A Bioinformatician's Guide to Metagenomics

Victor Kunin[1], Alex Copeland [2], Alla Lapidus [3], Konstantinos Mavromatis [4] and Philip Hugenholtz[1¶].

[1] Microbial Ecology Program, [2] Quality Assurance Department, [3] Microbial Genomics Department, [4] Genome Biology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA.

[¶]Corresponding author: fax 925-296-5720 • email: phugenholtz@lbl.gov

Summary

As random shotgun metagenomic projects proliferate and become the dominant source of publicly available sequence data, procedures for best practices in their execution and analysis become increasingly important. Based on our experience at the Joint Genome Institute, we describe step-by-step the chain of decisions accompanying a metagenomic project from the viewpoint of a bioinformatician. We guide the reader through a standard workflow for a metagenomic project beginning with pre-sequencing considerations such as community composition and sequence data type that will greatly influence downstream analyses. We proceed with recommendations for sampling and data generation including sample and metadata collection, community profiling, construction of shotgun libraries and sequencing strategies. We then discuss the application of generic sequence processing steps (read preprocessing, assembly, and gene prediction and annotation) to metagenomic datasets by contrast to genome projects. Different types of data analyses particular to metagenomes are then presented including binning, dominant population analysis and gene-centric analysis. Finally data management systems and issues are presented and discussed. We hope that this review will assist bioinformaticians and biologists in making better-informed decisions on their journey during a metagenomic project.

Introduction

For the purposes of this review, we define metagenomics as the application of random shotgun sequencing to DNA obtained directly from an environment sample or series of related samples. This is to distinguish it from functional metagenomics, reviewed elsewhere (54), whereby environmental DNA is cloned and screened for specific functional activities of interest. Metagenomics is a derivation of conventional microbial genomics, with the key difference that it bypasses the requirement for obtaining pure cultures for sequencing. Therefore metagenomics holds the promise of revealing the genomes of the majority of microorganisms that cannot be readily obtained in pure culture (58). In addition, since the samples are obtained from communities rather than isolated populations, the structure of and interactions in the communities can potentially

be elucidated. In this review we address the bioinformatic aspects of analyzing metagenomic datasets, stressing the differences with standard genomic analyses. Although our focus is on bioinformatics, we will begin by considering experimental planning and implementation of metagenomic projects as these can have major impacts on the subsequent bioinformatic analyses.

Throughout the review we will follow the workflow of a typical metagenomic project, summarized in **Fig. 1**. This process begins with sample and metadata collection, proceeds with DNA extraction, library construction, sequencing, read preprocessing and assembly. Genes are then called on either reads or contigs, or both, and binning is applied. Community composition analysis is employed at several stages of this workflow, and databases are used to facilitate the analysis. All of these stages will be discussed in detail below.

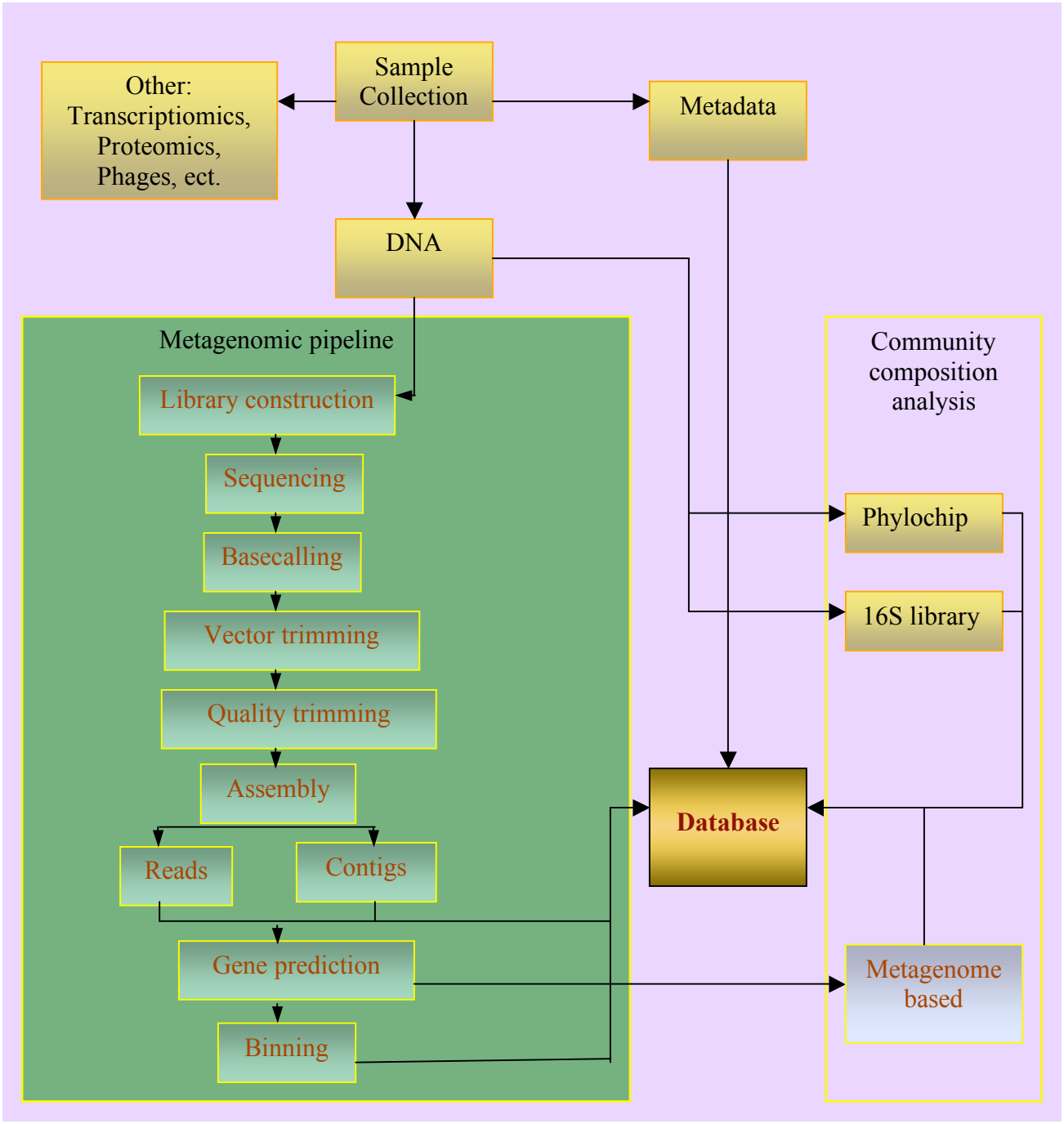


Fig. 1. A typical workflow for metagenomic projects at the JGI. The process begins with sample and metadata collection, proceeds with DNA extraction, library construction, sequencing, basecalling, vector and quality trimming, assembly, gene prediction and binning. Community composition analysis is applied in several forms, both prior and during a metagenomic project. See text for discussion.

Pre-sequencing considerations

Community composition

Metagenomic bioinformatics should begin before a single nucleotide of DNA has been sequenced. When a community is selected for metagenomic analysis, its species composition (number and relative abundance and if possible genome sizes) should be assessed with respect to the amount of allocated sequence. The community composition has a deciding influence on the types of analyses that can be performed on the sequence dataset. A complex microbial community usually includes bacteria, archaea, microbial eukaryotes and viruses. Historically however, microbiologists are trained to think of themselves as either bacteriologists or virologists or protistologists and ecological studies investigating more than one of these taxonomic groups are still remarkably uncommon (70). To be frank, the authors are no exception, therefore when we talk about community composition in the following sections, we are primarily referring to bacterial and archaeal species that have been the focus of most of our metagenomic studies.

At current sequencing capacity, metagenomic sequencing of communities containing eukaryotes, in particular protists, is mostly cost-prohibitive because of their enormous genome sizes and low gene-coding densities (127). Therefore selecting a community that does not contain eukaryotes, or from which eukaryotes or their DNA can be excluded, is an important consideration prior to embarking on a metagenomic analysis. For example, one of the main reasons that the hindgut of a higher, rather than lower termite was sequenced (138) is because the former lacks protist symbionts. When sequencing microbial communities that are found in tight symbiotic relationships with eukaryotic hosts, removal of host cells or extracted host DNA is important to avoid eukaryotic contamination. For example, in the analysis of a gutless worm microbial symbiont community, host cells were physically separated from bacterial endosymbiont populations using a nycodenz gradient (142).

Simply excluding eukaryotes from a metagenomic analysis is not ideal from an ecological perspective as it compromises our ability to assess a microbial community in its entirety. An alternative or complementary strategy could be to obtain molecular data at the RNA (metatranscriptomics) or protein (metaproteomics) level, thus bypassing the

problem of large amounts of non-coding eukaryotic sequence data. Emerging sequencing technologies such as pyrosequencing (84) may ultimately allow metagenomic sequencing of communities comprising eukaryotes, but the data is likely to present numerous challenges for many downstream bioinformatic analyses (see Selecting the sequencing technology).

Within the sequence-tractable bacterial, archaeal and viral components of a community, a key variable is species abundance distribution, in particular the presence or absence of dominant populations. Dominant populations that comprise more than a few percent of the total number of cells or virions in a community will have higher representation in a metagenomic dataset resulting in a greater likelihood of assembly and recovery of contigs (contiguous genomic stretches comprised of overlapping reads). Note that we define assembled contigs arising from a population as composite genomic fragments because each component read likely comes from a different individual within the population in which individuals are usually not clonal.

We will distinguish between two basic types of community composition throughout this review; “complex” and “simple”. Communities of the first type lack populations abundant enough to result in assembled contigs >10 kbp (**Fig. 2**). Such communities also tend to be species rich, for example soil (129). Communities of the second type have one or more dominant populations producing contigs >10 kbp up to several 100 kbp. Examples include simple communities that are mostly comprised of a few dominant species, such as acid mine drainage (132) or a gutless worm symbiont community (142). Some communities have hallmarks of both types, in which dominant populations are flanked by a long tail of low abundance species, such as Enhanced Biological Phosphorus Removing (EBPR) sludge (43).

Sequencing of a community with dominant species is likely to reproduce a significant part of the genomes of the dominant organisms, and in some cases near complete genomes (43, 132). Therefore, analysis of large genomic fragments is similar to conventional comparative genomics. In contrast, sequences obtained from a complex system without dominating species will not contain large genomic fragments of any component population using current technologies (129, 133). The analysis therefore will

normally be focused on averaged properties of the community, such as gene content and abundance, since information on any given component species will be sparse.

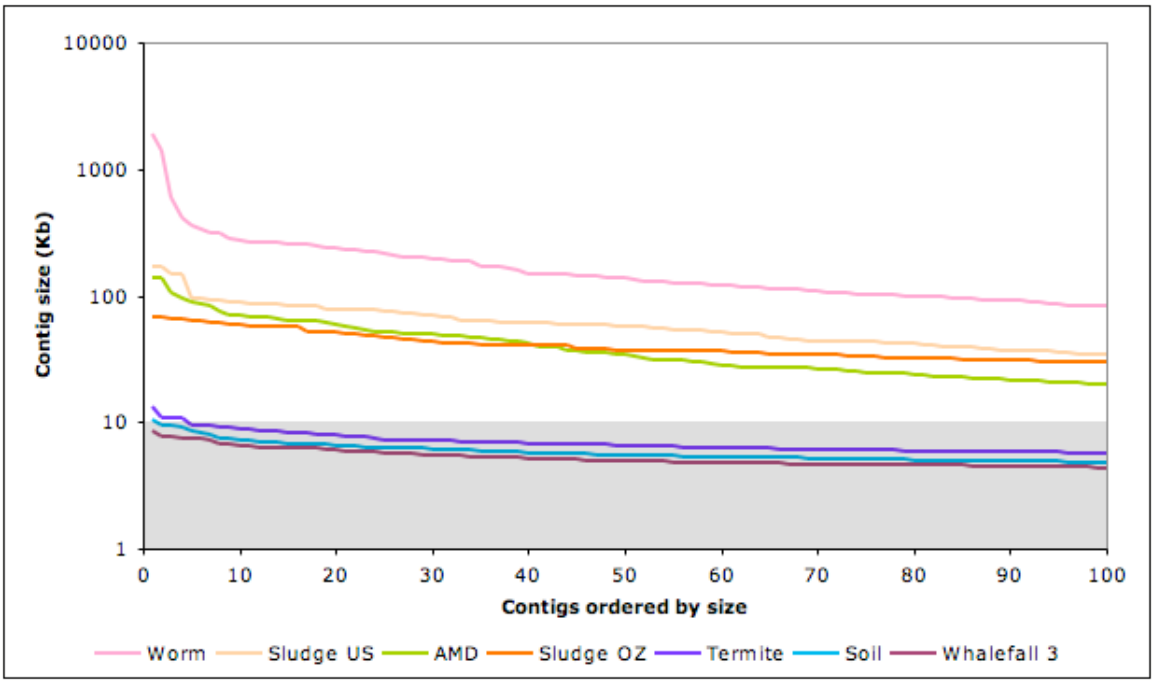


Fig. 2. Contig size distribution of assembled metagenomic datasets from seven microbial communities. The grey area denotes small contigs with a higher likelihood of chimeric assemblies (see Assembly). Communities with contigs found mostly in this zone (termite hindgut (138), soil and whalefall (129)) lack dominant populations whereas communities with larger contigs outside this zone have dominant populations; gutless worm (142), phosphorus-removing sludges from lab-scale bioreactors (43) and an acid mine drainage (AMD) biofilm (132).

Selecting the sequencing technology

The number of sequencing technologies is currently expanding, drawn by demand to bring down the cost of sequencing. While Sanger sequencing (112, 113) so far remains the major source of metagenomic sequence data, alternative strategies have also been used, namely pyrosequencing (84) which has been applied to viral (8) and bacterial (32) communities. Advantages of pyrosequencing over Sanger sequencing include much lower per base cost and no requirement for cloning (108). The latter is useful for both bacterial and virion communities because of demonstrated cloning bias of bacterial genes (121) and promoters (44) in *E. coli* and difficulties with cloning viral nucleic acids (13).

