MetaMapper (http://scums.sdsu.edu/Mapper), GAAS is able to graph the relative abundance of viral, microbial or eukaryotic species on phylogenetic trees such as the Viral Proteomic Tree (VPT) or Tree Of Life (http://itol.embl.de). The Viral Proteomic Tree was constructed using the approach introduced in the Phage Proteomic Tree and extending it to the >3,000 viral sequences present in the NCBI RefSeq viral collection (Edwards, R. A.; unpublished data, 2009).

Benchmark using simulated viral metagenomes

Simulated metagenomes were created to test the validity and accuracy of the GAAS approach

using the free software program Grinder (http://sourceforge.net/projects/biogrinder), which was developed in conjunction with GAAS. Grinder creates metagenomes from genomes present in a user-supplied FASTA file. Users can simulate realistic metagenomes by setting Grinder options such as community structure, read length and sequencing error rate. Over 9,500 simulated metagenomes based on the NCBI RefSeq virus collection were generated using Grinder. The viral database was chosen since its large amount of mosaicism and horizontal gene transfer represents a worst-case scenario. Therefore, benchmark results using the viral database are expected to be valid for higher-order organisms such as Bacteria, Archaea and eukaryotes. The parameters used were a coverage of 0.5 fold, and a sequencing error rate of 1% (0.9% substitutions, 0.1% indels). Half of the simulated metagenomes had a uniform rank-abundance distribution, while the other half followed a power law with model parameter 1.2. Sequence length in the artificial metagenomes was varied from 50 to 800 bp for the analysis of read length effects on GAAS estimates.

For each simulated viral metagenome, GAAS was run repeatedly with different parameter sets (relative alignment length and percentage of identity). The maximum E-value was fixed to 0.001 in order to remove similarities due to chance alone. Each set of variable parameters was tested on a minimum of 1,200 different Grinder-generated metagenomes. All computations were run on an 8-node Intel dual-core Linux cluster.

Due to the limited number of whole genome sequences available, a great majority of the sampled organisms in a metagenome cannot be assigned to a taxonomy. To evaluate the effect of sequences from novel organisms on GAAS estimates, the taxonomy of 80% randomly chosen organisms in the database was made inaccessible to GAAS rendering them “unknown”. A control simulation with 100% known organisms was run for comparison (Figure S2).

The accuracy of GAAS estimates was evaluated by comparing GAAS results to actual community composition and average genome size of the simulated metagenomes. The relative error for average genome size was calculated as $r = |x - x_e| / x$, where x and x_e are the true and estimated values respectively. For the composition, the cumulative error was calculated as $R = \frac{|r|_2}{n} = \frac{\sqrt{\sum_i^n r_i^2}}{n}$, where r_i is the relative error on the relative abundance of the target genome i and n is the total number of sequences in the database.

Because the benchmark results were not normal, non-parametric statistical tests were used for all pairwise (Mann-Whitney U test) and multi-factor comparisons (Friedman test) of average errors. Non-parametric correlations were calculated using Kendall's tau.

Benchmark using simulated microbial metagenomes

GAAS was also tested on the three simulated metagenomes available at IMG/m (http://fames.jgi-psf.org). Parameter setting and data processing were conducted as in viral benchmark experiments. Points on the IMG/m microbial benchmark graphs represent the average of 58 repetitions.

Microbial strains typically have a largely identical genome, with a fraction coding for additional genes and accounting for differences in genome length. An additional simulation was performed to investigate how the presence of closely related genomes influences the accuracy of the GAAS estimates. The 15 *Escherichia coli* strains present in the NCBI RefSeq database, ranging from 4.64 to 5.57 Mb in genome size, were used to produce ~4,500 shotgun libraries with Grinder. The parameters used were the same as for the simulated viral metagenomes, but with a coverage of 0.0014 fold (>1,000

sequences). Half of the simulated metagenomes were treated as in the viral benchmark, using the GAAS approach and assuming no unknown species. The other half were treated similarly but taking only the top similarity. Points on the graph of the microbial strain benchmark represent the average of >2,200 repetitions.

Meta-analysis of 169 metagenomes

The composition and average genome size for 169 metagenomes were calculated using GAAS. Most of these metagenomes were publicly available from the CAMERA [41], NCBI [42], or MG-RAST [43] (Table S2), and a few dozens were viromes and microbiomes newly collected from solar saltern ponds, chicken guts, different soils and an oceanic oxygen minimum zone (Protocol S1). The metagenomes used here therefore represent viral, bacterial, archaeal, and protist communities sampled from a diverse array of biomes and were categorized as one of the following: “aquatic”, “terrestrial”, “sediment”, “host-associated”, and “manipulated / perturbed”. The large number of aquatic metagenomes was further subdivided into: “ocean”, “hypersaline”, “freshwater”, “hot spring” and “microbialites”. Sampling, filtering, processing and sequencing methods differed among compiled metagenomes. Table 1 provides a summary of the number of metagenomes from each biome (a list of the complete dataset is presented in detail in Table S2).

For all metagenomes, GAAS was run using a threshold E-value of 0.001, and an alignment relative length of 60%. In addition, for bacterial, archaeal and eukaryotic metagenomes, similarities were calculated using BLASTN with an alignment similarity of 80%. Due to the low number of similarities in viral metagenomes using BLASTN, TBLASTX was used for viruses, with a threshold alignment similarity of 75%. All average genome length estimates produced from less than 100 similarities were discarded to keep results as accurate as possible. Manipulated metagenomes were

ultimately not used in the meta-analysis because they do not accurately represent environmental conditions. Statistical pairwise differences between average genome lengths across biomes were assessed using Mann-Whitney U rank-sum tests.

The average genome length and relative abundance results obtained for all metagenomes with our GAAS method were compared to the "standard" analytical approach where: 1) only the top similarity for each metagenomic sequence is kept, 2) there is no filtering by alignment similarity or relative length, and 3) no normalization by genome length is carried out. The virome from the Sargasso Sea was chosen to illustrate in detail the difference between the results obtained with the two methods (Figure 3).

Correlation between viral and microbial average genome length

Average genome lengths were calculated for 25 pairs of microbial and viral metagenomes sampled from the same location at the same time. The statistical relationship between viral and microbial average genome length in paired metagenomes was evaluated using Kendall's tau, since lengths were not normally distributed. Regression analysis was performed with Generalized Linear Models (GLM). Interactions between genome lengths and biome classifications were not significant and were not included in final models.

Statistical analyses

All statistical analyses of the GAAS benchmark results, environmental genome length and genome length correlations described above were performed using the free statistical software package R (http://www.R-project.org/) [44].

Acknowledgements

We want to acknowledge Dr. Ed Delong, Dr. Osvaldo Ulloa and Dr. Gadiel Alarcon for organizing the Oxygen Minimum Zone project. We are thankful to Linlin Li and John Buchanan for their assistance in the collection of fish metagenomes at the Kent SeaTech fish farm. Finally, we thank the J. Craig Venter Institute for making metagenomes of the Global Ocean Sampling Phase II and Antarctica Lakes publicly available.

Financial Disclosure

The Massachusetts Institute of Technology and the Agouron Institute for sequencing funded the Oxygen Minimum Zone project. The National High Technology Research and Development Program of China (2007AA09Z443 and 2007AA021301) and Knowledge Innovation Project of The Chinese Academy of Sciences (KSCX2-YW-G-022) supported the South China sediments microbiome project. The Antarctica Lakes research was supported by the Gordon and Betty Moore Foundation. NSF OPP 0124733 funded the Arctic microbiome sampling. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. PLoS Biology 4: e368.

2. Pignatelli M, Aparicio G, Blanquer I, Hernández V, Moya A, et al. (2008) Metagenomics reveals our incomplete knowledge of global diversity. Bioinformatics 24: 2124-2125.

3. Raes J, Foerstner KU, Bork P (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol 10: 490-498.

4. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377-86.

5. Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52: 540-542.

6. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, et al. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452: 340-343.

7. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629-632. doi:10.1038/nature06810

8. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6:41. doi:10.1186/1471-2105-6-41

9. Steward GF, Montiel JL, Azam F (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. Limnol Oceanogr 45: 1697-1706.

10. Holmfeldt K, Middelboe M, Nybroe O, Riemann L (2007) Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their Flavobacterium hosts. Appl Environ Microbiol 73: 6730-6739.

11. Sandaa R (2008) Burden or benefit? Virus-host interactions in the marine environment. Res Microbiol 159: 374-381.

12. Weinbauer MG, Rassoulzadegan F (2004) Are viruses driving microbial diversification and diversity? Environ Microbiol 6: 1-11.

13. Graves LM, Swaminathan B (2001) PulseNet standardized protocol for subtyping Listeria monocytogenes by macrorestriction and pulsed-field gel electrophoresis. Int J Food Microbiol 65:

55-62.

14. Diez B, Antón J, Guixa-Boixereu N, Pedrós-Alió C, Rodriguez-Valera F (2000) Pulsed-field gel electrophoresis analysis of virus assemblages present in a hypersaline environment. Int Microbiol 3: 159-164.

15. Raes J, Korbel J, Lercher M, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. Genome Biol 8: R10.

16. Sandaa R, Foss Skjoldal E, Bratbak G (2003) Virioplankton community structure along a salinity gradient in a solar saltern. Extremophiles 7: 347-351.

17. Wommack KE, Ravel J, Hill RT, Chun J, Colwell RR (1999) Population dynamics of Chesapeake bay virioplankton: total-community analysis by pulsed-field gel electrophoresis. Appl Environ Microbiol 65: 231-40.

18. Ranea JAG, Buchan DWA, Thornton JM, Orengo CA (2004) Evolution of protein superfamilies and bacterial genome size. J Mol Biol 336: 871-87.

19. Rohwer F, Edwards RA (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol 184: 4529-35.

20. Hugenholtz P, Tyson GW (2008) Microbiology: metagenomics. Nature 455: 481-483.

21. Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: read length matters. Appl Environ Microbiol 74: 1453-1463.

22. Rosario K, Duffy S, Breitbart M (2009) Diverse circovirus-like genome architectures revealed by environmental metagenomics. J Gen Virol 90: 2418-2424. doi:10.1099/vir.0.012955-0

23. Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, et al. (2009) The genomic basis of trophic strategy in marine bacteria. Proc Nat Acad Sci USA 106: 15527-15533.

24. Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Nat Acad Sci USA 101: 3160-3165.

25. Willner D, Thurber RV, Rohwer F (2009) Metagenomic signatures of 86 microbial and viral metagenomes. Environ Microbiol 11: 1752-1766.

26. Hewson I, Winget DM, Williamson KE, Fuhrman JA, Wommack KE (2006) Viral and bacterial assemblage covariance in oligotrophic waters of the West Florida shelf (Gulf of Mexico). J Mar Biol Assoc UK 86: 591-603.

27. Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, et al. (2008) Microbial ecology of four

coral atolls in the northern Line Islands. PLoS ONE 3: e1584.

28. DasSarma P, DasSarma S (2008) On the origin of prokaryotic "species": the taxonomy of halophilic Archaea. Saline Syst 4: 5.

