Published in final edited form as:

J Comput Sci Technol. 2009 January ; 25(1): 71–81. doi:10.1007/s11390-010-9306-4.

Metagenomics: Facts and Artifacts, and Computational Challenges*

John C. Wooley¹ and Yuzhen Ye²

John C. Wooley: jwooley@ucsd.edu; Yuzhen Ye: yye@indiana.edu

¹ Center for Research on BioSystems, Calit2, UC San Diego, La Jolla CA 92093

² School of Informatics and Computing, Indiana University, Bloomington, Indiana, 47408

Abstract

Metagenomics is the study of microbial communities sampled directly from their natural environment, without prior culturing. By enabling an analysis of populations including many (so-far) unculturable and often unknown microbes, metagenomics is revolutionizing the field of microbiology, and has excited researchers in many disciplines that could benefit from the study of environmental microbes, including those in ecology, environmental sciences, and biomedicine. Specific computational and statistical tools have been developed for metagenomic data analysis and comparison. New studies, however, have revealed various kinds of artifacts present in metagenomics data caused by limitations in the experimental protocols and/or inadequate data analysis procedures, which often lead to incorrect conclusions about a microbial community. Here, we review some of the artifacts, such as overestimation of species diversity and incorrect estimation of gene family frequencies, and discuss emerging computational approaches to address them. We also review potential challenges that metagenomics may encounter with the extensive application of next-generation sequencing (NGS) techniques.

Keywords

Metagenomics; next-generation sequencing (NGS); taxonomic/functional profiling; statistical approaches; comparative metagenomics

1 Introduction

Metagenomics, a term first tossed in 1998 [1], is a methodology that applies genome sequencing or assays of functional properties to the culture-independent analysis of complex and diverse (“meta”) populations of microbes. In sequencing studies, unlike traditional microbial genomic sequencing projects, metagenomics research attempts to determine directly the whole collection of genes within an environmental sample (i.e., the metagenome), and analyze their biochemical activities and complex interactions. Current metagenomics projects are facilitated by the rapid development of so-called Next-Generation Sequencing (NGS) techniques [2], which provide lowered cost experimental tools without the cloning process inherent in conventional capillary-based methods. Landmark progress in metagenomics occurred in 2004 [3,4] when two research groups published results from large-scale environmental sequencing projects. Metagenomics projects have very broad applications, from ecology and environmental sciences [5], to the chemical industry [6], and human health (e.g., the human gut microbiome metagenomics) [7,8]. Other than the large-scale shotgun metagenomics, there are small-scale approaches, such as the 16S rRNA-based surveys, and targeted metagenomics to study the microbes in the environment. 16S rRNA-based surveys of microbial communities have been conducted

well before the large-scale shotgun metagenomics started, and is now boosted by the development of the barcoded pyrosequencing methodology [9]. Targeted metagenomics, as the name indicates, aims at acquiring sequence reads with specific protein functions, such as glycoside hydrolases [10,11], bile salt hydrolase [12], and bleomycin resistance genes [13] (for a review see [14]).

Metagenomics, in principle, enables the study of any microbial organism, including the large number of microorganisms (more than 99%) that cannot be isolated or are difficult to grow in lab [15]. More importantly, microbes by nature live in communities, where they interact with each other by exchanging nutrients, metabolites and signaling molecules. Although the conventional pure-culture paradigm remains important for complete characterization of a species, its traditional exclusive usage limits exploring the microbial world. Traditional clonal culture microbiology needs to be complemented by culture-independent microbiology that can directly characterize microbes in natural environments, and can address important biological questions related to those microbial environments, such as the diversity of microbes in different environments [16], microbial (and microbe-host) interactions [17], and the environmental and evolutionary processes [18].

Taking advantage of several excellent, recent reviews on metagenomics [19–21], we will focus, in this review, on the introduction of new computational and statistical tools for metagenomic data analysis, and the artifacts that could lead to wrong conclusions if not handled appropriately.

2 Facts

Computational analysis has shown an even greater impact on metagenomic studies compared to traditional genomic projects, due not only to the large amount of metagenomic data, but also to the new complexity introduced by metagenomic projects (e.g., assembly of multiple genomes simultaneously is more challenging than the assembly of single genomes), and the new questions we are asking (e.g., host-microbe interaction).

2.1 Computational and statistical tools for metagenomic studies

2.1.1 Assembly and gene prediction—Once the sequences are collected, the first step in data analysis is to reconstruct the entire microbial genomes from metagenomic sequence reads using fragment assemblers. Unfortunately, due to the high species complexity and the short length of sequencing reads from NGS sequencers, the reconstruction goal is too difficult if not impossible to attain for samples from many microbial environments. As a result, metagenomic sequences are often subject to further analysis as a collection of short reads.

The early attempts for assembling metagenomic sequences used conventional whole genome assembly (WGA) pipelines, including whole genome assemblers and gene finding programs originally designed for conventional whole genome shotgun sequence (WGS) projects with only some small parameter modifications [22]. The development of genome assembly algorithms has been boosted recently by the development of NGS techniques. New genome assemblers for short reads, including Velvet (a Eulerian path assembler) [23], ALLPATHS [24], Euler-SR [25], have been developed that are specifically targeted at short and ultra-short reads (for examples, the 454 pyrosequencer, and the Illumina/Solexa and SOLiD sequencers). We note that none of these assemblers are designed for assembling mixture genomes (i.e., metagenomes), and refer the reader to a review [26] on the recent development of genome assemblers and the difficult challenges assembly approaches encounter in the metagenomics field.