However, the major disadvantage of pyrosequencing is the average read length, initially ~100 bp on the GS20 platform and currently ~200 bp on the GS FLX platform. Reads of this length present additional challenges for assembly and gene calling. Indeed, most studies that have used pyrosequencing for metagenomic analysis did not attempt assembly or gene calling, instead relying on similarity searches of the short reads against a reference database as the basis of the analysis (8, 32) (see also Table 1). Therefore, the bioinformatics processing sections below mostly refer to Sanger data. Notably however, 454 Life Sciences is currently evaluating 400-500 bp (XLR) pyroreads (http://www.454.com/) and if technical problems associated with longer reads, such as reagent dilution and maintaining nucleotide extension synchronization (108), can be adequately addressed to produce read quality comparable to Sanger data, then pyrosequencing will be able to supplant Sanger sequencing as the preferred data type for metagenomic analysis.

Combinations of different sequencing technologies have been evaluated for producing high quality draft assemblies of microbial isolates (47) that could be applied to metagenomes containing one or more dominant populations. The Illumina (http://www.illumina.com) and ABI SOLiD (http://www.appliedbiosystems.com) sequencing technologies have not yet been applied to environmental samples, but their application is likely to be limited to resequencing of dominant populations since reads are currently too short (25-35 bp) to be used for *de novo* assembly or gene calling. One next generation sequencing technology worth keeping an eye on is real time single molecule sequence determination that aims to produce multi-kilobase length reads at throughputs comparable to the short read technologies (67) and (http://visigenbio.com). If such an ambitious goal can be achieved with acceptable sequence quality and cost, then this platform will become the choice for metagenomic studies, since even single reads will contain contextual data of one or more neighboring genes and assembly will be simplified.

How much sequence data?

A common question asked by researchers embarking on their first metagenomic analysis is how much sequence data should they request or allocate for their project.

Unlike genome projects, metagenomes have no fixed end point, i.e. a completed genome. Therefore, decisions on how much sequence data to generate for an environmental sample have been based on pragmatic reasons, chiefly sequencing budget. For example, 100 Mbp is a typical Sanger sequencing request for a metagenomic project through the JGIs community sequencing program (http://www.jgi.doe.gov/CSP/index.html). However, with the per base cost of sequencing continuing to drop, other more objective criteria can be brought to the fore, such as estimates of sequence coverage (number of reads covering each base in a contig) of the community. Since species do not have uniform abundance in a community, it is simpler to address coverage of individual populations for which an approximate average genome size is known. For example if a dominant population represents 10% of the total community and 100 Mbp are obtained, then this population is expected to be represented by 10 Mbp, assuming completely random sampling of the community. If the average genome size of individuals in this population is 2 Mbp, then an average of 5X coverage of the composite population genome will be expected. To place this in perspective, 6-8X coverage of microbial isolates is a common target to obtain a draft genome suitable for finishing. Ultimately, the objectives of the study should guide sequence allocation. For example, if the aim is to determine the SNP frequency profile of a dominant population as part of a population genetic analysis (63), then, ideally, a coverage of 20X or higher will be needed for the dominant population. If the aim is to identify over-represented gene functions in the community as a whole (see gene-centric analysis), then much less sequence data will be needed. Indeed, we recently found that extremely low coverage of a highly complex and stratified hypersaline mat community (estimated dominant population coverage of <0.01X) was still sufficient to detect genetic gradients in the mat community using 10 Mbp per layer (Kunin et al., unpublished data).

Sampling and data generation

Sample collection for metagenomes and other molecular analyses

Metagenomes are sequence inventories of genomic DNAs from environmental samples. Extracting and purifying high quality DNA is still one of the main bottlenecks in

metagenomics, compounded by the fact that there is not a "one size fits all" extraction method for all environmental samples. Low biomass samples yield small quantities of DNA that may be insufficient for library construction. In general microgram quantities of genomic DNA are required for cloning (see clone libraries) and pyrosequencing. Whole genome amplification has been used on small yields of environmental DNAs to provide microgram quantities for sequencing (8), but relative representation of genomic DNAs may be compromised by this process (105). This is important for downstream comparative analyses, particularly between samples that used whole genome amplification and those that did not.

In many cases it may be beneficial to collect additional sample material for complementary analyses. Examples of additional molecular analyses that will leverage and enhance metagenomic data include metatranscriptomics (50, 70), metaproteomics (76) and viral metagenomics (33). While it is sometimes possible to resample many habitats, two temporally separated samples may not be directly comparable. For example, habitats that have seasonal patterns such as the marine water column (28) can not be considered equivalent at different times of the year. Even in habitats that do not show seasonal variation such as controlled lab-scale bioreactors, community composition may be influenced by predators, parasites or other variables that confound comparisons of metagenomic data. For example, from an initial metagenomic analysis of two lab-scale sequencing batch reactors, we implicated bacteriophage as important determinants in driving bacterial community composition (50). Unfortunately, we did not have appropriately stored material from the original sampling and characterized the virion community in a reactor sample taken 7 months after the initial metagenomic sampling. During this time, both the bacterial and viral communities had changed complicating the comparative analysis. It is of course impossible to store sample material in the appropriate manner for every conceivable downstream molecular analysis, but as a number of techniques become more routine, such as metatranscriptomics, metaproteomics, metabolomics and viral metagenomics, subsamples can be inexpensively stored in standardized ways to provide researchers with the potential to perform these analyses if needed.

Sample metadata collection

Collecting collateral, non-sequence data associated with an environmental sample greatly enhances the ability to interpret the sequence data, particularly for comparative analysis of temporal or spatial series (29, 133). Such "metadata" include biochemical data, such as pH, temperature, salinity; geographical data such as GPS (global positioning system) coordinates, depth, height and sample processing data, such as collection date, DNA extraction method and clone library details. The type of metadata can vary considerably depending on the sample type, for instance environmental and clinical samples historically have very different metadata. Databases housing metagenomic data already include varying degrees of metadata (86, 117), but cross-referencing such data is problematic due to a lack of consistency and standards. Initiatives are underway to standardize metadata collection, e.g. by use of a controlled vocabulary where possible (38). Such data are expected to prove invaluable once enough data is generated to compare communities along environmental, spatial or longitudinal gradients (133).

Pre-metagenome community composition profiling

To facilitate decisions on sequence allocation and processing, the community composition of the environmental sample under study should be assessed prior or at least in parallel to the metagenomic analysis using a conserved marker gene survey, ideally conducted on the same sample. Indeed, several samples could be prescreened using marker genes to aid in selection of a subset for metagenomic analysis. The small subunit ribosomal RNA (16S rRNA) gene is usually the marker gene of choice owing to its widespread use and consequent large reference database (21, 30). One drawback of the 16S rRNA gene is that copy number can vary by an order of magnitude between bacterial species that, along with PCR induced biases (124, 136), can skew estimates of community composition. PCR products are normally cloned and sequenced to provide a semi-quantitative phylogenetic profile of a community. At the JGI, we typically sequence one 384 well plate containing 16S clones (called a ribosomal panel) to provide a baseline estimate of community structure.

For most microbial communities however, 384 clones is a gross undersampling of diversity and highlights only relatively dominant taxa. Other approaches that have higher

resolution include microarrays to which fluorescently labeled 16S PCR amplicons or rRNAs are applied (15, 100, 103). For example, the Phylochip comprises 500,000 probes redundantly targeting ~9000 phylogenetic groups (operational taxonomic units) and has one to two orders of magnitude higher sensitivity than a PCR clone library sequenced to ~10^2 (15). On the downside, species that are not represented by probes on the microarray will be missed and relative abundance of sequence types cannot be easily estimated. This means that dominant populations are currently difficult to detect from Phylochip data alone.

Pyrosequencing has recently been applied to PCR-amplified 16S rRNA genes, providing 100 or 200 bp 16S "pyrotags" to evaluate community composition (57, 59, 120). This approach has the benefits of high resolution (due to the large number of pyrotags; ~500,000 per bulk 454-FLX run) comparable to a 16S microarray, while retaining relative amplicon abundance like a clone library. The main limitation of this approach is the reduced phylogenetic resolution afforded by 100-200 bp, so the method is dependent on a high quality reference 16S database for accurate classification of pyrotags.

Fluorescence *in situ* hybridization (FISH) using group-specific 16S rRNA-targeted oligonucleotide probes (7, 58) also can be used to profile community composition. Fluorescently labeled cells can be quantified by microscopy either manually or with the aid of image analysis software (24), or in combination with flow cytometry (115). In principle, FISH-based counting is the most accurate method for determining relative and absolute abundance of populations since it is not affected by 16S copy number variation. In practice, only a few phylogenetic groups can be targeted per sample due to logistical considerations (e.g. number of fluorochromes that can be visualized simultaneously, availability and cost of suitable probes) and for gross community composition estimates, these tend to target broader groups, such as domains or phyla. Therefore, complete or even widespread population-level characterization of communities using FISH has not been feasible to date.

Shotgun library preparation

Shotgun clone libraries for genome sequencing are typically prepared using three different average sizes of cloned DNA; 3, 8 and 40 kbp (fosmids). This primarily facilitates assembly and finishing since longer clones will have a greater likelihood of spanning gaps in the genome assembly. The JGI uses a ratio of 4:4:1 for 3, 8 and 40 kbp end sequence data to economically produce high quality draft assemblies (largest, correctly assembled contigs). We have more or less adopted the same insert size libraries and sequencing ratios for metagenomic projects, even though the end product may be vastly different from a genomic project. In the case of microbial communities with one or more dominant populations, the ratio of insert size sequencing will serve the same function of improving assembly (and occasionally finishing) of composite population genomes. For microbial communities lacking dominant populations, the main purpose of the larger size inserts is to provide gene neighborhood context, usually through complete sequencing of selected fosmids (36, 138). Bacterial artificial chromosomes (BACs) allow access to even larger pieces of contiguous genomic DNA from environmental samples (11), however they are technically more demanding to prepare than fosmids and small insert libraries.

Occasionally, the environmental sample will dictate which libraries can be created. For example, despite repeated attempts, DNA extracted from acid mine drainage biofilm samples could not be obtained in high enough purity and molecular weight to create an 8 kbp or fosmid clone library limiting the study to data from a 3 kbp library only (132). Preparation of clone libraries requires between 5 μg (for 3 kb library) and 20 μg (for a fosmid library) of DNA which often cannot be obtained directly from low biomass communities. Whole genome amplification via multiple displacement amplification can circumvent this problem, but the average size of the amplified DNA, ~15 kbp, is too short to allow fosmid library construction, although fosmid libraries have been reported from amplified environmental DNA (94).

Sequencing

At the JGI, metagenomic projects are sequenced in at least two stages for quality control (QC). The first stage is a 20 plate QC of a 3 kbp insert (pUC) library generating

365 approximately 10 Mbp of Sanger sequence data followed by a preliminary informatic analysis to guide allocation of the remainder (majority) of the sequence allotment. First and foremost the QC sequencing confirms that the shotgun clone libraries produce sequence data of sufficient quality to warrant further sequencing. For genome projects sufficient quality typically means that 95% of clones produce reads with at least 650 Q20 370 bases (see Sequence read preprocessing), i.e. a 95% pass rate. For metagenomic projects this bar is dropped sometimes to as low as an 85% pass rate because of the greater difficulty in making high quality libraries from environmental DNAs, and often precious nature of difficult to collect environmental samples. The preliminary analysis usually involves assembly but not gene prediction primarily to confirm initial community 375 composition estimates but also to determine if populations can be easily discriminated in the data. For example, similarity searches against public nucleotide and protein databases will identify populations via conserved marker genes and provide some indication of relative abundance according to the size and read depth of the contig that the marker genes were found on. A histogram of contig read depth will alert the researcher to the 380 presence of one or more dominant populations, since 10 Mbp is sufficient to result in assembly of genomic fragments from dominant populations. Plotting contig depth against another variable, such as GC content, often helps to discriminate populations. If a dominant population was expected based on community composition profiling and not noted by contig read depth then this could indicate greater than expected 385 microheterogeneity in the population hindering assembly (see Finishing) or a technical error in the experimental work-up. For example, QC sequencing of enhanced biological phosphorus removing (EBPR) sludge from a lab-scale bioreactor revealed that the primary target organism, *Candidatus* Accumulibacter phosphatis Type I, was grossly under-represented relative to the initial community composition estimate (4% vs 60%). 390 The discrepancy arose because this organism was poorly lysed in the DNA extraction, a fact that was missed because the community was profiled using a Type I specific FISH probe (Shaomei He and Katherine McMahon, personal communication). At this point, it was not too late to re-extract DNA from the EBPR sludge using a different method.

Sequence processing

Processing sequence data from metagenomic and genomic samples share many features in common namely read preprocessing, assembly including selected instances of finishing (dominant populations) and gene prediction and annotation. As mentioned earlier, the key difference between genomes and metagenomes is that the latter, with the exception of finishable dominant populations, do not have a fixed end point, i.e. one or more completed chromosomes as for microbial isolate genomes. This means that metagenomes rarely progress beyond draft assemblies and lack many of the quality assurance procedures associated with producing finished genomes. Therefore, greater care needs to be taken when processing sequences of metagenomic datasets than genomic datasets.

Sequence read preprocessing

Preprocessing of sequence reads prior to assembly, gene prediction and annotation is a critical and largely overlooked aspect of metagenomic analysis. Preprocessing comprises base calling of raw data coming off the sequencing machines, vector screening to remove cloning vector sequence, quality trimming to remove low quality bases (as determined by base calling) and contaminant screening to remove verifiable sequence contaminants. Errors in each of these steps can have greater downstream consequences in metagenomes than genomes and will be discussed in turn.

