Microbial diversity: let's tell it how it is

Created: March 04, 2004

An impressive number of bacteria—about 30,000 species—are represented in GenBank. However, our view of the microbial world is both scant and skewed. A recent estimate suggests that the sea may support as many as 2 million different bacteria, and a ton of soil might contain 4 million (1). Less than half of the bacteria represented in GenBank—about 13,000—have been formally described, and almost all of these (90%) lie within 4 of the 40 bacterial divisions (2). Similar or greater paucity of knowledge also exists for archaea and viruses (3).

Sampling "wild" microorganisms leads to the discovery of new species and novel metabolisms, which may be important from both a basic science and a practical perspective (for example, see Refs 4,5 [search PubMed]). For example, if we characterized the community in the human gut, it would be easier to spot non-native organisms in food poisoning outbreaks. Pathogens that may underlie neurological syndromes that present with features of infection would stand out against the background flora (1). Engineered communities of microorganisms might also be able to assist clean up of environmental disasters or create sustainable energy sources.

Exploring bacterial diversity is typically done by amplifying rRNA genes, in particular 16S rRNA genes, from DNA samples isolated from a habitat. The sequences are then compared to each other and to the 16S rRNA sequences from known species. If no close match to an existing 16S rRNA gene sequence is found, then the test sequence is thought to represent a new bacterium and is listed in GenBank as "uncultured bacterium". Even in well-studied, discrete places like the human mouth, new groups of uncultured bacteria continue to be discovered all the time. A newly identified organism has to be isolated and cultured in the lab to be described further; but many bugs are just not amenable to monoculture—they have adapted to living in a specific environment and may need to be part of a complex community to survive (1-3).

16S rRNA genes are considered standard because they are thought to be conserved across vast taxonomic distance (they are critical for protein translation), yet show some sequence variation between closely related species. However, one problem with using rRNA genes is that they are often present in multiple copy numbers; therefore, other representative genes may be used for sampling specific populations.

Whole Genome Shotgun Sequencing of Environmental Samples

New approaches to environmental sampling are emerging (6–8). One of these used a microarray to discover and assist in the isolation of new viruses (6). Two others have used whole genome shotgun (WGS) sequencing on a population of bacteria, obviating the need to isolate each organism before sequencing can begin (7,8). These methods, used

in combination with existing methods, may provide shortcuts to the discovery of new genes and give a holistic persective to microbial populations.

One recent study used a WGS approach to explore a sample from an acid mine drainage biofilm (7; AADL00000000). These investigators report that near-complete genomes for *Leptospirillum* Group II and *Ferroplasma* Type II were assembled, along with more fragmentory assemblies for *Leptospirillum* Group III, *Thermoplasmatales archaeon gpl*, and *Ferroplasma acidarmanus Type I*. Analysis of the results provided some insight into how such organisms survive in an extreme environment.

In another test case of the WGS method, Venter *et al.* (8) sampled water from the Sargasso Sea—one of the most well-characterized regions of ocean in the world. The major set of samples produced 1.66 million short sequences, some of which could be grouped together into larger genomic pieces. There remained about 400,000 paired-end reads and singleton reads.

Finding the Data

Using a WGS method to sequence an undefined population as opposed to a single organism adds significant complexity to the assembly process and to the identification of genes. About 25% of the assembled data from the Sargasso Sea had 3X coverage or greater; these well-sampled portions were used to cluster the sequence by “organism”.

The assembled sequences have been deposited in the WGS division [www.ncbi.nlm.nih.gov/Genbank/wgs.html] of GenBank, with the project Accession number AACY01000000; thus, there are 811,372 WGS contigs in GenBank with the Accession numbers AACY01000001–AACY01811372. 498,641 of the WGS contigs are assembled into 232,442 scaffolds, the rest remain “singleton” WGS contigs; all but 10,685 of the scaffolds are made up of two contigs only. For the organism genomes listed in Table 1, 301 of the total scaffolds plus 36 singleton WGS contigs were used; the remainder have not been associated with any particular organism.

