COMMUNITY GENOMICS IN MICROBIAL ECOLOGY AND EVOLUTION

Eric E. Allen and Jillian F. Banfield*‡*

Abstract | It is possible to reconstruct near-complete, and possibly complete, genomes of the dominant members of microbial communities from DNA that is extracted directly from the environment. Genome sequences from environmental samples capture the aggregate characteristics of the strain population from which they were derived. Comparison of the sequence data within and among natural populations can reveal the evolutionary processes that lead to genome diversification and speciation. Community genomic datasets can also enable subsequent gene expression and proteomic studies to determine how resources are invested and functions are distributed among community members. Ultimately, genomics can reveal how individual species and strains contribute to the net activity of the community.

CLONE LIBRARY
A collection of targeted DNA sequences, such as the 16S rRNA gene, most often derived from PCR amplification and subsequent cloning into a vector. Specifically, 16S rRNA gene clone libraries are often used in surveys of microbial diversity from environmental samples.

**Department of Environmental Science, Policy, and Management, ‡Department of Earth and Planetary Sciences, University of California, Berkeley, Berkeley, California 94720, USA. Correspondence to J.F.B. e-mail: jill@eps.berkeley.edu*
doi:10.1038/nrmicro1157

Microbial genomics has, until recently, been confined to individual, isolated microbial strains. Genome sequence information for isolates from phylogenetically diverse lineages has had a marked impact on our understanding of microbial physiology, biochemistry, genetics, ecology and evolution. However, this approach is limited because we do not know how to cultivate most microorganisms[1]. Consequently, many questions about the roles of uncharacterized organisms in natural ecosystems remain.

Our ability to survey the resident microbiota in a given community has been greatly expanded by various cultivation-independent methodologies, which include 16S rRNA gene CLONE LIBRARY collections and group-specific fluorescence *in situ* hybridization (FISH)[2]. Although the description and quantitation of the phylogenetic diversity of microbial communities is an important first step, linking these organisms to their ecological roles remains a significant challenge.

In the natural environment, individual organisms do not exist in isolation. Rather, microbial communities are dynamic CONSORTIA of microbial species populations. The understanding of consortia function will benefit from genomic information from all coexisting members. This cannot be adequately addressed by focused isolation and individual genome sequencing efforts, as isolates might not be representative of the full genetic and metabolic potential of their associated natural populations. Moreover, artificial cultivation conditions often do not replicate those found in nature. Therefore, there is a compelling impetus to move beyond the culture-centric realm of microbial sequencing and to begin focusing sequencing efforts on microbial communities *en masse*.

The analysis of genome sequence data that has been recovered directly from the environment is motivated by many objectives, which include the establishment of gene inventories and natural product discovery[3,4]. This approach is often referred to as metagenomics, which is defined as the functional and sequence-based analysis of the collective microbial genomes that are contained in an environmental sample[3]. Recent reviews have covered environmental and functional metagenomics[3,5–8].

Here we centre our discussion on the opportunities for analysis of ecological and evolutionary processes in natural microbial consortia using environmentally-derived genome sequence data. We

Box 1 | **Acid mine drainage community genomics**

A decade of research on the biogeochemistry[78] and microbiology[79] of the Richmond Mine at Iron Mountain, California, provided the important scientific foundation for the acid mine drainage (AMD) community genome sequencing project. Initial work determined the types of organism that were present and correlated community membership with geochemical conditions[80]. In 2002–2003, 76 Mb of environmental sequence data were obtained from a small-insert library from a single biofilm sample[9]. Using this data, it was possible to reconstruct the genomes of the dominant bacterium, *Leptospirillum* group II (10X coverage) and the dominant archaeon, *Ferroplasma* type II (10X coverage). Partial reconstruction was also possible for the bacterial genomes of *Leptospirillum* group III (3X coverage) and a *Sulfobacillus* species (0.5X coverage) that is closely related to *Sulfobacillus thermosulfidooxidans*. Archaeal genomes that were partially reconstructed include *Ferroplasma acidarmanus* Type I (very closely related to *F. acidarmanus* fer 1; 4X coverage) and 'G-plasma' (3X coverage) — a novel group within the *Thermoplasmatales*.

The sequencing allowed metabolic reconstructions of these organisms based on genome annotations and an analysis of functional partitioning among community members[9]. Importantly, it was revealed that a relatively minor community component, *Leptospirillum* group III, possessed the sole complement of nitrogen fixation (*nif*) genes. This subsequently led to the design of a selective isolation strategy to successfully cultivate this organism using genome sequence data[43]. Furthermore, carbon fixation pathways and gene products that are possibly involved in iron oxidation were revealed, which provided important insights into the intricate metabolism of these chemolithoautotrophic communities.

Genomic analyses also provided insights into population structure. These included evidence for genetic recombination among archaeal populations, which revealed a high degree of genome mosaicism in these species. Furthermore, comprehensive population genomic information has allowed analysis of factors that contribute to genomic heterogeneity within species populations and the ability to assess evidence of selection based on the analysis of nucleotide substitutions (E.E.A. *et al.*, unpublished observations). Finally, the community genomic dataset has provided the foundation for performing environmental proteomic surveys from a natural biofilm sample[10]. These studies have revealed the complement of genes that are expressed *in situ*, and therefore go beyond inferences based on 'genome-annotation gazing' to provide insights into how functions are distributed and which functions are important in natural microbial consortia.

focus on 'community genomics', which emphasizes the analysis of species populations and their interactions, recognizing that both species composition and interactions change over time, and in response to environmental stimuli. This requires that the system under investigation can be sampled repeatedly, and defined well enough to enable *in situ* ecological studies and the analysis of adaptive processes. Genomics can resolve the genetic and metabolic potential of communities and establish how functions are partitioned in and among populations, reveal how genetic diversity is created and maintained, and identify the primary drivers of genome evolution and speciation.

We draw upon experiences from our ongoing analyses of an extreme acid mine drainage (AMD) ecosystem[9,10] (BOX 1). We discuss the challenges that are associated with the assembly of near-complete, and potentially complete, genomes of uncultivated organisms, the documentation of genomic heterogeneity in populations and the use of these data to enable comprehensive functional studies.

Approaches to community genomics

Community genomics provides a platform to assess natural microbial phenomena that include biogeochemical activities, population ecology, evolutionary processes such as lateral gene transfer (LGT) events, and microbial interactions. Only by placing these processes in their environmental context can we begin to understand complex community structure and functions, and the evolutionary constraints that define and sustain them.

Insights into the metabolic functions of uncultivated microorganisms have been facilitated by exploiting phylogenetic anchors that are contained in environmental libraries (BOX 2). For example, in large-insert environmental libraries, contiguous DNA that flanks taxonomic-specific markers such as 16S rRNA genes can provide a glimpse into the genetic potential of sampled organisms[11–15]. Alternatively, random clones from shotgun libraries can be sequenced. In this review, we focus primarily on the shotgun sequencing method, which represents a relatively unbiased, non-directed approach to survey the structure and metabolic capacity of a community.

