METAGENOMICS: DNA SEQUENCING OF ENVIRONMENTAL SAMPLES

Susannah Green Tringe and Edward M. Rubin

Abstract | Although genomics has classically focused on pure, easy-to-obtain samples, such as microbes that grow readily in culture or large animals and plants, these organisms represent only a fraction of the living or once-living organisms of interest. Many species are difficult to study in isolation because they fail to grow in laboratory culture, depend on other organisms for critical processes, or have become extinct. Methods that are based on DNA sequencing circumvent these obstacles, as DNA can be isolated directly from living or dead cells in various contexts. Such methods have led to the emergence of a new field, which is referred to as metagenomics.

Complete genome sequences have been obtained from hundreds of organisms. In the well-studied, easily manipulated organisms that were targeted by early genome projects, genotypic and phenotypic data could be compared and genome-based hypotheses tested experimentally. Comparative genomics allowed experiment-based annotations to be transferred to novel genomes, and quickly gained prominence as a valuable tool for understanding both genes and genomes[1]. Protocols for DNA purification are well-established; although some optimization is usually required for DNA extraction from new organisms, the effort involved is generally much less than that required to develop techniques for genetic manipulation. As a result, the focus of some genomic sequencing has changed dramatically, so that DNA sequence is used to predict features and behaviours of otherwise poorly understood organisms, and to understand the genetic basis of characterized traits.

Barriers to genome sequencing range from a lack of sufficient material for the construction of sequencing libraries to the cost of sequencing. Improvements in cloning and sequencing technologies have consistently decreased the amount of starting material needed for library construction, making DNA sequencing feasible for a wide range of organisms that are otherwise difficult to study. Meanwhile, the progressive reduction in the cost of high-throughput sequencing has made it feasible to sequence libraries that are constructed from mixtures of organisms, even those that are 'contaminated' with genomes other than that of the targeted organism[2]. This has opened the door to sequence-based studies of organisms and environments that were previously thought to be inaccessible, including obligate pathogens and symbionts, which cannot survive outside their hosts; environmental microbes, most of which cannot be grown in pure culture; and ancient organisms for which fossilized remains are the only record. DNA for these studies is extracted directly from the organisms in their natural habitat, such as host tissue or soil, and cloned into sequencing vectors. The resulting libraries contain genome fragments from a heterogeneous mix of species, strains and subpopulations. Therefore, the sequence data from these libraries harbour a wealth of information on community dynamics, such as species interactions and selective processes.

Here we focus on the insights that have emerged from DNA sequencing of naturally occurring populations and communities. The first section describes methodological advances that have allowed the sequencing of natural populations. The second section gives examples of studies that have used these techniques. The third section suggests future directions that these studies could take.

Environmental nucleic-acid analysis

Natural samples contain DNA in various packages, including free DNA, virus particles, and prokaryotic and eukaryotic cells. These can be suspended in water,

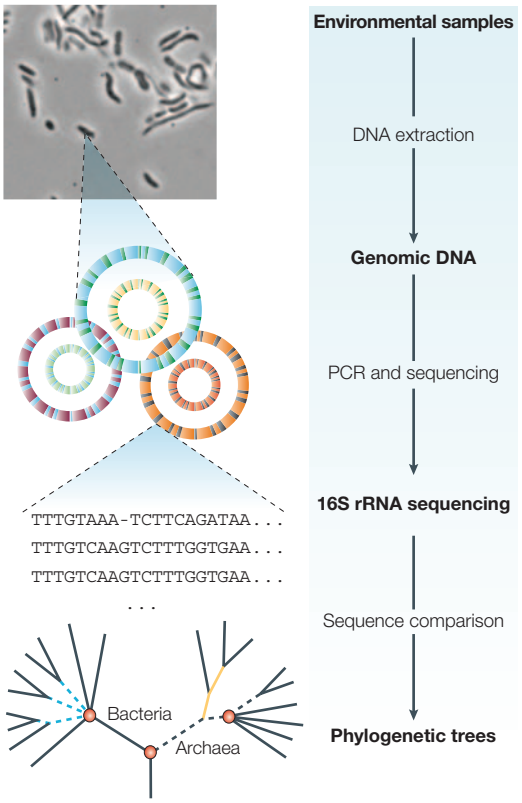
Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA.
Correspondence to E.M.R.
e-mail: emrubin@lbl.gov
doi:10.1038/nrg1709
Published online
11 October 2005

Box 1 | 16S ribosomal RNA analysis of microbial communities

In the 1980s environmental microbiologists realized that only a small fraction of the microscopically observable organisms in a sample were capable of colony formation. Pioneering experiments by Norman Pace and colleagues revealed, through ribosomal RNA (rRNA) sequencing, that these 'unculturable' microbes represented novel species that were often only distantly related to known, cultured lineages[14,87,88]. An rRNA sequence can provide a unique molecular 'bar code' for identifying an organism and placing it in an evolutionary context, providing a first glimpse into the broad diversity that is invisible to culture-based approaches. The labour-intensive methods that were initially used, such as direct sequencing of isolated 5S rRNA or screening of genomic libraries before sequencing, were eventually supplanted by PCR-based methods. This is because well-conserved sequences that participate in secondary-structure formation can be targeted for amplification by universal primers to generate clone libraries[89,90].

In these studies DNA is extracted directly from an environmental sample such as ocean water, soil or a biofilm, and the 16S genes of the community microbes are then amplified from the mixed genomic DNA using PCR (see figure) (for reviews see REFS 91,92). The PCR products are cloned into vectors and sequenced, which produces rRNA signatures for the microbes that were present in the sample. Comparison of these sequences with databases of 16S ribosomal RNA genes allows them to be phylogenetically classified. The frequencies of particular small subunit (SSU) rRNA clone sequences provide a rough preliminary estimate of the community structure, as sequences from dominant community members should be more abundant. In some cases the presence of SSU rRNA sequences from specialized clades, such as METHANOGENS, can also indicate functional activities. The downside, however, is that even species that are closely related on the basis of SSU rRNA sequence can have different lifestyles, and the phylogenetic position of organisms with no cultured close relatives frequently offers little or no insight into their phenotypic characteristics.

16S rRNA genes have been amplified, cloned and sequenced from thousands of distinct environmental niches, but these surveys routinely continue to identify unique new bacterial and archaeal taxa. Tools (such as ARB and EstimateS) and databases (such as the Ribosomal Database Project) have been developed to manage and analyse this flood of data.



METHANOGENS
A group of hydrogen-consuming Archaea that generate methane by reduction of carbon dioxide.

BIOFILM
A layered aggregate of microorganisms.

NORMAL FILTRATION
A process in which particles that are above a certain size are removed from a fluid by forcing the solution through a membrane containing pores of a defined size.

TANGENTAL FLOW FILTRATION
A process in which a fluid is pumped tangentially along the surface of a porous membrane and an applied pressure forces some of the fluid, as well as dissolved particles of sufficiently small size, across the membrane.