Basecalling is the procedure of identifying DNA bases from the readout of a sequencing machine. There are surprisingly few choices for basecallers and the differences between them for the purposes of metagenomics are small, therefore we have no specific recommendation from the ones described below. By far the dominant basecaller used today is phred (37). Phred initiated the widespread use of probabilistic-based quality scores, which all later basecallers adopted. Phred quality scores are estimates of per base error probabilities. The quality score q assigned to a base is related to the estimated probability p of erroneously calling the base by the following formula:

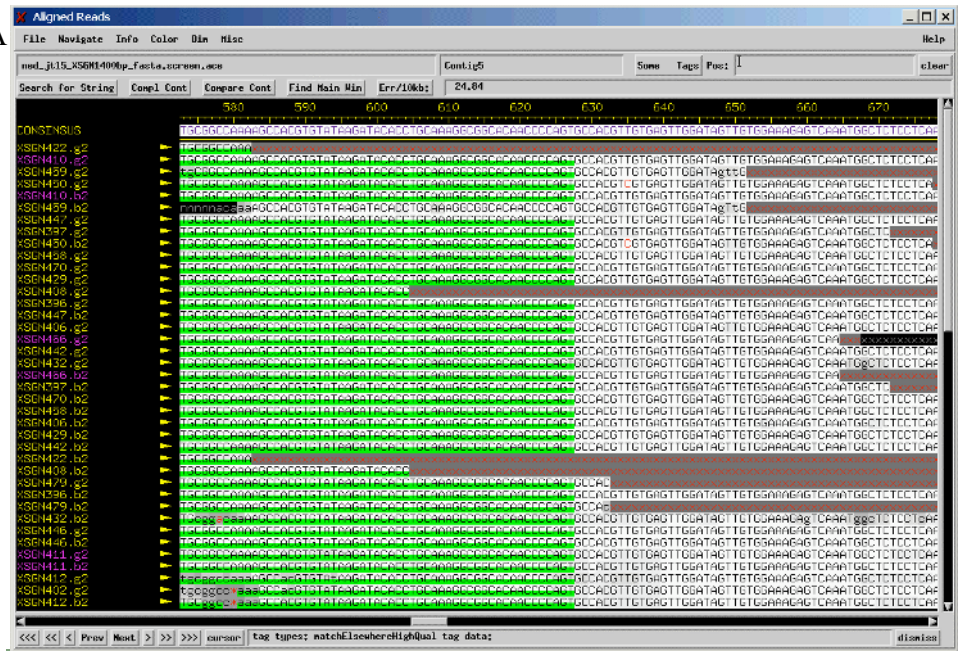
$$q = -10 * \log10(p)$$

Thus, a phred quality score of 20 corresponds to an error probability of 1%. Other frequently used basecallers are Paracel's TraceTuner (www.paracel.com) and ABI's KB (www.appliedbiosystems.com), which behave very similarly to phred converting raw data into accuracy probability base calls. In general however, metagenomic assemblies have lower coverage than genomes and therefore errors are more likely to propagate to the consensus. For complex communities, the majority of reads will not assemble into contigs, and base calling errors in these unassembled reads will appear directly in the final dataset.

Vector screening is the process of removing cloning vector sequences from basecalled sequence reads. Complete and accurate removal of cloning vector sequence is especially important in metagenomic datasets since these datasets often have large regions of very low coverage in which each read uniquely represents a part of a genome. Assembly of these data without vector trimming can produce chimeric contigs in which the vector sequence, being common to most reads, acts to draw together unrelated sequences (**Fig. 3**). Also, genes may be predicted on the vector sequence introducing phantom gene families into downstream analyses (see gene-centric analysis).

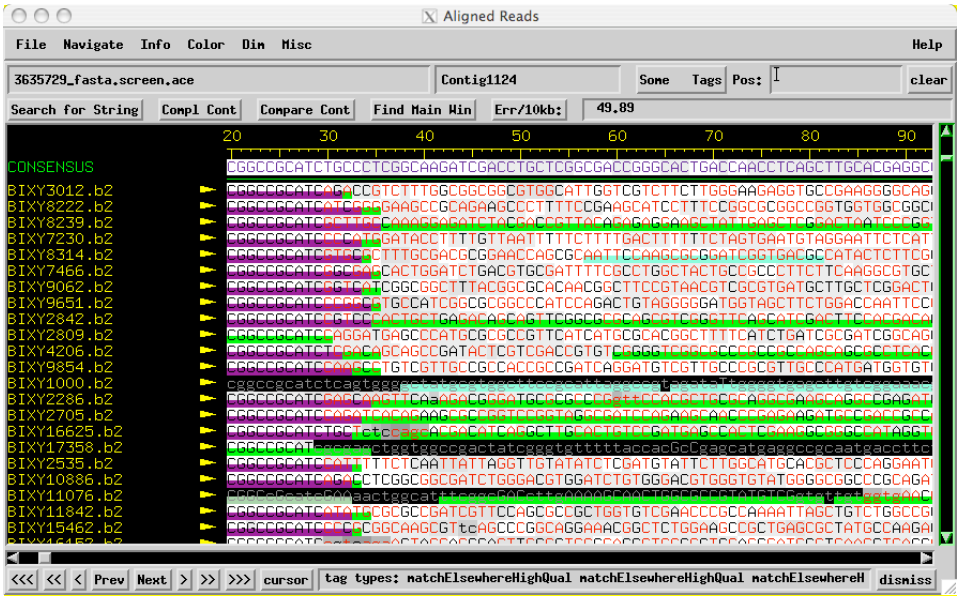
A number of different tools are available for vector screening including cross_match (www.phrap.org), LUCY (18) and vector_clip (122). Also, some assemblers include vector trimming as part of a preprocessing pipeline including PGA (http://www.paracel.com) and Arachne (10, 61). The most commonly used tool is cross_match, which uses a modified Smith-Waterman algorithm to identify matches to vector that are extended to produce optimal alignments. However, cross_match requires exact matches to vector sequences, and has no expectation for the location of vector sequence in a read. In our experience, this program frequently fails to remove vector sequence because of frequent basecalling errors on the edges of reads where vector sequence is found. Another vector trimming tool, LUCY, avoids this problem by specifying error rates as a function of sequence position. In every case that we have tested to date LUCY results are substantially better than those achieved with cross_match. The downstream effects of improved vector screening are fewer spurious protein predictions and fewer errors in prediction of real protein coding sequences, particularly open reading frames at the ends of reads (see gene prediction).

A


Aligned Reads
File Navigate Info Color Dim Misc Help
Search for String
Compl Cont
Compare Cont
Find Main Win
Err/10kb:
CONSENSUS
cursor
dismiss

455

B


Aligned Reads
File Navigate Info Color Dim Misc Help
3635729_fasta.screen.ace
Contig1124
Some
Tags
Pos:
clear
Search for String
Compl Cont
Compare Cont
Find Main Win
Err/10kb:
49.89
CONSENSUS
Prev
Next
cursor
dismiss

C

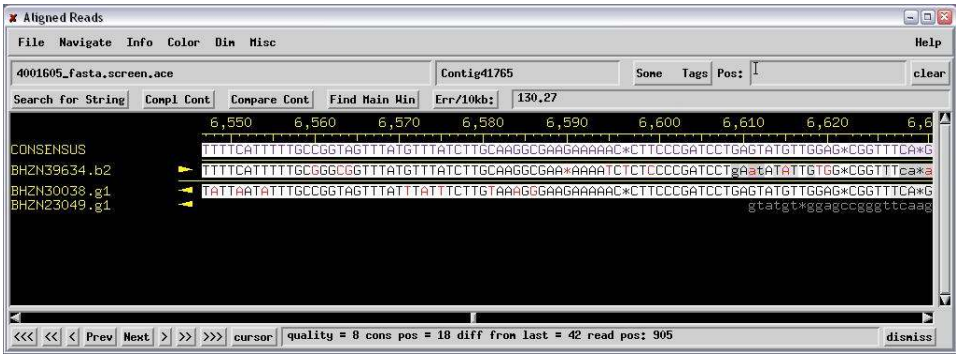

Aligned Reads
File Navigate Info Color Dim Misc Help
4001605_fasta.screen.ace
Contig41765
Some
Tags
Pos:
clear
Search for String
Compl Cont
Compare Cont
Find Main Win
Err/10kb:
130.27
CONSENSUS
BHZN39634.b2
BHZN30038.g1
BHZN23049.g1
Prev
Next
cursor
quality = 8 cons pos = 18 diff from last = 42 read pos: 905
dismiss

Fig. 3. Assembly screenshots from the Consed (49) program. The consensus sequence is shown at the top of the display and is derived from aligned reads shown below the consensus. Read identifiers and orientation (arrow heads) are shown on the left of the display. **A**. An example of a good quality assembly with high read depth. Note the consistent alignment of all residues. **B**. An example of a misassembled contig drawn together by a common repeat sequence (at left). Note the misaligned residues colored in red, and meaningless 'consensus' sequence that does not correspond to any single read below it. **C.** A chimeric contig produced by co-assembly of closely related strains (haplotypes) in a metagenomic dataset. Note that the consensus sequence is a chimera of the two haplotypes, and likely does not represent an extant organism.

Most post-processing pipelines appear to ignore base quality scores associated with reads and contigs and few take positional sequence depth into account as a weighting factor for consensus reliability. Therefore, low quality data will be indistinguishable to the average user from the rest of the dataset and should be removed. An extreme example of a poor quality read that inadvertently passed through to gene prediction is shown in **Fig. 4**. We recommend quality trimming to be performed after vector screening, described above. The reason is that trimming low quality bases might truncate vector sequence and impede the ability of vector-screening programs to recognize the remainder of the vector. In such cases significant parts of vector might still remain for the next stages of the pipeline. LUCY combines vector and quality trimming in one tool.

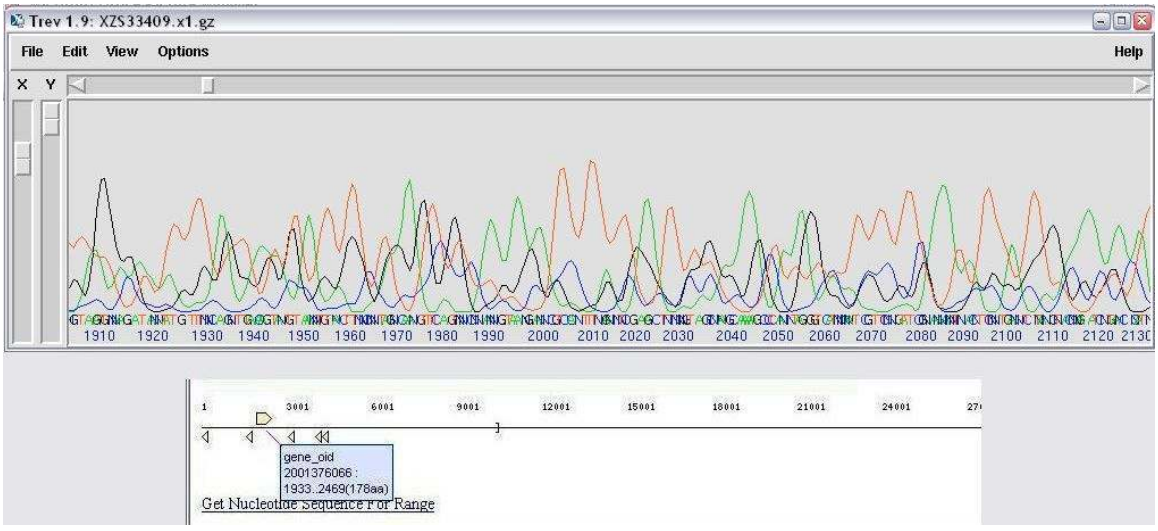


Fig. 4. Part of the chromatogram of a low quality read without quality trimming on which multiple non-existent genes were predicted (bottom panel). Visualized with the TreV program (122).

Recognizing sequence contamination of metagenomic datasets, other than vector sequence, is non-trivial. Sanger datasets from clonal organisms are routinely screened for *E. coli* genomic sequence because *E. coli* is the cloning vector host, and small amounts of its genome may get through plasmid purification. Pyrosequencing which does not rely on cloning DNA into *E. coli*, will not have this problem, however other types of contamination cannot be excluded. For metagenomic datasets, host contamination screening should be considered carefully because the environment under study may have *E. coli* or close relatives as *bona fide* members of the community and screening would therefore bias representation of these species in the dataset. Occasionally mislabeling of sequence plates occurs in production pipelines. These types of cross contamination between two datasets can usually be detected if one of the datasets is from an isolate by differences in GC content or BLAST. If plates from two metagenomic projects are mixed up, the contamination may be harder to detect since neither dataset is likely to be homogeneous. It is quite common that reads and even contigs are not incorporated into finished microbial genomes and these are usually dismissed as either low quality or contaminant sequences. In contrast, metagenomic projects will keep high quality contaminating reads and contigs as they will probably not be easily distinguishable from the rest of the dataset, and may therefore skew downstream analyses such as gene centric analysis depending on the degree of contamination. Presently, there is no solution to this quandary and suspected contaminant sequences would need to be investigated on a case-by-case basis.

Assembly

Assembly is the process of combining sequence reads into contiguous stretches of DNA called contigs, based on sequence similarity between reads. Contigs contain multiple reads linked together by overlapping sequence based on a minimum length of identical bases. The consensus sequence for a contig is either based on the highest quality nucleotide in any given read at each position or based on majority rule, i.e. the most frequently encountered nucleotide at each position. The number of reads underlying each consensus base is called depth or coverage. Sequencing is typically performed from both

sides of an insert in a vector plasmid, and such pairs are called paired reads or mate pairs. Knowledge of the approximate insert size of the library facilitates producing a more accurate assembly since mate pairs provide an external constraint to guide assembly. The presence of paired reads in two different contigs allows those contigs to be linked into larger non-contiguous DNA sequence called a scaffold whose inter-contig gap size can be estimated based on the insert size of the read pairs. For this reason, large insert clones such as fosmids are particularly useful for improving assemblies.

The major cause of misassembly in genomic projects is repetitive regions that can be resolved in the finishing process (74). Assembly of metagenomic projects will also be confounded by repeats, but pose additional assembly challenges in the form of non-uniform read depth due to non-uniform species abundance distribution and the potential for co-assembly of reads originating from different species. Therefore, not only can misassembled reads be retained in the final published dataset due to the absence of finishing, but reads from more than one species can be assembled together producing chimeric contigs. Co-assembly is more likely to happen with reads from closely related genomes where the sequence similarity is higher (we routinely observe homologous regions of two or more strains with up to 4% nucleotide sequence divergence co-assembling) but has been found between reads originating from phylogenetically distant taxa with conserved genes serving as the focal point for misassembly. For example, a contig from a surface seawater metagenome comprised reads originating from bacteria and archaea as evidenced by gene calls, with the 16S rRNA gene serving as the focal point in this instance (28). A recent simulation study found that chimeras are particularly prevalent among contigs sized below 10 kbp (89). High complexity microbial communities lacking dominant populations rarely produced contigs larger than 10 kbp (**Fig. 2**), prompting the recommendation that such datasets should not be assembled at all (89).

