

## Review

## A Primer on Metagenomics

John C. Wooley<sup>1</sup>, Adam Godzik<sup>1,2</sup>, Iddo Friedberg<sup>3,4\*</sup>

**1** Community Cyberinfrastructure for Marine Microbial Ecology Research and Analysis, California Institute for Telecommunications and Information Technology, University of California San Diego, La Jolla, California, United States of America, **2** Program in Bioinformatics and Systems Biology, Burnham Institute for Medical Research, La Jolla, California, United States of America, **3** Department of Microbiology, Miami University, Oxford, Ohio, United States of America, **4** Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio, United States of America

**Abstract:** Metagenomics is a discipline that enables the genomic study of uncultured microorganisms. Faster, cheaper sequencing technologies and the ability to sequence uncultured microbes sampled directly from their habitats are expanding and transforming our view of the microbial world. Distilling meaningful information from the millions of new genomic sequences presents a serious challenge to bioinformaticians. In cultured microbes, the genomic data come from a single clone, making sequence assembly and annotation tractable. In metagenomics, the data come from heterogeneous microbial communities, sometimes containing more than 10,000 species, with the sequence data being noisy and partial. From sampling, to assembly, to gene calling and function prediction, bioinformatics faces new demands in interpreting voluminous, noisy, and often partial sequence data. Although metagenomics is a relative newcomer to science, the past few years have seen an explosion in computational methods applied to metagenomic-based research. It is therefore not within the scope of this article to provide an exhaustive review. Rather, we provide here a concise yet comprehensive introduction to the current computational requirements presented by metagenomics, and review the recent progress made. We also note whether there is software that implements any of the methods presented here, and briefly review its utility. Nevertheless, it would be useful if readers of this article would avail themselves of the comment section provided by this journal, and relate their own experiences. Finally, the last section of this article provides a few representative studies illustrating different facets of recent scientific discoveries made using metagenomics.

## Introduction

For most of its history, life on Earth consisted solely of microscopic life forms, and microbial life still dominates Earth in many aspects. The estimated  $5 \times 10^{30}$  prokaryotic cells inhabiting our planet sequester some 350–550 Petagrams (1 Pg =  $10^{15}$  g) of carbon, 85–130 Pg of nitrogen, and 9–14 Pg of phosphorous making them the largest reservoir of those nutrients on Earth [1]. Bacteria and archaea live in all environments capable of sustaining other life and in many cases are the sole inhabitants of extreme environments: from deep sea vents with temperatures of  $340^\circ\text{C}$  to rocks found in boreholes 6 km beneath the Earth's surface. Bacteria, archaea, and microeukaryotes dominate Earth's habitats, compound recycling, nutrient sequestration, and, according to some estimates, biomass. Microbes are not only ubiquitous, they are essential to all life, as they are the primary source for nutrients, and the primary recyclers of dead matter back to available organic form. Along with all other animals and plants, the human condition is profoundly affected by microbes, from the scourges of

human, farm animal, and crop pandemics, to the benefits in agriculture, food industry, and medicine to name a few. We humans have more bacterial cells ( $10^{14}$ ) inhabiting our body than our own cells ( $10^{13}$ ) [2,3]. It has been stated that the key to understanding the human condition lies in understanding the human genome [4,5]. But given our intimate relationship with microbes [6], researching the human genome is now understood to be a necessary though insufficient condition: sequencing the genomes of our own microbes would be necessary too. Also, to better understand the role of microbes in the biosphere, it would be necessary to undertake a genomic study of them as well.

The study of microbial genomes started in the late 1970s, with the sequencing of the genomes of bacteriophages MS2 [7] and  $\phi$ -X174 [8]. In 1995 microbiology took a major step with the sequencing of the first bacterial genome *Haemophilus influenza* [9]. The genomes of 916 bacterial, 1,987 viral, and 67 archaeal species are deposited in GenBank release 2.2.6. Having on hand such a large number of microbial genomes has changed the nature of microbiology and of microbial evolution studies. By providing the ability to examine the relationship of genome structure and function across many different species, these data have also opened up the fields of comparative genomics and of systems biology. Nevertheless, single organism genome studies have limits. First, technology limitations mean that an organism must first be clonally cultured to sequence its entire genome. However, only a small percentage of the microbes in nature can be cultured, which means that extant genomic data are highly biased and do not represent a true picture of the genomes of microbial species [10–12]. Second, very rarely do microbes live in single species communities: species interact both with each other and with their habitats, which may also include host organisms. Therefore, a clonal culture also fails to represent the true state of affairs in nature with respect to organism interaction, and the resulting population genomic variance and biological functions.

**Citation:** Wooley JC, Godzik A, Friedberg I (2010) A Primer on Metagenomics. PLoS Comput Biol 6(2): e1000667. doi:10.1371/journal.pcbi.1000667

**Editor:** Philip E. Bourne, University of California San Diego, United States of America

**Published:** February 26, 2010

**Copyright:** © 2010 Wooley et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** The authors acknowledge funding from the Gordon and Betty Moore Foundation (<http://www.moore.org>), grant name Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA). IF acknowledges funding from Miami University start up funds. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: [i.friedberg@muohio.edu](mailto:i.friedberg@muohio.edu)

New sequencing technologies and the drastic reduction in the cost of sequencing are helping us overcome these limits. We now have the ability to obtain genomic information directly from microbial communities in their natural habitats. Suddenly, instead of looking at a few species individually, we are able to study tens of thousands all together. Sequence data taken directly from the environment were dubbed the metagenome [13], and the study of sequence data directly from the environment—metagenomics [14].

However, environmental sequencing comes with its own information-restricting price tag. In single organism genomics practically all of the microbe's genome is sequenced, providing a complete picture of the genome. We know from which species the DNA or RNA originated. After assembly, the location of genes, operons, and transcriptional units can be computationally inferred. Control elements and other cues can be identified to infer transcriptional and translational units. Consequently, we achieve a nearly complete and well-ordered picture of all the genomic elements in the sequenced organism. We may not recognize all the elements for what they are, and some errors may creep in, but we can gauge the breadth of our knowledge and properly annotate those areas of the genome we manage to decipher.

In contrast, the sequences obtained from environmental genomic studies are fragmented. Each fragment was obviously sequenced from a specific species, but there can be many different species in a single sample, for most of which a full genome is not available. In many cases it is impossible to determine the true species of origin. The length of each fragment can be anywhere between 20 base pairs (bp) and 700 bp, depending on the sequencing method used. Short sequence reads that are dissociated from their original species can be assembled to lengths usually not exceeding 5,000 bp; consequently, the reconstruction of a whole genome is generally not possible. Even the reconstruction of an entire transcriptional unit can be problematic. In addition to being fragmented and incomplete, the volume of sequence data acquired by environmental sequencing is several orders of magnitude larger than that acquired in single organism genomics.

For these reasons, computational biologists have been developing new algorithms to analyze metagenomic data. These computational challenges are new and very exciting. We are entering an era akin to that of the first genomic revolution almost two decades ago. Whole organism genomics allows us to examine the evolution not only of single genes, but of whole transcriptional units, chromosomes, and cellular networks. But more recently, metagenomics gave us the ability to study, on the most fundamental genomic level, the relationship between microbes and the communities and habitats in which they live. How does the adaptation of microbes to different environments, including host animals and other microbes, manifest itself in their genomes?

For us humans, this question can strike very close to home, when those habitats are our own bodies and the microbes are associated with our own well-being and illnesses: almost every aspect of human life, as well as the life of every other living being on the planet, is affected by microbes. We now have the experimental technology to understand microbial communities and how they affect us, but the sheer volume and fragmentary nature of the data challenge computational biologists to distill all these data into useful information.

In this article we shall briefly outline some experimental, technological, and computational achievements and challenges associated with metagenomic data, from sequence generation and assembly through the various levels of metagenomic annotation. We will also discuss computational issues that are unique to

environmental genomics, such as estimating the metagenome size and the handling of associated metadata. Finally, we will review some studies highlighting the advantages of metagenomic-based research, and some of the insights it has enabled.

## Sampling

### Sample Size and Number of Samples

The first step in a metagenomic study is to obtain the environmental sample. Samples should represent the population from which they are taken. The problem in microbial ecology is that we are unable to see the organisms we are trying to capture. How many samples are enough?

To estimate the fraction of species sequenced, rarefaction curves are typically used. A rarefaction curve plots the number of species as a function of the number of individuals sampled. The curve usually begins with a steep slope, which at some point begins to flatten as fewer species are being discovered per sample: the gentler the slope, the less contribution of the sampling to the total number of operational taxonomic units or OTUs. For microbial samples, different OTUs are typically characterized by 16S (prokaryotic) or 18S (eukaryotic) rDNA, and are also referred to as ribotypes. Classification is rarely done in the field, so some initial estimate of species diversity by a pilot study or previous studies is desirable to gauge the number of samples needed to get a comprehensive picture of the OTUs in the sampled habitat. More of this will be discussed in the “Species Diversity” section below.

### Filtering

When filtering an environmental sample, as with any kind of filtering, the goals are: (1) get as much as you can of what you want and (2) leave out as much as you can from what you do not want. So if we are interested in bacteria only, our goal would be to filter out the smaller viroid particles, and the usually larger protists. Of course, this process will leave in the lysogenic phages and prophages, which are integrated in bacterial genetic material, as well as mimivirus particles, which are as large as some bacteria. On the other side of the size scale, small protists and large bacteria may overlap in size, making a full size-based separation impossible. Also, filamentous forms of bacteria that grow in multicellular colonies may also be filtered out owing to colony size exceeding that of the filter's pores.