GRAM-POSITIVE BACTERIA
Members of the phyla Actinobacteria and Firmicutes, which have a single membrane and a thick cell wall that is made of crosslinked peptidoglycan and therefore can be stained with the Gram staining procedure.

bound to a solid matrix, such as soil, or encased in a BIOFILM or tissue. Extraction methods must be chosen carefully on the basis of the medium and the DNA population of interest.

Aquatic samples must be concentrated, typically by NORMAL or TANGENTIAL FLOW FILTRATION, and might also be pre-filtered to remove large cells or debris[3]. The choice of filter size is crucial, as cells that are smaller or larger than the size fraction that is targeted will be invisible to further analysis. In this way, filtration protocols can be chosen to enrich for eukaryotic cells, prokaryotic cells or viral particles[4,5]. Cells in soils and sediments are less easily concentrated than aquatic samples and often contain enzyme inhibitors, such as humic acids, that must be removed before amplification or cloning. Solid-matrix DNA isolation is either direct (cells are lysed within the sample material) or indirect (cells are separated from non-cellular material before lysis). In either case, contaminants that tend to co-purify with DNA from samples that are high in organic matter can be removed by methods such as agarose gel electrophoresis or column chromatography[6,7]. Direct isolation might also capture DNA from virus particles or free DNA from dead cells. When these non-cellular DNAs are the intended target, they can be directly solubilized and concentrated without lysis of cells in the sample[8,9].

The techniques that are used to lyse cells might also affect the composition of environmental DNA libraries, as the harsh lysis methods that are necessary to extract DNA from every organism will cause degradation of the DNA from some organisms[7]. Hard-to-lyse cells, such as GRAM-POSITIVE BACTERIA, might therefore be underrepresented or overrepresented in environmental DNA preparations[10]. The desire for complete lysis often must be balanced with the need for high-quality DNA, especially when preparing DNA of high molecular weight for large-insert libraries[11,12].

Once DNA has been obtained, it can be directly cloned into small-insert vectors for high-throughput sequencing (for example, see the Joint Genome

Institute Protocols web site in the Online links box). Alternatively, it can be cloned into large-insert libraries and screened for clones with activities or genes of interest, which are then subcloned and sequenced.

Insights into 'inaccessible' organisms

The first forays in sequencing natural samples aimed to characterize the genomes of organisms that occur in tight association with one or more species, and therefore cannot be easily studied in isolation. Here the challenge is to extract the relevant sequence from a mixed-species library, which might contain only a small fraction of clones from the target species. Various pre- and post-sequencing 'sifting' techniques have allowed the genomic characterization of organisms that cannot be cultivated, such as obligate pathogens and symbionts, and even long-extinct species.

16S ribosomal RNA: a launch pad for novel prokaryotic genomes. The genomic study of natural communities has been largely driven by interest in the ~99% of microbes that are not easily isolated in culture. These species are identified by their 16S/18S small subunit ribosomal RNA (SSU rRNA) genes, which are commonly used as phylogenetic markers because every cellular organism contains these genes and almost all gene variants can be amplified by standard sets of degenerate primers (BOX 1). Several investigators have used rRNA genes as a starting point to explore the genomes of uncultivated microbes through large-insert clone sequencing. One such 'PHYLOGENETIC ANCHORING' study led to the discovery of proteorhodopsin, a type of light-harvesting protein, in oceanic bacteria — a surprise not only because these microbes were previously believed to depend on organic matter, not light, as an energy source, but also because rhodopsin-like proteins had never before been identified in the bacterial domain[13–15]. Further studies have provided glimpses of the genomes of several other uncultivated prokaryotes, including Crenarchaeota[11,16–18] and Acidobacteria[10] from many habitats. In some cases these sequences have provided evidence for unexpected biological functions[11] or HORIZONTAL GENE TRANSFERS[19].

Host-associated bacteria: genomic insights into pathogenesis and symbiosis. Although discussion of uncultivated microbes most often brings environmental organisms to mind, several of the uncultivated microbes for which the genomes have already been sequenced are obligate pathogens or symbionts[20–32]. The amenability of these host-associated microbes to physical separation makes them well-suited to this approach (TABLE 1), which is in contrast to organisms that reside in complex environmental communities. The first complete genome of an uncultured bacterium, the syphilis spirochete *Treponema pallidum*, was released in 1998

Table 1 | **Assembled genomes of uncultivated microbes**

Species	Genome size	Host or habitat	Separation technique	Refs
Treponema pallidum	1.1 Mb	Human, rabbit	Dissection, differential lysis	20
Rickettsia prowazekii	1.1 Mb	Human, chicken	Differential centrifugation	21
Mycobacterium leprae	3.3 Mb	Human, armadillo	Gradient centrifugation	22
Tropheryma whipplei	0.9 Mb	Human	Differential centrifugation	23
Buchnera aphidicola str. APS	0.6 Mb	Aphid (*Acyrthosiphon pisum*)	Dissection, differential lysis, filtration	24
Buchnera aphidicola str. Sg	0.6 Mb	Aphid (*Schizaphis graminum*)	Gradient centrifugation	25
Wigglesworthia glossinidia brevipalpis	0.7 Mb	Tsetse fly (*Glossina brevipalpis*)	Dissection, differential lysis	26
Blochmannia floridanus	0.7 Mb	Carpenter ants	Differential lysis	27
Buchnera aphidicola str. Bp	0.6 Mb	Aphid (*Baizongia pistaciae*)	Differential lysis, filtration	28
Wolbachia pipientis wMel	1.27 Mb	Fly (*Drosophila melanogaster*)	Differential lysis, pulsed-field gel electrophoresis	29
Wolbachia pipientis wAna	1.4 Mb	Fly (*Drosophila ananassae*)	None	30
Wolbachia pipientis wBm	1.1 Mb	Parasitic nematode worm (*Brugia malayi*)	BAC library screening	31
Phytoplasma asteris, line OY-M	0.9 MB	Plants and leafhoppers	Differential lysis, pulsed-field gel electrophoresis	32
Nanoarchaeum equitans	0.5 Mb	*Ignicoccus* sp. co-culture	Differential centrifugation	59
Ferroplasma acidarmanus type II	1.8 Mb	Acid-mine biofilm	None	45
Leptospirillum sp. Group II	2.2 Mb	Acid-mine biofilm	None	45
Burkholderia sp.	~8.8 Mb	Sargasso Sea	Filtration	4
Shewanella sp.	~5 Mb	Sargasso Sea	Filtration	4

PHYLOGENETIC ANCHORING
A technique that involves screening large-insert libraries made from environmental DNA for clones that contain phylogenetic marker genes, and sequencing those clones in their entirety.

HORIZONTAL GENE TRANSFER
The transfer of genetic material between the genomes of two organisms that does not occur through parent–progeny routes.

— a landmark in genome sequencing[20]. Although the bacterial origin of syphilis was recognized a century ago, the infectious agent has never been isolated in continuous culture. The DNA that was used for sequencing the intracellular pathogen was obtained from the testes of infected rabbits — some 400 of them — by a series of lysis and centrifugation steps that eventually resulted in an essentially pure bacterial preparation (TABLE 1). Sequence analysis immediately identified potential contributors to virulence and aided the development of DNA-based diagnostics[33].