For finding genes in metagenomic sequences, one can certainly use the full genome scale “gene predictors” (i.e., those developed for gene prediction from whole genomes). If not, one has to deal with short reads (or small contigs). Most metagenomic studies use 6-frame translation when conducting a similarity search on the short reads. There has been very little development of gene prediction methods specifically for metagenomic sequences or for short contigs assembled from the reads. MetaGene [27] and Orphelia [28,29] were designed as gene predictors for short reads (~700bp). MetaGene utilizes dicodon frequencies estimated from the GC content of a given sequencing read and other measures, such as the length distribution of open reading frames (ORFs). Orphelia uses a two-stage machine learning approach for protein-coding gene prediction from the metagenomic reads: first it uses linear discriminants for monocodon usage, dicodon usage and translation initiation sites to extract features from DNA sequences, and then it combines these features with ORF length and fragment GC-content in an artificial neural network to compute the probability that an ORF encodes a protein [28,29]. Krause and colleagues [30] developed a gene finder for small contigs that may have in-frame stop codons as well as frame shifts. The gene finder works by initially conducting a similarity search using BLAST for all the contigs and sequences, and by next calculating the coding potential for each position in a contig based on BLAST-based search results (related by the number of synonymous and non-synonymous substitutions), and finally, by predicting the coding sequences via looking for a chain of nucleotide positions that maximizes the sum of coding potential using dynamic programming. The ORFome assembly approach [31] took a different path, which starts with open reading frames (ORFs) prediction followed by assembly of the predicted peptides. This approach can produce long peptides even when the sequence coverage of reads is extremely low (and the reads may have synonymous mutations). A potential application of the ORFome assembly is to increase the sensitivity of homology searching by using longer peptides, thus to improve the functional annotation of metagenomic sequences.

2.1.2 Tools for characterizing microbial diversity qualitatively and quantitatively—

The most extensive development of computational tools for metagenomics might well have occurred in conjunction with taxonomic studies. One of the primary goals of metagenomic projects is to characterize the organisms present in an environmental sample and, in turn, the taxonomic composition of environmental communities is an important indicator of their ecology and function. In contrast, the molecular approaches, such as genome assembly and function prediction, are relatively standard or well known, although their usage and extension provides new challenges in the metagenomics field. Many computational tools have been developed to infer species information from raw short reads directly, i.e., without the need for assembly. Here we only give a very brief review of the development in this aspect, and we refer the readers to a review [32] for more details.

Similarity-based and phylogeny-based phylotyping tools utilize similarity searches of the metagenomic sequences against a database of known genes/proteins. MEGAN applies a simple lowest common ancestor algorithm to assign reads to taxa, based on BLAST similarity search results [33]. Phylogenetic analysis of marker genes, including 16S rRNA genes [34], DNA polymerase genes [35], and 31 selected marker genes [36] have also been applied to determining taxonomic distribution. MLTreeMap [37] and AMPHORA [38] are two phylogeny-based phylotyping tools that use the phylogenetic analysis of marker genes for taxonomic distribution estimation. CARMA [39] searches for conserved Pfam domain and protein families [40] in the raw metagenomic sequences and classifies them into a higher-order taxonomy, based on the reconstruction of a phylogenetic tree of each matching Pfam family.

Other tools have also been developed to address a related problem; namely, the binning problem, which is to cluster metagenomic sequences into different bins (i.e., operational taxonomic units or species). Most existing computational binning tools simply utilize DNA composition. The basis of these approaches is that genome G+C content, dinucleotide frequencies, and synonymous codon usage vary among organisms, and are generally characteristic of evolutionary lineages [41]. TETRA [42] uses z-scores from tetramer frequencies to classify metagenomic sequences. MetaClust [43] uses a combination of k-mer frequency metrics to score metagenomic sequences, and was used to classify sequences from the endosymbionts of a gutless worm. CompostBin [44], a semi-supervised approach, uses a weighted PCA algorithm to project high dimensional DNA composition data into an informative lower-dimensional space, and then uses the normalized cut clustering algorithm to classify sequences into taxon-specific bins. Zhou et al [45] developed a tool that generates genomic barcodes from DNA composition, and then, uses the barcodes to classify sequences. However, the DNA composition-based approaches only work with long reads (~800–1Kbp); they fail with short reads because of the local variation of DNA composition across a genome. (The GC content across genomes has been observed to be far from uniform [41]).

More tools are emerging that target ultra-short reads of 100 bps or shorter for taxonomic studies [46]. However, computational tools with higher accuracy, especially for short reads, for phylotyping (both quantitative and qualitative) and binning problems still remain much needed.

2.1.3 Function prediction—There has been relatively less algorithmic and software development for function predictions for metagenomics, though such tools are extremely important and will be needed for understanding the function of a community. (This is especially so given the presence of so many protein families of unknown function in the microbial communities studied [47]). Function prediction on metagenomic data is generally more difficult than that for genomic studies, because short reads are more difficult to annotate, and some information (such as the gene neighborhood in a genome) that is very important for function prediction in sequence data from individual genomes may not be so useful in working with metagenomic data. Existing functional annotation pipelines (developed for annotating genomic sequences derived from traditional genome projects) have been extensively used for functional annotation of metagenomic sequences. An often used practice is to predict COG families, KEGG families [48], FIG families [49], and other functional categories for metagenomic short reads based on BLAST search results. Currently, BLAST is the only readily available or established similarity tool that can handle comprehensive analyses of a metagenomic dataset (with the support of a big computer cluster with hundreds or thousands of CPUs). Very often, an E-value cutoff (e.g., $1e-5$) has been adopted for annotating if a read has a certain function [7,50]. Biological pathways (or subsystems) can also be reconstructed for metagenome building from protein family predictions. (In most studies, the information obtained is on pooled biological pathways from various species in the same microbial community). MG-RAST is an automatic server for subsystem annotation for metagenomic datasets, based on an extension of the very successful microbial genome annotation server RAST [51]. Pooled biological pathways provide key insights into the functional/metabolic ability of a microbial community (e.g., the functional profiling by KEGG pathways in [16]), although in most metagenomic studies, many details, such as the contribution of individual species to the total metabolic ability of a microbial community, are not well defined.