Computational filtering can be used after sequencing. Genomic material that is obviously within the clades of interest can be filtered in using similarity searches against annotated sequence databases. Care must be taken, though, with false negatives: relevant genomic material may be filtered out in this fashion simply because homologs have never been deposited in existing databases. Another option would be to search for obviously false-positive sequence motifs, e.g., eukaryote material when only prokaryote material is to be analyzed. This technique can also be used to detect sample contamination.

### Recording Metadata

Keeping strict and comprehensive records of metadata is as important as the sequence data. Metadata are the “data about the data”: where the samples were taken from, when, and under which conditions. In microbial ecology, this commonly refers to physical, chemical, and other environmental characteristics of the sample's location. For example, an ocean sample metadata will typically include sampling date and time, depth, salinity, light intensity, geographical coordinates, pH, soluble gases, etc. In clinical microbiology, metadata would refer to the pathology,

medical history, and vital statistics of the patient as well as the exact location and tissue from which the sample was taken, the sampling conditions, and so on.

Many metagenomic studies are driven by discovery and data mining, rather than by hypothesis. These studies seek statistically significant correlations between the metagenomic data and the habitat-associated metadata, which may lead to biologically significant discoveries. There is therefore a need to provide metadata in a form that is standard, comprehensive, and amenable to computation. For example, semantic information should be provided, wherever possible, in ontological form. A description of the environmental context and the experimental methods used is vital to enable comparative studies. As we shall see, genes or even “gene-less” sequence signatures are linked to habitats rather than to species. Finally, sequencing technology is rapidly improving, and the adoption of new sequencing methods will require the adoption of descriptors of those methods such as sequence coverage, quality, assembly programs that were used, and so on.

The Genomic Standards Consortium (<http://gensc.org/>) is an international group working to standardize the description of genomes and metagenomes and the exchange of genomic data and metadata. In a recent publication, a standard for the Minimum Information of Genomic and Metagenomic (MIGS/MIMS) metadata was suggested for adoption [15], and an associated markup language, the Genomics Contextual Data Markup Language or GCDML is under active development [16]. It is the consortium’s aim that the MIGS/MIMS shall be adopted by journals as a publication requirement when genomic or metagenomic data are being deposited, akin to standards such as MIAME for microarray data [17] or PDB/mmCIF for structural biology data [18].

## Sequencing

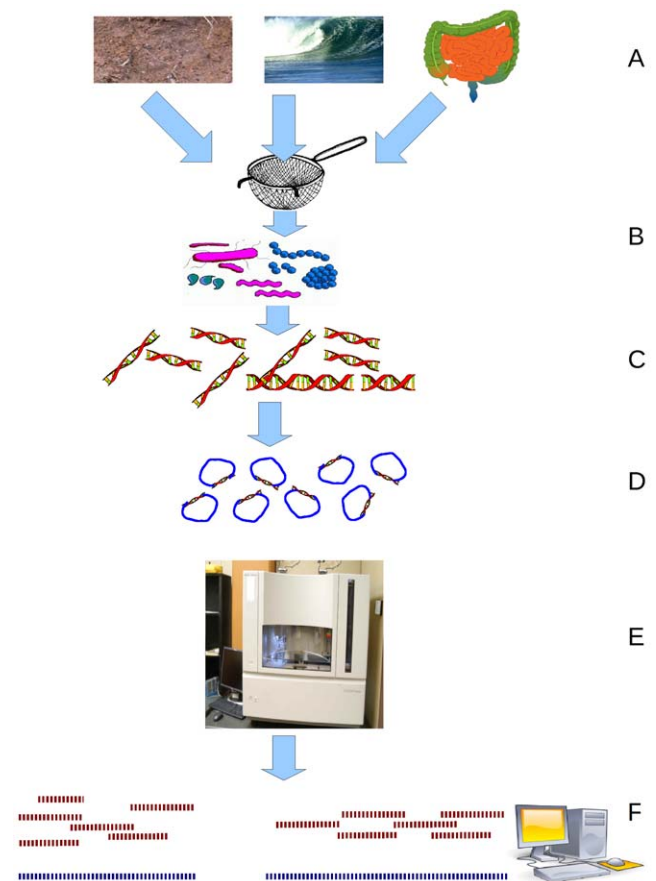
### First, Second, and Third Generation Sequencing

Until recently, prokaryotic genomes have been typically sequenced using Sanger shotgun [19,20] sequencing. The first step is shearing the DNA content of a genomic clone into random fragments, hence the “shotgun.” The fragments are then cloned into plasmid vectors that are grown in monoclonal libraries to produce enough genomic material for sequencing. The DNA is then sequenced using dye-termination methods. Repetition of this process ensures that all parts of the studied genome are sequenced, several times over. Assembly software is then used to assemble the sequence fragments into the whole genome. Theoretically any genome shorter than 5 Mbp can be assembled this way, although regions with large repeats tend to frustrate assembly algorithms. Therefore, regions with large repeats are often not incorporated into the whole genomic picture, leaving some gaps. Another disadvantage of shotgun sequencing is the “cloning bias.” Some genes cannot be incorporated into the library vector, usually because of toxicity to the vector expressing them [21]. This inability to be incorporated is typically mitigated by using more than one organism for cloning, or by using sequencing techniques that do not require cloning (see below) in second generation sequencing.

In metagenomics, shotgun sequencing is done in the same manner as in clonal culture genomics. However, the raw genomic material does not come from a single organism: it comes from a community of microbes, hence the name environmental shotgun sequencing or ESS. Depending on our ability to sample, this DNA may provide only a partial genomic picture of the organisms in the environment, since the genomic material from the more abundant species dominates the sample. To obtain a better picture of the

species composing the community, 16S rDNA or 18S rDNA for prokaryotic and eukaryotic samples, respectively, are sequenced separately using universal primers, see Figure 1. It should be noted that when using primers for rDNA to classify OTUs in an environmental sample, there are choices to be made regarding the primer sequence, especially when the studied OTU composition is expected to differ significantly from most known species, the so-called rare biosphere [22,23]. In this case, there is the possibility that the primers used will be too different from the rDNA in the sample, which would result in many OTUs not being identified [24,25].

Second generation sequencing methods have been rapidly gaining ground and are replacing Sanger sequencing for small sized genomes and environmental genomics. A common denominator among second generation methods is the generation of “polymerase colonies” or polonies [26,27]. Polonies are PCR amplicons derived from a single molecule of nucleic acid. Thousands to millions of polonies, each with an effective reaction size of  $10^{-9}$  l to  $10^{-12}$  l can be amplified simultaneously, generating templates for sequencing. Following that, enzymatic reactions can be performed in parallel to sequence the nucleic acid material in the polonies. Polony-based methods produce considerably more sequences than Sanger sequencing, but those sequences are much shorter. Furthermore, each polony-based method has its own anomalies that should be accounted for when



**Figure 1. Environmental Shotgun Sequencing (ESS).** (A) Sampling from habitat; (B) filtering particles, typically by size; (C) DNA extraction and lysis; (D) cloning and library; (E) sequence the clones; (F) sequence assembly.

doi:10.1371/journal.pcbi.1000667.g001

**Table 1.** Comparison of different sequencing technologies, taken from [34].

Sequencer	ABI 3730	Roche 454	Solexa <sup>a</sup>	SOLiD (mp, frag) <sup>b</sup>	HeliScope <sup>c</sup>
Read length	600–900	400–500	75–100	50	25–35
Run time	6–10 h	10 h	2–10 d	(4–7 d, 8–14 d)	h
Yield (Mbp)	0.01	1	2,300–3,500/d	(500, 1,000)	105–140/h
Cloning bias	Yes	No	No	No	No
Mate pair information	Yes	No	Yes	Yes	No

<sup>a</sup>Based on the GA IIx. See full specifications at: [http://www.illumina.com/systems/genome\\_analyzer.ilmn](http://www.illumina.com/systems/genome_analyzer.ilmn).

<sup>b</sup>mp, mate pair; frag, fragment. See <https://products.appliedbiosystems.com/SOLiD3PlusSystem>.

<sup>c</sup>See: <http://www.helicobio.com/Products/HelicoregGeneticAnalysisSystem/HeliScopetradeSequencer/tabid/87/Default.aspx>.

doi:10.1371/journal.pcbi.1000667.t001

processing the data. See Table 1 for a comparison between the yield, fragment length, and run times of the different sequencers.

In pyrosequencing (Figure 2) [28,29], methods such as Roche 454 [30] sequencing is performed by polymerase extension of a primed template. Single nucleotide species are added at each cycle. If the particular nucleotide species added to the polymerase reaction pairs with the one on the template, the incorporation causes luciferase-based light reaction. The reaction chamber is then washed, and the cycle repeated. Several hundreds of thousands of wells containing material for sequencing are typically used in a single reaction. Second is the inability to read long mononucleotide repeats correctly.

ABI SOLiD and Illumina GAI sequencers produce even shorter reads: 25–100 bp, but very large volumes of DNA per sequencing run. As we shall see in the “Assembly” section below, despite the individual short read lengths, these technologies provide a viable alternative for sequencing whole genomes, by sheer volume of DNA sequenced. For further reading on second generation sequencing see [31–34].

Third generation sequencing, loosely defined as technology that is capable of sequencing long sequences without amplification, is in advanced development. There are encouraging signs that this technology might be available as early as 2011 [35–37].