A year and a half of painstaking growth in co-culture with human fibroblasts was necessary to obtain sufficient DNA to sequence the genome of the Whipple disease bacterium *Tropheryma whipplei*. The sequence revealed deficiencies that indicated an explanation for the failure to propagate in AXENIC culture. Based on these genomic insights, Renesto *et al.* used a standard tissue-culture medium, supplemented with amino acids that were implicated by the sequence analysis, to successfully cultivate *T. whipplei* in the absence of host cells, shortening their doubling time by an order of magnitude[34]. This is one of many cases in which DNA sequence information has been used to improve culture techniques, diagnostics and therapies for fastidious organisms[35–37].

Several genomes of obligate intracellular symbionts, primarily from insect hosts, that could not be grown by conventional means have also been obtained by various separation and purification methods (TABLE 1). The first was *Buchnera aphidicola*[24], a relative of *Escherichia coli* that provides nutrients to supplement its aphid host's restricted diet of plant sap. Bacteriomes — specialized symbiont-harbouring organs — were isolated from 2,000 aphids by dissection before being crushed and filtered, resulting in virtually pure *B. aphidicola* cells for DNA isolation. Symbionts of tsetse flies, fruitflies, carpenter ants, a nematode and two other aphid species have since had their genomes sequenced, as has one uncultured plant pathogen[25–29,31,32]. In each project, techniques such as dissection, DIFFERENTIAL LYSIS and PULSED-FIELD GEL ELECTROPHORESIS, often in combination, have helped to enrich for prokaryotic material (TABLE 1); where reported, between 5% and 47% of the sequences were host-derived[27–29]. Another essentially complete symbiont genome from a member of the genus *Wolbachia* recently emerged as a by-product of a metazoan genome project, as the sequencing libraries were constructed from symbiont-harbouring whole embryos[30].

Palaeogenomics. Evolutionary biology depends heavily on DNA sequence data to reconstruct evolutionary pathways, but these molecular trees are limited to the modern species that lie at the ends of the branches for which DNA is readily available. Phylogenetic placement and proposed phenotypes of the organisms at the branching nodes, or the branches that terminate before the modern era, are based primarily on morphological examination of fossilized specimens. The ability to sequence genomes from ancient organisms would offer a 'genomic time machine' to study these poorly characterized species.

When an animal dies, its tissues are quickly exploited as an organic nutrient source by various creatures, particularly single-celled microbes. Rarely, conditions are such that the carcass escapes total decomposition and parts, particularly bone, remain preserved; however, the DNA contained therein is not only damaged and fragmented, but also mixed with the genomes of the abundant opportunistic microbes that have invaded the tissue. Nonetheless, gentle and rigorously sterile DNA-isolation procedures have allowed the generation of verifiable mitochondrial and nuclear sequence from material such as bones, teeth and coprolites (fossilized fecal material) dating back to as long as 50,000 years ago[38,39]. These studies, which rely mainly on PCR-amplified mitochondrial sequence, have been used to resolve phylogenetic relationships between extinct and modern animals[40]. Mitochondria are present in more than 1,000 copies per cell and are therefore relatively easily amplified; the single-copy nuclear genome, which could offer far more phenotypic information, has remained minimally explored owing to technical hurdles[41,42]. Low-cost high-throughput sequencing, coupled with a METAGENOMIC approach, now provides a means to access the nuclear genomes of extinct organisms without amplification. This was recently applied to the analysis of the cave bear, *Ursus spelaeus*, which is a relative of modern brown and black bears that lived in caves throughout Europe in the Late Pleistocene, but became extinct tens of thousands of years ago. The investigators exploited a metagenomic strategy to demonstrate the presence of verifiable cave bear sequence in libraries that were created by directly cloning DNA extracted from 40,000-year-old bones[43]. A library-construction protocol that involved neither lysis nor shearing allowed the cloning of end-repaired ancient DNA isolated from cave bear bone and tooth samples. Although the cave bear sequence constituted a mere 1–5% of the libraries that were described by Noonan *et al.* these sequences were readily identified by their high level of sequence identity to a related carnivore, the dog, for which the genome is fully sequenced[44] (FIG. 1). Approximately 27 kb of putative cave bear sequence was obtained, and PCR amplification of orthologous sequences from modern black, brown and polar bears verified their origin and allowed the reconstruction of a phylogenetic tree that is congruent with that based on mitochondrial sequences. Modern human contamination from laboratory personnel, which accounted for a surprisingly low 0.05% of clones, was easily identified as this proof-of-principle study focused on a species that is readily distinguishable from modern human.

These techniques open up the possibility of genome projects that target extinct species, and could revolutionize palaeobiology. Our closest hominid relatives, the Neanderthals, diverged from modern humans roughly 500,000 years ago but survived until the Late Pleistocene, and numerous Neanderthal remains that have ages that are comparable to the sequenced cave bear samples

AXENIC
A pure culture of a single species of microorganism.

DIFFERENTIAL LYSIS
A technique that uses conditions that will only lyse certain cells so that the DNA from those cells can be isolated from other cells in a community.

PULSED-FIELD GEL ELECTROPHORESIS
The use of pulsed electrical fields of alternating polarity to separate large fragments of DNA.

METAGENOMIC
A term used to describe techniques that characterize the genomes of whole communities of organisms rather than individual species.