In addition to mapping reads onto known protein families, discovering how many protein families are encoded in a metagenome (including known and unknown families) is important for understanding the functionality and functional diversity of a community and is

also important for expanding the protein (family) universe. For example, the Sorcerer II Global Ocean Sampling (GOS) expedition expanded the known protein space by more than doubling its size and adding thousands of new families [52]. Li et al [53] developed an approach for rapid analysis of the sequence diversity for very large metagenomic datasets using a clustering approach powered by CD-HIT algorithm [54]. Similarity search (by BLAST or by a more sensitive similarity search tool) helps predict the function of some of these protein families; yet, the functions of numerous small, rare protein families remain to be discovered. Though we have witnessed the successful application of non-homology based methods to predicting metagenomic sequences (e.g., gene neighborhood methods customized to metagenomics [37]), many of the existing non-homology based function prediction methods [55] cannot be applied for function prediction for metagenomic sequences. And the community is waiting for the new approaches that take advantage of the information provided through metagenomic projects but not in traditional genomic projects.

2.1.4 Comparative metagenomics—Just as comparative genomics, comparative metagenomics is important for functional and evolutionary studies of microbial communities living in different environments, with the hope that fingerprints reflecting known characteristics of the sampled environments can be extracted for interpreting and diagnosing environments [56]. Comparisons have been made among microbial communities collected from different environments (e.g., terrestrial and marine [56]) and different hosts (e.g., lean mouse versus obese mouse [7]). Comparative metagenomics studies have revealed significant variation in microbial composition [7], sequence composition [57], genome size [58], evolutionary rates [37], and metabolic capabilities [50,59].

UniFrac has been a very popular tool for comparing communities based on the lineages they contain ever since it was developed in 2005 [60], partially because of its simple yet useful distance measurement between two communities. Given the sets of taxa for the communities to be compared, UniFrac calculates the phylogenetic distances between two communities (i.e., the phylogenetic distances between their corresponding sets of taxa) as the fraction of the branch length of the phylogenetic tree (with all the taxa from both communities) that leads to the taxa from either one set, but not both. The UniFrac metric can be computed for all pairwise combinations of communities from different environments to make a distance matrix, which can then be used with multivariate statistics such as clustering and principal coordinate analysis to provide an overview of the overall phylogenetic relationship of the communities. Further analyses often include intuitive graphical representation of the relationship of the communities from different environments based on their taxonomic distributions (such as a tree from clustering analysis, and a 2D-graph showing the first two principal components in PCA analysis). UniFrac was first implemented as a tool considering only the qualitative information of lineage distribution of a community, and later incorporated a weighting schema to count for the quantitative information of the taxon distribution (so that more abundant taxa will be weighted more).

MEGAN is another tool that provides visual and statistical comparison of metagenomes based on what the lineages they contain. MEGAN was first developed for studying taxonomic distribution of single samples based on BLAST search result of metagenomic sequences (metagenomic reads are mapped to the tree of known species based on the BLAST top hits). MEGAN 2.0 further incorporated comparative analysis of functionality, and MEGAN 3.0 also included a statistical analysis component for pairwise comparison as discovered below [61,62]. MEGAN can provide a quick glance at the similarity between multiple datasets (if BLAST search results for all the metagenomic datasets are given), and can highlight taxa with statistically different numbers of assigned reads between two datasets.

Both UniFrac and MEGAN enable comparisons of microbial communities based on the lineages they contain. Microbial communities can also be compared based on other types of information, such as the functions encoded by metagenomes. Functional profiling of metagenomes are often used in metagenomic studies, and metagenomes (and their corresponding environments) can be compared based on their functional profiles [16,50], or compared by MG-DOTUR as shown in [63].

2.1.5 Statistical tools for metagenomics—There is no doubt that statistics should be heavily involved in analyzing and interpreting metagenomic data, considering that both sampling and sequencing in metagenomic surveys can be largely viewed as at best a stochastic process (i.e., if no systematic bias was introduced). If one observes some apparent difference between communities, one should determine if that difference is statistically significant or not. In addition, many metagenomic projects adopted NGS techniques—such as the 454 sequencers in its early development, and the current Illumina/Solexa sequencers—that produce ultra-short reads. Those short reads may introduce additional complexity to the interpretation of metagenomic sequences, because short reads are more difficult to annotate [64]. Another problem that obviously needs statistical attention is the estimate of species diversity based on one or many observations of a microbial community (i.e., a dataset of metagenomic shotgun sequences, or a set of 16S rRNAs). Unless there is an application of rigorous statistical analysis, the significance of many conclusions from metagenomic analysis will be discounted. Even a simple task like the estimation of relative abundance of different protein coding gene families in a community seems not so straightforward, as we show in the Artifacts session.