## Assembly

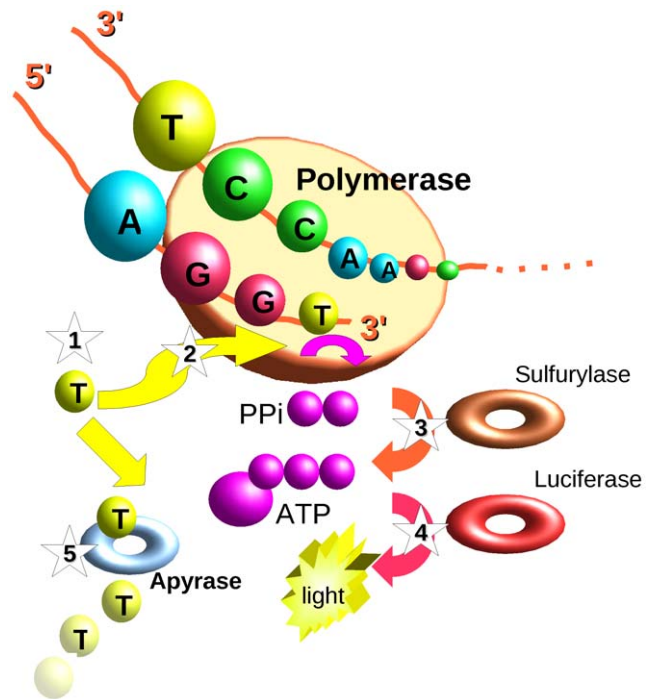
When sequencing a whole genome, the reads are assembled into progressively longer contiguous sequences or contigs, and finally to the whole genome. Dealing with genomic data, we are used to analyzing long stretches of contiguous sequence data. This analysis lets us find not only open reading frames, but also operons, operational transcriptional units, their associated promoter elements, and transcription factor binding sites. Longer elements such as pathogenicity islands, and other mobile genetic elements are evident only when large fractions of the genome are assembled. The gain of information correlates with the length of the genomic elements. Table 2 shows the length of a genomic sequence, and the information that may be gleaned from it.

In contrast, in all but the most species-poor metagenome, a full assembly is not possible—first, because the sampling is incomplete, and many if not all species’ genomes are partially sampled, if at all; second, because the species information itself is incomplete, and it is difficult to map individual reads to their species of origin. Therefore, the analysis of genomic elements using metagenomic data is generally limited to the first three or four rows in Table 2.

In this section, we will discuss assembly of metagenomic data, how information is extracted from partial assemblies, and how the extent of information gained can be estimated.

## Metagenomic Sample Coverage

**Coverage.** Coverage of a genome is defined as the mean number of times a nucleotide is being sequenced. Thus, 5× coverage means that each nucleotide in the genome is sequenced a mean number of five times. If we could sequence a genome in a single read, then 1× coverage would suffice for sequencing. Shorter read lengths (25–700, depending on sequencing technologies, see Table 1), necessitate more coverage, to ensure all reads overlap, and that those overlaps are unique enough to



**Figure 2. Pyrosequencing.** Single stranded DNA template is first hybridized with the sequencing primer and mixed with the enzymes along with the two substrates adenosine 5′-phosphosulfate (APS) and luciferin. In each cycle, (1) one of the four nucleotides (dTTPi, in this case) is then added to the reaction. (2) If the nucleotide is complementary to the base in the template strand then the DNA polymerase incorporates it into the growing strand. (3) Pyrophosphate (PPi)—in an amount equal in molarity to that of the incorporated nucleotide—is released and converted to ATP by sulfurylase in the presence of APS. (4) ATP then serves as a substrate to luciferase, causing a light reaction. Photon emission is in equimolar quanta to the amount of nucleotide incorporated in a given cycle. (5) The excess nucleotides are degraded by apyrase.

doi:10.1371/journal.pcbi.1000667.g002

**Table 2.** The information contained in different lengths of genomic DNA.

Sequence Length (bp)	Genome Element
25–75	SNPs, short frameshift mutations
100–400	Short functional signatures
500–1,000	Whole domains, single domain genes
1,000–5,000	Short operons, multidomain genes
5,000–10,000	Longer operons, some <i>cis</i> -control elements
>100,000	Prophages, pathogenicity islands, various mobile insertion elements
>1,000,000	Whole prokaryotic chromosome organization

doi:10.1371/journal.pcbi.1000667.t002

reconstruct the genome by assembling the fragments. If we treat DNA shearing and sequencing as random events, and our ability to detect and overlap between two truly overlapping reads does not vary between clones (when those are used), then we can use a Poisson distribution model to estimate the number of reads required to sequence an entire genome. This model is given by the Lander-Waterman equation [38]:

$$C = \frac{L \times N}{G}$$

Where  $L$  is the read length,  $N$  is the number of reads,  $G$  is the genome length, and  $C$  is coverage as described above. The fraction of sequence covered would be given as:

$$P_0 = 1 - e^{-C} = 1 - e^{-(LN/G)}$$

To get the number of reads sequencing fraction  $P_0$  of the genome

$$N = -\frac{\log(1 - P_0)}{L} \times G$$

In an environmental sample containing  $l$  species, the metagenome size  $G_m$  is:

$$G_m = \sum_{i=1}^l n_i G_i$$

Where  $G_i$  is the size of any given genome in a sample containing  $l$  genomes, and  $n_i$  the number of copies of genome  $g_i$ .

However the species that constitute the sample appear in different frequencies in the metagenome. Therefore a metagenome of size  $G_m$  composed of genomes of sizes  $G_1$  through  $G_k$  can be viewed as a sum of fractions. Each component genome of size  $G_i$  constitutes a fraction of  $G_m$ :

$$\hat{G}_m = p_1 G_m + p_2 G_m + \dots + p_l G_m$$

and:

$$\sum_{i=1}^l p_i = 1$$

Where  $p_i$  is the fraction of copies of the genome of species  $i$  in the sample and  $G_i$  is the size of the genome of species  $i$ .

Using species-specific gene markers, usually small ribosomal subunit rDNA, it is possible to estimate the species diversity in the sample, and provide an estimate of the different  $p_i$  values. Nevertheless, full or sometimes even adequate coverage (as judged by the rarefaction curve) of a species-rich environmental sample may be unattainable, especially for the genomes of the less represented species [39–42]. We expand upon this subject in the “Species Diversity” section.

Jeroen Raes and his colleagues have suggested an effective genome size or EGS measure that includes multiple plasmid copies, inserted sequences, and associated phages and viruses [43]. EGS uses the density (counts per megabase) of single copy marker genes to extrapolate the EGS.

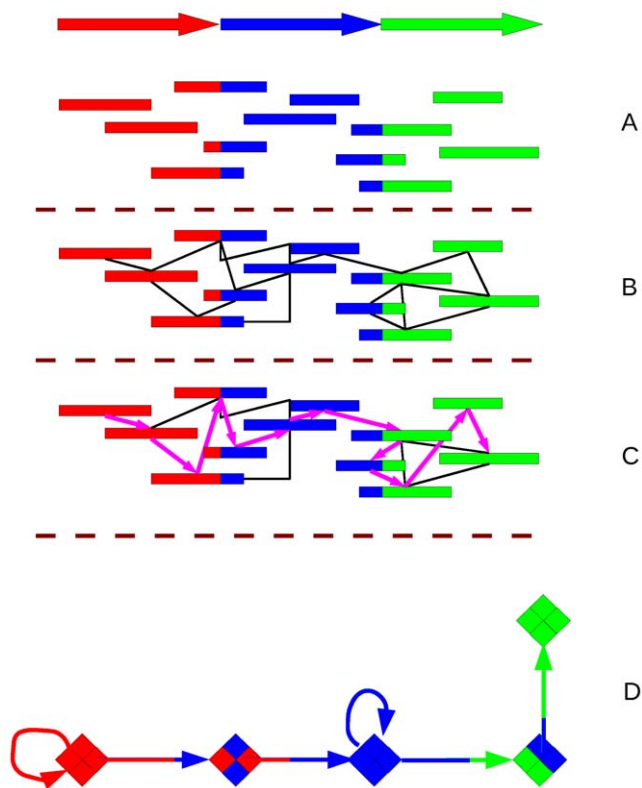
$$EGS = \frac{a + b \times L^{-c}}{x}$$

Where  $L$  is the read length,  $x$  is the marker gene density,  $a$ ,  $b$ , and  $c$  are empirical parameters empirically derived from 154 simulated metagenomes and found to be 21.2, 4,230, and 0.733, respectively. Raes and colleagues derived this formula from several different metagenomes, providing a useful measure of central tendency for genome size using a metagenomic sample. Note that  $a$ ,  $b$ , and  $c$  were derived from simulated metagenomes, Therefore, care must be taken in using the EGS formula above, since the parameters given only provide a snapshot of a particular simulation. It is probably better to use EGS as a framework, in conjunction with a metagenomic simulator such as MetaSim [44] to generate parameters more compatible with population estimates in one’s own research. MetaSim enables the creation of a simulated genome from regular genomic files; this makes it useful for testing and assessing the performance of other programs that manipulate and analyze metagenomic data, such as assembly or annotation programs.

## Metagenome Assembly

In a genome project of a single organism or clone we can be certain that all extracted DNA fragments belong to the same genome, barring contaminants and extrachromosomal DNA. That is not the case when a metagenome is concerned. As we have just seen, coverage is usually incomplete, since environmental sequence sampling rarely produces all the sequences required for assembly. Furthermore, there is also the danger of assembling sequences from different OTUs, creating interspecies chimeras. Phrap, Forge, Arachne [45], JAZZ [46], and the Celera Assembler [47] are all assembly programs that were developed for single genome assembly from Sanger sequencing. They seem to provide good results even when assembling metagenomic sequence data from Sanger sequencing [48]. Most of these algorithms use mate-pair information for the assemblies. This information is used in assembly to check the scaffolds or the assembled intermediaries between raw reads and whole chromosomes. These assembly algorithms represent each read as a vertex and each detected overlap as an edge between the overlapping vertices. Finding the correct assembly is cast as a Hamiltonian path finding problem, for finding a path in a graph where each vertex is visited once (see Figure 3A–3C).