All of the short sequence reads, including those that were not included in the assembly, can be found in the Trace Archive.

The assemblies were then further clustered into 30 tentative organism “bins” based on depth of coverage, oligonucleotide frequencies and similarities to previously sequenced genomes. Of these, 12 are of sufficient size to be considered a genome assembly, while the remaining 16 are relatively small single scaffolds (Table 1). All organism bins have been assigned a taxonomy ID, and have been placed in the taxonomic tree. Figure 1 shows the graphical representation of the cf. Shewanella SAR-1 “genome” sequence.

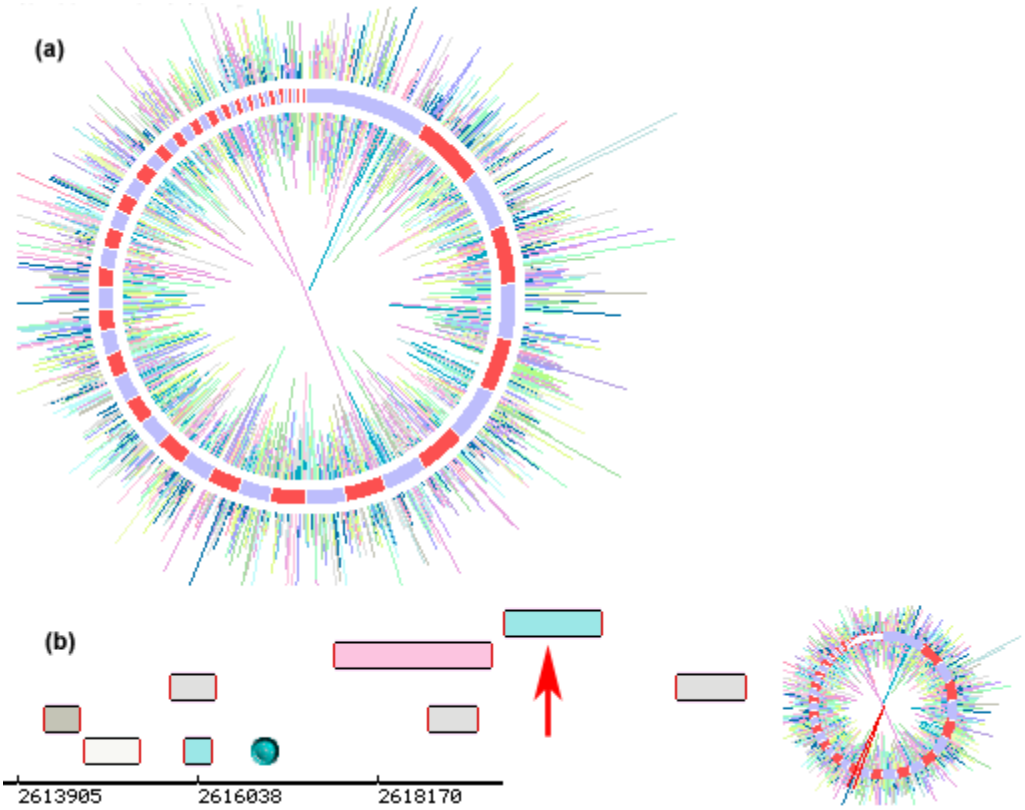

Figure 1: **(a)** Genome view of cf. Shewanella SAR-1, constructed from the whole genome shotgun sequence derived from Sargasso Sea environmental samples (8). Genes have been classified according to the COG functional categories [www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?fun=all] of the protein products, and color-coded accordingly. note that the actual order of the scaffolds is unknown, so in this representation they have been ordered by size. Clicking on the image reveals the gene sequences and approximate location. **(b)** Selecting one of the genes (in this case, the blue gene around position 2619000) shows the results of an automated BLAST search [http://www.ncbi.nlm.nih.gov/sutils/blink.cgi?pid=44354903&cut=95] (BLink). This gene is similar to L-sorbosone dehydrogenase from a variety of bacteria, archaea, and fungi. L-sorbosone dehydrogenase is an enzyme required for the biosynthesis of L-ascorbic acid, a product widely used in the food industry as a vitamin and antioxidant.