Since exhaustive inventories of microbial communities would be impractical or too expensive, the actual diversity of microbial communities based on metagenomic samples needs instead to be estimated. Nevertheless it is important to know how well a sample reflects a community's "true" diversity ("counting the uncountable" [65]). One class of methods is the *extrapolation* that uses the observed accumulation curve to fit an assumed functional form that models the process of observing new species as sampling effort increases. The species richness expected at infinite effort can then be projected (i.e., as the asymptote of the curve). Parametric estimators are another class of estimation methods, which estimate the number of unobserved species in the community by fitting sample data to models of relative species abundances. Thus, in these methods, an assumption about the true abundance distribution of a community is made. Power-law and the exponential functions are the basic functional forms of relative abundance distribution curves observed for biological populations. And these two forms have been tested to describe the observed distribution of q-contigs (i.e., the groups of q overlapping sequences) to model the viral community structure [66]. Nonparametric estimators have also been used in metagenomics studies, including the Chao1 and abundance-based coverage estimators (ACE) that add a correction factor to the observed number of species [67] or protein families [63]. (Chao1 uses singletons and doubletons, whereas ACE incorporates data from all species with fewer than 10 individuals.).

PHACCS is an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information [68]. PHACCS builds models of possible community structure using a modified Lander-Waterman algorithm to predict the underlying contig spectrum, a vector containing the number of q-contigs from the assembly of shotgun reads, and finds the most appropriate structure model by optimizing the model parameters until the predicted contig spectrum is as close as possible to the one derived experimentally. The resulted structure model provides the basis for making estimates of uncultured viral community richness, evenness, diversity index and abundance of the most abundant genotype.

Statistical techniques are also needed for comparing the structure of microbial communities based on *incomplete* observation/characterizations of microbial communities through metagenomic studies. Phylogeny-based statistical tools for comparing community structures (i.e., what are the members and how they are related to each other) include integral-LIBSHUFF, TreeClimber, UniFrac (see Subsection 2.1.4), analysis of molecular variance (AMOVA, which determines whether the genetic diversity within two or more communities is greater than their pooled genetic diversity), and homogeneity of molecular variance (HOMOVA, which determines whether the amount of genetic diversity in each community is significantly different) [69]. Tools have also been developed for community membership and structure comparison based on operational taxonomic units (OTUs). (DOTUR was developed by Schloss et al. for assigning sequences to OTUs based on the genetic distance between sequences [67]). SONS implements nonparametric estimators for the fraction and richness of OTUs shared between two communities [70]. Many of these tools (including DOTUR, SONS, and TreeClimber) are now all available in a package called MOTHER.

Recently, Metastat [71] was developed for detecting significantly different features (such as taxa, biological pathways, or gene families) between two populations (each population has multiple subjects; e.g., the lean mouse population which has multiple individuals), aiming to study *how* two populations are different from each other. The input of Metastat is a feature (relative) abundance matrix, with rows representing different features and columns representing subjects or replicates (from two populations). For significance estimation, Metastat utilized a nonparametric *t*-test, which was implemented so that it is not limited to the assumption that the underlying distribution is normal. For low frequency features, Metastat utilizes Fisher's exact test, which was demonstrated to outperform other methods for features with sparse counts.

Note that different statistical techniques are used for different tests, and all have their limitations. For example, Schloss has conducted a comparison of different statistical techniques and concluded that different techniques need to be employed correctly for different testing purposes [69].

2.2 Modeling interactions between microbes and their environment (or hosts)

Metagenomics has enabled us to look into hidden microbial world, and started to ask important questions such as how microbes adapt to the environment they live in, and how microbes co-evolve with their hosts (or reshape the environment). Studying the host-bacterial coevolution and microbe-environment interaction has both fundamental and practical values. (For example, these studies help to discover new drug targets [72] and biosensors [59]). Gianoulis et al. introduced a new concept, the metabolic footprint, to describe an ensemble of weighted biological pathways that maximally covaries with a combination of environmental variables (e.g., the temperature and the sample depth) by employing canonical correlation analysis and related techniques (which model many-to-many relationships) [59]. They identified footprints predictive of aquatic environments that can potentially be used as biosensors using available aquatic datasets. Turning to biological or host constrained microbes, Ley et al. [73] applied a network-based analysis of bacterial 16S ribosomal RNA gene sequences from the fecal microbiota of humans and 59 other mammalian species living in two zoos and in the wild. 16S ribosomal RNAs were used to define OTUs of the bacterial species. In the network, the nodes are the OTUs (representing bacterial species) and 60 mammals; the edges connect OTUs with the mammals in which their sequences were found. Probing the coevolution of mammals and their indigenous microbial communities, their analysis (supported by Cytoscape [74]; <http://www.cytoscape.org/>) indicates that host diet and phylogeny both influence bacterial diversity; that bacterial communities codiversified with their hosts; and that the gut microbiota of humans living a modern life-style is typical of omnivorous primates.

3 Artifacts

Superficially, there is no difference at the lowest or nucleotide level between DNA sequence data from conventional genomic projects and the metagenomics sequencing projects—both are just strings of the A, T, C, and G. But metagenomic data differ from conventional genomic data. The patterns and long-range sequence data are different in metagenomics projects. In a conventional genome project, whose goal was to recover a sequence of the complete genome of a single organism, only one genome was represented and the organization of its DNA was known. In contrast, metagenomic sequencing projects *randomly* sample the sequence data from the aggregate members of the community (yielding the sample's metagenome). We do not have a clear model for the organization of the DNA in any given organism, nor know how many different organisms are contained in the sequenced sample. Even the most extensive sequencing of a specific sample (like in the Global Ocean Sampling [4,75]) will provide only a partial sampling of the DNA in a given environment. Less abundant and particularly rare species will yield only incomplete information obtained from DNA fragments from these species. (One cannot get a complete genome or even contigs of reasonable size for those species.) Thus, metagenomic sequence data mining is challenging because of the large scale and complexity of the metagenomic sequence, which means one also could easily make mistakes in solving problems that look so simple at first glance (e.g., the estimation of gene family frequencies for a microbial community). Below we list several artifacts discovered in metagenomic studies that could impact the conclusion drawn from those studies, and the new computational and statistical tools developed with the aim of reducing the impact of these artifacts.