As a first step, consideration should be given to the community chosen for investigation. On the one hand, simple communities with low species diversity can be characterized thoroughly with modest sequencing effort. On the other hand, complex communities are more representative of most natural microbial assemblages, but their characterization presents myriad challenges that require special consideration. For example, it is necessary to address gaps in knowledge owing to incomplete sequence COVERAGE, and limitations that might arise owing to a lack of reproducibility that results from community heterogeneity.

Currently, both the cost of sequencing and the challenges that are associated with the management of vast datasets precludes comprehensive genomic studies of highly complex communities. Consequently, we favour an initial approach that is based on the analysis of simpler model communities. The technical and

CONSORTIUM
Physical association between cells of two or more types of microorganism. Such an association might be advantageous to at least one of the microorganisms.

COVERAGE
The average number of times a nucleotide is represented by a high-quality base in the sequence data; full genome coverage is usually attained at 8–10X coverage.

Box 2 | **Environmental libraries**

The extraction of high quality DNA is central to the success of any sequencing project. In the case of environmental samples that contain a mixed consortia of organism types, the objective is to obtain a quantitatively accurate representation of all community members during extraction and subsequent construction of shotgun sequencing libraries. Realizing this importance, microbial ecologists have invested substantial effort in optimizing DNA extraction procedures for various environmental samples[15,81].

The advantage of large-insert libraries (for example, ~40 kb for fosmids) is that they provide substantial contiguous genomic information that is representative of individual community members[15,82,83]. For community genome sequencing and assembly, paired-end sequences from large-insert libraries are particularly useful as they provide valuable linking information for orientation and scaffolding of assembled genome fragments. Furthermore, the complete sequences of large-insert clones can be used as reference sequences for the assembly and statistical analysis of environmental shotgun sequence data[22].

Despite the obvious utility of large-insert libraries, certain environments present a considerable challenge to obtaining the high-molecular-weight DNA that is suitable for large-insert library construction. For example, small-insert shotgun libraries (3- to 4-kb insert size) might be the only viable option for the acid mine drainage (AMD) biofilm community from Iron Mountain, as the many steps that are required for DNA purification result in excessive DNA shearing. Nevertheless, small-insert DNA libraries alone have been used in environmental genomic studies with considerable success[9,36].

scientific lessons that have been gleaned from these studies can then be extended to more complex systems and their generality evaluated.

System tractability. Extreme geological environments, such as acidic geothermal hotsprings, highly acidic, or hypersaline habitats, provide important geochemical and selective constraints on species diversity, which makes them ideal for high-resolution studies of microbial ecology and evolution. There are other system attributes that can enhance our ability to learn about the structure of communities and the degree to which they function as integrated synergistic assemblages. These include: first, self-sustaining systems, in which all essential metabolic functions are carried out *in situ* and which therefore represent a complete ecosystem microcosm; second, systems that are characterized by strong and clearly defined geochemical–microbiological feedbacks, which enables analysis of the interplay between organism function and environmental conditions; third, systems that are characterized by systematic fluctuations in environmental conditions, and that can be sampled over space and time, to understand how the community-level metabolic networks change during colonization and as a function of community membership and geochemical conditions; fourth, systems that are defined by well-established species interactions, as expected in extreme environments and other specialized habitats, such as host–pathogen and host–symbiont relationships, in which organisms have co-evolved over extended evolutionary time periods; and finally, systems that have sufficient biomass for post-genomic functional assays (such as proteomic surveys).

ABIOTIC
The non-living physical and chemical attributes of a system, which include pH, temperature, pressure, osmotic strength, and chemical composition.

Sampling and defining the biogeochemical framework. To understand the ecology of a community, it is necessary to describe the associations of organisms with each other and with their environment. Characterization of the ABIOTIC system attributes is important to understand the factors that control community membership. Spatial and temporal environmental heterogeneity is inextricably linked to successional changes in community composition and diversity[16–18]. Therefore, it is important to define physicochemical gradients such as pH, temperature, osmotic strength, mineralogy and nutrient levels, and to identify sources of energy, nutrient fluxes and feedbacks owing to microorganism–mineral interactions. For instance, geochemical patterns[19] might indicate important metabolic functions in the system. In combination with genomic information, the assessment of environmental conditions that contribute to spatial or temporal heterogeneity in species composition might enable the identification of traits that are important to microbial adaptation in the community.

The biological attributes of the system are also an important consideration. For example, the presence of a microbial species might depend upon the presence (or absence) of another species. This might be due to a metabolic dependence and is often suggested to be a phenomenon that limits the success of cultivation endeavours[20]. The interdependence of community members might also take the form of thermodynamic control, such as that observed in microbial consortia that can couple methane oxidation to sulphate reduction[21,22]. Biotic features, such as grazing and phage predation, also impact community structure. Grazing pressure that is imposed by eukaryotic protozoa, such as flagellates and ciliates, is one example of a top–down control[23–25]. Perhaps more important, however, is the well-documented contribution of phage to microbial mortality. The efficacy of phage predation can have profound effects on the composition of microbial assemblages by controlling dominant groups[26,27]. Phage-induced cell lysis can also release cellular contents into the environment, thereby influencing microbial food-web dynamics and biogeochemical processes[28]. Furthermore, the capacity for phage-mediated DNA transfer (transduction), or the direct release of free DNA during virus-induced host-cell lysis[29], can contribute to the overall mobile gene pool in natural communities. Laterally transferred genes and genome fragments can alter the metabolic properties of the host[30] and represent a primary driving factor that contributes to genomic heterogeneity, and therefore evolution, in natural species populations (REF. 31 and E.E.A. *et al.*, unpublished observations).

Estimating the community sequencing endeavour. It is possible to predict the amount of sequencing that will be needed to analyse a given community based on the desired degree of coverage of genomes and the available information about species number, relative species abundance and genome sizes. An approximation of community diversity can be made through the analysis

of 16S rRNA gene libraries, together with a quantitative assessment of relative species richness (number of species) and evenness (relative abundance of each species) using FISH. However, diversity estimates are complicated by PCR bias, *rrn* (ribosomal RNA gene) copy number per genome, and the fact that libraries are rarely sequenced to completion. Genome sizes can be estimated from known sizes of related species, if available, or approximated using the average prokaryotic genome size (~3.16 Mb ± 1.79 Mb; calculated from 215 prokaryotic genomes published in the Genomes Online Database at the time of writing; see the Online links box). Such estimates can prove imperfect, however, owing to marked variation in genome size in a microbial species[32] and the fact that current genome databases are biased towards pathogens and symbionts, which often have reduced genomes. Correlations that have been drawn between the ecology of an organism and its genome size might provide a more refined estimate of genomic complexity for community members[33].

To predict the amount of sequencing that will be required for community coverage, estimates of species richness and the abundance of the dominant organism(s) can be used with statistical methods to describe the species abundance distribution[34]. If the abundance of a given organism is 1%, with a genome size of 3 Mb, then 2.4 Gb of sequence would be required to obtain 8X (near complete) genome coverage of that organism. A sequencing effort of this magnitude would vastly over-sample the genomes of more dominant community members. Therefore, directed strategies to target low abundance organisms may be advantageous (see below).