A variety of assembly programs are publicly available, including Phrap (www.phrap.org), Arachne (10, 61), the Celera Assembler (92), PGA (http://www.paracel.com/), and cap3 (56). For a description and history of these assemblers we refer the reader to (74). Most currently available assemblers were designed to assemble individual genomes, or in some cases genomes of polyploid eukaryotes,

however they were not designed to assemble metagenomes comprising multiple species with non-uniform sequence coverage, and therefore their performance with metagenomic datasets varies significantly (89). For example, the Celera assembler does not assemble contigs with atypically high read depth (based on an expected Poisson distribution) because it interprets them as potential assembly artifacts due to co-assembly of repeats, whereas in metagenomic data, they may be *bona fide* contigs arising from dominant populations (133). A second example; Phrap is optimized for making maximal use of its input data using a "greedy" algorithm, and will extend contigs as far as possible. This is a good approach for assembling low-coverage non-repetitive regions from low quality reads as it makes the most of the available data, particularly if the assembly will be verified by finishing, but is not desirable for metagenomes since it is more likely to produce chimeras when data includes reads from multiple strains and species. More conservative assembly programs, such as Arachne have been shown to produce smaller but more reliable contigs than Phrap (89).

A useful auxiliary approach to *de novo* assembly is comparative assembly, that is, aligning reads and/or contigs to a reference genome of a closely related organism. The AMOS Comparative assembler has been developed specifically for this purpose (106). For metagenomic datasets, this can improve assembly of dominant populations since it provides a mechanism to span hypervariable regions in a composite population genome and is computationally much less expensive than *de novo* assembly (3). A major caveat of the approach however is that it will only be useful for a small subset of the average metagenomic dataset since reference genomes cover only a fraction, and a highly biased fraction at that, of microbial diversity (see Post-sequencing community composition estimates).

One thing is clear, there is no magic bullet for assembling metagenomic datasets and all assemblers will make numerous errors. Ideally, therefore, every metagenomic assembly should be manually inspected for errors before public release. Assembly errors can be easily identified with visualization tools such as Consed (**Fig. 3**) (49) which are used to facilitate genome finishing, however the sheer scale of most metagenomic datasets precludes manual inspection let alone correction of all identified assembly errors. One approach we have taken to address this limitation is to make two or more assemblies

of the same data using different assemblers (43) to facilitate identification of misassemblies during the downstream analysis phase following gene calling. It is however, feasible and worthwhile to resolve misassemblies of the largest contigs in a metagenomic assembly, especially contigs greater or equal in length to fosmids, using standard initial steps in the finishing process (74).

The final products of assembly, contigs and scaffolds, are submitted to public databases as flat text files, meaning that all information about the underlying reads is lost including sequencing depth and quality scores of each base, length and overlaps between reads, and quality of vector trimming. This is not ideal for two reasons. Firstly, quality of the contigs cannot be assessed and is also not taken into consideration by tools such as BLAST. Secondly, meaningful polymorphisms in the data due to co-assembled strains (haplotypes, see Analyzing dominant populations) are lost because a single consensus sequence is submitted. Methods for weighting consensus accuracy and preserving polymorphism information for subsequent analyses are needed. A first step in this direction has been taken by the public databases with the establishment of the Trace and Assembly Archives which archive raw read files and assemblies associated with submitted genomic and metagenomic datasets respectively (139). In practice however, most users will only work with the flat text consensus data and ignore read and consensus quality unless it is presented to them in a more convenient user interface. Such interfaces are beginning to be provided by dedicated comparative genome and metagenome platforms (see Data Analysis and Management).

Finishing

Genome closure and finishing is commonplace for microbial isolate projects, and part of the standard processing pipeline at sequence facilities such as JGI. For most metagenomes, finishing is not possible. However, for dominant populations within metagenome datasets that have draft level coverage, finishing may be an option. This is largely dependent on the degree of microheterogeneity within the population. Genome rearrangements such as insertions, deletions and inversions, will break assemblies, whereas point mutations usually will not. Even in instances where chromosomal walking along large insert clones is used instead of shotgun sequencing, microheterogeneity can

still complicate assembly (52). However, there are now several examples in the literature of complete or near-complete composite population genomes of uncultivated organisms derived from environmental sources including *Cenarchaeaum symbiosum*, the sole archaeal symbiont of a marine sponge (52), *Kuenenia stuttgartiensis*, an anaerobic ammonia-oxidizing planctomycete sequenced from a lab-scale bioreactor sludge (123), a Rice Cluster 1 methanogen from an enrichment culture (36), *Candidatus* Cloacamonas acidaminovorans, the first sequenced representative of candidate phylum WWE1, from an anaerobic digestor (102) and *Ferroplasma acidarmanus*, one of a handful of dominant populations in an acid mine drainage biofilm (4). In the last case, the assembly was facilitated by the availability of an isolate genome (fer1) obtained from the same habitat. The *Kuenenia*, Rice Cluster 1 methanogen and *Candidatus* Cloacamonas genomes, however, could be assembled without reference to an isolate genome because the populations were near clonal. We make the general observation that sequence microheterogeneity within populations often seems to reflect spatial heterogeneity within the ecosystem from which the populations were derived. Homogenized systems, such as bioreactors or enrichment cultures, have produced composite population genomes with very low levels of polymorphism (36, 102, 123) perhaps due to the higher likelihood of selective sweeps through the population curtailing genomic divergence (20). Therefore, if the goal is to assemble a complete population genome from an environmental sample, we recommend use of ecosystems with low spatial heterogeneity if at all possible, or by finer-scale sampling to reduce the effect of spatial heterogeneity.

Gene prediction and annotation

Gene prediction (or gene calling) is the procedure of identifying protein and RNA sequences coded on the sample DNA. Depending on the applicability and success of the assembly, gene prediction can be done on post-assembly contigs, on reads from the unassembled metagenome, and finally for a mixture of contigs and individual unassembled reads.

There are two main approaches for gene prediction. The ‘evidence-based’ gene calling methods use homology searches to identify genes similar to those observed previously. Simple BLAST comparisons against protein databases, as well as tools like

Critica (9) and Orpheus (42) use such an approach. Conversely, the second approach, '*ab initio*' gene calling, relies on intrinsic features of the DNA sequence to discriminate between coding and non-coding regions allowing the identification of genes without homologs in the available databases. The use of gene training sets, i.e. sets of parameters derived from known genes of the same or related organisms can enhance the quality of the predicted genes for some of those programs (e.g. fgenesB (http://www.softberry.com)), while others are self trained on the target sequence (Genemark (12), GLIMMER (27), metagene (95)).

Pipelines that use a combination of evidence-based and '*ab initio'* gene calling are frequently used for complete genomes. In the first step, genes are identified based on homology searches of sequence of interest versus public databases. Hits to genes in databases are considered to be real genes, and can be used as a training set for the *ab initio* gene calling programs. Subsequently an '*ab initio'* method fine-tuned for a particular genome is used to identify more genes that were missed in the previous step. One such pipeline called mORFind uses a combination of Orpheus, Critica and Glimmer.

Table 1. Gene prediction methods used in metagenomic projects.

Project	Institution	Reference	Gene prediction method
Acid Mine Drainage biofilm communities from Richmond mine	Univ of California, Berkeley Joint Genome Institute	(132)	Fgenesb
Aquatic microbial communities from Drinking-water networks	Univ of Goettingen	(114)	Blast
Aquatic microbial communities from Soudan Mine in Minnesota	SDSU	(32)	Blast
Fossil microbial community from Whale Fall at Santa Cruz Basin of the Pacific Ocean	Joint Genome Institute	(129)	Fgenesb
Gut microbiome of Human healthy adults	J. Craig Venter Institute Washington Univ Stanford univ	(46)	Blast
Gut microbiomen of Human healthy Japanese infants and adults	Univ of Tokyo	(72)	Metagene
Gut microbiome of Mouse lean and obese	Washington Univ	(130)	Blast
Gut viriome of Human healthy adults	Genome Institute of Singapore	(144)	Blast
Marine archaeal anaerobic methane oxidation (AOM) communities from Eel River sediments	Joint Genome Institute MBARI	(53)	fgenesB
Marine microbial communities from Bras del Port saltern in Santa Pola Spain crystallizer pond	Univ Miguel Hernandez	(75)	Glimmer
Marine microbial communities from Global Ocean Sampling (GOS)	J. Craig Venter Institute	(109)	Similarity searches and filtering of ORFs
Marine microbial communities from Sargasso Sea	J. Craig Venter Institute	(133)	Blast
Marine Plankton communities from deep Mediterranean sea Ionian station Km3	Univ Miguel Hernandez	(88)	Blast
Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)	Joint Genome Institute	(29)	Blast
Marine RNA viral communities from coastal samples	http://www.sfu.ca/ Univ of British Columbia	(23)	Blast
Marine viral communities from ocean environments	SDSU	(8)	Blast
Olavius algarvensis (gutless worm) microbiome from Mediterranean sea	Max Planck Institute Joint Genome Institute	(142)	mORFind
Oral TM7 microbial communities of Human healthy adults	Joint Genome Institute Stanford Univ	(83)	fgenesb
Soil microbial communities from Minnesota Farm	Joint Genome Institute	(128)	fgenesb
Wastewater EBPR microbial communities from Bioreactor	Joint Genome Institute	(43)	fgenesb

655

In metagenomic sequences, genes can originate from many, frequently diverse organisms. When dominant populations exist, their sequences can be separated from the rest of the dataset (see binning) and the pipeline generally used for complete genomes applied to this subset of the data. For communities or their parts that defy assembly or assemble poorly, no training is possible. In these cases "generic" gene prediction models can be used, or models fine-tuned to the closest phylogenetic group. For example, MetaGene (95), a gene prediction program developed specifically for metagenomic datasets, using two generic models, one for archaea and one for bacteria. Due to the fragmented nature of such datasets and the quality of the sequencing, gene prediction is further complicated by the fact that many genes are represented only by fragments, contain frameshifts or are chimeras due to errors in the assembly. Recently a tool that allows gene prediction despite these problems, even on short 454 reads, has been published (69) although its performance has yet to be evaluated in real applications. The method is based on similarity comparisons of the metagenomic nucleotide sequences either to the same metagenome or to other external sequences and subsequent discrimination of conserved coding sequences from conserved non-coding sequences by synonymous substitution rates. BLAST searches are conducted on the amino acid level to provide higher resolution than nucleotide searches.

Both evidence based and '*ab initio*' methods have been used for the prediction and analysis of metagenomic datasets (Table 1). Evidence-based gene calling has been used as the sole method of gene calling in at least one metagenomic study using Sanger reads (133) and all metagenomic studies using unassembled pyrosequence data due to the short read lengths (Table 1). Since this approach relies entirely on comparisons to existing databases it has two major drawbacks. Low similarity values to known sequences either due to evolutionary distance or due to the short length of metagenomic coding sequences and the presence of sequence errors prevent the identification of homologs. Moreover, novel genes without similarities are completely ignored. Despite these drawbacks this approach has been used in several studies, and can be useful for gene-centric comparisons of metagenomes, especially in cases where the size of the sequence fragments is not adequate for the *ab initio* gene prediction, such as high complexity metagenomes and metagenomes sequenced by high throughput parallel pyrosequencing.

Treating all open reading frames (ORFs) as putative genes usually produces prohibitive amounts of data, contains too much noise and therefore is very hard to use. Methods, based on features of the sequences, the size of the predicted ORFs, and the similarity to known sequences, have been used to lower the total number of candidate coding sequences from a population of ORFs (143).

At the JGI we are using two '*ab initio*' gene prediction pipelines for the analysis of metagenomic datasets. The first uses fgenesB with specific training models for sequences that can be assigned to phylogenetic groups and generic models for the unassigned sequences (Table 1). The second uses Genemark, which allows gene prediction without the need for training sets and classification of sequences. Both pipelines have proved to be quite accurate when used on simulated datasets (http://fames.jgi-psf.org). Other studies have employed Glimmer, metagene and the mORFind pipeline (Table 1).

RNA genes (tRNA, rRNA) are predicted using tools such as tRNAscan (77) for tRNAs, and similarity searches for ribosomal RNAs. Other types of non coding RNA genes (ncRNA) can be detected by comparison to covariance models (51) and sequence-structure motifs (79). However, searching covariance models and motifs is computationally expensive and it is prohibitively long for large metagenomic datasets. Overall the identification of other ncRNA genes is difficult, since their sequence is not conserved and reliable '*ab initio*' methods are lacking even for isolate genomes.

There are several types of errors that can be made by a gene-calling pipeline. A gene can be missed completely, or called on the wrong strand. A less severe mistake would call part of the gene correctly, but fail in estimating gene boundaries or call genes that are partly correct and partly wrong, due to chimeric assemblies or frameshifts (89). The quality of the gene prediction relies on the quality of read preprocessing and assembly. Gene calling methods used on accurately assembled sequences predict correctly more than 90% of the genes that are included in the dataset, as evidenced from studies on simulated datasets (http://fames.jgi-psf.org). This high number was achieved with training on generic models or self-trained algorithms. Gene prediction on unassembled reads exhibits lower accuracy than on contigs (~70% vs >80% respectively, (89), a result attributed to the small size and higher chance of sequencing errors for individual reads.

Often, even in low complexity communities, a large number of reads, belonging to less abundant organisms, remain unassembled. Although the genes predicted on the assembled sequences allow the metabolic reconstruction of the abundant organisms, a better representation of the metabolic capacity of the community is gained when genes from both contigs and reads are included in the subsequent analyses as a majority of the functionality may in fact be encoded in the unassembled reads (89). Therefore, it is advisable to perform gene calling both on reads and contigs. For high complexity communities, where assembly is minimal, gene calling on unassembled reads is the only possibility.

Gene prediction is usually followed by functional annotation. Functional annotation of metagenomic datasets is very similar to genomic annotation and relies on comparisons of predicted genes to existing, previously annotated sequences. The goal is to propagate accurate annotations to correctly identified othologs. However, there are additional complications in metagenomic data where predicted proteins are often fragmented and lack neighborhood context. Annotation of metagenomic data created by short-read methods, such as 454, is even more complicated since most reads contain only fractions of proteins.

At the JGI we use profile-to-sequence searches to identify functions. Protein sequences are compared to profiles from TIGRFAM (116), PFAM (39) and COGs (125), using RPS-BLAST (86). PFAMs allow the identification and annotation of protein domains. TIGRfams include models both for domain and full length proteins. COGs also allow annotation of the full length proteins. Unfortunately, although PFAMs and TIGRfams are updated regularly allowing annotation of new protein families COGs are still lacking such updates. As a rule, assignment of protein function solely based on Blast results should be avoided, mainly because of the potential for error propagation through databases (45, 71, 73).

In addition to annotation by homology, several methods are available for context annotation. These include genomic neighborhood (26, 98), gene fusion (34, 81), phylogenetic profiles (101) and co-expression (82). We are aware of one study that performed adapted neighborhood analysis on metagenomic data, which combined with homology searches inferred specific functions for 76% of the metagenomic datasets (83%

when nonspecific functions are considered) (55). It is possible that more context information will be used to predict protein function in metagenomic data in the future.