3.1 16S rRNA chimeras could lead to inaccurate estimation of the species diversity of a community

The application of barcoded pyrosequencing can produce large 16S rRNA datasets that contain hundreds and thousands of 16S RNAs, enabling deep views into hundreds of microbial communities simultaneously [9]. Chimera among the sequenced 16S rRNAs can affect the estimates of species diversity within a microbial community sample, if they are not filtered out properly in advance. (A chimeric sequence usually comprises two phylogenetically distinct parent sequences, and in rare cases, they may be from more than two parents.) Chimeras are well known in community profiling with 16S rRNA: chimeras occur when a prematurely terminated amplicon reanneals to a foreign DNA strand and is copied to completion in the following PCR cycles. Ashelford et al. reported that at least one in twenty 16S rRNA sequences held in public repositories is estimated to contain substantial anomalies [76]. Improved experimental techniques (protocols), e.g., the development of emulsion PCR [77] serve to minimize the generation of chimera. Computational techniques have also been developed for removing chimeric 16S rRNA data, including Bellerophon [78], Pintail [76], and Mallard [79]; most of such tools are based on comparison with reference 16S rRNAs. The scalability of these tools, though, needs to be improved so that they can work with large 16S rRNA datasets that are derived from barcoded pyrosequencing.

3.2 Artificial replicates may introduce systematic artifacts to the estimation of gene and taxon abundance

A recent communication by Gomez-Alvarez et al [80] reports a systematic error in metagenomes generated by 454 pyrosequencing and introduced by artificial replicates that may cause incorrect estimation of the gene and taxon abundance. They discovered that between 11% and 35% of sequences in a typical metagenome are artificial replicates, which need to be removed before further analysis. In most metagenomic studies, only the exact duplicates were removed, which however, only account for 3–18% of the artificial

replicates. Geomez-Alvarez et al implemented a pipeline for removing these artifacts by identifying not only exact duplicates, but also other types of artificial replicates, i.e., reads that began at the same position but varied in length or contained a sequencing discrepancy. (A simple statistical analysis would show that the probability of multiple reads occurring at the same position at random, given independent sampling with replacement, is very low.) They also discussed that failure to remove replicated sequences can lead to incorrect conclusions. One example is that the fraction of genes identified as being involved in denitrification did not differ much between agricultural and forested sites before and after removing duplicate reads (exact replicates), but a statistically significant difference was detected after removing *all* replicate reads.

3.3 Gene family frequencies derived based on read counts in metagenomic data may be unreliable due to different gene family lengths

Computing gene family frequencies seems not so straightforward since Sharon et al. addressed such calculations [81]. We use here an extreme (yet simple) example to explain the problem. Suppose functional annotation exists for all of the reads in a metagenomic dataset. Suppose 1000 and 500 reads are found with function A and B, respectively. It would seem very obvious that function A is more abundant than function B in the corresponding microbial community. And based on this simple read counting, one might guess that function A is likely to be more important than function B in this environment, which unfortunately could be wrong. The hidden information that was ignored is that proteins with function A may be significantly longer than proteins with function B. Even though there are the exactly the same number of copies of the genes encoding function A and B in the environment, *randomly* sampling and sequencing of the community DNA will result in a significantly higher number of reads that encode function A (which is longer) than function B. As this example shows, even a very simple analysis, such as the functional distribution studies, could turn wrong if we are not very careful with the interpretation. Sharon et al. proposed a statistical framework for assessing gene family frequencies by considering the different lengths of various proteins [81]. Their statistical model is based on the Lander-Waterman model [82], proposed in 1988, that describes the shotgun sequencing of a genome as a Poisson process. By applying this statistical model, the gene family frequencies (estimated by read-count approach) will be adjusted, with frequencies for short gene families increased while frequencies for long gene families decreased.

3.4 Be aware of artificial pathways

Since biological pathways are important contexts for understanding the functionality of a microbial community, biological pathway based analysis are frequently attempted in metagenomic studies. The random sampling of metagenomic sequences from microbial community genomes will certainly leave gaps in the reconstructed pathways for any metagenome. The “optimistic” way to handle this situation is to consider that a pathway is encoded by the metagenome whenever a functional role (such as an enzyme) can be found in the metagenomic sequences, independent of the presence of pathway gaps or holes (missing enzymes). Such an optimistic pathway reconstruction by a naïve mapping strategy, as termed by the authors [83], may leave us with many artificial pathways because of the redundant nature of the mapping from functional roles to pathways; that is, some functional roles could be involved in multiple pathways. The entire cellular network is partitioned into several hundreds of biological pathway entities (such as documented in the KEGG database), which is extremely important for an understanding of biological processes, even though all pathways are to some extent connected and overall, there is a singular, integrated network within any cell [84]. Thus, It is not surprising that many pathways defined in the pathway databases overlap. Also, some proteins carry out multiple biological functions [85] through, for example, different protein domains, active sites, or substrate specificities.