It is likely that sequencing projections will be imprecise. For example, although species abundance impacts on the relative proportion of DNAs that are present in sequencing libraries, cloning bias might skew species representation. Furthermore, there might be multiple genome types per species. Therefore, predictions should be refined after the assembly of an initial sequencing increment. One simple approach is to use the coverage statistics of the assembly based on a version of the Lander–Waterman equation[35] that is modified to take into consideration the relative abundance of species in the community. If the equation predicts fewer contigs than are observed, the representation of organisms in the library or effective genome sizes can be refined. The prediction should be performed iteratively as more sequence data is analysed. This approach was used to successfully predict the outcome of the community sequencing project undertaken by Tyson *et al.*[9] Specifically, estimates based solely on community characterization by 16S rRNA gene library and quantitative FISH analyses, with an average genome size of ~2 Mb, estimated that ~80 Mb of sequence would be sufficient to cover the five dominant genome types, with individual genome coverages ranging from 0.4 to 30X. Analyses post-assembly of sequencing increments (2, 10, 15 and 25 Mb) revealed that cloning bias in sequencing libraries resulted in the significant overrepresentation of the Archaea, which prompted a reappraisal of actual genome coverages[9].

Community genomics. Perhaps the primary challenge of any community genomic study that aims to extract ecological insights is to correctly assign genome fragments to organism types. In our experience, the weight of this requirement falls most heavily on genome assembly. Various genome assembly programmes are currently available (ARACHNE, CAP, CELERA, EULER, JAZZ, PHRAP and TIGR assemblers, to name but a few). However, the relative efficacy with which most of these programs handle mixed community DNAs has yet to be determined (JAZZ, PHRAP and CELERA have been used previously[9,36]).

Conventional shotgun sequencing of microbial isolates is simplified by the fact that the sequenced clones are derived from organisms with a single genome type. In environmental samples, however, each clone represents a unique sequence that is probably derived from an individual in the community, and the genomes that are sampled come from a pool of both distinct and related genome types. This might pose challenges that currently available genome assembly programs are not designed to deal with. So far, studies have revealed that the genomes of different species have sufficient nucleotide-level sequence divergence (as well as changes in gene order) to prevent co-assembly[9,36]. Hurdles do arise, however, owing to genomic variation within species populations.

The resolution of strain-level differences is a fundamental goal of community genomic analysis (FIG. 1). Although many comparative genomic studies of strain variants indicate a highly conserved gene order[37–39], extensive genome rearrangements in members of the same species[40,41] will confuse genome assembly and can preclude the assembly of environmental shotgun sequence data[22]. In the AMD community, genome rearrangements that involve more than two genes were extremely rare in archaeal populations, and breakdown of conserved SYNTENY occurs primarily after species divergence (E.E.A. *et al.*, unpublished observations). In regions where single-nucleotide polymorphisms are the predominant form of genomic heterogeneity, it is possible to define composite species genome sequences (that is, an aggregate sequence comprised of multiple strain sequence types). However, assembly is problematic in regions where members of the strain population have different gene contents (FIG. 2).

It is important to develop mixed genome assembly methods to deal with differences in gene content and gene sequence, because these phenomena can artificially terminate SCAFFOLDS or separate sequencing reads into multiple scaffolds at regions of strain genome 'confusion'. This results in separate, but homologous, DNA fragments that can be mapped onto the composite species genome dataset (FIG. 1). Other complications owing to strain assembly include inaccurate (over)estimation of genome sizes and artificial duplication of open reading frames (ORFs) in community genome datasets. If assembly heuristics can overcome complications owing

SYNTENY
Refers to the presence of two or more genes on the same chromosome. However, the term is often used to refer to the shared colinearity in orthologous gene content and gene order between genomes.

SCAFFOLD
A genome fragment constructed by the ordering and orienting of sets of unlinked contigs generated from raw shotgun sequence data by using additional information (such as paired-end sequence information or homology data) to determine proper contig linkage and placement along the chromosome. Scaffolds can be comprised of multiple contigs.

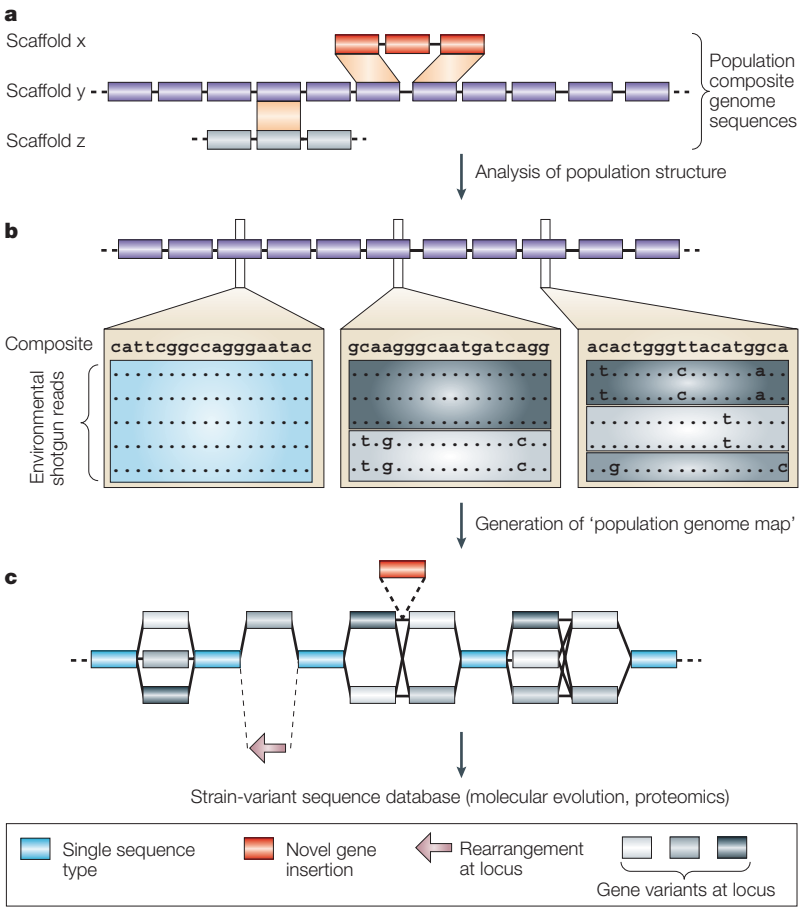

Figure 1 | **Resolving strain-level heterogeneity.** Schematic depicting the forms of genomic heterogeneity that lead to separation of genomic regions during co-assembly of sequences from strain populations. **a** | Alignment of three composite genome fragments for the same genomic region. Two small scaffolds were separated during genome assembly because of sequence divergence (shown in grey) and differences in gene content (shown in red). Shading between sequences indicates sequences that are homologous. **b** | Single-nucleotide polymorphism patterns in environmentally-derived sequencing reads contribute to the composite sequence and reveal distinct variants, which are indicative of strain heterogeneity. **c** | Quantitation of the form and distribution of strain sequence types at each locus allows for the generation of a 'population genome map'. Some loci have a single sequence type (shown in blue) whereas other loci have multiple strain sequence types (shades of gray). The map also incorporates information about gene insertions (shown in red) and gene rearrangements (pink arrow). Compiling strain-variant sequence databases should enable analyses of molecular evolution to be carried out and detection of strain-level expression patterns through microarray analyses and proteomic surveys.

to strain-level heterogeneity, it is probable that complete (closed) composite genomes can be reconstructed for uncultivated organisms.