29. Dyall-Smith M, Tang S, Bath C (2003) Haloarchaeal viruses: how diverse are they? Res Microbiol 154: 309-313.

30. Vega Thurber RL, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, et al. (2008) Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. Proc Nat Acad Sci USA 105: 18413-18418.

31. Thurber RV, Willner-Hall D, Rodriguez-Mueller B, Desnues C, Edwards RA, et al. (2009) Metagenomic analysis of stressed coral holobionts. Environ Microbiol 11: 2148-2163.

32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-10.

33. Rasko D, Myers G, Ravel J (2005) Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics 6: 2.

34. Sadreyev R, Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 326: 317-336.

35. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Nat Acad Sci USA 87: 2264-2268.

36. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, et al. (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J Proteome Res 5: 2339-2347.

37. Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, et al. (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. Methods 40: 303-311.

38. Paoletti AC, Parmely TJ, Tomomori-Sato C, Sato S, Zhu D, et al. (2006) Quantitative proteomic analysis of distinct mammalian mediator complexes using normalized spectral abundance factors. Proc Nat Acad Sci USA 103: 18928-18933.

39. Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. Annu Rev Genet 38: 771-92.

40. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127-8.

41. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. PLoS Biol 5: e75.

42. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, et al. (2000) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 28: 10-4.

43. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

44. R Foundation for Statistical Computing, Vienna, Austria (n.d.) R: A language and environment for statistical computing.

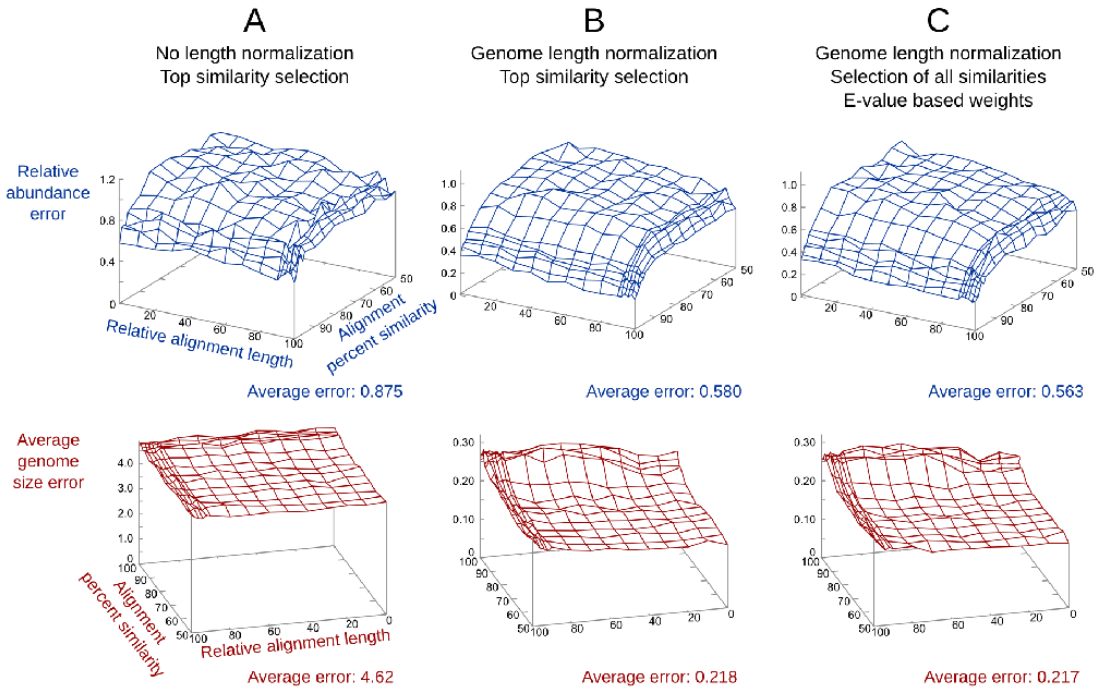


Figure 1: Effects of length normalization and similarity weighting on the accuracy of GAAS estimates. Different methods were used: (A) the standard method (no length normalization, selection of the top similarity only), (B) a combination of genome length normalization and top similarity selection only, and (C) the GAAS method (genome length normalization, selection of all significant similarities, and E-value based weights). Decreases in average error indicate increased accuracy. In the simulated viral metagenomes, 100 bp sequences were used and 80% of the species were considered unknown.

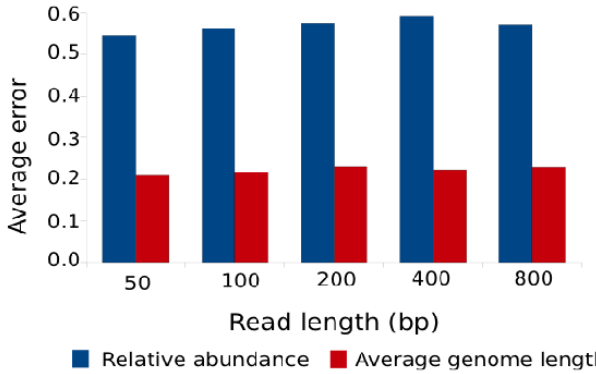


Figure 2: Effects of metagenomic read length on average error of GAAS estimates. Decreases in average error indicate increased accuracy. In the simulated metagenomes, 80% of the species were considered unknown. See Figure S5 and Figure S6 for full details. considered unknown. See Figure S5 and Figure S6 for full details.

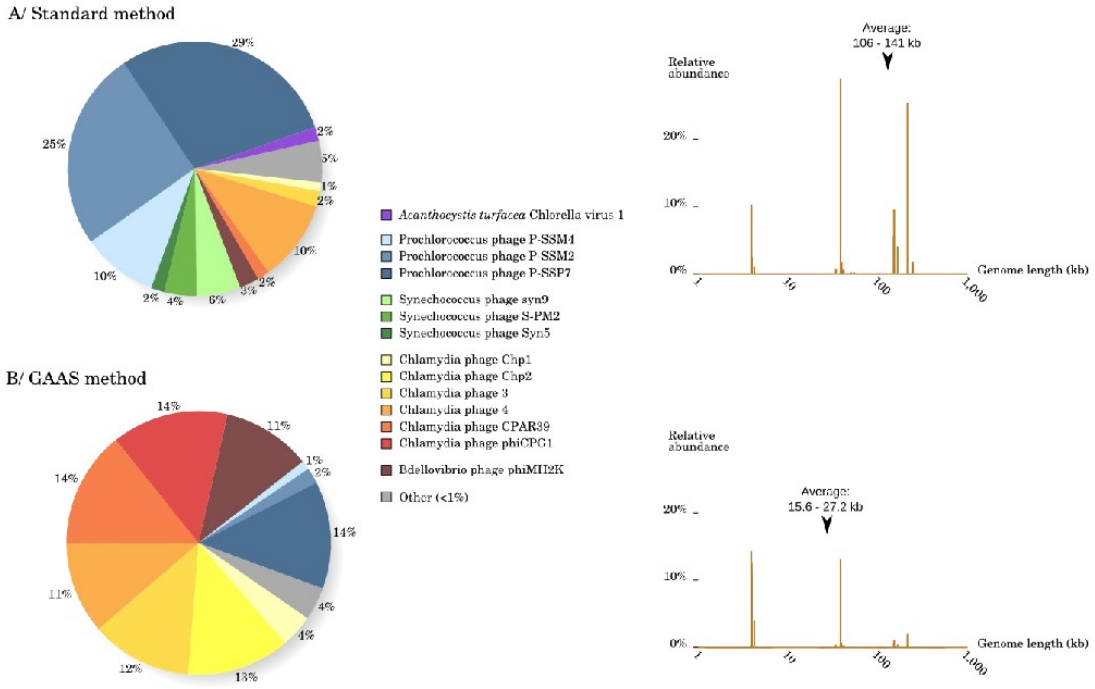


Figure 3: Re-analysis of the Sargasso Sea viral community. Genome relative abundance in the Sargasso Sea (left) and size spectrum with 95% confidence interval for the average genome length (right) were calculated using the standard method (A) and GAAS (B).

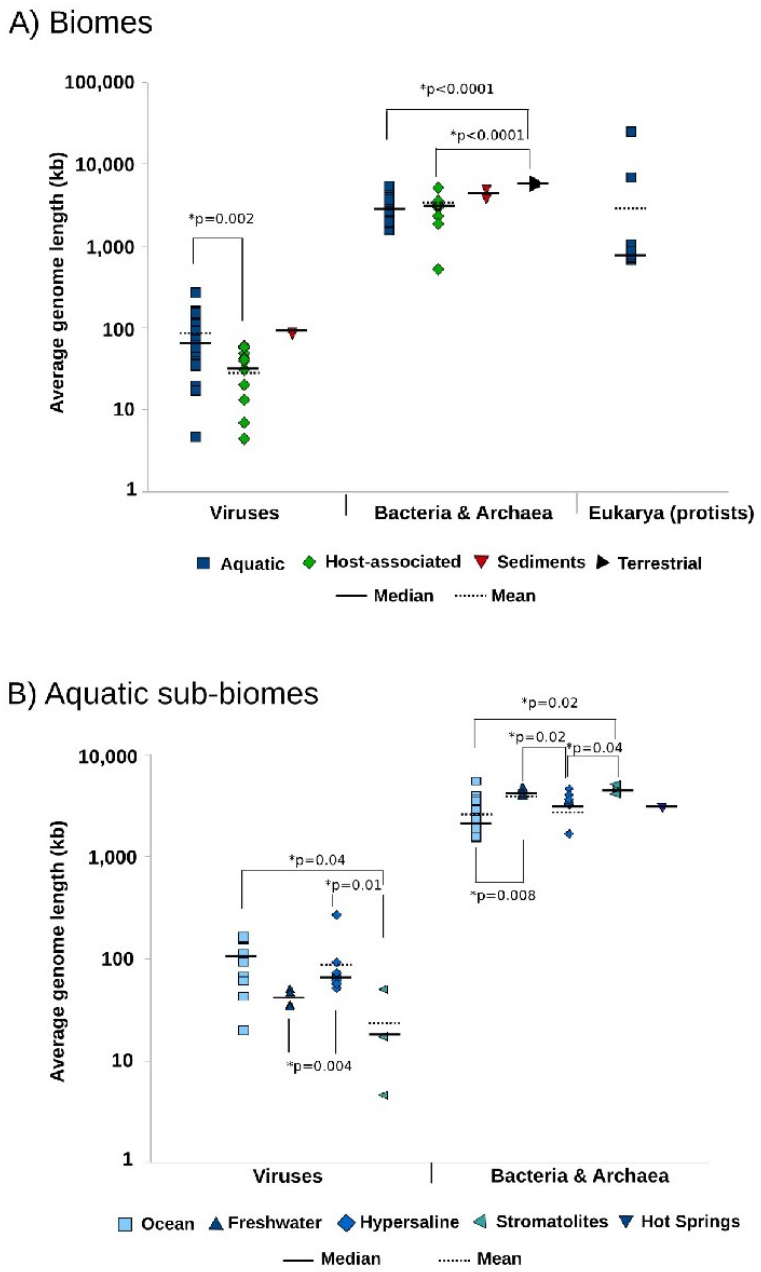


Figure 4: Average genome length of viruses, Bacteria and Archaea, and protists in metagenomes. Different biomes (A) and marine sub-biomes (B) were analyzed using GAAS. Non-parametric Mann-Whitney U tests were used to compare biomes. Metagenomes from sediments and hot springs were excluded from the statistical analysis due their small number. All protist metagenomes were from the ocean and could not be sub-classified further.