Each of the 28 organism “genomes” can be viewed in a similar mannner (see Table 1).

Table 1. The organism bins assembled from the Sargasso Sea WGS environmental sample dataset (8).

Organism Bin	Description	Data	Further Reading
Genome Assemblies			
cf. Alphaproteobacteria SAR-1	Oligotrophic Typical of marine bacterioplankton	Genome | GenBank	PubMed | Books
cf. Archaea SAR-1	One of the three major domains of life Often inhabit extreme environments	Genome | GenBank	PubMed | Books
cf. Bacteria SAR-1	One of the three major domains of life	Genome | GenBank	PubMed | Books
cf. Burkholderia SAR-1	Gram-negative bacilli Aerobic Found in a variety of aquatic environments	Genome | GenBank	PubMed | Books
cf. Gammaproteobacteria SAR-1	Purple bacteria Some plant pathogens	Genome | GenBank	PubMed | Books
cf. Microbulbifer SAR-1	Marine bacteria that degrade and recycle complex carbohydrates	Genome | GenBank	PubMed | Books
cf. Prochlorococcus SAR-1	Smallest known photosynthetic organism The most abundant in the ocean	Genome | GenBank	PubMed | Books

Organism Bin	Description	Data	Further Reading
cf. Proteobacteria SAR-1	Phylum includes nitrogen-fixing bacteria and enteric bacteria	Genome | GenBank	PubMed | Books
cf. Pseudomonadaceae SAR-1	Gram-negative rods Often motile Includes many plant and a few animal pathogens	Genome | GenBank	PubMed | Books
cf. Shewanella SAR-1	Versatile metabolism Potential biotech applications such as heavy metal or chlorinated solvent reduction	Genome | GenBank	PubMed | Books
cf. Shewanella SAR-2*	Versatile metabolism Potential biotech applications such as heavy metal or chlorinated solvent reduction	Genome | GenBank	PubMed | Books
cf. Streptomyces SAR-1	Superficially similar to fungi (filaments and spores) Common in many habitats	Genome | GenBank	PubMed | Books
Single Scaffolds			
cf. Actinobacteria SAR-1	High G+C group of Gram-positive bacteria Most found in soil Some pathogens	GenBank	PubMed | Books
cf. Bordetella SAR-1	Gram-negative coccobacilli Strict aerobes	GenBank	PubMed | Books
cf. Burkholderiaceae SAR-1	Occupy diverse ecological niches May have potential for biotech applications but also involved in human infections	GenBank	PubMed | Books
cf. Caulobacter SAR-1	Found in oligotrophic environments Prosthecate (having appendages)	GenBank	PubMed | Books
cf. Crenarchaeota SAR-1	Archaeal Most species are motile Tolerant of extreme acidity and temperature	GenBank	PubMed | Books
cf. Cyanobacteria SAR-1	Aquatic and photosynthetic Often called ”blue-green algae”	GenBank	PubMed | Books
cf. Enterobacteriaceae SAR-1	Large Gram-negative rods Facultative anaerobes	GenBank	PubMed | Books
cf. Haemophilus SAR-1	Gram-negative rods Like to grow on blood agar Some pathogens	GenBank	PubMed | Books
cf. Magnetococcus SAR-1	Gram-negative coccus Magnetic bacteria Usually located at sediment-water interface	GenBank	PubMed | Books
cf. Magnetospirillum SAR-1	Magnetic bacteria	GenBank	PubMed | Books
cf. Ralstonia SAR-1	Includes medically and economically important plant and animal pathogens	GenBank	PubMed | Books
cf. Rhizobiales SAR-1	Involved in nitrogen fixation, often in symbiotic relationships with plants	GenBank	PubMed | Books
cf. Sinorhizobium SAR-1	Symbiotic nitrogen fixation in plant root nodules	GenBank	PubMed | Books
cf. Spirochaetales SAR-1	Spiral rods Some pathogens (e.g. *Borrelia burgdorferi* - Lyme disease)	GenBank	PubMed | Books
cf. Streptomycetaceae SAR-1	Typically aerobic and found in soil Some parasitic forms	GenBank	PubMed | Books
cf. Vibrionaceae SAR-1	Gram-negative, non-sporing rods Generally motile Many strains of *Vibrio* genus cause infection	GenBank	PubMed | Books