Once large scaffolds are generated, they can be assigned to the correct organism ('binned') based on various parameters, which include: the phylogeny of conserved marker genes, such as the 16S rRNA gene, *recA*/*radA*, *rpoB*/*C*, Hsp60/70, EF-Tu, and aIF2β; % GC content; depth of sequence coverage (number of reads per unit length of DNA); codon usage; and di-, tri-, and tetra-nucleotide frequencies[42]. Moreover, the alignment of scaffolds with available sequenced genomes as a reference also provides a means to categorically assign assembled fragments. The effectiveness of these methods increases with increasing scaffold size.

For the AMD community, >95% of the individual genomes of all five dominant members were reconstructed[9]. This allowed the inference of metabolic maps for these organisms and revealed species that might have key roles in community-essential tasks such as nitrogen fixation and biofilm polymer production. As only one organism, *Leptospirillum* group III, has genes that are required for nitrogen fixation, it was inferred that this relatively low abundance community member is a keystone species in the AMD system owing to a lack of a significant input of externally derived nitrogen[9]. The analysis of the functional partitioning of essential roles of community members has led to the successful isolation of this previously uncultivated organism[43]. These results highlight the utility of genomic information for the design of cultivation strategies that are based on the presence, absence or distribution of metabolic capacities in a community.

The dissection of composite genomes provides a wealth of information about the degree of genetic variability that exists in a species population. In the AMD study, bacterial representatives of the genus *Leptospirillum* were essentially clonal. By contrast, the archaeal species populations showed significant strain heterogeneity, both at the level of sequence divergence and in terms of gene content. This variation often results from insertions of probable phage origin and transposable-element dynamics (E.E.A. *et al.*, unpublished observations). Some genome regions represent 'hot spots', where genes have been gained and lost in coexisting species and strains (FIG. 2).

Multilocus sequence typing of environmental isolates has provided evidence for recombination in natural populations[44,45]. However, because these methods are based only on the subset of organisms that can be isolated, they cannot comprehensively evaluate the importance of recombination in natural environments. Comparison among DNA fragments for specific regions in the genomes of archaeal species in the AMD system revealed a mosaic genome structure that is inferred to have arisen through homologous recombination between closely related strain variants. More recently, the PANMICTIC character of natural archaeal *Halorubrum* populations has also been reported[45]. The maintenance of populations with genomes that are combinatorial mosaics on the basis of subtly different sequence types might be a strategy for the fine-tuning of environmental adaptation, and might provide a cohesive mechanism that prevents species divergence. Loss of the ability of members of a population to undergo genetic exchange might initiate irreversible species divergence[46] and has been proposed as a method for the definition of a species boundary[9]. Community genomics can provide the opportunity to assess the mechanisms that contribute to genomic heterogeneity in a species population, thereby allowing the evaluation of processes that can lead to speciation.

Assembly of environmental shotgun reads by mapping to the finished sequence of a related strain is a rapid and efficient means to reconstitute the environmental genome type. Using the genome sequence

PANMICTIC
Characterized by a lack of restriction in genetic exchange within the population; all individuals within the species population are potential recombination partners.

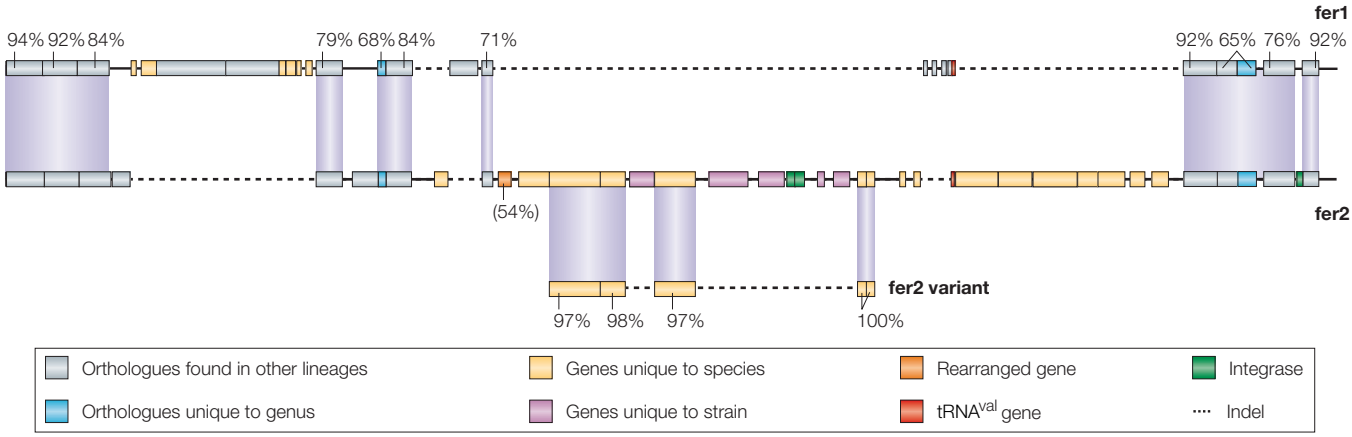

Figure 2 | **Genomic heterogeneity in *Ferroplasma* species and strains**. Comparison of a genomic region between *Ferroplasma acidarmanus* fer1 and two *Ferroplasma* Type II (fer2) strains, which shows an example of the distribution of heterogeneity at the species and strain level. Orthologues (grey, linked by purple shading) are shown with the corresponding % amino-acid identity. Some genes do not have orthologues in the most closely related species, but do have orthologues in other lineages, which might indicate species-specific gene loss at these loci (grey without purple shading). Other orthologues are specific to the genus *Ferroplasma* (blue). Differences in gene content in the fer2 variant result in assembly of a separate genomic fragment. The region shown contains numerous species-specific genes (yellow) and strain-specific genes (pink) and evidently represents a 'hot spot' for gene gain and loss. Strain-specific genes are of interest as they probably represent recently acquired genes, possibly associated with integrated elements such as phage. In this region, two phage integrases (green) and a gene rearrangement (orange) localize in the vicinity of the large fer2 insertion. Indel refers to insertions and deletions.

of the archaeon *Ferroplasma acidarmanus* fer1, we have reconstructed >90% of the environmental strain composite genome (E.E.A. *et al.*, unpublished observations). As more isolate genome sequences are completed, comparative genome assembly will find increased utility in community genomic ventures. Moreover, as the genome of one isolate incompletely defines the genomic potential of a species, it makes sense that future isolate genome sequencing efforts should be coupled to the sequencing of the associated environmental populations.