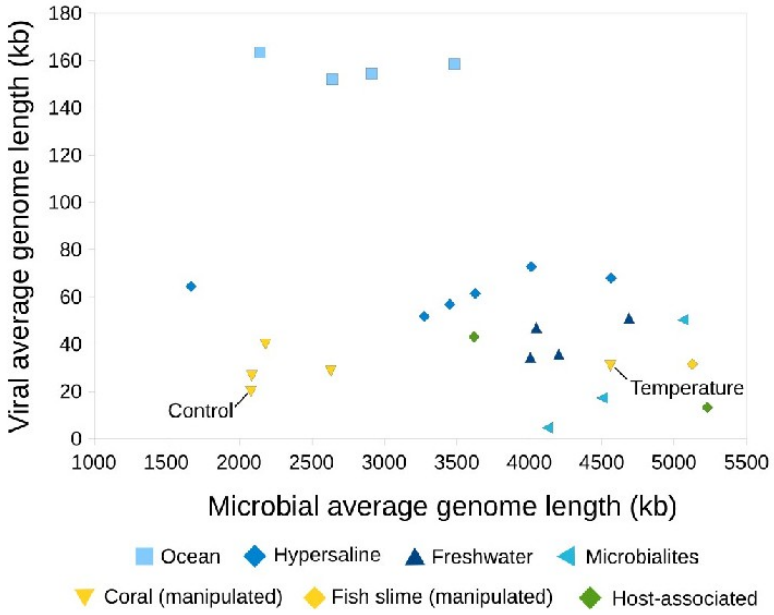


Figure 5: Relationship between average microbial and viral genome lengths in paired metagenomes.

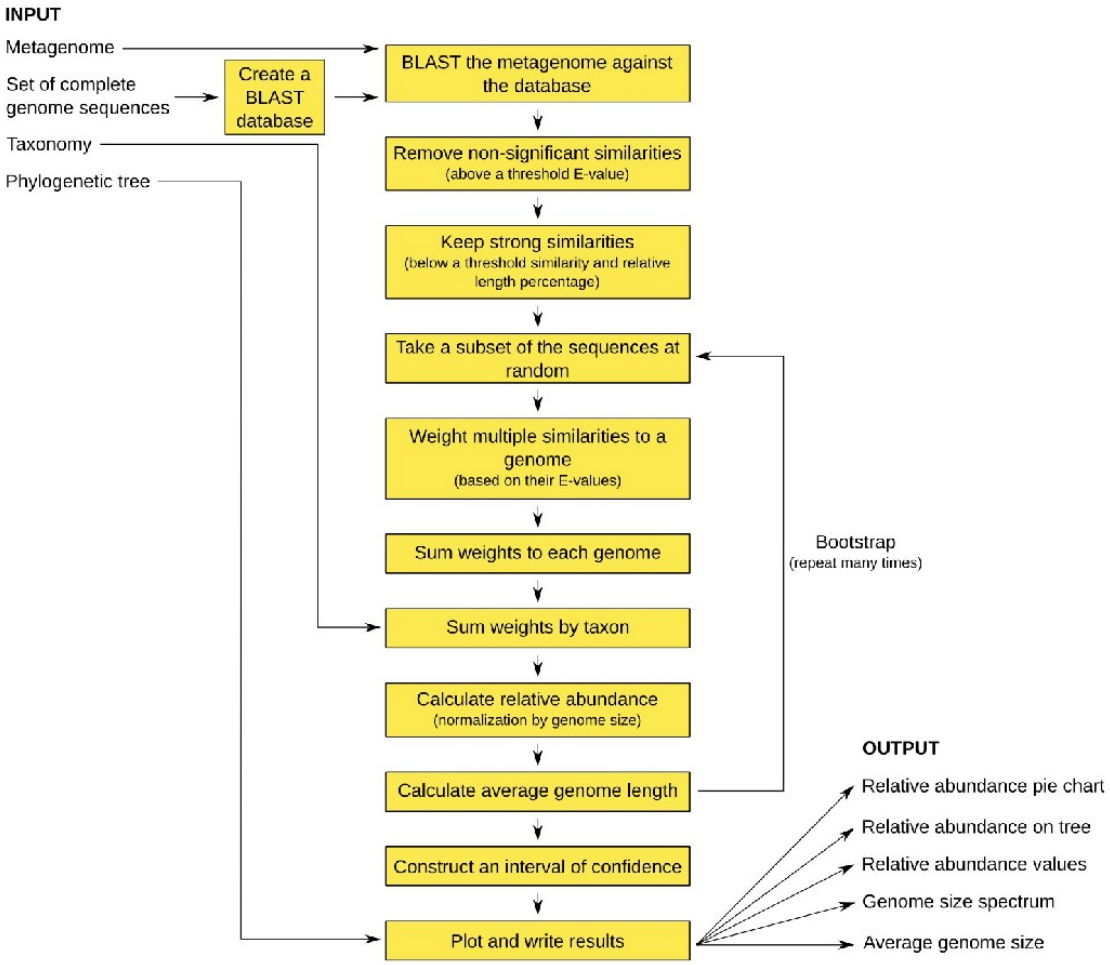


Figure 6: Flowchart of GAAS to calculate relative abundance and average genome size. GAAS runs BLAST and uses various corrections to obtain accurate estimations.

Tables

Table 1: Summary of metagenomes by type used in the meta-analysis

Biome	Sub-biome	Number of viral metagenomes	Number of bacterial and archaeal metagenomes	Number of protist metagenomes
Aquatic (total)	-	34	45	17
Aquatic	Ocean	15	26	17
Aquatic	Hypersaline	10	10	0
Aquatic	Freshwater	4	4	0
Aquatic	Hot spring	2	2	0
Aquatic	Microbialites	3	3	0
Sediments	-	3	2	0
Terrestrial (soil)	-	4	19	2
Host-associated	-	17	11	0
Manipulated / perturbed	-	7	8*	0

* The five manipulated coral metagenomes also contained sequences from eukaryotic genomes as described in [31].

Supplementary Material

Protocol S1: sample collection and metagenome sequencing

Oxygen minimum zone viromes:

The oceanic oxygen minimum zone samples were collected in June 2008 off Iquique, Chile, (20.104° S and 70.404° W). Oxygen minimum zone viral metagenomes were constructed by filtering 40 l of water collected using a CTD rosette lowered to a sampling depth of 90 and 200 m (OxMinZoneVir200806-90 and OxMinZoneVir200806-200 respectively). Samples were concentrated through a 100 kDa tangential flow filter to retain viral particles. The concentrate was passed through a 0.45 µm sterivex filter to remove larger cells and treated with chloroform. The viruses were purified using cesium chloride (CsCl) step gradients to remove free DNA and any cellular material. Viral samples were visually checked for microbial contamination using epifluorescence microscopy. Viral DNA was extracted using CTAB/phenol:chloroform extractions and amplified using Genomiphi reactions. These reactions were pooled and purified using silica columns (Qiagen Inc, Valencia, CA). The DNA was precipitated with ethanol and re-suspended in water at a concentration of approximately 300 ng μl^{-1}. Sequencing was performed using pyrosequencing on Roche Applied Sciences/454 Life Sciences GS-FLX platforms with a practical limit of 250 bp. Duplicate sequences were removed from the obtained dataset and the submitted to NCBI (Genome Project 40791 and 40793).

Runting-stunting chicken gut viromes:

One day old specific pathogen-free broilers (USDA-ARS, SEPRL, Athens, GA) were orally

infected with 1 ml of gut content from 12-day-old commercial broiler chickens which showed the typical signs of runting-stunting-syndrome (RSS) in chicken (growth retardation > 40%, cystic lesions in the small intestine). Before inoculation, the gut content of RSS affected chicken was centrifuged at 4°C for 30 min at 3000 x g. the obtained supernatant was filtered first through a 0.45 µm filter followed by filtration through a 0.22 um filter. A second group of broilers was mock-infected with phosphate buffered saline. Five days, 8 d, and 12 d after infection, 10 birds of each group were euthanized and necropsy was performed. The duodenal loop was taken for histological examination. The analysis of the sections showed that cystic lesions were only present in the infected group. The highest number of lesions was observed at 8 d after infection. Based on this result the gut content harvested at 5 d after infection was used for subsequent experiments. The purification of the gut content was performed following a multi-step centrifugation protocol. In a first step, the samples were centrifuged at 16000 x g to remove cellular organelles and debris. The obtained supernatant was filtered twice as described above. Next the filtrate was centrifuged through a 10% sucrose cushion made in TEN buffer (10 mM Tris-HCl, 100 mM NaCl, 1 mM EDTA, pH 7.5) at 174899 x g for 3h. The obtained pellets (RSS+, RSS-) were resuspended in 400 µl TEN buffer. To purify nucleic acids, the RNA and DNA localized outside of viral particles needed to be degraded. To this end, 40 µl of 10x DNA I buffer (Roche), 20 units of DNase I (Roche) and 10 ug of RNase I (Roche) was added to 360 µl of the viral suspension. The mixture was incubated for 1 h at 37C. Both samples (RSS+, Con) were then split and 200 µl of each sample was used for purification of either the DNA (QIAamp DNA Blood Mini Kit, Qiagen) or RNA (High Pure RNA isolation Kit, Roche) following the manufacturers instructions. The resulting samples (RSS+ RNA, RSS- RNA, RSS+ DNA and RSS- DNA) were amplified separately using two different protocols to amplify the metagenome (called respectively

ChickenRuntingStuntingPRnaVir2008, ChickenRuntingStuntingMRnaVir2008, ChickenRuntingStuntingPDnaVir2008 and ChickenRuntingStuntingMDnaVir2008). The RNA containing samples were amplified using the Transplex Whole Transcriptome Amplification Kit (Sigma). The amplification of the DNA library for both DNA samples was performed using GenomiPhi V2 DNA Amplification Kit (GE Healthcare). Both protocols were applied as recommended by the manufacturer. The resulting cDNA library was submitted to 454 Life Science for sequencing using the GS-FLX platform. Duplicate sequences were removed and submitted to NCBI (Genome Project 40789, 40785, 40787 and 40783).

Solar saltern microbiome:

A water sample from the solar saltern of South Bay Salt Works (Chula Vista, CA) was collected in July 2004 from a pond with high salinity (28-30%, measured using a hand refractometer). The microbial fraction was isolated from the water sample by passage through a 0.2 µm tangential flow filter (TFF, Millipore). The retentate was kept and the microbial fraction was collected from the 0.2 µm TFF retentate by centrifugation at ~ 2000 xg for 10 min. Microbial DNA was extracted using the Ultra Clean Soil DNA Kit (Mo Bio Laboratories, CA). The microbial DNA samples was amplified using the strand-displacement Φ29 DNA polymerase (GenomiPhi Amersham Biosciences, NJ). The resulting metagenomic DNA was pyrosequenced on the GS20 sequencer (454 Life Sciences, CT). The raw metagenomic sequences were screened to remove duplicate sequences. The metagenome, referred to as HighSalternSDbayMicD200407, was submitted to NCBI (Genome Project 40795).