It is common practice that all gene predictions and annotations for microbial genomes are manually checked as part of informatic quality control pipelines. Such manual curation is not feasible for metagenomic projects, although, as for the assembly, we recommend manual curation of larger contigs. Therefore, the quality of gene calling and annotation for the majority of metagenomic data rests solely on automated procedures. A recent benchmarking study using simulated metagenomic datasets suggests that there is significant room for improvement in existing gene prediction and annotation tools (89). One final note of caution; some vector screening and trimming programs only mask out rather than remove vector and low quality sequences, resulting in runs of Ns at the ends of reads and contigs. When sequences are submitted to the public databases, terminal runs of Ns are removed as part of the submission process which can introduce systematic errors in the start-stop coordinates of any genes predicted on the untrimmed reads and contigs. Therefore all reads and contigs should be trimmed of terminal N runs prior to gene prediction and annotation.

Data analysis

Gene prediction and annotation completes the list of procedures that are routinely applied to both genomic and metagenomic data. While there is still great room for improvement in applying a number of these steps to metagenomic data, they constitute part of the standard data processing pipeline at sequencing centers such as the JGI. Beyond this point, the data analysis methods apply specifically to metagenomes.

Post-sequencing community composition estimates

One of the first analyses that can be performed on metagenomic data following standard processing steps is a re-evaluation of the community composition estimate, this time directly from the metagenomic data itself. This is important for interpretation of the data since biases in the initial estimates, such as PCR skewing (124, 136), are different from biases introduced during metagenomic data generation (described below). Mapping

conserved phylogenetically-informative marker genes, such as 16S and 23S rRNA (ribosomal RNAs), recA (DNA repair protein), EF-Tu, EF-G (elongation factors), HSP70 (heat shock protein) and rpoB (RNA polymerase subunit), onto their reference trees has been used to assess both organism identity and relative abundance (133). Single-copy, mostly ribosomal, genes have been applied for the same purpose (19, 43, 134). Ubiquitous single copy genes have the advantage of being present once in all microbial genomes and therefore are thought to provide more accurate estimates of community composition than markers such as 16S rRNA with variable copy number (134).

Marker gene analyses are performed as follows. An alignment of each gene is prepared from a reference dataset, usually from all available complete genomes. The marker genes are identified in the metagenomic dataset of interest, and included in the reference alignment. For quantification of populations, the depth of contigs containing the marker genes should be taken into account (129, 135). Trees are calculated, and the relative positions of metagenomic genes are identified in the tree. There are several limitations to community composition estimates based on phylogenetic inference of single copy genes identified in metagenomic datasets:

i) the reference genome database is currently incomplete and highly biased towards just three bacterial phyla (Proteobacteria, Firmicutes and Actinobacteria) out of at least 50 phyla (58). This means that accurate placement of metagenomic genes is compromised if they originate from organisms not belonging to the three well-represented phyla, with the exception of the 16S rRNA gene which is broadly used to define taxonomic groups (30). Initiatives to improve genome sequence representation of the tree of life should help to rectify this problem, such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA) pilot project at the JGI (http://www.jgi.doe.gov/programs/GEBA/). Even so, the majority of microbial lineages still lack cultured representatives (58) complicating our ability to obtain representative genome sequences.

ii) genes derived from metagenomic datasets, particularly those with minimal assembly, are often fragmented and produce incomplete alignments. Indeed it is often the case that metagenomic gene fragments from the same protein family are entirely non-overlapping. This precludes the use of evolutionary distance methods as infinite distances are created in the pairwise distance matrix severely compromising the resulting tree (14).

Discrete character inference methods, particularly maximum likelihood, can tolerate incomplete alignments to a certain extent. Alternative approaches to address the problem include making separate trees for each metagenomic gene only in the context of the reference dataset, subdividing the alignment into smaller parts to produce more complete subalignments that can still contain multiple metagenome-derived genes, or inserting partial sequences into a reference tree of full-length sequences using for example probabilistic maximum likelihood placement (135) or the ARB parsimony insertion tool (78).

iii) erroneous gene calls, in particular ribosomal proteins are sometimes missed by automatic gene callers because of their small size (89).

iv) finally, and perhaps most importantly, conserved phylogenetically-informative genes represent only a small fraction of the total metagenomic dataset. For example, 100 Mbp of Sanger sequence will typically yield about a dozen mostly partial length sequences of any given marker gene. In addition, it has recently come to light that single copy genes are particularly prone to under-representation in shotgun libraries due to their toxicity to the *E.coli* host (121). Furthermore, since the toxicity is due to expression of the introduced gene, it varies between organisms depending on the ability of *E.coli* to transcribe and translate the introduced gene (121). Therefore low numbers of incompletely overlapping marker sequences, together with the toxicity effect compromise the ability to reliably infer community composition from single copy genes.

Sequence similarity tools such as BLAST (6) can be used to identify homologs in reference sequences (60). Such an analysis results in a much higher fraction of the dataset being involved in the composition estimate, but suffers from other effects. Potentially, bigger genomes are expected to generate more matches than smaller genomes (119), and therefore the assessment is of gene rather than organism abundance. The closest BLAST hit is not necessarily the nearest phylogenetic neighbor (68), and therefore classifying by BLAST hits can be misleading particularly if only distantly related homologs are available in the reference database. Additionally, the potential for horizontal gene transfer between sympatric populations can cause the recipient organism to be identified as the donor organism. Presently the biggest problem for BLAST-based composition estimation is the poor representation of microbial diversity by sequenced isolates (58, 62) often

840 resulting in remote matches to phylogenetically distant organisms or absence of any hits. In our experience, BLAST-based methods overestimate the abundance of highly covered taxa such as Proteobacteria and Firmicutes, especially if only the top hit is taken into consideration. One recent implementation of BLAST-based community composition profiling, MEGAN (60), addresses this problem by assigning sequence fragments to the 845 lowest common ancestor of the set of taxa that it hit in the comparison, thereby reducing false matches. Unfortunately this often results in the bulk of a dataset being assigned to very high level groupings, such as Bacteria, or being unclassified altogether. Again the problem lies with the reference genome database rather than the tool, and can be expected to improve as the bias in the database is addressed.

850 Finally, given that fundamental upstream processes such as DNA extraction can produce an equal or greater skewing of community representation as any bioinformatic analysis, researchers should if possible calibrate their data against the original intact community using methods such as 16S rRNA-targeted fluorescence *in situ* hybridization.

Binning

855 A metagenomic sequence pipeline produces a collection of reads, contigs and genes. Associating these data with the organisms from which they were derived is highly desirable for interpretation of the ecosystem. This process of association between sequence data and contributing species (or higher level taxonomic groups) is called binning or classification. The most reliable binning is assembly, that is, in a good 860 assembly all reads in a contig are derived from the same species with the optimal binning being a closed chromosome. As described above, this is often not the case and some level of co-assembly is usually encountered in metagenomic datasets, particularly between strains (see Assembly). However, binning methods rarely have the resolution to discriminate between strains of the same species, so strain co-assembly is not a practical 865 concern when it comes to binning. In fact much coarser level assignment of sequences can be useful for interpreting microbial communities, such as the classification of fragments from a termite hindgut analysis into two dominant class-level groups, the treponeme spirochetes and fibrobacter-like bacteria, each group comprising numerous but functionally related species (138). In this regard, less stringent “extreme” assemblies

(109), which certainly produce chimeric and misassembled contigs, may be a useful binning approach.

In many ways binning and community composition estimates share a common goal; classification of sequence data into taxonomic groups, and so there is overlap in the methods to achieve this goal. Phylogenetic marker genes can be used to bin sequence fragments but this approach suffers from the same problems as for community profiling, namely an incomplete and biased reference database, difficulties with tree building and low overall incidence of marker genes (~1%) in the metagenomic dataset. Similarly, sequence comparison and visualization tools such as BLAST and MEGAN (60) can also be used to bin a larger cross section of sequence fragments to phylogenetic groups, with the associated problems described above.

An entirely different binning approach is based on genome sequence composition. Cellular processes such as codon usage, restriction-modification systems and DNA repair mechanisms produce sequence composition signatures, primarily oligonucleotide (word) frequencies, that are distinct in different genomes (31, 65, 66). This property of genomes has been exploited by a variety of methods to identify groups of sequences with similar composition features and to determine their phylogenetic origin (2, 25, 93, 111, 126) which can not only be used to bin metagenomic data, but also to identify atypical regions within genomes, such as laterally transferred genes. The words can be of any length – usually from 1 (GC content) to 4 and usually no higher than 8. Typically longer words give better resolution but also require longer sequences and are more computationally expensive, with the best results provided by words between 3 and 6 nucleotides long.

Composition-based methods can be divided into supervised and unsupervised (clustering) procedures. Unsupervised procedures cluster metagenomic fragments in composition signature space without the need to train models on reference sequences, and include Self-Organizing Maps (1) and the program TETRA (126). An advantage of unsupervised classification is that phylogenetically novel populations lacking close or even distantly related sequenced taxa can potentially be binned by shared sequence composition features, although identification of the clustered fragments still relies on sequence similarity to reference organisms. Such populations, even when well represented in metagenomes, cannot be binned directly by homology-based methods. A

drawback of unsupervised methods is that they tend to focus on major classes in a dataset and will not perform well on low abundance populations. Supervised methods classify metagenomic fragments against models trained on classified reference sequences and in principle can assign fragments from low abundance populations if there is a model learned from reference data. Examples of supervised approaches include Bayesian classifiers (25) and the support vector machine-based phylogenetic classifier Phylopythia (90). As they are able to learn the relevant features that distinguish a particular population from others using the labeled reference sequences, supervised methods usually achieve higher classification accuracy (sensitivity and specificity) than unsupervised methods, and therefore are preferable if training data are available. Further details on the underlying principles and relevant merits of different binning methods can be found in a recent opinion article on metagenomic binning (91).

At the JGI, we have had most experience with the supervised classifier, Phylopythia (90). This program uses generic or sample-specific models, the former usually derived from reference genomes and the latter usually derived from the metagenomic dataset itself. Perhaps not surprisingly, sample-specific models based on training data from the metagenome under study produced the most specific and sensitive binning of the available approaches as determined by simulated datasets (89) or subsequent assembly of the targeted population (90), often increasing the amount of classified sample data by an order of magnitude over the training set. Ideally, at least 100 kbp of training data is required to make a sample-specific model (91). For dominant populations this amount of target population data can often be found using a single phylogenetic marker gene identified on a large contig that can be extended to other contigs by mate pair information. For low abundance populations, identifying 100 kbp of training data may not be possible based on marker genes, particularly if the population is not closely related to sequenced reference genomes. However, higher-level taxonomic models may still be feasible in which multiple species contribute to the training set. This approach was used successfully for sample-specific binning of treponeme spirochete species that were collectively the dominant group in a termite hindgut symbiont community (138).

Finally, sequence length is a critical parameter for all composition-based classifiers, with no method convincingly classifying sequences less than 1 kb long due to the limited

number of words that are contained in short sequences (91). This precludes the classification of individual Sanger and pyrosequence reads meaning that largely or completely unassembled complex communities cannot be binned at all by composition-based methods.

Analyzing dominant populations

In several aspects, the analysis of low-complexity communities resembles the analysis of isolate genomes. As with isolate genomes, draft-level composite genomes of dominant populations have sufficient coverage and gene context to allow a reasonably comprehensive metabolic reconstruction in which most major pathways can be elucidated. If more than one dominant population is sequenced then potential metabolic interplay of those populations may also become apparent. For example, a metagenomic study of an acid mine drainage biofilm revealed that while all dominant bacterial and archaeal populations were potentially capable of iron oxidation (the main energy generating reaction in this habitat), only *Leptospirillum* group III had genes for nitrogen fixation, suggesting a keystone function for this species since the habitat is limited in externally derived fixed nitrogen (132). Similarly, a metabolic reconstruction of the dominant bacterial symbiont populations in a gutless worm suggested a model for how these organisms together satisfy the nutritional requirements of their host (142). As with draft genomes of isolates, caution needs to be exercised in inferring the absence of metabolic traits since the relevant genes may be present in sequencing gaps, particularly if the trait is encoded by only one or two genes. For example, respiratory nitrate reductase necessary for denitrification was not found in the draft composite population genome of *Candidatus* Accumulibacter phosphatis Type II despite circumstantial experimental evidence suggesting that this organism is capable of denitrification (43).

The major difference between isolate genomes and composite dominant population genomes is that the latter are usually not clonal due to genetic variation inherent in natural populations (140). Genomic differences between individuals and strains within a population can take the form of single nucleotide polymorphisms (SNPs) and rearrangements (insertions, deletions, inversions, transitions, duplications). Co-assembly of genetically distinct strains (haplotypes) will produce high quality discrepancies (SNPs)

in the consensus, that finishing would normally try to resolve. However, in metagenomic datasets, SNPs can be mined in a number of ways to provide insights into population structure and evolution. For example, total SNP frequency provides a quantitative estimate of the degree of genetic variation within a species population which has been found to range from virtually clonal in enrichment cultures (123) and activated sludges (102, 123)to highly polymorphic in acid mine drainage archaeal populations (132). The ratio of non-synonymous to synonymous SNPs in protein-coding genes within a population provides an estimate of the fraction of genes under selective pressure. Furthermore, the ratio of haplotypes for individual SNPs (site-frequency spectra) can be used to estimate important parameters in population genetics, such as the scaled mutation rate and scaled exponential growth rate (64). SNPs also highlight junctions of homologous recombination between strains allowing the degree of sexuality within a population to be estimated (140). In all cases, the clear advantage of using environmental shotgun sequence data for these analyses over isolate sequence data is a broader and less biased sampling of genetic variation within a population (3, 140).

A complication associated with interpreting these data is sequencing error. Setting base quality thresholds too low will introduce noise into the analysis, while setting it too high will discard potentially useful information. The latter may be an important consideration when read depth is low. A conservative approach to avoid mistaking errors as polymorphisms is to only score SNPs with haplotypes represented by at least two different reads requiring a minimum read depth of four. A second complication is the inability to easily distinguish between orthologous from paralogous regions. Unless repeats occur on the same (manually verified) contig or scaffold, such as in the case of a neighboring gene duplication, it is difficult to distinguish repeats from orthologous regions in different organisms. This problem is alleviated if the composite population is finished.