Genomics-enabled functional analyses

The recovery of genes and genomes from environmental samples is the first step towards determining how organisms function in a consortium and the mechanisms by which community-essential tasks are partitioned among organisms. Genome sequence information can be used to design probes for gene expression profiling and to generate the databases that are required to identify proteins in proteomic surveys. This allows functional analyses to be carried out directly on natural samples from the field. *In situ* studies are an important goal if we are to understand regulatory networks, functional organization and community dynamics in their true ecological context (FIG. 3). Such studies would be further enhanced if changes in gene expression in response to perturbation could be monitored *in situ*.

Community microarray analysis. Applying community genomic data to DNA microarrays allows the analysis of global gene expression patterns and regulatory networks in a rapid, parallel format. Microarray technologies have been used to monitor gene expression[47], assess functional gene diversity[48], and to screen metagenomic libraries[49] from environmental samples. Community microarray analyses can uncover apparent linkages between different genes and gene families and the distribution of metabolic functions in the community. However, as the utility of microarrays is markedly reliant on the identity between the probe sequence (derived from the community genome dataset) and the environmentally-derived target (RNA or DNA), sequence divergence (or similarity) in the community might bias the interpretation of results. Similarly, if genomic heterogeneity in species populations is not assessed prior to array fabrication, this might prevent the detection of variant signals. If strain-resolved community genomic data can be obtained, microarrays that incorporate strain variants could be used to monitor population structure dynamics in the environment.

Community proteomics. Mass-spectroscopy-based proteomic methods are rapid and sensitive means to identify proteins in complex mixtures[50]. When applied to environmental samples, 'shotgun' proteomic analyses can produce surveys of prevalent protein species, which allows inferences of biological origin and metabolic function[51,52]. However, because peptides are assigned to proteins on the basis of similarity to database sequences, the environmental proteomic fingerprints are limited by the completeness of the database[53].

In a study that highlighted the potential of integrated environmental genomics and proteomics, Ram *et al.*[10] characterized the protein complement of a natural AMD biofilm utilizing genome sequence data from a closely related community[9]. 17% of the proteins encoded by the genomes of the five dominant members were detected, which enabled evaluation of functional

partitioning. Furthermore, products were identified for 48% of the genes from the dominant community member, which allowed assessment of the relative metabolic activity of this organism on the basis of the functional classification of proteins. However, even if it were possible to detect all of the relatively abundant proteins in the community, biological insights are restricted because we do not know the function of a significant fraction (~31% of the detected proteins in the biofilm studied by Ram *et al.*[10]) Although the magnitude of this problem is immense, a solution must be found if we are to approach the desired level of understanding of microorganisms in their environments. Initiatives that are directed at improved genome annotation and experimental elucidation of hypothetical gene function hold promise in this regard[54].

Strain heterogeneity. Strain heterogeneity complicates environmental proteomic analyses because a single amino-acid substitution is sufficient to prevent matching of a peptide to a protein database entry if stringent identification criteria are used. This limits the utility of using community genomic data that have been derived from one sample for the proteomic analysis of other samples that differ in strain composition. For example, this might restrict the use of genome sequences from isolates for characterization of environmental samples. Integrated strain-resolved community genomic and proteomic approaches largely remedy this problem by providing a consistent database for protein identification, and these approaches can begin to reveal how metabolic functions are distributed among community members *in situ*.

It is important to address the problem of strain heterogeneity to evaluate the roles of distinct protein variants or strain-specific laterally acquired genes, because these might be important in species adaptation and in the emergence of new species. Consequently, it is important to understand how these variants arise (for example, by mechanisms such as phage lysogeny, interspecific lateral-gene transfer or by mutation) and the factors that maintain heterogeneity in natural populations. Our preliminary analyses indicate that virtually all gene variants in one of the AMD archaeal species populations are under strong stabilizing selection. It has been inferred that the mosaic genome pool is subjected to selective sweeps that generally result in a limited number of gene variants at any one locus[9] (E.E.A. *et al.*, unpublished observations). Therefore, it seems that modifications in protein sequence either become dominant or are removed (or reduced to low abundance) in the population, which implies that even subtle protein variants are under strong selective pressures. As selection acts to distinguish strain-level heterogeneity, an understanding of these communities requires documentation of activity at the strain (rather than species) level. Consequently, it is crucial that we develop methods to generate strain-variant databases to enable population genetic studies of selection and to enhance post-genomic functional analyses (FIG. 1).

Community genomics and complex systems

The extent of microbial diversity — while still a matter of debate[55] — is undeniably immense. In complex systems, datasets with little or no sequence assembly can provide functional insights. Habitat-specific metabolic demands can be inferred on the basis of the identification of predominant gene families (functional inventories). For example, Schmeisser *et al.*[56] surveyed the metagenome of a model drinking-water biofilm community by shallow sequencing of clones from a small-insert library and by analysing the protein-coding information. Tringe *et al.*[57] analysed a soil environment and three deep-ocean whale skeletons to highlight metabolic 'fingerprints' that were specific to each environment. However, such fingerprinting is probably limited by our current inability to assign functional roles to a large fraction of the predicted proteins, many of which are lineage- and environment-specific.

In environmental shotgun sequencing projects, high abundance organisms will be greatly over-sampled to obtain adequate coverage of lower abundance organism types. For example, extensive shotgun sequencing of the Sargasso Sea[36] produced ~1.36 Gb of sequence data, estimated to derive from >1,000 genomic species. Over 1.2 million gene products that were encoded both on assembled fragments and on individual reads provided information about function and also provided a basis for the estimation of diversity. Despite this massive effort, near-complete genome assembly was only possible for a few organisms, with the dominant organism sampled, *Burkholderia* spp., represented at ~21X coverage.

If adequate funding, sequencing and data management resources are available, it is conceivable to apply shotgun-sequencing community genomic approaches to complex natural systems such as soils, where estimates range from 10^3 to $>10^4$ species per g[1,58], and to obtain significant genomic reconstructions. Furthermore, procedures to normalize DNA content in libraries might provide a way to obtain a more balanced genomic representation of the community by removing over-sampled organisms prior to sequencing. Alternatively, 'selective-isolation genome sequencing' approaches that are aimed at targeted species populations in a community can provide important insights into the ecology and evolution of low abundance microorganisms. Such approaches involve the enrichment or isolation of target organisms from complex consortia prior to sequencing. Fluorescent tagging using specific molecular recognition probes (for example 16S rRNA genes) combined with flow-cytometric separation and fluorescence-activated cell sorting can concentrate and isolate a target species from its associated assemblage[59]. Once collected, total DNA can be recovered and sequenced to provide a genomic snapshot of a natural species population in a cultivation-independent manner. Selective-capture genome sequencing might also target individual microorganisms that have specific functional attributes, by the FISH-based labelling