cf. is used to designate an unidentified species of the genus. Therefore, “cf. Burkholderia” means “something that is like the genus Burkholderia“ (in this case, by sequence similarity).
As each organism bin could actually represent several different unidentified species, a strain name cannot be assigned, so instead, the suffix "SAR-#" identifies each bin as a “Sargasso Sea cyber-species”.
* cf. Shewanella SAR-2: two distinct Shewanella genomes were constructed from the dataset.

A variety of approaches suggested that there are at least 1000 species represented in the Sargasso Sea samples (8). *Burkholderia* species were represented in a high proportion (a genus that includes human and plant pathogens and some environmentally important bacteria), as were two distinct

species closely related to *Shewanella oneidensis*. Both of these genera require a more nutrient-rich environment than the open ocean can offer, suggesting that they originated from microhabitats such as marine snow. The cyanobacterium *Prochlorococcus* was also relatively abundant in some samples.

Although the primary focus of this study was on bacterial populations, WGS environmental sampling may be an equally valid approach for exploring plasmids (Table 2), phage, viruses, and eukaryotic microbes.

Table 2. The plasmid bins assembled from the Sargasso Sea WGS environmental sample dataset (8).

Plasmid Bin	Data
Plasmid pSAR-1	GenBank
Plasmid pSAR-2	GenBank
Plasmid pSAR-3	GenBank
Plasmid pSAR-4	GenBank
Plasmid pSAR-5	GenBank
Plasmid pSAR-6	GenBank
Plasmid pSAR-7	GenBank
Plasmid pSAR-8	GenBank
Plasmid pSAR-9	GenBank

References

1. Curtis TP, Sloan WT, Scannell JW. Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci USA 99:10494–10499; 2002. (PubMed)(Full text in PMC)
2. DeLong EF. Microbial seascapes revisited. Curr Opin Microbiol 4:290–295; 2001. (PubMed)
3. Roossinck MJ. Plant RNA virus evolution. Curr Opin Microbiol 6:406–409; 2003. (PubMed)
4. Kazor CE, Mitchell PM, Lee AM, Stokes LN, Loesche WJ, Dewhirst FE, Paster BJ. Diversity of bacterial populations on the tongue dorsa of patients with halitosis and healthy patients. J Clin Microbiol 41:558–563; 2003. (PubMed) (Full text in PMC)
5. Béejà O, Aravind L, Koonin EV. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289:1902–1906; 2000. (PubMed)
6. Wang D, Urisman A, Liu YT. Viral Discovery and Sequence Recovery Using DNA Microarrays. PLoS Biol 1:2003. (PubMed)(Full text in PMC)
7. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF Nature advance online publication, 01 Feb 2004 (doi:10.1038/nature 02340)(PubMed)
8. Venter JC, Remington K, Heidelberg J. Environmental whole genome shotgun sequencing: the Sargasso Sea. Science 2004.

Important Links

Genomes

Browse “genomes” by organism bin [/books/bv.fcgi?call=bv.View..ShowSection&rid=coffeebrk.table.634]

Taxonomy

View taxonomic tree of all organism bins

GenBank

Genome assemblies in GenBank

Trace Archive

All Sargasso Sea traces