Ignoring the pathway redundancy may also leave us with artificial pathways. Examples of potential artificial pathways reconstructed from metagenomes include the inositol metabolism pathway, the androgen and estrogen metabolism pathway, and the caffeine metabolism pathway annotated for a coral biome [50], which obviously do not sound likely for this particular biome.

A new computational tool, MinPath, was proposed to reduce these artificial pathways based on a parsimony approach [83]. In MinPath, the goal was to define the minimal set of pathways that can explain all annotated functions, instead of using all the pathways that have at least one function mapped in the sequences. For all of the datasets we tested, MinPath reduced the total number of annotated pathways (or subsystems) significantly. For example, for the metagenome sampled from a coral microbial community, there are in total 232 KEGG biological pathways annotated in at least one of its 7 sequencing datasets. Based on MinPath, however, only 160 KEGG biological pathways are sufficient to explain all of the functions predicted for these datasets. These results indicate that the naïve mapping of the biological pathways from predicted functions may overestimate the biological pathways (and in turn, the functional diversity) of those microbial communities, and one needs to be cautious when interpreting the results from such an analysis.

4 Challenges

The growth of public DNA sequence data over the last two decades has been exponential, with a doubling time of about 14 months; the doubling time appears likely to be substantially shortened due to the addition of metagenomic data [15], as well as the impact from NGS approaches. The preliminary data from the Global Ocean Survey (GOS), with 6,109,770 ORFs, more than quadrupled the number of known existing (predicted) protein-coding ORFs (data from CAMERA database [86]; <http://camera.calit2.net/>). Similarly, as of Aug 25, 2009, the IMG/M system (<http://img.jgi.doe.gov/cgi-bin/m/main.cgi>) had collected 67 metagenomes, and the GOLD genomes online database (<http://www.genomesonline.org/>) listed a total of 168 completed or ongoing metagenomic projects. The increasingly longer read lengths from 454 runs and the rapidly falling cost of sequencing will expedite this process. Note that recent second-generation sequencing techniques, such as the Solexa/Illumina sequencer, still generate extremely short reads. The massive metagenomic data poses great challenges in many areas involving data management and data mining.

4.1 Scalability

There have recently been initial developments of faster and scalable tools toward the requirements of metagenomic projects. The developments include a faster similarity search tool, FastBLAST [87]; chimera removal tools that can handle 16S rRNA datasets from barcoded pyrosequencing; and a clustering method that can handle large scale data: ESPRIT [88]. We expect more developments of efficient and powerful computational tools that can deal with (or, even better, take advantage of) the huge amount of metagenomic sequences. Besides, these tools need to be able to deal with the high species complexity of the metagenomic datasets.

4.2 Integration of metaproteomic, metatranscriptomic and metagenomics data sets

Several recent researches targeted at the expressed genetic information and proteins of a microbial community, aiming to reveal the regulation and dynamics of genes in an environment. A pysequencing of cDNA prepared from ocean water samples revealed a large number of microbial small RNAs in the ocean water sample, and many of these RNAs have not seen before [89]. A shotgun mass spectrometry-based whole community proteomics approach resulted in a metaproteome with thousands of proteins in human fecal samples,

which showed a skewed distribution relative to the metagenome when compared to what was earlier predicted from metagenomics [90]. There is no doubt that metaproteomic and metatranscriptomic studies, together with metagenomic studies offer an unprecedented opportunity to explore both the organization and function of microbial communities. Integration of various datasets will be challenging since there is often lack of direct cross-reference. A research group may only look at one functional aspect of a microbial community, like the gene expression, but not all. A direct result of this is the knowledge gap between different projects. (For example, only ~50% of the transcripts from an ocean sample are highly similar to genes previously detected in ocean metagenomic survey [91].) However, it is clear that there will be significant reward to integrating all of these different pieces of information to gain a more comprehensive picture of the organization and function of microbial communities.

4.3 And use your imagination!

Metagenomics (and related research) provides us with enormous opportunities to dig into the hidden world of microbes and their relation with biological hosts and/or physical environments. Probably, the only thing that might slow down our scientific march is our imagination, since producing genomic sequences from microbial communities is no longer a stumbling block in our journey to revealing the secrets of our microbial planet.

Acknowledgments

The authors would like to acknowledge the support from NIH grant 1R01HG004908-01 and NSF grant DBI-0845685 (YY), and from the Gordon and Betty Moore Foundation for the Community Cyberinfrastructure for Marine Microbial Ecological Research and Analysis (CAMERA) project (JW).

References

- Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* 1998;5:R245–R249. [PubMed: 9818143]
- Mardis E. Anticipating the 1,000 dollar genome. *Genome Biol* 2006;7:112. [PubMed: 17224040]
- Tyson G, Chapman J, Hugenholtz P, Allen E, Ram R, Richardson P, Solovyev V, Rubin E, Rokhsar D, Banfield J. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 2004;428:37–43. [PubMed: 14961025]
- Venter J, Remington K, Heidelberg J, Halpern A, Rusch D, Eisen J, Wu D, Paulsen I, Nelson K, Nelson W, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004;304:66–74. [PubMed: 15001713]
- Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, Hatay M, Hall D, Brown E, Haynes M, et al. Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE* 2008;3:e158.
- Lorenz P, Eck J. Metagenomics and industrial applications. *Nat Rev Microbiol* 2005;3:510–516. [PubMed: 15931168]
- Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al. A core gut microbiome in obese and lean twins. *Nature* 2009;457:480–484. [PubMed: 19043404]
- Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. *Nature* 2007;449:804–810. [PubMed: 17943116]
- Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 2008;5:235–237. [PubMed: 18264105]
- Li L, McCorkle S, Monchy S, Taghavi S, van der Lelie D. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnol Biofuels* 2009;2:10. [PubMed: 19450243]