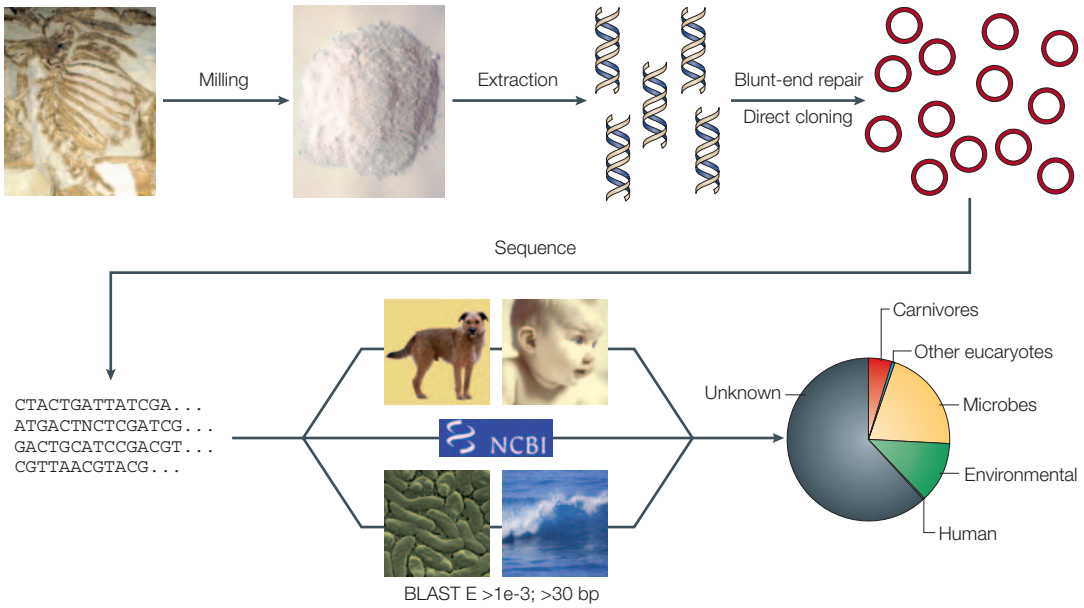


Figure 1 | **Sequencing of ancient DNA.** Genomic sequence from extinct organisms can be obtained from the DNA in ancient remains such as bone. Bones are milled into powder and immersed in a solution to extract the DNA. The damaged ends of the DNA molecules are then repaired enzymatically and cloned into a sequencing vector. The clones are sequenced using standard protocols, and the probable species of origin is determined by BLAST analysis. In the study of Pleistocene cave bears by Noonan *et al.*[43], up to 5% of the clones found their closest match in the genome of the dog, a carnivore that is closely related to bears. Only a few (~0.05%) of the reads were of human origin, whereas 10–20% had significant matches only to environmental sequences.

have been found. By providing sequence from another hominid, the Neanderthal genome could define human-specific sequences and expand our knowledge of the biology of both *Homo sapiens* and Neanderthals.

High-throughput shotgun sequencing of environmental samples. Environmental samples are much more complex than single organisms, as they might contain tens, hundreds or even thousands of distinct species, and were therefore until recently widely considered unsuitable for high-throughput sequencing. Similar concerns once accompanied the application of WHOLE-GENOME SHOTGUN (WGS) sequencing to large genomes, as it was thought that assembly of WGS reads into chromosomes and genomes would prove too computationally complex. However WGS has proved to be the most efficient and effective approach to generating complete genomes, both large and small, mainly as a result of computational advances. In the case of environmental genomics, tools for analysis have once again risen to the task, allowing the simultaneous study of whole ensembles of genomes through high-throughput sequencing. A new perspective, in which genes and genomes are viewed as subunits of a larger whole, is changing the way that we understand evolution and adaptation.

The first large-scale environmental shotgun-sequencing project interrogated the organisms that make up an acid-mine biofilm[45]. Acid-mine drainage is an environmentally devastating consequence of commercial mining that results from the production of sulphuric acid when pyrite (FeS_2) is exposed to air and water during mining operations. Microorganisms have long been recognized as important players in this process, as the rate-limiting step of ferric (Fe^{3+}) ion regeneration is slow under sterile conditions but can be greatly accelerated by microbes that derive energy from the reaction (chemolithotrophs)[46]. Microbial communities flourish under these seemingly hostile conditions, forming extensive underwater streamers and floating biofilms that are anchored in pyritic sediments, but are typically of relatively low diversity as few organisms can tolerate the extreme acidity.

To address the physiology of the uncultivated microbes in one mine, Tyson *et al.* built a short-insert genomic library from biofilm DNA and generated 76.2 million base pairs of sequence from the resident bacteria and archaeans[45]. From this they assembled near-complete genomes for two community members and partial genomes for three more, allowing metabolic reconstruction to assess the role of each individual organism. Interestingly, one organism, an uncultivated *Leptospirillum* group III species, was the only member of this community that possessed the genes for nitrogen fixation. As this process is essential in such a nutrient-limited environment, this low-abundance species seems to be a linchpin for the whole community and, theoretically, a potential biological target for clean-up efforts.

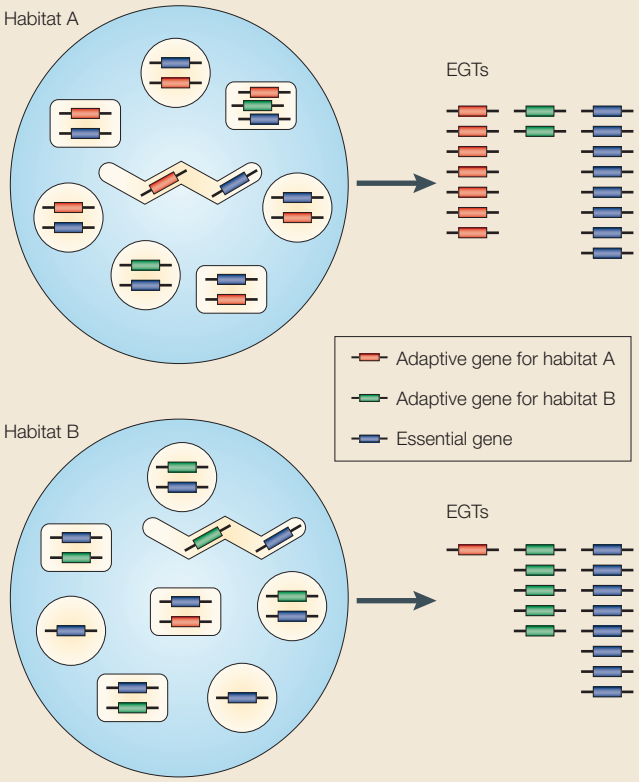
Another study reported the metagenomic sequencing of the surface-water microbial community of the Sargasso Sea, which is a body of low-nutrient water in the North Atlantic[4]. Planktonic microbes were collected from many locations and extracted DNA was used to construct 7 independent libraries, from which a total of more than 1.6 Gb of DNA sequence

WHOLE-GENOME SHOTGUN
An approach to genomic sequencing that involves breaking the DNA into small pieces and cloning them into vectors, followed by sequencing the clones at random.

Box 2 | **Environmental gene tags**

Each organism in a community has a unique set of genes in its genome; the combined genomes of all the community members make up the metagenome. Essential genes are present in each individual genome, regardless of environment, and will therefore occur frequently in the metagenome. Among the non-essential genes, those that are adaptive for a particular niche will occur in the genomes of many organisms in that environment, whereas those that are not adaptive might occur at low abundances.

Environmental gene tags (EGTs) are short sequences from the DNA of microbial communities that contain fragments of functional genes. Each EGT derives from a different member of the community, but genes that are important for survival and adaptation will be present in many genomes (possibly in more than one copy) and will therefore occur repeatedly in the EGT data. When the gene abundances in the EGT data are compared between habitats, genes that are adaptive in only one context are more abundant in that habitat (see figure).



was generated. Reflecting the unexpected complexity of the Sargasso Sea planktonic community, just 3% of this sequence was accounted for at 3x coverage or more — although this quantity of raw sequence would be sufficient to complete as many as 50 prokaryotic genomes at 8x coverage. More than 1.2 million genes were found to have significant similarity to database entries. Although less than a third could be assigned tentative cellular roles, some functions stood out, including numerous rhodopsin-related genes and genes involved in phosphorus uptake and metabolism, which is consistent with the need to efficiently use the plentiful sunlight and limited phosphate that is available in this environment[47]. However, most of the predicted genes in the Sargasso Sea data could not be definitively linked to particular phylogenetic groups, much less individual species. These data have since been mined to identify a range of genes, including those that encode iron–sulphur proteins, chitinases, proteorhodopsins and electron-transport proteins[48–51]. Each of these studies has identified genes that are highly divergent from known family members, highlighting the novelty of environmental sequences compared with genome sequences from cultured isolates.