For short reads, however, this technique is not suitable. To establish adequate coverage, short reads need to be produced in large quantities, and their short lengths means that there are many identical, or nearly identical, reads. The plethora of reads makes



**Figure 3. Fragment assembly.** (A–C) Hamiltonian. (A) A sequence with overlapping reads; (B) Each read is represented as a vertex, with edges connecting the overlapping vertices; (C) the assembly solution is a Hamiltonian path (all vertices are visited, no vertex is visited more than once) through the resulting graph; (D) For short reads assembly, each vertex is a  $k$ -mer (or a hashed collection of  $k$ -mers), and the reads are threaded between vertices as edges. The solution is a Eulerian path, where each edge is visited once. Repeats are merged into a single edge. For detailed algorithms see [49,50,53–55]. doi:10.1371/journal.pcbi.1000667.g003

representing the vertices as single reads impossible. Another problem is that the sheer volume of reads makes the graph large and unmanageable. The solution to a Hamiltonian path is an NP-complete problem, meaning that the time necessary for a solution grows exponentially with the number of nodes. So while it is possible to solve for a relatively low number of reads as are produced using Sanger sequencing, the problem becomes intractable with the large amounts of sequence data from second generation sequencers.

One solution is for the vertices to represent  $k$ -mer words with the reads themselves being the edges connecting the vertices. Since the vertices represent  $k$ -mers rather than reads, the high number of reads and their redundancy does not affect the number of nodes. Repeats exist in the graph only once, with links to the different start and end points. Searches for overlaps are simplified, as overlapping reads are mapped onto the same edge and can easily be followed simultaneously. Finally, since the reads are represented as edges rather than vertices, the solution is a Eulerian path, where each edge is visited once. Unlike a Hamiltonian path, a linear-time algorithm to solve a Eulerian path does exist, making the assembly problem tractable for large number of reads.

The EULER assembler [49,50] was the first to present this technique using de Bruijn graphs. De Bruijn graphs are  $n$ -dimensional graphs of  $m$  symbols. For metagenomic assembly,  $m = 4$  (A,T,G,C) and  $n \approx k$ -mer length. Theoretically, there are  $m^n$

vertices, but the dimensionality can be greatly reduced by hashing the reads in the dataset to be assembled (see Figure 3). Other variations have since been published, adapting to short (100–200) [51,52] and very short read lengths [53–55]. EULER and VELVET are available for download. Recently, Ye and Tang developed an assembly method that finds putative open reading frame (ORF) regions first, and then assembles those regions. This method, dubbed ORFome assembly, increases assembly accuracy for ORF regions at the expense of losing noncoding regions. Nevertheless, for many practical purposes this method is very useful, because it appears to have a better recovery rate, for coding regions only, than regular, whole genome assemblers [56]. For recent reviews on computational assembly methods see [57,58].

## Gene Calling

Genes are the basic functional unit in the genome, which may constitute larger functional units such as operons, transcriptional units, and functional networks. Again, the incomplete and fragmentary nature of metagenomic data presents challenges to identifying genes. With Sanger random shotgun sequencing, whole genomes are rarely assembled, and in species-rich environments, many reads remain as singletons rather than being joined in contigs. In the Global Ocean Sampling (GOS) data, which were Sanger-sequenced, the mean number of whole reading frames per assembly is 4.7 [59].

Gene finding algorithms are trained to find whole ORFs and take into account information gleaned from large genomic stretches. For metagenomic data, however, this information is unavailable. Despite such drawbacks, Mavromatis and colleagues have shown that for a high complexity metagenomic dataset, gene prediction on assemblies can be as accurate as 85% of the originally predicted genes in the constituting genomes. For a low complexity set this goes up to 90% [48].

For genes with known homologs, BLASTing (using the Basic Local Alignment Search Tool) [60,61] against known databases is a common approach. This approach informs of the existence of gene family members within a metagenome. BLAST cannot be used to find new families and new genes that have no homologs in known databases. For that, *ab initio* gene prediction tools are used. Those tools are mostly based on supervised learning and statistical pattern recognition methods. Most models use Markov models or Hidden Markov models. Genemark.hmm is a program that uses inhomogeneous Markov models based on monocodon frequency analysis for gene calling [62]. When applied to metagenomic data, however, those methods lose sensitivity, because they often fail to identify partial ORFs that may be part of true genes. This is especially true when conventional gene calling methods are applied to raw Sanger fragments rather than to assemblies. Unsupervised methods are therefore required.

Yooseph and colleagues [59,63] have used a different approach to gene finding when analyzing the global ocean survey data. They began with simple ORF identification of consecutive translatable regions that translate to at least 60 amino acids (aa). They then clustered those sequences using an all-against-all BLAST search, identifying clusters containing nonredundant sequences. In the next step, shadow ORFs were eliminated. Shadow ORFs are false ORFs in a different reading frame than the true ORF, but they overlap the true ORF and hence may be mistaken for a coding region. Yooseph and colleagues handled this by clustering all ORF candidates in the same reading frame and selecting the larger cluster as the one containing true ORFs, discarding the other ones as shadow ORFs. Finally, they removed ORF families with a Ka/Ks  $K_a/K_s$  ratio that is close to 1. The rationale for this step is

that putative proteins that are seemingly under no selective pressure (positive or negative) are probably falsely identified. Gene families coding for proteins under selective pressure are expected to have a  $K_a/K_s \gg 1$  or  $K_a/K_s \ll 1$ .

It has been argued that one drawback of the incremental clustering method is that it increases specificity at the expense of sensitivity; that is, it may have an excess of false negatives due to the removal of putative ORFs that do not cluster well or do not cluster at all in the database [64]. As of today, however, there has not been a thorough comparative evaluation of gene calling methods on first or second generation sequence data.

## Species Diversity

### Measuring Diversity

In the “Sample Size” we discussed using 16S/18S rDNA for phylotyping and assessing species coverage using a rarefaction curve. Microbial ecology has many tools for assessing species diversity. Rarefaction curves are used to estimate the coverage obtained from sampling, see Figure 4.  $\alpha$ -diversity,  $\beta$ -diversity, and  $\gamma$ -diversity are all well-established diversity indices used in ecology, including microbial ecology.  $\alpha$ -diversity is the biodiversity in a defined habitat or ecosystem;  $\beta$ -diversity compares species diversity between habitats;  $\gamma$ -diversity is the total biodiversity over a large region containing several ecosystems. Here we will discuss the application of these indices to metagenomic data.

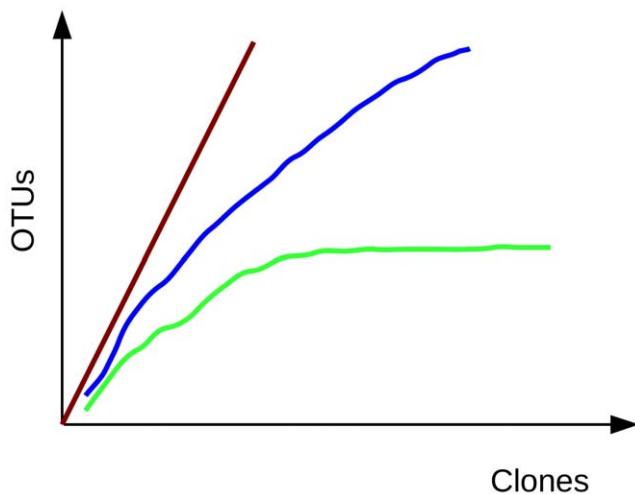
One way to calculate  $\alpha$ -diversity is by using Shannon’s index:

$$H'_\alpha = - \sum_{i=1}^S p_i \ln p_i$$

where:

$$p_i = \frac{n_i}{N}$$

Where  $S$  is the total number of OTUs,  $n_i$  is the number of clones in each OTU, and  $N$  is the total number of individuals.  $p_i$  is the relative abundance of each OTU.  $H_{max} = \ln S$ .



**Figure 4. Rarefaction curves.** Green, most or all species have been sampled; blue, this habitat has not been exhaustively sampled; red, species rich habitat, only a small fraction has been sampled. doi:10.1371/journal.pcbi.1000667.g004

### Using different sequence markers for OTU identification.

It should be noted that using 16S/18S rDNA as a proxy for OTU identification and counting is not without problems. First, rDNA has been criticized as an OTU marker, and evidence of horizontal gene transfer involving rDNA may confound its reliability even more [65]. Second, 16S rDNA may exist in multiple different sequence copies in a single bacterium: this would cause a variance in both the estimated individual bacterial count, and OTU numbers. It is commonly accepted that the mean number of bacterial ribosomal operons per genome is 4.1 [66], but in a recent publication it has been shown that 16S rDNA gene copy numbers may vary between 1 and 15 [67,68]. Alternative markers, such as single copy housekeeping genes have been suggested as alternative or complementary species and population tally markers for bacterial genomes. The *rpoB* gene is a strong candidate [69], but *amoA*, *pmoA*, *nirS*, *nirK*, *nosZ*, and *pufM* have also been suggested in different contexts [67,70]. The housekeeping functionality of these genes makes them less susceptible to horizontal gene transfer. However, these studies have shown that on a finer level the use of housekeeping genes does improve upon 16S rDNA alone, the use of 16S rDNA as a marker for OTU identification and count is still sufficiently accurate for many purposes. The use of housekeeping genes for OTU classification is primarily for those cases when 16S rDNA provides a lower resolution than when a high diversity of species is expected. Another case where a housekeeping gene is preferable to 16S rDNA is when the variation in the housekeeping gene matches the acceptable taxonomy better than the variation in the rDNA sequences. The use of non-rDNA phylogenetic markers has been applied to metagenomic data, showing that certain microbial communities evolve faster than others [71].