South China sediments microbiome:

A marine sediment sample was collected using a gravity piston corer during a March 2006 Marine Expedition at the BD7-2 station of the South China Sea at a depth of 778.5 m below seafloor. The sample was stored onboard at 4°C and then divided into 5-cm sediment subsamples below seafloor. and stored in −80°C. The 5 to 10 cm layer was used for the library construction in this study. Prior to the metagenome DNA extraction, marine sediments were washed following the protocol previously described by Fortin [1] to remove contaminants: three washes, each wash with 100ml washing buffer (50mM Tris-HCl, pH 9.0, 100mM Na2EDTA, 1.0% PVP, 100mM NaCl, 0.05% Triton X-100), after vortexing for 1 min, the sample was incubated in 55°C for 3 min, and then centrifuged at 3,000×g for 5 min [1]. After washing steps, 5g pellet was mixed by vortexing with 13.5 ml of extraction buffer (100 mM Tris-HCl, pH 8.0, 100 mM sodium EDTA, pH 8.0, 100 mM sodium phosphate, pH 8.0, 1.5 M NaCl, 1% CTAB). Three cycles of thawing and freezing in liquid nitrogen were then applied to the suspension and the sample was then incubated at 37°C with 50 µl of proteinase K (20 mg/ml) for 30 min [2,3]. The extracted metagenomic DNA was repaired using Epicentre's repair enzyme mix and size-selected on 1% agarose PFGE with CHEF-DRIII system (Bio-Rad). Pulsed-field gel electrophoresis was carried out at 5 V/cm voltages with a ramping time of 0.1s to 40s at 14°C in 0.5×TBE buffer for 16 h. The metagenomic DNA with size of 36 to 48 kb was cut off from the gel and recovered by electro elution and then ligated to Epicentre's pCC2 FOS fosmid vector. This metagenomic library, named SouthChinaSeaSedimentsMic, was constructed using Epicentre' CopyControl fosmid library production kit. Over 1000 fosmid clones were randomly selected from the IMCAS-F003 library for end sequencing using T7 primer (5'-TAATACGACTCACTATAGGG-3') and pCC2 reverse sequencing primer (5'-CAGGAAACAGCCTAGGAA-3'). All the fosmid end sequences were revised and trimmed using Lasergene package, version 7.10 (DNA star, USA) before submission to

NCBI (Genome Project 33581).

Pacific Beach sand metagenome:

DNA was extracted from a sample of sand at Pacific Beach, San Diego, California, USA, in august 1999, cloned and sequenced. The protocol was described in detail by Naviaux [4]. Here, over 2,300 additional clones from this metagenomic library, named PacificBeachSandEuk here, were sequenced following the same procedure as before and the full set of sequences (~4900) was made publicly available through the NCBI (Genome Project 13729).

Fish gut viromes:

Adult hybrid striped bass were collected in April 2006 from a 5x2 m open-air aquaculture pond in San Diego, California, USA. Each fish was classified as healthy or morbid by veterinarians upon visual inspection. Fish were sacrificed with an overdose of MS-222 (Finquel, Argent Laboratories), and examined for the presence of gross external and internal lesions to confirm preliminary diagnoses. Symptomatic fish had empty gut contents. Five symptomatic and five asymptomatic fish were selected and gut contents were collected by flushing aseptically with 10 mL of SM buffer. Samples were sonicated (15 seconds, 3 times) and then centrifuged at 150 x g for 20 minutes at 4°C. The supernatant was then filtered (0.45 µm and 0.2 µm) to separate the microbial fraction (attached onto the filter) from the viruses (filtrate). Viral particles in the filtrate were purified using a CsCl step gradient and viral DNA was extracted as described by Thurber [5]. Viral DNA was amplified with GenomiPhi (GE Healthcare, Piscataway, NJ) and ethanol precipitated. Approximately 10 µg of each DNA sample was submitted for GS20 pyrosequencing at 454 Life Sciences to produce the FishHealGutKentSTVir20060504 and FishMorGutKentSTVir20060504 metagenomes (from healthy

fish and morbid fish respectively). Duplicate sequences were removed and the metagenomes were submitted to NCBI (Genome Project 28397 and 28399).

Arctic marine microbial metagenome:

The Arctic sample (ArcticMic) was collected from a depth of 10 m at 72:19.33N, 151:59.07W [6]. Environmental DNA was extracted from the bacterial size fraction obtained by pumping 500 L of seawater sequentially through a 1 µm nominal pore size polypropylene string-wound filter (Cole Parmer) and a 0.8 µm polycarbonate filter (Nuclepore). Bacteria were collected from the filtrate by tangential flow filtration using a 0.1 µm hollow fiber filter (A/G Technology) and an Amicon DC10 gear pump. The sample was concentrated to 2 L, diafiltered with a buffer (0.5 M NaCl, 0.1 M EDTA, 10 mM Tris pH 8.0), and stored frozen. Cells were later collected from the thawed concentrate and lysed by treatment with SDS and lysozyme. Nucleic acids were extracted from the lysate using phenol and chloroform, sequenced, and released (Genome Project 29035).

Soil microbiomes:

Soil cores were taken to a depth of 5 cm from a random location in proximity of the land-use type associated with primary and secondary tower locations at selected National Ecological Observatory Network (NEON) primary sites. NEON soils were collected and stored at -20 °C. Prior to downstream analysis soil samples were passed through an 8 mm sieve in order to remove roots and any associated surface litter. After sieving, remaining fine roots were hand-picked from the soil with tweezers. Metagenomic DNA was isolated from 5-10 g of soil using the UltraClean® Mega Soil DNA

Isolation Kit (MOBIO, Carlsbad, CA). Concentration and quality assessment was determined by fluorometry (Qubit Quantitation Platform, Invitrogen, Carlsbad, CA) and agarose gel electrophoresis. Metagenomic DNA (5 ug) was used to construct shotgun libraries and prepared for sequencing using the standard GS FLX emPCR protocol and LR70 sequencing chemistry (Roche Applied Science, Indianapolis, IN). Sequencing was performed by the High-Throughput Genome Analysis Core (HGAC), Institute for Genomics and Systems Biology at Argonne National Laboratory. MG-RAST accession numbers are:

Metagenome	MG-RAST ID
SoilSJ1Mic	MG 4441557.3
SoilWF1Mic	MG 4441556.3
SoilHF1Mic	MG 4441642.3
SoilKP3Mic	MG 4441643.3
SoilLF2Mic	MG 4441644.3
SoilSJ2Mic	MG 4441645.3
SoilKW1Mic	MG 4441664.3
SoilWF2Mic	MG 4441665.3
SoilYN2Mic	MG 4441687.3
SoilTF1Mic	MG 4441688.3
SoilCP1Mic	MG 4441689.3
SoilCC1Mic	MG 4441690.3
SoilCP3Mic	MG 4441691.3
SoilKW2Mic	MG 4441691.4
SoilKP1Mic	MG 4441994.3
SoilTF2Mic	MG 4442452.3
SoilYN1Mic	MG 4442453.3
SoilLF1Mic	MG 4442455.3

Microbial metagenomes of the Indian Ocean and Antarctica lakes:

These metagenomes were collected during phase II of the Global Ocean Sampling effort [7,8] and during an Antarctica expedition. While these data is unpublished, the sequences and metadata for

these samples are available on CAMERA and NCBI:

Metagenome	NCBI Genome Project ID
AntarcticaLakeMic	GP 33179
GS000a11Mic	GP 13694
GS000a13Mic	GP 13694
GS000b11Mic	GP 13694
GS000b13Mic	GP 13694
GS000cMic	GP 13694
GS000dMic	GP 13694
GS001aEuk	GP 13694
GS001bEuk	GP 13694
GS011Mic	GP 13694
GS012Mic	GP 13694
GS016Mic	GP 13694
GS020Mic	GP 13694
GS023Mic	GP 13694
GS025Euk	GP 19735
GS034Mic	GP 13694
GS048aMic	GP 13694
GS048bEuk	GP 13694
GS108bEuk	GP 13694
GS110bEuk	GP 13694
GS112bEuk	GP 13694
GS117bEuk	GP 13694
GS122bEuk	GP 13694
Move858Vir	GP 13694

References:

1. Fortin N, Beaumier D, Lee K, Greer CW (2004) Soil washing improves the recovery of total community DNA from polluted and high organic content sediments. J Microbiol Methods 56: 181-191.

2. Kauffmann IM, Schmitt J, Schmid RD (2004) DNA isolation from soil samples for cloning in different hosts. Appl Microbiol Biotechnol 64: 665-670.

3. Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. Appl

Environ Microbiol 62: 316-322.

4. Naviaux RK, Good B, McPherson JD, Steffen DL, Markusic D, et al. (2005) Sand DNA - a genetic library of life at the water's edge. Mar Ecol Prog Ser 301: 9-22.

5. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protocols 4: 470-483.

6. Cottrell MT, Yu L, Kirchman DL (2005) Sequence and expression analyses of Cytophaga-like hydrolases in a Western Arctic metagenomic library and the Sargasso Sea. Appl Environ Microbiol 71: 8506-8513.

7. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.

8. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5: e77.

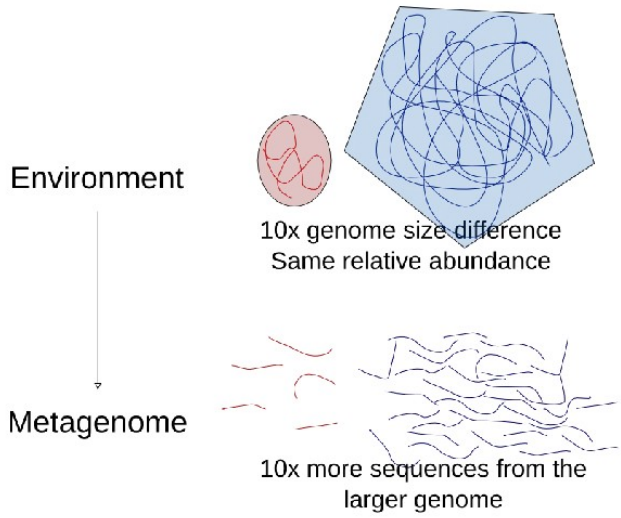


Figure S1: Sampling bias toward larger genomes in metagenomic libraries. Larger genomes will produce more fragments of a given size, and are more likely to be sampled even if they occur in the same abundance as small genomes.