Whereas acid mines and the Sargasso Sea represent relatively nutrient-poor environments, a recent study by Tringe *et al.*[52] explored two nutrient-rich environments: agricultural soil and deep-sea whale skeletons, also known as 'whale falls', which sustain thriving communities of microorganisms and macroorganisms as they decompose[53]. The combination of these environments with the previously sequenced samples spans a wide range of environmental variables such as temperature, pH and illumination, providing a rich testing ground for comparative analysis. Just as comparative genomics forms the foundation for most genome annotation efforts, it was reasoned that patterns of gene abundance between environments would enhance understanding of both the environments and the gene products.

Genomic sequencing of complex, nutrient-rich samples did not result in assembled genomes — indeed, it was estimated that several billion bases of sequence would need to be generated from a complex environment, such as soil, before genomes would begin to assemble. It did however identify gene families that are important for survival in the environments that were sampled. In this gene-centric approach, each sequence obtained was called an environmental gene tag (EGT), because it contained a snippet of sequence that potentially encoded a protein adapted to that environment (BOX 2). Predicted genes on the EGTs from each sample were compared with each other and with sequences from previous environmental sequencing projects[4,45]. Several characterized and uncharacterized orthologous groups, functional modules or biochemical processes emerged that were unevenly distributed across the samples[52]. This provided an EGT 'fingerprint' of each environment and demonstrated that similar environments, such as two whale skeletons that are 8,000 miles apart on the ocean floor, have similar gene contents.

The analysis of functions that are overrepresented in particular niches provided unique insights into the demands that are placed on the organisms living there. One of the most significant disparities in gene distribution to emerge from this analysis was the overabundance of rhodopsin-like proteins in the Sargasso Sea compared with non-illuminated environments. Similarly, as might be predicted in hindsight, many homologues of cellobiose phosphorylase, an enzyme that is involved in the breakdown of plant material, were found in the soil sample, taken near a silage bunker, but not in the other samples. A preponderance of sodium transport and osmoregulation proteins in all the marine samples, both surface and deep sea, was consistent with the high sodium content of seawater. The soil sample, by contrast, contained more potassium transporters. Biochemical analysis revealed that potassium ions outnumbered sodium in the sample seven to one. Overall, variations in gene distribution were most evident for transporters and metabolic enzymes — the molecules that are most involved in interacting with, and presumably adapting to, the environment. The many uncharacterized orthologous groups that have highly skewed distributions across samples might function in niche adaptation, and might therefore be promising candidates for future investigations. With these comparative tools in hand, researchers can now investigate the factors that influence microbial colonization or the changes that occur in environments under stress, without the constraints on diversity that are created by the need to assemble genomes.

Future directions

The goals of metagenomic projects vary considerably, from characterizing one particular species to understanding the dynamics of a whole community. Although the 'difficult-to-access' genome projects described here might seem to share little in common with environmental projects that examine complex communities, many of the methods and challenges overlap. These two previously separate fields are rapidly converging in several metagenomic projects that now target either individual members of free-living communities, such as the marine Crenarchaeota[54], or entire communities of symbiotic organisms, such as the syntrophic consortium that inhabits the marine oligochaete *Olavius olgarvensis*[55]. (For information on these and other ongoing projects at the JGI, see the Community Sequencing Program Sequencing Plans for 2005 web page in the Online links box.) A 'second human genome project' has even been proposed to sequence the genomes of the human-associated microbiota[56].

We have described many innovations that have improved our ability to study inaccessible genomes. However, the current methods of DNA isolation, library construction, sequence assembly and bioinformatic analysis are all still optimized for single-genome analysis and will probably need to be modified for application to metagenomic projects.

DNA isolation and library construction. The methods that are used to isolate DNA from mixed samples and construct libraries substantially affect the results obtained, as cells differ in their sensitivity to lysis and DNA 'cloneability' varies widely. For environmental samples, particular effort has been devoted to obtaining DNA that is representative of all the organisms present to best study the community as a whole. These representative libraries are effective tools for community overviews and for characterizing the dominant activities in an environment[52]. However, when complete genome sequences are desired, representative libraries are an inefficient means of sequencing non-dominant community members; for example, one organism in the Sargasso Sea study was sequenced at 21x coverage[4].

Several techniques have been used to normalize or enrich environmental libraries for various applications, based on generic properties such as cell size or DNA composition. Filtration has already been mentioned as a means of separating cells on the basis of size, particularly for separating prokaryotes from eukaryotes; it has also been used to separate multicellular consortia from individual cells[57]. The separation of DNA on bisbenzimidazole gradients allows fractionation that is based on GC content, which exploits the change in buoyant density that occurs when bisbenzimidazole binds to adenine and thymidine[58]. Other techniques that have been applied to host-associated microbes include differential centrifugation[59], DENSITY GRADIENTS[25,57], differential lysis[20], pulsed-field gel electrophoresis[32] and selective use of restriction enzymes[60].

In some cases, a particular organism or group of organisms in a community is of interest; for example, those that carry out a particular metabolic process or members of an uncharacterized phylogenetic group. Successful targeting of these organisms could significantly reduce the amount of sequence needed for genome coverage and simplify assembly. STABLE ISOTOPE PROBING (SIP) holds promise as a means to obtain DNA from organisms that can metabolize a particular substrate, and might be a valuable method for community fractionation[61]. FLOW CYTOMETRY is a highly specific method for isolating organisms on the basis of viability[62], membrane properties[63], surface protein expression[64] or SSU rRNA sequence[65]. Finally, AFFINITY PURIFICATION might hold promise for separating some groups[66,67] on the basis of cell-wall characteristics or extracellular markers. Building libraries from such enriched DNA will greatly improve sequencing efficiency compared with whole-community libraries.

Whole-genome amplification through ISOTHERMAL STRAND DISPLACEMENT could dramatically increase the possibilities for sequencing unculturable organisms by significantly reducing the amount of starting material required for library construction. DNA from prokaryotic and eukaryotic cells has been amplified by this technique and used for various PCR-based and hybridization-based genomic analyses[68]. Encouraging results were recently reported for a metagenomic

DENSITY GRADIENT
This occurs in a solution in which the concentration of the solute is lowest at the top and gradually becomes more dense towards the bottom.

STABLE ISOTOPE PROBING
A technique that relies on the incorporation of a substrate that is enriched in a stable isotope, such as ^{13}C, to identify microorganisms that can metabolize that substrate.

FLOW CYTOMETRY
A technique that measures the fluorescence of individual cells as they pass through a laser beam in an individual stream.

AFFINITY PURIFICATION
A technique for purifying cells or molecules that is based on specific binding to a protein or other molecule that has been immobilized on a solid substrate, such as beads or a column.