Epidemiologists classify bacterial serovars for pathogen verification using Multilocus Sequence Typing (MLST) [72,73]. MLST is a technique by which several standardized housekeeping genes are selected for OTU typification. There is an online resource for MLST, including a database for OTU identification (<http://www.mlst.net/>). MLST has been used successfully in some metagenomic studies [74]. However, MLST appears to be more useful for a finer level substrain typification, rather than OTUs.

In the same vein, 18S rDNA can have different count numbers in microeukaryotes, with an even larger copy number variation between species than 16S rDNA counts in prokaryotes. Care must be taken to account for this copy number variation when assessing the cell count in eukaryotic samples [75,76].

There are several software packages we found very useful for biodiversity analysis. The first is a general purpose population analysis software, EstimateS (8.0) [77]. EstimateS contains a rich set of biodiversity analysis modules, but for microbial analysis it requires preprocessing of sequence data to transform it into generic population data. MOTHUR [78] is tailored towards microbial diversity analysis and provides tools for transforming sequence data to population data. It is not as rich in functional modules as EstimateS, but for most diversity analyses (rarefaction curves, standard estimate indices) it is more than adequate. QIIME, an extension of PyCogent [79], is in beta, but testing by one of us (IF) has shown it to be a very powerful and versatile package for analysis of genomic and metagenomic microbial ecology data (<http://qiime.sourceforge.net>). A more specialized software geared to the analysis of viral metagenomic data is PHACCS [80].

### Binning

We wish to know not only who populates the sample, but also what the different OTUs are doing. We must therefore associate

sequence data with the OTU of its origin. This analysis is called binning (placing the sequence in its correct “bin” or OTU). In many cases, suitable phylogenetic marker genes are missing either because rDNA sequences may be unsuitable (as in virus analyses), or may have been undersampled.

Here we will examine two binning strategies: composition-based binning and phylogenetic binning.

**Composition-based binning.** The GC content of bacterial genomes is being used routinely for higher-level systematics [81]. With the advent of ESS data, a finer resolution for classifying or binning sequences is called for. Markov models based on  $k$ -mer frequencies have shown to be quite powerful for statistical analyses of DNA sequences [82]. For example, tetranucleotides are being used by the TETRA [83] program in the following fashion. There are  $4^4 = 256$  possible DNA tetranucleotides. For each tetranucleotide  $t_i = [n_1, n_2, n_3, n_4]$ , an expected frequency  $E(t_i)$  can be calculated by means of a maximal-order Markov model:

$$E(t_i) = \frac{O([n_1, n_2, n_3]) \times O([n_2, n_3, n_4])}{O([n_2, n_3])}$$

Where  $O$  is the observed count of the sub-trimers and dimer of the tetramer.

The level of over- and underrepresentation of each tetranucleotide is evaluated using  $z$ -scores:

$$\begin{aligned} \sigma(O(t_i)) \\ = E(t_i) \frac{(O([n_1, n_2]) - O([n_1, n_2, n_3])) \times (O([n_2, n_3]) - O([n_2, n_3, n_4]))}{O([n_2, n_3])^2} \\ Z(t_i) = \sqrt{\frac{(O(t_i) - E(t_i))^2}{\sigma(O(t_i))}} \end{aligned}$$

Where  $\sigma(O(t_i))$  is the variance in the tetranucleotide  $t_i$ .

Composition-based binning is not error-free. The closer the OTUs in the studied metagenome and the more numerous they are, the higher is the frequency of misclassification errors. The strength of  $k$ -mer-based binning is that there are no reference sequences required for the actual binning: all the information is intrinsic. This makes  $k$ -mer a powerful tool for binning ORFan sequences: sequences that have few or no homologs and therefore no known function. Therefore, TETRA is independent of existing genomic data, since it does not require any training. PhyloPythia [84] is a supervised method that trains a set of support vector machines (SVMs) to bin sequences of a length greater than 1 kb, and thus not suitable for binning second generation sequences. It performs best when a training set is similar in phylotypic composition to the training set. Growing Self Organizing Maps or GSOM [85] and Seeded GSOM or S-GSOM [86] use a variant of the machine learning algorithm self-organizing maps. S-GSOM improves upon GSOM by extracting the flanking sequences of highly conserved 16S rDNA from the metagenome and using them as seeds to assign other reads on the basis of their compositional similarity. Both use frequencies of di- to pentanucleotides for binning assignment.

Another composition-based method is codon-usage. An old technique in genomics, codon usage, can also be used for binning metagenomic data. Different species use different codon frequencies to encode the same amino acids, and this observation can be

exploited to classify ORF sequences. Shani Tzahor and colleagues have developed a composite supervised method that uses both TETRA and codon usage statistics to classify fragments in the 100–300-bp range [87].

TETRA is available for download, and PhyloPythia is available as a Web site, with a downloadable version available by request. GSOM/S-GSOM does not seem to be available at this time.

**Similarity-based binning.** Another way to bin sequences is to find similarities to reference sequences that can be used to build a tree. This technique is useful when most sequences in the sample have significant similarities to reference sequences from known OTUs. Given an unannotated sequence  $A$ , and two annotated reference sequences  $B$  and  $C$ , and using the similarity function  $sim$ , let us consider the case where we have  $sim(A, B) > sim(A, C)$ ; then, the sequence  $A$  will be placed on a node in the tree between  $B$  and  $C$ , and, in the case considered, closer to  $B$ . MEGAN [88] implements this method by reading a BLAST file output. Typically, the output is from the metagenomic reads or assemblies against nr, or any other sequence database that has a phylogenetic tree associated with it. MEGAN then assigns each read to the lowest common ancestor on the phylogenetic tree. This allows all sequences that have a homolog in nr to be assigned. Predicted gene sequences, having no homologs, are aggregated into their own single node on the tree. CARMA [89] is somewhat similar to MEGAN, but uses Pfam [90] as its source for taxonomic classification. It should be noted that a precise assignment to an OTU may not be possible in many cases. Nevertheless, unless it is an ORFan, the sequence can be placed in the species tree. The resulting picture of sequences on the species tree can provide an overview of the dominant species in the sample. Phymm [91] uses interpolated Markov models to characterize variable length DNA sequences by their phylogenetic grouping, unlike other methods. Phymm is trained on existing OTUs and learns which nucleotide length is best for classification. Also, Phymm does not leave reads unclassified, although that may impact its overall accuracy if there are many reads that cannot be accurately binned to any phylogenetic group.

As far as the usability of these software, CARMA will run on Unix-like environments, and its installation requires some third party software, and a rudimentary knowledge of Perl and MySQL. MEGAN runs in a Java virtual machine, and thus runs on almost out of the box Java-enabled platforms; it does require an installation of National Center for Biotechnology Information (NCBI)-formatted taxonomic reference database for lowest common ancestor mapping. Also, CARMA can run its own BLAST, whereas MEGAN requires a previously generated BLAST output as its input.

## Functional Annotation

Having assembled the metagenome and identified putative ORFs we would now like to understand the functional potential of the microbial community from where we derived the metagenome: what are these microbes capable of doing as a community? The first level of functional annotation is assigning biological functions to the ORFs. This task is highly challenging when applied to regular genomic data [92], and the challenge is compounded in metagenomic data where many ORFs are partial, and a large fraction have no annotated homologs. The second level would be discovering genes that constitute biological networks, such as metabolic pathways, in the data. The latter task is hampered by our inability to accurately associate each annotated ORF with a single species, which means it is sometimes hard to determine which component of a network comes from



which organism. Nevertheless, binning can help to some extent. As we shall see in the “Case Studies” section below, several studies have been carried out and led to the successful discovery of complementary metabolic pathways from microbes that constitute a community.

In metagenomic samples the probability of not calling all genes is higher than in a fully assembled genome, since many ORFs may be partial, and thus invisible to regular gene calling software that require a full ORF. Therefore, one strategy for functional annotation would be to skip the gene calling step altogether. Instead, simply use six-frame translations on the reads provided. If the translations are reasonably long they may be ORFs. Even if they are short, but they are cut short because of being at the edge of a contig, they may still be partial ORFs. Now these putative partial ORFs can be searched for motifs, HMM profiles, and other sequence signatures that may indicate functionality. The rationale is that the probability of calling a false ORF that also includes a known sequence signature is negligible. Some metagenomic annotation programs use this rationale. For example, Motif EXtraction (MEX) [93] is an unsupervised motif creation method that is successful in identifying enzymes in genomic and metagenomic data [94,95]. Short, enzyme-specific peptides are identified in an unsupervised learning stage. They are subsequently associated with certain functions, in the supervised learning stage. The reason an unsupervised stage takes place is because, in many cases, new motifs can be identified within ORFs, even though their functional association may be unknown.

Even unassembled single reads (singletons) may be used to infer functional information, being long enough to find short motifs or significant BLAST hits. BLASTing singletons and annotating the results without assembly or postassembly has its use. Two versatile and useful annotation pipelines for metagenomics that implement the annotation principles outlined above are MG-RAST [96] and RAMMCAP [97]. MG-RAST accepts a 454 dataset as input, normalizes it (removes artefactual duplicate sequences, a known problem with 454 sequencing), and then performs gene calling and annotation by a variety of sequence similarity searches (mainly BLAST) against various sequence databases, including 16S rDNA. It then produces statistics on species associations and on metabolic pathway associations using the SEED subsystems database as its guideline. RAMMCAP uses the fast clustering algorithm CD-HIT [98] to cluster translated ORFs by high sequence similarity. The rationale is that many similar putative ORFs strengthen the hypothesis that they are indeed real ORFs. Optionally, CD-HIT also serves to reduce the volume of data to be annotated by picking representatives from identical or nearly identical sequences and annotating only the representative sequences. The annotation is then transferred to the highly similar sequence in each similarity-based cluster. The sequences are then compared to the profile HMM databases Pfam [90] and TIGRFam [99] using HMMer (<http://hmmer.janelia.org/>) for functional annotation.