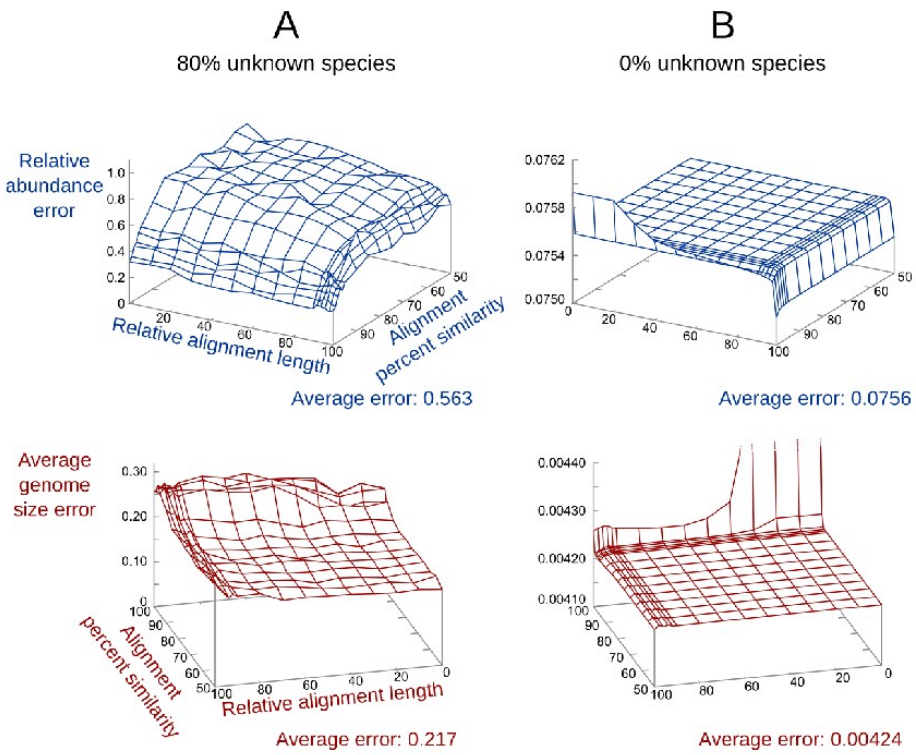


Figure S2: Accuracy of the GAAS estimates when no species are unknown. Error on the relative abundance (top) and average genome size estimates (bottom) when: (A) 80% of the species were treated as unknown, (B) no species were assumed to be unknown. The simulated viromes were made of 100 bp sequences.

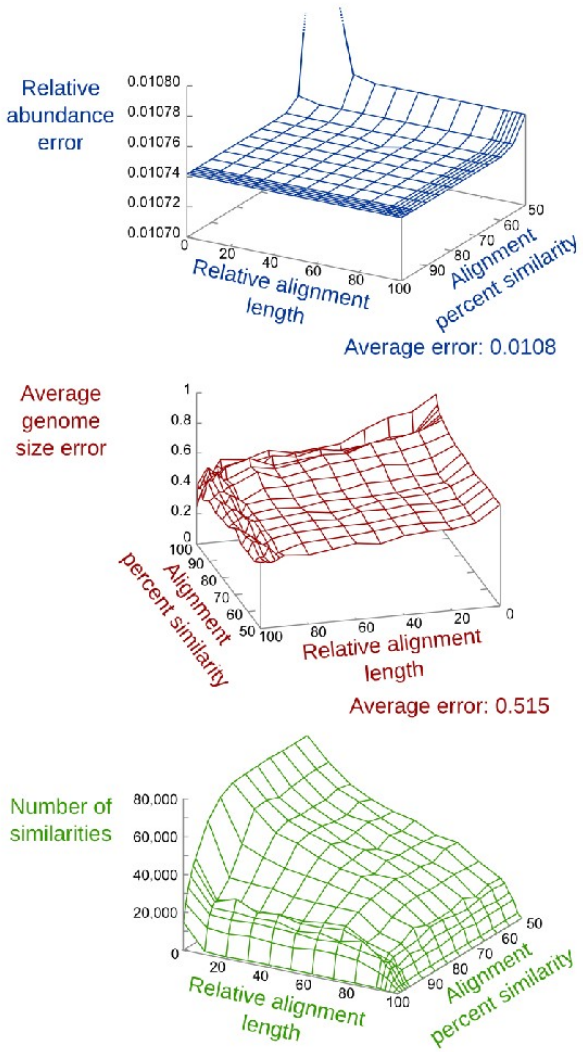


Figure S3: Accuracy of GAAS estimates for microbial metagenomes. GAAS relative abundance error (top), average genome size error (middle) and number of similarities (bottom) for the JGI simulated microbial metagenomes (~1,200 bp/read). 80% of the species were treated as unknown.

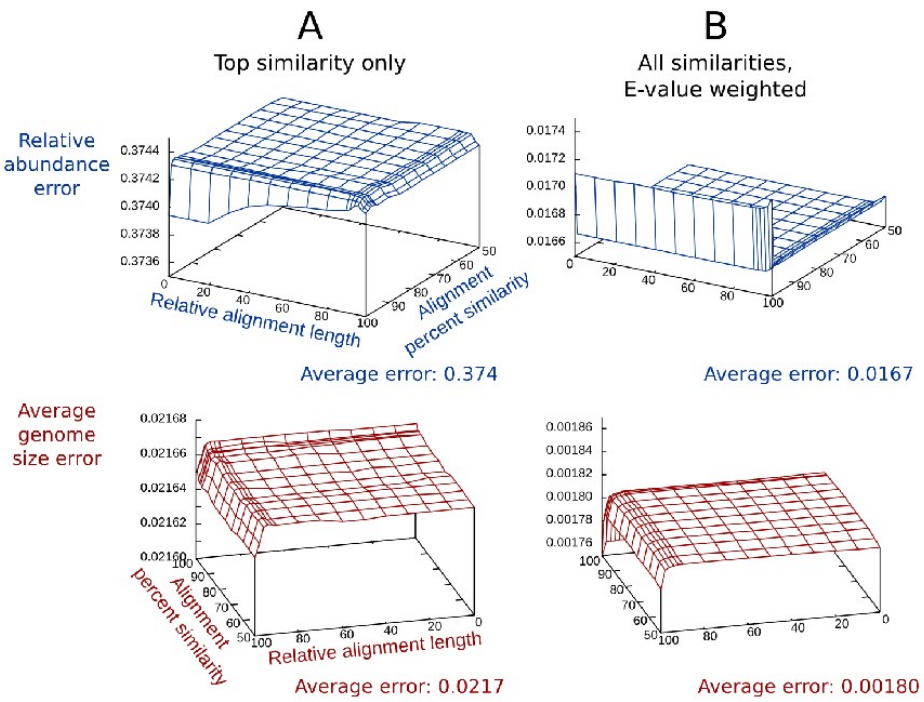


Figure S4: Effect of using all similarities for microbial strains. The error on community composition (top) and average genome length (bottom) for simulated metagenomes made of 15 *Escherichia coli* strains was estimated by GAAS. Sequence length was 100 bp and no strains were treated as unknown.

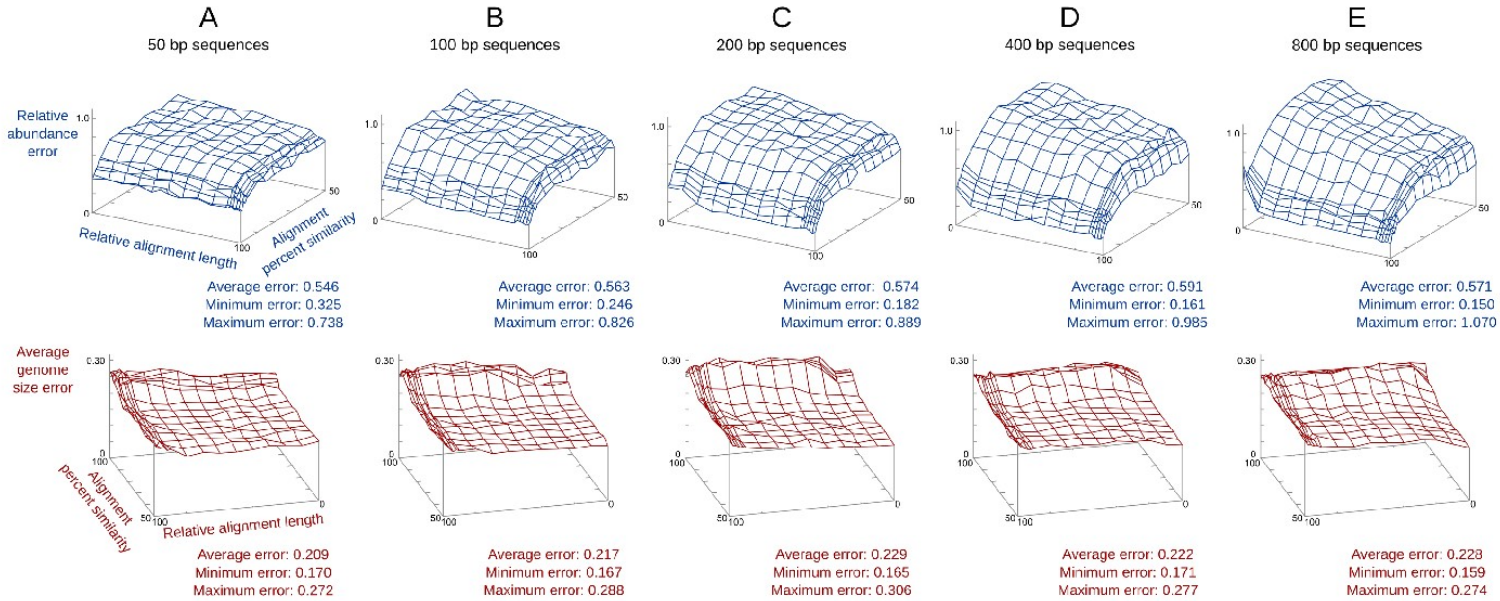


Figure S5: Effect of metagenomic sequence length on the accuracy of GAAS estimates. Error was calculated for the relative abundance (top) and average genome length (bottom) estimates. 80% of the species in the viral simulated metagenomes were treated as unknown.

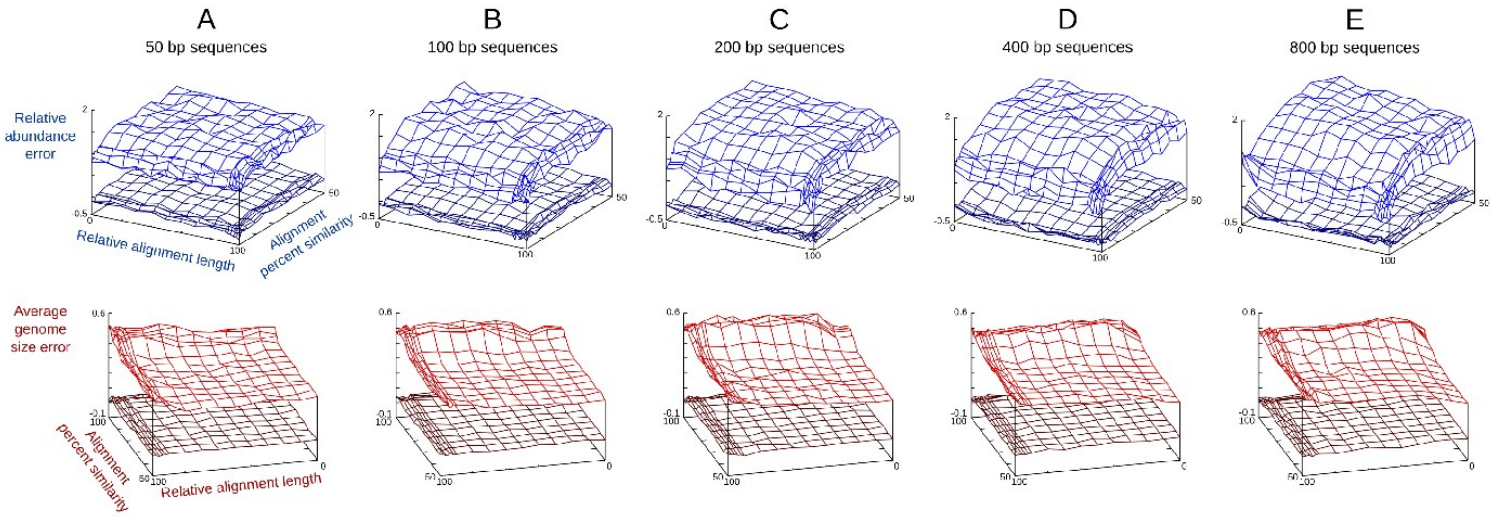


Figure S6: Error surfaces for Figure S5. The two surfaces of each graph correspond to the average error ± the standard deviation for the >1,200 simulated metagenomes.