ISOTHERMAL STRAND DISPLACEMENT
A DNA amplification technique that uses rolling-circle amplification with ϕ29 DNA polymerase to generate large quantities of DNA without thermal cycling.

sample, in which PCR results from amplified and unamplified DNA were comparable[69]. Short-insert shotgun-sequencing libraries have also been constructed from whole-genome amplified samples[70,71]. However, a high rate of sequencing artefacts has so far precluded genome assemblies on the basis of these libraries (P. Richardson, personal communication).

Library construction is a potentially important source of bias, as some genome segments are uncloneable and/or lethal to *E. coli*. New, highly parallel non-Sanger sequencing technologies that are already being marketed, such as PYROSEQUENCING, obviate the need for libraries of any sort[72]. By eliminating this source of bias and also decreasing time, effort and expense, these technologies could have an important effect on the field; although other obstacles such as short read lengths will still need to be overcome.

Data analysis. One of the most pressing issues in metagenomics is genome assembly, which is crucial for some types of genomic analysis. The most basic obstacle to assembly is simply the cost of achieving sufficient sequence coverage of a single microbe in a community that might contain hundreds of species. However, given the decreasing cost of sequencing, this might soon be less of a problem. Another concern is how assembly algorithms will perform when confronted with mixed data from multiple species. Fortunately, experience suggests that cross-species assemblies are not a common occurrence[45], except in the case of highly conserved sequences such as those of rRNAs[73]. Perhaps the most serious challenge in assembling genomes from metagenomic data is population heterogeneity, in the form of sequence polymorphisms and genomic rearrangements. Assembly algorithms seem to be robust to sequence polymorphisms[28,45,74], although very high polymorphism can interfere with proper assembly, especially in complex genomes[75]. Genomic rearrangements, however, might require serious rethinking of the meaning and purpose of genome assembly[76]. It is not yet clear what level of heterogeneity is 'typical': in the limited set of communities that have been explored, some populations are almost clonal, some are highly polymorphic, and some contain extensive insertions, deletions and translocations[4,45,57]. It will be interesting to see whether heterogeneity correlates with features such as growth rate, competition or community stability.

Once sequences have been generated, be they whole genomes, large scaffolds or individual reads, we often want to assign them to phylogenetic groups. For closed or nearly closed genomes, scaffold assignment is straightforward because functional genes are directly linked to phylogenetic markers such as 16S rRNA. But even under optimal conditions each genome might be divided into multiple scaffolds, and many sequences, particularly those from low-abundance community members, will remain in small CONTIGS or unassembled reads that lack obvious marker genes. The simplest method of taxonomic assignment, best BLAST hit, should be used with caution: it is only reliable when close relatives are available for comparison, and is essentially useless when no relatives have been fully sequenced[77]. Other features that have been used to 'bin' scaffolds or contigs into taxonomic groups include GC content and oligonucleotide frequency, coverage depth, and similarity to sequenced genomes[4,43,45,78].

Gene calling is in its infancy in metagenomic data because the data are fragmented, heterogeneous and abundant. Homology-based methods are accurate but not sensitive, particularly for genomes that lack sequenced relatives, and will always miss novel genes, which are potentially the most interesting. *Ab initio* methods (based on DNA sequence alone) can predict novel genes, but training of these methods is optimally carried out on complete genomes, and false-positive rates might be high even for assembled genomes[79]. One way to circumvent this problem is to use a sample of sequenced genomes as a training set, preferably one that has a similar phylogenetic range as the species in the sample[52]. Further improvements to gene prediction methods are urgently required. Validation of potential novel genes can be obtained by sequence clustering: predicted proteins that have homologues within the data set are likely to be valid[4].

Gene annotation is also a challenge for metagenomic projects, as the amount of data generated is likely to be too large for manual annotation. Fortunately, there are several high-quality automated annotation tools for complete microbial genomes, such as ERGO[80], GenDB[81] and PRIAM[82]. In general, these can be adapted with minimal effort to metagenomic data sets. However, accuracy is a concern as no automated methods can fully replace manual annotation. The greatest improvements in accuracy are likely to result from the further production of high-quality complete genomes, particularly in phylogenetic groups, such as Chloroflexi and Acidobacteria, that are well-represented in the environment but poorly represented in sequence databases[83]. Such high-quality genome data will provide better substrates for homology searches.

Conclusions

Genome sequencing has made invaluable contributions to evolutionary biology, medicine and agricultural science, and is rapidly being adapted to studies of organisms in their natural habitats. Such studies offer several unique benefits beyond those of traditional genomic studies of clonal laboratory strains.

The most obvious benefit of sequencing DNA from natural samples is the ability to access a much wider range of genomes. Many organisms fail to 'reproduce in captivity' and therefore cannot be subjected to laboratory manipulation and genomic study. These include not only exotic groups (for example, Nanoarchaeota), but many close relatives of cultivable microbes. Other species are extinct and therefore cannot provide clean material for DNA isolation — most notably, ancient hominids such as the Neanderthals, which might soon be the target of their own human genome project.

A less immediately apparent advantage of this technique is the ability to capture the genomic diversity within a natural population. Although DNA sequence

PYROSEQUENCING
A DNA sequencing technique that relies on detection of pyrophosphate release on nucleotide incorporation rather than chain termination with dideoxynucleotides.

CONTIG
A continuous stretch of DNA sequence that is assembled from multiple independent sequencing reads.

from a clonal strain is easier to generate and assemble, an individual genome represents a single snapshot of the population from which it derives. Both clonal strain sequencing and environmental studies reveal that there can be substantial variation in gene content, gene order and nucleotide sequence even within populations that are thought of as a single species[84–86]. Sequences from natural samples reflect this variation and reveal the prevalence of specific subgroups.

By offering access to genomes of hard-to-study organisms, metagenomics and its offshoots have advanced our understanding of species interrelationships, environmental-niche adaptation and human evolutionary history. Technologies that are now under development will continue to lower the barriers to genome sequencing, allowing the study of ever more scarce and complex samples, and vastly expanding the range of species on the genomics radar.