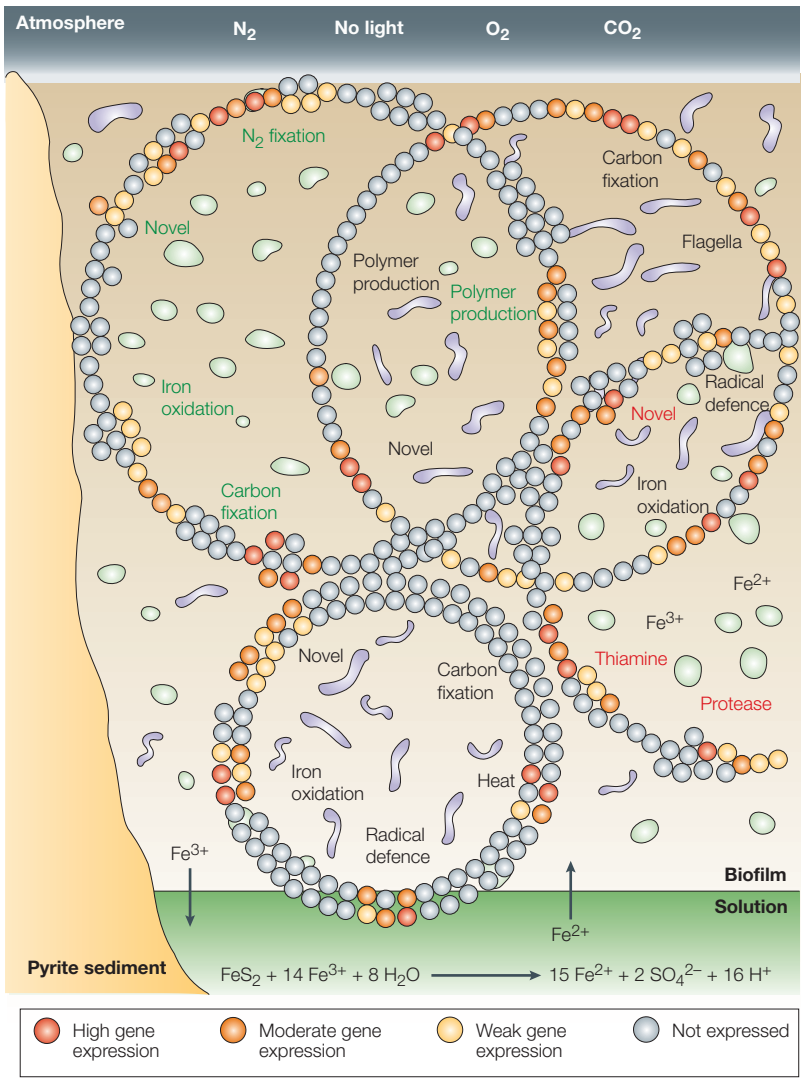

Figure 3 | **Integrating community genomics and functional assays *in situ*.** This schematic illustrates the genetic potential of four populations that comprise a simple community. A snapshot of gene expression *in situ* as revealed by community genome-enabled proteomic analysis is shown. The genome of each organism is shown as a ring comprised of open reading frames (ORFs), which are represented as coloured spheres. The colour of each sphere indicates the level of gene expression (for example, red represents a high level of gene expression).The ring structure represents population structure — a single ring indicates a clonal species population and a double ring indicates a species with two dominant clonal variants. Rings with 1–3 variants per ORF indicate a mosaic genome structure — combinatorial variants formed by intra-species recombination — which seems to be typical of some archaeal populations[9,45]. These integrated approaches can begin to address strain-level gene expression patterns and to reveal how functions are distributed among community members. Included is the pyrite dissolution reaction that defines the interplay between metabolism, minerals and solution chemistry. The goal of community genomic research is to render this cartoon into a real model that addresses the main physical, chemical and biological parameters that define the ecosystem and its dynamics.

of functional gene transcripts (mRNA FISH[60] or RING-FISH[61]). Such analyses can be applied to microorganisms whose metabolism is beneficial in the bioremediation of environmental contamination, or to microorganisms with genes that contribute to major elemental cycling processes, such as denitrification (*napA* or *narG*), oxidation of ammonia (*amo*), or C1 metabolism such as anaerobic methane oxidation (*mcrA*). Technologies that are capable of high-level, representational amplification of genomic DNA (for example, through strand displacement amplification using bacteriophage phi29 DNA polymerase[62,63]) can also be exploited to enable genomic analysis of minor community members for which source DNA is scarce.

Microbial ecology and evolution

It is difficult to determine how an organism's behaviour is modified as the result of cooperative and competitive interactions without the ability to survey the metabolic activity of all community members simultaneously. The formation of microbial consortia presumably results in an optimization that is not achieved through solitary existence, and might provide a high degree of plasticity to respond to environmental perturbation. It is now possible to design experiments that integrate genomics, gene expression and proteomics in an environmental context to determine how roles are distributed among different members of populations and communities. Using this approach, we might be able to resolve how an individual functions as a member of a community.

A related question is the degree to which inter-organism communication influences microbial activity. Although signalling between some specific organism types is well characterized[64], many questions about cooperative interspecies signalling and antagonistic interactions such as antibiotic production, remain. The combination of metabolomics — the profiling of metabolites, including small molecules[65] — with analyses of gene expression and community membership might begin to unravel the nature of organism–organism interactions in communities.

Using community genomic tools, it is now possible to begin to determine the dynamic interplay between organisms and their environment, particularly for relatively simple systems. This may be achieved by correlating microbial activity with an environmental event, such as environmental acidification, thermal oscillation or anaerobicity, and by monitoring changes in metabolism and community membership that result from environmental change.

The relationship between area and species number, which is well-established in the ecology of plant and animal communities, also applies to microorganisms[16,66]. Genomic sampling across environmental gradients might reveal clues about the differences in fitness that account for this relationship. Studying genomic heterogeneity from the micrometer scale to the macro scale will show us how spatial diversity is generated and maintained in the natural environment.

Comparative community genomic analysis in a spatio-temporal context can paint a dynamic portrait of the forces that shape community diversity and stability. We are just beginning to learn about the degree of heterogeneity in gene content in natural populations, the rate of gain or loss of novel genes and the significance of these processes. Determination of the form, distribution and tempo of genomic variation in populations

might point to processes that lead to diversification and speciation. For example, metabolic functions that contribute to or deter genetic-exchange mechanisms, such as recombination, may be particularly important in microbial diversification. Genomic and proteomic analyses may be able to document the presence and activity of such functions, including mismatch repair[46] or restriction-modification[67] systems, which have been shown to influence homologous recombination and, possibly, speciation.

An important question is what fraction of genetic variability is due to drift and what fraction is due to selection? Modelling the ratio of non-synonymous to synonymous changes (K_a/K_s) for all genes in and among species populations is one route to answering this question[68,69]. Moreover, a host of molecular evolutionary analyses based on comparative sequence analysis can be applied to such datasets[70].

Comparative genomics of coexisting, closely related organisms will enable identification of genes that were acquired after speciation or strain divergence, and might enable identification of the sources of these genes. The importance of phage as sources of novel genes that differentiate species and strains has been documented previously[40,71] and the significance of phage in microbial diversification and evolution has been suggested[31,72]. It should be possible to quantify the number of genes that have been derived from lateral transfer that contribute to population genomic heterogeneity and to evaluate the distribution and persistence of these genes over time. These data will be more informative when interpreted in conjunction with complementary community genomic analyses, such as description of the associated phage populations from the same systems. Proteomic analyses can further clarify the extent to which genes of putative phage origin are expressed, thereby delineating the relative significance of phage conversion to the organism.