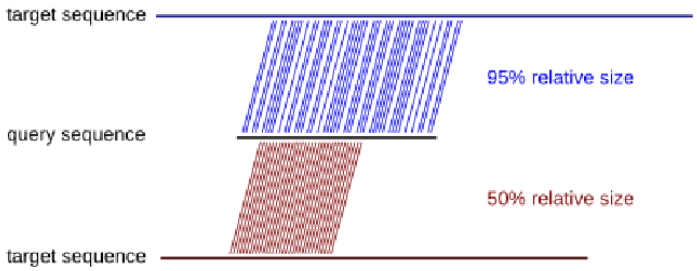


Figure S7: The relative alignment length filtering parameter. The relative alignment length is defined as the ratio of the length of the alignment over the length of the query sequence length, expressed in percent.

Table S1: Biome-averaged genome length estimated by GAAS for the metagenomes of each environment. The numbers reported are: mean (median) ± standard deviation

Biome	Sub-biome	Average viral genome length (kb)	Average bacterial and archaeal genome length (kb)	Average protist genome length (kb)
Aquatic (total)	-	79.9 (61.4) ± 59.4	3,020 (2,970) ± 1,150	2,690 (807) ± 6,060
Aquatic	Ocean	102 (103) ± 53.3	2,580 (2,150) ± 1,120	-
Aquatic	Hypersaline	91.9 (66.2) ± 72.2	3,250 (3,430) ± 925	-
Aquatic	Freshwater	42.0 (41.4) ± 8.23	4,240 (4,130) ± 313	-
Aquatic	Hot spring	-	2,820 (2,560) ± 1,340	-
Aquatic	Stromatolites	24.1 (17.3) ± 23.6	4,560 (4,500) ± 157	-
Sediments	-	85.6 (85.6) ± 4.28	4,370 (4,370) ± 803	-
Terrestrial (soil)	-	-	5,910 (5,930) ± 218	-
Host-associated	-	33.2 (39.8) ± 21.2	3,150 (3,190) ± 1,420	-

Table S2: Detail of the 169 metagenomes used for the meta-analysis and their average genome size estimated by GAAS. Accession numbers: CA, CAMERA Accession; GB, NCBI GenBank; GP, NCBI Genome Project; GSS, NCBI Genome Survey Sequence; MG: MG-RAST Accession; SRA, NCBI Short Read Archive.

Metagenome name	Type	Biome	Sub-biome	Microbial-viral pairing	Accession number	Reference	Est. avg. genome length (kb)
AlaskanSoilFungalEuk	eukaryal	Terrestrial	-	-	GP 28853	Allen et al., ISME J., 2009 [1]	-
AlicanteSaltern Mic	microbial	Aquatic	Hypersaline	-	GSS DU824018-DU826964	Legault et al., BMC Genomics 2006 [2]	3,190.26
AlvinellaWorm Mic	microbial	Host-associated	-	-	GP 17241	Grzymski et al., Proc. Nat. Acad. Sci. USA 2008 [3]	2,351.67
AntarcticaLake Mic	microbial	Aquatic	Hypersaline	-	GP 33179	-	1,657.26
ArcticMic	microbial	Aquatic	Ocean	Arctic Ocean	GP 29035	-	1,606.13
ArcticVir2002	viral	Aquatic	Ocean	Arctic Ocean	GP 17769	Angly et al., PLoS Biology 2006 [4]	66.67
BabyFecesSDVir	viral	Host-associated	-	-	GB ED651217-ED651693	Breitbart et al., Res. Microbiol. 2008 [5]	-
BBCVir1996to2004	viral	Aquatic	Ocean	-	GP 17767	Angly et al., PLoS Biology 2006 [4]	42.75
BearpawHotSpringVir	viral	Aquatic	Hot spring	-	GP 18929	Pride et al., BMC Genomics 2008 [6]	-
CheseapeakeBayVir	viral	Aquatic	Ocean	Chesapeake Bay	GP 16522	Bench et al., Appl. Environ. Microbiol. 2007 [7]	112.38
ChickenCecum CJejuniMic2007	microbial	Host-associated	-	Chicken Cecum	GP 28599	Qu A et al., PLoS One, 2008 [8]	3,695.38
ChickenCecum UninfectedMic2007	microbial	Host-associated	-	Chicken cecum	GP 28597	Qu A et al., PLoS One, 2008 [8]	-
ChickenRunting StuntingMDnaVir2008	viral	Host-associated	-	Chicken cecum	GP 40783	-	39.77
ChickenRunting StuntingMRnaVir2008	viral	Host-associated	-	Chicken cecum	GP 40785	-	41.99
ChickenRunting StuntingPDnaV	viral	Host-associated	-	Chicken cecum	GP 40787	-	31.10

ir2008							
ChickenRuntingStuntingPRnaVir2008	viral	Host-associated	-	Chicken cecum	GP 40789	-	6.93
ConPorCompHawMic200602	microbial	Manipulated / Perturbed	-	Control Pcomp	GP 28429	Vega Thurber et al., Env. Mic. 2009 [9]	2,077.79
ConPorCompHawVir200602	viral	Manipulated / Perturbed	-	Control Pcomp	GP 28417	Vega Thurber et al., Proc. Nat. Acad. Sci. USA 2008 [10]	19.90
DesertSoilJoshuaTreeVir	viral	Terrestrial	-	-	GSS ER781257-ER785833	Fierer et al., Appl. Environ. Microb. 2007 [11]	-
DOCPorCompHawMic200602	microbial	Manipulated / Perturbed	-	DOC Pcomp	GP 28433	Vega Thurber et al., Env. Mic. 2009 [9]	2,630.08
DOCPorCompVirHaw200602	viral	Manipulated / Perturbed	-	DOC Pcomp	GP 28421	Vega Thurber et al., Proc. Nat. Acad. Sci. USA 2008 [11]	28.61
FannLIMic20050811	microbial	Aquatic	Ocean	Fanning island	GP 28367	Dinsdale et al., PLoS One 2008 [12]	2,638.17
FannLIVir20050811	viral	Aquatic	Ocean	Fanning island	GP 28369	Dinsdale et al., PLoS One 2008 [12]	152.09
FishHealGutKentSTMic20060504	microbial	Host-associated	-	Fish gut	GP 28389	Dinsdale et al., Nature 2008 [13]	5,234.77
FishHealGutKentSTVir20060504	viral	Host-associated	-	Fish gut	GP 28397	-	-
FishHealSlimKentSTMic20060504	microbial	Host-associated	-	Fish slime	GP 28393	Dinsdale et al., Nature 2008 [13]	5,229.78
FishHealSlimKentSTVir20060504	viral	Host-associated	-	Fish slime	GP 28401	Dinsdale et al., Nature 2008 [13]	13.23
FishMorGutKentSTMic20060504	microbial	Manipulated / Perturbed	-	Fish Morbid Gut	GP 28391	Dinsdale et al., Nature 2008 [13]	5,260.79
FishMorGutKentSTVir20060504	viral	Manipulated / Perturbed	-	Fish Morbid Gut	GP 28399	-	-
FishMorSlimKentSTMic20060504	microbial	Manipulated / Perturbed	-	Fish Morbid Slime	GP 28395	Dinsdale et al., Nature 2008 [13]	5,126.12
FishMorSlimKentSTVir200605	viral	Manipulated / Perturbed	-	Fish Morbid	GP 28403	Dinsdale et al., Nature 2008 [13]	31.57

04				Slime			
GOMVir1994to2001	viral	Aquatic	Ocean	Gulf of Mexico	GP 17765	Angly et al., PLoS Biology 2006 [4]	61.04
GS000a11Mic	microbial	Aquatic	Ocean	Sargasso Sea	GP 13694	Venter et al., Science 2004 [14]	5,465.57
GS000a13Mic	microbial	Aquatic	Ocean	Sargasso Sea	GP 13694	Venter et al., Science 2004 [14]	5,465.57
GS000b11Mic	microbial	Aquatic	Ocean	Sargasso Sea	GP 13694	Venter et al., Science 2004 [14]	1,948.84
GS000b13Mic	microbial	Aquatic	Ocean	Sargasso Sea	GP 13694	Venter et al., Science 2004 [14]	1,948.84
GS000cMic	microbial	Aquatic	Ocean	Sargasso Sea	GP 13694	Venter et al., Science 2004 [14]	2,292.36
GS000dMic	microbial	Aquatic	Ocean	Sargasso Sea	GP 13694	Venter et al., Science 2004 [14]	1,599.07
GS001aEuk	eukaryal	Aquatic	Ocean	-	GP 13694	Venter et al., [14] Science 2004	24,690.91
GS001bEuk	eukaryal	Aquatic	Ocean	-	GP 13694	Venter et al., Science 2004 [14]	6,921.47
GS011Mic	microbial	Aquatic	Ocean	-	GP 13694	Rush et al., PLoS Biol. 2007 [15]	1,543.51
GS012Mic	microbial	Aquatic	Ocean	Chesapeake Bay	GP 13694	Rush et al., PLoS Biol. 2007 [15]	1,658.76
GS016Mic	microbial	Aquatic	Ocean	Gulf of Mexico	GP 13694	Rush et al., PLoS Biol. 2007 [15]	1,564.53
GS020Mic	microbial	Aquatic	Ocean	-	GP 13694	Rush et al., PLoS Biol. 2007 [15]	2,896.19
GS023Mic	microbial	Aquatic	Ocean	-	GP 13694	Rush et al., PLoS Biol. 2007 [15]	1,671.19
GS025Euk	eukaryal	Aquatic	Ocean	-	GP 19735	Rush et al., PLoS Biol. 2007 [15]	895.31
GS034Mic	microbial	Aquatic	Ocean	-	GP 13694	Rush et al., PLoS Biol. 2007 [15]	2,002.14
GS048aMic	microbial	Aquatic	Ocean	-	GP 13694	Rush et al., PLoS Biol. 2007 [15]	1,718.17
GS048bEuk	eukaryal	Aquatic	Ocean	-	GP 13694	Rush et al., PLoS Biol. 2007 [15]	691.96
GS108bEuk	eukaryal	Aquatic	Ocean	-	GP 13694	-	686.97
GS110bEuk	eukaryal	Aquatic	Ocean	-	GP 13694	-	935.95
GS112bEuk	eukaryal	Aquatic	Ocean	-	GP 13694	-	782.20
GS117bEuk	eukaryal	Aquatic	Ocean	-	GP 13694	-	765.82
GS122bEuk	eukaryal	Aquatic	Ocean	-	GP 13694	-	719.78