1. Boffelli, D., Nobrega, M. A. & Rubin, E. M. Comparative genomics at the vertebrate extremes. *Nature Rev. Genet.* **5**, 456–465 (2004).
2. Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nature Rev. Genet.* **5**, 335–344 (2004).
3. Somerville, C. C., Knight, I. T., Straube, W. L. & Colwell, R. R. Simple, rapid method for direct isolation of nucleic acids from aquatic environments. *Appl. Environ. Microbiol.* **55**, 548–554 (1989).
4. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
This project to sequence the entire metagenome of the Sargasso Sea surface waters revealed unexpected community complexity and sequence diversity.
5. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* **99**, 14250–14255 (2002).
6. LaMontagne, M. G., Michel, F. C. Jr, Holden, P. A. & Reddy, C. A. Evaluation of extraction and purification methods for obtaining PCR-amplifiable DNA from compost for microbial community analysis. *J. Microbiol. Methods* **49**, 255–264 (2002).
7. von Wintzingerode, F., Gobel, U. B. & Stackebrandt, E. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.* **21**, 213–229 (1997).
8. Kolman, C. J. & Tuross, N. Ancient DNA analysis of human populations. *Am. J. Phys. Anthropol.* **111**, 5–23 (2000).
9. Breitbart, M. *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223 (2003).
10. Liles, M. R., Manske, B. F., Bintrim, S. B., Handelsman, J. & Goodman, R. M. A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl. Environ. Microbiol.* **69**, 2684–2691 (2003).
11. Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & DeLong, E. F. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.* **178**, 591–599 (1996).
12. Berry, A. E., Chiocchini, C., Selby, T., Sosio, M. & Wellington, E. M. Isolation of high molecular weight DNA from soil for cloning into BAC vectors. *FEMS Microbiol. Lett.* **223**, 15–20 (2003).
13. Suzuki, M. T., Beja, O., Taylor, L. T. & Delong, E. F. Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environ. Microbiol.* **3**, 323–331 (2001).
14. Schmidt, T. M., DeLong, E. F. & Pace, N. R. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* **173**, 4371–4378 (1991).
15. Beja, O. *et al.* Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902–1906 (2000).
A seminal paper in metagenomics, this study identified a novel protein on a BAC from the uncultivated SAR86 group of bacterioplankton that was later revealed to represent a previously unknown, widespread group of ecologically important light-harvesting proteins.
16. Beja, O. *et al.* Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl. Environ. Microbiol.* **68**, 335–345 (2002).
17. Lopez-Garcia, P., Brochier, C., Moreira, D. & Rodriguez-Valera, F. Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ. Microbiol.* **6**, 19–34 (2004).
18. Quaiser, A. *et al.* First insight into the genome of an uncultivated crenarchaeote from soil. *Environ. Microbiol.* **4**, 603–611 (2002).
19. Quaiser, A. *et al.* Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Mol. Microbiol.* **50**, 563–575 (2003).
20. Fraser, C. M. *et al.* Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388 (1998).
This paper reported the first genome sequence of a microbe that could not be grown in continuous pure culture.
21. Andersson, S. G. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
22. Cole, S. T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).
23. Bentley, S. D. *et al.* Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. *Lancet* **361**, 637–644 (2003).
24. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**, 81–86 (2000).
This paper reported the first complete genome sequence of an uncultivated intracellular symbiont and revealed significant genome reduction.
25. Tamas, I. *et al.* 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376–2379 (2002).
26. Akman, L. *et al.* Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nature Genet.* **32**, 402–407 (2002).
27. Gil, R. *et al.* The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc. Natl Acad. Sci. USA* **100**, 9388–9393 (2003).
28. van Ham, R. C. *et al.* Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl Acad. Sci. USA* **100**, 581–586 (2003).
29. Wu, M. *et al.* Phylogenomics of the reproductive parasite *Wolbachia pipientis wMel*: a streamlined genome overrun by mobile genetic elements. *PLoS Biol.* **2**, e69 (2004).
30. Salzberg, S. L. *et al.* Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. *Genome Biol.* **6**, R23 (2005).
31. Foster, J. *et al.* The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* **3**, e121 (2005).
32. Oshima, K. *et al.* Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nature Genet.* **36**, 27–29 (2004).
33. Liu, H., Rodes, B., Chen, C. Y. & Steiner, B. New tests for syphilis: rational design of a PCR method for detection of *Treponema pallidum* in clinical specimens using unique regions of the DNA polymerase I gene. *J. Clin. Microbiol.* **39**, 1941–1946 (2001).
34. Renesto, P. *et al.* Genome-based design of a cell-free culture medium for *Tropheryma whipplei*. *Lancet* **362**, 447–449 (2003).
Using information on *T. whipplei's* metabolic deficiencies revealed by its genome sequence, these investigators successfully created the first pure culture system for this organism and reduced its *in vitro* generation time by a factor of 15.
35. Fenollar, F. & Raoult, D. Molecular genetic methods for the diagnosis of fastidious microorganisms. *Apmis* **112**, 785–807 (2004).
36. Ogata, H. & Claverie, J. M. Metagrowth: a new resource for the building of metabolic hypotheses in microbiology. *Nucleic Acids Res.* **33** (database issue), D321–D324 (2005).
37. Lemos, E. G., Alves, L. M. & Campanharo, J. C. Genomics-based design of defined growth media for the plant pathogen *Xylella fastidiosa*. *FEMS Microbiol. Lett.* **219**, 39–45 (2003).
38. Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Paabo, S. Ancient DNA. *Nature Rev. Genet.* **2**, 353–359 (2001).
39. Cooper, A. *et al.* Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* **409**, 704–707 (2001).
40. Hofreiter, M. *et al.* Evidence for reproductive isolation between cave bear populations. *Curr. Biol.* **14**, 40–43 (2004).
41. Poinar, H., Kuch, M., McDonald, G., Martin, P. & Paabo, S. Nuclear gene sequences from a Late Pleistocene sloth coprolite. *Curr. Biol.* **13**, 1150–1152 (2003).
42. Greenwood, A. D., Capelli, C., Possnert, G. & Paabo, S. Nuclear DNA sequences from Late Pleistocene megafauna. *Mol. Biol. Evol.* **16**, 1466–1473 (1999).
43. Noonan, J. P. *et al.* Genomic sequencing of Pleistocene cave bears. *Science* **309**, 597–599 (2005).
The first report of a DNA sequence from an extinct species that was generated without PCR amplification.
44. Kirkness, E. F. *et al.* The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898–1903 (2003).
45. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
This paper reports the first assembled genomes to emerge from shotgun sequencing of environmental samples, allowing metabolic reconstruction of community members.
46. Johnson, D. B. & Hallberg, K. B. The microbiology of acidic mine waters. *Res. Microbiol.* **154**, 466–473 (2003).
47. Wu, J., Sunda, W., Boyle, E. A. & Karl, D. M. Phosphate depletion in the western North Atlantic Ocean. *Science* **289**, 759–762 (2000).
48. McDonald, A. E. & Vanlerberghe, G. C. Alternative oxidase and plastoquinol terminal oxidase in marine prokaryotes of the Sargasso Sea. *Gene* **349**, 15–24 (2005).
49. Sabehi, G., Beja, O., Suzuki, M. T., Preston, C. M. & DeLong, E. F. Different SAR86 subgroups harbour divergent proteorhodopsins. *Environ. Microbiol.* **6**, 903–910 (2004).
50. Meyer, J. Miraculous catch of iron–sulfur protein sequences in the Sargasso Sea. *FEBS Lett.* **570**, 1–6 (2004).
51. LeCleir, G. R., Buchan, A. & Hollibaugh, J. T. Chitinase gene sequences retrieved from diverse aquatic habitats reveal environment-specific distributions. *Appl Environ Microbiol* **70**, 6977–83 (2004).
52. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).
This study revealed that differences in gene content among communities are apparent even in unassembled genomic data.
53. Smith, C. R. & Baco, A. R. in *Oceanography and Marine Biology: an Annual Review* Vol. 41 (eds Gibson, R. N. & Atkinson, R. J. A.) 311–354 (Taylor & Francis, London, 2003).
54. Karner, M. B., DeLong, E. F. & Karl, D. M. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**, 507–510 (2001).
55. Dubilier, N. *et al.* Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. *Nature* **411**, 298–302 (2001).
56. Relman, D. A. & Falkow, S. The meaning and impact of the human genome sequence for microbiology. *Trends Microbiol.* **9**, 206–208 (2001).
57. Hallam, S. J. *et al.* Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science* **305**, 1457–1462 (2004).
A genomic analysis of uncultured Archaea from deep-sea sediments that provided evidence for a 'reverse-methanogenesis' mechanism of anaerobic methane oxidation.