## Comparative Metagenomics

Comparing two or more metagenomes is necessary to understand how genomic differences affect, and are affected by, the abiotic environment. There are several sequence-based traits that can be compared: GC content was compared between marine and soil samples [59], microbial genome size [43], taxonomic [71], and functional content (e.g., [100]). Many comparative analyses, pairwise or multiple, make use of ordination statistics as when several metagenomic datasets are involved, or when several types of metadata are hypothesized to affect the observed compositions of the metagenomic populations. Principal component analysis

(PCA) and nonmetric multidimensional scaling (NM-MDS) are typically used to visualize the data and to reveal which factors affect the observed data most (e.g., [101,102]).

We mentioned MEGAN before as a binning software. MEGAN can also be used to compare the OTU composition of two or more frequency-normalized samples [103,104]. MG-RAST provides a comparative functional and sequence-based analysis for uploaded samples, whereas IMG/M provides similar analysis for metagenomes that exist in the IMG/M site [105]. RAMMCAP also provides the ability to compare metagenomes. Other software used for the comparison of microbial populations based on phylogenetic data are UniFrac [106] and MetaStats [107], the latter being suitable for preprocessed clinical metagenomic data. Galaxy, an online workbench for the analysis of genomic data, can also perform some comparative metagenomic analysis, as well as taxonomic mapping [108]. ShotgunFunctionalizeR [109] is a stand-alone analysis tool for metagenomics samples written in R [110]. The megx.net resource includes include MetaMine [111] for annotating genes using neighboring ORF information, and MetaLook [112] for organization of sequences using customized habitat criteria. CAMERA (<http://camera.calit2.net>) offers to BLAST the user's sequences against 40 existing genomic and metagenomic datasets. CAMERA also serves as an archive for select metagenomic datasets generated by marine microbial research funded by the Gordon and Betty Moore Foundation. All of these sites appear to be in a state of flux, with promised new functionalities to be added soon and with datasets constantly being updated.

We mentioned the importance of standardized recording of metadata in the “Recording Metadata” section above. Comparative analysis is where the importance of metadata comes into play: in order to properly compare between different environments, we need a common vocabulary describing the abiotic components. To date we do not know of software that provides a comparison between metadata or a comparative correlation between metadata and sequence data, although several such comparisons have been performed (see “Case Studies” section below).

## Applications

In this section we will discuss a few studies involving metagenomics. We chose these studies because each one illustrates a different insight that is derived from using metagenomics.

### Correlations between Environmental Data and Metadata

The study of the effects of the environment on microbes is as old as microbiology itself. Antoni van Leeuwenhoek noted that the “animalcules” scraped from his mouth and that he viewed under his microscope were gone or were immobile after he drank hot coffee. Leeuwenhoek was the first to describe a correlation between temperature change and organism viability [113]. Ever since then, microbe species distribution, genetics, pathogenicity, virulence, colonization—indeed every aspect of microbial life—has been correlated with habitat traits such as temperature, salinity, pH, nutrient content, etc. Traits of host-borne microbes have been correlated with the host species, age, habitat, behavior, feeding habits, host organs chosen for settlement/pathogenicity, and, of course, clinical symptoms and many other traits.

With the advent of metagenomics, we are now able to study the genomic potential of a bacterial community and how it is affected by and affects its habitat. Many metagenomic studies have looked to some extent at correlations between sequence data, environment, and environmental attributes in an attempt to gain

biological insight. One notable study by Turnbaugh and colleagues looked at the connection between the gut microbiome and obesity. The authors discovered that the metagenome in obese mice was enriched in carbohydrate active enzymes over that of lean mice. A separate biochemical experiment confirmed that the microbiome in obese mice has a larger energy harvesting capacity than in lean mice. They concluded that the gut microbiome contributes to obesity through this feed-forward cycle [100].

Studies such as those presented above looked at bivariate correlations: obesity and carbohydrate active enzyme enrichment. One recent study by Gianoulis and colleagues suggests how to locate multivariate correlations between metagenomic data and environmental attributes [114]. At the same time, environmental factors may combine in unexpected ways revealing new insights. Gianoulis and colleagues have identified covariation in amino acid transport and cofactor synthesis in nutrient-poor ocean areas, suggesting that limiting amounts of cofactor can (partially) explain increased import of amino acids in nutrient-limited conditions.

### Understanding Symbiosis

In many cases, symbiotic bacteria living in an animal host consist of a small number of species, which are often phylogenetically distant. Because they are few species and the phylogenetic distance makes their sequences relatively easy to bin, metagenomics is useful for studying symbionts. Eisen and his colleagues sequenced ESS data from bacterial symbionts living in the glassy-winged sharpshooter, which is an insect that lives solely on tree sap, a nutrient poor diet. By binning the ESS data they inferred that one symbiont synthesizes amino acids for the host insect, while another synthesizes cofactors and vitamins [115]. Not only that, but the symbiont providing the vitamins lacks some amino acid synthetic pathways, and the symbiont providing the amino acid synthetic lacks the ability to synthesize the vitamins. Thus, both symbionts complement each other's metabolic deficiencies, as well as feeding their host. Another study of the marine gutless worm *Olavius algarvensis* has revealed the different roles of its four symbionts in generating nutrients and processing the worm's waste [116]. None of the symbionts in the insect or in the worm study could be cultured under the reported conditions. Metagenomics thus became the chosen avenue for these studies.

### Enriching Gene Families

Another type of study enabled by metagenomics is the search for new members of a gene family. Metagenomics has opened up the floodgates of genomic material. Consequently the laborious hen-pecking for exemplars to enrich a studied gene family from known cultured species, has been replaced by the laborious computational filtering of appropriate exemplars from millions of environmental sequences. The previously small bacterial Eukaryotic Protein Kinase Like (ELK) family has been enriched several folds by the Global Ocean Sampling (GOS) project. Many new members of known families were identified, as well as new families. Within the protein sequences, four new residues of unknown function were found to be conserved, setting the stage for future functional studies of this family [117].

### Metagenomics and Environmental Virology

Outnumbering living microbes, viruses are the most abundant biological entity on Earth: there are an estimated  $10^{30}$  tailed bacteriophages in the biosphere [118]. In marine environments, viruses constitute 94% of all nucleic-acid containing particles, although owing to their small size they are estimated to constitute only 5% of the biomass. Metagenomic studies have enriched our knowledge of viral diversity and the role viruses play as facilitators

of microbial genetic diversity. Sequence similarity analyses of viral metagenomic data have shown that approximately 90% of the sequences have no similarity to GenBank sequences, telling of an underrepresentation of viral sequence data in sequence databases [119].

Transduction—the transfer of genetic material via a viral vector—is known to be a strong contributor to genetic diversity in prokaryotes. Metagenomic studies help us assess the magnitude of virally contributed genetic diversity. For example, the existence of photosynthetic genes in cyanophages—viruses infecting cyanobacteria—has been known for some time [120,121]. However, metagenomic studies have revealed the extent of this phenomenon: it is estimated that 60% of the *psbA* genes, a component of Photosystem I, in surface water are of phage origin. Another metagenomic study revealed the existence of whole photosynthetic cassettes in cyanophages, which may increase host fitness by supplementing and enhancing existing cyanobacterial photosystems. The latter findings were enabled by the metagenomic data from Global Ocean Sampling (GOS). Surveying these data using simple sequence similarity analyses and chromosomal gene location have revealed the existence of Photosystem I genes in cyanophages, and the extent of their distribution [94,122].

Clinical virology also stands to benefit from metagenomic analysis [123]. Indeed, recent molecular-based discoveries of highly prevalent viral infections caused by anellovirus [124] and GBV-C [125] highlight the need for a better understanding of the human viral flora.

The computational analysis of viral metagenomic data is particularly challenging. First, viruses may exist as a chromosomal insert, such as prophages, which are incorporated in the host genome. This incorporation confuses the ability to distinguish viral genomic elements from the host. Furthermore, when filtering exclusively for viral particles, prophage elements are lost. Second, viruses have no distinct phylogenetic marker gene, equivalent to the small ribosomal subunit rRNA in prokaryotes or eukaryotes. The lack of a consensual marker gene hampers phylogenetic and diversity analysis. Third, as stated above, most viral genes have no annotated homolog in sequence databases, which impedes functional analysis and indeed the identification of viral genes for what they are. Indeed, by some estimates the majority of ORFans in the biosphere is due to lateral gene transfer of viral origin [126] and the fact that phage-induced lateral gene transfer contributes in a major way to microbial diversity [127].

### The Future

We are in the midst of the fastest growing revolution in molecular biology, perhaps in all of life science, and it only seems to be accelerating. Sanger sequencing has been with us for over three decades. High-throughput 3730 sequencing has been around for 8 years, Roche 454 instrumentation has been available for 6 years, and Illumina GA for 3 years. The latter two methods have enabled us to generate more sequence data than Sanger sequencing has. We are still coming to grips with the large volume of data, and how to analyze it. Assembly, quality control, binning, and annotation all require ingenious algorithms combined with the latest computational power. It appears that sequencing technology is changing almost faster than the associated computational techniques can keep up. There are many indications that within a few years, short-read second generation sequencing may be outdated. Third generation sequencing that will enable the sequencing of a single chromosome in a single pass with few or no fragments should be established very soon [35,36]. Does this plausible obsolescence of second

generation sequencing change current metagenomic computational challenges? For some applications, assembly algorithms may be less warranted, but for species-rich samples, we may not be able to rely solely on third-generation sequencing for good sampling. Coverage assessment, gene finding, binning, and annotation will still be necessary.