Several tools are available for visualization and analysis of polymorphisms in composite population assemblies. Consed developed to assist in the finishing process, is a generically useful graphical tool for viewing assemblies at the nucleotide level (49). A note of caution however, Consed sometimes masks stretches of nucleotide sequence with Xs, and when SNP analysis is performed it identifies these X characters as SNPs.

Therefore manual post-processing is required for Consed results. SNP-VISTA (118) is an adaptation of the comparative genomics tool VISTA (40) developed specifically to visualize SNPs in alignments. Input for this program is BLASTn output for user-friendliness. Reads are ordered by haplotype using a clustering algorithm calculated for sliding windows. Putative recombination sites are detected by sudden changes in cluster composition between adjacent windows (**Fig. 5**). Strainer is also dedicated software for the analysis of genetic variation in populations (35). As the name suggests, it facilitates the reconstruction of individual strains from co-assembled sequences, clusters reads by haplotype from which it predicts gene and protein variants, identifies conserved regulatory sequences and quantifies and displays homologous recombination sites along contigs.

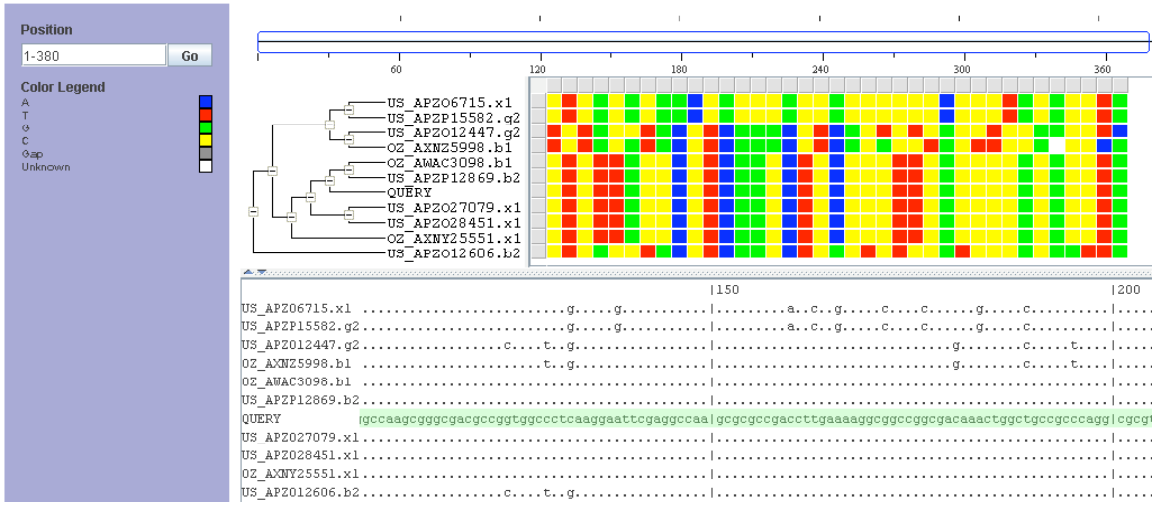


Fig. 5. Screenshot of SNP-VISTA, showing SNPs in individual reads relative (and aligned) to a reference contig belonging to *Candidatus* Accumulibacter phosphatis (50) (labeled query in the lower panel and highlighted in pale green). The upper panel shows the alignment condensed to show only polymorphic columns color-coded by base (see left panel for color-coding), while the expanded alignment is presented in the lower panel. Note that reads are ordered dynamically by similarity for the window under investigation to facilitate SNP pattern recognition.

As for fine-scale genetic variation, methods for visualizing and analyzing gross within-population variation caused by rearrangements are beginning to emerge. For example, recruitment plots display alignments of environmental reads against a reference

sequence such as an isolate genome with one axis showing read location along the reference and the other axis showing sequence identity to the reference. The depth of alignment at each point is a measure of frequency of occurrence of the particular genomic region. Genomic regions that are present in all members of the species will be covered by multiple reads, while strain-specific regions will have shallow or no coverage (**Fig. 6**) effectively highlighting hypervariable regions in a population. A number of important biological insights have been made using this type of analysis including the discovery of genomic islands encoding ecologically-important genes (22) and that phage-defence mechanisms, notably CRISPRs are amongst the fastest evolving elements in the genome (131).

Recruitment plots can be enhanced by displaying data from multiple metagenomes against a reference sequence distinguished by color-coding. This is particularly effective for spatial series where differences between allopatric populations can be highlighted and correlated with metadata (109). Rearrangements such as inversions or indels can be specifically visualized using a variant of recruitment plotting. Instead of plotting all reads, only reads with inconsistently distanced end pairs are shown which draws attention to rearrangements (109). Similarly, individual reads that do not map 1:1 onto the reference genome can be plotted to highlight inversion, insertion or deletion boundaries. As has been discussed in the context of several other analyses, recruitment plots can be limited by the availability of reference genomes unless reference sequences are forthcoming from the metagenomic dataset itself.

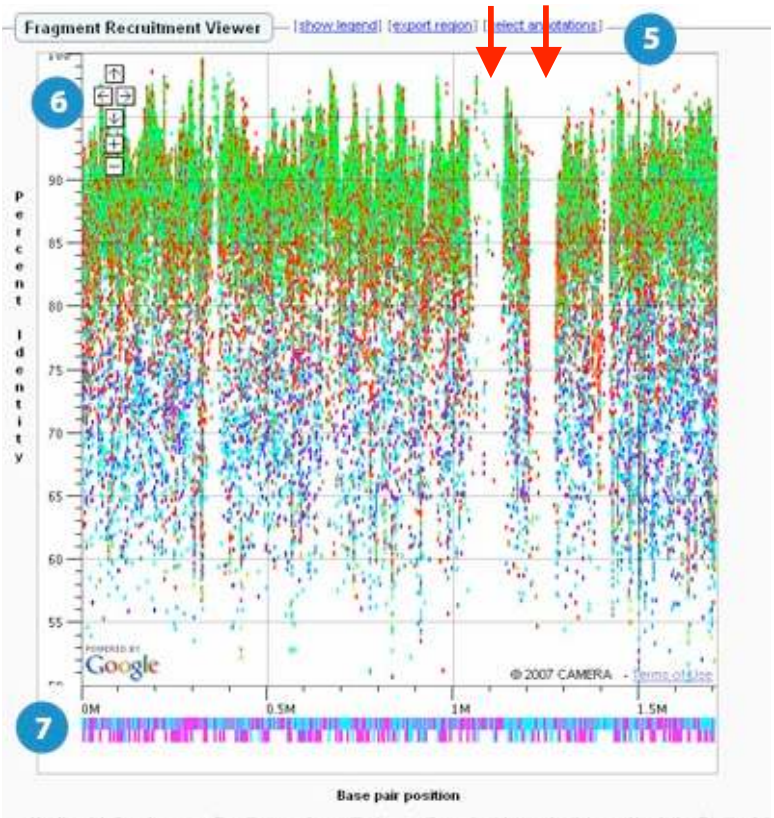


Fig. 6. Screenshot of the CAMERA fragment recruitment viewer taken from the explanatory notes (http://camera.calit2.net/about-camera/frv_help.php). A reference contig or genome, in this case *Prochlorococcus marinus* str. MIT 9312 is shown on the X-axis against which metagenomic reads are aligned and arrayed by similarity to the reference sequence on the Y-axis. Reads have been color-coded according to sampling site to highlight site-to-site variations in *Prochlorococcus* populations. Genomic islands peculiar to strain MIT 9312 are easily identified as gaps in the read coverage (arrows). This viewer also allows users to zoom into regions of interest for higher resolution.

Gene-centric analysis

Metagenomic sequencing of high-complexity microbial communities result in little or no assembly of reads (129), which precludes the use of microheterogeneity analyses described above for dominant populations. The high coding density of bacterial and archaeal genomes and average gene size does, however, mean that most reads will capture coding sequence. This allows a gene-centric analysis of the data that treats the community as an aggregate largely ignoring the contribution of individual species. Genes and gene fragments from a given metagenomic dataset are mapped to gene families providing an estimate of relative representation (**Fig. 7**). The power of the method lies in comparing relative gene family or subsystem abundances between metagenomes to

highlight functional differences. Since determining relative gene family frequencies within and between metagenomic datasets is a key aspect of the method, it is important that the frequencies are not masked by assembly. Either the analysis should be conducted on unassembled reads, or read depth of contigs should be taken into account (89). The approach was first described by Tringe et al. (128, 129) in which they coined the term environmental gene tags (EGTs) because of the fragmentary nature of the data akin to expressed sequence tags (ESTs). Other groups published similar but distinct approaches in quick succession (46, 107) (29).

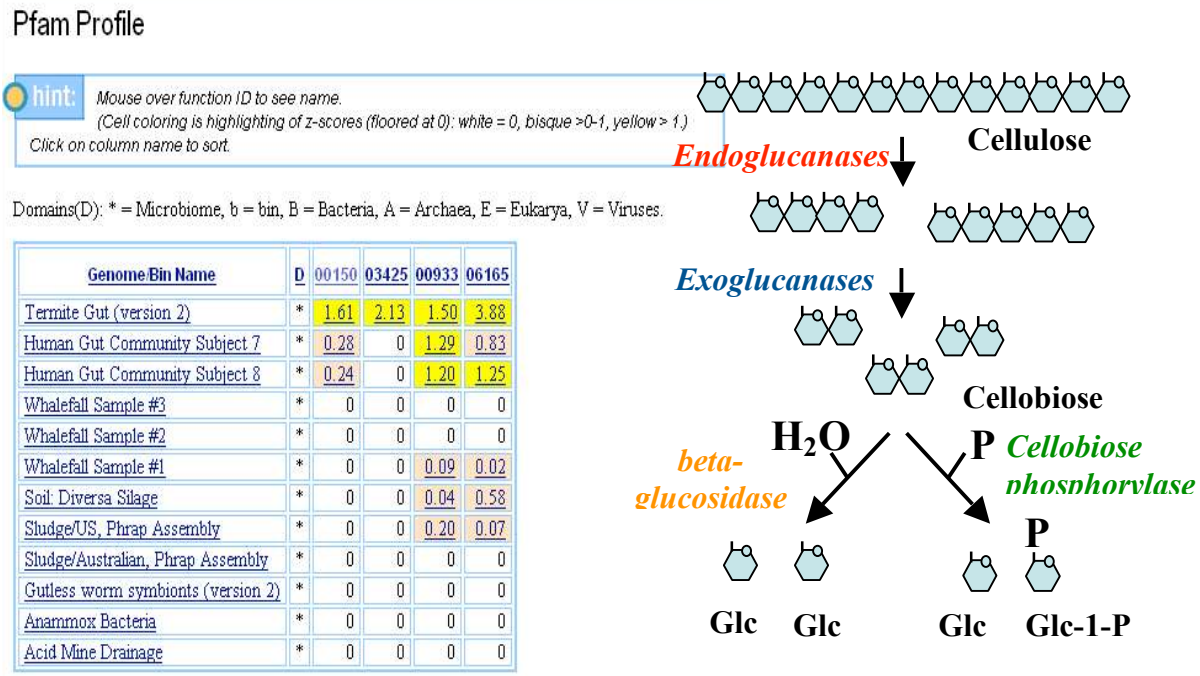


Fig. 7. A screenshot (at left of figure) from the IMG/M database (86) showing one implementation of gene-centric analysis available through this system. Four PFAM families involved in cellulose hydrolysis are shown in columns, color-coded to match the pathway schematic to the right of the figure. The relative representation of these families in twelve metagenomic datasets (rows) is shown as fractions normalized for dataset size. Over-represented families are further highlighted by color; bisque (moderately over-represented) and yellow (highly over-represented). This figure shows that termite hindgut followed by human gut samples have the greatest over-representation of genes involved in cellulose hydrolysis and indeed are the only communities that appear to have the enzymatic potential to breakdown cellulose, of the compared datasets. It also shows that one whale fall sample, a soil sample from the drainage path of a silage storage bunker, and one lab-scale phosphorus-removing sludge are moderately over-represented in genes for processing the dimer, cellobiose.

The implicit assumption of gene-centric analysis is that high relative abundance equates to metabolic and ecological significance. Knowledge of the ecosystem is required for simple sanity checks. For example, one of the most over-represented gene families in ocean surface waters relative to soil and whale fall (deep ocean) samples is the proteorhodopsin family that function as light-driven proton pumps (128), a function that is receiving great attention as a major missed energy flux in surface waters (110). A recent RNA-based study of a pikoplankton community in the photic zone confirmed that proteorhodopsins are indeed highly expressed, however other over-represented gene families, such as DNA repair photolyase, were not highly expressed bringing into question the metabolic or ecological significance of their high copy number in the community (41). Conversely other gene families that were poorly represented in the metagenomic data, such as *puf*B encoding a subunit of a light harvesting protein, were highly expressed (41) indicating that potentially important functions will be overlooked or underestimated by DNA-based gene-centric analysis.

In addition to violations of the implicit assumption, the method has a number of technical limitations. Chen and Pachter estimated that 6 Gbp of sequence data would be required to sample half the genes in a simulated soil community (17), whereas a typical metagenome project is on the order of 100 Mbp. Therefore, only genes present in high copy number in higher abundance organisms will be sampled meaning that the method is actually very low resolution. EGT data is also noisy due to uneven cloning efficiency of different genes (121), differences in gene length (longer genes will be detected more often on reads than short genes) and gene calling and annotation errors. A more pervasive problem may be the inability to normalize gene prediction between datasets. For example, read length will affect the ability to call genes, the shorter the read the lower the gene prediction resolution. Therefore, Sanger (~750 bp reads) and pyrosequence (100-200 bp reads) datasets cannot be directly compared using gene-centric analysis because of the differences in gene calling sensitivity between the two data types (141). A final word of caution on technical considerations; whole genome amplification of environmental DNAs is becoming a more common method, particularly for low biomass microbial communities (8, 32). The degree to which the amplification may skew relative

gene frequencies is presently unknown, but should be kept in mind when interpreting gene-centric analyses particularly between amplified and non-amplified datasets.

To differentiate between signal and noise, statistical tests to estimate the confidence of over- and under-representation of gene families have been reported (46, 107). Despite these statistical reassurances, simulated metagenomic datasets show that up to 20% of COGs may have incorrect frequency calls and should be interpreted with caution (89). However, the error rate reduces when gene family frequencies are grouped by metabolic pathway, because error in any given gene family will be averaged out in a multi-gene pathway. One important potential source of error when mapping gene family frequencies onto pathways is uneven coverage of the pathway. For example, broad gene families such as oxidoreductases can be non-specifically mapped to a pathway via incomplete EC numbers and give the false appearance that the pathway is over-represented. In the extreme case, the pathway may be entirely absent from the community and only the non-specific gene family is mapped to the pathway. This type of error can be overcome by weighting pathways for gene coverage or excluding incomplete EC numbers from the analysis. In addition, to avoid spurious prediction there is no substitution for manual inspection by experts of all results obtained by automatic data mining.