To arrive at an integrated view of microbial ecology and evolution, genomic studies must be complemented with field-level observations. For example, biogeophysical methods can be used to analyse microenvironmental geometry (such as pore space), physical connectivity and transport processes[73]. This may be augmented by methods to analyse substrate utilization in the environment, such as stable isotope probing[74–76] or FISH-based microautoradiographic assays[77]. High-resolution geochemical measurements can reveal microbial activity levels and define spatial heterogeneity. Geochemical measurements can also provide critical time constraints for geological systems. This might allow the estimation of absolute rates of evolution by enabling correlation between environmental and genome change.

To this end, community genomics will undoubtedly establish itself as an integral component of ecological and evolutionary studies as we strive to enrich our understanding of the microbial processes that sustain our biosphere.

1. Torsvik, V., Gokoyr, J. & Daae, F. L. High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* **56**, 782–787 (1990).
2. Ammann, R. R., Ludwig, W. & Schleifer, K. H. Phlyogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.* **59**, 143–169 (1995).
3. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**, 669–685 (2004).
4. Cowan, D. A. *et al.* Metagenomics, gene discovery, and the ideal biocatalyst. *Biochem. Soc. Trans.* **32**, 298–302 (2004).
5. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**, 525–552 (2004).
6. Streit, W. R. & Schmitz, R. A. Metagenomics — the key to the uncultured microbes. *Curr. Opin. Microbiol.* **7**, 492–498 (2004).
7. Eyers, L. *et al.* Environmental genomics: exploring the unmined richness of microbes to degrade xenobiotics. *Appl. Microbiol. Biotechnol.* **66**, 123–130 (2004).
8. Handelsman, J. Sorting out metagenomes. *Nature Biotechnol.* **23**, 38–39 (2005).
9. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
This report describes the first near-complete reconstruction of uncultivated microbial genomes using shotgun sequencing of a natural microbial community.
10. Ram, R. J. *et al.* Community proteomics of a natural microbial biofilm. *Science* (in the press).
This paper reports the first 'shotgun' proteomic investigation of a natural microbial community performed in conjunction with community genome sequence data from the same location.
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 genomic fragment from a planktonic marine archaeon. *J. Bacteriol.* **178**, 591–599 (1996).
12. Beja, O. *et al.* Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902–1906 (2000).
13. Quaiser, A. *et al.* First insight into the genome of an uncultivated crenarchaeote in soil. *Env. Microbiol.* **4**, 603–611 (2002).
14. 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).
15. Treusch, A. H. *et al.* Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. *Environ. Microbiol.* **6**, 970–980 (2004).
16. Horner-Devine, M. C., Carney, K. M. & Bohannan, B. J. M. An ecological perspective on bacterial biodiversity. *Proc. Biol. Sci.* **271**, 113–122 (2003).
17. Kassen, R. The experimental evolution of specialists, generalists, and the maintenance of diversity. *J. Evol. Biol.* **15**, 173–190 (2002).
18. Rosenzweig, M. L. *Species Diversity in Space and Time* (Cambridge University Press, Cambridge, 1995).
19. Shock, E. L., McCollom, T. & Schulte, M. D. Geochemical constraints on chemolithoautotrophic reactions in hydrothermal systems. *Orig. Life Evol. Biosph.* **25**, 141–159 (1995).
20. Lorenz, P. & Schleper, C. Metagenome — a challenging source of enzyme discovery. *J. Mol. Catalysis B: Enzymatics* **19**, 13–19 (2002).
21. Hoehler, T. M. & Alperin, M. J. in *Microbial Growth on C1 Compounds* (eds Lindstrom, M. E. & Tabita, F. R.) 326–333 (Kluwer Academic, Dordrecht, 1996).
22. Hallam, S. J. *et al.* Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science* **304**, 1457–1462 (2004).
23. Huws, S. A., McBain, A. J. & Gilbert, P. Protozoan grazing and its impact upon population dynamics in biofilm communities. *J. Appl. Microbiol.* **98**, 238–244 (2005).
24. Kiorboe, T., Tang, K., Grossart, H. P. & Ploug, H. Dynamics of microbial communities on marine snow aggregates: colonization, growth, detachment, and grazing mortality of attached bacteria. *Appl. Environ. Microbiol.* **69**, 3036–3047 (2003).
25. Boenigk, J., Stadler, P., Wiedlroither, A. & Hahn, M. W. Strain-specific differences in the grazing sensitivities of closely related ultramicrobacteria affiliated with the *Polynucleobacter* cluster. *Appl. Environ. Microbiol.* **70**, 5787–5793 (2004).
26. Thingstad, T. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45**, 1320–1328 (2000).
27. Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000).
28. Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181 (2004).
29. Jiang, S. C. & Paul, J. H. Viral contribution to dissolved DNA in the marine environment as determined by differential centrifugation and kingdom probing. *Appl. Environ. Microbiol.* **61**, 317–325 (1995).
30. Cheetham, B. & Katz, M. A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. *Mol. Microbiol.* **18**, 201–208 (1995).
31. Weinbauer, M. G. & Rassoulzadegan, F. Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* **6**, 1–11 (2004).
32. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl Acad. Sci. USA* **102**, 2567–2572 (2005).
This report provides a thorough comparative analysis of 70 microbial genomes that highlight the extent of genomic variability that exists within and between microbial species.
33. Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl Acad. Sci. USA* **101**, 3160–3165 (2004).
34. Hughes, J. B., Hellmann, J. J., Ricketts, T. H. & Bohannan, B. J. M. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* **67**, 4399–4406 (2001).
This review describes and evaluates the utility of statistical approaches in assessing microbial diversity in natural communities.

35. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).
36. Venter, J. C. *et al.* Environmental shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
37. Tamas, I. *et al.* 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376–2379 (2002).
38. Nelson, K. E. *et al.* Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species. *Nucleic Acids Res.* **32**, 2386–2395 (2004).
39. Bolotin, A. *et al.* Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. *Nature Biotechnol.* **22**, 1554–1558 (2004).
40. Deng, W. *et al.* Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.* **184**, 4601–4611 (2002).
41. Wu, M. *et al.* Phylogenomics of the reproductive parasite *Wolbachia pipientis* mMel: a streamlined genome overrun by mobile genetic elements. *PLoS Biol.* **2**, e69 (2004).
42. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glockner, F. O. TETRA: a web-service and stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequence. *BMC Bioinformatics* **5**, 163 (2004).
43. Tyson, G. W. *et al.* Genome-directed isolation of the key nitrogen fixer, *Leptospirillum ferrodiazotrophum* sp. nov., from an acidophilic microbial community. *Appl. Environ. Microbiol.* (in the press).
44. de Las Rivas, B., Marcobal, A. & Munoz, R. Allelic diversity and population structure in *Oenococcus oeni* as determined from sequence analysis of housekeeping genes. *Appl. Environ. Microbiol.* **70**, 7210–7219 (2004).
45. Papke, R. T., Koenig, J. E., Rodriguez-Valera, F. & Doolittle, W. F. Frequent recombination in a saltern population of *Halorubrum*. *Science* **306**, 1928–1929 (2004).
46. Vulic, M., Lenski, R. E. & Radman, M. Mutation, recombination, and incipient speciation of bacteria in the laboratory. *Proc. Natl Acad. Sci. USA* **96**, 7348–7351 (1999).
47. Dennis, P., Edwards, E. A., Liss, S. N. & Fulthorpe, R. Monitoring gene expression in mixed microbial communities by using DNA microarrays. *Appl. Environ. Microbiol.* **69**, 769–778 (2003).
48. Wu, L. *et al.* Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.* **67**, 5780–5790 (2001).
49. Sebat, J. L., Colwell, F. S. & Crawford, R. L. Metagenomic profiling: microarray analysis of an environmental genomic library. *Appl. Environ. Microbiol.* **69**, 4927–4934 (2003).
50. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
51. Schulze, W. X. *et al.* A proteomic fingerprint of dissolved organic carbon and of soil particles. *Oecologia* **142**, 335–343 (2005).
52. Powell, M. J., Sutton, J. N., Del Castillo, C. E. & Timperman, A. T. Marine proteomics: generation of sequence tags for dissolved proteins in seawater using tandem mass spectrometry. *Marine Chem.* (in the press).
53. Habermann, B., Oegema, J., Sunyaev, S. & Shevchenko, A. The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. *Mol. Cell. Proteomics* **3**, 238–249 (2004).
54. Roberts, R. J., Karp, P., Kasif, S., Linn, S. & Buckley, M. R. *An Experimental Approach to Genome Annotation.* Critical Issues Colloquia Report, Washington DC, USA: American Academy of Microbiology (Jan 2005).
55. Curtis, T. P. & Sloan, W. T. Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. *Curr. Opin. Microbiol.* **7**, 221–226 (2004).
56. Schmeisser, C. *et al.* Metagenome survey of biofilms in drinking-water networks. *Appl. Environ. Microbiol.* **69**, 7298–7309 (2003).
57. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).
58. Curtis, T. P., Sloan, W. & Scannell, J. Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci. USA* **99**, 10494–10499 (2002).
59. Wallner, G., Fuchs, B., Spring, S., Beisker, W. & Amann, R. Flow sorting of microorganisms for molecular analysis. *Appl. Enivron. Microbiol.* **63**, 4223–4231 (1997).
60. Pernthaler, A. & Amann, R. Simultaneous fluorescence *in situ* hybridization of mRNA and rRNA in environmental bacteria. *Appl. Environ. Microbiol.* **70**, 5426–5433 (2004).
61. Zwirglmaier, K., Ludwig, W. & Schleifer, K. H. Recognition of individual genes in a single bacterial cell by fluorescence *in situ* hybridization — RING-FISH. *Mol. Microbiol.* **51**, 89–96 (2004).
62. Lizardi P. M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nature Genet.* **19**, 225–232 (1998).
63. Gadkar, V. & Rillig M. C. Application of Phi29 DNA polymerase mediated whole genome amplification on single spores of arbuscular mycorrhizal (AM) fungi. *FEMS Microbiol. Lett.* **242**, 65–71 (2005).
64. Henke, J. M. & Bassler, B. L. Bacterial social engagements. *Trends Cell Biol.* **14**, 648–656 (2004).
65. Kell, D. B. Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* **7**, 296–307 (2004).
66. Horner-Devine, M. C., Lage, M., Hughes, J. B. & Bohannan, B. J. M. A taxa–area relationship for bacteria. *Nature* **432**, 750–753 (2004).
67. Jeltsch, A. Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? *Gene* **317**, 13–16 (2003).
68. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–26 (1986).
69. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–502 (2000).
70. Liberles, D. A. & Wayne, M. L. Tracking adaptive evolutionary events in genomic sequences. *Genome Biol.* **3**, 1018 (2002).
71. Wick, L. M., Qi, W., Lacher D. W. & Whittam, T. S. Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7. *J. Bacteriol.* **187**, 1783–1791 (2005).
72. Ohnishi, M., Kurokawa, K. & Hayashi, T. Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? *Trends Microbiol.* **9**, 481–485 (2001).
73. Scheibe, T. D., Chien, Y. J., & Radtke, J. S. Use of quantitative models to design microbial transport experiments in a sandy aquifer. *Ground Water* **39**, 210–222 (2001).
74. Orphan, V. J., House, C. H., Hinrichs, K.-U., McKeegan, K. D. & DeLong, E. F. Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis. *Science* **293**, 484–487 (2001).
75. Radajewski, S. *et al.* Identification of active methylotroph populations in an acidic forest soil by stable-isotope probing. *Microbiol.* **148**, 2331–2342 (2002).
76. Wellington, E. M., Berry, A. & Krsek, M. Resolving functional diversity in relation to microbial community structure in soil: exploiting genomics and stable isotope probing. *Curr. Opin. Microbiol.* **6**, 295–301 (2003).
77. Ouverney, C. C. & Fuhrman, J. A. Combined microautoradiography-16S rRNA probe technique for determination of radioisotope uptake by specific microbial cell types *in situ*. *Appl. Environ. Microbiol.* **65**, 1746–1752 (1999).
78. Druschel, G. K., Baker, B. J., Gihring, T. H. & Banfield, J. F. Acid mine drainage biogeochemistry at Iron Mountain, California. *Geochem. Trans.* **5**, 13–32 (2004).
79. Baker, B. J. & Banfield, J. F. Microbial communities in acid mine drainage. *FEMS Microbiol. Ecol.* **44**, 139–152 (2003).
80. Bond, P. L., Dreuschel, G. K. & Banfield, J. F. Comparison of acid mine drainage microbial communities in physically and geochemically distinct ecosystems. *Appl. Environ. Microbiol.* **66**, 4962–4971 (2000).
81. Miller, D. N., Bryant, J. E., Madsen, E. L. & Ghiorse, W. C. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl. Environ. Microbiol.* **65**, 4715–4724 (1999).
82. DeLong, E. F. Microbial population genomics and ecology. *Curr. Opin. Microbiol.* **5**, 520–524 (2002).
83. Rondon, M. R., *et al.* Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**, 2514–2547 (2000).

Acknowledgements

We thank G. W. Tyson and anonymous reviewers for helpful comments. Support for our work from the Department of Energy Microbial Genome Program, National Science Foundation (NSF) Biocomplexity Program, NASA Astrobiology Institute, and the NSF Postdoctoral Research Fellowship Program in Microbial Biology (E.E.A.) is gratefully acknowledged.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

Ferroplasma acidarmanus fer1 | *Leptospirillum* group III

FURTHER INFORMATION

Jillian Banfield's laboratory:
http://seismo.berkeley.edu/~jill/banfield.html

Genomes Online Database: http://www.els.net/

Access to this interactive links box is free online.