The BASE technology from Oxford Nanopore is able to differentiate between cytosine and methyl-cytosine during sequencing [37]. Methylation acts as a primitive immune system in bacteria [128], and as an expression control mechanism in eukarya [129]. This additional epigenetic information has been mostly unavailable in sequencing projects due to an inability to obtain it in a high-throughput fashion. Pyrosequencing already offers a capability for quantitative methylation [130] and in all likelihood methylation data will be soon made available routinely along with the four base data, and the associated bioinformatics would need to address that.

Another growing problem is that of data management. Sequencing centers are working to equip themselves with computational infrastructure to meet the flow of sequence data. However, many research institutes who request the sequencing do not have the computational infrastructure needed to deal with analysis and long-term storage of these data. The sheer volume of data raises new constraints on its transfer and analysis. These challenges would have to be met by concerted efforts of life scientists, computer scientists, engineers, and funding agencies [131,132].

Genomic data tell us what an organism is capable of doing, i.e., its genomic potential. What it is actually doing at a given time-frame is discovered by examining transcription (mRNA) and translation (protein) data. In the world of microbial communities, those studies have been dubbed metatranscriptomics and metaproteomics, respectively. These two fields are outside the scope of this review, but note that they too are very much in a development boom, technologically and computationally [133–135].

We hope this primer has been useful and informative. Because computational metagenomics is changing rapidly, we call upon the readers of this article who are knowledgeable in the subject to use the comment section of *PLoS Computational Biology* to provide updated information.

## Acknowledgments

We would like to thank Linda Amaral-Zettler, Yanay Ofran, Daniel Huson, Jack Gilbert, Mufit Ozden, Frank Oliver Glöckner, and Renzo Kottmann for their critique and input. We would also like to thank the referees for contributing their time and effort towards making this a better article. IF dedicates his contribution to the memory of Ilan Friedberg.

## References

- Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* 95: 6578–6583.
- Savage DC (1977) Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol* 31: 107–133.
- Berg R (1996) The indigenous gastrointestinal microflora. *Trends Microbiol* 4: 430–435.
- Collins FS, McKusick VA (2001) Implications of the human genome project for medical science. *JAMA* 285: 540–544.
- Kaput J, Cotton RGHG, Hardman L, Watson M, Al Aqeel AH, et al. (2009) Planning the human variome project: the spain report. *Hum Mut* 30: 496–510.
- O'Hara AM, Shanahan F (2006) The gut flora as a forgotten organ. *EMBO Rep* 7: 688–693.
- Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, et al. (1976) Complete nucleotide sequence of bacteriophage ms2 RNA: primary and secondary structure of the replicase gene. *Nature* 260: 500–507.
- Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, et al. (1978) The nucleotide sequence of bacteriophage phix174. *J Mol Biol* 125: 225–246.
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of *haemophilus influenzae* rd. *Science* 269: 496–512.
- Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59: 143–169.
- Pace NR (1997) A molecular view of microbial diversity and the biosphere. *Science* 276: 734–740.
- Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. *Annu Rev Microbiol* 57: 369–394.
- Handelsman J, Rondon MR, Brady SF, 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.
- 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.

## Box 1. Glossary of terms.

**Binning** Clustering sequences based on their nucleotide composition or similarity to a reference database

**Contig** A set of overlapping DNA segments

**Coverage (in sequencing)** The mean number of times a nucleotide is sequenced in a genome

**ESS** Environmental Shotgun Sequencing

$K_a/K_s$  The ratio of the rate of nonsynonymous substitutions ( $K_a$ ) to the rate of synonymous substitutions ( $K_s$ ), which can be used as an indicator of selective pressure acting on a protein-coding gene

**Mate pairs** Sequences known to be in the 3' and 5' of a contig from a single clone

**Metadata** Definitional data that provide information about or documentation of other data

**Metagenome** The DNA obtained from uncultured microorganisms

**Metagenomics** The study of genomic DNA obtained from uncultured microorganisms

**Metaproteomics** The study of protein molecular data obtained from environmental samples using proteomics techniques

**Metatranscriptomics** The study of transcription sequence data obtained from environmental samples

**ORFan** An ORF that has no (or few, depending on definition) homologs in other organisms

**OTU** Operational taxonomic unit, species distinction in microbiology. Typically using rDNA and a percent similarity threshold for classifying microbes within the same, or different, OTUs

**Ontology** A formal representation of a set of concepts and the relationships between them. Ontologies are used to create a consensual unambiguous controlled vocabulary

**Polony** Discrete clonal amplifications of a single DNA molecule, grown in a gel matrix. The clusters can then be individually sequenced, producing short reads. Polony-based sequencing is the basis of most second generation sequencers

**Rarefaction curve** A curve describing the growth of a number of species discovered as a function of individuals sampled

**Ribotype** A phylotypic classification based on rDNA sequences

**Scaffold** A series of contigs that are in the right order but not necessarily connected in one contiguous stretch

**Shadow ORF** An incorrectly identified ORF that overlaps the coding region of the true ORF

15. Field D, Garrity G, Gray T, Morrison N, Selengut J, et al. (2008) The minimum information about a genome sequence (migs) specification. *Nat Biotech* 26: 541–547.
16. Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, et al. (2008) A standard MIGS/MIMS compliant xml schema: toward the development of the genomic contextual data markup language (gcdml). *OMICS* 12: 115–121.
17. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* 29: 365–371.
18. Westbrook JD, Fitzgerald PM (2003) The PDB format, mmCIF, and other data formats. *Methods Biochem Anal* 44: 161–179.
19. Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94: 441–448.
20. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74: 5463–5467.
21. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, et al. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318: 1449–1452.
22. Pedros-Alio C (2007) Ecology: Dipping into the rare biosphere. *Science* 315: 192–193.
23. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc Natl Acad Sci U S A* 103: 12115–12120.
24. Hamp TJ, Jones WJ, Fodor AA (2009) Effects of experimental choices and analysis noise on surveys of the “rare biosphere”. *Appl Environ Microbiol* 75: 3263–3270.
25. Neufeld JD, Li J, Mohn WW (2008) Scratching the surface of the rare biosphere with ribosomal sequence tag primers. *FEMS Microbiol Lett* 283: 146–153.
26. Mitra RD, Church GM (1999) In situ localized amplification and contact replication of many individual DNA molecules. *Nucl Acids Res* 27: e34.
27. Porreca GJ, Shendure J, Church GM (2006) Polony DNA sequencing. *Current protocols in molecular biology*. Frederick M. Ausubel, et al., editors. Chapter 7 Hoboken (New Jersey): John Wiley and Sons, Inc.
28. Nyrén P, Pettersson B, Uhlén M (1993) Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem* 208: 171–175.
29. Ronaghi M, Uhlén M, Nyrén P (1998) A sequencing method based on real-time pyrophosphate. *Science* 281: 363–365.
30. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 7057: 376–380.
31. Holt RA, Jones SJM (2008) The new paradigm of flow cell sequencing. *Genome Res* 18: 839–846.
32. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26: 1135–1145.
33. Harismendy O, Ng P, Strausberg R, Wang X, Stockwell T, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 10: R32.
34. McPherson JD (2009) Next-generation gap. *Nat Methods* 6: S2–S5.
35. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, et al. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nano* 4: 265–270.
36. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. (2008) Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133–138.
37. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, et al. (2008) The potential and challenges of nanopore sequencing. *Nat Biotech* 26: 1146–1153.
38. Lauder ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2: 231–239.
39. Torsvik V, Goksoyr J, Daae FL (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol* 56: 782–787.
40. Youssef NH, Elshahed MS (2008) Species richness in soil bacterial communities: a proposed approach to overcome sample size bias. *J of Microb Meth* 75: 86–91.
41. Fierer N, Jackson RB (2006) The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci U S A* 103: 626–631.
42. Countway PD, Gast RJ, Pratik Sava I, Caron DA (2005) Protistan diversity estimates based on 18s rDNA from seawater incubations in the western north atlantic. *J Euk Microbiol* 52: 95–106.
43. Raes J, Korb J, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* 8: R10.
44. Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008) Metasima sequencing simulator for genomics and metagenomics. *PLoS ONE* 3: e3373. doi:10.1371/journal.pone.0003373.
45. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, et al. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Res* 12: 177–189.
46. Aparicio S, Chapman J, Stupka E, Putnam N, Ming Chia J, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of *fugu rubripes*. *Science* 297: 1301–1310.
47. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. (2000) A whole-genome assembly of *Drosophila*. *Science* 287: 2196–2204.
48. Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* 4: 495–500.
49. Pevzner PA, Tang H, Waterman MS (2001) An eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* 98: 9748–9753.
50. Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. *Genome Res* 18: 324–330.
51. Myers EW (2005) The fragment assembly string graph. *Bioinformatics* 21 Suppl 2: ii79–ii85.
52. Chaisson M, Pevzner P, Tang H (2004) Fragment assembly with short reads. *Bioinformatics* 20: 2067–2074.
53. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res* 18: 821–829.
54. Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE* 2: e484. doi:10.1371/journal.pone.0000484.
55. Warren RL, Sutton GG, Jones SJ, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 23: 500–501.
56. Ye Y, Tang H (2009) An orfome assembly approach to metagenomics sequences analysis. *J Bioinform Comput Biol* 7: 455–471.
57. Pop M (2009) Genome assembly reborn: recent computational challenges. *Brief Bioinform* 4: 354–366.
58. Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 6: S6–S12.
59. Yooshef S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The Sorcerer II global ocean sampling expedition: expanding the universe of prokaryotic families. *PLoS Biol* 5: e16. doi:10.1371/journal.pbio.0050016.
60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
61. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
62. Azad RK, Borodovsky M (2004) Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. *Brief Bioinform* 5: 118–130.
63. Yooshef S, Li W, Sutton G (2008) Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. *BMC Bioinformatics* 9: 182.
64. Hoff KJ, Tech M, Lingner T, Daniel R, Morgenstern B, et al. (2008) Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinformatics* 9: 217.
65. Schouls LM, Schot CS, Jacobs JA (2003) Horizontal transfer of segments of the 16s rRNA genes between species of the streptococcus anginosus group. *J Bacteriol* 185: 7241–7246.
66. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with arb. *Appl Environ Microbiol* 72: 5069–5072.
67. Case RJ, Boucher Y, Dahllöf I, Holmstrom C, Doolittle FW, et al. (2007) Use of 16s rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol* 73: 278–288.
68. Klappenbach JA, Saxman PR, Cole JR, Schmidt TM (2001) rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucl Acids Res* 29: 181–184.
69. Walsh DA, Bapteste E, Kamekura M, Doolittle FW (2004) Evolution of the RNA polymerase  $\beta'$  subunit gene (rpoB) in halobacteriales: a complementary molecular marker to the ssu rRNA gene. *Mol Biol Evol* 21: 2340–2351.
70. Achenbach LA, Carey J, Madigan MT (2001) Photosynthetic and phylogenetic primers for detection of anoxygenic phototrophs in natural environments. *Appl Environ Microbiol* 67: 2922–2926.
71. von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, et al. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315: 1126–1130.
72. Enright MC, Spratt BG (1999) Multilocus sequence typing. *Trends Microbiol* 7: 482–487.
73. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, et al. (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* 95: 3140–3145.
74. Mahenthalingam E, Baldwin A, Drevinck P, Vanlaere E, Vandamme P, et al. (2006) Multilocus sequence typing breathes life into a microbial metagenome. *PLoS ONE* 1: e17. doi:10.1371/journal.pone.0000017.
75. Zhu F, Massana R, Not F, Marie D, Vaulot D (2005) Mapping of picococaryotes in marine ecosystems with quantitative PCR of the 18s rRNA gene. *FEMS Microbiol Ecol* 52: 79–92.
76. Loram JE, Boonham N, O’Toole P, Trapido-Rosenthal HG, Douglas AE (2007) Molecular quantification of symbiotic dinoflagellate algae of the genus *Symbiodinium*. *Biol Bull* 212: 259–268.
77. Colwell RK (2005) EstimateS - statistical estimation of species richness and shared species from samples.
78. Schloss PK (2008) Mothur - the one-stop source for your computational microbial ecology needs.
79. Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, et al. (2007) PyCogent: a toolkit for making sense from sequence. *Genome Biol* 8: R171.
80. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. (2005) Phacss, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* 6: 41.