Data management

Shotgun sequencing of environmental samples produces massive amounts of data, that already dwarf the existing genomic sequences in public databases. This trend will not only continue, but will accelerate as the cost of sequencing continues to fall and more researchers enter into the field drawn by the promise of metagenomics and greater access to high throughput sequencing via new sequencing technologies. For the average researcher to make sense of this mountain of data, dedicated data management resources are required. There is a variety of web-based and standalone computational resources available for comparative genomic analysis, including ACT (16), MicrobesOnLine (5), CMR (104), ERGO (99), PUMA2 (80), COGENT++ (48) and IMG (87) but only recently have data management systems been developed specifically for metagenomic analysis, notably CAMERA (117), IMG/M (86) and SEED (97).

These systems allow comparison of a metagenome of interest to other genomes and metagenomes on multiple levels, including gene, protein family, pathway, scaffold, or complete genome and all include variants of the metagenome-specific tools described in the preceding sections (85). Most systems also allow some degree of curation by users to improve annotation. Although the same type of analyses can be performed without the aid of such systems, pre-packaged tools with transparent user interfaces can save considerable amounts of time even for expert users. Custom analyses need to be performed externally and the main use of dedicated metagenomic databases in these cases is improved curation over generic databases.

It is fair to say that all developers of metagenomic data management and analysis systems are struggling to keep pace with new data. This acute problem is manifest at two levels.

i) Data volume. Genomic data is more compressed than metagenomic data by virtue of assembly and underlying read data is typically not incorporated into comparative genome systems. By contrast, metagenomic systems not only keep read information, but quality data associated with reads for population analysis and quality control. The problem is expected to accelerate in the future as new sequencing technologies produce much larger volumes of data than traditional Sanger sequencing. For example, a single Illumina run produces ~ 1 Gb of sequence data compared to 700 kb for a Sanger run. While trace quality information may be important for quality assessment, their storage together with the sequence, and incorporation of quality information into sequence search methods might not be feasible.

ii) Pairwise comparisons. The cornerstone of comparative analysis is all-against-all comparisons. Ideally these should be pre-computed to prevent lengthy on-the-fly calculations for users. Unfortunately all-against-all comparisons scale poorly (quadratically) and for metagenomic data can become extremely computationally expensive. For example, 28.6 million protein sequences were compared using all-against-all BLAST searches in the Global Ocean Survey (GOS) study, which required more than 1 million CPU hours (143). The sheer size of the computational effort needed for this metagenomic dataset was unprecedented in sequence analysis. A parallelized implementation of BLAST, ScalaBLAST (96) is used to pre-compute all pairwise gene

similarities at the amino acid level for IMG/M reducing the computation time by ~30 fold (85). ScalaBLAST uses a combination of database sharing and task scheduling to achieve high computational performance (96). Computationally intensive tasks can also be bypassed by profile scans, using profile databases such as TIGRFAM, PFAM, COGS, and InterProScan. Because the number of profiles is constant, computational complexity scales linearly with the growth of the data, as opposed to quadratically in the case of all-against-all comparisons. One drawback of profile searches is that new families will not be identified, but such novel families will have unknown functions (hypothetical families) and in the first instance will not contribute to metabolic reconstruction efforts.

It remains to be seen if any data management system will be capable of incorporating all metagenomic data, and present the data in a pre-computed format for comparative analyses. More likely is that subsets of the data united by common phylogenetic or functional themes will be made into separate databases for analyses.

The final stage of any sequencing project is submission of the data to public repositories such as GenBank. Metagenomic data submission is more problematic than isolate genome submission because it is usually not discrete. For example, should a metagenomic dataset be described as a single entry or as multiple entries? On one hand, the data is a collection of sequence fragments from multiple species, which argues towards multiple entries. On the other hand, there is often a single sampling site and single study performed on the sequence, although this too is changing as single studies incorporate spatial or temporal sampling. At the JGI, we submit the data as one entry, and whenever possible subdivide it into bins of organisms. For example the metagenome of the *Olavius algarvensis* symbionts was submitted under accession number AASZ00000000, with scaffolds ranging between AASZ01000001 and AASZ01005597. The scaffolds assigned to particular genome bins were then assigned to sub-accession numbers, such as DS021107-DS021197 for the O. algarvensis Gamma 1 symbiont.

Concluding remarks

We hope that this review will serve as a useful primer for researchers embarking on their first metagenomic project. The field is moving forward rapidly driven by changes in

sequencing technology and the availability of many complementary technologies (137). We therefore anticipate that methodological details presented in this review will rapidly change and improve, particularly if (when) Sanger sequencing is no longer the main source of metagenomic data. The discussed methodological considerations and approaches for analyzing communities and populations, however, will no doubt persist for much longer enabling interpretation of metagenomic datasets and likely contributing many more profound insights into the microbial world.

Acknowledgements

We thank Alice McHardy, Susannah Tringe, Tanja Woyke and Gene Tyson for useful input and feedback during the course of preparing this review. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

References

1. **Abe, T., S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura.** 2003. Informatics for unveiling hidden genome signatures. Genome Res **13:**693-702.
2. **Abe, T., H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura.** 2005. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Res **12:**281-90.
3. **Allen, E. E., and J. F. Banfield.** 2005. Community genomics in microbial ecology and evolution. Nat Rev Microbiol **3:**489-98.
4. **Allen, E. E., G. W. Tyson, R. J. Whitaker, J. C. Detter, P. M. Richardson, and J. F. Banfield.** 2007. Genome dynamics in a natural archaeal population. Proc Natl Acad Sci U S A **104:**1883-8.
5. **Alm, E. J., K. H. Huang, M. N. Price, R. P. Koche, K. Keller, I. L. Dubchak, and A. P. Arkin.** 2005. The MicrobesOnline Web site for comparative genomics. Genome Res **15:**1015-22.
6. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-402.
7. **Amann, R., B. M. Fuchs, and S. Behrens.** 2001. The identification of microorganisms by fluorescence in situ hybridisation. Curr Opin Biotechnol **12:**231-6.
8. **Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer.** 2006. The marine viromes of four oceanic regions. PLoS Biol **4:**e368.
9. **Badger, J. H., and G. J. Olsen.** 1999. CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol **16:**512-24.
10. **Batzoglou, S., D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, and E. S. Lander.** 2002. ARACHNE: a whole-genome shotgun assembler. Genome Res **12:**177-89.
11. **Beja, O., M. T. Suzuki, E. V. Koonin, L. Aravind, A. Hadd, L. P. Nguyen, R. Villacorta, M. Amjadi, C. Garrigues, S. B. Jovanovich, R. A. Feldman, and E. F. DeLong.** 2000. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ Microbiol **2:**516-29.
12. **Besemer, J., and M. Borodovsky.** 1999. Heuristic approach to deriving models for gene finding. Nucleic Acids Res **27:**3911-20.
13. **Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer.** 2002. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A **99:**14250-5.
14. **Brocchieri, L.** 2001. Phylogenetic inferences from molecular sequences: review and critique. Theor Popul Biol **59:**27-40.
15. **Brodie, E. L., T. Z. Desantis, D. C. Joyner, S. M. Baek, J. T. Larsen, G. L. Andersen, T. C. Hazen, P. M. Richardson, D. J. Herman, T. K. Tokunaga, J. M. Wan, and M. K. Firestone.** 2006. Application of a high-density

oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl Environ Microbiol **72:**6288-98.

16. **Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill.** 2005. ACT: the Artemis Comparison Tool. Bioinformatics **21:**3422-3.
17. **Chen, K., and L. Pachter.** 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol **1:**106-12.
18. **Chou, H. H., and M. H. Holmes.** 2001. DNA sequence quality trimming and vector removal. Bioinformatics **17:**1093-104.
19. **Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork.** 2006. Toward automatic reconstruction of a highly resolved tree of life. Science **311:**1283-7.
20. **Cohan, F. M.** 2002. What are bacterial species? Annu Rev Microbiol **56:**457-87.
21. **Cole, J. R., B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje.** 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res **35:**D169-72.
22. **Coleman, M. L., M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. Delong, and S. W. Chisholm.** 2006. Genomic islands and the ecology and evolution of Prochlorococcus. Science **311:**1768-70.
23. **Culley, A. I., A. S. Lang, and C. A. Suttle.** 2006. Metagenomic analysis of coastal RNA virus communities. Science **312:**1795-8.
24. **Daims, H., S. Lucker, and M. Wagner.** 2006. daime, a novel image analysis program for microbial ecology and biofilm research. Environ Microbiol **8:**200-13.
25. **Dalevi, D., D. Dubhashi, and M. Hermansson.** 2006. Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. Bioinformatics **22:**517-22.
26. **Dandekar, T., B. Snel, M. Huynen, and P. Bork.** 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci **23:**324-8.
27. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res **27:**4636-41.
28. **DeLong, E. F.** 2005. Microbial community genomics in the ocean. Nat Rev Microbiol **3:**459-69.
29. **DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N. U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl.** 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science **311:**496-503.
30. **DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol **72:**5069-72.
31. **Deschavanne, P. J., A. Giron, J. Vilain, G. Fagot, and B. Fertil.** 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol Biol Evol **16:**1391-9.

32. **Edwards, R. A., B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D. M. Peterson, M. O. Saar, S. Alexander, E. C. Alexander, Jr., and F. Rohwer.** 2006. Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics **7:**57.
33. **Edwards, R. A., and F. Rohwer.** 2005. Viral metagenomics. Nat Rev Microbiol **3:**504-10.
34. **Enright, A. J., I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis.** 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature **402:**86-90.
35. **Eppley, J. M., G. W. Tyson, W. M. Getz, and J. F. Banfield.** 2007. Strainer: software for analysis of population variation in community genomic datasets. BMC Bioinformatics **8:**398.
36. **Erkel, C., M. Kube, R. Reinhardt, and W. Liesack.** 2006. Genome of Rice Cluster I archaea--the key methane producers in the rice rhizosphere. Science **313:**370-2.
37. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8:**175-85.
38. **Field, D., and N. Kyrpides.** 2007. The positive role of the ecological community in the genomic revolution. Microb Ecol **53:**507-11.
39. **Finn, R. D., J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman.** 2008. The Pfam protein families database. Nucleic Acids Res **36:**D281-8.
40. **Frazer, K. A., L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak.** 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Res **32:**W273-9.
41. **Frias-Lopez, J., Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W. Chisholm, and E. F. Delong.** 2008. Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A **105:**3805-10.
42. **Frishman, D., A. Mironov, H. W. Mewes, and M. Gelfand.** 1998. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res **26:**2941-7.
43. **Garcia Martin, H., N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. McHardy, C. Yeates, S. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. Blackall, K. D. McMahon, and P. Hugenholtz.** 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol **24:**1263-9.
44. **Gase, K., J. J. Ferretti, C. Primeaux, and W. M. McShan.** 1999. Identification, cloning, and expression of the CAMP factor gene (cfa) of group A streptococci. Infect Immun **67:**4725-31.
45. **Gilks, W. R., B. Audit, D. De Angelis, S. Tsoka, and C. A. Ouzounis.** 2002. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics **18:**1641-9.
46. **Gill, S. R., M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson.**

2006. Metagenomic analysis of the human distal gut microbiome. Science **312:**1355-9.

47. **Goldberg, S. M., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y. H. Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier, and J. C. Venter.** 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A **103:**11240-5.

48. **Goldovsky, L., P. Janssen, D. Ahren, B. Audit, I. Cases, N. Darzentas, A. J. Enright, N. Lopez-Bigas, J. M. Peregrin-Alvarez, M. Smith, S. Tsoka, V. Kunin, and C. A. Ouzounis.** 2005. CoGenT++: an extensive and extensible data environment for computational genomics. Bioinformatics **21:**3806-10.

49. **Gordon, D., C. Abajian, and P. Green.** 1998. Consed: a graphical tool for sequence finishing. Genome Res **8:**195-202.

50. **Grant, S., W. D. Grant, D. A. Cowan, B. E. Jones, Y. Ma, A. Ventosa, and S. Heaphy.** 2006. Identification of eukaryotic open reading frames in metagenomic cDNA libraries made from environmental samples. Appl Environ Microbiol **72:**135-43.

51. **Griffiths-Jones, S., S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman.** 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res **33:**D121-4.

52. **Hallam, S. J., K. T. Konstantinidis, N. Putnam, C. Schleper, Y. Watanabe, J. Sugahara, C. Preston, J. de la Torre, P. M. Richardson, and E. F. DeLong.** 2006. Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. Proc Natl Acad Sci U S A **103:**18296-301.

53. **Hallam, S. J., N. Putnam, C. M. Preston, J. C. Detter, D. Rokhsar, P. M. Richardson, and E. F. DeLong.** 2004. Reverse methanogenesis: testing the hypothesis with environmental genomics. Science **305:**1457-62.

54. **Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman.** 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol **5:**R245-9.

55. **Harrington, E. D., A. H. Singh, T. Doerks, I. Letunic, C. von Mering, L. J. Jensen, J. Raes, and P. Bork.** 2007. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc Natl Acad Sci U S A **104:**13913-8.

56. **Huang, X., and A. Madan.** 1999. CAP3: A DNA sequence assembly program. Genome Res **9:**868-77.

57. **Huber, J. A., D. B. Mark Welch, H. G. Morrison, S. M. Huse, P. R. Neal, D. A. Butterfield, and M. L. Sogin.** 2007. Microbial population structures in the deep marine biosphere. Science **318:**97-100.

58. **Hugenholtz, P.** 2002. Exploring prokaryotic diversity in the genomic era. Genome Biol **3:**REVIEWS0003.

59. **Huse, S. M., J. A. Huber, H. G. Morrison, M. L. Sogin, and D. M. Welch.** 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol **8:**R143.