58. Nusslein, K. & Tiedje, J. M. Characterization of the dominant and rare members of a young Hawaiian soil bacterial community with small-subunit ribosomal DNA amplified from DNA fractionated on the basis of its guanine and cytosine composition. *Appl. Environ. Microbiol.* **64**, 1283–1289 (1998).
59. Waters, E. *et al.* The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proc. Natl Acad. Sci. USA* **100**, 12984–12988 (2003).
60. Garcia-Chapa, M., Batlle, A., Rekab, D., Rosquete, M. R. & Firrao, G. PCR-mediated whole genome amplification of phytoplasmas. *J. Microbiol. Methods* **56**, 231–242 (2004).
61. Dumont, M. G. & Murrell, J. C. Stable isotope probing — linking microbial identity to function. *Nature Rev. Microbiol.* **3**, 499–504 (2005).
62. Bernard, L. *et al.* A new approach to determine the genetic diversity of viable and active bacteria in aquatic ecosystems. *Cytometry* **43**, 314–321 (2001).
63. Park, H. S., Schumacher, R. & Kilbane, J. J. 2nd. New method to characterize microbial diversity using flow cytometry. *J. Ind. Microbiol. Biotechnol.* **32**, 94–102 (2005).
64. Gu, F. *et al. In situ* and non-invasive detection of specific bacterial species in oral biofilms using fluorescently labeled monoclonal antibodies. *J. Microbiol. Methods* **62**, 145–160 (2005).
65. Sekar, R., Fuchs, B. M., Amann, R. & Pernthaler, J. Flow sorting of marine bacterioplankton after fluorescence *in situ* hybridization. *Appl. Environ. Microbiol.* **70**, 6210–6219 (2004).
66. Lin, Y. S., Tsai, P. J., Weng, M. F. & Chen, Y. C. Affinity capture using vancomycin-bound magnetic nanoparticles for the MALDI-MS analysis of bacteria. *Anal. Chem.* **77**, 1753–1760 (2005).
67. Bundy, J. L. & Fenselau, C. Lectin and carbohydrate affinity capture surfaces for mass spectrometric analysis of microorganisms. *Anal. Chem.* **73**, 751–757 (2001).
68. Hawkins, T. L., Detter, J. C. & Richardson, P. M. Whole genome amplification — applications and advances. *Curr. Opin. Biotechnol.* **13**, 65–67 (2002).
69. Erwin, D. P. *et al.* Diversity of oxygenase genes from methane- and ammonia-oxidizing bacteria in the Eastern Snake River Plain aquifer. *Appl. Environ. Microbiol.* **71**, 2016–2025 (2005).
70. Detter, J. C. *et al.* Isothermal strand-displacement amplification applications for high-throughput genomics. *Genomics* **80**, 691–698 (2002).
71. Kwon, Y. M. & Cox, M. M. Improved efficacy of whole genome amplification from bacterial cells. *Biotechniques* **37**, 40, 42, 44 (2004).
72. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* (2005).
73. DeLong, E. F. Microbial community genomics in the ocean. *Nature Rev. Microbiol.* **3**, 459–469 (2005).
74. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157–2167 (2002).
75. Holt, R. A. *et al.* The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149 (2002).
76. Allen, E. E. & Banfield, J. F. Community genomics in microbial ecology and evolution. *Nature Rev. Microbiol.* **3**, 489–498 (2005).
77. Koski, L. B. & Golding, G. B. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52**, 540–542 (2001).
78. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glockner, F. O. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163 (2004).
79. McHardy, A. C., Goesmann, A., Puhler, A. & Meyer, F. Development of joint application strategies for two microbial gene finders. *Bioinformatics* **20**, 1622–1631 (2004).
80. Overbeek, R. *et al.* The ERGO genome analysis and discovery system. *Nucleic Acids Res.* **31**, 164–171 (2003).
81. Meyer, F. *et al.* GenDB — an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* **31**, 2187–2195 (2003).
82. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* **31**, 6633–6639 (2003).
83. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3**, REVIEWS0003 (2002). **A provocative discussion of the problems of culture bias and the need for genomic investigation of underrepresented bacterial and archaeal phyla.**
84. Thompson, J. R. *et al.* Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**, 1311–1313 (2005).
85. Spencer, D. H. *et al.* Whole-genome sequence variation among multiple isolates of *Pseudomonas aeruginosa*. *J. Bacteriol.* **185**, 1316–1325 (2003).
86. Rocap, G., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S–23S ribosomal DNA internal transcribed spacer sequences. *Appl. Environ. Microbiol.* **68**, 1180–1191 (2002).
87. Stahl, D. A., Lane, D. J., Olsen, G. J. & Pace, N. R. Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science* **224**, 409–411 (1984).
88. Stahl, D. A., Lane, D. J., Olsen, G. J. & Pace, N. R. Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. *Appl. Environ. Microbiol.* **49**, 1379–1384 (1985).
89. Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63 (1990).
90. Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* **173**, 697–703 (1991).
91. Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.* **59**, 143–169 (1995).
92. Theron, J. & Cloete, T. E. Molecular techniques for determining microbial diversity and community structure in natural environments. *Crit. Rev. Microbiol.* **26**, 37–57 (2000).

Acknowledgements

This work was carried out under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory, Lawrence Berkeley National Laboratory and Los Alamos National Laboratory. S.G.T. was supported by a National Institutes of Health National Research Service Award Training and Fellowship grant. We would like to thank P. Hugenholtz and T. Woyke for helpful comments on the manuscript.

Competing interests statement

The authors declare no competing financial interests.

Online links

FURTHER INFORMATION

ARB: http://www.arb-home.de/
Community Sequencing Program Sequencing Plans for 2005: http://www.jgi.doe.gov/sequencing/cspseqplans.html
Dog Genome Project: http://mendel.berkeley.edu/dog.html
ERGO: http://ergo.integratedgenomics.com/ERGO/login.cgi
EstimateS: http://purl.oclc.org/estimates
GenDB: https://www.cebitec.uni-bielefeld.de/software/gendb/cgi-bin/login.cgi
Joint Genome Institute Protocols web site: http://www.jgi.doe.gov/sequencing/protocols
PRIAM: http://bioinfo.genopole-toulouse.prd.fr/priam
Ribosomal Database Project: http://rdp.cme.msu.edu/index.jsp
***Wolbachia* Genome Project:** http://tools.neb.com/wolbachia/about.html
Access to this interactive links box is free online.