81. (1988) International committee on systematic bacteriology announcement of the report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *J Appl Bacteriol* 64: 283–284.
82. Schbath S, Prum B, de Turckheim E (1995) Exceptional motifs in different markov chain models for a statistical analysis of DNA sequences. *J Comput Biol* 2: 417–437.
83. Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO (2004) Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 5: 163.
84. Mchardy AC, Martín HG, Tsigros A, Hugenholtz P, Rigoutsos I (2006) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 4: 63–72.
85. Chan CKKK, Hsu AL, Tang SLL, Halgamuge SK (2008) Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. *J Biomed Biotech*: 513701.
86. Chan CKK, Hsu AL, Halgamuge SK, Tang SL (2008) Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics* 9: 215.
87. Tzahor S, Aharonovich DM, Kirkup B, Yogev T, Frank IB, et al. (2009) A supervised learning approach for taxonomic classification of core-photosystem-ii genes and transcripts in the marine environment. *BMC Genomics* 10: 229.
88. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17: 377–386.
89. Krause L, Diaz NN, Goessmann A, Kelley S, Nattkemper TW, et al. (2008) Phylogenetic classification of short environmental DNA fragments. *Nucl Acids Res* 36: 2230–2239.
90. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJJ, et al. (2008) The pfam protein families database. *Nucleic Acids Res* 36: D281–D288.
91. Brady A, Salzberg SL (2009) Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nat Methods* 6: 673–676.
92. Friedberg I (2006) Automated protein function prediction—the genomic challenge. *Brief Bioinform* 7: 225–242.
93. Kunik V, Meroz Y, Solan Z, Sandbank B, Weingart U, et al. (2007) Functional representation of enzymes by specific peptides. *PLoS Comput Biol* 3: e167. doi:10.1371/journal.pcbi.0030167.
94. Sharon I, Tzahor S, Williamson S, Shmoish M, Man-Aharonovich D, et al. (2007) Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J* 1: 492–501.
95. Meroz Y, Horn D (2008) Biological roles of specific peptides in enzymes. *Proteins* 72: 606–612.
96. Meyer F, Paarmann D, D'Souza M, Olson RD, 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.
97. Li W (2009) Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* 10: 359.
98. Li W, Godzik A (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659.
99. Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. *Nucl Acids Res* 31: 371–373.
100. 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: 1027–1031.
101. Brule JM, Antonopoulos DA, Berg Miller ME, Wilson MK, Yannarell AC, et al. (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A* 106: 1948–1953.
102. 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. doi:10.1371/journal.pone.0007370.
103. Mitra S, Klar B, Huson DH (2009) Visual and statistical comparison of metagenomes. *Bioinformatics*: btp341.
104. Huson D, Richter D, Mitra S, Auch A, Schuster S (2009) Methods for comparative metagenomics. *BMC Bioinformatics* 10: S1–S12.
105. 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: D534–D538.
106. Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71: 8228–8235.
107. White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 5: e1000352. doi:10.1371/journal.pcbi.1000352.
108. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 15: 1451–1455.
109. Kristiansson E, Hugenholtz P, Dalevi D (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* 25: 2737–2738.
110. R Development Core Team (2009) R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, Available: <http://www.R-project.org>. ISBN 3-900051-07-0.
111. Bohnbeck U, Lombardot T, Kottmann R, Glockner F (2008) Metamine - a tool to detect and analyse gene patterns in their environmental context. *BMC Bioinformatics* 9: 459.
112. Lombardot T, Kottmann R, Giuliani G, de Bono A, Addor N, et al. (2007) Metalook: a 3d visualisation software for marine ecological genomics. *BMC Bioinformatics* 8: 406.
113. Egerton FN (2006) A history of the ecological sciences, part 19: Leeuwenhoek's microscopic natural history. *Ecol Appl* 87: 47–58.
114. Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, et al. (2009) Quantifying environmental adaptation of metabolic pathways in metagenomes. *Proc Natl Acad Sci U S A* 106: 1374–1379.
115. Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, et al. (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biol* 4: e188. doi:10.1371/journal.pbio.0040188.
116. 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.
117. Kannan N, Taylor SSS, Zhai Y, Venter JCC, Manning G (2007) Structural and functional diversity of the microbial kinome. *PLoS Biol* 5: e17. doi:10.1371/journal.pbio.0050017.
118. Brussow H, Hendrix RW (2002) Phage genomics: small is beautiful. *Cell* 108: 13–16.
119. Edwards RA, Rohwer F (2005) Viral metagenomics. *Nat Rev Microbiol* 3: 504–510.
120. Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003) Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* 424: 741.
121. Millard A, Clokie MRJ, Shub DA, Mann NH (2004) Genetic organization of the psbd region in phages infecting marine *synechococcus* strains. *Proc Natl Acad Sci U S A* 101: 11007–11012.
122. Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, et al. (2009) Photosystem i gene cassettes are present in marine virus genomes. *Nature* 461: 258–262.
123. Delwart EL (2007) Viral metagenomics. *Rev Med Virol* 17: 115–131.
124. Nishizawa T, Okamoto H, Konishi K, Yoshizawa H, Miyakawa Y, et al. (1997) A novel dna virus (ttv) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. *Biochem Biophys Res Commun* 241: 92–97.
125. Simons JN, Leary TP, Dawson GJ, Pilot-Matias TJ, Muerhoff AS, et al. (1995) Isolation of novel virus-like sequences associated with human hepatitis. *Nat Med* 1: 564–569.
126. Yin Y, Fischer D (2006) On the origin of microbial orfans: quantifying the strength of the evidence for viral lateral transfer. *BMC Evol Biol* 6: 63.
127. Hamblly E, Suttle C (2005) The virosphere, diversity, and genetic exchange within phage communities. *Curr Opin Microbiol* 8: 444–450.
128. Boyer HW (1971) Dna restriction and modification mechanisms in bacteria. *Annu Rev Microbiol* 25: 153–176.
129. Kass S (1997) How does DNA methylation repress transcription? *Trends Genet* 13: 444–449.
130. Tost J, Gut IG (2007) Dna methylation analysis by pyrosequencing. *Nat Protoc* 2: 2265–2275.
131. Batley J, Edwards D (2009) Genome sequence data: management, storage, and visualization. *BioTechniques* 46: 333–336.
132. Richter BG, Sexton DP (2009) Managing and analyzing next-generation sequence data. *PLoS Comput Biol* 5: e1000369. doi:10.1371/journal.pcbi.1000369.
133. Bailly J, Fraissinet-Tachet L, Verner MC, Debaud JC, Lemaire M, et al. (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. *ISME J* 1: 632–642.
134. Wilmes P, Bond PL (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ Microbiol* 6: 911–920.
135. Wilmes P, Bond PL (2006) Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol* 14: 92–97.