11. Brulc J, Antonopoulos D, Miller M, Wilson M, Yannarell A, Dinsdale E, Edwards R, Frank E, Emerson J, Wacklin P, et al. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A* 2009;106:1948–1953. [PubMed: 19181843]
12. Jones B, Begley M, Hill C, Gahan C, Marchesi J. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc Natl Acad Sci U S A* 2008;105:13580–13585. [PubMed: 18757757]
13. Mori T, Mizuta S, Suenaga H, Miyazaki K. Metagenomic screening for bleomycin resistance genes. *Appl Environ Microbiol* 2008;74:6803–6805. [PubMed: 18791008]
14. Steele H, Jaeger K, Daniel R, Streit W. Advances in recovery of novel biocatalysts from metagenomes. *J Mol Microbiol Biotechnol* 2009;16:25–37. [PubMed: 18957860]
15. Handelsman, J.; Tiedje, JM.; Alvarez-Cohen, L.; Ashburner, M.; Cann, IKO.; Delong, EF.; Doolittle, WF.; Fraser-Liggett, CM.; Godzik, A.; Gordon, JI., et al. The new science of metagenomics: Revealing the secrets of our microbial planet. The National Academies Press; 2007.
16. Tringe S, von Mering C, Kobayashi A, Salamov A, Chen K, Chang H, Podar M, Short J, Mathur E, Detter J, et al. Comparative metagenomics of microbial communities. *Science* 2005;308:554–557. [PubMed: 15845853]
17. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 2006;444:1027–1131. [PubMed: 17183312]
18. Hooper SD, Raes J, Foerstner KU, Harrington ED, Dalevi D, Bork P. A molecular study of microbe transfer between distant environments. *PLoS ONE* 2008;3:e2607. [PubMed: 18612393]
19. Raes J, Foerstner KU, Bork P. Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* 2007;10:490–498. [PubMed: 17936679]
20. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* 2008;72:557–578. Table of Contents. [PubMed: 19052320]
21. Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* 2009;19:1141–1152. [PubMed: 19383763]
22. Galperin M. Metagenomics: from acid mine to shining sea. *Environ Microbiol* 2004;6:543–545. [PubMed: 15142241]
23. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* 2008;18:802–809. [PubMed: 18332092]
24. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* 2008;18:810–820. [PubMed: 18340039]
25. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. *Genome Res* 2008;18:324–330. [PubMed: 18083777]
26. Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinform* 2009;10:354–366. [PubMed: 19482960]
27. Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* 2006;34:5623–5630. [PubMed: 17028096]
28. Hoff KJ, Tech M, Lingner T, Daniel R, Morgenstern B, Meinicke P. Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinformatics* 2008;9:217. [PubMed: 18442389]
29. Hoff KJ, Lingner T, Meinicke P, Tech M. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res* 2009;37:W101–105. [PubMed: 19429689]
30. Krause L, Diaz NN, Bartels D, Edwards RA, Puhler A, Rohwer F, Meyer F, Stoye J. Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics* 2006;22:e281–289. [PubMed: 16873483]
31. Ye Y, Tang H. An orfome assembly approach to metagenomics sequences analysis. *J Bioinform Comput Biol* 2009;7:455–471. [PubMed: 19507285]

32. Cardenas E, Tiedje J. New tools for discovering and characterizing microbial diversity. *Curr Opin Biotechnol* 2008;19:544–549. [PubMed: 18984052]
33. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res* 2007;17:377–386. [PubMed: 17255551]
34. Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods* 2007;69:330–339. [PubMed: 17391789]
35. Monier A, Claverie JM, Ogata H. Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* 2008;9:R106. [PubMed: 18598358]
36. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. *Science* 2006;311:1283–1287. [PubMed: 16513982]
37. von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 2007;315:1126–1130. [PubMed: 17272687]
38. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 2008;9:R151. [PubMed: 18851752]
39. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* 2008;36:2230–2239. [PubMed: 18285365]
40. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al. Pfam: clans, web tools and services. *Nucleic Acids Res* 2006;34:D247–251. [PubMed: 16381856]
41. Bentley SD, Parkhill J. Comparative genomic structure of prokaryotes. *Annu Rev Genet* 2004;38:771–792. [PubMed: 15568993]
42. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 2004;5:163. [PubMed: 15507136]
43. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, et al. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 2006;443:950–955. [PubMed: 16980956]
44. Chatterji, S.; Yamazaki, I.; Bai, Z.; Eisen, J. CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. The 12th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2008; Singapore; Springer. 2008. p. 17-28.
45. Zhou F, Olman V, Xu Y. Barcodes for genomes and applications. *BMC Bioinformatics* 2008;9:546. [PubMed: 19091119]
46. Brady A, Salzberg SL. Phymm and Phymm BL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods*. 2009
47. Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, Joint I. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One* 2008;3:e3042. [PubMed: 18725995]
48. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30. [PubMed: 10592173]
49. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 2005;33:5691–5702. [PubMed: 16214803]
50. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, et al. Functional metagenomic profiling of nine biomes. *Nature* 2008;452:629–632. [PubMed: 18337718]
51. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008;9:386. [PubMed: 18803844]

52. Yoosseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* 2007;5:e16. [PubMed: 17355171]
53. Li W, Wooley JC, Godzik A. Probing metagenomics by rapid cluster analysis of very large datasets. *PLoS One* 2008;3:e3375. [PubMed: 18846219]
54. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 2001;17:282–283. [PubMed: 11294794]
55. Marcotte EM. Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol* 2000;10:359–365. [PubMed: 10851184]
56. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, et al. Comparative metagenomics of microbial communities. *Science* 2005;308:554–557. [PubMed: 15845853]
57. Foerster KU, von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. *EMBO Rep* 2005;6:1208–1213. [PubMed: 16200051]
58. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. Prediction of effective genome size in metagenomic samples. *Genome Biol* 2007;8:R10. [PubMed: 17224063]
59. Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, et al. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* 2009;106:1374–1379. [PubMed: 19164758]
60. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005;71:8228–8235. [PubMed: 16332807]
61. Huson DH, Richter DC, Mitra S, Auch AF, Schuster SC. Methods for comparative metagenomics. *BMC Bioinformatics* 2009;10 (Suppl 1):S12. [PubMed: 19208111]
62. Mitra S, Klar B, Huson DH. Visual and Statistical Comparison of Metagenomes. *Bioinformatics*. 2009
63. Schloss PD, Handelsman J. A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* 2008;9:34. [PubMed: 18215273]
64. Wommack KE, Bhavsar J, Ravel J. Metagenomics: read length matters. *Appl Environ Microbiol* 2008;74:1453–1463. [PubMed: 18192407]
65. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* 2001;67:4399–4406. [PubMed: 11571135]
66. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* 2002;99:14250–14255. [PubMed: 12384570]
67. Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* 2005;71:1501–1506. [PubMed: 15746353]
68. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* 2005;6:41. [PubMed: 15743531]
69. Schloss PD. Evaluating different approaches that test whether microbial communities have the same structure. *ISME J* 2008;2:265–275. [PubMed: 18239608]
70. Schloss PD, Handelsman J. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol* 2006;72:6773–6779. [PubMed: 17021230]
71. White J, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 2009;5:e1000352. [PubMed: 19360128]
72. Zaneveld J, Turnbaugh PJ, Lozupone C, Ley RE, Hamady M, Gordon JJ, Knight R. Host-bacterial coevolution and the search for new drug targets. *Curr Opin Chem Biol* 2008;12:109–114. [PubMed: 18280814]

73. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, Gordon JI. Evolution of mammals and their gut microbes. *Science* 2008;320:1647–1651. [PubMed: 18497261]
74. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498–2504. [PubMed: 14597658]
75. Rusch DB, Halpern AL, Sutton G, et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 2007;5(3):e77. [PubMed: 17355176]
76. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol* 2005;71:7724–7736. [PubMed: 16332745]
77. Williams R, Peisajovich S, Miller O, Magdassi S, Tawfik D, Griffiths A. Amplification of complex gene libraries by emulsion PCR. *Nat Methods* 2006;3:545–550. [PubMed: 16791213]
78. Huber T, Faulkner G, Hugenholtz P. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* 2004;20(14):2317–2319. [PubMed: 15073015]
79. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. New Screening Software Shows that Most Recent Large 16S rRNA Gene Clone Libraries Contain Chimeras. *Appl Environ Microbiol* 2006;72:5734–5741. [PubMed: 16957188]
80. Gomez-Alvarez V, Teal T, Schmidt T. Systematic artifacts in metagenomes from complex microbial communities. *ISME J*. 2009
81. Sharon, I.; Pati, A.; Markowitz, VM.; Pintter, RY. RECOMB 2009; Tucson, AZ. Springer; Berlin/Heidelberg: 2009. A Statistical Framework for the Functional Analysis of Metagenomes; p. 496–511.
82. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 1988;2:231–239. [PubMed: 3294162]
83. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol* 2009;5:e1000465. [PubMed: 19680427]
84. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res* 2008;36:W423–426. [PubMed: 18477636]
85. Rosin FM, Watanabe N, Lam E. Moonlighting vacuolar protease: multiple jobs for a busy protein. *Trends Plant Sci* 2005;10:516–518. [PubMed: 16202638]
86. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. CAMERA: a community resource for metagenomics. *PLoS Biol* 2007;5:e75. [PubMed: 17355175]
87. Price MN, Dehal PS, Arkin AP. FastBLAST: homology relationships for millions of proteins. *PLoS One* 2008;3:e3589. [PubMed: 18974889]
88. Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W, Farmerie W. ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* 2009;37:e76. [PubMed: 19417062]
89. Shi Y, Tyson GW, DeLong EF. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* 2009;459:266–269. [PubMed: 19444216]
90. Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, Lefsrud MG, Apajalahti J, Tysk C, Hettich RL, Jansson JK. Shotgun metaproteomics of the human distal gut microbiota. *ISME J* 2009;3:179–189. [PubMed: 18971961]
91. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF. Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A* 2008;105:3805–3810. [PubMed: 18316740]

Biographies

John Wooley is Associate Vice Chancellor of Research at UCSD. He now largely focuses on structural genomics (SG) and metagenomics (MG), along with various bioinformatics

and computational methods for probing SG and MG data and for community engagement via web services technology.

Yuzhen Ye is an Assistant Professor at the School of Informatics and Computing, Indiana University, Bloomington. Ye received her Ph.D degree in computational biology from Shanghai Institute of Biochemistry, Chinese Academy of Sciences in 2001. Ye's research interests are in the areas of bioinformatics (especially structural bioinformatics), and computational metagenomics. Ye received a NIH grant for developing computational tool for human microbiome project in 2008, and NSF CAREER Award in 2009.