GutlessWormMic	microbial	Host-associated	-	-	GB AASZ01000000	Woyke et al., Nature 2006 [16]	522.88
HBCStromBahamasMic20050111	microbial	Aquatic	Microbialites	Bahamas	GP 28383	Dinsdale et al., Nature 2008 [13]	4,125.02
HBCStromBahamasVir20050111	viral	Aquatic	Microbialites	Bahamas	GP 28381	Desnues et al., Nature 2008 [17]	4.63
HealSputRep3SDVir20060707	viral	Host-associated	-	-	GP 28439	Dinsdale et al., Nature 2008 [13]	-
HighSalternSDbayMic20051128	microbial	Aquatic	Hypersaline	High saltern SD 200511	GP 28453	Dinsdale et al., Nature 2008 [13]	3,407.29
HighSalternSDbayMicD200407	microbial	Aquatic	Hypersaline	High saltern SD	GP 40795	-	3,275.39
HighSalternSDbayVir20051116	viral	Aquatic	Hypersaline	High saltern SD	GP 28457	Dinsdale et al., Nature 2008 [13]	51.73
HighSalternSDbayVir20051128	viral	Aquatic	Hypersaline	High saltern SD 200511	GP 28451	Dinsdale et al., Nature 2008 [13]	267.95
HighSalternSDbayVir20051207	viral	Aquatic	Hypersaline	High saltern SD	GP 28447	Dinsdale et al., Nature 2008 [13]	-
Hot10Mic20021007	microbial	Aquatic	Ocean	-	GSS DU731018-DU796676, DU800850-DU800864	DeLong et al., Science 2006 [18]	1,838.16
Hot130Mic20021006	microbial	Aquatic	Ocean	-	GSS DU731018-DU796676, DU800850-DU800864	DeLong et al., Science 2006 [18]	3,849.65
Hot200Mic20021006	microbial	Aquatic	Ocean	-	GSS DU731018-DU796676, DU800850-DU800864	DeLong et al., Science 2006 [18]	1,868.80
Hot4000Mic20031221	microbial	Aquatic	Ocean	-	GSS DU731018-DU796676, DU800850-DU800864	DeLong et al., Science 2006 [18]	3,664.53
Hot500Mic10021006	microbial	Aquatic	Ocean	-	GSS DU731018-DU796676,	DeLong et al., Science 2006 [18]	3,981.13

					DU800850-DU800864		
Hot70Mic20021007	microbial	Aquatic	Ocean	-	GSS DU731018-DU796676, DU800850-DU800864	DeLong et al., Science 2006 [18]	2,157.18
Hot770Mic20031221	microbial	Aquatic	Ocean	-	GSS DU731018-DU796676, DU800850-DU800864	DeLong et al., Science 2006 [18]	2,597.30
HumanFecesSDVir	viral	Host-associated	-	-	GSS CC820769-CC821300	Breitbart et al., J. Bacteriol. 2003 [19]	-
KingLIMic20050821	microbial	Aquatic	Ocean	Kingman island	GP 28343	Dinsdale et al., PLoS One 2008 [12]	2,911.81
KingLIVir20050821	viral	Aquatic	Ocean	Kingman island	GP 28345	Dinsdale et al., PLoS One 2008 [12]	154.43
LeanMouseCecumMic2005	microbial	Host-associated	-	-	GP 17401	Turnbaugh et al., Nature 2006 [20]	3,144.82
LowSalternSDbayMic200407	microbial	Aquatic	Hypersaline	Low saltern 200407	GP 28359	Dinsdale et al., Nature 2008 [13]	4,564.80
LowSalternSDbayMic20051128	microbial	Aquatic	Hypersaline	Low saltern 200511	GP 28461	Dinsdale et al., Nature 2008 [13]	1,664.29
LowSalternSDbayVir200407	viral	Aquatic	Hypersaline	Low saltern 200407	GP 28353	Dinsdale et al., Nature 2008 [13]	67.90
LowSalternSDbayVir20051110	viral	Aquatic	Hypersaline	-	GP 28373	Dinsdale et al., Nature 2008 [13]	92.07
LowSalternSDbayVir20051128	viral	Aquatic	Hypersaline	Low saltern 200511	GP 28455	Dinsdale et al., Nature 2008 [13]	64.41
MarineBacterioplanktonS35131Euk	eukaryal	Aquatic	Ocean	-	CA BACTERIOPLANKTON_SMPL_S_35131	Hewson et al., Limnol Oceanogr 2009 [21]	804.70
MarineBacterioplanktonS35139Euk	eukaryal	Aquatic	Ocean	-	CA BACTERIOPLANKTON_SMPL_S_35139	Hewson et al., Limnol Oceanogr 2009 [21]	741.86
MarineBacterioplanktonS35147Euk	eukaryal	Aquatic	Ocean	-	CA BACTERIOPLANKTON_SMPL_S_35147	Hewson et al., Limnol Oceanogr 2009 [21]	809.65
MarineBacterio	eukaryal	Aquatic	Ocean	-	CA	Hewson et al.,	971.31

planktonS35155 Euk					BACTERIOPL ANKTON_SM PL_S_35155	Limnol Oceanogr 2009 [21]	
MarineBacterio planktonS35163 Euk	eukaryal	Aquatic	Ocean	-	CA BACTERIOPL ANKTON_SM PL_S_35163	Hewson et al., Limnol Oceanogr 2009 [21]	1,055.38
MarineBacterio planktonS35171 Euk	eukaryal	Aquatic	Ocean	-	CA BACTERIOPL ANKTON_SM PL_S_35171	Hewson et al., Limnol Oceanogr 2009 [21]	779.40
MarineBacterio planktonS35179 Euk	eukaryal	Aquatic	Ocean	-	CA BACTERIOPL ANKTON_SM PL_S_35179	Hewson et al., Limnol Oceanogr 2009 [21]	864.10
MediterraneanB athypelagicEuk	eukaryal	Aquatic	Ocean	-	GSS EI942868-EI951915	Martin-Cuadrado et al., PLoS ONE 2007 [22]	-
MedSalternSDb ayMic20051110	microbial	Aquatic	Hypersaline	Med saltern 20051110	GP 28377	Dinsdale et al., Nature 2008 [13]	4014.17
MedSalternSDb ayMic20051111	microbial	Aquatic	Hypersaline	Med saltern	GP 28379	Dinsdale et al., Nature 2008 [13]	3,451.03
MedSalternSDb ayMic20051116	microbial	Aquatic	Hypersaline	Med saltern 20051116	GP 28459	Dinsdale et al., Nature 2008 [13]	3,627.08
MedSalternSDb ayMic20051128	microbial	Aquatic	Hypersaline	Med saltern 20051128	GP 28449	Dinsdale et al., Nature 2008 [13]	3,613.68
MedSalternSDb ayVir20051110	viral	Aquatic	Hypersaline	Med saltern 20051110	GP 28375	Dinsdale et al., Nature 2008 [13]	72.77
MedSalternSDb ayVir20051116	viral	Aquatic	Hypersaline	Med saltern 20051116	GP 28465	Dinsdale et al., Nature 2008 [13]	61.41
MedSalternSDb ayVir20051122	viral	Aquatic	Hypersaline	Med saltern	GP 28445	Dinsdale et al., Nature 2008 [13]	56.80
MedSalternSDb ayVir20051128	viral	Aquatic	Hypersaline	Med saltern 20051128	GP 28463	Dinsdale et al., Nature 2008 [13]	-
Mosq1SDVir20 060125	viral	Host-associated	-	-	GP 28413	Dinsdale et al., Nature 2008 [13]	4.39
Mosq2SDVir20 60606	viral	Host-associated	-	-	GP 28469	Dinsdale et al., Nature 2008 [13]	4.46
Move858Vir	viral	Aquatic	Ocean	Chesapeake Bay	GP 13694	Rush et al., PLoS Biol. 2007 [15]	-
MushroomHotS pringMic	microbial	Aquatic	Hot spring	-	GP 20953	Bhaya et al., ISME J., 2007 [23]	2,966.33
Norm3LungVir 20080407	viral	Host-associated	-	-	GP 39545	Willner et al., PLoS ONE 2009 [24]	61.23

Norm4LungVir20080407	viral	Host-associated	-	-	GP 39545	Willner et al., PLoS ONE 2009 [24]	20.51
Norm5LungVir20080407	viral	Host-associated	-	-	GP 39545	Willner et al., PLoS ONE 2009 [24]	57.05
Norm6LungVir20080407	viral	Host-associated	-	-	GP 39545	Willner et al., PLoS ONE 2009 [24]	49.29
Norm7LungVir20080407	viral	Host-associated	-	-	GP 39545	Willner et al., PLoS ONE 2009 [24]	58.24
NutPorCompHawMic200602	microbial	Manipulated / Perturbed	-	Nutrient Pcomp	GP 28437	Vega Thurber et al., Env. Mic. 2009 [9]	2,083.43
NutPorCompHawVir200602	viral	Manipulated / Perturbed	-	Nutrient Pcomp	GP 28425	Vega Thurber et al., Proc. Nat. Acad. Sci. USA 2008 [10]	26.50
OctopusHotSpringMic	microbial	Aquatic	Hot spring	Yellowstone Octopus	GP 20953	Bhaya et al., ISME J., 2007 [17]	3,006.73
OctopusHotSpringVir	viral	Aquatic	Hot spring	Yellowstone Octopus	GP 20821	Pride et al., BMC Genomics 2008 [6]	-
OxMinZoneVir200806-200	viral	Aquatic	Ocean	-	GP 40793	-	-
OxMinZoneVir200806-90	viral	Aquatic	Ocean	-	GP 40791	-	93.13
PacificBeachSandEuk	eukaryal	Terrestrial	-	-	GP 13729	Naviaux et al., Mar Ecol Prog Ser 2005 [25]	-
PalmLIMic20050818	microbial	Aquatic	Ocean	Palmyra island	GP 28363	Dinsdale et al., PLoS One 2008 [12]	2,140.61
PalmLIVir20050818	viral	Aquatic	Ocean	Palmyra island	GP 28365	Dinsdale et al., PLoS One 2008 [12]	163.41
PANoVectorsBocasMic20050921	microbial	Host-associated	-	-	GP 28371	Wegley et al., Env. Mic. 2007 [26]	3,236.71
PAStromBahamasMic20050722	microbial	Aquatic	Microbialites	Pozas	GP 28385	Dinsdale et al., Nature 2008 [13]	4,504.51
PAStromCCMexVir20050722	viral	Aquatic	Microbialites	Pozas	GP 28355	Desnues et al., Nature 2008 [17]	17.26
PeruvianCoastalMarginMic	microbial	Aquatic	Ocean	-	SRA 001015	Biddle et al., Proc. Nat. Acad. Sci. USA 2008 [27]	2,521.81
pHPorCompHawMic200602	microbial	Manipulated / Perturbed	-	pH Pcomp	GP 28435	Vega Thurber et al., Env. Mic. 2009 [9]	2,178.71
pHPorCompHawVir200602	viral	Manipulated / Perturbed	-	pH Pcomp	GP 28423	Vega Thurber et al., Proc. Nat. Acad. Sci. USA 2008 [10]	39.92