60. **Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster.** 2007. MEGAN analysis of metagenomic data. Genome Res **17:**377-86.
61. **Jaffe, D. B., J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh, J. P. Mesirov, M. C. Zody, and E. S. Lander.** 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res **13:**91-6.
62. **Janssen, P., B. Audit, I. Cases, N. Darzentas, L. Goldovsky, V. Kunin, N. Lopez-Bigas, J. M. Peregrin-Alvarez, J. B. Pereira-Leal, S. Tsoka, and C. A. Ouzounis.** 2003. Beyond 100 genomes. Genome Biol **4:**402.
63. **Johnson, P. L., and M. Slatkin.** 2008. Accounting for bias from sequencing error in population genetic estimates. Mol Biol Evol **25:**199-206.
64. **Johnson, P. L., and M. Slatkin.** 2006. Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res **16:**1320-7.
65. **Karlin, S., and C. Burge.** 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet **11:**283-90.
66. **Karlin, S., J. Mrazek, and A. M. Campbell.** 1997. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol **179:**3899-913.
67. **Korlach, J., P. J. Marks, R. L. Cicero, J. J. Gray, D. L. Murphy, D. B. Roitman, T. T. Pham, G. A. Otto, M. Foquet, and S. W. Turner.** 2008. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci U S A **105:**1176-81.
68. **Koski, L. B., and G. B. Golding.** 2001. The closest BLAST hit is often not the nearest neighbor. J Mol Evol **52:**540-2.
69. **Krause, L., N. N. Diaz, D. Bartels, R. A. Edwards, A. Puhler, F. Rohwer, F. Meyer, and J. Stoye.** 2006. Finding novel genes in bacterial communities isolated from the environment. Bioinformatics **22:**e281-9.
70. **Kunin, V., S. He, F. Warnecke, S. B. Peterson, H. Garcia Martin, M. Haynes, N. Ivanova, L. L. Blackall, M. Breitbart, F. Rohwer, K. D. McMahon, and P. Hugenholtz.** 2008. A bacterial metapopulation adapts locally to phage predation despite global dispersal. Genome Res **18:**293-7.
71. **Kunin, V., and C. A. Ouzounis.** 2005. Clustering the annotation space of proteins. BMC Bioinformatics **6:**24.
72. **Kurokawa, K., T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. K. Sharma, T. P. Srivastava, T. D. Taylor, H. Noguchi, H. Mori, Y. Ogura, D. S. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi, and M. Hattori.** 2007. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res **14:**169-81.
73. **Kyrpides, N. C., and C. A. Ouzounis.** 1999. Whole-genome sequence annotation: 'Going wrong with confidence'. Mol Microbiol **32:**886-7.
74. **Lapidus, A.** In press. Genome Sequence Databases (overview): Sequencing and Assembly. *In* M. Schaechter (ed.), The Encyclopedia of Microbiology. Elsevier.
75. **Legault, B. A., A. Lopez-Lopez, J. C. Alba-Casado, W. F. Doolittle, H. Bolhuis, F. Rodriguez-Valera, and R. T. Papke.** 2006. Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. BMC Genomics **7:**171.

76. **Lo, I., V. J. Denef, N. C. Verberkmoes, M. B. Shah, D. Goltsman, G. DiBartolo, G. W. Tyson, E. E. Allen, R. J. Ram, J. C. Detter, P. Richardson, M. P. Thelen, R. L. Hettich, and J. F. Banfield.** 2007. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature **446:**537-41.

77. **Lowe, T. M., and S. R. Eddy.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:**955-64.

78. **Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer.** 2004. ARB: a software environment for sequence data. Nucleic Acids Res **32:**1363-71.

79. **Macke, T. J., D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath.** 2001. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res **29:**4724-35.

80. **Maltsev, N., E. Glass, D. Sulakhe, A. Rodriguez, M. H. Syed, T. Bompada, Y. Zhang, and M. D'Souza.** 2006. PUMA2--grid-based high-throughput analysis of genomes and metabolic pathways. Nucleic Acids Res **34:**D369-72.

81. **Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg.** 1999. Detecting protein function and protein-protein interactions from genome sequences. Science **285:**751-3.

82. **Marcotte, E. M., M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg.** 1999. A combined algorithm for genome-wide prediction of protein function. Nature **402:**83-6.

83. **Marcy, Y., C. Ouverney, E. M. Bik, T. Losekann, N. Ivanova, H. G. Martin, E. Szeto, D. Platt, P. Hugenholtz, D. A. Relman, and S. R. Quake.** 2007. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A **104:**11889-94.

84. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376-80.

85. **Markowitz, V. M.** 2007. Microbial genome data resources. Curr Opin Biotechnol **18:**267-72.

86. **Markowitz, V. M., N. Ivanova, K. Palaniappan, E. Szeto, F. Korzeniewski, A. Lykidis, I. Anderson, K. Mavromatis, V. Kunin, H. Garcia Martin, I. Dubchak, P. Hugenholtz, and N. C. Kyrpides.** 2006. An experimental metagenome data management and analysis system. Bioinformatics **22:**e359-67.

87. **Markowitz, V. M., F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X. Zhao, I. Dubchak, P. Hugenholtz, I. Anderson, A. Lykidis, K. Mavromatis, N. Ivanova, and N. C. Kyrpides.** 2006. The integrated microbial genomes (IMG) system. Nucleic Acids Res **34:**D344-8.

88. **Martin-Cuadrado, A. B., P. Lopez-Garcia, J. C. Alba, D. Moreira, L. Monticelli, A. Strittmatter, G. Gottschalk, and F. Rodriguez-Valera.** 2007. Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. PLoS ONE **2:**e914.

89. **Mavromatis, K., N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides.** 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods **4:**495-500.

90. **McHardy, A. C., H. G. Martin, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos.** 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods **4:**63-72.

91. **McHardy, A. C., and I. Rigoutsos.** 2007. What's in the mix: phylogenetic classification of metagenome sequence samples. Curr Opin Microbiol **10:**499-503.

92. **Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter.** 2000. A whole-genome assembly of Drosophila. Science **287:**2196-204.

93. **Nakashima, H., M. Ota, K. Nishikawa, and T. Ooi.** 1998. Genes from nine genomes are separated into their organisms in the dinucleotide composition space. DNA Res **5:**251-9.

94. **Neufeld, J. D., Y. Chen, M. G. Dumont, and J. C. Murrell.** 2008. Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. Environ Microbiol.

95. **Noguchi, H., J. Park, and T. Takagi.** 2006. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res **34:**5623-30.

96. **Oehmen, C., and J. Nieplocha.** 2006. ScalaBLAST: A Scalable Implementation of BLAST for High-Performance Data-Intensive Bioinformatics Analysis. IEEE Transactions on Parallel and Distributed Systems **17:**740-749.

97. **Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V.**

Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res **33:**5691-702.

98. **Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev.** 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A **96:**2896-901.

99. **Overbeek, R., N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov, Jr., K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, N. Mikhailova, O. Vasieva, A. Osterman, V. Vonstein, M. Fonstein, N. Ivanova, and N. Kyrpides.** 2003. The ERGO genome analysis and discovery system. Nucleic Acids Res **31:**164-71.

100. **Palmer, C., E. M. Bik, M. B. Eisen, P. B. Eckburg, T. R. Sana, P. K. Wolber, D. A. Relman, and P. O. Brown.** 2006. Rapid quantitative profiling of complex microbial populations. Nucleic Acids Res **34:**e5.

101. **Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates.** 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96:**4285-8.

102. **Pelletier, E., A. Kreimeyer, S. Bocs, Z. Rouy, G. Gyapay, R. Chouari, D. Riviere, A. Ganesan, P. Daegelen, A. Sghir, G. N. Cohen, C. Medigue, J. Weissenbach, and D. Le Paslier.** 2008. "Candidatus Cloacamonas acidaminovorans": genome sequence reconstruction provides a first glimpse of a new bacterial division. J Bacteriol **190:**2572-9.

103. **Peplies, J., S. C. Lau, J. Pernthaler, R. Amann, and F. O. Glockner.** 2004. Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. Environ Microbiol **6:**638-45.

104. **Peterson, J. D., L. A. Umayam, T. Dickinson, E. K. Hickey, and O. White.** 2001. The Comprehensive Microbial Resource. Nucleic Acids Res **29:**123-5.

105. **Podar, M., C. B. Abulencia, M. Walcher, D. Hutchison, K. Zengler, J. A. Garcia, T. Holland, D. Cotton, L. Hauser, and M. Keller.** 2007. Targeted access to the genomes of low-abundance organisms in complex microbial communities. Appl Environ Microbiol **73:**3205-14.

106. **Pop, M., A. Phillippy, A. L. Delcher, and S. L. Salzberg.** 2004. Comparative genome assembly. Brief Bioinform **5:**237-48.

107. **Rodriguez-Brito, B., F. Rohwer, and R. A. Edwards.** 2006. An application of statistics to comparative metagenomics. BMC Bioinformatics **7:**162.

108. **Ronaghi, M.** 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res **11:**3-11.

109. **Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean

Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol **5:**e77.

110. **Sabehi, G., A. Loy, K. H. Jung, R. Partha, J. L. Spudich, T. Isaacson, J. Hirschberg, M. Wagner, and O. Beja.** 2005. New insights into metabolic properties of marine bacteria encoding proteorhodopsins. PLoS Biol **3:**e273.
111. **Sandberg, R., G. Winberg, C. I. Branden, A. Kaske, I. Ernberg, and J. Coster.** 2001. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res **11:**1404-9.
112. **Sanger, F., and A. R. Coulson.** 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol **94:**441-8.
113. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A **74:**5463-7.
114. **Schmeisser, C., C. Stockigt, C. Raasch, J. Wingender, K. N. Timmis, D. F. Wenderoth, H. C. Flemming, H. Liesegang, R. A. Schmitz, K. E. Jaeger, and W. R. Streit.** 2003. Metagenome survey of biofilms in drinking-water networks. Appl Environ Microbiol **69:**7298-309.
115. **Sekar, R., B. M. Fuchs, R. Amann, and J. Pernthaler.** 2004. Flow sorting of marine bacterioplankton after fluorescence in situ hybridization. Appl Environ Microbiol **70:**6210-9.
116. **Selengut, J. D., D. H. Haft, T. Davidsen, A. Ganapathy, M. Gwinn-Giglio, W. C. Nelson, A. R. Richter, and O. White.** 2007. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res **35:**D260-4.
117. **Seshadri, R., S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier.** 2007. CAMERA: A Community Resource for Metagenomics. PLoS Biol **5:**e75.
118. **Shah, N., M. V. Teplitsky, S. Minovitsky, L. A. Pennacchio, P. Hugenholtz, B. Hamann, and I. L. Dubchak.** 2005. SNP-VISTA: an interactive SNP visualization tool. BMC Bioinformatics **6:**292.
119. **Snel, B., P. Bork, and M. A. Huynen.** 1999. Genome phylogeny based on gene content. Nat Genet **21:**108-10.
120. **Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A **103:**12115-20.
121. **Sorek, R., Y. Zhu, C. J. Creevey, M. P. Francino, P. Bork, and E. M. Rubin.** 2007. Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer. Science.
122. **Staden, R., K. F. Beal, and J. K. Bonfield.** 2000. The Staden package, 1998. Methods Mol Biol **132:**115-30.
123. **Strous, M., E. Pelletier, S. Mangenot, T. Rattei, A. Lehner, M. W. Taylor, M. Horn, H. Daims, D. Bartol-Mavel, P. Wincker, V. Barbe, N. Fonknechten, D. Vallenet, B. Segurens, C. Schenowitz-Truong, C. Medigue, A. Collingro, B. Snel, B. E. Dutilh, H. J. Op den Camp, C. van der Drift, I. Cirpus, K. T. van de Pas-Schoonen, H. R. Harhangi, L. van Niftrik, M. Schmid, J. Keltjens, J. van de Vossenberg, B. Kartal, H. Meier, D. Frishman, M. A. Huynen, H. W. Mewes, J. Weissenbach, M. S. Jetten, M. Wagner, and D. Le Paslier.** 2006.

Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature **440:**790-4.

124. **Suzuki, M. T., and S. J. Giovannoni.** 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol **62:**625-30.
125. **Tatusov, R. L., E. V. Koonin, and D. J. Lipman.** 1997. A genomic perspective on protein families. Science **278:**631-7.
126. **Teeling, H., A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glockner.** 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol **6:**938-47.
127. **Thomas, C. A., Jr.** 1971. The genetic organization of chromosomes. Annu Rev Genet **5:**237-56.
128. **Tringe, S. G., and E. M. Rubin.** 2005. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet **6:**805-14.
129. **Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin.** 2005. Comparative metagenomics of microbial communities. Science **308:**554-7.
130. **Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon.** 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444:**1027-31.
131. **Tyson, G. W., and J. F. Banfield.** 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol **10:**200-7.
132. **Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37-43.
133. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66-74.
134. **von Mering, C., P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork.** 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. Science **315:**1126-30.
135. **von Mering, C., L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork.** 2007. STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res **35:**D358-62.
136. **von Wintzingerode, F., U. B. Gobel, and E. Stackebrandt.** 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev **21:**213-29.
137. **Warnecke, F., and P. Hugenholtz.** 2007. Building on basic metagenomics with complementary technologies. Genome Biol **8:**231.
138. **Warnecke, F., P. Luginbuhl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. C. McHardy, G. Djordjevic, N. Aboushadi, R.**

Sorek, S. G. Tringe, M. Podar, H. G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N. C. Kyrpides, E. G. Matson, E. A. Ottesen, X. Zhang, M. Hernandez, C. Murillo, L. G. Acosta, I. Rigoutsos, G. Tamayo, B. D. Green, C. Chang, E. M. Rubin, E. J. Mathur, D. E. Robertson, P. Hugenholtz, and J. R. Leadbetter. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature **450:**560-5.

139. **Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko.** 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **36:**D13-21.

140. **Whitaker, R. J., and J. F. Banfield.** 2006. Population genomics in natural microbial communities. Trends Ecol Evol **21:**508-16.

141. **Wommack, K. E., J. Bhavsar, and J. Ravel.** 2008. Metagenomics: Read length matters. Appl Environ Microbiol.

142. **Woyke, T., H. Teeling, N. N. Ivanova, M. Huntemann, M. Richter, F. O. Gloeckner, D. Boffelli, I. J. Anderson, K. W. Barry, H. J. Shapiro, E. Szeto, N. C. Kyrpides, M. Mussmann, R. Amann, C. Bergin, C. Ruehland, E. M. Rubin, and N. Dubilier.** 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature **443:**950-5.

143. **Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. PLoS Biol **5:**e16.

144. **Zhang, T., M. Breitbart, W. H. Lee, J. Q. Run, C. L. Wei, S. W. Soh, M. L. Hibberd, E. T. Liu, F. Rohwer, and Y. Ruan.** 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol **4:**e3.