PrarieSoilKonzaVir	viral	Terrestrial	-	-	GSS ER781257-ER785833	Fierer et al., Appl. Environ. Microb. 2007 [11]	-
PrePondKentSTMic20060504	microbial	Aquatic	Fresh water	Kent SeaTech Pre 200605	GP 28407	Dinsdale et al., Nature 2008 [13]	4,049.11
PrePondKentSTVir20060504	viral	Aquatic	Fresh water	Kent SeaTech Pre 200605	GP 28411	Dinsdale et al., Nature 2008 [13]	47.00
RainforestSoilPeruVir	viral	Terrestrial	-	-	GSS ER781257-ER785833	Fierer et al., Appl. Environ. Microb. 2007 [11]	-
RicePaddySoilVir	viral	Terrestrial	-	-	GP 20811	Kim et al., Appl. Environ. Microb. 2008 [28]	-
RMStromCCMexMic20050722	microbial	Aquatic	Microbialites	Mexico	GP 28351	Dinsdale et al., Nature 2008 [13]	5,061.41
RMStromCCMexVir20050722	viral	Aquatic	Microbialites	Mexico	GP 28357	Desnues et al., Nature 2008 [13]	50.27
Rumen640FMic20051215	microbial	Host-associated	-	-	GP 28607	Dinsdale et al., Nature 2008 [13]	3,331.03
Rumen710FMic20051215	microbial	Host-associated	-	-	GP 28609	Dinsdale et al., Nature 2008 [13]	2,938.60
Rumen80FMic20051215	microbial	Host-associated	-	-	GP 28605	Dinsdale et al., Nature 2008 [13]	1,867.34
SaltonSea1Vir20060823	viral	Sediments	-	Salton Sea	GP 28613	Dinsdale et al., Nature 2008 [13]	88.67
SaltonSea2Vir20060823	viral	Sediments	-	Salton Sea	GP 28613	Dinsdale et al., Nature 2008 [13]	82.61
SaltonSeaMic20060823	microbial	Sediments	-	Salton Sea	GP 28613	Dinsdale et al., Nature 2008 [13]	3,806.14
SARVir20050630	viral	Aquatic	Ocean	Sargasso Sea	GP 17771	Angly et al., PLoS Biology 2006 [4]	19.95
SeawaterMissionBaySDVir	viral	Aquatic	Ocean	-	GSS BH898061-BH898933	Breitbart et al., Proc. Nat. Acad. Sci. USA 2002 [29]	-
SeawaterScrippsSDVir	viral	Aquatic	Ocean	-	GSS AY079522-AY080585	Breitbart et al., Proc. Nat. Acad. Sci. USA 2002 [29]	-
SedimentsMissionBaySDVir	viral	Sediments	-	-	GB CC821301-CC822456	Breitbart et al., Proc. R. Soc. B. 2004 [30]	-
SkanBayAlaskaVir20060927	viral	Aquatic	Ocean	-	GP 28619	Dinsdale et al., Nature 2008 [13]	-
SoilCC1Mic	microbial	Terrestrial	-	-	MG 4441690.3	-	6,063.12

SoilCP1Mic	microbial	Terrestrial	-	-	MG 4441689.3	-	5,927.76
SoilCP3Mic	microbial	Terrestrial	-	-	MG 4441691.3	-	5,776.93
SoilHF1Mic	microbial	Terrestrial	-	-	MG 4441642.3	-	5,887.77
SoilKP1Mic	microbial	Terrestrial	-	-	MG 4441994.3	-	5,952.04
SoilKP3Mic	microbial	Terrestrial	-	-	MG 4441643.3	-	5,777.56
SoilKW1Mic	microbial	Terrestrial	-	-	MG 4441664.3	-	5,741.19
SoilKW2Mic	microbial	Terrestrial	-	-	MG 4441691.4	-	5,766.49
SoilLF1Mic	microbial	Terrestrial	-	-	MG 4442455.3	-	6,181.55
SoilLF2Mic	microbial	Terrestrial	-	-	MG 4441644.3	-	6,131.82
SoilSJ1Mic	microbial	Terrestrial	-	-	MG 4441557.3	-	5,995.31
SoilSJ2Mic	microbial	Terrestrial	-	-	MG 4441645.3	-	5,974.06
SoilTF1Mic	microbial	Terrestrial	-	-	MG 4441688.3	-	6,119.44
SoilTF2Mic	microbial	Terrestrial	-	-	MG 4442452.3	-	6,312.57
SoilWF1Mic	microbial	Terrestrial	-	-	MG 4441556.3	-	5,812.88
SoilWF2Mic	microbial	Terrestrial	-	-	MG 4441665.3	-	5,632.30
SoilYN1Mic	microbial	Terrestrial	-	-	MG 4442453.3	-	5,960.67
SoilYN2Mic	microbial	Terrestrial	-	-	MG 4441687.3	-	5,899.56
SouthChinaSeaSedimentsMic	microbial	Sediments	-	-	GP 33581	-	4,941.46
T0PortComHawMic20060223	microbial	Host-associated	-	Coral tissue P. compressa	GP 28427	Vega Thurber et al., Env. Mic. 2009 [9]	3,618.00
T0PortComHawVir20060223	viral	Host-associated	-	Coral tissue P. compressa	GP 28415	Vega Thurber et al., Proc. Nat. Acad. Sci. USA 2008 [10]	43.07
TempPorCompHawMic200602	microbial	Manipulated / Perturbed	-	Temperature Pcomp	GP 28431	Vega Thurber et al., Env. Mic. 2009 [9]	4,560.68
TempPorCompHawVir200602	viral	Manipulated / Perturbed	-	Temperature Pcomp	GP 28419	Vega Thurber et al., Proc. Nat. Acad. Sci. USA 2008 [10]	30.80
TilPondKentSTMic200511	microbial	Aquatic	Fresh water	Kent SeaTech Tpond 200511	GP 28387	Dinsdale et al., Nature 2008 [13]	4,204.22
TilPondKentSTMic20060504	microbial	Aquatic	Fresh water	Kent SeaTech Tpond 200605	GP 28405	Dinsdale et al., Nature 2008 [13]	4,007.80
TilPondKentSTMic200608	microbial	Aquatic	Fresh water	Kent SeaTech Tpond 200608	GP 28603	Dinsdale et al., Nature 2008 [13]	4,689.22

TilPondKentSTVir20060504	viral	Aquatic	Fresh water	Kent SeaTech Tpond 200605	GP 28409	Dinsdale et al., Nature 2008 [13]	34.45
TilPondKentSTVir200608	viral	Aquatic	Fresh water	Kent SeaTech Tpond 200608	GP 28601	Dinsdale et al., Nature 2008 [13]	51.03
TpondKentSTVir200511	viral	Aquatic	Fresh water	Kent SeaTech Tpond 200511	GP 28361	Dinsdale et al., Nature 2008 [13]	35.71
WasecaFarmSoilMic	microbial	Terrestrial	-	-	GP 13699	Tringe SG et al., Science, 2005 [31]	5,350.40
XmasLIMic20050805	microbial	Aquatic	Ocean	Christmas island	GP 28347	Dinsdale et al., PLoS One 2008 [12]	3,483.52
XmasLIVir20050805	viral	Aquatic	Ocean	Christmas island	GP 28349	Dinsdale et al., PLoS One 2008 [12]	158.48

References:

1. Allen HK, Moe LA, Rodbumrer J, Gaarder A, Handelsman J (2008) Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. ISME J 3: 243-251.

2. Legault B, Lopez-Lopez A, Alba-Casado J, Doolittle WF, Bolhuis H, et al. (2006) Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. BMC Genomics 7: 171.

3. Grzymski JJ, Murray AE, Campbell BJ, Kaplarevic M, Gao GR, et al. (2008) Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility. Proceedings of the National Academy of Sciences 105: 17516-17521.

4. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. PLoS Biology 4: e368.

5. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, et al. (2008) Viral diversity and dynamics in an infant gut. Res Microbiol 159: 367-373.

6. Pride D, Schoenfeld T (2008) Genome signature analysis of thermal virus metagenomes reveals Archaea and thermophilic signatures. BMC Genomics 9: 420.

7. Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, et al. (2007) Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol 73: 7629-7641.

8. Qu A, Brulc JM, Wilson MK, Law BF, Theoret JR, et al. (2008) Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. PLoS ONE 3: e2945.

9. Thurber RV, Willner-Hall D, Rodriguez-Mueller B, Desnues C, Edwards RA, et al. (2009) Metagenomic analysis of stressed coral holobionts. Environ Microbiol 11: 2148-2163.

10. Vega Thurber RL, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, et al. (2008) Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. Proc Nat Acad Sci USA 105: 18413-18418.

11. Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, et al. (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of Bacteria, Archaea, Fungi, and viruses in soil. Appl Environ Microbiol 73: 7059-7066.

12. Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, et al. (2008) Microbial ecology of four coral atolls in the northern Line Islands. PLoS ONE 3: e1584.

13. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629-632.

14. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.

15. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5: e77.

16. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, et al. (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443: 950-955.

17. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, et al. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452: 340-343.

18. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496-503.

19. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185: 6220-6223.

20. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-131.

21. Hewson I, Paerl RW, Tripp HJ, Zehr JP, Karl DM (2009) Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. Limnol Oceanogr 54: 1981-1994.

22. Martín-Cuadrado A, López-García P, Alba J, Moreira D, Monticelli L, et al. (2007) Metagenomics of the deep mediterranean, a warm bathypelagic habitat. PLoS ONE 2: e914.

23. Bhaya D, Grossman AR, Steunou A, Khuri N, Cohan FM, et al. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J 1: 703-713.

24. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis Individuals. PLoS ONE 4: e7370.

25. Naviaux RK, Good B, McPherson JD, Steffen DL, Markusic D, et al. (2005) Sand DNA - a genetic library of life at the water's edge. Mar Ecol Prog Ser 301: 9-22.

26. Wegley L, Edwards R, Rodriguez-Brito B, Liu H, Rohwer F (2007) Metagenomic analysis of the microbial community associated with the coral Porites astreoides. Environmental Microbiology 9: 2707-2719.

27. Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, House CH (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proc Nat Acad Sci USA 105: 10583-10588.

28. Kim K, Chang H, Nam Y, Roh SW, Kim M, et al. (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl Environ Microbiol 74: 5975-5985.

29. Breitbart M, Salamon P, Andresen B, Mahaffy J, Segall A, et al. (2002) Genomic analysis of uncultured marine viral communities. Proc Nat Acad Sci USA 99: 14250-14255.

30. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, et al. (2004) Diversity and population structure of a near-shore marine-sediment viral community. Proc R Soc B 271: 565–574.

31. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554-557.